UNIVERSITÀ DEGLI STUDI DI PADOVA

DOCTORAL THESIS

# Natural Language Processing for Technology Foresight

## Summarization and Simplification: the case of patents

*Author:*
Silvia CASOLA

*Supervisor:*
Alberto LAVELLI

*Co-supervisor:*
*Sabrina* CIPOLLETTA

*A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy
in Brain, Mind and Computer Science*

May 30, 2023

UNIVERSITÀ DEGLI STUDI DI PADOVA

# *Abstract*

**Natural Language Processing for Technology Foresight**

by Silvia CASOLA

Technology foresight aims to anticipate possible developments, understand trends, and identify technologies of high impact. To this end, monitoring emerging technologies is crucial. Patents – the legal documents that protect novel inventions – can be a valuable source for technology monitoring.

Millions of patent applications are filed yearly, with 3.4 million applications in 2021 only. Patent documents are primarily textual documents and disclose innovative and potentially valuable inventions. However, their processing is currently underresearched. This is due to several reasons, including the high document complexity: patents are very lengthy and are written in an extremely hard-to-read language, which is a mix of technical and legal jargon.

This thesis explores how Natural Language Processing – the discipline that enables machines to process human language automatically – can aid patent processing. Specifically, we focus on two tasks: patent summarization (i.e., we try to reduce the document length while preserving its core content) and patent simplification (i.e., we try to reduce the document's linguistic complexity while preserving its original core meaning).

We found that older patent summarization approaches were not compared on shared benchmarks (making thus it hard to draw conclusions), and even the most recent abstractive dataset presents important issues that might make comparisons meaningless. We try to fill both gaps: we first document the issues related to the BigPatent dataset and then benchmark extractive, abstraction, and hybrid approaches in the patent domain. We also explore transferring summarization methods from the scientific paper domain with limited success.

For the automatic text simplification task, we noticed a lack of simplified text and parallel corpora. We fill this gap by defining a method to generate a silver standard for patent simplification automatically. Lay human judges evaluated the simplified sentences in the corpus as grammatical, adequate, and simpler, and we show that they can be used to train a state-of-the-art simplification model.

This thesis describes the first steps toward Natural Language Processing-aided patent summarization and simplification. We hope it will encourage more research on the topic, opening doors for a productive dialog between NLP researchers and domain experts.

# *Acknowledgements*

I'm not good at thank yous. I worry to sound too mellifluous, and I end up speaking like a bad-programmed machine. So, this page will only contain a clumsy attempt at acknowledging and tanking all the people that were there for me in these three years.

First of all, I want to thank my advisor, Alberto Lavelli. In these three years, he has been constantly motivating me with his uncommon patience in reading my three-page-long confusing emails, his eagle eye for typos (sorry), and his passion for clean bibliographies. Most importantly, he has always been there for me, to give a good piece of advice, talk ideas, or just listen while keeping my feet on the ground. Nevertheless, he has still given me the freedom to explore, learn, experiment, and fail, which I am very grateful for.

I also wanted to thank the whole Fondazione Bruno Kessler, where I conducted most of my work. Thanks to the whole Natural Language Processing group, with a special mention to Bernardo Magnini and Roberto Zanoli, and the whole FBK staff, from the technical to the admin personnel.
I spent countless hours in Povo, and I would not have had the same experience without all the people I connected with: thanks to Samuel, with whom I drank a worrying quantity of hot lemon tea while discussing anything from football to research and life goals – in the office or in an Indian restaurant; to Yi-Ling, Vevake, and all the people who welcomed me to the group.
Thanks to Sara, Marco, Alina, and all the Machine Translation Group with whom I have shared food, breaks, and thoughts.
Thanks to Tommaso, Alessandro, and all the people who knocked at my door for another cigarette-long break or an 8-hour hike.

Every person I have crossed paths with has impacted me and this work.

While I have not been around as much as I would have wanted, I also wanted to thank people at the University of Padua, with a special mention to my fellow Ph.D. students, my co-supervisor Prof. Sabrina Cipolletta and the whole Ph.D. teaching and admin personnel (which I often bored with my very specific and out-of-time requests).

During the Ph.D. I also had the opportunity to connect with people outside my University and research institute.
I want to thank the Huawei Research Ireland teams I worked with during my internship and especially my supervisor, Tri Kurniawan Wijaya.
I also want to thank the TALN group at UPF, who have made me feel welcome and helped me in my journey. Thanks to Prof. Horacio Saggion for hosting and guiding me, and to Santiago, Alba, Kim, Luis, Ishmael, and everybody else who has cheered my stay in one of my favorite cities. And since there is more to work, thank Sofia for sharing a part of the stay and the Sitges' sun.

Theses can become cold list of experiments, successes, and justifications, and I wanted to acknowledge the people behind and around the numbers. Thank you, especially to those I forgot to mention.

I also wanted to thank the people without whom I would literally not be able to write this thesis and exist at all: my father and my mother, who have always encouraged and supported me during my University years. One day I will explain to you what I do in detail, I promise. Thanks to my brothers and all my extended family for being there and for the love you have always given me.

Thanks to my friends for things that this page is too short to contain – and that are honestly too embarrassing to write.

Finally, thanks to Riccardo, with whom I have shared all the ups and downs of this journey. His care, love, and understanding made everything easier, even on bad days. Thanks for all the places we visited together, all the plans, all the weird discussions (and the spare technical talks). Thanks for being able to be with me even if you are somewhere else. Your support, patience, and love at my best and at my worst continue to be a gift I will always treasure.

# Contents

# Common mathematical formulas

| | |
|---|---|
| Term frequency (TF) | $TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$ |
| Inverse Document Frequency (IDF) | $IDF(t,d) = log\frac{\lvert D \rvert}{\lvert\{d \in D : t \in d\}\rvert}$ |
| Term Frequency - Inverse Document Frequency (TF-IDF) | TF-IDF $= \frac{TF}{IDF}$ |
| Cosine similarity (cos(.,.)) | $cos(\vec{x},\vec{y}) = \frac{\vec{x}\vec{y}}{\Vert\vec{x}\Vert\Vert\vec{y}\Vert} = \frac{\sum_i^n x_i y_i}{\sqrt{\sum_i^n x_i^2}\sqrt{\sum_i^n y_i^2}}$ |
| Kullback–Leibler divergence | $D_{\mathrm{KL}}(P \parallel Q) = -\sum_{x \in \mathcal{X}} P(x) \log\left(\frac{Q(x)}{P(x)}\right)$ |
| Singular Value Decomposition | $A = U\Sigma V^T$ |
| Attention | $Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$ |
| Coverage | $Coverage(D,S) = \frac{1}{\lvert S \rvert}\sum_{f \in F(D,S)} \lvert f \rvert$ |
| Density | $Density(D,S) = \frac{1}{\lvert S \rvert}\sum_{f \in F(D,S)} \lvert f \rvert^2$ |

# Chapter 1

# Introduction

## 1.1 Technology foresight: understanding and predicting the technological landscape

The technological landscape evolves continuously and at an increasing speed. New research results, technological paradigms, and applications are announced daily, enhancing opportunities for investors, stakeholders, and society.

Understanding the technological trends (and foreseeing such trends for the future) is critical to define strategies for social problems, enacting public policies, and selecting investment measures. Technology foresight tries to explore these issues systematically. According to a classical definition, it

> *"Seeks to look into the longer term future of science, technology, economy and society with the aim of identifying the areas of strategic research and the emerging generic technologies likely to yield the greatest economic and social benefit." [119]*

While technology foresight cannot predict the future, it aims to anticipate possible developments and trends. It involves approaches to identifying promising scientific discoveries and emerging technologies, assessing their potential impact, and defining strategies for their development and adoption. To achieve these goals, technology foresight experts adopt a systematic approach, aiming at limiting biases and acknowledging assumptions. The goal of the exercise is to provide policymakers, businesses, and other stakeholders with a long-term perspective on the possible futures of technology and the opportunities and challenges these scenarios raise; starting from these assumptions, they can better prepare and make informed decisions.

Technology foresight, however, does not only aim at forecasting but has a proactive component at its core. In foresight exercises, researchers, scientists, experts, policy-makers and other stakeholders from different disciplines are asked to define the most desirable future outcomes. Then, they try to identify which decisions must be taken to make a desired outcome more likely, by systematically evaluating the impact of all possible choices.

For example, consider telemedicine, which is becoming increasingly popular in today's medical practice. The possibility of consulting a physician through a laptop or a smartphone can mitigate problems related to medicine accessibility; however, it also opens

questions about ethics or data privacy. Moreover, telemedicine practices are only possible in an ecosystem with a strong technological component, excellent internet access, proper data infrastructures, and clear legal regulations.
Foresight experts might thus try to monitor telemedicine-related studies and inventions and systematically shape their development (including interventions on the whole technological and legal ecosystem) to optimize the societal benefits.
Telemedicine was one of the issues identified as of interest in the long-term future ($\geq$ 10 years) by the World Health Organization (WHO) in a recent technological foresight exercise[1]. Experts involved in the exercise discussed today's remote clinical care practices from the one hand, and other embryonic elements that might become increasingly relevant in the future from the other. For example, they discussed telemonitoring practices needs and risks, and the legal and technical requirement for telecollaboration between on-site carers and remote ones. Experts identified potential obstacles in conflicting national regulation, privacy issues, possible misuse of personal data, and exclusion of populations with poor technological access. Moreover, cross-disciplinary links were analyzed, for example, regarding the use of software systems for the analysis of data, the use of machine learning, and related problems, e.g., ethics and interoperability.
The process was conducted using a Delphi-like [107] method. Experts were asked to identify which issues "will shape the future of global health"; then, the group was asked to score each of the proposed issues by their impact and plausibility on a 1 to 100 scale. Thus, the most voted issues were short-listed and discussed; the scoring process was then repeated.

Technology foresight methodologies have a relatively long history [125] and were first adopted in Japan in the 1970s as part of their national technology planning efforts; the process was then called "forecasting". In the 1980s, Irvine and Martin [74] introduced the term "foresight" to highlight that the process aims at integrating the predictive component with strategic operations and policies to influence the future rather than solely prognosticate it. Other countries (e.g., France, Sweden, Australia, and Canada) later started to perform technology foresight exercises. In the 1990s, these exercises were adopted in other industrialized countries such as the United States, the United Kingdom, the Netherlands, and Germany. More recently, foresight has also spread to developing countries [143].

Technological Foresight can be pursued from the local to the supranational level, involves various activities, and often requires collaboration between stakeholders, including researchers, policymakers, industry leaders, and civil society organizations.

Methodologies vary and include qualitative (e.g., monitoring and expert-based methods), quantitative (e.g., trend forecasting), and mixed methods (e.g., scenario forecasting).

[1] Emerging trends and technologies: a horizon scan for global public health
https://www.who.int/publications/i/item/9789240044173 [Last accessed: March 2023]

## 1.2 Monitoring emerging technologies and how Natural Language Processing can help

One of the core assumptions in the technology foresight effort is that the technologies that will have a significant impact in the future exist in an embryonic form today. Thus, technology foresight aims at monitoring ideas and products at all stages of development, from basic research to post-commercialization.

This effort is sometimes referred to as "horizon scanning". Horizon scanning typically involves gathering data from various sources, including scientific publications, news reports, industry reports, and expert opinions. This information is then analyzed and synthesized to identify patterns, trends, and potential implications for the future.

At each stage of the technological cycle, information is published in written form: funding agencies publish grants and reports; basic and applied scientists publish their findings in conferences and journals through specialized scientific papers; inventors file patent applications, stored in huge patent datasets; Research and Development (R&D) laboratories write blogs, reports, and white papers; media agencies cover hyped new products in general and specialist newspapers; customers use social media and consciously or unconsciously give feedback.
It would thus be advantageous to access all the information these sources contain for technological monitoring, either for human consultation or in an automatic tool. In the first scenario, information derived from these sources can be used to provide factual data and inform experts and decision-makers that take part in the technology foresight exercise; in the latter case, they could be used in a specialized tool, e.g., gathering information from written sources and providing human-friendly interfaces.

However, the volume and complexity of such sources are so enormous that it becomes impossible to find, select, categorize, and process them to make informed decisions.

Trying to transform an unmanageable amount of data into actionable information is a vast area of research, with contributions from psychology, management, information studies [149], computer science and engineering.

When coming to text, Natural Language Processing (NLP) has played a significant role: taking from computer science, statistics, linguistics, and other fields, it aims at enabling machines to process human language. To make sense of an unprecedented amount of information, subfields of Natural Language Processing have thus specialized in helping users retrieve valuable content through search (information retrieval), in automatically extracting more structured information from documents (information extraction), or in providing simpler outlines of the available pieces of text.

In particular, automatic text summarization – the task of automatically extracting or generating a summary from one or more documents – can have a crucial role in avoiding information overload. Automatic text summarization can enable users to grasp the essence of a text without having to read its whole (possibly noisy) content; moreover, when used in an automatic pipeline, it can help condense documents' content for further processing.

If summarization aims at revealing the core of a text, simplification aims at making it more accessible. Automatic text simplification turns hard-to-read text into content that is easier to understand and process for its intended reader. For example, it can help

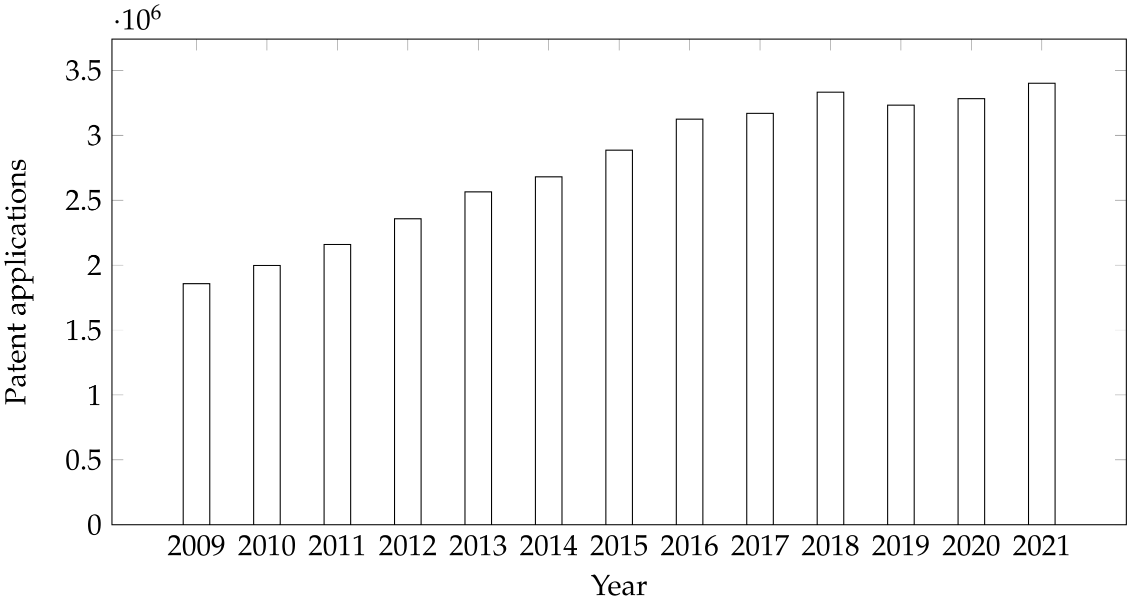


FIGURE 1.1: Total number of patent applications worldwide

people with specific conditions and disorders, second-language learners, people with a low literacy level, or any lay reader when approaching technical text (for example, in the medical or legal domain). The latter is the case we will explore in this thesis.

Summarization and simplification techniques have been applied to various types of documents: non-technical text (e.g., news articles [129, 49, 156]), business-related documents [169, 45], medical notes [83, 145], and scientific articles [38, 58], to name a few. In this thesis, we explore a much less researched class of documents: patents.

## 1.3 Natural Language processing for patent summarization and simplification

Patents are peculiar documents. Many items that pervade our daily lives have been protected through patents, from the lightbulbs (Edison, 1880 & Swan, 1880) to plastic (Baekeland, 1906), from the ballpoint pen (Biro, 1945) to Lego building blocks (Christiansen, 1958).

Patents protect inventions that their holders consider important enough to take legal action to obtain the monopoly in using, making, and selling them – and thus, profit from their wit. Thus, they help in valuing intellectual work. At the same time, inventors must disclose the invention and its characteristics in detail to file a patent application: thus, patents are intended to benefit society as they help new knowledge spread – correcting the tendency to keep valuable technical details secret. Thus, patents are valuable documents that preserve and spread technical information in a similar way as academic papers preserve and spread scientific knowledge.

On the other hand, patents are long and hard to read; they are difficult to process both for machines and humans, and their knowledge is masked by a mix of legal, technical, and extremely vague language. Moreover, the volume of patent applications is enormous.

Figure 1.1 reports the total number of patent applications worldwide from 2009 to 2021: note the highly increasing trend[2].

One of the patents' aims is to make knowledge circulate, and they might be a precious source for technology foresight practitioners. However, due to the issues with their length and complexity, the knowledge patents contain has vastly remained hidden.

As discussed in the previous sections, summarization and simplifications can be valuable tools: they help patent agents, R&D groups, professionals, and technology foresight experts; they can also improve the performance of automatic processes.
This stands true in the patent domain as well. For example, commercial solutions currently provide new patent Abstracts in plain English; however, these solutions are vastly based on manual work, to our best knowledge.

In other domains, summarization and simplification tools and methodologies have shown promising results to assist or completely automatize such processes; applications to the patent domain are, however, limited.
In this thesis, we try to fill this gap by applying Natural Language Processing techniques to the patent domain, focusing on summarization and simplification.

We do so primarily from a Natural Language Processing perspective; we hope to be able to collaborate more with patent and domain experts in the future to gain a more rounded perspective and include more domain insights.

## 1.4 Thesis contributions

The main contributions of this thesis are the following:

- We analyze patent documents and describe characteristics that make them particularly challenging for state-of-the-art Natural Language Processing systems. We comprehensively survey the literature on summarization, simplifications, and other generation techniques in the patent domain.
- We show a number of issues related to the most popular dataset for patent summarization, BigPatent [158]. We discover that the dataset exists in two versions, one of which is flawed as it contains a summary of the invention in the input itself. We also show how, as a consequence, a direct comparison with previous literature is impossible. We clearly describe these issues that were previously completely undocumented.
- Since we noticed a lack of direct method comparisons in the previous literature, we benchmark and compare extractive, abstractive and hybrid summarization methods in the patent domain and discuss their strengths and limitations. We do so both by using automatic metrics and by providing qualitative insights.
- We explore ways to adapt methodologies from scientific papers to the patent domain, focusing on the peculiar patent length. Specifically, we adapt a method to summarize scientific papers that exploits structure and summarizes sections independently. We find, however, that since patents' structure is less predictable and

[2] Source: WIPO IP Statistics Data Center
`https://www3.wipo.int/ipstats/index.htm?tab=patent` [Last accessed: March 2023]

their abstract less compositional, the approach is less successful when changing the target domain.

- Since no data for patent simplification exists, we propose a methodology to create a (more noisy) bronze standard and a (cleaner) silver standard for patent simplification. We show that the silver corpus is considered grammatical, adequate, and simpler from a layperson's perspective. We also show that the corpus can be used to train a sequence-to-sequence system. We make the dataset available for future research.

## 1.5 Thesis outline

**Chapter 2** After this brief Introduction, in Chapter 2, we describe patent documents' challenges and survey the state of the art on approaches to patent summarization and simplification in the Natural Language Processing literature. We find that most techniques that are at the state of the art in other domains have yet to be applied to patent documents. Thus, we identify promising directions, some of which we will follow in this thesis.

**Chapter 3** From Chapter 3, we deepen into patent summarization. Specifically, in this chapter we briefly describe the general automatic text summarization task and reference methods and systems we will use in the following.

**Chapter 4** Chapter 4 details the challenges we faced when comparing our work with existing systems: we found that the most popular dataset for patent summarization, BigPatent [158], exists in at least two very different versions leading to incomparable model performance. In this chapter, we describe the differences between the two datasets, which were previously undocumented, and the issues they rise.

**Chapter 5** In Chapter 5, we benchmark extractive, abstractive, and hybrid models and systems to the patent domain. We evaluate the outputs using both an automatic and a qualitative approach.

**Chapter 6** In this chapter, we try to adapt a method for scientific paper summarization to the patent domain; however, we show that using the patent structure does not seem to improve over baselines; we explore and discuss why the approach does not seem to transfer well to our domain.

**Chapter 7** The patent simplification section contains a brief overview of models and techniques used in the field. However, we find that the most popular systems have yet to be applied to the patent domain; this gap depends on the lack of a parallel dataset for simplifying patent sentences.

**Chapter 8** Given the need for simplification data in the patent domain, in this chapter we explore techniques to automatically generate a silver standard of patent sentences. We conduct a human evaluation campaign and show how the corpus can be used for patent summarization.

**Chapter 9** In this final chapter, draw our Conclusions: we first summarize our main cotributions; then, we discuss the limitations of our work and reflect on any ethical concerns that might arise. Finally, we discuss possible future works.

## List of publications

The following publications were produced over the PhD period:

- S. **Casola**, A. Lavelli, H. Saggion. *Creating a Silver Standard for Patent simplification* (2023), Association for Computing Machinery's Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval (SIGIR)
- S. **Casola**, A. Lavelli, H. Saggion, *What's in a (dataset's) name? The Case of BigPatent* (2022), in Proceedings of the Generation, Evaluation & Metrics Workshop. [31]
- S. **Casola**, A. Lavelli, *Summarization, simplification, and generation: The case of patents* (2022), in Expert Systems with Applications. [29]
- I. Obonyo, S. **Casola**, H. Saggion, *Exploring the limits of a base BART for multi-document summarization in the medical domain* (2022), in Proceedings of the Third Workshop on Scholarly Document Processing. [136]
- S. **Casola**, A. Lavelli, *WITS: Wikipedia for Italian Text Summarization* (2021), in Proceedings of the Italian Conference on Computational Linguistics. [30]
- S. Louvan, S. **Casola**, B. Magnini, *Investigating Continued Pretraining for Zero-Shot Cross-Lingual Spoken Language Understanding* (2021), in Proceedings of the Italian Conference on Computational Linguistics. [113]
- S. **Casola**, I. Lauriola, A. Lavelli, *Pre-trained transformers: an empirical comparison.* (2022), in Machine Learning with Applications. [27]
- S. **Casola**, A. Lavelli, *FBK@SMM4H2020: RoBERTa for detecting medications on Twitter.* (2020) in Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task. [28]

In these three years, we have explored several sources, including patent documents, scientific papers, and social media data. We have also researched various other Natural Language Processing tasks. In this thesis, we chose to only include research related to patents to produce a more compact and coherent document. Please refer to the publications listed above if you are curious about research conducted in other domains.

# Chapter 2

# Natural Language Processing for patent summarization, simplification, and generation: state of the art and open directions

Patents disclose what their creators consider valuable inventions – so valuable, in fact, that they spend a nontrivial amount of time and money on protecting them legally. Not only do patents define the extent of legal protection, but they also describe in detail the invention and its embodiments, its relation to the prior art, and contain metadata. It was common wisdom among patent professionals that up to 80% of the information in patents cannot be found elsewhere [11].

As a result, patents have been widely studied with various aims. Recently, Natural Language Processing (NLP) approaches to patent mining are emerging.
In this chapter, we explore the application of NLP techniques to patent summarization, simplification, and generation.

First, we present an analysis of patents' linguistic characteristics and focus on the idiosyncrasies that negatively affect the use of off-the-shelf Natural Language Processing tools (Section 2.1). After defining the patent summarization, simplification, and generation tasks (Section 2.2), we then describe the few available datasets and the evaluation approaches (Sections 2.3 and 2.4). Next, we review previous work in Sections 2.5, 2.6, and 2.7.

Our review is rather comprehensive, and covers works from the early 2000s to date. We pay special attention to the algorithms and models used from a Natural Language Processing perspective. Note that, however, since patent processing has historically been application-oriented, previous work often used project-specific datasets, making it difficult to compare approaches directly in terms of performance. Finally, we present interesting lines of investigation and open questions, some of which we try to answer in this thesis.

This chapter is based on Casola and Lavelli [29].

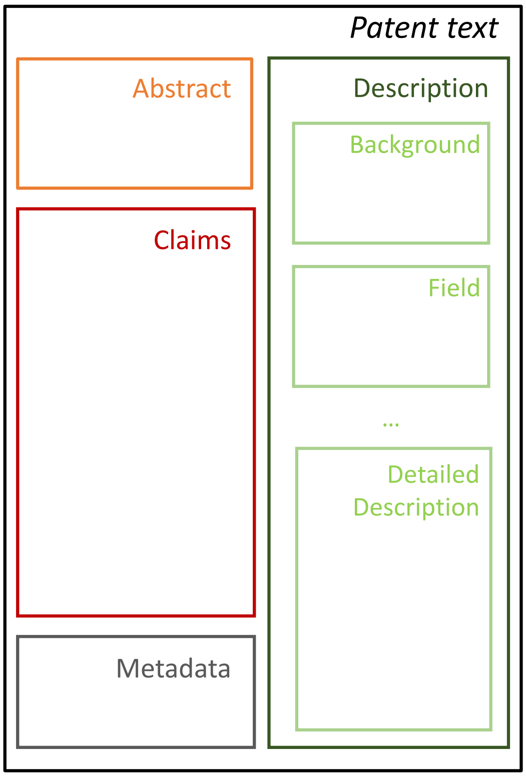


FIGURE 2.1: An example of the textual part of patent documents. Note that, in reality, patent documents tend to span several pages. Figures are also generally included in the full text. See Appendix B for an example of a patent full text.

## 2.1 A primer on patents

Patents are primarily legal documents. Their owner controls the use of an invention for a limited time (usually 20 years) in a given geographic area and thus excludes others from making, using, or selling it without previous authorization. In exchange, the inventor discloses the invention to facilitate the transfer of technology.

This section defines some domain-specific concepts that we will reference in the following; we use patent US4575330A[1] (the antecedent of a 3D printer, designed by Hull in 1989) as a running example. For reference, we include the full text of the patent in Appendix B.

### 2.1.1 Patent documents

Patent documents are highly structured and must follow strict rules[2]. Figure 2.1 sketches the structure of the textual part of patent documents.

Typically, they contain the following textual sections:

[1] https://patents.google.com/patent/US4575330A/en [Last accessed: March 2023]

[2] WIPO Patent Drafting Manual (2007).
URL: https://www.wipo.int/publications/en/details.jsp?id=297 [Last accessed: January 2023].

**Title** E.g., *Apparatus for production of three-dimensional objects by stereolithography*

**Claim** Specifies the extent of legal protection. This section can include multiple claims[3] with a hierarchical structure.

> 1. *A system for producing a three-dimensional object from a fluid medium capable of solidification when subjected to prescribed synergistic stimulation, said system comprising: means for drawing upon and forming successive cross-sectional laminae of said object at a two-dimensional interface; and means for moving said cross-sections as they are formed and building up said object in step wise fashion, whereby a three-dimensional object is extracted from a substantially two-dimensional surface.*
> 2. *An improved system for producing a three-dimensional object from a fluid medium capable of solidification when subjected to prescribed synergistic stimulation, said system comprising: [...]*
> 3. *A system as set forth in claim 2, and further including: programmed control means for varying the graphic pattern of said reaction means operating upon said designated surface of said fluid medium.*

Claims 1 and 2 are independent, while claim 3 is dependent on claim 2, which it further specifies. The document comprises 47 claims, which this thesis is too small to contain. Following patent rules, each claim consists of a single sentence, therefore long, complex, and highly punctuated. The language is abstract to obfuscate the invention's limitations and full of legal jargon.

**Description** This section contains a description detailed enough for a person skilled in the art[4] to make and understand the invention.

> *Briefly, and in general terms, the present invention provides a new and improved system for generating a three-dimensional object by forming successive, adjacent, cross-sectional laminae of that object at the surface of a fluid medium capable of altering its physical state in response to appropriate synergistic stimulation, the successive laminae being automatically integrated as they are formed to define the desired three-dimensional object.*
>
> *In a presently preferred embodiment, by way of example and not necessarily by way of limitation, the present invention harnesses the principles of computer generated graphics in combination with stereolithography, i.e., the application of lithographic techniques to the production of three dimensional objects, to simultaneously execute computer aided design (CAD) and computer aided manufacturing (CAM) in producing three-dimensional objects directly from computer instructions. [...]*

While the Claim section aims at legally protecting the invention (the construct in the mind of the inventor, with no physical substance), the Description discloses one or more embodiments (physical items). Drawings are standard in this section. The Description illustrates the invention to the public, on the one hand, and supports

[3] We will refer to the whole document section using the cased form *Claim*, while the individual *claims* contained in such section will be lowercase.

[4] A "person skilled in the art" has ordinary skills in the invention technical field. For a formal definition, refer to the PCT International Search and Preliminary Examination Guidelines [Last accessed: March 2023]

the Claim, on the other. Notice how, while the language is still convoluted, it is less abstract.

**Abstract** This section summarizes the invention.

*A system for generating three-dimensional objects by creating a cross-sectional pattern of the object to be formed at a selected surface of a fluid medium capable of altering its physical state in response to appropriate synergistic stimulation by impinging radiation, particle bombardment or chemical reaction, successive adjacent laminae [...].*

**Other metadata** These include standard classification codes, prior art citations, relevant dates, inventors', assignees', and examiners' information.

**Patent classification codes** Patents are classified using standard codes. The Patent Classification (IPC)[5] and the Cooperative Patent Classification (CPC)[6] are the most widespread. Patent examiners assign codes manually depending on the invention's technical characteristics.

Patent US4575330A has 14 IPC classification codes. For example, code *G09B25/02* indicates that the patent is in the Physics (G) section and follows to specify the class (*G09 - EDUCATION; CRYPTOGRAPHY; DISPLAY; ADVERTISING; SEALS*), sub-class (*G09B - EDUCATIONAL OR DEMONSTRATION APPLIANCES; APPLIANCES FOR TEACHING, OR COMMUNICATING WITH, THE BLIND, DEAF OR MUTE; MODELS; PLANETARIA; GLOBES; MAPS; DIAGRAMS*), group (*G09B25/00 - Models for purposes not provided for in group, e.g. full-sized devices for demonstration purposes*), and sub-group (*G09B25/02 - of industrial processes; of machinery*).

### 2.1.2 Patent language

In this section, we describe what makes patent documents unique from a linguistic perspective. Few documents are, in fact, as hard to process (for both humans and automatic systems) as patents, with their obscure language and complex discourse structure.

**Long sentences** According to patents' rules, each claim must be written in a single sentence, which is, therefore, particularly long. Verberne et al. [193] examined over 67 thousand Claim sections and found a median length of 22 and a mean of 55; note that this figure is highly underestimated, as the authors segment sentences using semicolons in addition to full stops. In contrast, they found that the British National Corpus median length (when segmented using the same methodology) is less than 10; the British National Corpus contains samples from several sources (news, novels, letters, essays) and is thus used as a corpus for "general" English. For comparison, the first claim in patent US4575330A (a "rather short" one) is 69 words long, while claim 2 contains 152 words. Shinmori et al. [164] found similar characteristics in Japanese. While most quantitative work focuses on the Claim, sentences in other sections are also remarkably long.

[5] wipo.int/classifications/ipc/en/ [Last accessed: March 2023]

[6] cooperativepatentclassification.org [Last accessed: March 2023]

**Words' distribution and vocabulary** Claims do not use much lexicon not covered in general discourse, but their word frequency is different, and novel technical multiword terms are created *ad hoc* [193]. Moreover, many words are used unusually: *said*, for example, typically refers back to a previously mentioned entity, repeated to minimize ambiguity (e.g., *A system for [...], said system comprising [...]*, in claim 1); transitions (e.g., *comprising*, *including*, *wherein*, *consisting*) have specific legal meanings. The Claim's language is abstract (*system*, *object*, *medium* in claim 1), not to limit the invention's scope, while the Description is more concrete [37].

**Complex syntactic structure** Patent claims are built out of noun phrases instead of clauses, making it nontrivial to use standard NLP resources. As a result, previous work has tried to adapt existing parsers with domain-specific rules [25] or simplify the claim before parsing [127]. Figure 2.2 shows the dependency tree of Claim 1 of patent US4575330A.

## 2.2 Tasks description

Chapter 3 and Chapter 7 describe the tasks of summarization and simplification from a Natural Language Processing perspective. Here, we will discuss their specific applications in the patent domain.

### 2.2.1 Summarization

Loosely speaking, a summary is a piece of text that, based on one or more source documents, a) contains the main information in such document(s) and b) is shorter, denser, and less redundant. A possible taxonomy of text summarization approaches is provided in Chapter 3. Here, we will categorize previous work according to the following dimensions:

**Extractive vs. abstractive** Most previous work relies on extractive approaches, given the legal nature of patent documents – i.e., it directly selects sentences from the source and concatenates them. Recently, some general-purpose abstractive models – i.e., which generate the summary as a new piece of text – have also been tested on the BigPatent dataset [158], which we will discuss in the following.

**Generic vs. query-based** While most patent summarization approaches are generic, query-based models [57, 55, 56] – i.e., models that primarily focus on content related to a given query – might also be relevant. For example, during a prior art search, the user might only be interested in aspects of the retrieved documents that might invalidate their patent.

**Human- vs. machine-focused** Patent summaries are typically intended for humans, but machine-focused approaches have also been explored. Tseng, Lin, and Lin [189] and Tseng et al. [190], for example, perform summarization in view of patent map creation and classification. A patent map is a visualization of patents in a given technology field. It aims to show patents in a given technology space, verify their characteristics and relations, understand trends, and possible gaps, perform market research, and show possible infringements.

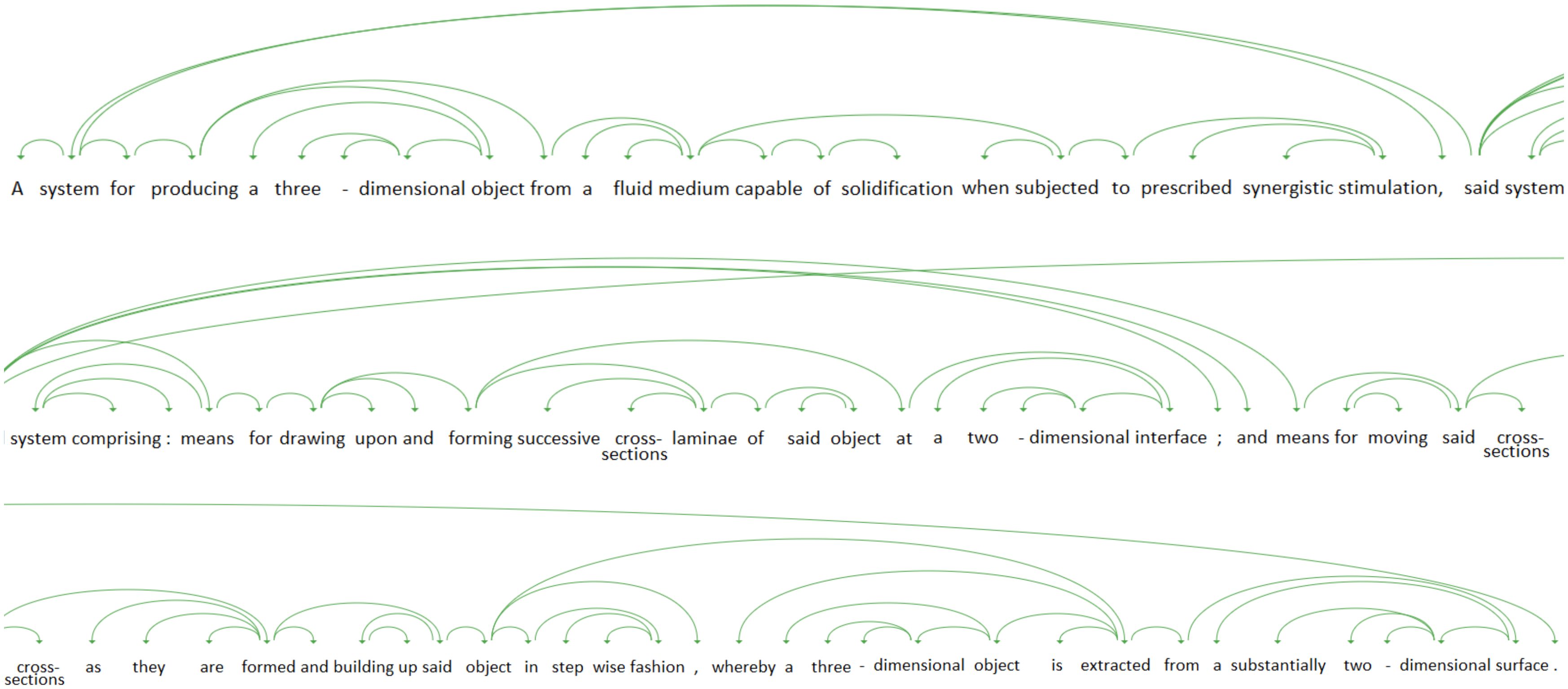

A system for producing a three - dimensional object from a fluid medium capable of solidification when subjected to prescribed synergistic stimulation, said system
system comprising : means for drawing upon and forming successive cross- sections laminae of said object at a two - dimensional interface ; and means for moving said cross- sections
cross- sections as they are formed and building up said object in step wise fashion , whereby a three - dimensional object is extracted from a substantially two - dimensional surface .

FIGURE 2.2: The dependency tree of Claim 1 of patent US4575330A. We report the tree on multiple lines given its length to improve its readability.

**Language-specific vs. multilingual** While published research has primarily been anglocentric, some works in other languages and language-independent techniques have been proposed.

As expected, patent summarization comes with its challenges. For example, while in some domains (e.g., news), the essential facts are typically in the first paragraphs, this assumption does not hold for patents, whose important content is spread in the whole input. Patent Abstracts contain a high percentage of n-grams not in the source and shorter extractive fragments. Finally, the summaries' discourse structure is complex, and entities recur in multiple sentences. All these characteristics make patents an interesting testbed for summarization, for which a real semantic understanding of the input would be crucial [158].

In addition to the research interest, patent summaries are practically relevant for R&D teams, companies, and stakeholders. A brief search of online services showed that some companies sell patent summaries and related data as a paid service. For example, Derwent[7] produces patent abstracts distilling the novelty, use, and advantages of the invention in plain English; to the best of our knowledge, the abstract is manually compiled by domain experts.

### 2.2.2 Simplification

Automatic simplification reduces the linguistic complexity of a document to make it easier to understand. In contrast with summarization, the simplified text does not necessarily lack the details from the original text. For general text, approaches vary depending on the system's target user (e.g., second-language learners, people with reading disorders, and children).

Given patents' complexity – lexically and syntactically – the challenge lies in making their content accessible to the lay reader (which justifiably gets scared away from patents) and in simplifying the experts' work. Moreover, simplification approaches might improve the performance of other Natural Language Processing tasks, as we will see in Chapter 7.

We will consider the following aspects:

**Expert vs. lay target reader** Patents' audience ranges from specialists (e.g., attorneys and legal professionals), to laypeople (including academics) who might be interested, for example, in the invention's technical features. Depending on the target user (and, in turn, on the target task), the degree of simplification might vary. When considering the legal nature of patents, for example, special attention should be given to keeping their scope unchanged. The first claim of patent US4575330A, for example, states: *"A system for producing [...] comprising: means for drawing [...]; and means for moving [...]."*. A system "comprising" a feature might include additional ones; thus, replacing the term with "consisting of" – which, in patent jargon, excludes any additional component – would be problematic, even if thesauruses treat the terms as synonyms[8]. Obviously, the attention to the jargon can be loosened if the target

[7] `https://clarivate.com/derwent` [Last accessed: March 2023]

[8] Compare, for example, the Collins Online Thesaurus and the European Patent Office guidelines. [Last accessed: March 2023]

user is more interested in the overall technical characteristics of the embodiments rather than in the invention's legal scope.

**Unstructured vs. structured output** The simplification system's output can be either a text or a more complex data structure. A textual output can be formatted appropriately (e.g., coloring essential tokens [137]), annotated with explanations (e.g., with links from a claim to a Description passage [163]), or paraphrased [21]. Alternatively, a graphical representation, in the form of trees or graphs – which, e.g., highlights the relation among the invention components – can be used.

**Application** The simplification system can be designed with a specific application in mind: Okamoto, Shan, and Orihara [137], for example, designed an interface to help patent experts in comparing documents from the same patent family[9].

As in the case of summaries, designing appropriate simplification systems has interesting use cases. Suominen et al. [180] performed a user study with both experts and laypeople: most of their participants considered patents difficult to read. When presented with various reading aids, most considered them useful. Even law scholars have called for the use of a simpler language in patents [50]. Commercially, companies that provide patent reports do so in plain language. Somewhat ironically, Derwent goes as far as replacing the document title with a less obscure one of more practical use.

### 2.2.3 Generation

Natural Language Generation is a Natural Language Processing branch that aims to generate new, original text automatically. This definition might include summarization and simplification as text-to-text instances. Here, we will use Patent Generation to refer to methods that aim at generating a patent or part of it. To the best of our knowledge, this line of research is relatively new and is likely inspired by the recent success of modern generative models (e.g., GPT and its evolutions [147, 146, 23]) in various domains, including law [71], health [7] and journalism [166], to name a few.

Some approaches only produce "patent-like" text (i.e., employing technical terminology and respecting patents' writing rules): their generation is unconstrained or constrained to a short user prompt – the first words of a text that the system needs to extend coherently. Their practical use is likely limited, but their success shows that even patents' obscure language can be mastered by machines, at least at a very superficial level. Another class of approaches conditions the generation of a fragment of the patent to produce a coherent output. For example, one might want to produce a plausible patent Abstract given its Title or a set of coherent claims with a given Description. In this case, the generation is constrained to the whole input section (e.g., the Title text) and the type of output section (e.g., the Abstract).

While patent generation is still in its early days, researchers dream of "augmented inventing" [99], assisting inventors in redefining their ideas and helping with patent drafting. To this end, some hybrid commercial solutions are already in the market[10].

[9] A patent family is a set of patents that relate to the same invention.

[10] See, for example,
https://bohemian.ai/case-studies/automated-patent-drafting/,
https://www.patentclaimmaster.com/automation.html,
https://harrityllp.com/services/patent-automation/ [Last accessed: March 2023]

## 2.3 Datasets

Patent documents are issued periodically by the responsible patent offices. The United States Patent and Trademark Office (USPTO), for example, publishes patent applications and grants weekly, along with other bibliographic and legal data[11]. To access the documents programmatically, Application Programming Interfaces (APIs) are available. PatentsView[12], for example, is a visualization and mining platform to search and download USPTO patents, updated every three months. It provides several endpoints (patent, inventor, assignees, location, CPC, etc.) and a custom query language. Google also provides public datasets[13], accessible through BigQuery.

While it is relatively easy to obtain raw patent text, few curated datasets exist. These data are of the greatest importance: having a set of shared benchmarks allows to directly compare approaches, which is much more difficult otherwise. A large-scale dataset for patent summarization is BigPatent[14] [158]. The dataset was built for abstractive summarization and contains 1.3 million patents' Descriptions and their Abstracts. We will describe BigPatent (and its issues) in Chapter 4.

In 2022, Suzgun et al. [183] published the Harvard USPTO Patent Dataset. The dataset contains more than 4.5 million patents (in their inventor-submitted version) with their metadata and is built to be used in a multiplicity of tasks, including summarization.

While most previous work focuses on the Claims section, no comparable Claim to summary dataset exists (nor would it be easy to obtain), and authors resort to expert-written summaries for evaluation.

For patent simplification, no simplified corpus or parallel corpus exists to date. We propose a methodology to create a silver standard in Chapter 8.

## 2.4 Evaluation

In this section, we will only discuss how previous work has evaluated summarization, simplification, and generation approaches in the patent domain; for a more general view, please see Chapter 3 and Chapter 7.

Qualitative evaluation of patent summarization might involve experts or non-experts; Mille and Wanner [127], for example, assess summaries intelligibility, simplicity, and accuracy on a Likert scale [104]. Quantitatively, ROUGE [105] is often used, as in the general automatic summarization field. However, not all studies follow this convention: some measured the similarity between the generated text and the reference summary in uni-gram Precision, Recall, and $F_1$, while some report the Compression Ratio (the ratio among the length of the source and that of the summary) and the Retention Ratio (the percentage of original information kept in the summary) only. When summarization is part of a pipeline, the relative improvement of the downstream task is considered.

[11] `https://developer.uspto.gov/data` [Last accessed: March 2023]

[12] `www.patentsview.org/` [Last accessed: March 2023]

[13] `https://console.cloud.google.com/marketplace/browse?q=google%20patents%20public%20datasets&filter=solution-type:dataset` [Last accessed: March 2023]

[14] `https://evasharma.github.io/bigpatent` [Last accessed: March 2023]

When evaluating simplification approaches, two complementary points of view exist. The first only considers the method's correctness: if the algorithm needs to segment text in a certain way, one can manually annotate a segmented gold standard and measure accuracy. However, assessing the readability improvement requires qualitative studies.

Suominen et al. [180], for example, use a questionnaire for quantifying patents' complexity and test simplification solutions. Following their work's findings, experts' and laypeople's opinions should be analyzed separately, as they are concerned with different issues. For instance, experts worry that the simplified patent might be misrepresented and its legal scope changed, while laypeople demand strategies to understand the invention and find information.

Finally, measuring the quality of the generated patent text is generally tricky. When no reference exists, some authors have introduced *ad hoc* measures (see, for example [100]); when a human-written reference exists, metrics such as ROUGE can be used. Note that some studies have openly criticized the use of ROUGE; Lee [98], for example, also reports the results using the Universal Sentence Encoder [33] representation, which they speculate handles semantics better in their use case.

## 2.5 Approaches for patent summarization

In this section, we describe extractive and abstractive approaches to patent summarization. As we discussed already, their direct comparison is difficult, as publications tend to use slightly different tasks on unshared data. The approaches discussed in this chapter are summarized in Table 2.1.

### 2.5.1 Extractive summarization

Extractive approaches select the most informative sentences in the original document.

A typical pipeline comprises the following steps:

1. Document segmentation: documents are split into sentences or paragraphs using punctuation or heuristics. While many approaches work at the sentence level, Codina-Filbà et al. [37] argued that patent sentences are too long to be used directly and further segmented them. In many cases, only some Sections (e.g., the Description or the Claims) are considered.

2. Sentence preprocessing: depending on needs, this step might include standard text preprocessing, e.g., removing stopwords or stemming. Given the peculiar patent style, patent-specific stopwords (cured by experts) also need to be taken into account if removing stopwords. Some approaches [184, 187] only keep specific Parts of Speech.

3. Feature extraction: for each sentence, general-domain features include keywords, title words, cue words (from expert-designed lists), and sentence position. In particular, patents contain several multi-word entities that need to be identified. To this end, Tseng, Lin, and Lin [189] propose an algorithm that merges nearby unigrams words and extracts maximally repeated strings as multi-word terms. Given that patent text is often full in technical terms, Trappey, Trappey, and Wu [187] and

| Study | Approach | Main contribution | Limitations | Dataset |
|---|---|---|---|---|
| Tseng, Lin, and Lin [189] and Tseng et al. [190] | Extractive | Domain-specific considerations;<br>key-phrase extraction algorithm | Extrinsic eval. only (classification surrogates) | National Science Council Patent Set (612 patents) |
| Trappey, Trappey, and Kao [184] and Trappey and Trappey [185] | Extract information-dense paragraphs | Application of general-domain techniques | Evaluation | 111 patents |
| Trappey, Trappey, and Wu [187] and Trappey, Trappey, and Wu [186] | Extract information-dense paragraphs | Ontology for key-phrase extraction | | 200 patents |
| Mille and Wanner [127] | Abstractive (Deep-Syntactic Structs) | Multilinguality | Complexity | 50 patents |
| Brügmann et al. [24] and Codina-Filbà et al. [37] | Hybrid | Patent-specific approach (lexical chain, Claim-Description alignment, sentence fragmentation) | Complexity | 26 patents (test) |
| Girthana and Swamynathan [56, 57, 55] | Extractive (query-oriented) | Query-oriented approach<br>Query expansion strategies | | Smartphone -related patents |
| Sharma, Li, and Wang [158] | | Dataset | Complex Abstract style | 1.3M patents |
| Souza et al. [171] | Extractive, semantic similarity | Summarization to name patent groups | | 733 patents (test) |
| Trappey et al. [188] | Hybrid (abstractive to extractive) | Attention-based method for extracting keywords | Complexity | 1708 (train)<br>30 (test)<br>patents |
| Zhang et al. [212]<br>He et al. [64]<br>Zaheer et al. [209] | Abstractive (transformer-based) | Analysis of SOTA general-domain NLP systems in the patent domain | Data requirements<br>Computational cost | BigPatent |
| Souza, Meireles, and Almeida [172] | Abstractive (LSTM), semantic similarity | Summarization to name patents group | Abstractive approaches inferior to extractive ones | 41,527 (train),<br>733 patents (test) |

TABLE 2.1: Surveyed studies for Patent Summarization.

Trappey, Trappey, and Wu [186] use a domain ontology for identifying domain-specific key phrases. The approaches above try to customize general-discourse features to the patent domain; in contrast, Codina-Filbà et al. [37] propose a linguistically-motivated domain-specific approach. They consider the lexical chain length as a measure of entity importance: i.e., invention components that appear many times in the Claim and Description are particularly relevant. Given the abnormal patents' sentence length, they further segment sentences and use fragments as extractive candidates.

In most approaches, the segment position is also considered (favoring sentences at the beginning of a paragraph or paragraphs at the beginning or end of a Section). Query-oriented approaches also measure the sentence similarity to the query (e.g., with overlapping words [56]), which can be further expanded using a domain ontology [55] or general-domain resources [57] like WordNet. Query expansion can be particularly important as different patent documents can purposely use completely different wording for similar components. Table 2.2 includes some frequent features in extractive patent summarization.

4. Sentence weighting: the extracted features are used to score the sentence relevance in the summary. For example, Tseng, Lin, and Lin [189] heuristically score sentences as:

   $$score(S) = \left( \sum_{w \in key_w, title_w} TF_w + \sum_{w \in clue_w} mean(TF) \right) \times FS \times P$$

   where TF is the term frequency of word w in sentence S, mean(TF) is the average term frequency over keywords and title words in S, and FS and P are the sentence position weights. In particular, FS is set to 1.5 if the sentence is the first in the paragraph and to 1 otherwise; P is the position weight of the sentence with respect to the Section and is set to 2 or 4 if the sentence is in the first or last two paragraphs of the Section respectively, and to 1 otherwise.

   Another option is to learn weights from data directly: for example, Codina-Filbà et al. [37] score each segment as $score(S) = \sum_i^n w_i f_i$; they use linear regression to learn features weights based on textual segments and their cosine similarity to the gold standard.

   Lastly, sentences can be directly classified as relevant or not relevant: to this end, Girthana and Swamynathan [55, 56] train a Restricted Boltzmann Machine [97] without supervision. To minimize repetitions, Trappey, Trappey, and Kao [184], Trappey, Trappey, and Wu [187], and Trappey and Trappey [185] cluster semantically similar sentences and only select one sentence per cluster.

5. Summary generation: most commonly, the final summary consists of the union of the extracted sentences. Trappey, Trappey, and Wu [187] and Trappey, Trappey, and Wu [186] also draw a summary tree linked to the domain ontology.

While popular, the above pipeline is not the only route to extractive summarization. Alternatively, Bouayad-Agha et al. [21] exploit the patent's complex discourse structure, which they prune following predefined domain-specific rules. Finally, Souza et al. [171]

| Features | Description |
|---|---|
| **Entity features** | |
| Term frequency - Inverse Document Frequency | Measures a keyword importance |
| Ontology-based | Concepts from a domain-specific ontology; specific concepts are more relevant |
| Coreference-chain based | Entities coreferenced repeatedly are more central |
| **Segment features** | |
| Title similarity<br>Abstract similarity<br>Claim similarity | Computed by considering either word overlap or semantic similarities |
| Query similarity | Relevance to the query |
| Position | Patent section (Claim, Description, etc) and sentence position within the section |
| Length | Overly long segments might be discouraged |
| Number of keywords | |
| Number of cue-words | |

TABLE 2.2: Extractive features. We use the term *entity* to generically refer to keywords, phrases, or other mentions in the document. Similarly, *segment* indicated both complete sentences and fragments.

discuss applying general-domain algorithms to patent sub-groups naming[15]: in that context, Latent Semantic Analysis (LSA) [44] performs best compared to LexRank [46] and to a TF-IDF approach.

### 2.5.2 Abstractive models

Abstractive models aim at generating a stand-alone text whose content is not directly extracted from the source.

In the patent domain, the first approaches used deep syntactic structures. Given patents' linguistic structure, Mille and Wanner [127] first simplify the claims (see [20]) to achieve adequate parsing performance; then, they map the shallow syntactic structures to deep ones, using rules. Deep syntactic structures are closer to a semantic representation and thus used for summarization: to this end, the least relevant chunks are removed using handcrafted rules. Finally, they transfer the summarized deep structures to the target language (English, French, Spanish, or German) and use a generator to convert them to text.

While neural models have widely been used in automatic text summarization of text, they require a large training set, which is probably why they have only spread very recently in the patent domain. No large-scale benchmark, in fact, existed before 2019, when BigPatent [158] was published. Sharma et al. proposed several baselines: an LSTM [182] with attention [12], a Pointer-Generator [157] with and without coverage, and SentRewriting [34] (a hybrid approach).

Given its differences with the previously available datasets (mostly in the news domain) – in terms of style, content distribution, and discourse structure –, BigPatent became an

[15] Patent sub-groups are the most specific level of the patents' classification hierarchy and are named with a representative name, e.g., *"Extracting optical codes from image or text carrying said optical code"*.

| Model | R-1 | R-2 | R-L |
|---|---|---|---|
| TextRank [124] | 35.99 | 11.14 | 29.60 |
| LexRank [46] | 35.57 | 10.47 | 29.03 |
| SumBasic [134] | 27.44 | 7.08 | 23.66 |
| RNN-ext RL [34] | 34.63 | 10.62 | 29.43 |
| LSTM seq2seq [182] + attention | 28.74 | 7.87 | 24.66 |
| Pointer-Generator [157] | 30.59 | 10.01 | 25.65 |
| Pointer-Generator + coverage [157] | 33.14 | 11.63 | 28.55 |
| SentRewriting [34] | 37.12 | 11.87 | 32.45 |
| TLM [144] | 36.41 | 11.38 | 30.88 |
| TLM + Extracted sentences | 38.65 | 12.31 | 34.09 |
| CTRL$_{sum}$ [64] | 45.80 | 18.68 | 39.06 |
| Pegasus$_{base}$ [212] (no pretraining) | 42.98 | 20.51 | 31.87 |
| Pegasus$_{base}$ | 43.55 | 20.43 | 31.80 |
| Pegasus$_{large}$ (C4) | 53.63 | 33.16 | 42.25 |
| Pegasus$_{large}$ (HugeNews) | 53.41 | 32.89 | 42.07 |
| BIGBIRD-RoBERTa (base, MLM) [209] | 55.69 | 37.27 | 45.56 |
| BIGBIRD-Pegasus (large, Pegasus pretrain) | 60.64 | 42.46 | 50.01 |
| LongT5 [63] | **76.87** | **66.06** | **70.76** |

TABLE 2.3: Results on the BigPatent dataset. TextRank, LexRank, SumBasic, and RNN-ext RL are extractive baselines. TLM uses a GPT-like transformer (TLM) and concatenates extracted sentences to the Description (TLM + Extracted sentences). Results reported for CTR refer to unconditioned summarization. For Pegasus, we report results for the base model (223M parameters) with and without pre-training and a larger model (568M parameters) independently pre-trained on a dataset of web pages (C4) and a dataset of news articles (HugeNews). For BIGBIRD, results using RoBERTa's (MLM) and Pegasus' (Gap Sentence Generation) pre-training are considered. All results are from the models' papers. Note that direct comparison of the models is not possible, as explained in Chapter 4.5.

interesting testbed for general domain NLP summarization models: this is, for example, the case of Pegasus [212].

One of the significant challenges of the dataset is the input length, which is very large (with a 90% percentile of 7,693 tokens in its original version) and is problematic for standard transformers (whose attention mechanism scales quadratically in the input size): to this end, BIGBIRD [209] proposes a sparse attention mechanism which allowed dealing with very large input sequences and performed well on the dataset. To the best of our knowledge, the best-performing model on the dataset is currently LongT5 [63], which couples the T5 [148] model with an efficient attention mechanism.

Summarization models' performance on the BigPatent dataset is shown in Table 2.3. Transformer models obtain the best results, in line with the general trend in Natural Language Processing; note, however, that a direct comparison among results is not possible, as explained in Chapter 4.

Finally, summarization methods could also be used for solving domain-specific tasks. CTRLsum [64], for example, is a system that allows controlling the generated text by interacting through keywords or short prompts. The authors experiment with inputting *[the purpose of the present invention is]* to retrieve the patent aim. Finally, Souza, Meireles, and Almeida [172] have compared extractive and abstractive models in naming patents' subgroups. When used to "summarize" the Abstract to produce a patent Title – which should contain, similarly to its subgroup name, the essence of the invention – extractive methods were found superior. This result highlights the challenges met by abstractive models, which are likely to be magnified in the legal domain.

### 2.5.3 Hybrid models

Hybrid models integrate elements of extractive and abstractive summarization. For example, the TOPAS workbench [24] included a module that first selects segments in an extractive manner and then paraphrases them using deep syntactic structures. A similar approach was adopted in [37]. In these approaches, a sentence fragment is the unit of extraction (sentences are too long to be used directly); extracted fragments are then paraphrased. More recently, Pilault et al. [144] have shown that adding previously extracted sentences to the input when training a language model helps with long dependencies and improves the model's abstractiveness. While the models described so far train the extractive and the abstractive components separately, SentRewriting [34] uses reinforcement learning for selecting salient sentences and training the model end to end. The last two mentioned models are general-domain and also test their results on patents.

In contrast with the previous works, Trappey et al. [188] explore an abstractive to extractive approach. They use an LSTM with attention to guide the extraction of relevant sentences: it receives a set of English and Chinese documents (Title, Abstract, and Claim) and is trained to produce a human-written summary (abstractive component). After the training, the words with the highest attention weights are retrieved and treated as automatically-extracted keywords; sentences are then scored and extracted accordingly (extractive component). This approach is domain-specific and is used as a way to simplify keyword extraction, which is complex in the patent domain.

## 2.6 Approaches for Patent simplification

Patents' claims are the hardest section of an overall hard-to-read document. As such, a lot of effort has been spent in improving the accessibility and readability of the Claim. Table 2.4 summarized previous work.

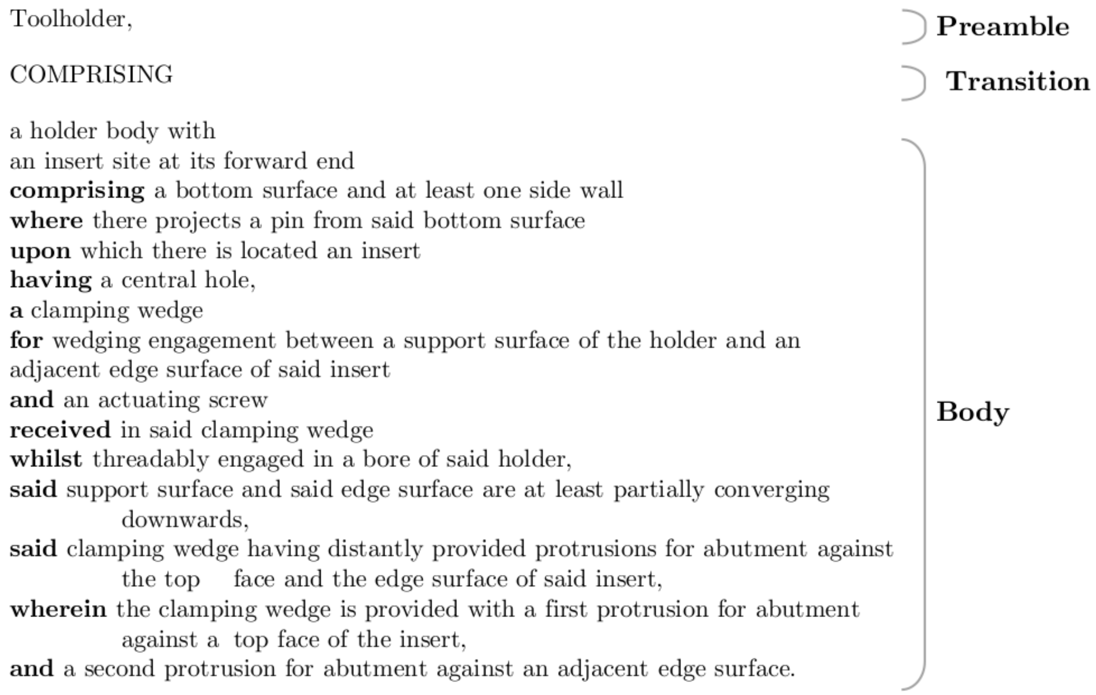
Given the Claim's legal nature, however, the extent of the modification is crucial, and previous approaches' views to the task have varied widely.

Ferraro, Suominen, and Nualart [51], for example, aim at improving the Claim's presentation without modifying its text. They segment each claim into a preamble, a transition text, and a body (rule-based) and then further divide the body into clauses using a Conditional Random Field. Knowing the elements' boundaries, the claim can then be formatted more clearly, e.g., by adding line breaks. See Figure 2.3 for an example.

A somewhat opposite approach was taken in the PATExpert project [198], which developed a rewriting and paraphrasing module [21]. The researchers considered two levels

| Study | Approach | Main contribution | Limitations | Dataset |
|---|---|---|---|---|
| Shinmori et al. [165, 164] | Rhetorical Structure Theory<br>Linguistic analysis<br>Linguistic rules | Claim explanation through Description segments | | NTCIR3 data<br>59,956 patents |
| Bouayad-Agha et al. [21] | Discourse-based and Deep Syntactic Structure-based simplification | Shallow and deep strategies | Legal scope can be modified<br>Complexity | 30 patents (test) |
| Mille and Wanner [126] | Deep Syntactic Structure-based simplification | | | 500 sentences (test) |
| Bouayad-Agha et al. [20] | Discourse-structure simplification | | | 29 patents (test) |
| Andersson, Lupu, and Hanbury [8] | Claim Dependencies Graph | Adaptation of general NLP tools to the patent domain | Errors in PoS tagging can lead to graph collapse | EN CLEF–IP 2012 Passage Retrieval topic set (40 train, 600 test claims) |
| Ferraro, Suominen, and Nualart [51] | Text segmentation | Increase readability without modifying the text | Body segmentation can be improved | 821 train, 80 test patents |
| Sheremetyeva [160] | Rules, linguistic knowledge, statistics | Text highlighting, claims diagram | Complexity<br>Linguistic knowledge is domain-specific | 25 patents |
| Okamoto, Shan, and Orihara [137] | Claim structure analysis through Information Extraction | Relation extraction techniques for highlighting<br>Claim aspects | | 12,972 patents on AI |
| Kang, Souili, and Cavallucci [82] | Rule-based<br>Improve the readability of an extracted graph | Machine-oriented simplification for information extraction and graph visualization | Simplification does not improve extraction performance | 30 patents (test) |
| Suominen et al. [180] | User Study | Evaluation of users attitude toward patents and simplification solutions | | |

TABLE 2.4: Surveyed studies for Patent Simplification.



Toolholder, **Preamble**

COMPRISING **Transition**

**Body**

a holder body with
an insert site at its forward end
**comprising** a bottom surface and at least one side wall
**where** there projects a pin from said bottom surface
**upon** which there is located an insert
**having** a central hole,
**a** clamping wedge
**for** wedging engagement between a support surface of the holder and an adjacent edge surface of said insert
**and** an actuating screw
**received** in said clamping wedge
**whilst** threadably engaged in a bore of said holder,
**said** support surface and said edge surface are at least partially converging downwards,
**said** clamping wedge having distantly provided protrusions for abutment against the top face and the edge surface of said insert,
**wherein** the clamping wedge is provided with a first protrusion for abutment against a top face of the insert,
**and** a second protrusion for abutment against an adjacent edge surface.

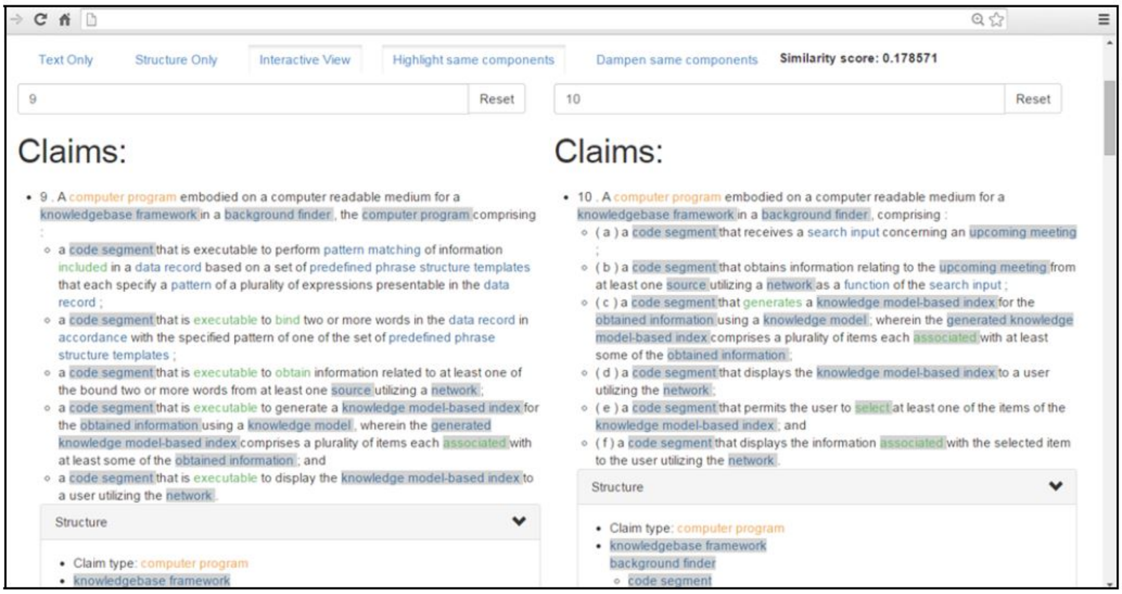
FIGURE 2.3: A segmented patent. Adapted from [51].

of simplification: one uses surface criteria to segment the input and reconstructs chunks into shorter, easier-to-read sentences [20]. The other [126] is conceptually similar to [127] for multilingual summarization: after shallow simplification and segmentation, patents are parsed and projected to Deep Syntactic Structures. This representation is, in turn, used to rewrite a text that is simpler to process for the reader (possibly in another language). Both approaches modify the patent text. Note how, in this framework, rewriting and summarization are essentially unified, with the key difference that no content is removed for simplification.

Instead of relying on linguistic techniques, Okamoto, Shan, and Orihara [137] use an Information Extraction engine that detects entity types and their relations using distant supervision. They provide a visualization interface that a) formats each patent claims to improve readability: color is used to highlight the claim type (e.g., apparatus, method), the transaction, and technical components in the patent body; b) shows the Claim structure: for each claim they include its type, dependencies, and references to other technologies and components. See Figure 2.4.

They target patent experts, which might use the system to compare claims (e.g., in the same patent family) and search for similar documents.

The approaches described so far output a simplified and easier-to-read textual version of the original Claim. Another option is to visualize them in a structured way. Andersson, Lupu, and Hanbury [8], for example, obtain a connected graph of the claim content; each node contains a noun phrase (NP) and is linked through a verb, a preposition, or a discourse relation. An example is shown in Figure 2.5 (top). Similarly, Kang, Souili, and Cavallucci [82] constructs a graph for visualizing the patent content in the contest of an Information Retrieval pipeline. Sheremetyeva [160] uses visualization on two levels: they



Claims:

- 9 . A computer program embodied on a computer readable medium for a knowledgebase framework in a background finder , the computer program comprising :
  - a code segment that is executable to perform pattern matching of information included in a data record based on a set of predefined phrase structure templates that each specify a pattern of a plurality of expressions presentable in the data record ;
  - a code segment that is executable to bind two or more words in the data record in accordance with the specified pattern of one of the set of predefined phrase structure templates ;
  - a code segment that is executable to obtain information related to at least one of the bound two or more words from at least one source utilizing a network ;
  - a code segment that is executable to generate a knowledge model-based index for the obtained information using a knowledge model , wherein the generated knowledge model-based index comprises a plurality of items each associated with at least some of the obtained information ; and
  - a code segment that is executable to display the knowledge model-based index to a user utilizing the network .

Structure

- Claim type: computer program
- knowledgebase framework

Claims:

- 10 . A computer program embodied on a computer readable medium for a knowledgebase framework in a background finder , comprising :
  - ( a ) a code segment that receives a search input concerning an upcoming meeting ;
  - ( b ) a code segment that obtains information relating to the upcoming meeting from at least one source utilizing a network as a function of the search input ;
  - ( c ) a code segment that generates a knowledge model-based index for the obtained information using a knowledge model ; wherein the generated knowledge model-based index comprises a plurality of items each associated with at least some of the obtained information ;
  - ( d ) a code segment that displays the knowledge model-based index to a user utilizing the network ;
  - ( e ) a code segment that permits the user to select at least one of the items of the knowledge model-based index ; and
  - ( f ) a code segment that displays the information associated with the selected item to the user utilizing the network .

Structure

- Claim type: computer program
- knowledgebase framework
  background finder
  - code segment

FIGURE 2.4: Interface for comparing two patents, from [137].

first construct a hierarchical tree of the whole Claim section (highlighting dependency relations) and simplify each claim. In this phase, a tailored linguistic analysis is used [161]; the simplified claim is segmented into shorter phrases (whose NPs are highlighted and linked to the Description) and visualized as a forest of trees. An example is shown in Figure 2.5 (bottom).

Note that most approaches do not measure the improvement in readability so it is not clear how effective they are in enhancing intelligibility.

Finally, the Claim simplification problem was also studied for the Japanese language. In particular, Shinmori et al. [165] propose a method to expose patent structure using manually-defined cue phrases and explain invention-specific terms using the Description [164]. In [163], Description chunks are used to paraphrase corresponding sentences in the Claim and improve readability.

## 2.7 Approaches for Patent generation

The task of Patent generation has recently been investigated by Lee and Hsiang, which try to leverage state-of-the-art NLP models to generate patent text. Table 2.5 reports their main results.

Their early work [99] fine-tunes GPT-2 – a language model which demonstrated unprecedented results in generating text from a wide range of domains – using patents' first claims. Interestingly, only a small number of fine-tuning steps are sufficient to adapt the general domain model and produce patent-like text. However, the quality of the generation is not measured. This gap is partially filled in [100], where a BERT classifier is used to measure if two consecutive spans, generated automatically, are consistent. They train the classifier on consecutive spans from the same patent (positive examples) and from

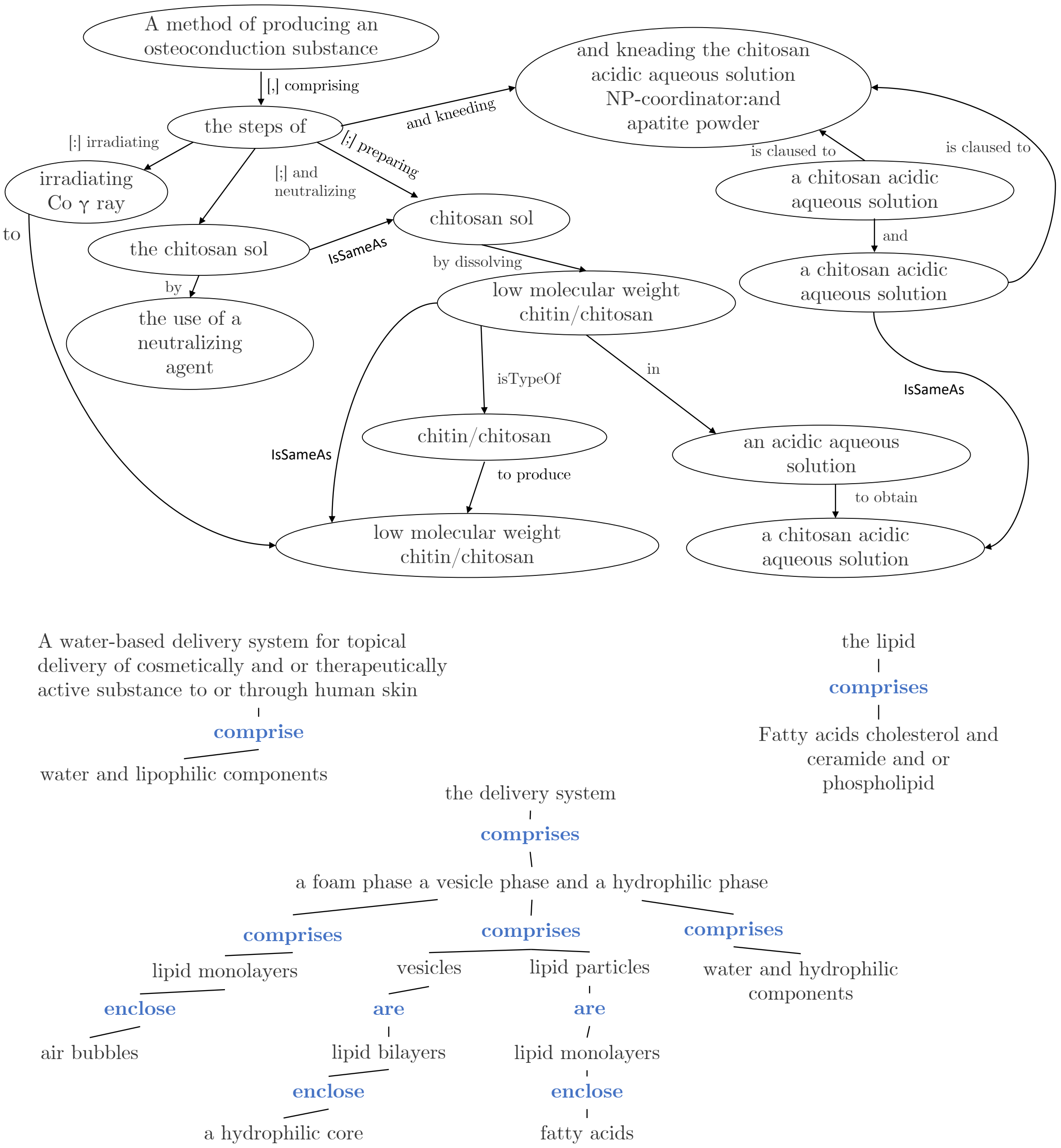

A method of producing an osteoconduction substance
[,] comprising
the steps of
and kneeding
and kneading the chitosan acidic aqueous solution NP-coordinator:and apatite powder
is claused to
is claused to
a chitosan acidic aqueous solution
and
a chitosan acidic aqueous solution
[:] irradiating
[:] preparing
[;] and neutralizing
irradiating Co γ ray
to
the chitosan sol
IsSameAs
chitosan sol
by dissolving
by
the use of a neutralizing agent
low molecular weight chitin/chitosan
isTypeOf
in
IsSameAs
IsSameAs
chitin/chitosan
to produce
an acidic aqueous solution
to obtain
low molecular weight chitin/chitosan
a chitosan acidic aqueous solution
A water-based delivery system for topical delivery of cosmetically and or therapeutically active substance to or through human skin
comprise
water and lipophilic components
the lipid
comprises
Fatty acids cholesterol and ceramide and or phospholipid
the delivery system
comprises
a foam phase a vesicle phase and a hydrophilic phase
comprises
comprises
comprises
lipid monolayers
vesicles
lipid particles
water and hydrophilic components
enclose
are
are
air bubbles
lipid bilayers
lipid monolayers
enclose
enclose
a hydrophilic core
fatty acids

FIGURE 2.5: Top: connected graph for visualizing a patent claim, adapted from [8]; bottom: diagram of a claim, adapted from [160].

| Study | Approach | Main contribution | Limitations | Dataset |
|---|---|---|---|---|
| Lee and Hsiang [99] | GPT-2 fine-tuning | Adaptation of a general-domain LM to patent text | Evaluation | 555,890 patent |
| Lee and Hsiang [100] | Span-pair classification (BERT) | Automatic evaluation of generation relevancy | Negative examples can have unrelated vocabulary | 14M span pairs |
| Lee [98] | GPT-2 -based | Conditional generation of patent Sections | | Google Patents Datasets (1976 2017-08 Utility patents) |
| Lee and Hsiang [101] | Similarity and reranking | Ranking of most similar training samples to the generated text | Mixed results | Huge |

TABLE 2.5: Surveyed studies for Patent Simplification.

non-overlapping classes and subclasses (negative examples), which might make the classification not particularly difficult (e.g., the model could relay in shallow lexical features). The generation process is further investigated in [101], which, given a generated text, tries to find the most similar example in the generator's fine-tuning data.

The models described above try to generate consistent text resembling a patent without specific constraints. A different approach is explored in a following work [98], where authors train the model to generate a patent's Section (Title, Abstract, or claims) given other parts of the same patents. The model uses GPT-2, which receives as input the text on which to condition and learns to produce a section of the same patent accordingly. For example, one can input the Title of a patent and train the model to generate the corresponding Abstract. Two things should be noted: first, the authors frame the problem as self-supervised and use patents' sections as gold-standard, which simplifies evaluation; second, the problem generalizes abstractive patent summarization so that it might be interesting to study the performance obtained, e.g., generating the Abstract from the Description.

## 2.8 Current and future directions

This chapter aimed to show that patents are an interesting domain both for their practical importance and their linguistic challenges. While generative approaches for patents are still relatively niche topics, with few active groups, the domain is drawing attention from general NLP practitioners for its unique characteristics.

In the following, we present some open issues which might be worthy of future research.

**Data, data, data** Labeled and annotated data are few in the patent domain. For summarization, the first available large-scale benchmark was BigPatent [158] (published in 2019), while no simplified corpus (let alone parallel corpora) exists, to the best of our knowledge. Moreover, while BigPatent represented a milestone for patent

| Task | Input → Output | Evaluation | Challenges |
|---|---|---|---|
| Summarization | Patent or Section → Summary | Human evaluation, ROUGE, $F_1$, compression, retention ratio | • Long input<br>• Long sentences<br>• Spread content<br>• Factuality |
| Simplification | Patent text (usually Claim) → Simplified text, visual interface | Human evaluation | • Maintain legal scope<br>• Lack of simplified data |
| Generation | None, seed or Section → Patent text | Human evaluation, ROUGE | • Peculiar language<br>• Domain mismatch<br>• Evaluation |

TABLE 2.6: The tasks described in this chapter and their challenges in the patent domain. In addition, all tasks are challenged by the patents' peculiar linguistic characteristics described in Section 2.1.

summarization, the target Abstract is written in the typical arcane patent language; thus, the practical usefulness of systems trained on these data is probably scarce for laypeople – who would rather read a "plain English" abstract, like those provided by commercial companies. A dataset that targets a clearer summary (unifying summarization and simplification) would also help in understanding models' capabilities in going beyond shallow features and having a global "understanding" of the source. Finally, while no public corpora of simplified patent text exist to date, other domains have exploited creative ploys to minimize human effort: in the medical domain, for example, Pattisapu et al. [141] uses social media content to create a simplified corpus.

**Benchmarks** There are many approaches to summarization and simplification. However, it is difficult to compare them given the absence of shared benchmarks. For extractive summarization, for example, many studies have only compared their results with a baseline or a general-domain commercial system. However, directly comparing the performance of different approaches is difficult, as they solve slightly different tasks on different datasets and often fail to report implementation details. Even models trained on BigPatent cannot be compared directly, as we will explain in Chapter 4.5.

**Evaluation metrics** Generative approaches for patents often resort to general-domain metrics for evaluation (e.g., ROUGE). However, it is not clear how suitable these measures are for the patent domain, given its peculiarities. In the context of abstractive summarization and patent generation, some works [172, 98] highlight that ROUGE is unable to find semantically similar sentences expressed in different wording. In the context of Natural Language Generation, some new measures have been proposed to solve these issues. BERTScore [213], for example, evaluates the similarity among the summary and gold-standard tokens instead of their exact match, while QAGS [195] uses a set of questions to evaluate factual consistency between a summary and its source (a reference is not needed). It is yet to be explored if these metrics could be applied to the patent domain successfully. Finally, note that even human studies are difficult in the patent domain, as they require high expertise, which most people lack.

**Factuality** While neural abstractive models have shown impressive performance in summarization, they tend to fabricate information. Cao et al. [26] studied the phenomenon in the news domain and found that around 30% of documents included fabricated facts. This behavior is particularly problematic in a legal context; ROUGE, however, is a surface metric and is unable to detect factual inconsistencies; model-based metrics, on the other hand, might need to be fine-tuned or adapted to work properly in the patent domain.

**Domain adaptation** Patents' language hardly resembles general-discourse English (used in models' pre-training), but the domain adaptation problem has not been studied in detail. Among the previous works, Aghajanyan et al. [1] propose a second multitask pre-training step, Chen et al. [35] studies models cross-domain performance and Fabbri et al. [48] evaluates zero- and few-shot settings; all these works described applications to the patent domain, among the others.

**Input length** Patent documents are extremely long. For summarization, there are few datasets with comparable or longer inputs, for example, the arXiv and the PubMed dataset [38], which summarize entire research papers. While solutions to allow the processing of long inputs have been proposed, the in-depth study of methods and performance for such long documents is still in its early days. For neural models, a very long input translates into prohibitive computational requirements (e.g., several GPUs), which researchers have recently tried to mitigate by modifying the underlying architectures.

# Chapter 3

# Automatic Text Summarization

This section introduces Automatic Text Summarization. After briefly describing the task (Section 3.1), we classify summarization systems based on several dimensions (Section 3.2), describe relevant datasets (Section 3.3), and explain how candidate summaries are evaluated (Section 3.4).

## 3.1 The task of Automatic Text Summarization

Automatic text summarization is the task of automatically creating a summary of one or more documents. A precise definition of the task is difficult to formulate.
Hovy and Lin [68] complain *"no-one seems to know exactly"* what a summary is, and provide their definition:

> *"A summary is a text that is produced out of one or more (possibly multimedia) texts, that contains (some of) the same information of the original text(s), and that is no longer than half of the original text(s)."*

The definition above is clearly only one of the possible ones. Putting aside the hubris of a formal comprehensive definition, we will describe a good summary as a piece of text that condenses the central ideas of its source documents and does so minimizing repetition. A summary aims at being concise, coherent, fluent, easy to read, and consistent with the original document while being informative and having good content coverage.

Text summarization is one of the most challenging tasks in Natural Language Processing. However, the task is relatively old: in 1958, Luhn described automatic methods to extract abstracts from technical articles [115]. The method computed sentences' "significance" based on their words frequencies and relative position and can be considered the ancestor of modern extractive methods.

While we will focus on text only, automatic summarization can be performed on many media, e.g., speech [196], videos [9, 85] or even source code [62, 162], and genomic data [197]; exploring non-standard sources and jointly summarizing from many modalities [208] is currently an active and exciting direction of research.

## 3.2 Summarization methods

In the following, we will describe a brief taxonomy of summarization methods and systems.

### 3.2.1 Extractive and abstractive summarization

Suppose a student wants to summarize a research paper. There are two main approaches they can take: some people prefer to directly highlight key points in the original text: when revising, this approach allows them to skim the paper and only read the marked text; in this way, a set of sentences from the original text become a "summary" of the whole article. Other people prefer to take notes: they read the whole document, try to understand it, and write down a new text containing the key concepts from what they read. These two ideas roughly correspond to two classes of summarization algorithms: extractive and abstractive methods.

In the following, we will formalize each of the two categories and present the most important baseline and state-of-the-art methods.

#### Extractive summarization

Extractive summarization is the task of automatically creating a summary of a text by selecting a subset of fragments from the original source.

More formally, given a document $\mathcal{D}$ composed by $|N|$ fragments

$$\mathcal{D} = \langle n_0, \ldots, n_i, \ldots, n_{|N|} \rangle$$

we what to obtain a summary $\mathcal{S}$ composed by $|M|$ fragments $m_i$, which are in the original document.

$$\begin{gathered} S = \langle m_0, \ldots, m_j, \ldots, m_{|M|} \rangle \\ |M| < |N| \\ m_i \in D \end{gathered}$$

Generally, full sentences are chosen as fragments, so we will loosely use the word *sentence* in place of *fragment* in the following; note, however, that some methods extract full paragraphs [75, 128] or use a different granularity.

Thus, in a nutshell, an extractive summary is a text obtained by the concatenation of representative sentences extracted from the source.

Extractive summarization is commonly modeled as a classification or a ranking task. Specifically, after a possible preprocessing step, the document is transformed into a useful representation, and its sentences are ranked according to a score (possibly taking into account relationships among fragments); top-ranked sentences are then extracted and possibly post-processed.

### Methods in extractive summarization

Without any aim at completeness, in this section, we will describe some summarization methods that we will use in the following of this thesis.

**Graph-based methods** The core idea of this "classical" class of methods is to represent the original document $D$ as a graph having sentences as nodes and their similarity as edges. For example, **TextRank** [124] uses the number of shared words among two sentences, normalized by the length of the sentences, while **LexRank** [46] uses the cosine similarity of their Term Frequency–Inverse Document Frequency (TF-IDF) representation. Edges in the complete graph are then pruned using a threshold, and the most central sentences according to PageRank [139] are extracted.

Recently, embedding-based similarities have been used instead of token-based ones. **PacSum** [216], for example, is a method that makes two main modifications to classical graph-based algorithms: it uses the cosine similarity among BERT-based [42] embeddings as similarity metrics and makes the original undirected graph directed by exploiting the sentence order in the source.

**Other "classical" methods** Many methods represent sentences as a predefined set of explicit features. Typical features derive from the vocabulary (e.g., based on n-grams or TF-IDF), words' casing, number and type of named entities, Parts of Speech, and sentence position, to name a few. Once a numerical representation of sentences is established, sentences are ranked either by heuristics (e.g., predefined weights) or through a learned function of the features (e.g., learned through a supervised binary classification algorithm from a training set).

Some methods also try to minimize the repetition of information explicitly; for example, **SumBasic** [134] performs the extraction in several sequential steps: at each step, one sentence is extracted based on its words probability distribution; then, to minimize repetition, the probability assigned to the words in the chosen sentence is squared (and thus reduced). The process is repeated until the target number of tokens is obtained.

Another class of methods uses techniques that are popular for semantic analysis and topic modeling in the context of text summarization. **Latent Semantic Analysis** (LSA) [73], for example, aims at exploiting the latent semantic structure of the document. The algorithm decomposes the term-sentence matrix constructed from the source document using SVD [88]. The $t \times s$ terms-by-sentence matrix $A$ is thus decomposed as $A = U\Sigma V^T$. Thus, the original matrix is decomposed into a matrix of term distributions over latent topics, a diagonal matrix of topic importance (the singular values), and a matrix of topic distributions across sentences. For each of the $K$ most salient latent topics (i.e., those corresponding to the largest singular values), the sentence with the largest index value is included in the summary [59, 175].

**Transformer encoders for sentence ranking** Pre-trained transformers have shown their potential in a wide range of tasks [27]. They owe their luck to a new architecture [192] based on multi-head attention and to their self-supervised pre-training that improves text representation.

Liu and Lapata [109] were among the first to explore the use of transformers' encoders for extractive summarization. Their extractive system, **BERTSum**, is a BERT-based encoder. Its input sequence is composed of concatenated sentences from the source, interspersed with a special token; the whole document is implicitly represented hierarchically in higher layers of the transformer. The model is trained in a supervised manner as a classifier to simultaneously predict whether each sentence should or should not be included in the summary. At inference, sentences are ranked according to their score, and the top $K$ are typically extracted, where $K$ is chosen in advance. The model is often used as a state-of-the-art baseline for extractive summarization.

Zhong et al. [217] proposed **MatchSum**, which tries to overcome the common paradigm for extractive summarization. They argue that sentence-level systems are suboptimal since they are prone to redundancy and do not consider the semantics of the summary as a whole. In contrast, they propose a summary-level approach: they compose several multi-sentence summary candidates (with a varying number of sentences). Then, they train a Siamese-BERT model using contrastive learning: their core idea is to teach the model to assign higher matching scores to gold summaries (in contrast to other extracted summaries) while also scoring better candidates (as measured by ROUGE) higher than unqualified ones. At inference, they rank candidate summaries according to their matching score.

**Advantages and disadvantages of extractive summarization**

Extractive methods are relatively simple and generally faster than abstractive methods both at training and inference time. Moreover, since they directly extract sentences from the original document, they are factual to the source. The original sentence-level style is also preserved.

However, since the obtained summaries are a concatenation of sentences, they lack cohesion and sound unnatural. No discourse structure is preserved. Pronouns might lose their references, anaphoras might either hang or seem to refer to a wrong antecedent, and temporal expressions be incoherent or misplaced. These problems become even more evident when the summary is extracted from multiple documents. Some of these problems might be attenuated with a post-processing step.

From a content point of view, redundancy usually needs to be considered explicitly, or the extracted sentences might be very similar to each other. Moreover, on the one hand, relevant content is usually spread across sentences and on the other, sentences might contain both important and non-central information – but extractive approaches are generally only able to extract complete sentences.

On a more technical note, since gold-standard summaries are human-written (thus not composed by a subset of sentences), the maximum automatic score an extracted summary can obtain has an upper bound; this bound is given by the set of sentences that maximize the metric with the gold standard, as extracted by an oracle. Depending on the nature of the gold standard, this bound can be relatively low.

### Abstractive summarization

Abstractive summarization systems aim at generating a summary that is not composed of fragments of the original text, similar to what a person would do.

More formally, given a document $\mathcal{D}$ we want to obtain a summary $\mathcal{S}$

$$\mathcal{S} = f(\mathcal{D})$$

where $f(.)$ is a function that generates the summary from the document instead of simply extracting its sentences.

### Methods in abstractive summarization

**Non-neural methods** In contrast to the large amount of work in extractive summarization, less work on generative models for non-extractive summarization has been published until the last decades. For example, Witbrock and Mittal [200] proposed a method to generate a headline-style summary that can be shorter than any sentence in the original source. Knight and Marcu [89] focused on sentence compressions and used the sentence tree to remove subsets of words. Other methods relied on learning human-like transformations to turn the source into the summary. For example, Jing and McKeown [77] proposed a Hidden Markov Model [15] to divide sentences in the source and later used them to generate a summary; somewhat similarly, Jing and McKeown [76] proposed a method that mimics human-like operations (e.g., deletion and merging) on sentences and their fragments while Saggion [151] learned a set of transformations to turn the source into the summary. Other systems used linguistically-motivated compression techniques and unsupervised topic detection [210] or non-neural machine translation-like algorithms based on term selection and ordering [13].

**Neural methods** Most recent works learn $f(.)$ through a neural network, specifically a sequence to sequence (seq2seq) architecture [182]. The network uses the tokens in the document as input and is taught to generate the gold standard's sequence of tokens, usually in an autoregressive way. A popular training approach uses the Maximum Likelihood Estimation (MLE) framework. Thus, the parameters $\theta$ of the neural network are estimated as:

$$\theta^* = argmax_\theta \sum_i logp_{f\theta}(\mathcal{S}^i_{gold}|\mathcal{D}^i)$$

where $\theta^*$ are the estimated parameters, $p_{f\theta}$ is the probability distribution entailed by $f$ with parameters $\theta$. $\mathcal{S}^i_{gold}$ and $\mathcal{D}^i$ are the ith reference summary and the ith document in the training set, respectively.

For a specific pair of training documents and references, this is equivalent to minimizing the sum of the negative loglikelihoods of the tokens in the reference summary. The training loss for each training example is thus:

$$\mathcal{L} = -\sum_{j=1}^{|T_{gold}|} \sum_s p_{correct}(s|\mathcal{D}, \mathcal{S}_{gold_{<j}}) logp_{f\theta}(s|\mathcal{D}, \mathcal{S}_{gold_{<j}}; \theta)$$

where $|T_{gold}|$ is the total length of the reference summary $S_{gold}$ in terms of tokens, $S_{gold_{<j}}$ is the reference summary up to token $j$, and $\mathcal{S} = \langle s_0, \ldots, s_j, \ldots s_T \rangle$ is the generated summary.

$p_{correct}$ is one-hot under the standard framework, but label smoothing is generally used:

$$p_{correct} = \begin{cases} 1-\beta & \text{if } s = s_j^* \\ \frac{\beta}{|V|-1} & \text{if } s \neq s_j^* \end{cases}$$

where $\beta$ is a small positive number and $|V|$ is the size of the vocabulary.

Decoding is generally autoregressive, with a new token being generated conditioned to the source and to the previously generated tokens.

$$p_{f\theta} = (s|\mathcal{D}, \mathcal{S}_{<j}; \theta)$$

Since enumerating all the possible summaries is intractable, decoding methods [211] (e.g., beam search) are used to define how to handle the search space over potential output tokens when generating a candidate summary.

Rush, Chopra, and Weston [150] were the first to use neural networks for abstractive summarization and sentence compression. Their model learned the probability of the next tokens using a standard feed-forward language model and an encoder that provides a representation of the source. They experimented with several encoders (Bag of Words-based, convolutional, and attention-based, which was shown to perform best).

Other seminal works relied on **Recurrent Neural Networks**. Recurrent Neural Networks are a class of neural networks that allow previous outputs to be used as inputs through hidden states and were very popular with sequential inputs, including text. Chopra, Auli, and Rush [36], for example, improved over [150] by replacing the feed-forward decoder with a recurrent neural network in order to generate arbitrarily long contexts. Nallapati et al. [130] used Recurrent Neural Networks both as encoders and as decoders, with encoder-decoder attention, and proposed modifications to the standard machine-translation model to account for summarization-specific requirements. Their decoder also adopted the pointer mechanism to deal with rare and Out Of Vocabulary (OOV) tokens and relied on hierarchical attention to deal with longer documents.

**Convolutional Neural Networks**, a class of neural networks that extract meaningful sub-structures from a structured input (a 1-dimensional array, in the case of text), were also used for the task. For example, Narayan, Cohen, and Lapata [132] proposed a topic-conditioned neural model which is based entirely on convolutional neural networks: they use a convolutional encoder and a convolutional decoder and assume to know the word and document topic distributions (which they obtain through Latent Dirichlet Allocation).

See, Liu, and Manning [157] introduced the **Pointer Generator** network for summarization. The pointing mechanism allowed the model to "point" at individual tokens (and then copy them) to improve the factuality of the summary and further diminish problems with Out Of Vocabulary tokens. They also introduced coverage to avoid repetition: the coverage vector stores how much attention to each word in the source document already received, and discourages the network from further attending to tokens that have already been covered.

Most of the current state-of-the-art models are based on pre-trained encoder-decoder **Transformers** [192].

An example of this class of models is **BART** [102], which uses a standard transformer-based architecture. BART's training procedure adopts the standard pre-training, fine-tuning paradigm: during pre-training, the input text is corrupted through some noising functions, and the model has to reconstruct the original text. The noising functions include random token masking, random token deletion, text infilling, sentence permutation, and sentence rotation. Its pre-training is not tailored to summarization but is compatible with many sequence-to-sequence tasks. After the self-supervised pre-training, the model can be fine-tuned with a summarization dataset.

**Pegasus** [212] has an architecture similar to that of BART but a different pre-training objective specific to summarization. In Gap Sentences Generation (GSG), sentences are masked, and the model has to learn to generate them according to the context, which is a closer task to abstractive summarization.

Finally, standard transformers are limited when handling long texts: BART, for example, has a standard maximum input length of 1024 subtokens. In fact, standard attention is quadratic (in time and memory) in the input length, thus making it difficult to use these models for long sources. Interestingly, a new class of transformers aims at solving the issue by using non-quadratic attention. To this end, the attention architecture has recently been modified to deal with long sentences. Examples of these models include the **Longformer** [16] and **BigBird** [209], to name a few.

**Advantages and disadvantages of abstractive summarization**

Abstractive summarization aims at solving a very complex task – similar to what most people understand as human-like summarization. Since the summary is generated as a stand-alone new piece of text, concepts in different parts of the original document can be fused; since no constraints on extracting full sentences exist, the summary can contain only the important aspects. Grammaticality and redundancy reduction are generally dealt with intrinsically.

However, abstractive methods are generally more computationally expensive than extractive ones. Their performance closely depends on the richness of their internal representation – they cannot summarize what their representations can not capture.

Moreover, since the summary is based on generation and not on information selection, generated summaries are not always factual to the source. They might, in fact, contain hallucinations, i.e., information that is not present in the source or that directly contradicts it.

#### Hybrid summarization

Some approaches to summarization are based on ideas from both the extractive and the abstractive worlds.

The most common approach to hybrid summarization is extractive-to-abstractive. The model extracts relevant sentences first and then blends them together and paraphrases them through an abstractive step. This approach could be useful, for example, in the case of very long sources, for which using an abstractive model directly is suboptimal [70]. Using extractive fragments to guide the abstraction has also been found useful to minimize hallucinations [144].

Conversely, an abstractive-to-extractive method uses abstraction to improve the extraction of sentences. Huang et al. [70], for example, first generate salient text snippets from groups of sentences in the source using T5 [148], and then extract the sentences that are more similar to those snippets.

### 3.2.2 Single-document and multi-document summarization

While most of the current research focuses on single-document summarization (i.e., cases in which the summary must be obtained based on a single source document), multi-document summarization is becoming more popular. In this case, a single summary must be obtained based on a set of source documents, e.g., to give a brief digest of documents on the same topic or news on the same event. Multi-source summarization models can be used to obtain a more rounded perspective of facts (e.g., for news) or to summarize knowledge on some topic (e.g., for automatic literature review). The task comes with its own challenges, e.g., how to represent multiple sources, how to deal with conflicting information, or how to order the extracted information.

### 3.2.3 Generic and query-based summarization

Standard summarization produces a short version of a source without a focus on any specific aspect. In contrast, query-based models receive a query in addition to the source and summarize information of relevance to such query. This is useful if one is only interested in specific aspects of the original document. For example, given a Wikipedia article on the *The Hitchhiker's Guide to the Galaxy* and the query "towel", a query-based system should produce a summary describing why a towel is the most important item a Hitchhiker can carry.

### 3.2.4 Independent summarization vs. summarization for a downstream task

Most summarization research focuses on generic summarization, and the document is summarized per se. However, summarization can also be used as a tool to improve performance on other tasks: during the TIPSTER Text Summarization Evaluation (SUMMAC) [118], for example, subjects were asked to judge if a document were relevant to a topic; in a first setting they had to read the full document, while in a second setting they used an automatically generated summary. Results revealed that summarization is very effective in the relevance assessment of news.

While summaries are typically intended for humans, producing a shorter dense representation might also be relevant for other automatic tasks. For example, Sakai and Sparck-Jones [152] studied generic summarization for information retrieval. They found that indexing based on automatically-generated summaries can be better for precision-oriented search than using the whole document. Summarization can also be useful when the input is too long or contains too much noisy information to be processed directly, e.g., by another machine learning algorithm. In all these cases, summarization constitutes a building block of a more complex pipeline.

Note that when a summarization system is intended in a downstream fashion, it should be evaluated extrinsically – i.e., how useful it is on the downstream task – rather than with the usual intrinsic metrics.

## 3.3 Datasets

Manually creating a large-scale dataset of sources and summaries is a very hard, expensive, and time-consuming task. Historically, researchers have thus explored textual venues where documents are naturally accompanied by a piece of text that "summarizes", in some sense, the original source.

In practice, this might mean many different things. One classical dataset for summarization is the CNN/DM dataset [130], which exploits online news articles; the webpages that were scraped to generate the dataset also displayed "highlights", i.e., a short list of stand-alone sentences containing one core piece of information in the articles each. The highlights were concatenated and used as the summary.
Another example in the news domain is XSum (Extreme Summarization) [132]. The dataset is constructed from BBC articles and uses a one-sentence summary answering the question "What is the article about?" as target.
In the case of Gigawords [150], the summarization task is actually a header generation task: given the first sentence of the article, generate its headline.
Many other summarization datasets are in the news domain, e.g., the New York Times dataset (NYT) [153], and the NewsRoom [60] dataset, to name a few. In general, these datasets have some characteristics one should be aware of: the input text is generally short (one news article), and news articles are written with an "inverted pyramid" schema, such that the most relevant information is generally at the very beginning of the article.

Considering other domains, scientific papers are also popular for the summarization task. In this approach, either the full article text or some of its sections (usually the Introduction) are used as input, and its abstract is used as the target summary. Scientific articles are longer documents and suffer less (or rather differently) from the positional bias news article present. Examples from this set of datasets in the medical domain are ArXiv and Pubmed [38], while PeerRead [81] contains papers in the computer science domain.

Many other summarization datasets in English exist, e.g., Reddit TIFU (TL;DR) [86], wikiHow [92], and BillSum [91].

The vast majority of summarization datasets only present a single reference. A notable example is the DUC (Document Understanding Conference) dataset [210]. The DUC2004 Task 1 dataset consists of 500 news articles paired with four human-written references. The dataset is small and is thus mainly used for testing.

Human-written datasets are, by their nature, abstractive in the sense that they do not contain extracted sentences but are rather novel pieces of text. To train an extractive model, the most common approach is thus to select the sentences that collectively maximize the ROUGE score (that we will discuss in the next session) with the human-written summary.

While there is a plethora of datasets in English, those in other languages are fewer.
To contribute to the Italian community, we created WITS [30], a large-scale dataset for abstractive summarization in Italian built by using Wikipedia articles as sources and their leads as targets. Please see Appendix A for details.

## 3.4 Measuring summarization quality

Measuring the quality of a summary is currently an open problem in Natural Language Processing [32, 111]. This is mainly because summarization is an open-ended task: a document can be summarized in different ways, all of which can be "good" from different perspectives. Therefore, evaluation performed by humans remains the gold standard – despite its limitation. However, human evaluation is expensive, and automatic metrics are practically used to understand progress and performing automatic system optimization. They can be roughly divided into:

**Untrained reference-based metrics** Given a gold-standard, untrained reference-based metrics compute the similarity among the candidate summary and the reference using a known (untrained) representation with a known (untrained) similarity function. Token-based measures belong to this class: the most prominent example is ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [105], a package of metrics for the evaluation of automatic text summarization.

ROUGE-N is n-gram based and is measured as:

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{Reference}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \text{Reference}} \sum_{gram_n \in S} Count(gram_n)} \tag{3.1}$$

Another flavor of ROUGE is the Longest Common Subsequence (LCS). Words of the Longest Common Subsequence must appear in the same relative order but not necessarily be contiguous. The metric can be computed either on the whole summary (this is normally referred to as ROUGE-L) or on individual sentences (the mean score over individual sentences is normally referred to as ROUGE-LSum).

ROUGE-1, ROUGE-2, and ROUGE-L/ROUGE-LSum are generally used in practice, as they best correlate with human judgment. The higher the values, the better the scores.

**Model-based reference-based metrics** Model-based metrics compute the similarity between the reference and the generated summary either by using a learned representation or through a learned similarity function.

An example of this class of models is BERTScore [213]. It uses contextual embeddings to obtain a contextual representation of tokens; after computing the cosine similarity among each pair in the generated sequence and the gold standard, the maximum similarities over the gold-standard tokens (Recall) and the generated tokens (Precision) are summed, and normalized; they are later used to compute f1-like metrics. Using an embedding-based representation allows going beyond the token-matching approach, which, for example, is not robust to the use of synonyms. The higher the value, the better the score.

**Reference-free metrics** Writing or obtaining gold standards is a complex and expensive task. This is why a line of research has tried to evaluate the summary correctness without a reference.

For example, some previous work has investigated the use of similarity metrics (that are normally used in a reference-based context) among the source and the generated summaries directly. Louis and Nenkova [112], for example, explored information-theoretic metrics (e.g., the Kullback–Leibler divergence [78] between the vocabulary distributions of the input and of the generated summaries), vector similarity (e.g., the cosine similarity on the TF-IDF representations of the input and the generate summaries), topic-related metrics and combinations of these similarities. They also found that the quality of a candidate summary can be estimated (with a high correlation to human judgment) by measuring its similarity to a set of automatic-generated summaries.

Another dimension often measured in a reference-free manner is factuality. For example, FacCC [93] used a model trained to directly score if a summary is factual and to extract spans from the source and/or the summary to support the consistency claim. Wang, Cho, and Lewis [195] took a different approach and proposed a metric based on question answering: they automatically generate relevant questions and compare the answers obtained from the whole source and the generated summary: the core idea is that the same question should be answered in the same way from both if the summary is factual. DeYoung et al. [43] propose to quantify factuality with a domain-specific approach: in the contest of summarization of medical papers, they measure the disagreement of (Is, Os, EI) triplets between the input documents and the generated summary, where Is are the Interventions, Os are the Outcomes and EI is Evidence Inference.

While automatic summarization metrics are useful as a proxy of the quality of the summary, the gold-standard remains human evaluation, where annotators are used as judges of the summary quality. Depending on the task, lay people or experts might be necessary for judgment.

Practically, a group of humans is asked to judge the quality of a summary, often using a Likert scale [104]. The evaluation can be either on an overall dimension or on several dimensions (e.g., fluency, quality of the information extracted, factuality, etc.). Another option is to ask participants to rank two or more summaries. While regarded as optimal, researchers have highlighted that human evaluation results are often difficult to reproduce (sometimes for the lack of details in published research), and comparability is an issue [69].

Recently, online crowd-sourcing platforms (e.g., Amazon Mechanical Turk or Prolific) have been used for large-scale experiments; however, the quality of annotation is not always easy to control, and a large number of annotators might be necessary.

Finally, semi-automatic metrics also exist. Pyramid [133], for example, compares the generated summary with the human-written ones using Summary Content Units (SCUs). From a set of model summaries, the authors manually identify similar sentences, use them to create Summary Content Units, and rank them in a pyramid model. A summary is considered appropriate if it contains a large number of higher-level Summary Content Units.

In case the summary is not aimed at humans but rather at improving another Natural Language Processing task, the summarization should be evaluated extrinsically, e.g., by measuring the relative improvement of the downstream task.

# Chapter 4

# The BigPatent Dataset(s)

This Section describes our analysis of the BigPatent dataset and the issue related to its versions and experiment reproducibility. In fact, the dataset exists in two versions, so different in their content and characteristics that they should be considered as two different datasets.

In this chapter, we analyze the two versions of the dataset and how previous work has ignored their differences, making it impossible to directly compare with the reported results. Starting from these considerations, we also explain our choice for the dataset that we will use in the rest of this thesis.

This chapter is based on Casola, Lavelli, and Saggion [31].

## 4.1 BigPatent: what's in a (dataset's) name?

Sharing models and datasets is essential for Natural Language Processing.

With the rise of transfer learning in the last few years, releasing large pre-trained models has become standard practice. Consequently, several libraries have provided Application programming interfaces (APIs) to access and work with those models efficiently. Datasets have followed a similar trend: they are often shared by their authors and stored in hubs that expose APIs. Two notable examples of this trend are the TensorFlow Datasets collection[1] and the Hugging Face dataset library[2] [103]. These libraries allow accessing published data, often with just a few lines of code. They drastically ease the experimentation loop, and allow users to download, experiment with, and probe existing resources. There is, however, another side to the coin: the dataset documentation is sometimes insufficient, which might lead to inconsistencies when performing experiments and comparing results to previous work.

The BigPatent [158] dataset is an extreme example of this issue.

BigPatent was first published in 2019. Patents have many peculiar characteristics that might be challenging for standard Natural Language Processing systems: they span multiple pages, have very long sentences, contain a mix of legal and technical vocabulary, and are built out of noun phrases instead of clauses, with a long lexical chain [29], as

[1] `https://www.tensorflow.org/datasets` (Last accessed: February 2023)

[2] `https://huggingface.co/docs/datasets/` (Last accessed: February 2023)

explained in Chapter 2. Given its challenging characteristics, the dataset has also become popular as a general benchmark for summarization.

However, the two popular TensorFlow and Hugging Face dataset hubs expose different versions of BigPatent. These differences are not only superficial (e.g., casing, tokenization) but regard the very content of the source documents.

In this chapter, we first briefly describe this difference and its impact on the dataset features (Section 4.2); then, we examine previous work and show it hardly ever clarifies the version of the dataset used in experiments (Section 4.3); we also show how the difference substantially impacts models' performance (Section 4.4). Finally, starting from these considerations, we explain and justify our choice for the dataset we used for patent summarization (Section 4.5).

While strongly advocating for resource sharing and infrastructure that make them easier to use, we hope that the discussion of this extreme case can shed light on the importance of careful resource documentation.

## 4.2 The BigPatent dataset

Patents are structured legal documents containing several sections.

The Description section reports the technical characteristics of the invention and its preferred embodiments so that a person skilled in the art can understand and reproduce it. The Description can be further divided into subsections (e.g., Background, Field of the Invention, Summary of the invention, Detailed Description, Description of the Drawings, etc.). The patent document also contains a human-written Abstract. It is thus somewhat natural to construct a summarization dataset using the Descriptions (or part of them) as the source texts and the Abstracts as the gold-standard summaries.

The dataset is not only interesting for a niche of patent mining researchers: in fact, patent documents show several linguistic characteristics worth investigating (e.g., long sentences, unusual vocabulary, specific syntactic structure). Moreover, the dataset contains data from a domain that is not covered by the previously available corpora, with challenging characteristics. For example, patent documents are very long, and their Abstract is not very extractive with respect to the Detailed Description, as the original dataset paper shows [158].

In its original version, published by BigPatent's authors and accessible on GitHub[3], only part of the Description (typically the Detailed Description) is included in the input document, and the source does not contain any of the other subsections. The published dataset is also cased and tokenized. The Hugging Face dataset library exposes this version of the dataset (described in the original dataset description paper)[4].

With the advent of sequence-to-sequence transformer models for summarization (e.g., BART [102] or Pegasus [212])), however, using a strongly preprocessed dataset is not ideal. It is common practice to process the raw text with a model-specific tokenizer. This

[3] `https://evasharma.github.io/bigpatent/` [Last accessed: January 2023]

[4] `https://huggingface.co/datasets/big_patent` [Last accessed: January 2023]

is likely why the TensorFlow Datasets collection contains a different version of the dataset that is cased and untokenized, with limited preprocessing over the original raw text[5].

However, a deeper look at the data reveals another difference: the TensorFlow source documents contain a superset of the text in the original version. All subsections in the patent Description are included. Thus, the input not only contains the Detailed Description but often also the Background, the Field of the invention, and other subsections (refer to Figure 2.1); interestingly, the Summary of the invention[6] is also present. The Summary of the Invention is a condensed description of the invention, worded in a way that is, in general, different from the Abstract but has a great overlap in its content. We will refer to this summary included in the document (input) as Summary of the Invention and to the dataset gold-standard as Abstract or gold standard. Table 4.1 shows the first tokens of the input of some entries in the corpus.

In the following, we compute some statistics on the two dataset versions and their different characteristics. We call the original version $BigPatent_{Original}$ and the subsequent modified cased version $BigPatent_{New}$.

The dataset is divided into several subsets, following the Cooperative Patent Classification (CPC) codes. Due to the large dataset size (over 1.3 million examples), we restrict our analysis to its G (Physics) subset, which includes patents of information systems devices and processes; however, our considerations are general.

### 4.2.1 Dataset characteristics

Table 4.2 reports some statistics[7] over BigPatent/G. Note that the dataset split is identical in the two versions (i.e., the train, validation, and test splits contain the same documents).

While the summaries characteristics are very similar between the original and the new version (we attribute the difference to errors in the tokenization, since $BigPatent_{Original}$ is pre-tokenized, while $BigPatent_{New}$ is not), $BigPatent_{New}$ clearly contains more text than the original version (38% more tokens, on average, in the training set), and more sentences (68% more, on average, in the training set). The compression ratio (i.e., the ratio between the number of tokens in the source and the number of tokens in the Abstract) is also higher in $BigPatent_{New}$.

To get a closer look at the datasets' abstractiveness, we compute their coverage and density, following Grusky, Naaman, and Artzi [60].

Given a document $D = \langle d_1, d_2, \ldots, d_n \rangle$ where $d_i$ is a token of $D$ and a summary $S = \langle s_1, s_2, \ldots, s_m \rangle$, with $m \leq n$, where $s_j$ is a token in the summary, $F(D,S)$ is the set of their shared fragments (shared sequences of tokens). The extractive fragment coverage measures the proportion of tokens in the summary belonging to an extractive fragment and qualitatively describes how much a summary vocabulary is derivative of a text.

$$Coverage(D,S) = \frac{1}{|S|} \sum_{f \in F(D,S)} |f| \tag{4.1}$$

[5] `https://www.tensorflow.org/datasets/catalog/big_patent` [Last accessed: January 2023]

[6] Note that this difference is not explicitly discussed on the dataset page.

[7] We use NLTK for sentence and word tokenization.

| publication num. | Description$_{Original}$ | Description$_{New}$ |
|---|---|---|
| US-2007088503-A1 | referring now to fig1 and 2 , a service technician visiting a customer service location is provided with a technician input device 2 for receiving and transmitting information related to a disruption or interruption of service at the service location . the input device 2 can be a wireless pc , for example , a laptop , a personal digital assistant ( pda ), a wireless pager or any other device suitable for receiving and transmitting data associated with providing service at the customer service location . [+2858 tokens] | This is a continuation of application Ser. No. 10/445,861 filed May 27, 2003, which is a continuation of application Ser. No. 10/032,853 filed Oct. 25, 2001 and now U.S. Pat. No. 6,772,064. The present methods and systems generally relate to processing and transmitting information to facilitate providing service in a telecommunications network. [+986 tokens] Referring now to FIG1 and 2 , a service technician visiting a customer service location is provided with a technician input device 2 for receiving and transmitting information related to a disruption or interruption of service at the service location. [+2427 tokens] |
| US-2011144953-A1 | in the following , the invention is described in more detail referring to the attached figures by means of exemplary embodiments , wherein same reference signs refer to same components . fig1 schematically shows the system for compensating electromagnetic interfering fields . an object 2 to be protected against effects of the interfering field 1 is permeated by the interfering field 1 . here , the interfering field 1 is assumed to be a gradient field . the amplitude of the interfering field 1 is measured by two real magnetic field sensors 3 , and 4 . the first real sensor 3 provides an output signal right arrow over ( s ) 1 =[ x 1 ( t ), y 1 ( t ), z 1 ( t )], and the second real sensor 4 provides an output signal right arrow over ( s ) 2 =[ x 2 ( t ), y 2 ( t ), z 2 ( t )]. [+1855 tokens] | This application claims benefit under 35 U.S.C. (a) of German Patent Application No. 10 2009 024 826.9-32, filed Jun. 13, 2009, the entire contents of which are incorporated herein by reference.The invention relates generally to a system for compensating electromagnetic interfering fields, and in particular to a system for magnetic field compensation having two sensors and a digital processor. [+16010 tokens] In the following, the invention is described in more detail referring to the attached figures by means of exemplary embodiments, wherein same reference signs refer to same components.FIG1 schematically shows the system for compensating electromagnetic interfering fields. [+1427 tokens] |
| US-4830479-A | referring now to fig1 of the drawings , there is depicted a ray 12 entering the paper plane perpendicularly along an axis z orthogonal to axes x and y . ray 12 is deflected into the paper plane by a mirror 16 which is located at the origin and is oriented upwardly at a forty five degree angle from the paper plane . mirror 16 rotates with an angular velocity $\omega$ around axis z which is in line with the arriving ray 12 . [+1579 tokens] | The invention described herein may be manufactured and used by or for the Government for governmental purposes without the payment of any royalty thereon.At radio frequencies, superheterodyne receivers typically have sensitivities that are orders of magnitude higher than those of direct detection receivers. [+1044 tokens] Referring now to FIG1 of the drawings, there is depicted a ray 12 entering the paper plane perpendicularly along an axis Z orthogonal to axes X and Y. Ray 12 is deflected into the paper plane by a mirror 16 which is located at the origin and is oriented upwardly at a forty five degree angle from the paper plane. [+1380 tokens] |

TABLE 4.1: Some examples from the two versions of the dataset. We report the first tokens from the input in the original version, and the first tokens in the new version of the dataset. Note that the new version might contain many paragraphs before the content of the original input.

| | | $BigPatent_{Original}$ | $BigPatent_{New}$ |
|---|---|---|---|
| # docs (train, val, test) | | 258,935 | 258,935 |
| | | 14,385 | 14,385 |
| | | 14,386 | 14,386 |
| Summary | # tokens (avg) | 123.9 | 121 |
| | | 123.7 | 120.9 |
| | | 124.1 | 121.2 |
| | # sents (avg) | 3.7 | 3.6 |
| | | 3.6 | 3.6 |
| | | 3.7 | 3.7 |
| | sent len (avg) | 44.3 | 43.4 |
| | | 44.2 | 43.3 |
| | | 44.5 | 43.7 |
| Source | # tokens (avg) | 3,959.2 | 5,488.3 |
| | | 3,953.3 | 5,517.5 |
| | | 3,976.8 | 5,501.9 |
| | # sents (avg) | 105.6 | 177.6 |
| | | 105.5 | 178.4 |
| | | 106.3 | 178.3 |
| | sent length (avg) | 42.6 | 31.8 |
| | | 42.6 | 31.8 |
| | | 42.5 | 31.8 |
| compression ratio | | 36.1 | 51.2 |
| | | 36.0 | 51.5 |
| | | 35.8 | 50.9 |

TABLE 4.2: Length statistics on the two BigPatent versions. The number of tokens, sentences, tokens per sentence, and the compression ratio are computed per document and then averaged. The compression ratio is the ratio between the number of tokens in the source and the number of tokens in the Abstract.

| | $BigPatent_{Original}$ | $BigPatent_{New}$ |
|---|---|---|
| Coverage (avg) | 0.87 | 0.95 |
| Density (avg) | 2.40 | 20.8 |

TABLE 4.3: The extractive fragment coverage and the density for the two versions of the dataset. Measures are computed per document and then averaged.

where $|S|$ is the number of tokens in the summary.

The density also takes into account the length of the extractive fragments: the higher the density, the better a summary can be described as a series of extractions.

$$Density(D, S) = \frac{1}{|S|} \sum_{f \in F(D,S)} |f|^2 \tag{4.2}$$

Table 4.3 shows the measures computed for the two versions of the dataset, while Table 4.5 shows their percentage of novel n-grams.

Note that both datasets have relatively high coverage (the increase in $BigPatent_{New}$ might be partially motivated by the increased length of the source). However, the extractive density is an order of magnitude higher in $BigPatent_{New}$, suggesting that the reference summaries are significantly more extractive than the original version.

$BigPatent_{New}$ has also a lower number of novel n-grams in the summary (and the difference with $BigPatent_{Original}$ stays high even when accounting for the length of the source). We attribute this difference to the presence of subsections such as the Summary of the invention, the Background, and the Field of the invention in the input; these subsections already abstract the core features of the claimed invention.

To investigate how similar the Abstract is to each subsection in $BigPatent_{New}$, we compute their ROUGE scores [105] with the summary[8]. We report both ROUGE F1 and ROUGE recall since we want to quantify how much "information in the Abstract" each subsection contains.

$BigPatent_{New}$ includes all subsections in the Description, but it does not include the name of the headers (an uppercase short header in the raw patent text). This is because short sentences (including subsection headers) are removed from the patent text during the preprocessing when generating the dataset. To divide the text into subsections, we take the raw data and regenerate the dataset using the original TensorFlow preprocessing script; however, we remove the portion of the code that gets rid of short sentences and new lines. We use a regular expression to divide the text into subsections and extract their headers. Subsection headers are not normalized. For example, the Background's header might be indicated as "Background", "Background of the invention", etc. Thus, we use a simple key-based method to classify them into 9 groups. The complete pipeline is pictured in Figure 4.1. The keywords we used are in Table 4.4. Note that not all patents include all subsection types.

[8] All ROUGE scores are computed using the Hugging Face version of the metric, with stemming.

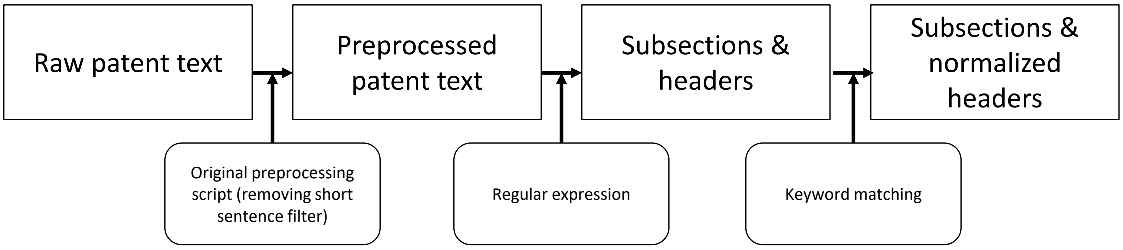

Raw patent text
Preprocessed patent text
Subsections & headers
Subsections & normalized headers
Original preprocessing script (removing short sentence filter)
Regular expression
Keyword matching

FIGURE 4.1: Pipeline to extract and normalize individual subsections and headers from the patent text. We first regenerate the dataset from the original raw data to preserve the section headers, then use a regular extraction and a simple keyword-matching algorithm.

| Section type | Keywords |
|---|---|
| SUMMARY OF THE INVENTION | summary, essence, overview |
| FIELD | field |
| BACKGROUND | background |
| DRAWINGS | drawing, figure |
| EMBODIMENTS | embodiment, example |
| REFERENCES | reference |
| RELATED ART | art |
| OBJECTIVE | problem, object |
| DETAILED DESCRIPTION | description, disclosure |

TABLE 4.4: Keywords used for the normalization of patent subsections. For each patent subsection, we check whether their headers contain any of the keywords (from top to bottom, in the table) and assign the text to the first subsection type for which at least one keyword matches. If no subsection type matches, we classify the patent subsection as "OTHER".

| | $BigPatent_{Original}$ | $BigPatent_{New}$ |
|---|---|---|
| Novel 1-grams (avg) | 10.9% | 4.21% |
| Novel 2-grams (avg) | 46.9% | 23.46% |
| Novel 3-grams (avg) | 74.0% | 42.25% |
| Novel 4-grams (avg) | 87.1% | 53.58% |

TABLE 4.5: Percentage of new n-grams in the summary in the two versions of the dataset. All percentages are computed per document and then averaged.

| | ROUGE-1 | | ROUGE-2 | | ROUGE-L | | #Tokens | % |
|---|---|---|---|---|---|---|---|---|
| | R | f1 | R | f1 | R | f1 | | patents |
| SUMMARY | 84.68 | 35.97 | 60.76 | 25.97 | 69.07 | 29.36 | 744.56 | 93.79% |
| FIELD | 23.62 | 28.66 | 10.17 | 11.92 | 16.14 | 19.44 | 73.73 | 38.27% |
| BACKGROUND | 66.04 | 24.45 | 25.38 | 8.60 | 41.42 | 14.70 | 710.04 | 94.85% |
| DRAWINGS | 38.96 | 28.36 | 10.35 | 7.39 | 24.52 | 17.55 | 243.43 | 97.6% |
| EMBODIMENTS | 81.39 | 8.58 | 42.44 | 4.14 | 59.21 | 5.92 | 3168.25 | 53.07% |
| REFERENCES | 10.82 | 11.40 | 1.48 | 1.35 | 07.38 | 7.94 | 92.10 | 28.18% |
| RELATED ART | 52.47 | 20.33 | 18.48 | 6.36 | 32.13 | 12.04 | 644.27 | 4.12% |
| OBECTIVE | 44.35 | 32.31 | 16.05 | 10.93 | 27.49 | 19.58 | 256.95 | 2.09% |
| DET. DESCRIPTION | 84.39 | 8.27 | 4.10 | 4.08 | 61.90 | 5.78 | 3404.91 | 55.23% |

TABLE 4.6: The ROUGE score (recall (R), f1) between the different subsections of the patents and the patent ABSTRACT. The subsections are obtained from the $BigPatent_{New}$ raw data. The scores are computed per document and normalized by the number of documents that contain each subsection. The average length of each subsection and the percentage of patents that contain the subsection are also reported.

Table 4.6 reports the obtained ROUGE score, the subsection average length, and the percentage of patents that include each subsection type. Note that the Summary of the invention (in 94% of the inputs in $BigPatent_{New}$) has the highest scores; compared to the Detailed Description, the Summary of the Invention has a higher ROUGE-recall even though it is much shorter.

In a nutshell, our analysis shows that the additional text in $BigPatent_{New}$ decreases the need for an abstractive model for the task. The additional Summary of the Invention – which is itself a summary of the rest of the patent – contain the most information in the patent Abstract.

## 4.3 How to compare to the previous literature?

While the two versions of the dataset have different characteristics, the vast majority of previous literature using BigPatent does not explicitly mention the version used.

Zhang et al. [212] mention they *"updated the BIGPATENT dataset to preserve casing, some format cleanings are also changed"* when performing experiments on the Pegasus model; this operation might have led to the creation of the new dataset version now exposed by TensorFlow (whose differences with the original version are, however, not limited to casing and minor format cleaning). Some previous work [65] noticed a substantial performance gap between models trained with the original version and Pegasus and speculated this difference might be due to the different preprocessing (and, we add, possibly to the additional content); these findings are compatible with our experiments in the next Section.

In the vast majority of cases, the reported statistics are directly taken from the original publication and not recomputed; in a few cases, the values computed (e.g., in terms of document lengths) are compatible with the use of the cased version (e.g., in Guo et al. [63]).

BigPatent is widely used when testing systems, generally to measure how well models behave in the case of very long sources. The dataset was cited 136 times, according to Google Scholar[9]. Since the used dataset version is unknown, and authors are unaware of the two different versions, it is impossible to understand if comparing results to previous work is fair. Since the Tensorflow version was updated on the 31st Jan 2020[10], papers published after that date could potentially use the new version of the dataset, with likely better results. In fact, a simple BART model results in a very different performance on the two versions of the dataset, as shown in the next Section.

## 4.4 Experiments

**Dataset full text** To understand if the version of the dataset impacts the models' performance, we fine-tuned a pre-trained BART [102] base model on the two versions of the dataset.

We train using the Hugging Face library with early stopping on the evaluation loss (patience: 5) and the following hyperparameters: max source length: 1000; max target length: 150; number of beams: 5; eval steps: 10k; max steps: 500M. We leave all other parameters to their default values.

Table 4.7 reports the results. Note how results on $BigPatent_{New}$ are more than 11 points of ROUGE-L over $BigPatent_{Original}$.

**Using the Summary of the Invention only** To corroborate the idea that the Summary of the invention in the input improves the performance on $BigPatent_{New}$, we trained a model using, as input, only the text in the Summary of the Invention subsection. In the few cases in which the patent did not include the Summary of the Invention, we used the Detailed Description or the Description of the Embodiments instead. As described in Section 4.2, we resorted to the raw data to extract the text in the Summary of the Invention subsection.

This setting further improves the performance, with an increase of almost 16 and almost 5 points of ROUGE-L with respect to the original and the new version; note, however, that since $BigPatent_{New}$ does not contain the subsection headers, it is not directly possible to train models using the Summary of the Invention only as input.

[9] As of 29/03/2023

[10] See this github commit: a708d506748870237eafa2bbb659dc64cd7cf04a

| | $BigPatent_{Original}$ | | | $BigPatent_{New}$ | | |
|---|---|---|---|---|---|---|
| | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-1 | ROUGE-2 | ROUGE-L |
| Lead-3 | 29.54 | 7.95 | 18.15 | 23.15 | 7.27 | 15.42 |
| Summary Lead-3 | - | - | - | 48.11 | 30.16 | 36.66 |
| BART-base | 42.25 | 15.99 | 27.58 | 50.18 | 29.46 | 38.64 |
| BART-base (Summary) | - | - | - | 55.16 | 34.85 | 43.56 |

TABLE 4.7: Results (test set) on the two dataset versions for a BART-base model. The Lead-3 baseline considers the first three sentences of the input text as a proxy for the generated summary. Summary Lead-3 uses the first 3 sentences of the SUMMARY OF THE INVENTION (obtained from the SUMMARY OF THE INVENTION as described in Section 4.2.1). We also trained a BART model that only uses the SUMMARY OF THE INVENTION as input. The split is identical, i.e., the train, validation, and test splits contain the same documents in both versions.

## 4.5 Choosing the dataset for patent summarization

As we showed above, BigPatent exists in two very different versions. We have shown that the updated version of the dataset lacks some of the original characteristics (e.g., the high level of abstraction in the reference summaries and their high percentage of novel n-grams) and leads to much higher results with a simple transformer.

To our best knowledge, this difference is not reported elsewhere, either in published research or in the dataset's online documentation. In fact, previous work tends to ignore the difference between the original and the new version, making it virtually impossible to understand experimental results, reproduce, and compare them.

Starting from these considerations, the choice of which of the two versions to use for our experiments is not obvious.

We considered two aspects:

- We prefer to work with a dataset that is not tokenized and cased.

  This is in line with most of the current work. The motivation is that cased and untokenized text contains more signal that can be exploited by models. The preprocessing performed when using pre-trained transformers is often minimal and mostly related on the specific format the transformer architechture was pre-trained on. BART [102], for example, uses byte-level Byte-Pair-Encoding [53] and was trained to treat spaces as tokens parts, so that a token is encoded differently whether it is at the beginning of the sentence (without space) or not.

- We believe that performing summarization having the Summary of the Invention in the input is not fair, as motivated by the experiments discussed above.

Finally, we are aware that some domain experts consider the Summary of the Invention (included in the Description) as the most suitable target for the summarization process. However, we prefer to use the Abstract as our gold standard to follow the general rationale behind the BigPatent dataset.

| Dataset | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-LSums |
|---|---|---|---|---|
| BigPatent$_{Original}$ | 41.93 | 15.94 | 27.37 | 25.77 |
| BigPatent$_{New}$ | 48.49 | 28.68 | 37.61 | 43.60 |
| BigPatent$_{New}$ + subsection headers | 51.05 | 30.01 | 39.18 | 46.03 |
| BigPatent$_{New}$ + subsection headers without the Summary of the Invention (ours) | 41.70 | 17.52 | 28.38 | 36.00 |

TABLE 4.8: Result of BART on the validation set for the two versions of the dataset.

Given these considerations, we decided to use the BigPatent version exposed by Tensorflow (that we called BigPatent$_{New}$) but removed the portion of text that can be attributed to the Summary of the invention.

This operation was performed as described above: starting from the raw data, we used a regular expression to divide the text into subsections and extract their headers, classified them, excluded the text that refers to the Summary of the invention, and reconstructed the dataset. Differently from the BigPatent$_{New}$ dataset, we kept the subsection headers as we will use them in our experiments. Doing so, we obtain a dataset that is cased and untokenized but does not contain the Summary of the Invention in the inputs (though it might contain other subsections like the Field of the Invention or the Background).
A manual check of 100 documents processed this way showed that none of them contained the Summary of the Invention in the input.

We also trained a BART model to compare its performance on the different versions of the dataset. Results are in Table 4.8. We train using the Hugging Face library with early stopping on the evaluation loss (patience: 5) and the following hyperparameters: max target length: 250; number of beams: 5; eval steps: 10k; max steps: 500M. We leave all other parameters to their default values. Note that the hyperparameters are slightly different than those used in Section 4.4, hence some minor differences.

We already discussed the differences between the Original and the New version. Here, we add that:

- Adding the headers to the New version gives an advantage to the model over using the same dataset without said headers. This is likely since the new version of the dataset is relatively suitable for extractive summarization and the model might, for example, use them to better locate the content of the Summary of the Invention subsection.
- Removing the Summary of the Invention from the new version of the dataset, we obtain a performance that is very similar to that of the original version. The differences (increase in all metrics except ROUGE-1) might be due to the absence of preprocessing and to the presence of other Description subsections (e.g., the Background) and of the subsection headers in the input.

Concluding, we will always use the above-described version of the dataset in the continuation of this thesis and will always replicate experiments to compare to previous work

| | | *BigPatent*/$G_{Ours}$ |
|---|---|---|
| # docs<br>(train, val, test) | | 258,935 |
| | | 14,385 |
| | | 14,386 |
| Summary | # tokens (avg) | 121.0 |
| | | 120.9 |
| | | 121.2 |
| | # sents (avg) | 3.6 |
| | | 3.6 |
| | | 3.7 |
| | sent len (avg) | 43.4 |
| | | 43.3 |
| | | 43.7 |
| Source | # tokens (avg) | 4,893.6 |
| | | 4,884.5 |
| | | 4,913.8 |
| | # sents (avg) | 161.2 |
| | | 160.8 |
| | | 161.8 |
| | sent length (avg) | 31.3 |
| | | 31.3 |
| | | 31.3 |
| compression ratio | | 45.8 |
| | | 45.7 |
| | | 45.5 |

TABLE 4.9: Length statistics on the dataset versions we will use in the following. The number of tokens, sentences, tokens per sentence, and the compression ratio are computed per document and then averaged. The compression ratio is the ratio between the number of tokens in the source and the number of tokens in the Abstract.

(and use the same version of the dataset when training models) when appropriate. Table 4.9 shows the characteristics of the dataset.

# Chapter 5

# Benchmarking general-domain methods for patent summarization

Our analysis of previous work has shown that while a number of classical approaches have been tested in the patent domain, different datasets and evaluation metrics were used, making a direct comparison infeasible.
In this chapter, we will benchmark extractive, abstraction, and hybrid summarization systems in the patent domain. We will train and evaluate on the BigPatent/G [158] dataset in a modified form, as described in Section 4.5.

## 5.1 Evaluation protocol

Evaluating patent summarization results is challenging. The reason is twofold.
Regarding the automatic metrics, they have known limitations, as we described in Chapter 3. In the patent domain in particular, some previous work [98, 171] has anecdotally questioned the metric validity (and its correlation to expert's opinion), even if no quantitative studies were performed. More complex metrics, e.g., model-based methods, should be fine-tuned with domain-specific data.
Human evaluation is particularly hard in the patent domain for two main reasons:

- The best way to evaluate a summarization output is to read the whole source document. However, patents are extremely long; considering a mean speed of 200 tokens per minute — which is likely optimistic esteem for such a complex text – it would take around 25 minutes to read a 5000 tokens document (roughly the length of a source document in the BigPatent dataset).
- Patent documents and Abstracts are extremely complex and should be evaluated by patent and technical experts; however, hiring such experts is very expensive and unpractical in most scenarios.

Conscious of their limitations, we will use two main methods:

**Automatic evaluation** We will select hyper-parameters and automatically evaluate outputs using ROUGE [105]. Specifically, we will report ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-LSum.
We also conducted some preliminary experiments with factuality-related metrics, e.g., QAEval [195]; the metric assumes the summary is factual if automatically-generated questions are answered in the same ways using the summary or the

source. However, it does not seem to adapt well to the patent domain, and both the question generation and the question answering portions of the model show suboptimal performance. We noticed a similar domain-shift problem with other model-based metrics and decided not to include them in our evaluation.

**Qualitative evaluation** Since we did not have the opportunity to involve patent or technical experts, we will perform a qualitative evaluation of a subset of the candidate summaries. We will mainly consider the patent fluency, consistency, and similarity to the patent Abstract.

For the most promising methods, we will also show a few samples of the generated summaries, together with the patent title and the related gold standard.

## 5.2 Extractive methods

### 5.2.1 Extractive graph-based systems

In this section, we will consider unsupervised extractive graph-based systems, which we described in Chapter 3. Specifically, we will consider two algorithms:

**TextRank** We used the *summa*[1] implementation. In this implementation, the user chooses the target summary length in terms of tokens, and the number of sentences that best approximate that number is extracted. We cross-validated the number of tokens and left any other parameters to their default values.

**LexRank** We used the *sumy* implementation[2]. We validated the number of extracted sentences per patent and left any other parameters to their default value.

The algorithms are unsupervised and tend to work well with long documents. We experimented with PacSum but found it extremely computationally demanding in our use case.

**Automatic evaluation.** ROUGE scores are shown in Table 5.1. As expected, performance is similar for the two systems, with TextRank being marginally superior. Unsurprisingly, the best-performing systems are those that select a number of tokens or sentences similar to that of the gold standard.

**Qualitative assessment.** Table 5.2 shows the outputs obtained by the best-performing TextRank system on the first instances of the test set. Table 5.1 reports the corresponding outputs obtained using the best LexRank system.
We also qualitatively evaluated a subset of 20 outputs.

The outputs obtained using the two algorithms are relatively similar. Technically, we notice that the sentence tokenization is not always perfect: for example, the extracted summary of patent US-2005152022-A1 contains the sentence *"The mixed color display [...] by the type of processes described in the aforementioned U.S. Pat. No."*, where the patent number has been incorrectly considered as a stand-alone sentence. This is in accordance with

[1] `https://summanlp.github.io/textrank/`

[2] https://github.com/miso-belica/sumy

| | Set | #T. | #S | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-LSum |
|---|---|---|---|---|---|---|---|
| TextRank [124] | Val | 50 | | 28.20 | 8.52 | 18.08 | 18.20 |
| | Val | 100 | | 37.06 | 11.40 | 21.99 | 23.33 |
| | Val | 150 | | **38.60** | **12.33** | **22.33** | **24.57** |
| | Val | 250 | | 35.39 | 12.27 | 20.69 | 23.76 |
| | Val | 500 | | 25.74 | 10.37 | 16.11 | 19.30 |
| | Val | 1000 | | 16.22 | 7.65 | 11.00 | 13.43 |
| | **Test** | 150 | | 38.59 | 12.30 | 22.33 | 24.50 |
| LexRank [46] | Val | | 1 | 26.03 | 8.12 | 17.40 | 17.46 |
| | Val | | 2 | 34.72 | 10.93 | 21.14 | 21.23 |
| | Val | | 3 | 37.48 | 12.02 | **21.89** | **21.99** |
| | Val | | 4 | **37.76** | 12.40 | 21.71 | 21.81 |
| | Val | | 5 | 36.92 | **12.46** | 21.16 | 21.26 |
| | Val | | 6 | 35.62 | 12.36 | 20.48 | 20.58 |
| | **Test** | | 4 | 37.76 | 12.46 | 21.76 | 21.86 |

TABLE 5.1: Result using classical graph-based algorithms. We selected the number of extracted tokens (#T. in the table) or sentences (#S in the table) on the validation test and run the most promising model on the test set.

previous work [25, 8], which showed that general-domain Natural Language Processing resources tend to have suboptimal performance in the patent domain.

Moreover, sentences naturally contain references to other parts of the original text[3] e.g., *"as described below"* in US.2005152011-A1 or *"according to claim 1"* in US-9478115-B2. We also notice that all the extracted sentences tend to be extremely long and naturally contain core and peripheral information (e.g., included in parenthesis). These are known limitations of naive extractive models and are very common problems of our extracted summaries. Extracted sentences do not seem too similar to each other, which is sometimes described as a limitation of graph-based systems.

Even with their limitations, the algorithms seem to perform reasonable content selection (with TextRank being superior to LexRank also from a qualitative perspective); when compared to their references, the extracted summaries often contain most of their core elements and, in many cases, are very similar to the reference in terms of content. This is evident in some specific cases (e.g., patent US-9478115-B2 and US-2003016244-A1) and is interesting, considering the algorithm is unsupervised.

If we assume the final target of the extracted summaries is human readers, however, the lack of discourse structure and length of the extracted sentences might make the outputs too hard to read. It might however be possible to use the outputs in an ad hoc interface, e.g. where core sentences are highlighted.

[3] Sentences also tend to contain numerical references to the figures, e.g., in US-2003016244-A1

| pub_num | Gold standard | TextRank |
|---|---|---|
| US-2005152022-A1 | Electro-optic media are disclosed containing a plurality of types of electro-optic media within a single layer. The media provide blinking displays, displays in which images appear transiently and displays having a wider color gamut than conventional electro-optic displays | For example, as described in the aforementioned 2002/01 80688 and WO 2004/088002, differing types of electrophoretic capsules capable of differing extreme optical states may be deposited in alignment with multiple sets of electrodes to produce color displays. The mixed color display described in this Example used two different types of opposite charge dual particle encapsulated electrophoretic media, with all four types of electrophoretic particles polymer-coated by the type of processes described in the aforementioned U.S. Pat. No. The polymer-coated pigment particles thus produced were incorporated into electrophoretic internal phases (i.e., mixtures of pigment particles dispersed in a suspending fluid) and these internal phases encapsulated and the resultant capsules formed into displays as described below. |
| US-2003016244-A1 | In a method for intermediating images, user group information concerning at least one user group to which the user belongs is extracted from a user group table recording a relationship between the user and the user group based on the authentication information of a user. The user group information extracted is sent to the user and obtaining information of a selected user group selected by the user and an image corresponding to the selected user group selected by the user is extracted. Then, the image corresponding to the selected user group is sent to the user. | When the receiving program 103 received user group selection information, which is information of the user group selected by the user in the step S 202 , the data extracting program 106 extracts an image set as images of the user group selected by the user from images registered in the image intermediating site 102 based on the user group selection information. When the group selecting process in the step S 202 is normally completed, the data extracting program 106 refers to the image managing file 110 and extracts a reduced image under the group selected by the user (step S 501 ). When the receiving program 103 receives the purchase information sent in the step S 701 (step S 702 ), the image intermediating system 102 checks whether or not the user is permitted to inquire about the desired image selected by the user based on the image managing file 803 . |
| US-9478115-B2 | An operator system for a machine, in particular for a beverage processing machine, the system comprising a mobile operator device for the machine, a signal emitter for reporting alarm and/or warning signals and safely glasses for protecting the eyes of a user. The safety glasses have a display system that is designed in particular as a head-mounted display, or a virtual retina display, or a projector and the operator device and/or the signal emitter and/or the safety glasses have a respective data transmitter for exchanging machine information and/or alarm and/or warning signals. | The object is achieved by an operator system according to claim 1 , according to which the operator system for a machine, in particular for a beverage processing machine, comprises a mobile operator device for the machine, a signal generator for reporting alarm and/or warning signals, and safety glasses for protecting the eyes of a user, wherein the safety glasses comprise a display system that is configured in particular as a head-mounted display, or a virtual retina display, or a projector, and wherein the operator device and/or the signal generator and/or the safety glasses each comprise a data transmitter for exchanging machine information and/or alarm and/or warning signals. Due to the fact that the operator device and/or the signal generator and/or the safety glasses each comprise a data transmitter both data for the display of information and alarm and/or warning signals can be exchanged between these units in an easy manner. |
| US-9082262-B2 | Methods and apparatus for providing an award based on a multiplicity of game outcomes. In one implementation the invention provides a method for providing a game. The method includes placing a bet by a player, playing of a game to produce a game outcome, evaluating a primary game outcome, paying a direct award if the primary outcome merits it, advancing a bonus accumulator if the primary outcome merits it, and if the bonus accumulator has advanced sufficiently, paying a bonus award to the player and clearing the bonus accumulator. | If the device determines that the game outcome advances the bonus accumulator, then the device advances the bonus accumulator (step 931 ) and determines whether a bonus award threshold has been reached (step 932 ). Where gaming is implemented over a network, players can compete against each other to reach n outcomes in order to attain a bonus award or bonus round play. For the above described games that provide variable outcome bonuses, a bonus award can be a fixed payout amount, a random payout amount, a payout amount that is based on the outcome, a multiplier of a non-bonus award, a bonus round, advancement to a bonus round, an increase in an accumulated awards such as a jackpot, and any combination of the examples of bonus awards described. |
| US-9208317-B2 | A plurality of untrusted digital files are run simultaneously in fewer sandboxes than there are files, while monitoring for malicious activity. Preferably, only one sandbox is used. If the monitoring detects malicious activity, either the files are run again in individual sandboxes, or the files are divided among subsets whose files are run simultaneously in one or more sandboxes, while monitoring for malicious activity. | As an untrusted digital file is being run in a sandbox session, the sandbox code monitors the session for attempted malicious activity. A session of simultaneously running several files may be terminated as soon as malicious activity is detected (which indicates that at least one of the files contains malware), or may be allowed to run to completion in the conventional manner, followed by log inspection. Under some favorable circumstances, the process ID of the parent process indicates with certainty which file is the malicious file, so that the simultaneous run may be terminated upon the appearance of such attempted malicious activity in the log and the identification of the responsible file (and it is the unpredictability of such circumstances that mandates the iterative nature of the present invention); but allowing the run to run to completion provides further clues about new malware, as described above. |

TABLE 5.2: Extracted summaries using TextRank (150 tokens, test set)

| pub_num | Gold standard | LexRank |
|---|---|---|
| US-2005152022-A1 | Electro-optic media are disclosed containing a plurality of types of electro-optic media within a single layer. The media provide blinking displays, displays in which images appear transiently and displays having a wider color gamut than conventional electro-optic displays | As discussed above, both the single and dual particle types of electrophoretic display normally only display two colors at each pixel, the colors of the particle and the suspending fluid in a single particle display, and the colors to the two types of particles in an dual particle display. On the other hand, the positively charged particles 118 B are attracted to the rear electrode 104 , so that the second areas of the display (i.e., the areas in which the capsules contain the positively charged particles 118 B) display the blue color of the suspending fluid 116 . In such a mixed color display, the electro-optic medium comprises a mixture of at least two types of electro-optic elements, each type having at least two different optical states. The mixed color display described in this Example used two different types of opposite charge dual particle encapsulated electrophoretic media, with all four types of electrophoretic particles polymer-coated by the type of processes described in the aforementioned U.S. Pat. |
| US-2003016244-A1 | In a method for intermediating images, user group information concerning at least one user group to which the user belongs is extracted from a user group table recording a relationship between the user and the user group based on the authentication information of a user. The user group information extracted is sent to the user and obtaining information of a selected user group selected by the user and an image corresponding to the selected user group selected by the user is extracted. Then, the image corresponding to the selected user group is sent to the user. | Moreover, the external storage unit of the image intermediating system 102 includes a user group file 108 for storing user group information, a user file 109 for storing user information, and an image managing file 110 for managing images of the user group. When the receiving program 103 received user group selection information, which is information of the user group selected by the user in the step S 202 , the data extracting program 106 extracts an image set as images of the user group selected by the user from images registered in the image intermediating site 102 based on the user group selection information. When the group selecting process in the step S 202 is normally completed, the data extracting program 106 refers to the image managing file 110 and extracts a reduced image under the group selected by the user (step S 501 ). When the receiving program 103 receives the purchase information sent in the step S 701 (step S 702 ), the image intermediating system 102 checks whether or not the user is permitted to inquire about the desired image selected by the user based on the image managing file 803 . |
| US-9478115-B2 | An operator system for a machine, in particular for a beverage processing machine, the system comprising a mobile operator device for the machine, a signal emitter for reporting alarm and/or warning signals and safely glasses for protecting the eyes of a user. The safety glasses have a display system that is designed in particular as a head-mounted display, or a virtual retina display, or a projector and the operator device and/or the signal emitter and/or the safety glasses have a respective data transmitter for exchanging machine information and/or alarm and/or warning signals. | The invention relates to an operator system for a machine, in particular for a beverage processing machine, to a mobile operator device, a signal generator and safety glasses. The object is achieved by an operator system according to claim 1 , according to which the operator system for a machine, in particular for a beverage processing machine, comprises a mobile operator device for the machine, a signal generator for reporting alarm and/or warning signals, and safety glasses for protecting the eyes of a user, wherein the safety glasses comprise a display system that is configured in particular as a head-mounted display, or a virtual retina display, or a projector, and wherein the operator device and/or the signal generator and/or the safety glasses each comprise a data transmitter for exchanging machine information and/or alarm and/or warning signals. In the operator system, the safety glasses may comprise a talk-listen unit which, in particular, is connected to the data transmitter of the safety glasses for transmitting speech information. In the operator system, the display system may be configured to display virtual operator devices to the user. |
| US-9082262-B2 | Methods and apparatus for providing an award based on a multiplicity of game outcomes. In one implementation the invention provides a method for providing a game. The method includes placing a bet by a player, playing of a game to produce a game outcome, evaluating a primary game outcome, paying a direct award if the primary outcome merits it, advancing a bonus accumulator if the primary outcome merits it, and if the bonus accumulator has advanced sufficiently, paying a bonus award to the player and clearing the bonus accumulator. | A device executing the game receives a wager from the player (step 820 ) and initiates a current game (step 821 ). Otherwise, the player is not. Accumulated win poker counts the number of time a player achieves a particular winning outcome and provides a bonus award when the count (i.e., the bonus accumulator) reaches a certain threshold number. Apparatus of the invention can be implemented in a computer program product tangibly embodied in a machine-readable storage device for execution by a programmable processor; and method steps in the invention can be performed by a programmable processor execution a program of instructions to perform functions of the invention by operating on input data and generating output. |
| US-9208317-B2 | A plurality of untrusted digital files are run simultaneously in fewer sandboxes than there are files, while monitoring for malicious activity. Preferably, only one sandbox is used. If the monitoring detects malicious activity, either the files are run again in individual sandboxes, or the files are divided among subsets whose files are run simultaneously in one or more sandboxes, while monitoring for malicious activity. | For example, if the sandbox operating system is a Microsoft Windows operating system then the process ID of the parent of the process that attempted malicious activity may indicate which file is most likely to be the malicious file. A set 12 of one or more untrusted digital files is inspected in a sandbox 14 .for running untrusted digital files. So, for example, a sandbox running on a host computer with a Linux operating system could set up and run a virtual computer whose operating system is Microsoft Windows. |

TABLE 5.3: Extracted summaries using LexRank (4 sentences, test set)

| | Set | #S | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-LSums |
|---|---|---|---|---|---|---|
| SumBasic [134] | Val | 1 | 16.06 | 4.78 | 12.24 | 12.26 |
| | Val | 2 | 24.98 | 7.08 | 16.76 | 16.80 |
| | Val | 3 | 30.28 | 8.35 | 18.71 | 18.77 |
| | Val | 4 | 33.38 | 9.07 | 19.52 | 19.60 |
| | Val | 5 | 35.10 | 9.46 | **19.76** | **19.85** |
| | Val | 6 | 35.96 | 9.70 | 19.71 | 19.80 |
| | Val | 7 | **36.18** | 9.80 | 19.46 | 19.57 |
| | Val | 8 | 36.00 | **9.83** | 19.12 | 19.22 |
| | Val | 9 | 35.56 | 9.80 | 18.73 | 18.84 |
| | **Test** | 5 | 35.04 | 9.42 | 19.72 | 19.81 |

TABLE 5.4: Result using SumBasic. We selected the number of extracted sentences (#S) on the validation test and run the most promising model on the test set.

### 5.2.2 SumBasic

SumBasic is a classical unsupervised summarization algorithm that extracts sentences based on their word frequencies while minimizing repetition. We used the *sumy* implementation. We validate the number of extracted sentences and leave any other parameters to their default value.

**Automatic evaluation.** Table 5.4 shows the ROUGE scores. SumBasic tends to perform worse than graph-based algorithms. In contrast to the graph-based methods, it tends to work best with a higher number of (short) extracted sentences.

**Qualitative assessment.** Table 5.5 shows some random examples of summaries generated by the best-performing SumBasic system (according to ROUGE) on the test set. We qualitatively evaluated a selection of 20 outputs.

In general, the outputs are visibly worse than the ones obtained through extractive summarizers. Specifically, while errors in sentence tokenization had a minor impact in the previously-generated systems, their impact is magnified when using SumBasic.[4]. For example, patent US-7206119-B2's summary contains a long list of references to previous art – which are not stand-alone sentences. Content-wise, the summaries are very noisy and difficult to interpret. The lack of discourse structure is magnified, and grasping the essence of the invention from the extracted summaries is practically impossible. Sentences are generally much shorter that those extracted from graph-based algorithms but often non-central in the patent (e.g., *"Each process in FIG2 will be described in detail."* for patent US-2003016244-A1).

Even though the difference in ROUGE when compared to graph-based systems is small, the summaries extracted using SumBasic are completely unusable for any practical application.

[4] Note that both algorithms use the same tokenization algorithm

| pub_num | Gold standard | SumBasic |
|---|---|---|
| US-2005152022-A1 | Electro-optic media are disclosed containing a plurality of types of electro-optic media within a single layer. The media provide blinking displays, displays in which images appear transiently and displays having a wider color gamut than conventional electro-optic displays | WO 99/67678; WO 00/05704; WO 00/38000; WO 00/38001; WO00/36560; WO 00/67110; WO 00/67327; WO 01/07961; WO 01/08241; WO 03/107,315; WO 2004/023195; WO 2004/049045; WO 2004/059378; WO 2004/088002; WO 2004/088395; and WO 2004/090857. A single particle medium has only a single type of electrophoretic particle suspended in a suspending medium, at least one optical characteristic of which differs from that of the particles. However, the capsules are of two different types. Obviously, the range of colors available in mixed color displays of the present invention are limited by the color ranges achievable by the constituent types of electro-optic elements. For example, the displays etc. |
| US-2003016244-A1 | In a method for intermediating images, user group information concerning at least one user group to which the user belongs is extracted from a user group table recording a relationship between the user and the user group based on the authentication information of a user. The user group information extracted is sent to the user and obtaining information of a selected user group selected by the user and an image corresponding to the selected user group selected by the user is extracted. Then, the image corresponding to the selected user group is sent to the user. | In step S 302 , the receiving program 103 receives the login information sent in the step S 301 .. . for the user groups, respectively. An example of the window for registering the image is shown as an image registering window 1101 in FIG1 . of the image managing file 803 . The members permitted to inquire are transferred. |
| US-9478115-B2 | An operator system for a machine, in particular for a beverage processing machine, the system comprising a mobile operator device for the machine, a signal emitter for reporting alarm and/or warning signals and safely glasses for protecting the eyes of a user. The safety glasses have a display system that is designed in particular as a head-mounted display, or a virtual retina display, or a projector and the operator device and/or the signal emitter and/or the safety glasses have a respective data transmitter for exchanging machine information and/or alarm and/or warning signals. | The machine may comprise a computer-based machine controller. The data transmitter may comprise a receiver unit and/or a transmit unit. 3D information may be three-dimensional information. In the operator system, the display system may be configured to display virtual operator devices to the user.Also, the safety glasses 4 may be connected to the signal generator 3 of FIG1 . |
| US-9082262-B2 | Methods and apparatus for providing an award based on a multiplicity of game outcomes. In one implementation the invention provides a method for providing a game. The method includes placing a bet by a player, playing of a game to produce a game outcome, evaluating a primary game outcome, paying a direct award if the primary outcome merits it, advancing a bonus accumulator if the primary outcome merits it, and if the bonus accumulator has advanced sufficiently, paying a bonus award to the player and clearing the bonus accumulator. | Optionally, the device obtains game outcomes of other players (step 114 ) The game outcomes of other players can be previous and/or current game outcomes. If the three matching symbols are the chimp symbols or the ape symbols, this is normally a losing game outcome. Alternatively, only the count for the achieved event is reset to zero and the counts for each other outcome event remain as they were prior to the achieved count. For the disposition of the game after a bonus award is given, all bonus accumulators is reset after any bonus awards given. A number of implementations of the invention have been described. |
| US-9208317-B2 | A plurality of untrusted digital files are run simultaneously in fewer sandboxes than there are files, while monitoring for malicious activity. Preferably, only one sandbox is used. If the monitoring detects malicious activity, either the files are run again in individual sandboxes, or the files are divided among subsets whose files are run simultaneously in one or more sandboxes, while monitoring for malicious activity. | A file that attempts malicious activity is therefore known to have malicious code embedded therein. Second, a complete log provides clues to new malware patterns. The principles and operation of computer security according to the present invention may be better understood with reference to the drawings and the accompanying description. As described above, the execution of sandbox code 44 sets up one or more sandboxes within system 30 for the simultaneous inspection of the files of each set. for running untrusted digital files. |

TABLE 5.5: Extracted summaries using SumBasic (7 sentences, test set)

| | Set | #S | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-LSums |
|---|---|---|---|---|---|---|
| LSA [175] | Val | 1 | 20.09 | 4.38 | 13.54 | 13.58 |
| | Val | 2 | 28.51 | 6.48 | 17.15 | 17.21 |
| | Val | 3 | 32.37 | 7.70 | 18.43 | 18.52 |
| | Val | 4 | 33.93 | 8.38 | **18.80** | **18.90** |
| | Val | 5 | **34.28** | 8.78 | 18.70 | 18.81 |
| | Val | 6 | 34.00 | 9.02 | 18.43 | 18.54 |
| | Val | 7 | 33.30 | 9.14 | 18.05 | 18.15 |
| | Val | 8 | 32.44 | **9.20** | 17.63 | 17.73 |
| | **Test** | 5 | 34.26 | 8.72 | 18.66 | 18.76 |

TABLE 5.6: Result using LSA. We selected the number of extracted sentences (#S) on the validation test and run the most promising model on the test set.

### 5.2.3 Latent Semantic Analysis

Latent Semantic Analysis [73] aims at exploiting the latent semantic structure of the document and extracts sentences that best represent the most important latent topics. We use the *sumy* implementation, validate the number of sentences, and leave all other parameters to their default values.

**Automatic evaluation.** Table 5.6 shows the ROUGE scores. LSA tends to perform marginally better than SumBasic, but worse than the graph-based algorithms. In contrast to the graph-based methods and similar to SumBasic, it tends to work best when extracting several short sentences.

**Qualitative assessment** Examples obtained through LSA are reported in Table 5.7. Results are much more intelligible than those obtained through SumBasic. Note that ROUGE seems completely unable to measure the high qualitative difference.

Even with the known limitations of extractive systems (references, structure, sentences needing compression, etc.), some reasonable content selection is performed. For example, they often extract the sentence that describes the invention's nature, as in *"The present invention is based on the object to provide an operator system for a machine, which is ergonomic with regard to the handling thereof and offers sufficient work protection."* for US-9478115-B2 or *"The present invention relates to computer security and, more particularly, to an efficient method of screening untrusted digital files."* for US-9208317-B2.

Sentences are generally shorter than those extracted by graph-based systems.

Souza et al. [171] noticed LSA showed a better quality when compared to TextRank in the generation of patent titles. Our results do not confirm this finding for Abstract generation from the Description as measured automatically; qualitatively, the results are relatively different and might be used for different purposes.

| pub_num | Gold standard | LSA |
|---|---|---|
| US-2005152022-A1 | Electro-optic media are disclosed containing a plurality of types of electro-optic media within a single layer. The media provide blinking displays, displays in which images appear transiently and displays having a wider color gamut than conventional electro-optic displays | Such a display uses a large number of small bodies (typically spherical or cylindrical) which have two or more sections with differing optical characteristics, and an internal dipole. Numerous patents and applications assigned to or in the names of the Massachusetts Institute of Technology (MIT) and E Ink Corporation have recently been published describing encapsulated electrophoretic media. In a microcell electrophoretic display, the charged particles and the suspending fluid are not encapsulated within microcapsules but instead are retained within a plurality of cavities formed within a carrier medium, typically a polymeric film. By encapsulating such oppositely-charged particles separately with a colored suspending fluid, in the manner described in many of the aforementioned E Ink and MIT patents and applications, one can provide the double medium display (generally designated 100 ) shown in FIG1 A and 1B of the accompanying drawings. The coated films were allowed to oven dry at 60 C. for 15 minutes to produce an electrophoretic medium approximately 30 cm thick containing essentially a single layer of capsules (see the aforementioned 2003/0137717). |
| US-2003016244-A1 | In a method for intermediating images, user group information concerning at least one user group to which the user belongs is extracted from a user group table recording a relationship between the user and the user group based on the authentication information of a user. The user group information extracted is sent to the user and obtaining information of a selected user group selected by the user and an image corresponding to the selected user group selected by the user is extracted. Then, the image corresponding to the selected user group is sent to the user. | Also, the user and the visitors can make a request of the service site for an extra development of a silver film of the images. However, in a case in which a plurality of users as a group individually attempt to register images to the same album, several problems occur.These programs 103 through 107 are stored in an external storage unit of the image intermediating system 102 by the installer 11 installing from a CD-ROM (Compact Disk Read Only Memory) 2 , and read and temporarily stored in an internal storage unit when each of the programs 103 through 107 is being executed. The receiving program 107 performs a service of the extra development (hard copy) of the desired image. When the user makes an instruction for completing a request of the service, the image intermediating system is terminated. |
| US-9478115-B2 | An operator system for a machine, in particular for a beverage processing machine, the system comprising a mobile operator device for the machine, a signal emitter for reporting alarm and/or warning signals and safely glasses for protecting the eyes of a user. The safety glasses have a display system that is designed in particular as a head-mounted display, or a virtual retina display, or a projector and the operator device and/or the signal emitter and/or the safety glasses have a respective data transmitter for exchanging machine information and/or alarm and/or warning signals. | The present invention is based on the object to provide an operator system for a machine, which is ergonomic with regard to the handling thereof and offers sufficient work protection. Due to the fact that the safety glasses comprise a display system, the eyes of the user are, on one hand, protected from dangerous foreign objects. Due to the fact that the talk-listen unit is connected to the data transmitter of the safety glasses for transmitting speech information same does not require an own transmission interface. The active sound suppression allows the generation of an acoustic counter-signal, so that disturbing ambient noise is extinguished in the ear by interference. Similar to a legend in a drawing the user 7 is displayed, for instance, the tightening torque for a screw located in the field of vision and/or the tool type suited therefor. |
| US-9082262-B2 | Methods and apparatus for providing an award based on a multiplicity of game outcomes. In one implementation the invention provides a method for providing a game. The method includes placing a bet by a player, playing of a game to produce a game outcome, evaluating a primary game outcome, paying a direct award if the primary outcome merits it, advancing a bonus accumulator if the primary outcome merits it, and if the bonus accumulator has advanced sufficiently, paying a bonus award to the player and clearing the bonus accumulator. | The features described can be applied to a wide variety of computer program applications in which awards can be based on multiple game outcomes. For example, a player can receive a bonus award or round if seven of the previous ten plays produced eligible outcomes. In another example, a bonus award or round can be earned when the player achieves three hands totaling twenty-one before the dealer gets two Blackjacks. In another example, multiple pays can be included such that getting three totals of twenty-one prior to the dealer getting two Blackjacks, as in the previous example, results in a certain bonus award or round but each additional total of twenty-one attained prior to the dealer getting a Blackjack results in an additional bonus award or round. The invention can be implemented as a traditional table game, or in digital electronic circuitry, or in computer hardware, firmware, software, or in combinations of them. |
| US-9208317-B2 | A plurality of untrusted digital files are run simultaneously in fewer sandboxes than there are files, while monitoring for malicious activity. Preferably, only one sandbox is used. If the monitoring detects malicious activity, either the files are run again in individual sandboxes, or the files are divided among subsets whose files are run simultaneously in one or more sandboxes, while monitoring for malicious activity. | The present invention relates to computer security and, more particularly, to an efficient method of screening untrusted digital files. A sandbox provides a tightly controlled set of resources, such as scratch space in a hard disk, for running untrusted digital files. The principles and operation of computer security according to the present invention may be better understood with reference to the drawings and the accompanying description.The general concept of the present invention is to run several untrusted digital files simultaneously in one sandbox. FIG2 is a high-level partial block diagram of an exemplary computer system 30 configured to implement the present invention. |

TABLE 5.7: Extracted summaries using LSA (5 sentences, test set)

| | Set | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-LSums |
|---|---|---|---|---|---|
| BART [102] | Validation | 41.70 | 17.52 | 28.38 | 36.00 |
| | Test | 41.53 | 17.25 | 28.18 | 35.80 |
| | Δ extractive | +2.94 | +4.95 | +5.85 | +11.30 |

TABLE 5.8: Result of the baselines on the validation and test sets. We also report the score difference with TextRank, the best-performing extractive system.

## 5.3 Abstractive sequence-to-sequence systems

**BART** BART is a sequence-to-sequence system that we use as a baseline for abstractive summarization. We fine-tune a BART-base model ($\sim$ 140 million parameters) on the BigPatent/G datasets. We train using the Hugging Face library with early stopping on the evaluation loss (patience: 5) and the following hyperparameters: max target length: 250; number of beams: 5; evaluation steps: 10k; max steps: 500M. We leave all other parameters to their default values.

**Automatic evaluation.** Table 5.8 shows the results. As expected, the results improve over all extractive systems, with an increase of almost 5 ROUGE-2 points over the best extractive system.

**Qualitative assessment.** Table 5.10 shows some random examples of summaries generated by the best-performing BART system on the test set.

Qualitatively, we notice that summaries are generally grammatical, with very rare local problems. Text is coherent and much easier to read and understand than those composed through extracted sentences. In all cases, summaries seem adequate and convey the main points of their gold standard counterparts.

However, we noticed that the generated summaries are largely extractive, with sentences taken from the source with no or few modifications.

Compare, for example, the summary generated for patent US-2005152022-A1 with this block of text from its source from the following Background of the Invention subsection, where we marked extractive fragments in bold.

*More specifically, in one aspect* ***this invention relates to electro-optic displays with simplified backplanes, and methods for driving such displays. In another aspect, this invention*** *relates* ***to electro-optic displays in which multiple types of electro-optic units are used to improve the colors available from the displays. The present invention is especially, though not exclusively, intended for use in electrophoretic displays.***

To quantify how extractive the generated summaries are with respect to the source, we compute the coverage (Equation 4.2) and the density of the generated summaries (Equation 4.1), which we report in Table 5.9. We notice that the generated summaries tend to have much longer abstractive fragments with respect to the gold standard.

| | Generated (BART) | Gold standard |
|---|---|---|
| Coverage (avg) | 95.75 | 90.68 |
| Density (avg) | 11.84 | 3.82 |

TABLE 5.9: Extractivity metrics on the summaries generated by the fine-tuned BART model. We also report the corresponding metrics on the gold-standard summaries. The metrics are computed per document and then averaged.

| pub_num | Gold standard | BART |
|---|---|---|
| US-2005152022-A1 | Electro-optic media are disclosed containing a plurality of types of electro-optic media within a single layer. The media provide blinking displays, displays in which images appear transiently and displays having a wider color gamut than conventional electro-optic displays | This invention relates to electro-optic displays with simplified backplanes, and methods for driving such displays. In another aspect, this invention is directed to electrophoretic displays in which multiple types of electro -optic units are used to improve the colors available from the displays. The present invention is especially, though not exclusively, intended for use in electrophoresis displays. |
| US-2003016244-A1 | In a method for intermediating images, user group information concerning at least one user group to which the user belongs is extracted from a user group table recording a relationship between the user and the user group based on the authentication information of a user. The user group information extracted is sent to the user and obtaining information of a selected user group selected by the user and an image corresponding to the selected user group selected by the user is extracted. Then, the image corresponding to the selected user group is sent to the user. | The present invention relates to a method for intermediating images that provides a service providing images via a network. The method includes the steps of: providing an image intermediating system that intermediates the images provided by the service providing service on the network; and registering the images registered by the user as belonging to a user group. |
| US-9478115-B2 | An operator system for a machine, in particular for a beverage processing machine, the system comprising a mobile operator device for the machine, a signal emitter for reporting alarm and/or warning signals and safely glasses for protecting the eyes of a user. The safety glasses have a display system that is designed in particular as a head-mounted display, or a virtual retina display, or a projector and the operator device and/or the signal emitter and/or the safety glasses have a respective data transmitter for exchanging machine information and/or alarm and/or warning signals. | An operator system for a machine, in particular for a beverage processing machine, includes a mobile operator device for the machine, a signal generator for reporting alarm and/or warning signals, and safety glasses for protecting the eyes of a user, wherein the safety glasses comprise a display system that is configured in particular as a head-mounted display, or a virtual retina display. |
| US-9082262-B2 | Methods and apparatus for providing an award based on a multiplicity of game outcomes. In one implementation the invention provides a method for providing a game. The method includes placing a bet by a player, playing of a game to produce a game outcome, evaluating a primary game outcome, paying a direct award if the primary outcome merits it, advancing a bonus accumulator if the primary outcome merits it, and if the bonus accumulator has advanced sufficiently, paying a bonus award to the player and clearing the bonus accumulator. | Methods and apparatus for providing an award based on a multiplicity of game outcomes. In one implementation the invention provides a method for providing a game. The method includes placing a bet by a player, playing of a game to produce a game outcome, evaluating a primary game outcome and paying a direct award if the primary outcome merits it. |
| US-9208317-B2 | A plurality of untrusted digital files are run simultaneously in fewer sandboxes than there are files, while monitoring for malicious activity. Preferably, only one sandbox is used. If the monitoring detects malicious activity, either the files are run again in individual sandboxes, or the files are divided among subsets whose files are run simultaneously in one or more sandboxes, while monitoring for malicious activity. | A method of screening untrusted digital files is disclosed. The method includes the steps of: dividing the set of files being inspected into two or more subsets; inspecting the subsets for malicious activity; and providing a complete report of attempted malicious activity and providing clues to new malware patterns. |

TABLE 5.10: Generated summaries using BART-base fine-tuned on Big-Patent/G

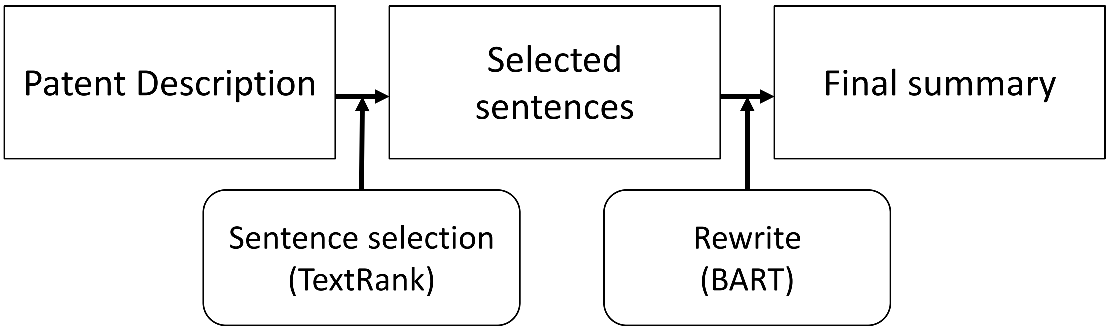


FIGURE 5.1: Pipeline of the select and rephrase system. Sentences are first extracted using TextRank, and then rephrased using a a simple BART model

## 5.4 Extractive to abstractive system: select and rephrase

Results in the previous sections show that graph-based extractive methods can select central content, but lack any discourse structure. Using BART solved some of these issues; however, the model summarizes the first part of the patent document only, as its input is limited to 1024 subtokens.

In this section, we explore a hybrid approach. We first select sentences using an unsupervised graph-based algorithm and then rewrite the content using an abstractive system. Specifically, we use TextRank as it performed best among our analyzed extractive models. We considered three extracted lengths, namely 1000, 500, and 250 tokens, i.e. a) a number of tokens that is close to the maximum the abstractive model can handle, b) an intermediate number with more content selection, and c) a number of tokens that is close to the target length. Then, we train a BART system to rephrase the selected sentences and generate the target summary: we fine-tuned the model using the selected sentences as input and the original gold standard as target.

**Automatic evaluation** Table 5.11 reports the ROUGE scores. Extracting 1000 tokens through TextRank and then rephrasing the summary using BART results in the highest ROUGE, surpassing the vanilla BART approach on all metrics. The obtained metrics are the highest among those of all the extractive and abstractive models we considered.
Note that, even for the approaches where a smaller number of tokens is extracted, relatively good performances are obtained. Extracting 500 tokens results in scores only marginally worse than those obtained by a model BART fed with the first 1024 subtokens. While results obtained by extracting 250 tokens only score worse in term of ROUGE, the rewriting component is crucial. In fact, an improvement of 5 ROUGE-1, 3.3 ROUGE-2, 5.3 ROUGE-L points is observed over the results obtained using TextRank only.

**Qualitative assessment** Table 5.12 shows some examples obtained by extracting 1000 tokens and then rewriting the result using BART, on the test set. The outputs are fluent, and relatively similar to those obtained through the vanilla BART. The coverage (96.12) and density (8.83) also show a marginally lower extractivity of the generated summaries.

| | Set | #T | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-LSums |
|---|---|---|---|---|---|---|
| Hybrid | Val | 1000 | **42.79** | **17.92** | **28.79** | **36.75** |
| | Val | 500 | 41.54 | 16.74 | 27.88 | 35.55 |
| | Val | 250 | 40.33 | 15.60 | 27.01 | 34.61 |
| | **Test** | 1000 | 42.47 | 17.74 | 28.59 | 28.69 |
| | $\Delta$ extractive | | +2.88 | +5.44 | +6.26 | +4.19 |
| | $\Delta$ abstractive | | +0.94 | +0.49 | +0.41 | -7.21 |

TABLE 5.11: Result using the previously described hybrid approach. We selected the number of extracted tokens (#T) on the validation test and run the most promising model on the test set. We also report the score difference with TextRank, the best-performing extractive system, and with BART.

| pub_num | Gold standard | Hybrid |
|---|---|---|
| US-2005152022-A1 | Electro-optic media are disclosed containing a plurality of types of electro-optic media within a single layer. The media provide blinking displays, displays in which images appear transiently and displays having a wider color gamut than conventional electro-optic displays | The present invention relates to a method for driving an electro-optic display. The method comprises the steps of: encapsulating a plurality of particles of the type having substantially the same optical characteristic and a charge of the same polarity in a colored suspending fluid under the influence of an electric field; and driving the electrophoretic display by applying an electrical field to the suspending fluid by the electrode, the at least one second area changing from its first to its second state at a rate different from the first area. |
| US-2003016244-A1 | In a method for intermediating images, user group information concerning at least one user group to which the user belongs is extracted from a user group table recording a relationship between the user and the user group based on the authentication information of a user. The user group information extracted is sent to the user and obtaining information of a selected user group selected by the user and an image corresponding to the selected user group selected by the user is extracted. Then, the image corresponding to the selected user group is sent to the user. | An image intermediating system includes: a receiving program for receiving user group selection information from a user terminal; a data extracting program for extracting an image set as images of the user group selected by the user from images registered in an image intermediation site based on the received information; and a sending program for sending the extracted image set to the user terminal. |
| US-9478115-B2 | An operator system for a machine, in particular for a beverage processing machine, the system comprising a mobile operator device for the machine, a signal emitter for reporting alarm and/or warning signals and safely glasses for protecting the eyes of a user. The safety glasses have a display system that is designed in particular as a head-mounted display, or a virtual retina display, or a projector and the operator device and/or the signal emitter and/or the safety glasses have a respective data transmitter for exchanging machine information and/or alarm and/or warning signals. | An operator system for a machine, in particular for a beverage processing machine, includes a mobile operator device for the machine, a signal generator for reporting alarm and/or warning signals, and safety glasses for protecting the eyes of a user, wherein the safety glasses include a display system that is configured in particular as a head-mounted display, or a virtual retina display. |
| US-9082262-B2 | Methods and apparatus for providing an award based on a multiplicity of game outcomes. In one implementation the invention provides a method for providing a game. The method includes placing a bet by a player, playing of a game to produce a game outcome, evaluating a primary game outcome, paying a direct award if the primary outcome merits it, advancing a bonus accumulator if the primary outcome merits it, and if the bonus accumulator has advanced sufficiently, paying a bonus award to the player and clearing the bonus accumulator. | Methods and apparatus for providing an award based on a multiplicity of game outcomes. In one implementation, a method for providing a game includes placing a bet by a player, playing of a game to produce a game outcome, evaluating a primary game outcome against a first set of criteria for winning an award, paying a direct award if the primary outcome merits it, advancing a bonus accumulator if a bonus event occurs, and if the bonus event has not yet occurred, paying an bonus award to the player and clearing the accumulator. |
| US-9208317-B2 | A plurality of untrusted digital files are run simultaneously in fewer sandboxes than there are files, while monitoring for malicious activity. Preferably, only one sandbox is used. If the monitoring detects malicious activity, either the files are run again in individual sandboxes, or the files are divided among subsets whose files are run simultaneously in one or more sandboxes, while monitoring for malicious activity. | A method of screening untrusted digital files for the presence of embedded malware is disclosed. The method includes the steps of dividing a set of one or more digital files into at least two subsets, testing the subsets for malicious activity, and, if malicious activity is detected in a subset of the set of digital files, running the subset in a sandbox. |

TABLE 5.12: Extracted summaries using the hybrid approach

## 5.5 Conclusions

Summarization approaches in the patent domain were rarely compared in previous work. Thus, in this chapter, we have benchmarked several extractive, abstraction, and hybrid models with a patent dataset.

Among the extractive models, graph-based approaches seem to perform best at content selection; sentence segmentation is, however, still often imperfect. Moreover, the extracted sentences are very long and hard to read, with many lost references and no structure. Other approaches show limited success. We have not considered supervised extractive systems due to resource constraints.

BART generates fluent summaries with relevant content, which are much easier to read and often similar to the targets on a surface level. The generated summaries show, however, a limited level of abstractivity. Experts should be involved in evaluating the factuality of the generated content.

A simple extractive-to-abstractive hybrid approach based on combining TextRank and BART obtained the best results in terms of automatic metrics and seems to encourage the use of ad hoc methods for long document summarization.

# Chapter 6

# From papers to patents: does a divide and conquer approach help patent summarization?

The summarization of long documents is challenging, as most standard systems can only handle inputs that are limited in size. To this end, previous research has explored the document structure. For scientific papers, for example, independently generating the summary from individual sections was found advantageous.

In this chapter, we apply this methodology to the patent domain, as patent documents have a structure similar to that of papers. We find, however, that patents' structure is less predictable, and their Abstracts are less compositional.

## 6.1 Using structure for patent summarization

Abstractive summarization has traditionally targeted relatively short text, e.g., news articles. Research summarizing long documents is more recent and faces additional challenges [90]. For example, models must deal with long-distance relations, and asserting relevance is more complex. Moreover, as we discussed, the maximum length of the input of standard sequence-to-sequence models – which rely on full attention – is often limited: BART [102], for example, can only handle up to 1024 subtokens in its default configuration.

In order to solve this issue without modifying the model architecture, previous work has exploited the document structure. Gidiotis and Tsoumakas [54], for example, proposed DANCER, an approach to summarize scientific paper sections independently.

This chapter explores the use of compositional approaches in the patent domain. Similarly to scientific papers, patents are divided into sections and subsections and have an Abstract that summarizes the characteristics of the invention.

We find that patents' structure is less consistent and Abstracts are less compositional than papers', making the approach challenging to transfer.

### 6.1.1 Long document summarization

The summarization of long documents is very challenging.

For example, since the compression ratio is much greater for long documents – and thus, the original content must be selected and condensed more with respect to general summarization – the chance that the gold standard is only one of the possible reasonable summaries of the source increases.

Secondly, datasets for long document summarization normally include more complex content with respect to classical ones (e.g., in the news domain): they feature scientific articles, business and financial reports, etc, which are in general harder to process. Moreover, these documents are often more structured, and thus more coherent at the local than at the global level; however, the target summaries generally need to be fluent and globally consistent.

Finally, standard state-of-the-art models (e.g., pre-trained transformers) cannot be used out of the shelf for long document summarization. Full attention is quadratic in time and memory in the input length: thus, standard state-of-the-art transformers can only deal with limited inputs, and summarizing long documents requires tailored approaches.

A first line of research modifies the model architecture, complementing the full attention with sliding window attention [16], local attention, random attention, or a mix [209]. These efficient attention mechanisms are generally pseudo-linear in the typical case with respect to the input length; thus, processing longer documents requires less time and memory.

Another approach uses hybrid methods, e.g., selecting relevant sentences and then rewriting them [34, 108, 144, 117]. We used a simple select and rewrite approach in Chapter 5.

Finally, when the document is structure, this characteristics can be exploited, e.g., by summarizing relevant sections independently, as proposed in Gidiotis and Tsoumakas [54] for scientific papers. We are inspired by this last line of research and aim to evaluate whether a similar method transfers to the patent domain.

### 6.1.2 Patents as structured long documents

Patent Descriptions are the longest section of a very long document.

Table 6.1 shows a comparison among the BigPatent/G dataset (in the version we are using) and other summarization datasets. Most "classical" datasets – particularly those in the news domain – contain relatively short documents and summaries. BigPatent, along with other datasets for the summarization of scientific papers and technical documents, contains much longer summaries and sources.

The textual part of patent documents contains the Claim section (its legal portion, which defines the extent of the legal protection) and a Description of the Invention (which discloses the invention in detail). The invention is summarized in an Abstract. We use the Description as input and the Abstract as target. The Description is further divided into subsections: some (e.g., the Field of the invention and the Background) include high-level information on the general technical domain the invention belongs to, the invention objectives, and its high-level characteristics (somewhat similar to the papers' Introduction

| | Summary | | Source | |
|---|---|---|---|---|
| Dataset | # tokens | # sentences | # tokens | Compression ratio |
| CNN/DM [130] | 55.6 | 3.8 | 789.9 | 13.0 |
| NYT [153] | 44.9 | 2.0 | 795.9 | 12.0 |
| NewsRoom [60] | 30.4 | 1.4 | 750.9 | 43.0 |
| XSum [132] | 23.3 | 1.0 | 431.1 | 18.8 |
| ArXiv [38] | 292.8 | 9.6 | 6,913.8 | 39.8 |
| Pubmed [38] | 214.4 | 6.9 | 3,224.4 | 16.2 |
| **BigPatent/G** [158] | 121.0 | 3.6 | 4894.2 | 45.8 |

TABLE 6.1: Length comparison between the BigPatent/G dataset and other summarization datasets. Measures for the BigPatent dataset were recomputed (using NLTK for tokenization). Other measures are from [158]

section). Related inventions and references (i.e., Previous Work) can be described in the Background, only be disclosed through metadata, or appear in the Related Art or References subsections. Finally, the invention is described in detail. This description can be included in a Detailed Description section or in a Description of the Embodiment section (some patents contain only one of the two, some both).

## 6.2 Modifying the DANCER method for the patent domain

Our method is strongly inspired by DANCER [54], with some modifications due to the difference in domains.

We first align sentences in the Abstract with Description subsections (Figure 6.1). Then, we use aligned instances to train a sequence-to-sequence model. At inference (Figure 6.4), we explore some subsection selection strategies and generate the Abstract by summarizing each selected subsection individually using the previously trained model. We also experiment with a final abstractive step.

Specifically, we perform the steps described in the following.

**Dividing and normalizing subsections** To divide the Description text into subsections, we use simple regular expressions, exploiting the fact that section headers are lines including fully cased tokens only. Patent headers can follow different naming conventions[1]. Thus, we normalize the headers through a simple keyword-matching algorithm into nine classes (the process is identical to the one described in Chapter 4 and Figure 4.1). The classes are shown in Table 6.2. Subsections that did not match with any of the keywords were left in a default category and ignored.

**Dividing the Abstract into sentences** We use a general-domain sentence tokenized for this step.

[1] For example, subsections with similar content can be named FIELDS, FIELDS , FIELD OF THE INVENTION, etc.

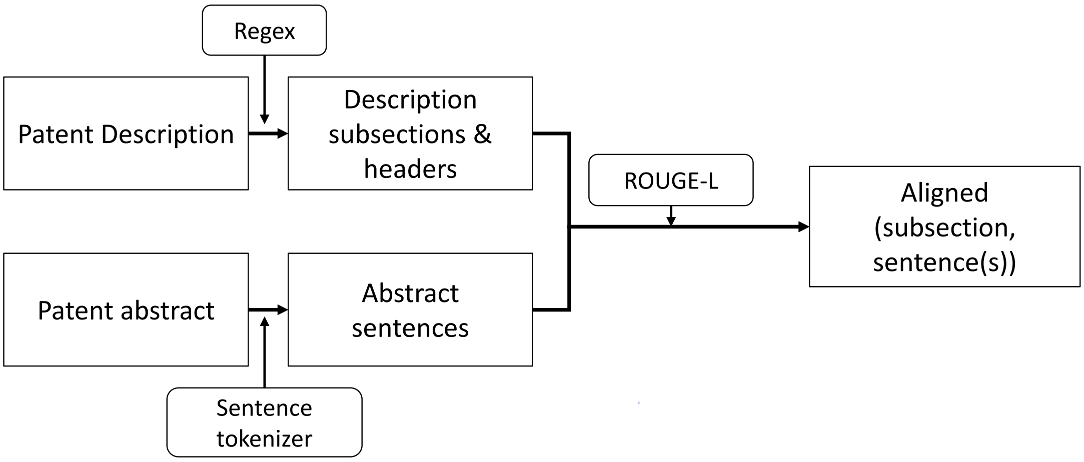


FIGURE 6.1: Pipeline to create training data for the sequence-to-sequence model. We first divide the Description into subsections and the Abstract into sentences and then use ROUGE-L to align sentences in the Abstract with the subsection that "contains most of its information". The pairs are then used to train a sequence-to-sequence model.

| | #Tokens | % patents |
|---|---:|---:|
| FIELD | 73.73 | 38.27% |
| BACKGROUND | 710.04 | 94.85% |
| DRAWINGS | 243.43 | 97.60% |
| EMBODIMENTS | 3168.25 | 53.07% |
| REFERENCES | 92.10 | 28.18% |
| RELATED ART | 644.27 | 4.12% |
| OBJECTIVE | 256.95 | 2.09% |
| DETAILED DESCR. | 3404.91 | 55.23% |

TABLE 6.2: Average length of each subsection type and percentage of patents that contain the subsection.

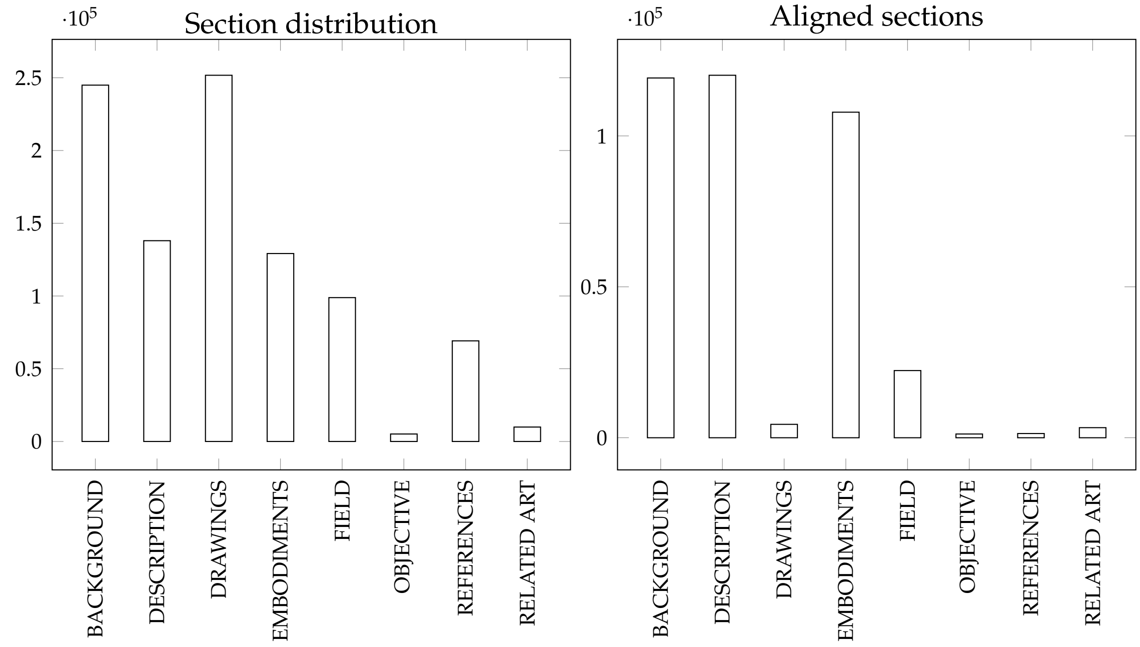


FIGURE 6.2: Section distribution in train patents

**Alignment between abstract sentences and subsections** We use ROUGE-L [105] to align sentences in the abstract to patent subsections. ROUGE-L uses the longest co-occurring n-grams sequence, i.e., the longest sequence of tokens (in the same order but not necessarily consecutive) that is shared between two sequences.

Specifically, for each sentence in the Abstract, we compute its ROUGE-L recall with all individual paragraphs in all subsections; we then align the sentence with the subsection containing the paragraph with the maximum score[2]. Figure 6.6 shows the percentage of subsections that, when present, align with at least one sentence in the patent's Abstract.

**Using paired elements as training data** Following the previous steps, each Abstract sentence is aligned with a Description subsection. Thus, for each $(\text{Description}, \text{Abstract})_i$ pair, we created $N$ $(\text{Subsection}, \text{Abstract sentence(s)})_{i_n}$ pairs, where $N$ is the number of unique subsections that are aligned with at least one sentence in the Abstract. If multiple sentences align with the same patent subsection, the target contains all the aligned sentences in their original order.

We then trained a BART-base model [102] using the subsection as input and the aligned sentence(s) as target; we set the maximum generated length to 250, the number of beans to 5, and left all other hyper-parameters to their default values. We trained with early stopping on the validation set.

In the original DANCER publication, authors experimented with several sequence-to-sequence methods, including a RNN based Pointer-Generator model [157], a RUM-based one [41], and Pegasus [212]. We used BART, to be able to directly compare with our baselines.

Table 6.3 reports the metrics obtained by the model on the sentence generation step. We

[2] We retrieve the subsection containing most of the sentence content, regardless of any possible additional text (that the summarization model will learn to filter out).

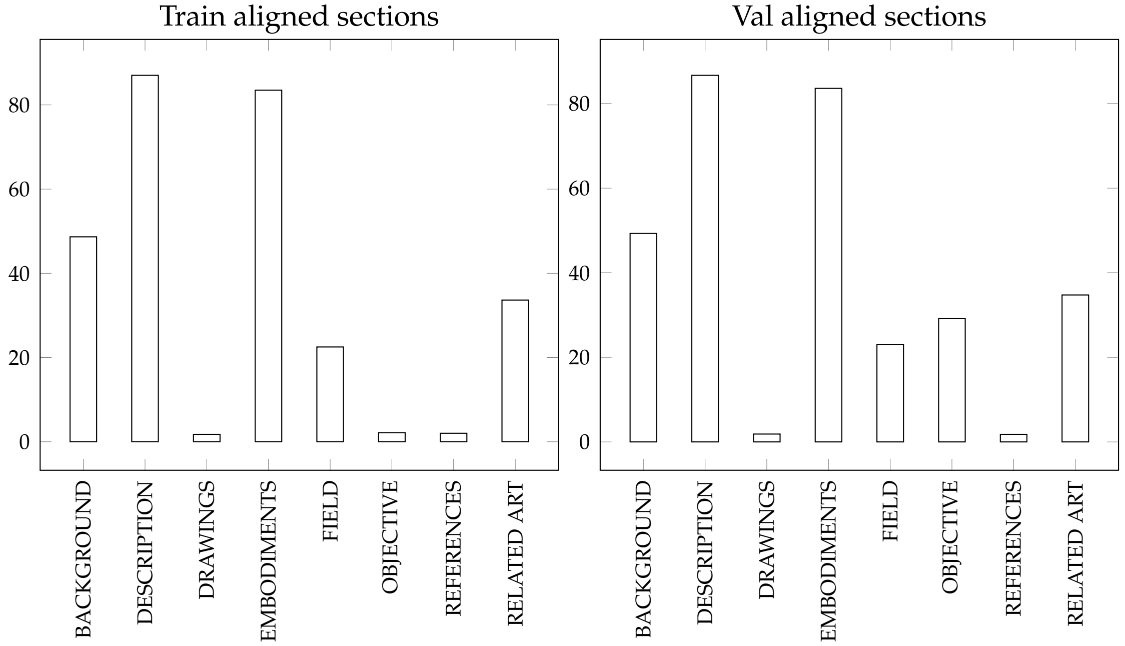


FIGURE 6.3: Percentage of subsections that, when present, are aligned to at least one sentence in the Abstract in the train (left) and validation (right) sets.

| Model | R1 | R2 | RL |
|---|---|---|---|
| BART | 35.00 | 15.74 | 26.63 |
| BART$_{(+\text{ subs. type})}$ | 33.28 | 14.81 | 25.66 |

TABLE 6.3: Model trained on generating the Abstract sentence(s) given the subsection. We also experimented with prepending the subsection text with its type.

also experimented with prepending the subsection type (as a special token) to its text but with no improvement.

**Inference** At inference, we generate the final summary by concatenating the sentences generated from the individual subsections. Patent structure is less coherent than that of papers; in fact, not all subsections appear in all patents. We thus consider several strategies for subsection selection:

(i) Pre-selection: We heuristically pre-select subsections based on their role and fed them to the trained model in their original order. We selected the subsections of type FIELD, BACKGROUND, EMBODIMENTS, OBJECTIVE, DESCRIPTION. We then concatenated the results.

(ii) Generate from M subsections: We retrieve all subsections in the patent and sort them according to how likely they are to be aligned in the whole dataset (Figure 6.6). We generate from the first M most commonly aligned subsection, where M goes from 1 to the total number of subsections in the patent. The final summary is a concatenation of the generated sentences.

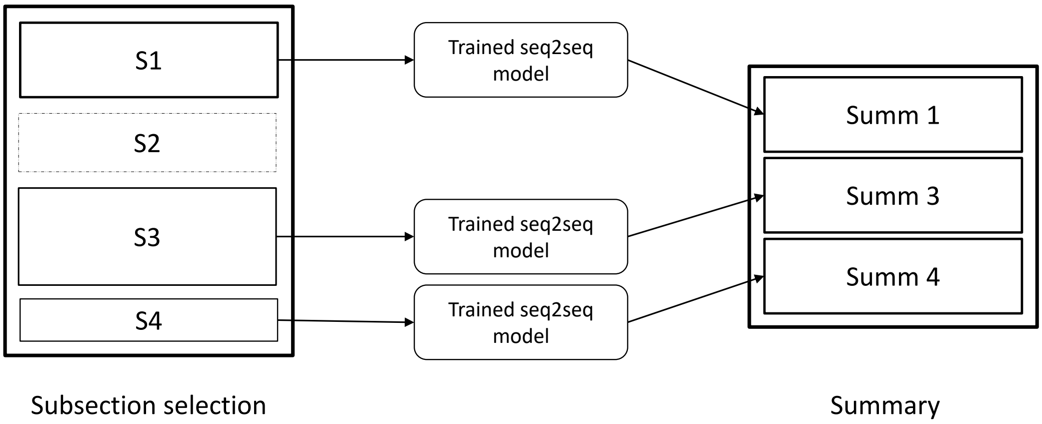


FIGURE 6.4: Pipeline for generating summaries at inference. We first use various methods to select relevant subsentences. Then, we use the previously-trained sequence-to-sequence system to independently summarize selected subsections. The final summary is a concatenation of the generated summaries.

(iii) Generate from all subsections in the patent: we use all subsections in their original order and concatenate the results.

**Final abstractive step** The final abstract obtained as a concatenation of sentences lacks any discourse structure and might not be coherent; in particular, we notice that it often contains repeated information, which is a known limitation of DANCER. Thus, we explore if performing a second abstractive step can improve performance. To this end, we train a second BART model that, given the output of the previous step (i.e., the summary as a concatenation of sentences), is trained to paraphrase it to be more similar to the target Abstract.

## 6.3 Results

Table 6.4 reports the final results on the validation set.

We consider two baselines: TextRank [124] (extractive) and BART-base fed with the first 1024 subtokens from the Description (abstrative).

We report results obtained by generating from pre-selection, using the best-aligned section only (as a baseline), the best result with a varying number of sections (and Figure 6.5 shows ROUGE-L as a function of the number of summarized subsections), and the result obtained by summarizing all sections. We also report the results after the second abstractive step. Note that none of the configurations surpasses the simple BART baseline.

Table 6.6 contains some of the generated outputs. Inspecting the outputs, we noticed that many of the sentences generated from various subsections are very similar and describe what the invention is and its goal. While the second abstractive step helps limit repetition, the resulting output is often short and contains too little information compared to the gold standard.

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| TextRank | 38.59 | 12.34 | 22.33 |
| BART | **41.70** | **17.52** | **28.38** |
| DANCER $_{\text{(preselection)}}$ | 38.73 | 16.03 | 25.63 |
| DANCER $_{\text{(best aligned, M=1)}}$ | 27.39 | 10.64 | 19.83 |
| DANCER $_{\text{(best M, M=3)}}$ | 40.70 | 16.45 | 25.08 |
| DANCER $_{\text{(all)}}$ | 40.68 | 16.38 | 25.90 |
| DANCER + abstractive | 38.88 | 15.89 | 26.99 |

TABLE 6.4: Results on the validation set.

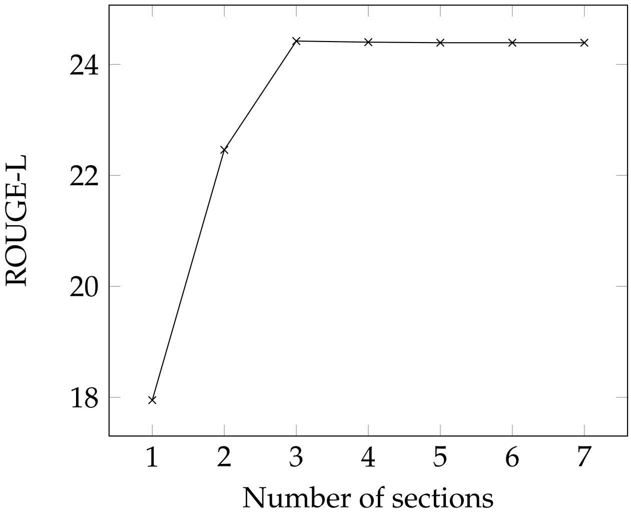


FIGURE 6.5: ROUGE-L results as a function of the number of subsections used for the generation.

| Sections | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-LSums |
|---|---|---|---|---|
| 1 | 27.41 | 10.66 | 19.85 | 19.90 |
| 2 | 36.45 | 14.66 | 23.59 | 25.87 |
| 3 | 40.11 | 16.29 | 24.95 | 28.16 |
| 4 | **40.70** | **16.45** | **25.08** | **28.53** |
| 5 | 40.60 | 16.41 | 25.04 | 28.50 |
| 6 | 40.66 | 16.40 | 25.03 | 28.49 |
| 7 | 40.66 | 16.40 | 25.03 | 28.49 |
| 8 | 40.65 | 16.40 | 25.02 | 28.49 |
| 9 | 40.65 | 16.40 | 25.02 | 28.49 |
| 10 | 40.65 | 16.40 | 25.02 | 28.49 |
| 11 | 40.65 | 16.40 | 25.02 | 28.49 |

TABLE 6.5: Result of the DANCER base model when generating a varying number of sections in the original patent, on the validationt set.

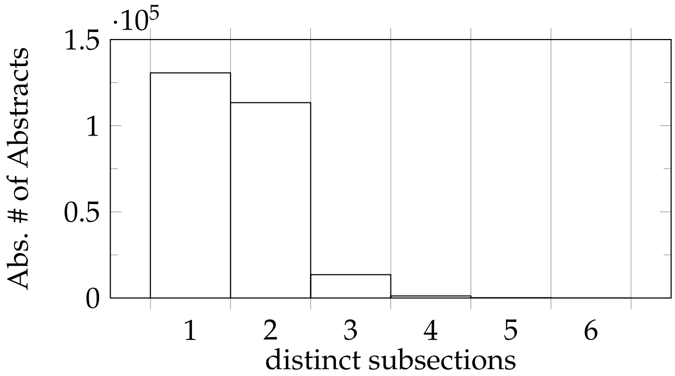


FIGURE 6.6: Number of unique subsections types to which the Abstract aligns

## 6.4 Discussion

**Less predictable structure and session headers** Scientific papers have a very coherent structure as a) they tend to roughly follow a fixed schema (e.g., Introduction, Previous Work, Method, Conclusions), and b) each section has a clear fixed role. While, on a superficial level, patent documents have a similar structure with sections and subsections, they are less coherent. As Table 6.2 shows, the subsections of the Description tend to vary. Moreover, the role of each subsection is less determined. For example, the difference between the Field and the Background subsections is not always well defined: sometimes, the two subsections appear in the same patent (the Field defining the general field of the invention and the Background the detailed technological background); in some cases, the Field subsection is not present, and the Background includes some of its typical content; in a few cases, only the Field of the Invention subsection appears.
Thus, the Description structure is much less predictable than that of papers. Therefore, it might be necessary to define more abstract subsection roles and to classify subsections into types using more complex algorithms that classify the sections' content rather than their headers.

**Less compositional Abstracts** An analysis of the Abstract compositionality shows that many of the sentences in the Abstract align with the same patent subsections. Figure 6.6 represents the number of unique sentences to which each Abstract aligns. Note most patent Abstracts only align to one or two different subsections.
Moreover, a qualitative analysis of the Abstract shows that while paper abstracts tend to follow a fixed structure (first describing the background, then the goal and methods, then the results and conclusions), patent Abstracts seem to lack the compositional nature of scientific papers. The lack of a fixed flow in the Abstract might also explain the relatively low results obtained by the abstractive model when generating the Abstract sentence(s) from the original subsections. As the alignment is more random, finding a pattern and correctly generating the aligned sentences is more challenging.

| pub_num | Gold standard | DANCER |
|---|---|---|
| US-2005152022-A1 | Electro-optic media are disclosed containing a plurality of types of electro-optic media within a single layer. The media provide blinking displays, displays in which images appear transiently and displays having a wider color gamut than conventional electro-optic displays | The present invention provides a double medium display in which a single electro-optic layer contains a plurality of capsules, each of which comprises a capsule wall within which is encapsulated a colored suspending fluid, which is typically dyed blue. The mixed color display of the present invention is useful for both charge dual particle encapsulated electrophoretic media, with all four types of particles polymer-coated by the type of processes described herein. This invention relates to electro-optic displays and to methods for driving such displays. |
| US-2003016244-A1 | In a method for intermediating images, user group information concerning at least one user group to which the user belongs is extracted from a user group table recording a relationship between the user and the user group based on the authentication information of a user. The user group information extracted is sent to the user and obtaining information of a selected user group selected by the user and an image corresponding to the selected user group selected by the user is extracted. Then, the image corresponding to the selected user group is sent to the user. | The present invention provides an image intermediating system including: a receiving program for receiving information from a user; a sending program for sending information to the user; and a data extracting program for extracting an image set as images of the user selected by the user based on the information received from the receiving program. A method for intermediating images that provides a service providing images via a network. There is provided a system and method for providing a user authentication process. |
| US-9478115-B2 | An operator system for a machine, in particular for a beverage processing machine, the system comprising a mobile operator device for the machine, a signal emitter for reporting alarm and/or warning signals and safely glasses for protecting the eyes of a user. The safety glasses have a display system that is designed in particular as a head-mounted display, or a virtual retina display, or a projector and the operator device and/or the signal emitter and/or the safety glasses have a respective data transmitter for exchanging machine information and/or alarm and/or warning signals. | According to the invention, the safety glasses are configured as a head-mounted display and are semipermeable. The invention relates to an operator system for a machine, in particular for a beverage processing machine, comprising a mobile operator device for the machine, a signal generator for reporting alarm and/or warning signals, and safety glasses for protecting the eyes of a user, wherein the safety glasses comprise a display system that is configured in particular as a head-mounted display, or a virtual retina display or a projector, and wherein the operator device and/ or the signal generator are configured in such a way as to be able to communicate machine information or alarm signals. An operator system for a machine, in particular for a beverage processing machine, is disclosed. |
| US-9082262-B2 | Methods and apparatus for providing an award based on a multiplicity of game outcomes. In one implementation the invention provides a method for providing a game. The method includes placing a bet by a player, playing of a game to produce a game outcome, evaluating a primary game outcome, paying a direct award if the primary outcome merits it, advancing a bonus accumulator if the primary outcome merits it, and if the bonus accumulator has advanced sufficiently, paying a bonus award to the player and clearing the bonus accumulator. | A method and apparatus for providing an award based on a multiplicity of game outcomes. A method and apparatus for determining a primary game outcome of a video poker game. Methods and apparatus for providing an award based on a multiplicity of game outcomes. |
| US-9208317-B2 | A plurality of untrusted digital files are run simultaneously in fewer sandboxes than there are files, while monitoring for malicious activity. Preferably, only one sandbox is used. If the monitoring detects malicious activity, either the files are run again in individual sandboxes, or the files are divided among subsets whose files are run simultaneously in one or more sandboxes, while monitoring for malicious activity. | A method and apparatus for scanning a set of digital files simultaneously in one or more sandboxes within a computer system. A method of screening untrusted digital files. Disclosed is a computer system. |

TABLE 6.6: Generated summaries using DANCER. At inference, we select the first three most frequently-aligned subsections.

# Chapter 7

# Automatic Text Simplification

This chapter is an introduction to Text Simplification.

After defining the task in Section 7.1, we briefly describe some translation-inspired methods, with a particular focus on sequence-to-sequence ones, that we will use in the following of the thesis (Section 7.2). We then comment on available datasets (Section 7.3) and on the most common evaluation metrics and procedures (Section 7.4).

## 7.1 The task of Automatic Text Simplification

Automatic Text Simplification is a text-to-text task: given a hard-to-read piece of text, automatic text simplification aims to make it easier to read and understand for its target users. In contrast to summarization, the whole original meaning is generally preserved, even if some details or technicalities might be removed.

For example, given the following original text[1]

> *"Owls are the order Strigiformes, comprising 200 bird of prey species. Owls hunt mostly small mammals, reptiles, insects, and other birds though some species specialize in hunting fish."*

a candidate simplified version might be:

> *"An owl is a bird. There are about 200 kinds of owls. Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits)."*

In the example above, several transformations are performed: some content is removed, sentences are split, the lexicon is simplified, and some examples are added to explain concepts. In general, all or a subset of such transformations are possible: replacing complex words with simpler terms, explaining technical jargon, or performing other syntactical changes.

The question of which modifications should be performed to "simplify" a piece of text is relatively open. Previous studies determined that humans perform simplifications by

[1] This is an example from a talk by Mirella Lapata `http://videolectures.net/esslli2011_lapata_simplification/` [Last accessed March 2023]

many different transformations: splitting long sentences, removing redundancies, rewriting, reordering, transforming verbal voices from passive to active, substituting difficult words, or deleting content [167].

### 7.1.1 Unit of simplification

Text simplification can be performed at different text levels.

As of today, most research on text simplification is at the sentence level (sentence simplification), i.e., sentences are simplified individually and independently of any other context. This will also be our approach for exploring patent simplification in Chapter 8.

In their survey on data-driven approaches to simplification, Alva-Manchego, Scarton, and Specia [2] argue that true text simplification cannot be achieved by sentence simplification only and call for more document-based approaches.
Document-based simplification is, however, still relatively limited. Sun, Jin, and Wan [179], for example, have recently explored the use of articles leads from Wikipedia and Simple English Wikipedia for the task and proposed modifications to SARI – the most popular metric for simplification, that we will discuss in Section 7.4. They find that document-level simplification is still an open problem, and sentence-level models have strong limitations when used at the document level.

### 7.1.2 Simplification target

Automatic text simplification aims at making text easier to read and understand for a target group (e.g., people who are neurodivergent [14] or have a disability [207] or people with a low literacy level [138]).
Moreover, having a simplified version of an input text could be beneficial in a more complex Natural Language Processing pipeline. For example, Silveira and Branco [168] showed that adding a rule-based simplification component to an extractive summarization system helps in producing better outputs in Portuguese. Many other tasks can be improved through a simplification component, e.g., semantic role labeling [194], question generation [66], and information extraction [47].

Siddharthan [167] published a survey detailing how simplification has been used for different target users and downstream tasks.

## 7.2 Methods for automatic text simplification

Several methodologies exist for automatic text simplification, from rule-based to data-driven ones. Refer to Alva-Manchego, Scarton, and Specia [2] for a comprehensive survey of data-driven sentence simplification methods.

Recently, automatic text simplification has largely been framed as a monolingual translation task: in practice, text in the complex language is "translated" into simple language. The transformations performed are learned intrinsically from the training data.

To this end, many methods from machine translation have been explored in the last few decades. We describe here some of the most popular approaches.

### 7.2.1 Non-neural systems

Many of the first approaches for data-driven simplification were based on the noisy channel framework. These approaches were phrase-based [173, 40, 202] or syntax-based.
Zhu, Bernhard, and Gurevych [218], for example, proposed TSM (Tree-based Simplification Model tailored for simplification), a probabilistic, syntax-based approach able to perform several transformations (splitting, deletion, reordering, and substitution) on the basis of the sentence parse tree. Their model was further improved by Coster and Kauchak [39], who added the possibility of performing phrasal deletion.
Wubben, Bosch, and Krahmer [202] argue that dissimilarity from the original sentence is important and rerank the candidate outputs using a heuristic based on their Levenshtein distance with the input.

Other approaches are based on grammar induction, where the task is modeled as a tree-to-tree rewriting problem. Methods rely on parallel corpora to learn a set of rules to transform the tree of the complex sentences into those of the simple sentences [201].

### 7.2.2 Sequence-to-sequence systems

Recently, text simplification (particularly at the sentence level) has been mainly performed using neural approaches, particularly using sequence-to-sequence machine translation models [182].

Sequence-to-sequence methods for automatic text simplification have, in their default setting, a simple encoder-decoder architecture. The base architecture is the same we discussed for neural abstractive methods in Chapter 3. Nisioi et al. [135] were the first to explore sequence-to-sequence translation models for text simplification and adopted an LSTM-based architecture.

As of today, pre trained transformers are generally the default choice. For example, BART [102] is commonly used as a baseline for sentence simplification after being fine-tuned on the target dataset.

Recently, sequence-to-sequence systems have been improved using tricks to control the level of simplification; this includes guiding the extent to which the output should be compressed and rephrased and how "simpler" the tokens that substitute "complex" ones must be. The level of simplification can be controlled at inference time; thus, the same model can produce text with different levels of simplification without needing to be retrained. Taking inspiration from learning-based methods for controlled text generation, ad hoc control tokens can be used for this goal.
Martin et al. [120], for example, proposed **ACCESS** (AudienCe-CEntric Sentence Simplification), a model for sentence simplification that enhances a sequence-to-sequence transformer [192] with control tokens that allow to explicitly manipulate the level of compression and paraphrasing, and the lexical and syntactic complexity. Authors compute several scores (character length ratio, character-level Levenshtein similarity, a lexical complexity score based on frequency, and a syntactic complexity score based on the dependency tree) for each training pair and prepend the input with tokens containing these values. At training time, the model is thus conditioned on these tokens. Authors show that, by modifying these tokens at inference, they can control the simplification characteristics and that the same trained models can thus be adapted for different targets.

Inspired by ACCESS, Sheang and Saggion [159] proposed a similar T5-based model [148] and added a new control token to help the model replace long, complex words with shorter alternatives.

Sequence-to-sequence systems are a "black box" that tries to mimic the simplification transformations in the training data. They have the advantage that they can be trained end-to-end without explicitly extracting features or estimating individual model components (e.g., the language model). This is in contrast to many previous approaches in simplification, where systems try to specifically reproduce some specific transformations. Note that these classical approaches might still be preferred when one wants more control over the specific transformations performed; however, in human-performed simplification, different types of transformations interact, so a sequence-to-sequence rewriting approach might be preferred.

## 7.3 Datasets

Current simplification approaches are data-driven. They generally rely on parallel corpora, similar to those used in machine translation, containing the original and the simplified text. In this section, we describe the most commonly used datasets for text simplification. Corpora and their characteristics are summarized in Table 7.1.

### 7.3.1 Corpora based on the Simple English Wikipedia

Many of the datasets for sentence simplification are based on aligned documents from the English and the Simple English Wikipedia. The Simple English Wikipedia is designed for children, people who speak English as a second language or have learning difficulties. Authors are encouraged to use simple language (inspired by the Basic English rules, with some modifications)[2]: guidelines include using a limited vocabulary when possible, avoiding idiomatic expressions and jargon, and using active voice and basic verbal forms. Syntactically, simple subject-verb-object sentences are preferred, compound sentences are discouraged, and sentences with many subordinate clauses avoided; guidelines also encourage sentence splitting.

Articles in the Simple English Wikipedia can easily be paired with their corresponding English ones. Paired abstracts from the English and the Simple English Wikipedia can thus be used as proxies for document-level simplification [179].
However, Simple English articles are written independently of their English counterpart and are not word-by-word translations; thus, sentence simplification corpora are based on using various strategies to automatically align sentences in the corresponding articles. Since these datasets have been automatically generated, they can contain noisy data and might thus not be ideal for evaluation. To this end, corpora based on crowd-sourced simplifications have been proposed.
The TurkCorpus [205] is built by collecting data using the Amazon Mechanical Turk; workers were instructed to paraphrase Wikipedia sentences and keep the original meaning unchanged as much as possible. Each complex sentence is paired with 8 human

[2] Guidelines for Simple English Wikipedia authors are at `https://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages#Guidelines`

| | Source | Unit | Method | Instances |
|---|---|---|---|---|
| PWKP [218] | Wikipedia | Sentence | Auto. alignment | 108K |
| C&K-1 [40] | Wikipedia | Sentence | Auto. alignment | 137K |
| RevisionWL [201] | Wikipedia | Sentence | SEW revisions | 15K |
| AlignedWL [201] | Wikipedia | Sentence | Auto. alignment | 142K |
| EW-SEW [72] | Wikipedia | Sentence | Auto. alignment | 392K |
| C&K-2 [84] | Wikipedia | Sentence | Auto. alignment | 392K |
| sscorpus [79] | Wikipedia | Sentence | Auto. alignment | 493K |
| HSplit [176] | Wikipedia | Sentence | Manual simpl. | 359 |
| WikiLarge [214] | Wikipedia | Sentence | Pairs from [218, 84, 201] | 286K |
| TurkCorpus [205] | Wikipedia | Sentence | Crowdsourcing | 2,350 |
| ASSET [4] | Wikipedia | Sentence | Crowdsourcing | 2,350 |
| Doc-Level EW-SEW [179] | Wikipedia | Document | | 143K |
| Newsela [204] | Newsela | Document | Manual simpl. | 1,130 |
| Newsela-SS [204, 174, 6] | Newsela | Sentence | Auto. alignment | |
| SimPA [155] | Public adm. | Sentence | Manual simpl. | 1,100 |
| PLOS Goldsack et al. [58] | PLOS | Document | Manual | 27K |
| eLife [58] | eLife | Document | Manual | 4K |
| PLOS [116] | PLOS | Document | Manual | 28k |

TABLE 7.1: Datasets for text simplification in English. We call "Unit" the unit of simplification, i.e., sentence or document. We use "Method" to refer to how the dataset has been constructed: large Wikipedia corpora derive from human-written articles written for different targets; leads are used in document-level datasets, while sentence-level corpora are built either by automatically aligning sentences or by exploiting revisions. Smaller Wikipedia-related sentence-simplification corpora (mostly used for evaluation) take complex sentences from the English Wikipedia and use either use crowdsourcing to obtain simplifications or simplify the complex sentences manually. We also report some dataset from domains different than Wikipedia, namely the public administration (sentence-level) and the scientific literature (document-level).

simplifications. In accordance to the instruction, the performed simplification does not involve much syntactic simplification, concept explanation, or content reduction.

To overcome these limitations, ASSET (Abstractive Sentence Simplification Evaluation and Tuning) [4] was proposed. ASSET consists of crowdsource simplifications of the same original sentences from TurkCorpus; 10 simplifications are associated with each complex sentence. Workers were explicitly encouraged to consider different types of transformations and shown examples involving lexical paraphrasing, sentence splitting, compression, and a mix. Authors show the dataset is more abstractive than TurkCorpus and other datasets.

Finally, while the datasets we described so far tend to contain various types of simplifications, Sulem, Abend, and Rappoport [176] proposed HSplit, a small dataset obtained by manually simplifying the test set of the TurkCorpus that specifically focuses on sentence splitting.

### 7.3.2 Newsela

The Newsela dataset contains 1,130 news articles at 5 different levels of complexity, corresponding to different levels of education. The simplification is performed manually by professional editors.

Newsela is generally considered of higher quality that Wikipedia-based datasets.

However, it comes with a very restrictive license: it requires signing a data usage agreement, can be used for non-commercial purposes only, and requires articles using the dataset to be approved. Moreover, the license prevents redistributions: thus, no public train, validation and test splits exists. This makes the dataset not particularly popular in the research community, since it is difficult to use and problematic for reproducibility.

### 7.3.3 Datasets from other domains

Datasets not related to Wikipedia or the news are less common.

Scarton, Paetzold, and Specia [155] proposed SimPA, a dataset that simplifies documents in the public administration domain. The corpus contains manual simplifications performed in two stages: lexical simplification first, followed by syntactic simplification. The simplifications were performed by human experts, and the dataset is thus regarded as high-quality. However, the dataset is small, with 1,100 original sentences.

Recently, Goldsack et al. [58] proposed two datasets for lay summarization of biomedical journal articles. Thus intended for lay summarization, the datasets contain the original article, the original technical abstract, and a lay abstract; thus, they can be explored for document-level text simplification (from the technical to the lay abstract). The first dataset is obtained from PLOS[3], The Public Library of Science, an open-access publisher,

[3] https://journals.plos.org/plosone/ [Last accessed: March 2023]

using the technical and the author's submitted lay summary [4]. The other dataset is obtained from eLife[5], a journal focusing on biomedical and life sciences. The journal publishes a digest with a selection of publications that contains a simplified summary of the work written by a professional editor.
PLOS was also exploited by Luo, Xie, and Ananiadou [116] as a source for a similar dataset.

## 7.4 Measuring simplification quality

Similarly to summarization, simplification is an open-ended task, and multiple simplification candidates might be acceptable. Thus, human evaluation is the gold standard, while automatic metrics are used as proxies of model performances.

### 7.4.1 Automatic metrics

In this section, we will consider two main automatic metric types: we call readability scores metrics that only assess the simplicity of readability of text without a reference. Instead, we refer as simplification metrics to metrics that are commonly used to evaluate the output of simplification systems (and are largely reference-based).

#### Reading scores

**Flesch Reading Ease score [52]** computes sentence simplicity as a function of the mean word length (in syllables) and the mean sentence length (in words). A higher score indicates a text that is easier to read.

$$FRE = 206.845 - 1.015\frac{\text{total words}}{\text{total sentences}} - 84.6\frac{\text{total syllables}}{\text{total words}}$$

**Flesch Kincaid Grade Level [87]** is very similar to the Flesch Reading Ease score, but it is weighted to correspond to U.S. education grade levels. We report it for completeness. A lower score indicates a text that is easier to read.

$$FKGL = \frac{\text{total words}}{\text{total sentences}} + 11.8\frac{\text{total syllables}}{\text{total words}} - 15.59$$

#### Reference-based metrics

**SARI [205]** (System output Against References and Input sentence) compares the candidate simplification against both the reference(s) and the original sentence. It is currently by far the most popular metric for automatic simplification. It measures how "good" the tokens that are added, deleted, and kept in the simplification are.

The metric rewards adding tokens not in the input but present in the reference. Considering the model output O, the input sentence I, and the references R, one can compute add precision and recall as:

[4] The authors' summary guidelines are accessible at https://journals.plos.org/plosgenetics/s/ submission-guidelines [Last accessed: March 2023]

[5] https://elifesciences.org/

$$
\begin{aligned}
p_{\text{add}} &= \frac{\sum_{g \in O} min(\#_g(O \cap \bar{I}), \#_g(R))}{\sum_{g \in O} \#_g(O \cap \bar{I})} \\
r_{\text{add}} &= \frac{\sum_{g \in O} min(\#_g(O \cap \bar{I}), \#_g(R))}{\sum_{g \in O} \#_g(R \cap \bar{I})}
\end{aligned}
\tag{7.1}
$$

Where $\#_g()$ is a binary indicator of the occurrence of the n-gram g in a given set, $\#_g(O \cap \bar{I}) = max(\#_g(O) - \#_g(I), 0)$ and $\#_g(R \cap \bar{I}) = max(\#_g(R) - \#g(I), 0)$.

The metric also rewards tokens that are kept in both the output and references.

$$
\begin{aligned}
p_{\text{keep}} &= \frac{\sum_{g \in I} min(\#_g(I \cap O), \#_g(I \cap R'))}{\sum_{g \in I} \#_g(I \cap O)} \\
r_{\text{keep}} &= \frac{\sum_{g \in I} min(\#_g(I \cap O), \#_g(I \cap R'))}{\sum_{g \in I} \#_g(I \cap R')}
\end{aligned}
\tag{7.2}
$$

Where $\#g(I \cap O) = max(\#_g(I), \#_g(O))$ and $\#g(I \cap R') = max(\#_g(I) - \#g(R)/r, 0)$ and e $R'$ marks the n-gram counts over $R$ with fractions.

Finally, for deletion:

$$
p_{\text{del}} = \frac{\sum_{g \in I} min(\#_g(I \cap \bar{O}), \#_g(I \cap \bar{R}'))}{\sum_{g \in I} \#_g(I \cap \bar{O})}
\tag{7.3}
$$

Where $\#g(I \cap \bar{O}) = max(\#_g(I) - \#_g(O), 0)$.

The precision and recall are used to compute and F1 metric, and the final SARI is computed as

$$
\text{SARI} = \frac{F_{\text{add}} + F_{\text{keep}} + F_{\text{del}}}{3}
\tag{7.4}
$$

**BLEU [140]** (BiLingual Evaluation Understudy) is the most popular metric in machine translation. It is computed as the product of a brevity penalty term and a harmonic mean of n-gram precisions. The highest, the better.

$$
\begin{aligned}
\text{Brevity penalty (BP)} &= \begin{cases} 1 & \text{if } |C| > |R| \\ e^{1-|R|/|C|} & \text{otherwise} \end{cases} \\
\text{BLEU} &= \text{BP} \times e^{\sum_1^N w_n \times ln(p_n)}
\end{aligned}
\tag{7.5}
$$

Where $|C|$ is the length of the candidate, $|R|$ is the length of the reference. In general, $N = 4$ and $w_n = \frac{1}{N}$

BLEU has been shown to correlate well with grammaticality and meaning preservation, but has been previously criticized as a metric for simplification as it is unable to evaluate structural and syntactical simplification [176]; moreover, it rewards simplifications that are very similar to the complex sentence [205], and thus strongly favors meaning preservation over simplicity.

**BERTScore** [213], which we discussed in Chapter 3, can also be used for automatic simplification. Specifically, Alva-Manchego, Scarton, and Specia [3] show that it correlates well with human evaluations and suggest using the metric for simplification that involves multiple transformations in reference-based settings.

**SAMSA** [177] (Simplification Automatic evaluation Measure through Semantic Annotation) aims at measuring structural sentence simplification only, without a reference. SAMSA rewards sentences that are split so that each sentence contains a single event and the main relations of such events are preserved. The metric has been shown to correlate with experts on structural simplicity.

Other token-based metrics are often machine-translation inspired and include BLEU variants (e.g., iBLUE [178]), or edit metrics (e.g., TER [170] (Translation Edit Rate)), among others. Though other model-based approaches have been proposed, they still have to gain traction in the simplification community.

### 7.4.2 Human evaluation

Given the task complexity, human evaluation is considered the gold standard for evaluating simplification outputs, despite its limitations (that we described in Chapter 3 in the context of summarization).

In general, judges are asked to consider the following dimensions:

**Grammaticality** and fluency of the evaluation output (commonly on a Likert scale)

**Adequacy** (or meaning preservation), i.e., if the core original meaning is preserved in the simplification output, commonly on a Likert scale.

**Simplicity** i.e., if the simplification output is, in fact, easier to read and understand. Simplicity can be measured on a positive Likert scale (e.g., 0 to 5) or on a negative-to-positive scale centered in 0, where negative values indicate that the output is, in fact, more complex than the original [135]. In most cases, the concept of "simplicity" is not explicitly defined, and judges are encouraged to use their intuition; in a few cases, authors have asked to quantify the gain, e.g., by counting the positive changes introduced in the simplification output [205].

In the best case, judges are recruited among the simplification target. Recently, human evaluators have often been required through online services like Amazon Mechanical Turk or Prolific.

# Chapter 8

# Patent Simplification

This chapter describes the work carried out on the simplification of patent documents. Patents are legal documents that aim at protecting inventions on the one hand and at making technical knowledge circulate on the other. However, as described in Chapter 2, their style is very complex, as they include a mix of legal, technical, and extremely vague language, making their knowledge hard to access for humans and machines.

In this chapter, we propose an approach to automatically simplify patent text through rephrasing. As we saw in Chapter 7, state-of-the-art methods for sentence simplification rely on supervised monolingual translation systems, which require parallel data with both complex and simplified text. Although parallel simplification data exist, e.g., in general domains or the news, they are challenging to obtain in many technical domains, including patents. To alleviate this issue, we show how to construct a silver standard for in-domain patent simplification by filtering a set of candidates obtained through a general-domain paraphrasing system. While the process of obtaining the candidates is difficult to control and error-prone, we can pair it with proper filters and construct a cleaner corpus that can successfully be used to train a controllable simplification system. Human evaluation of the synthetic silver corpus shows that it is considered grammatical, adequate, and contains simple sentences.

## 8.1 A silver standard for patent simplification

As discussed in Chapter 2, patent documents are extremely complex: they contain long sentences (especially in the Claim section), novel multiterm entities, and complex syntax built out of noun phrases instead of clauses with recurring entities. Suominen et al. [180] performed a user study on the readability of patent text: Most participants strongly agreed (35%) or somewhat agreed (23%) that improving the readability of patents was important, as they considered documents extremely (29%) or somewhat (29%) difficult to read. Moreover, 39% of the participants reported difficulties in finding information they were looking for.

Even scholars in the law domain have advocated for the use of simpler language in patent documents. Feldman [50], for example, argues that "When the subject of the case is wrapped in complex and unfamiliar terms, it is tremendously difficult for legal actors to grapple with the theoretical content of the dispute. [...] Communication at the intersection of law and science will always be tremendously challenging. Nevertheless, there are elements of the current patent system that substantially exacerbate the problem".

As we saw in Chapter 2, previous work on patent simplification has mainly targeted experts and other figures involved in the patent filing process itself: systems search ways to help patent attorneys understand the structure of the claims (the legal part of the document, which defines the scope of the legal protection) through ad hoc visualizations or compare documents in the same patent family[1]. Since the scope of the legal language needs to remain unchanged, modifying the text presentation is preferred to rephrasing.

In this chapter, we take a different stance and aim at creating a silver standard for simplifying the Description of the invention — the section that describes the invention embodiments in detail — through rephrasing. Simplifying the Description makes the technical knowledge more accessible to society and its theoretical target, i.e., the "person skilled in the art" (practitioners in the field, engineers, academics, and other laypeople); moreover, it can improve the performance of automatic systems for the processing of text, as shown in other domains [47].

The current state of the art in sentence simplification through rephrasing relies on parallel corpora of complex and simple sentences. It frames the problem as a monolingual translation task: using a sequence-to-sequence network, complex sentences get translated into their corresponding simple versions. However, parallel data for simplification are difficult to obtain, and many of the existing large-scale datasets derive from the automatic alignment of sentences in the English and Simple English Wikipedia [218, 214], as we saw in Chapter 7. Creating parallel datasets requires considerable human effort for other domains, making the process slow and expensive.

To the best of our knowledge, no simplified corpora exist for the patent domain, and manually creating one would require considerable effort and likely involve legal and technical experts. In this chapter, we propose a method to automatically create a parallel simplification corpus of patent sentences and show that the corpus can be used to train a controllable system. In particular, we generate a huge noisy corpus of simplification candidate pairs (that we call "bronze") and clean it to obtain a higher-quality silver corpus. For generating the candidates, we adopt a paraphrasing system trained on general-domain text and show that using a zero-shot approach on patent sentences generates text that is often simpler, shorter, and easier to read. The process is, however, hard to control and might be error-prone. Thus, we discuss filters to select reliable candidates only and show how the silver corpus can be used for training controllable simplification systems for patent sentences.

Our contributions are the following:

1. We propose the use of a paraphrasing system to obtain simplification candidates from complex patent sentences. The paraphraser is used in a zero-shot fashion and is only trained on out-of-domain general English sentences. We show that, while far from perfect, this zero-shot approach usually produces text which is more compressed and thus simpler to read. We call the simplifications obtained this way a "bronze corpus".
2. We discuss filtering to select candidates that are appropriate for simplification only. By doing so, we generate the first large-scale parallel silver standard for patent

[1] A patent family is a collection of patent applications covering the same or similar technical content.

sentence simplification. We make the bronze and silver parallel datasets public for future research[2].

3. We show that the silver standard can be used to train a sequence-to-sequence state-of-the-art system for controllable patent simplification that we release.

4. Finally, we perform a human evaluation of the results and make the unaggregated data public. During the evaluation, we also collect non-expert-generated simplification, which can be used in future research.

This chapter is organized as follows: in Section 8.2, we discuss previous work in the general domain of sentence simplification and in the specific domain of patent simplification. In Section 8.3, we describe our proposed Method. Section 8.4 reports automatic metrics computed on the corpora and a qualitative error analysis. Finally, we detail our human evaluation campaign on the generated silver standard in Section 8.5. We draw our Conclusions in Section 8.6.

## 8.2 Previous work

### 8.2.1 Silver standards for sentence simplifications

Modern sentence simplification architectures require a large number of parallel complex and simple sentences to train.
As we saw in Chapter 7, most popular simplification datasets are obtained by automatically aligning sentences from corresponding documents in the English and Simple English Wikipedia.
However, obtaining corpora in other domains and languages is more complex, as no natural alignment between documents usually exists. Only a few manually curated datasets have been created (e.g., the Newsela corpus [204]), as they imply a huge human effort, which is expensive and requires considerable time.

Researchers have thus proposed ways to automatically generate silver standards for simplification. Starting from a Japanese corpus of mixed complexity, Kajiwara and Komachi [80] first identified a complex and a simple subcorpus based on readability scores; then, they aligned sentences from the two corpora based on their word embeddings nearest neighbors; they experimented with several similarity metrics.
Similarly, Martin et al. [121] used the LASER [10] sentence embedding and retrieved nearest neighbors with some filters to ensure quality.
While interesting, these methods are hard to apply to the patent domain, whose textual content has largely the same (high) complexity, as strict rules and common patterns govern the style.

In contrast, Lu et al. [114] proposed a method to turn a translation corpus into a simplification corpus. Given a pair of sentences in two languages, one is used as a bridge and translated into the other (target) language. Authors argue that the two sentences will likely have a different complexity level because machine translation models tend to output high-frequency tokens [61], and there is often a difference in complexity between languages in translation corpora [17]. If the translation is satisfactory, and if there is a

[2] The corpora can be accessed at: `https://github.com/slvcsl/patentSilverStandard`

difference in the complexity level, the sentence pair is added to the simplification silver standard. Authors show that using a large filtered silver corpus obtained this way outperforms a smaller, cleaner corpus. This approach shifts the burden from a parallel simplification corpus to a parallel translation corpus, which is typically easier to obtain.

All described methods were proposed for general-domain simplification, and we are not aware of applications to the patent domain.

Our method is inspired by previous literature on generating silver data for simplification in that we propose to select relevant pairs from a larger corpus of possible candidates. However, we do not rely on in-domain simple data nor on external parallel data.

### 8.2.2 Patent simplification

An in deep discussion of patent simplification is available in Chapter 2. Here, we briefly discuss some of such works to better frame our contribution.

As patent claims are hard to read even for patent professionals, most effort in previous work has been spent on improving the accessibility and readability of the Claim section, targeting patent experts.

Ferraro, Suominen, and Nualart [51], for example, aimed at improving each claim presentation by segmenting it into fragments that are then formatted more clearly, e.g., by adding new lines. Okamoto, Shan, and Orihara [137] used an Information Extraction engine that detects entity mentions, their type, and relations through distant supervision. They built an interface to show the most salient elements in the Claim section to understand the patent structure and compare patents in the same family. These works target patent attorneys and other experts involved in the patent filing process. In contrast, Suominen et al. [180] proposed ways to improve patent visualization to lay people, all of which were considered at least as good as the original patent text by users. Some previous research has also tried to improve the understanding of entities in the claims by linking them to the Description, where they are mentioned in the context of actual embodiments [164]. Finally, other previous work has visualized claims in a more structured way, e.g., through graphs [8] or trees [160].

Previous work on simplification through rephrasing is much more limited. A rewriting and rephrasing system was built as part of the PATExpert project [198]. Researchers considered two levels of simplification: one uses surface criteria to segment the input and reconstructs chunks into shorter, easier-to-read sentences [20]. The other [126] represented patents by their Deep Syntactic Structures. This representation is, in turn, used to rewrite a text that is simpler to process for the reader (possibly in another language). Both methods modify the patent text.

## 8.3 Method

This section discusses our method to obtain a parallel silver standard of patent sentences for simplification. We also show how the corpus can be used as a training corpus for a simplification system.

### 8.3.1 Dataset

We use data in the Patent Translation Resource (PatTR) [203, 199] corpus. The data is available under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. PatTR is a sentence-level parallel corpus extracted from the MAREC patent collection[3]. It consists of sentence pairs for translation (German-English, French-English, and German-French). For sentences in the Description (that we will use), patent families were exploited to create sentence pairs in different languages; specifically, German and French documents from the EPO corpus were aligned to documents in English from the United States Patent and Trademark Office (USPTO) corpus, following Utiyama and Isahara [191]. In all cases, sentences in corresponding sections were automatically aligned using the Gargantua aligner [22]. Sentence pairs are indexed by language and patent sections, i.e., Title, Abstract, Description, and Claims.

In the following experiments, we will focus on the English sentences from the German-English pairs and specifically on sentences extracted from the Description only. The PatTR German-English Description dataset contains almost 12 million sentence pairs. For computational reasons, we sample 500 thousand sentences.

As a preprocessing step, we removed sentences shorter than 5 tokens or longer than 55 tokens. We also filter out sentences where alphabetic characters account for less than 60% of the total. These sentences mainly contain long lists of references or complex chemical formulas. This leaves us with 425,148 sentences. Finally, sentences in the Description often contain references to the figures in the form of numbers in brackets. We use regular expressions to remove these references.

We chose the PaTR as it contained pre-tokenized sentences, and finding sentence delimiters is an error-prone task for patents. Moreover, the PaTR dataset allows us to compare our method with the general-domain method proposed by Lu et al. [114], which requires a parallel translation corpus.

Table 8.1 contains some statistics on the English subset of the corpus, while Figure 8.1 shows their distribution. We use the following metrics as simplicity proxies:

- Flesch Reading Ease score [52], which we discussed in Chapter 7.
- Flesch Kincaid Grade Level [87], which we discussed in Chapter 7.
- WordRank score: this measure uses word frequency as a proxy of lexical simplicity (as text containing more common words is considered easier to understand). It is computed by taking the third quartile of log ranks (inverse frequency order) of all words in the sentence. We use the implementation proposed by Martin et al. [120].

  Specifically, it uses the word rank as computed from a word embedding (FastText [19], in our implementation), computes the sentence complexity as

[3] See: `http://www.ifs.tuwien.ac.at/imp/marec.shtml` [Last accessed January 2023]

| Metric | Mean ± std |
|---|---|
| Flesch Reading Ease [52] | 32.5 ± 26.3 |
| Flesch–Kincaid Grade Level [87] | 61.1 ± 9.5 |
| WordRank [120] | 9.8 ± 1.1 |
| Max dependency tree depth | 6.7 ± 2.5 |
| Length (chars) | 170.3 ± 75.3 |

TABLE 8.1: Statistics on the original English patent sentences (after preliminary filtering)

$$Q_3\Big(\{log(1 + rank(w_i))\}_{i=0}^{N}\Big)$$

where $Q_3$ indicates the third quartile of the set, $rank(.)$ computes the word ranking and $w_i$ is a token in sentence $S$, of length $N$.

- Maximum Dependency Tree: this score uses the height of the dependency tree as a proxy for syntactic complexity. We adopt the implementation proposed by Martin et al. [120], which uses spaCy[4] [67] for computing the dependency tree.

All metrics are computed per sentence and then averaged.

### 8.3.2 Generating simplification candidates using a general-domain paraphrasing system

Previous work has explored using models trained on out-of-domain simple text to simplify the complex text. Surya et al. [181], for example, propose an architecture with two decoders (one trained on complex text only and the other trained on simple text only) to control the level of simplification without supervision.

Inspired by this line of work, we investigate using a sequence-to-sequence system trained on general-domain text only for simplifying complex patent text. Specifically, we employ a Pegasus-based [212] paraphrasing system for general text. The system was fine-tuned on a custom set of 60k examples from multiple datasets, including PAWS [215]. The model is available on Huggingface[5]. We use the trained model as a black box and apply the model to the preprocessed patent sentences.

Table 8.2 reports some random sentence pairs (no cherry-picking) generated using Pegasus. The simplification is mainly by sentence compression and lexical; some other syntactical changes (e.g., use of the active voice instead of the passive voice) are also common.

While the generated candidates tend to be simpler than the original sentences, using the Pegasus model directly for simplification has two main limitations:

1. Some of the generated sentences contain errors and unknown tokens or are too similar or excessively compressed with respect to the original sentence.

[4] `https://spacy.io/` [Last accessed: March 2023]

[5] `tuner007/pegasus_paraphrase` [Last accessed: January 2023]

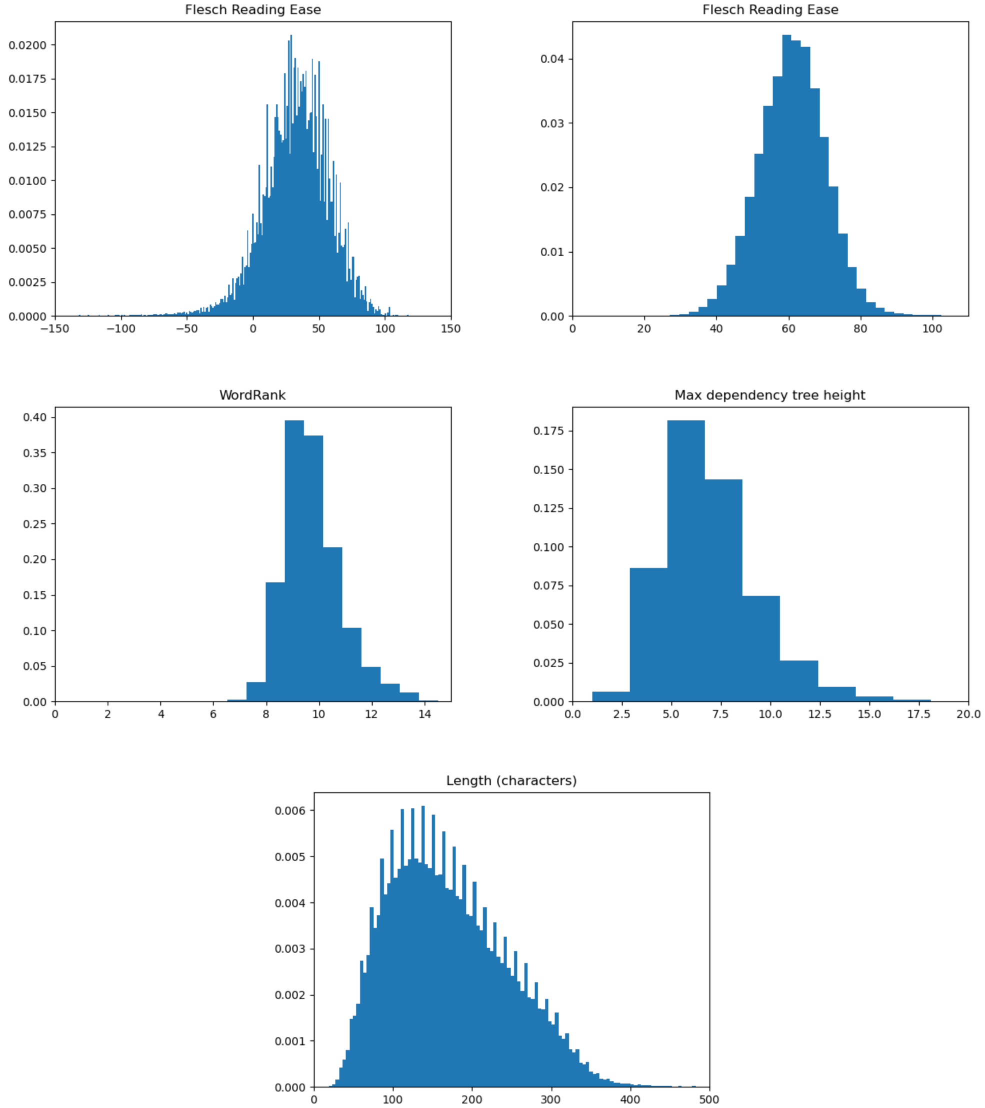


FIGURE 8.1: Metric distribution of the original dataset

2. The process is hard to control. For example, while state-of-the-art simplification systems allow controlling the level of compression through appropriate tokens, the Pegasus model has a strong tendency to compression. However, a too-strong compression of the original sentence can degenerate into text that does not contain the original core elements, loses too much content, or whose meaning is completely different than the original. The level of paraphrasing, similarity, and lexical simplification is also hard to control.

Due to these issues, in the following, we will call the candidates generated by the Pegasus model our "bronze corpus".

#### Comparying the candidates with Lu et al. [114]

For comparison, we also report the same candidates obtained through translation (prior to any filtering) using the model proposed by Lu et al. [114] in Table 8.3.

Their proposed method works in several steps:

1. Given a parallel translation corpus, translate the sentences in the bridge language (German, in our case) to the target language (English, in our case).
2. Verify that the output is not too short, is not equal to the original sentence, does not contain unknown tokens; verify that the original alignment is correct and the translation is satisfactory (i.e., its BLEU score with the original sentence is greater than a threshold), and there is a difference in the complexity between the original English sentence and the translated sentence (i.e., the absolute value of their difference in Flesch Reading Ease score is above a threshold).
3. In case the translated sentence is considerably simpler than the original one, consider the original one as the "complex" sentence and the translated one as the "simple" one and add them to the corpus; in case the translated sentence is considerably more complex than the original one, consider the translated sentence as the "complex" one and the original as the "simple" one, and add them to the corpus. If there is no difference in complexity, discard the pair.
4. Use the corpus for training a seq2seq system.

In this section, we explore results obtained after step 1 of the process, which generated candidates to be later filtered. To generate the candidates, given the PatTR German-English sentence pairs, we translate the German sentences to English and obtain sentence pairs in English.

We use the model proposed in [114] for this step: the model is an ensemble of 4 models trained on the WMT19 German-English data. We do not fine-tune the models with patent data due to our limited computational resources.

Lu et al. [114] claim translating from another language often produces sentences at a different level of simplification for two main reasons:

- Machine translation models tend to produce high-frequency tokens [61], and word frequency is often used as a proxy for lexical simplicity [61].

| | Original | Candidate |
|---|---|---|
| ~~0~~ | In interaction with the component secured on the cylinder head, the radial shoulder of the pressure medium distributor now prevents the camshaft from migrating axially further into the cylinder head. | The radial shoulder of the pressure medium distributor prevents the camshaft from moving further into the cylinder head when interacting with the component secured on the cylinder head. |
| 1 | Film 12 had holes, and these could be attributed to moisture in the blend material. | The blend material could be to blame for the holes in Film 12. |
| ~~2~~ | According to the invention, the formulations advantageously comprise one or more antioxidants. | The invention states that the formulations comprise one or more antioxidants. |
| 3 | In the pressure medium flow path between the consumers and the variable displacement pump a respective meter-in orifice is provided. | A meter-in orifice is provided in the pressure medium flow path between the consumers and the pump. |
| ~~4~~ | The yield of TMCH relative to the feed KIP was 27%. | The yield was 27%. |
| 5 | A throttle valve angle sensor for internal combustion engines which is described therein has a permanent magnet that is attached to one end of the throttle valve shaft and produces a constant magnetic flux having a direction which is turned as a function of the rotation of the throttle valve shaft. | A permanent magnet that is attached to one end of the throttle valve shaft produces a constant magnetic flux which is turned as a function of the rotation of the throttle valve shaft. |
| 6 | They have to be so thin in comparison with the wavelengths produced that they may be caused to vibrate by the ultrasonic waves. | They have to be so thin that they can't be seen by the naked eye |
| ~~7~~ | Mechanisms are provided which allow the guide rollers to yield under spring tension so that the force with which the rollers are pressed on the pane is held within a prescribed range regardless of installation and vehicle body tolerances. | The force with which the rollers are pressed on the pane is held within a prescribed range, regardless of installation and vehicle body tolerances, because of mechanisms provided which allow the guide rollers to yield under spring tension. |
| 8 | The line includes an outer conductor tube which is equipped on both ends with connection flanges . | The line has an outer conductor tube on both ends. |
| 9 | Thick pressure-sensitive adhesive coats may be present in unfilled, all-acrylate form or resin-blended form or filled with organic or inorganic fillers. | Thick pressure-sensitive adhesive coats can be filled with organic or inorganic fillers. |
| ~~10~~ | The seamless tubular casing preferably has a thickness of from 40 to 200 \u03bcm, when it is made by a blown tube process, and a thickness of from 25 to 75 \u03bcm, when it was obtained by biaxial orientation (double bubble process). | When it is made using a blown tube process and a thickness of 25 to 75?m is obtained, the seamless tubular casing should have a thickness of 40 to 200?m |
| 11 | Furthermore, with this type of integration it is mainly the classic MC functionalities, as are relevant in particular for machine tools, that are supported. | The classic MC functionalities are relevant in particular for machine tools that are supported by this type of integration. |
| ~~12~~ | Then, if a predetermined limiting pressure value is reached, the valve body member 19 is moved against the pretension force of the spring member 18, thus clearing a further conduit with respect to the liquid treatment apparatus 3. | If a limiting pressure value is reached, the valve body member 19 is moved against the pretension force of the spring member 18, thus clearing a further conduit with respect to the liquid treatment apparatus 3. |
| 13 | These process parameters for intermingling depend on the yarn titer and the lubricant applied to the yarn and must therefore be adapted accordingly. | The process parameters for intermingling are dependent on the yarn titer and the lubricant applied to the yarn. |
| 14 | The telephone network is formed as a digital telephone network (e.g. Integrated Services Digital Network, ISDN) or as an analogue telephone network (e.g. Public Switched Telephone Network, PSTN). | The telephone network can be formed as a digital telephone network or an analogue telephone network. |
| ~~15~~ | This may be perceived as disadvantageous by the consumer. | This may be seen as disadvantageous by the consumer. |
| 16 | On the other end of the coupling rods 20, 21, these are connected by way of second ball-and-socket joints 22, 23 as the second connection joint of the parallelogram linkage with the second axis 24 which extends in parallel to the first axis 19. | The second connection joint of the parallelogram linkage with the second axis 24 extends in parallel to the first axis 19 on the other end of the coupling rods. |

TABLE 8.2: Simplification candidate pairs obtained through the Pegasus paraphrasing model (zero-shot) prior to any filtering. Pairs with a crossed-out number are filtered.

- There is often a difference in complexity levels for sentences in different languages in translation corpora [17], which causes a difference in complexity when one sentence is translated to the target language. Note that, however, this effect is likely small for patents, given their legal nature.

Note that for a sentence pair to be "useful" for simplification, it is enough that there is a complexity difference between the original and the generated sentence in any direction: If the generated sentence is more complex than the original one, their role in the simplification corpus can be swapped (step 3 described above).
However, the sentences are practically paraphrased, and there is no evident difference in complexity. We attribute these characteristics to the lack of complexity differences in sentences in different languages in the patent domain. Our preliminary experiments show that even after filtering these instances for simplification level, e.g., maintaining only pairs with a high relative difference in reading scores, as the original method proposes, the level of simplification is poor.

**Automatic metrics** Figure 8.2 shows some automatic metrics for evaluation as computed on the whole corpus obtained through translation following Lu et al. [114], prior to any filtering. In addition to the simplicity proxies discussed in Section 8.3.1, we also compute similarity scores among the complex and the simple sentences:

- Normalized character-level Levenshtein similarity[6]: it is the number of insertions, deletions, and substitutions between the complex and the simple sentences normalized by the sum of their length.
- BLEU [140]: We use it here as a proxy of token-based similarity and meaning preservation. Higher is better. We used the NLTK [18] implementation of BLEU with a smoothing function. Specifically, we smooth by adding 1 to the numerators and the denominators of the n-gram precision terms, as proposed by Lin and Och [106]. Therefore, when no 4-gram matches are present (for example, for short generated sentences), we still get a positive smoothed BLEU score from shorter n-gram matches. BLEU is often used with modifications in previous simplification works.
- BERTScore [213]: We use the HuggingFace implementation. Higher is better.

All metrics are computed per sentence or sentence pairs and then averaged. When required, the original complex sentence is used as the reference.

We notice that the distribution of the simplification scores of the original and generated sentences are practically coincident. In practice, there is no appreciable difference in complexity: instead, all automatic measures indicate that the translated corpus is considered slightly more complex than the original one. Interestingly, the translated sentences are also, on average, slightly longer than the original ones.

[6] Implementation: `https://github.com/maxbachmann/RapidFuzz`

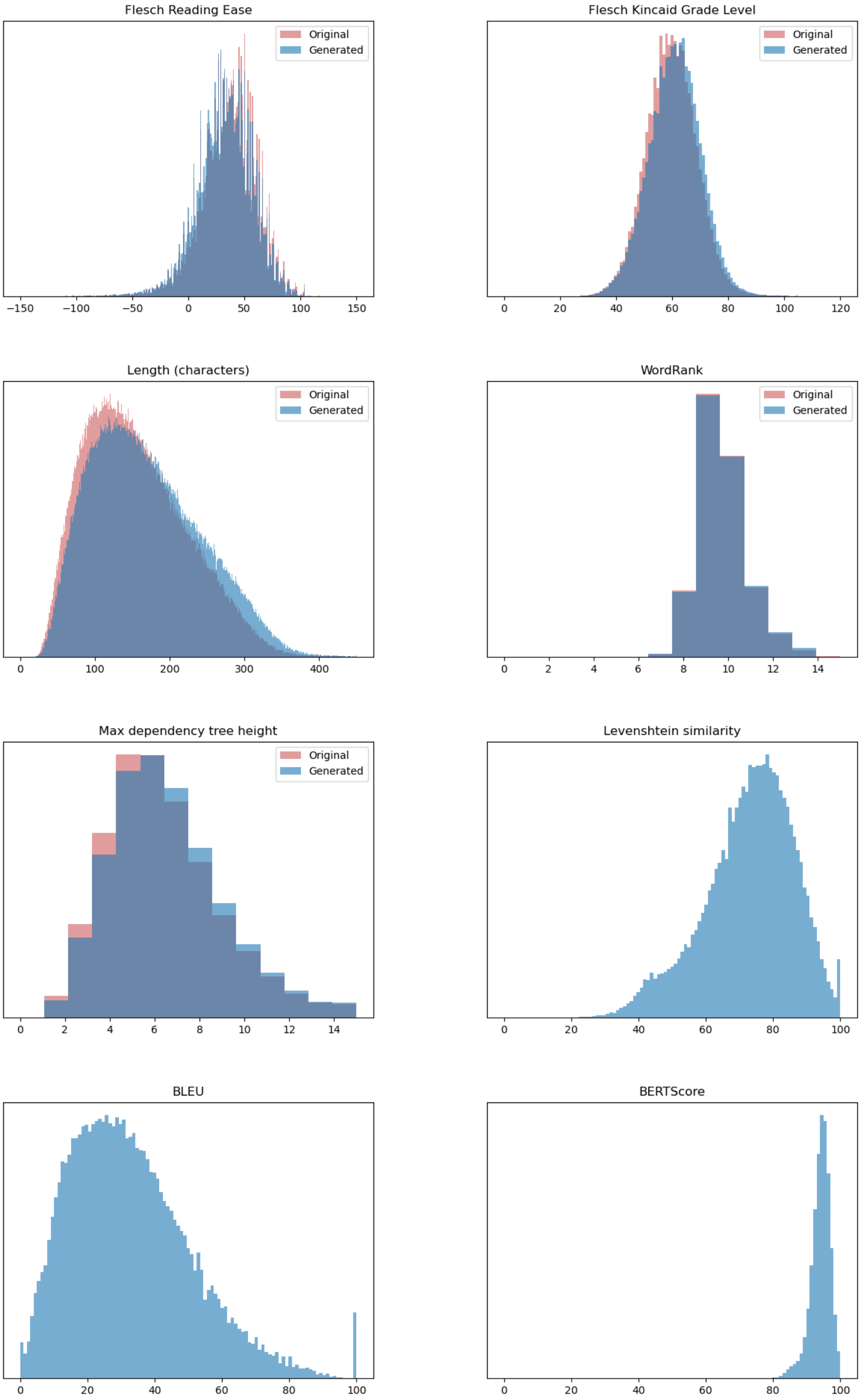


FIGURE 8.2: Automatic metrics as computed on the sentence pairs candidates obtained following Lu et al. [114].

**Qualitative analysis** We also manually annotated 100 original/translated pairs. We considered the following dimensions:

- **Grammaticality**: Grammaticality is not an issue and the vast majority of translated sentences are very fluent.
- **Compression** Since we are using a translation model, the generated sentence is generally complete and has no deletions. Some content deletion likely occurs since a few English and German pairs do not have the same content in the translation dataset.
- **Simplicity** We note that both the original and the translated sentences tend to be extremely complex. We do not note a clear pattern for which translated sentences contain simpler tokens. Only in 10% of the sentences we annotated, we noticed an appreciable difference in simplicity (in either direction), while in 25% we noticed a minor difference. In 65% of cases, we noticed no difference in simplification among the sentence pairs.

Concluding, in contrast to ours, the approach proposed by Lu et al. [114] does not seem to produce candidates with different complexity levels.

### 8.3.3 Filtering bad candidates and generating a silver standard of patent sentences.

In the previous step, we generated a big (bronze) corpus of simplification candidate pairs using a general-domain paraphrasing model, and we showed that our candidates are better than those obtained following [114]. The bronze corpus is, however, very noisy.

Thus, we remove candidates according to the following criteria:

- Bad tokens: we remove any pair that contains an unknown token. We also noticed that the generation can degenerate into a sequence of specific numerical tokens (e.g., 65561). We remove any pair where the generated sentence contained any of these tokens. For example, this removes pair 10 in Table 8.2.
- Non-alphabetical characters: we remove pairs where the generated sentence has less than 60% of alphabetical characters. These sentences are not necessarily the effect of a bad generation, but the pair is useless as a training example.
- Similarity: we remove sentences whose character-level Levenshteins similarity between the original sentences (normalized by the sum of their length) simplification is < 25% or > 90%. The low threshold removes simple sentences that are too different compared to the complex ones (generally very short); the high threshold removes sentences that are equal to the original one or only differ by a small number of characters. This removes, for example, pairs 12 and 15. We also remove pairs in which the generated sentence has a similarity > 99% with a continuous substring of the original sentence. Finally, we removed pairs where the strings obtained by sorting the tokens have a similarity > 90%. This removes, for example, pairs 0, 2, and 7.

| | Original | Candidate |
|---|---|---|
| 0 | In interaction with the component secured on the cylinder head, the radial shoulder of the pressure medium distributor now prevents the camshaft from migrating axially further into the cylinder head. | In this way, the radial collar of the distributor forms the axial bearing of the camshaft in conjunction with the cylinder-head-resistant component. |
| 1 | Film 12 had holes, and these could be attributed to moisture in the blend material. | Foil 12 had holes that could be traced back to moisture in the glare material. |
| 2 | According to the invention, the formulations advantageously comprise one or more antioxidants. | According to the invention, the preparations contain advantageous one or more antioxidants. |
| 3 | In the pressure medium flow path between the consumers and the variable displacement pump a respective meter-in orifice is provided. | A diaphragm is provided in each case in the pressure medium flow path between the consumers and the variable displacement pump. |
| 4 | The yield of TMCH relative to the feed KIP was 27%. | The yield of THCM in relation to the KIP used was 27%. |
| 5 | A throttle valve angle sensor for internal combustion engines which is described therein has a permanent magnet that is attached to one end of the throttle valve shaft and produces a constant magnetic flux having a direction which is turned as a function of the rotation of the throttle valve shaft. | Here a throttle angle sensor for internal combustion engines is described with a permanent magnet attached to one end of the throttle shaft, which generates a constant magnetic flux, the direction of which is rotated depending on the rotation of the throttle shaft. |
| 6 | They have to be so thin in comparison with the wavelengths produced that they may be caused to vibrate by the ultrasonic waves. | These must be so thin in comparison to the wavelengths produced that they can be oscillated by the ultrasonic waves. |
| 7 | Mechanisms are provided which allow the guide rollers to yield under spring tension so that the force with which the rollers are pressed on the pane is held within a prescribed range regardless of installation and vehicle body tolerances. | In order to ensure that the force with which rollers 56 are pressed to disc 4 remains within a predetermined range, independent of installation and body tolerances, means 42.70 are provided to enable guide rollers 56 to be cushioned. |
| 8 | The line includes an outer conductor tube which is equipped on both ends with connection flanges . | The cable comprises an outer conductor tube 1, which is equipped with connecting flanges 2 at both ends. |
| 9 | Thick pressure-sensitive adhesive coats may be present in unfilled, all-acrylate form or resin-blended form or filled with organic or inorganic fillers. | Thick adhesive coatings can be available unfilled as pure acrylate or mixed with resin or filled with organic or inorganic fillers. |
| 10 | The seamless tubular casing preferably has a thickness of from 40 to 200 \u03bcm, when it is made by a blown tube process, and a thickness of from 25 to 75 \u03bcm, when it was obtained by biaxial orientation (double bubble process). | The seamless, tubular sheath of the invention preferably has a thickness of 40 to 200 \u00b5m if manufactured by a tube blowing process, and a thickness of 25 to 75 \u00b5m if obtained by biaxial stretch orientation. |
| 11 | Furthermore, with this type of integration it is mainly the classic MC functionalities, as are relevant in particular for machine tools, that are supported. | In addition, this type of integration mainly supports the classic MC functionalities, which are particularly relevant for machine tools. |
| 12 | Then, if a predetermined limiting pressure value is reached, the valve body member 19 is moved against the pretension force of the spring member 18, thus clearing a further conduit with respect to the liquid treatment apparatus 3. | If a certain limit value is exceeded, valve body 19 is shifted against the prestressing force of spring 18, thus releasing a further line with respect to liquid treatment device 3. |
| 13 | These process parameters for intermingling depend on the yarn titer and the lubricant applied to the yarn and must therefore be adapted accordingly. | These process conditions for fishing depend on the yarn titer and the preparation applied to the yarn and must therefore be adapted to these parameters. |
| 14 | The telephone network is formed as a digital telephone network (e.g. Integrated Services Digital Network, ISDN) or as an analogue telephone network (e.g. Public Switched Telephone Network, PSTN). | Telephone Network 2 is designed as a digital telephone network (e.g. Integrated Services Digital Network, ISDN) or an analog telephone network (e.g. Public Switched Telephone Network, PSTN). |
| 15 | This may be perceived as disadvantageous by the consumer. | This can be detrimental to the consumer. |
| 16 | On the other end of the coupling rods 20, 21, these are connected by way of second ball-and-socket joints 22, 23 as the second connection joint of the parallelogram linkage with the second axis 24 which extends in parallel to the first axis 19. | At the other end of the coupling rods 20, 21, these are connected via second ball joints 22, 23, as the second joint of the parallelogram guide, to the second axis 24, which runs parallel to the first axis 19. |

TABLE 8.3: Candidate pairs obtained through translation, as proposed by Lu et al. [114].

- Compression: we remove pairs where the ratio between the generated and the original sentence length was > 1.5 or < 0.5. This step avoids excessive compression, which generally corresponds to losing important content or having a modified meaning. This removes, for example, pair 4.
- Simplicity: candidates are, in general, simpler than the original sentences. We remove candidates where the generated sentence is not simpler as measured by the Fresh Reading Ease score, the WordRank score, or the height of their dependency tree.

Table 8.4 shows the number of pairs removed by each filtering step, with examples. Note that the filters were applied in sequence. We chose the thresholds heuristically. Future work could investigate learning their most suitable values from a larger corpus of silver sentences annotated for errors or study the best threshold between noise and size of the corpus. The remaining sentences compose the silver standard we will use as a parallel corpus. After filtering, our corpus consists of 287,965 samples.

### 8.3.4 Using the corpus for training a controllable simplification system

A silver standard for simplification allows for experimentation with models and training processes. To demonstrate this point, we train ACCESS [120], a state-of-the-art system for automatic sentence simplification, which we described in Chapter 7. We randomly split the silver standard into a train (184,297 samples), validation (46,075 samples), and test (57,593 samples) set.

We train the model with early stopping on the validation SARI [205] (patience: 20). We also evaluate the model on the human annotations we collected from Prolific workers (described in Section 8.5.3). Table 8.5 shows the results, as computed using the EASSE library [5].

As expected, the metrics on the silver standard are particularly high. This is likely because the corpus is synthetic and presents patterns that are easy to learn for the simplification model. Note, however, that when the model is evaluated on human-written sentences, its performance is close to those obtained on other general-domain datasets. Specifically, we have reported the results on the human evaluation set when using it both as a validation set (allowing the ACCESS model to optimize its preprocessing parameters) and as a test set (only using the optimal parameters computed on the silver standard validation set).

Having a silver parallel corpus also allows experimenting with transfer learning. We plan to explore the impact of consecutive fine-tunings (for example, further fine-tuning a model already fine-tuned on Wikipedia-style data). On the other hand, the performance of a model first fine-tuned on a large-scale silver dataset can usually be improved with a small human-curated dataset; this approach would diminish the human effort and, thus, the time and money required for obtaining high-quality simplifications. We plan to investigate these aspects in future work.

| Filter | Removed | Examples |
|---|---|---|
| Bad tokens | 8,602 | **Simple:** *"It is1-65561-65561-65561-65561-65561-65561-65561-65561-65561-65561-65561-65561-65561-65561-65561-65561-65561-65561-65561-[...]-65561-65561-65561-6556"* |
| Non-alphabetical | 695 | **Simple:** *"R.sup.1, R.sup.2, R.sup.3, Z.sup.1, Z.sup.2 and Z.sup.3 are all related."* |
| Similarity | 38,217 | **Complex:** *"Example 23 (E,E)-3,3-[4-(4-Fluorophenyl)-6-isopropyl-1-methyl-pyrid-2one-3,5-diyl]-diprop-2-enal ##STR54## Analogously to Example 10, 2.70 g of crude product are obtained starting from 2.13 g (7.1 mmol) of the compound of Example 22."* **Simple:** *"There are 2.70 g of crude product."*<br>**Complex:** *"Another method of solution attempts to achieve optimum operating cracks by means of temperature-dependent shifting of a conical rotor in the tapered case diameter with the help of an adjustment device."* **Simple:** *"Another method of solution attempts to achieve optimum operating cracks by means of temperature dependent shifting of a conical rotor in the tapered case with the help of an adjustment device."* |
| Partial similarity | 5,937 | **Complex:** *"In the treatment of parts of plants, the active compound concentrations in the use forms can be varied within a substantial range."* **Simple:** *"The active compound concentrations in the use forms can be varied within a substantial range."* |
| Sorted similarity | 18,807 | **Complex:** *"The solution polymerization without addition of other auxiliaries is the preferred process for the ethylene-vinyl acetate copolymers to be used according to the present invention."* **Simple:** *"According to the present invention, the solution polymerization without addition of other auxiliaries is the preferred method for the production of the ethylene-vinyl acetate copolymers."* |
| Compression | 62,926 | **Complex:** *"he bottom end of this rod passes through a hole 46 which leads into the fourth circular chamber 20."* **Simple:** *"The hole 46 leads into the fourth chamber 20."* |
| Simplicity | 1,999 | **Complex:** *"The preparation and further processing of the catalyst supports used according to the invention are well known to the person skilled in the art."* **Simple:** *"The person skilled in the art knows about the preparation and further processing of the catalyst supports used in the invention."* |

TABLE 8.4: Number of instances removed by each filter from our original 426,963 bronze standard and examples. The filters were applied consecutively.

| | SARI | BLEU |
|---|---|---|
| Validation (silver) | 55.09 | 54.88 |
| Test (silver) | 55.22 | 54.99 |
| Human simplification (with param. search) | 39.25 | 56.86 |
| Human simplification (without param. search) | 36.99 | 57.78 |

TABLE 8.5: Results of the model trained using the silver standard

| Metric | Complex | Simplified |
|---|---|---|
| Flesch Reading Ease [52] | 33.7 ± 24.5 | 47.6 ± 24.2 |
| Flesch–Kincaid Grade Level [87] | 60.8 ± 9.1 | 56.6 ± 8.5 |
| WordRank [120] | 9.7 ± 1.1 | 9.5 ± 1.2 |
| Max dependency tree depth | 6.6 ± 2.5 | 5.6 ± 2.2 |
| Length (chars) | 157.5 ± 67.4 | 108.4 ± 46.7 |
| Levenshtein similarity | 67.7 ± 12.9 | |
| BLEU [140] | 38.6 ± 15.5 | |
| BERTScore (avg) [213] | 94.45 ± 1.82 | |

TABLE 8.6: Statistics on sentence pairs and simplified sentences from the silver corpus

## 8.4 Corpus quality estimation

### 8.4.1 Automatic metrics

To study the dataset characteristics of the bronze and silver corpora, we compute several automatic metrics.

Figure 8.3 reports the metric distribution; means and standard deviations are summarized in Table 8.6.

One can notice that sentences in the silver corpus are simpler than the original ones as measured by the reading scores. The effect on the lexicon is smaller as measured by WordRank, while the effect of simplification is again clear on the syntax complexity as measured by the maximum dependency tree height. Moreover, simple sentences are generally significantly shorter than the original. Note how the filtering process removes several sentences practically identical to the original one.

### 8.4.2 Qualitative error analysis

We manually analyzed a subset of 100 sentences to identify the remaining errors.

Generated sentences are grammatical and fluent. Errors in meaning preservation are mainly of two types: excessive or wrong compression or problems with compositionality. With respect to incorrect sentence compression, coordinate elements in a sentence can be removed. For example, the original sentence *"Both the solid and the corrugated sheets preferably exhibit on one or both outer sides a layer consisting of the compositions according to the invention."* is paired with the sentence *"The corrugated sheets have a layer consisting of compositions on one or both sides."* In this case, the solid sheets, which appear in the original sentence, disappear in the simplified one.

Another possible, even if rare, error is the removal of important adjectives or modifiers. For example, the sentence *"In some cases, it has proved advantageous to use emulsion polymers exhibiting reactive groups at the surface."* has a simplified pair *"In some cases, it has proved to be beneficial to use emulsion polymers with groups at the surface."*

Another class of errors derives from the high compositionality of some sentences. In these cases, the relations among elements might not be fully grasped and can be reversed. An example is the following sentence *"The contact lugs, projecting from the bearing plate of the*

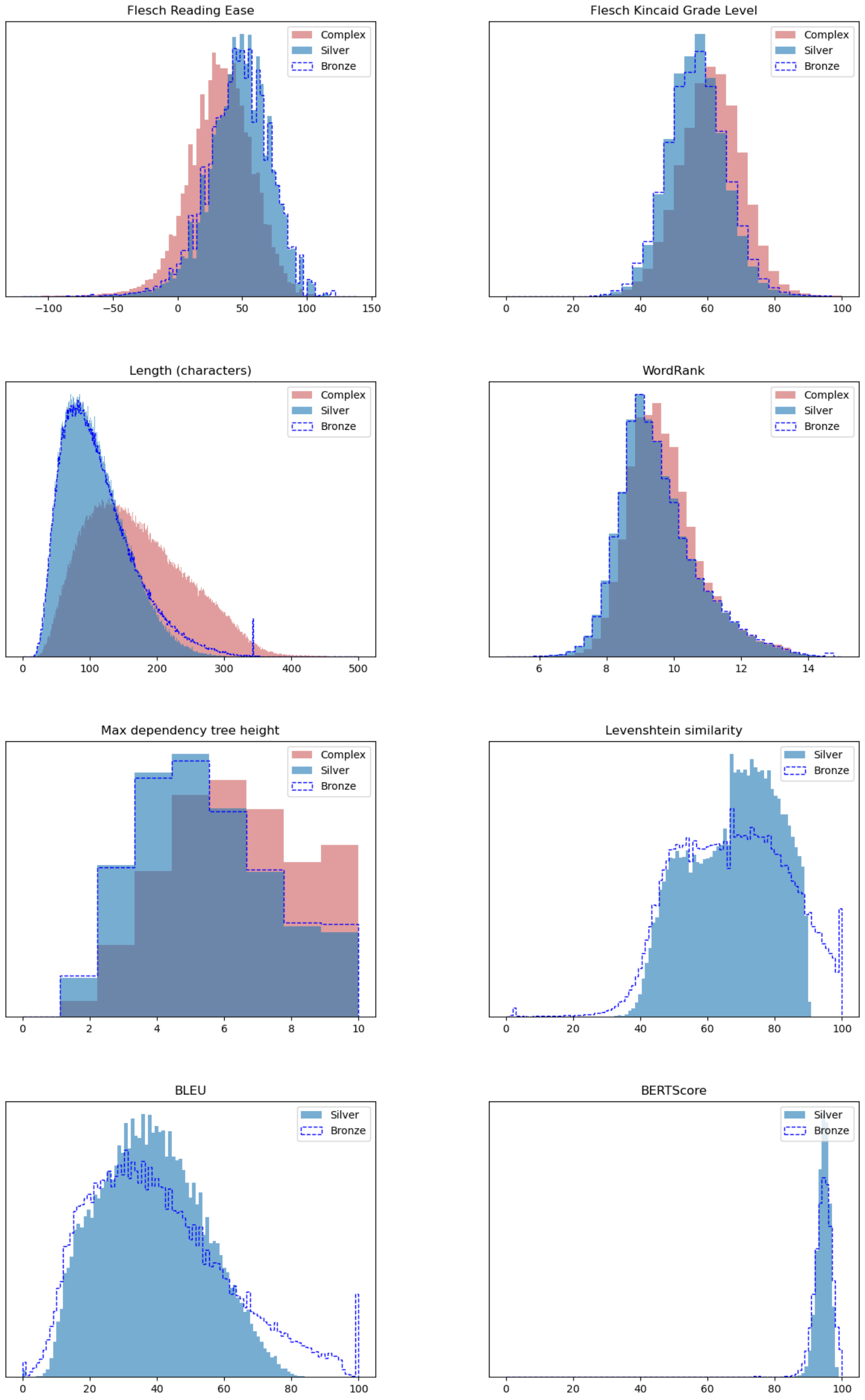


FIGURE 8.3: Automatic metrics as computed on the complex and simple sentences of the silver corpus. All histograms are normalized to have unit area.

*motor, for the electrical feed of the motor are easily accessible there for the connecting cabling in the housing installation orifice."* which is incorrectly paired with *"There is an easy way to connect the electrical feed of the motor to the housing installation orifice with the contact lugs projecting from the bearing plate."*

## 8.5 Human evaluation

### 8.5.1 Evaluation details

We recruited workers for human evaluation through Prolific[7] to quantify the quality of the silver standard as judged by laypeople. Prolific has been used for human evaluation or annotation, and some previous work has found results are more reliable in the analyzed case than those obtained with Amazon Mechanical Turk [142]. The survey interface was built using Qualtrics[8].

The survey consisted of a brief statement describing the goal of the study and the data collection process and asking for consent. The workers were informed that the data were collected for research purposes only but that they would be made available to other researchers and could be used for published work in scientific venues. Subjects were required to be at least 18 years old and native English speakers. After a brief demographic section asking for age, education, and current job, the users were introduced to the task.

Specifically, they had to judge the grammaticality and core meaning preservation (adequacy) on a 0-5 Likert scale [104]. Simplicity was measured in a -2 (the Simplified sentence is much more complex than the original one) to +2 (the Simplified sentence is much simpler than the original one) scale. Subjects were also asked to provide an overall 0-5 score and to write a simplification of the original sentence that they considered adequate. Numerical values had to be chosen through sliders.

Instructions (that we report in Figure 8.4) included a description of the meaning of Grammaticality, Core Meaning Preservation, and Simplicity, together with clarifications (e.g., regarding judging grammaticality by looking at the Simplified sentence only). We excluded subjects who did not conclude the survey or who did not report adequate simplifications.

After a small pilot study and filtering of inadequate responses, we collected evaluations for 96 sentence pairs. 78 workers participated in the study, so each sentence pair was evaluated, on average, by 3.93 participants.

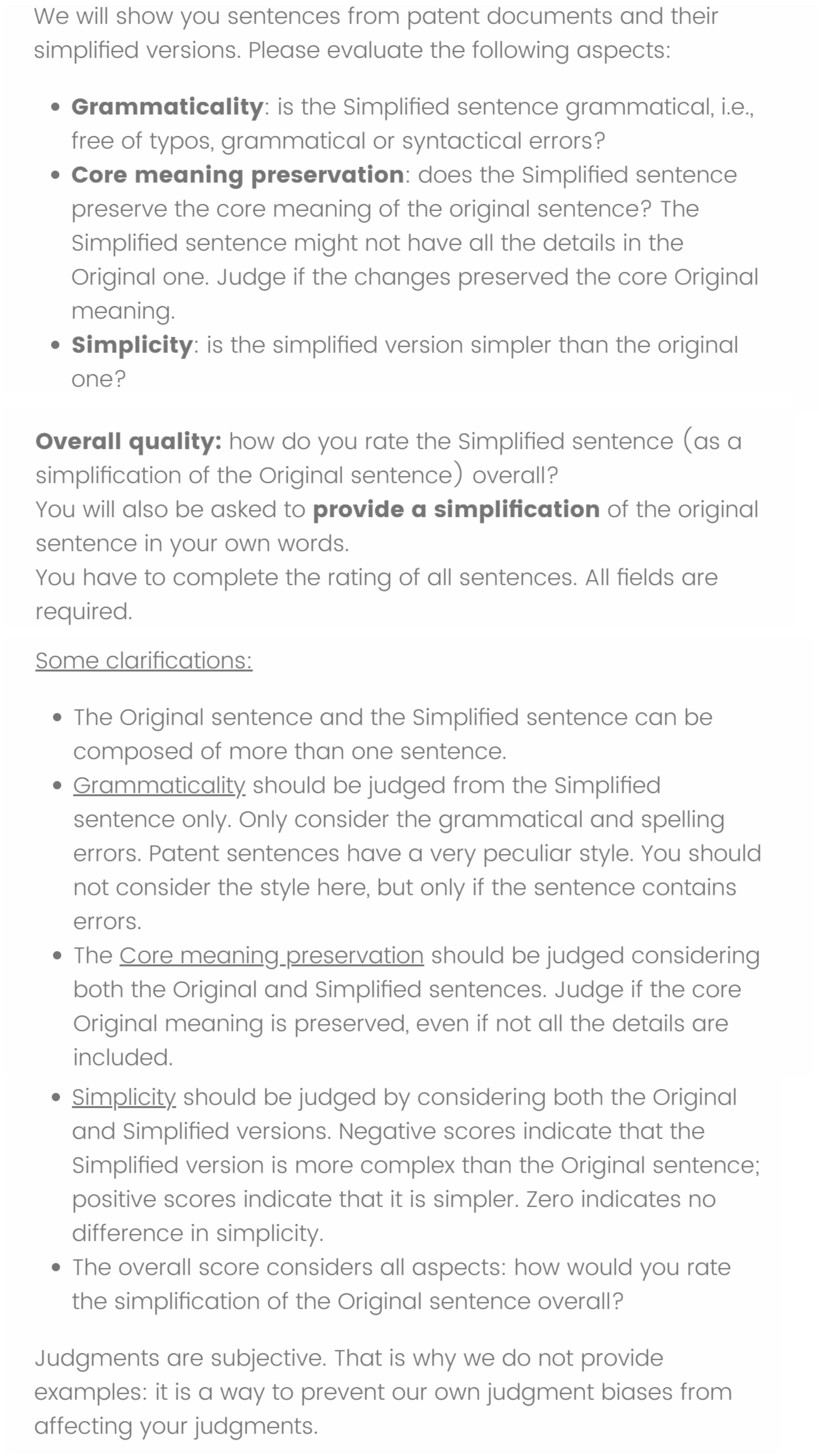
We chose to evaluate the silver standard with laypeople as they are the target of our simplification. In future work, we plan to further evaluate the dataset with experts of the patent domain and the target technical domain, which might be more reliable, particularly when judging the meaning preservation.

### 8.5.2 Numerical scores

Table 8.7 reports the results of the scores as assigned by the Prolific workers. Figure 8.6 reports the answer distribution for each dimension.

[7] `https://www.prolific.co/`

[8] `https://www.qualtrics.com/`



We will show you sentences from patent documents and their simplified versions. Please evaluate the following aspects:

- **Grammaticality**: is the Simplified sentence grammatical, i.e., free of typos, grammatical or syntactical errors?
- **Core meaning preservation**: does the Simplified sentence preserve the core meaning of the original sentence? The Simplified sentence might not have all the details in the Original one. Judge if the changes preserved the core Original meaning.
- **Simplicity**: is the simplified version simpler than the original one?

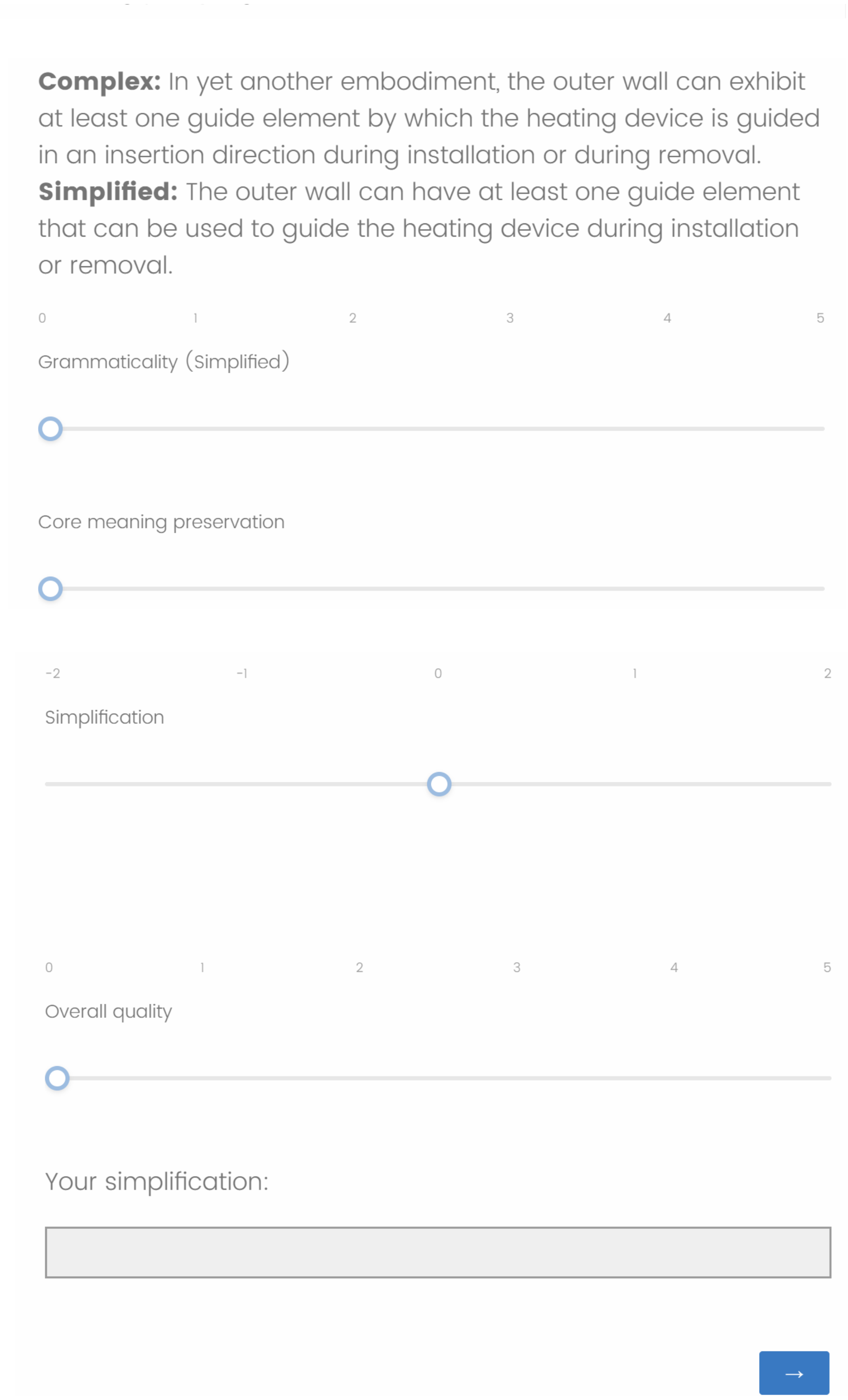
**Overall quality:** how do you rate the Simplified sentence (as a simplification of the Original sentence) overall?
You will also be asked to **provide a simplification** of the original sentence in your own words.
You have to complete the rating of all sentences. All fields are required.

Some clarifications:

- The Original sentence and the Simplified sentence can be composed of more than one sentence.
- Grammaticality should be judged from the Simplified sentence only. Only consider the grammatical and spelling errors. Patent sentences have a very peculiar style. You should not consider the style here, but only if the sentence contains errors.
- The Core meaning preservation should be judged considering both the Original and Simplified sentences. Judge if the core Original meaning is preserved, even if not all the details are included.
- Simplicity should be judged by considering both the Original and Simplified versions. Negative scores indicate that the Simplified version is more complex than the Original sentence; positive scores indicate that it is simpler. Zero indicates no difference in simplicity.
- The overall score considers all aspects: how would you rate the simplification of the Original sentence overall?

Judgments are subjective. That is why we do not provide examples: it is a way to prevent our own judgment biases from affecting your judgments.

FIGURE 8.4: Istructions for the human evaluation campaign.



**Complex:** In yet another embodiment, the outer wall can exhibit at least one guide element by which the heating device is guided in an insertion direction during installation or during removal. **Simplified:** The outer wall can have at least one guide element that can be used to guide the heating device during installation or removal.

Grammaticality (Simplified)

Core meaning preservation

Simplification

Overall quality

Your simplification:

FIGURE 8.5: Human evaluation interface. For each sentence, we also included instructions at the top of the page.

| Dimension | Mean ± std |
|---|---|
| Grammaticality | 3.59 ± 1.45 |
| Core meaning preservation | 3.27 ± 1.46 |
| Simplicity | 0.89 ± 1.15 |
| Overall | 3.07 ± 1.44 |

TABLE 8.7: Human evaluation results

Results show that sentences are considered rather grammatical. Grammatical sentences were described in the instructions as "free of typos, grammatical or syntactical errors" and workers were asked not to consider style when evaluating this dimension. However, by manually validating the answers, we noticed that complex sentences or sentences containing infrequent grammatical constructs (which are, however, frequent in patent documents) might have a low grammatical score despite no obvious errors.

Most Simplified sentences were considered to retain the original core meaning. Regarding simplicity, the vast majority of the sentences were considered somewhat or much simpler than the original.

As a first preliminary experiment, we investigated how each dimension correlated with the overall quality score. A simple linear model ($R^2 = 0.66$) scores the dimensions as:

$$\text{overall} = 0.51 + 0.20 \times \text{G} + 0.45 \times \text{M} + 0.40 \times \text{S}$$

where G is the grammaticality score, M is the core meaning score and S is the simplicity score.

### 8.5.3 Human-written simplifications

During the survey, we also asked participants to provide a simplification, in their own words, of the original sentence. Table 8.9 shows some examples generated by the workers.

Table 8.9 reports some metrics computed on the collected data.

Unsurprisingly, the human-written sentences are more similar to the simplified sentence than to the original ones; note, however, that their similarity is not as high as to consider them derivative. Moreover, they are shorter than synthetic simplified sentences.

Considering the simplification scores, they are in line with that generated by the model. The only exception is that of the dependency tree, which is sensibly less deep in the case of sentences generated by humans.

Human-written simplifications allow evaluating models trained on our silver standard with data that are not synthetic. However, users should be aware of their limitations, as they were produced by workers lacking expertise in the patent or technical domain.

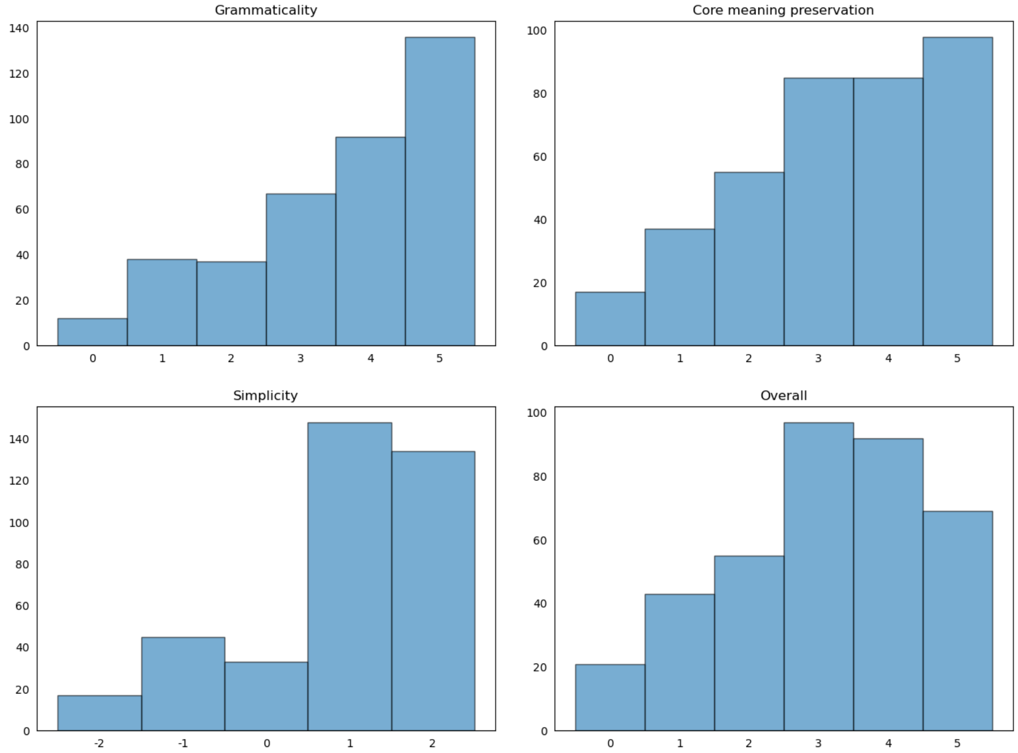


FIGURE 8.6: Distribution of votes for each dimension.

| Original | Simplified (automatic) | Simplified (human) |
|---|---|---|
| In yet another embodiment, the outer wall can exhibit at least one guide element by which the heating device is guided in an insertion direction during installation or during removal. | The outer wall can have at least one guide element that can be used to guide the heating device during installation or removal. | The outer wall will have at least 1 guide to help guide the heating device during installation or removal.<br>During installation or removal the outer wall has 1 or more guide element. |
| A lamp holder or fitting carrier for at least one lamp is arranged in a position in which it lies outwardly opposite a recess in the reflector and is releasably connectable with the base frame by second mounting means. | A lamp holder or fitting carrier for at least one lamp is arranged in a way that it is releasably connectable with the base frame by second mounting means. | A lamp holder or fitting carrier for at least one lamp is arranged in a way that it is connectable with the base frame by second mounting means.<br>A lamp holder or fitting carrier for at least one lamp is arranged in a position releasably connectable with the base frame by second mounting means.<br>A lamp holder is arranged in a way that lets the base frame be accessible for a second mount.<br>this product is easily changed from a one lamp to a 2 lamp item via the changeable base |
| In an emergency, the safety valve can also be opened by hand, when the pressure in the feed line is too high for unknown reasons. | When the feed line is too high for unknown reasons, the safety valve can be opened by hand. | In an emergency, the safety valve can be opened by hand, when the pressure is too high<br>When the feed line pressure is too high, the safety valve can be opened by hand.<br>In an emergency, the feed line is too high for unknown reasons, the safety valve can be opened by hand.<br>when feed line is too high open safety valve<br>if the line feed is too high, hand open the valve |

TABLE 8.8: Simplfications generated by the Prolific workers

| Metric | Complex | Simplified |
|---|---|---|
| Levenshtein similarity | 58.14 ± 20.35 | 65.18 ± 21.17 |
| BLEU [140] | 23.17 ± 19.44 | 35.10 ± 26.47 |
| BERTScore (avg) [213] | 92.10 ± 4.05 | 93.51 ± 4.14 |
| Flesch Reading Ease [52] | 46.12 ± 30.12 | |
| Flesch–Kincaid Grade Level [87] | 55.77 ± 13.24 | |
| WordRank [120] | 9.52 ± 1.54 | |
| Max dependency tree depth | 4.79 ± 2.08 | |
| Length (chars) | 87.53 ± 42.83 | |

TABLE 8.9: Statistics on simplified sentences produced by humans.

## 8.6 Conclusions

In this chapter, we have discussed a method to generate a parallel silver standard for simplifying patent sentences through rephrasing. To the best of our knowledge, this is the first parallel simplification corpus for patents. We have analyzed the corpus quantitatively and qualitatively, showing that it can be used to train a sequence-to-sequence simplification model. We have also conducted a large-scale human evaluation of the corpus and collected human-written simplifications for evaluation.

While not exempt from the shortcoming given by its automatic origin, we have shown that filtering out faulty candidates allows us to obtain a corpus that has been considered grammatical, adequate, and with a significant simplification.

In future work, we plan to further explore how the corpus can be used in relation to other simplification datasets and whether transfer-learning techniques can further improve patent simplification, making technical information more accessible. Moreover, we plan to investigate whether the method described here can be successfully adopted for other types of technical text.

# Chapter 9

# Conclusions

We started this thesis by describing the opportunities and challenges of converting a massive amount of technical text into information in the patent domain. We have mainly focused on how the length and complexity of patents make it challenging to understand and process their content. We have hypothesized Natural Language Processing could be a valuable tool for the task and have thus decided to explore whether automatic text summarization and simplification could effectively solve some of these issues.

In the following chapters, we have first surveyed how Natural Language Processing and Natural Language Generation methods are used in the patent domain. Then, we have experimented with applications of summarization and simplification.

In this chapter, we reflect on our contributions, the limitations of our work, the ethical implications, and future work we hope this thesis will inspire.

## 9.1 Contributions

In **Chapter 1**, we have described the background and motivation for this thesis. The work was initially conceived in the larger frame of Technology foresight. Technology foresight is an umbrella term that covers a variety of diverse techniques and approaches that typically require substantial human effort. Experts are central in the process and are, in a way, regarded as "human containers and processors" of a considerable amount of information in the technological landscape that they consume due to their work.
Given our background in Information Systems and Natural Language Processing, we have asked ourselves whether automatizing some of the information processing could be beneficial. We have particularly focused on the patent domain, where the amount of available text is astonishing. Patent data is understudied and underutilized outside of the patent process itself, partly due to the documents' complexity in terms of length, amount of details, and linguistic complexity.
We have thus decided to explore the application of summarization and simplification techniques in the patent domain.

In **Chapter 2**, we have surveyed previous related work from Natural Language Processing. We have specifically focused on patent document summarization, simplification, and generation tasks. We have found that most approaches we considered state-of-the-art in the general domain are yet to be explored for the patent domain. This is partly because patents are very peculiar documents, with several challenging aspects at the lexical and syntactic levels, which try to legally protect any possible variation of the described

invention and yet include as little (economically valuable) information as possible. Thus, we have found that most resources for general discourse Natural Language Processing tend to underperform in the patent domain.
We have then explored previous work for each of the target tasks individually. For summarization, we have noted that most of the classical approaches are extractive and use either general-domain techniques directly or adapt these techniques to account for the linguistic challenges and the technical domain. However, given the lack of shared benchmarks, we could not directly compare approaches (and thus sum up which techniques tend to work better). We have also found a relatively recent dataset for the abstractive summarization of patent documents, whose success encouraged researchers in the general summarization field to test large-scale general-purpose summarization models in the patent domain.
For simplification, we have found that most approaches aim to provide ways to better visualize the patent text (with a focus on its claims) with little or no textual simplification. The few publications that attempt to modify the text are based on rules and linguistic considerations. In most cases, the goal is to provide patent practitioners with tools to help them in their work.
Regarding generation, we have found an interesting line of work that tries to produce patent text automatically. The work is still in the research stages and tries to explore and adapt large-scale general-purpose systems to the domain.

In **Chapter 3**, we have provided an overview of the summarization task, describing methods, resources, and evaluation procedures. The chapter is intended as a short survey of recent summarization approaches in Natural Language Processing, and we have included it to better frame our work on patent summarization.

Since previous work often used the BigPatent dataset as a testbed in the patent domain (and since it was the only available curated dataset for patent summarization at the beginning of our work), we have decided to use it for our experiments. Our initial results, however, were hard to interpret. We had trained the exact same model with two versions that were described as identical except for their casing and tokenization; however, we had noticed a very high difference in metrics. In **Chapter 4**, we have decided to understand the issue better. We have found that the two versions are, in fact, *very* different, with the updated one containing a superset of the original input. We have also noted that the updated version contains the text of the "Summary of the Invention" *in the input*. We have decided to document the differences between the two versions' content and performance clearly and in detail. Moreover, we have found that due to the lack of any documentation prior to our work, published research uses the two versions of the dataset interchangeably, and it is thus tough – if at all possible – to understand if comparisons are fair or if variations are given by the differences in the dataset versions. We have also described the modifications to the dataset to use it in our experiments fairly.

In our preliminary explorations, we noticed a lack of shared benchmarks; thus, it is difficult to systematize what "works" or "does not work" in the patent domain. In **Chapter 5**, we have filled this gap by benchmarking existing extractive, abstraction, and hybrid summarization approaches in the patent domain. Despite our limited resources, we have found some interesting results: graph-based systems, for example, seem appropriate for

content selection and perform relatively well in metrics and outputs. However, the extracted outputs are subject to all the limitations of extractive systems, with dangling references being particularly common. The length of the sentences, the dangling references, and the lack of discourse structure make the outputs challenging to process for humans and possibly for machines as well. We have found other unsupervised extractive approaches we benchmarked to be generally less successful than the graph-based ones.
We have not considered supervised extractive systems for two reasons. First, we needed to transform our abstractive reference into an extractive one; while this is not technically complex, it requires resources and time, particularly with very long input documents. The second reason is again related to document length: many state-of-the-art supervised extractive systems tend to be able only to summarize documents with a limited input length.
Among the abstractive approaches, we have analyzed BART and have found that it performs best in automatic metrics compared to extractive algorithms. We have also found that the produced outputs are, in fact, not very extractive with respect to the input, with long chunks of texts identical to input passages; the model seems, however, very good in removing non-central content from the single sentences, which extractive systems are natively unable to do. We have also considered some simple select-and-rewrite approaches, which obtained the best automatic metrics.

From our previous experiments, it seemed that the lack of available off-the-shelf models and their suboptimal performances were vastly due to the considerable document length; in fact, the best-performing models on the datasets in the literature are those that use efficient attention mechanisms and allow models to process the entire document at once. In **Chapter 6**, we have deepened into the long document summarization problem and have tried to adapt DANCER to the patent domain. DANCER is a method created for scientific papers that automatically generates training data to summarize each section independently and then trains a sequence-to-sequence system for the task. However, we have found that patents are more variable in the sections they contain and in the sections' content itself, and their Abstracts tend to be less compositional than those of papers. Thus, the approach was not particularly successful when transferring to the patent domain.

After presenting our work on summarization, we have started discussing simplification in **Chapter 7**. The chapter overviews the task, including methods and resources available. We found that many recent approaches are sequence-to-sequence monolingual translators and thus require parallel data. However, no simple text nor parallel corpus was available for patents.

Thus, in **Chapter 8**, we have filled this gap and have explored a method to automatically create a silver corpus for sentence simplification in the patent domain. We have proposed to use a paraphrasing system only trained on general-domain text to create potential complex-simple pairs candidates. The system performs compression and lexical simplification, removing the long roundabout expressions common in patent text. However, only some model outputs were satisfactory, while others contained unknown tokens, were practically equal to the original ones, or were too compressed. Thus, we have filtered out these pairs and have obtained a silver standard for simplification that we validated through human evaluation; we have collected human "gold" simplifications in the process. We have shown that the corpus is considered grammatical, adequate, and contains simple sentences and that it can be used to train a state-of-the-art simplification system.

## 9.2 Challenges and limitations

This thesis takes the first steps toward applying summarization and simplification techniques in the patent domain. In the following, we will describe some of the challenges we faced and the limitations of our work.

### 9.2.1 Interdisciplinary collaboration

Since, when starting the thesis, our background in the patent domain was limited, we have closely studied the Intellectual Property landscape, how the patenting process works, and its core rules and laws. However, we are not domain experts, and having the relations and resources to collaborate with lawyers, patent experts, and other stakeholders would have been of help; for example, it would have allowed us to explore more applications and would have probably anticipated some of the issues we faced during the project.

Unfortunately, this work started at a time when collaboration (especially in person) was not easy, and our resources were relatively limited. However, we filled these gaps with published work and other available resources.

In the future, based on the experiments and results of this thesis, we hope to establish closer collaborations with domain experts.

### 9.2.2 Patent-specific NLP resources

We have reported numerous times that, from a linguistic perspective, patent documents are very peculiar; with a bit of an overstatement, we could consider *"patentize"* a similar yet different language than the portion of English we use in our everyday life. These considerations call for the use of patent-specific resources. Since most of the tools used in Natural Language Processing result from training on general-domain English, they often do not transfer well to the patent domain. Thus, even the tasks that are practically considered "solved" for general text can become problematic for patents.

### 9.2.3 Limitations of the summarization approaches

For summarization, our work had two main themes: the first one had to do with "making order" in the summarization of patent documents from a Natural Language Processing perspective. We have found that, due to the lack of benchmarks, most previous "older" work is either impossible to replicate or does not allow us to draw general considerations; we have also found out that the trend might continue in the future as the existence of two undocumented versions of BigPatent does not make it possible to compare approaches directly. The second line of research had to do with how well Natural Language Processing approaches transfer to the patent domain and with which issues make the problem of patent simplification hard to solve.

We have treated patents as an instance of an interested (yet very complex) domain and, when possible, enriched our approaches with domain knowledge.

We have always presented a small subset of model outputs so that the reader can make their own consideration of the merits and limitations of the approaches and evaluate them quantitatively and qualitatively. However, we could not involve experts in the evaluation process.

### 9.2.4 Limitations of the simplification approaches

The simplification approach has aimed at making the content of patent documents more accessible. When we started our work, previous experiments on textual simplification in the patent domain were very limited, and practically no resources were available.

We have thus decided to focus on sentence simplification, specifically on the automatic creation of a silver corpus for sequence-to-sequence models. We believe that, for performing "real" simplifications, document-level approaches are required. However, following the vast majority of previous work in automatic simplification, we have decided to first explore sentence simplification, for which approaches are more mature.

For the evaluation of the corpus, we have conducted a large-scale human evaluation campaign. While our system is targeted at laypeople rather than experts, we believe that getting feedback from experts could have been helpful, especially for assessing meaning preservation.

## 9.3 Ethical considerations

### 9.3.1 Data and artifacts

For summarization, we have used a public dataset built on top of public documents, and we are not aware of any ethical concerns associated with the data. We have built the simplification parallel corpus out of public data. We mention here that, while we inspected a large subset of the generated corpora for our experiments and analysis, we did not manually review the whole generated content. It is thus possible in principle – even though highly unlikely, in our opinion – that some sentences contain toxic or offensive, or other inappropriate content. Moreover, since the dataset is generated automatically, it might contain some errors and non-factual simplifications. Patents are legal documents, and any automatic text processing output should thus be inspected by experts and corrected depending on context.

### 9.3.2 Models

We have explored and trained a number of models, including various large-scale neural networks. We are aware of the environmental issues associated with their carbon footprint, and we have put our best effort into minimizing computing when possible. Content-wise, we again want to underline that our models rely on probability and patterns and can thus be subject to errors. They should thus be used as human aid rather than as an automatic means to make decisions.

### 9.3.3 Human evaluation

We have conducted a human evaluation campaign for the evaluation of our simplification corpus. We have resorted to Prolific, an online resource that hires workers to complete small tasks. Prolific has high ethical standards and requires workers to be paid a minimum of £6.00 per hour. We conducted a first pilot study to better estimate the time needed for each task we proposed. We also manually revised all the possible instances (selected at random from our corpus) to make sure no inappropriate content was present. We explicitly required consent, including to publish the annotations and use them for other research projects.

## 9.4 Future works

Despite the challenges that we faced, we believe patents are extremely underused as a source of information.

In our future work, we first plan to improve on the limitations that we described above and build a closer relationship with domain experts. For example, we plan to perform expert evaluations of the outputs.

Validating the use of automatic metrics and understanding their correlation with expert evaluation was not a topic of this thesis. However, we believe the topic is interesting and important both for generative tasks in the general domain and for the patent domain specifically. We believe that the validity of the automatic metrics (already widely discussed in general) might be even more unclear in some domains, including the patent domain. In the future, we plan to perform a meta-evaluation of summarization (and possibly simplification) metrics to assess whether they are affected by the text domains and to which extent.

In this thesis, we mostly evaluated model outputs from a human user perspective. Previous work showed that both summarization and simplification can be used as preprocessing tasks to improve models' performance on downstream tasks. We believe this aspect could be even more relevant in the patent domain, considering the text complexity. We plan to explore this aspect in future work.

Overall, we hope this thesis will inspire more work at the intersection of Natural Language Processing and patent mining, aid experts and practitioners in their work, and make the information contained in the patent documents more accessible by the whole society.

# Appendix A

# Appendix: The WITS dataset for abstractive text summarization in Italian

Performance on abstractive text summarization has recently improved due to the use of sequence-to-sequence models. However, while these models are extremely data-hungry, datasets in languages other than English are few.

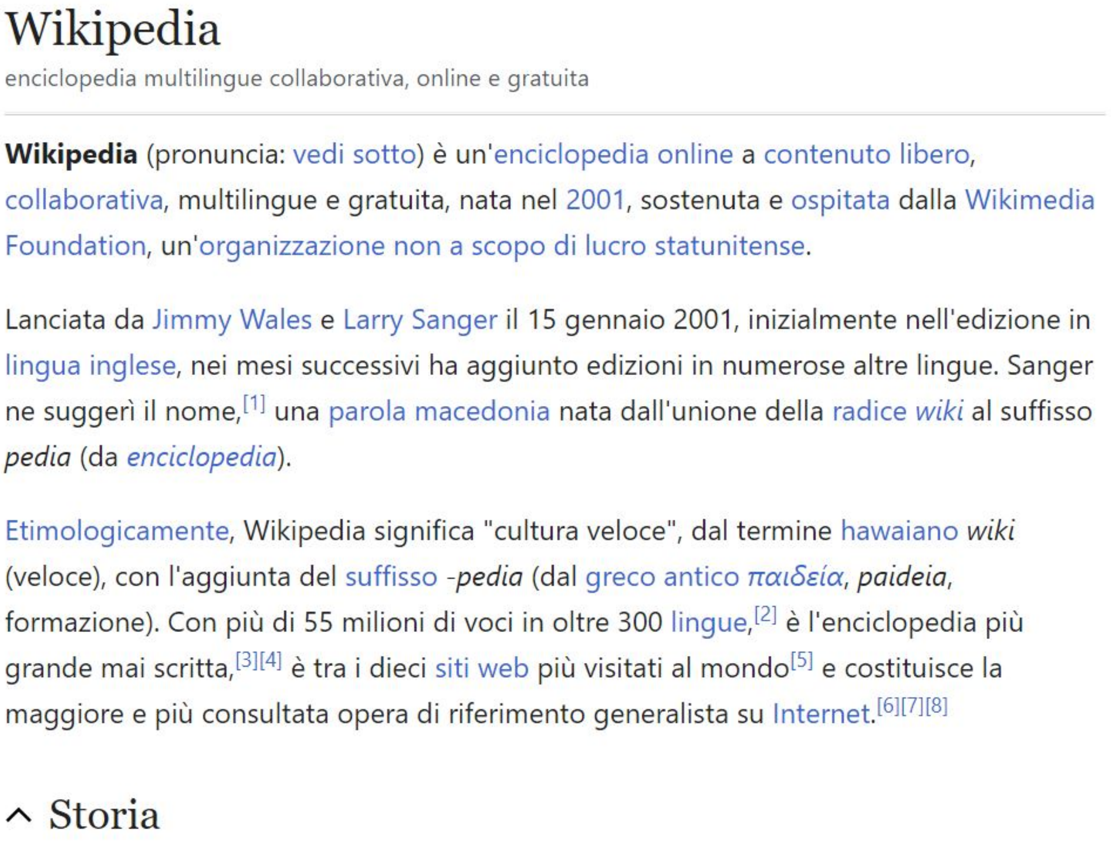
Here, we describe WITS (Wikipedia for Italian Text Summarization), a large-scale dataset we built by exploiting Wikipedia articles' structure.
WITS contains almost 700,000 Wikipedia articles, together with their human-written summaries. Compared to existing data for text summarization in Italian, WITS is more than an order of magnitude larger and more challenging, given its lengthy sources. We explore WITS characteristics and present some baselines for future work.

This chapter is based on Casola and Lavelli [30].

## A.1 WITS: Wikipedia for Italian Text Summarization

Recently, abstractive summarization has attracted a growing interest in the Natural Language Processing (NLP) community. Sequence-to-sequence models have been increasingly used for the task, with pre-trained encoder-decoder transformers becoming the de facto state of the art for abstractive text summarization. Normally pre-trained in an unsupervised manner, these models are then fine-tuned in a supervised way on the downstream dataset; during fine-tuning, the model learns to generate the summary from the source document.

While various datasets for abstractive summarization exist for English, resources in other languages are limited. Here, we describe WITS (Wikipedia for Italian Text Summarization), a large-scale dataset for abstractive summarization in Italian that we built by exploiting Wikipedia. Taking advantage of the structure of Wikipedia pages, which contain a lead section (Figure A.1) – giving an overview of the article's topic –, followed by the full-length article – describing the topic in details –, we create a large and challenging dataset for abstractive summarization in Italian, which we made publicly available.



Wikipedia

enciclopedia multilingue collaborativa, online e gratuita

**Wikipedia** (pronuncia: vedi sotto) è un'enciclopedia online a contenuto libero, collaborativa, multilingue e gratuita, nata nel 2001, sostenuta e ospitata dalla Wikimedia Foundation, un'organizzazione non a scopo di lucro statunitense.

Lanciata da Jimmy Wales e Larry Sanger il 15 gennaio 2001, inizialmente nell'edizione in lingua inglese, nei mesi successivi ha aggiunto edizioni in numerose altre lingue. Sanger ne suggerì il nome,[1] una parola macedonia nata dall'unione della radice *wiki* al suffisso *pedia* (da *enciclopedia*).

Etimologicamente, Wikipedia significa "cultura veloce", dal termine hawaiano *wiki* (veloce), con l'aggiunta del suffisso -*pedia* (dal greco antico *παιδεία*, *paideia*, formazione). Con più di 55 milioni di voci in oltre 300 lingue,[2] è l'enciclopedia più grande mai scritta,[3][4] è tra i dieci siti web più visitati al mondo[5] e costituisce la maggiore e più consultata opera di riferimento generalista su Internet.[6][7][8]

^ Storia

FIGURE A.1: The lead section (from Wikipedia's own page), which we consider as the article summary. We use the remaining article as the source.

WITS is particularly challenging, given its large source length and its highly abstractive summaries. Here, we describe the dataset, its statistics, and its characteristics and report some preliminary experiments that might be used as baselines for future work.

This chapter is organized as follows: in Section A.2, we describe the state of the art in text summarization, focusing on resources for Italian. We later present the dataset and its related task (Section A.3.1); we describe the data collection and preprocessing step in Sections A.3.2 and A.3.3. In Section A.4, we show baseline performance. Finally, we draw our conclusions in Section A.5.

## A.2 Resources in Italian for text summarization

Automatic text summarization has recently attracted increasing attention from the Natural Language Processing community. However, the majority of the research work still focuses on English, and resources in other languages are few.

As a matter of example, out of all the papers published in the Association for Computational Linguistics (ACL) conference in 2021, 46 explicitly refer to summarization in their title; 38 of these dealt with English only, while 7 presented experiments with one or more other languages (including 2 on source code summarization). For reference, only one paper [122] on text summarization (in English) was published at the Italian Conference on Computational Linguistics (CLiC-it) since its first edition, and none experimented with Italian.

In this section, we present the state of the art in abstractive text summarization. We first present the available datasets; then, we discuss some relevant models. We focus on the significant gap between English and Italian, for which very few resources exist.

### A.2.1 Datasets for automatic text summarization in Italian

As we described in Chapter 3, a typical dataset for text summarization contains source documents (which need to be summarized) and their corresponding summaries, used as the gold standard. A minority of datasets (e.g., the DUC 2004 dataset[1]) provide multiple gold standards; however, such datasets tend to be small and are mostly used for evaluation.

To the best of our knowledge, WikiLingua [95][2] was the only summarization dataset containing data in Italian published before WITS. WikiLingua is a cross-lingual dataset for abstractive text summarization built on top of WikiHow. WikiHow contains tutorials on how to perform specific tasks in the form of step-by-step instructions. The dataset constructs a summary by concatenating the first sentence for each step and using the remaining text as the source. WikiLingua contains data in 18 languages, including Italian (50,943 source-summary pairs). Both summaries and sources are relatively short (on average, 44 and 418 tokens, respectively, for the Italian split).
Since the publication of WITS, two news datasets for Italian have been published [96]: Il Post and Fanpage, which are built out of news articles exploiting the short "summary" presenting the news after the title.

[1] https://duc.nist.gov/duc2004/

[2] https://huggingface.co/datasets/wiki_lingua

### A.2.2 Models for abstractive text summarization in Italian

Models for abstractive text summarization are generally sequence-to-sequence: they encode the input and then generate the output through a neural network.

Summarization models either exploit encoders and decoders previously trained for other tasks or are pre-trained from scratch on a specific objective tailored for summarization.

Following a shared practice, most summarization models have first been trained and evaluated for English only. In some cases, a subsequent multilingual version of the model was also created [206, 154, 110].
To the best of our knowledge, few sequence-to-sequence models in Italian exist to date. IT5-base [3] was the only available model with an encoder-decoder architecture when this work was published, and no full-scale evaluation was performed yet[4].
In 2023 (after the publication of WITS), a BART-based model was also published by La Quatra and Cagliero [94].
GPT-2 (a decoder-only model) has also been adapted for Italian [123]. The model might be explored for summarization, e.g., by concatenating the source and the summary (separated by a special token) at training time and letting the model autoregressively predict the tokens in the summary at inference.

## A.3 WITS

### A.3.1 Task and rationale

Given a Wikipedia article, we extract the lead section (which we sometimes refer to as "Summary" in the remaining of the paper) and propose the following task:

> Given all article sections, summarize its content to produce its lead section.

The task is rather natural, given the page structure. According to the Wikipedia Manual of Style[5], the lead section is, in fact, a high-quality summary of the body of the article. The lead serves as an introduction to the article and a summary of its most important contents" and "gives the basics in a nutshell and cultivates interest in reading on—though not by teasing the reader or hinting at what follows". Moreover, it should "stand on its own as a concise overview of the article's topic".

As for the content, according to Wikipedia, the lead must define the topic, explaining its importance and the relevant context; then, it must summarize the most prominent points of the article, emphasizing the most important material.

Moreover, the lead should only cover information that is contained in the article: "significant information should not appear in the lead if it is not covered in the remainder of the article". This is particularly relevant for abstractive summarization, as models are more prone to produce summaries that are not factual to the source (often called hallucinations) when they are trained to generate summaries containing information not in the

[3] `https://huggingface.co/gsarti/it5-base`

[4] The paper describing the model is now available [154]

[5] `https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style`

| | **WITS** | | **IT-Wikilingua** | |
|---|---|---|---|---|
| # docs | 699,426 | | 50,943 | |
| | Summary | Source | Summary | Source |
| # sentences (avg) | 3.75 | 33.33 | 5.01 | 23.52 |
| # tokens (avg) | 70.93 | 956.66 | 23.52 | 418.6 |
| Comp. ratio (avg) | 16.14 | | 11.67 | |

TABLE A.1: Datasets statistics. spacy is used for text and sentence tokenization. The number of tokens and sentences is computed for all documents and then averaged.

source [131]. The problem of factuality in abstractive summarization is currently an active area of research, as previous work has shown that up to 30% of generated summaries contain non-factual information [26].

Linguistically, the lead "should be written in a clear, accessible style with a neutral point of view". It is worth noting that, in contrast to WikiLingua, where the summary is constructed as a concatenation of sentences from different parts of the articles, the summary in WITS is a stand-alone piece of text, with a coherent discourse structure.

### A.3.2 Data collection

This section describes the process of data collection and preprocessing.

We downloaded the latest available XML dump of Wikipedia in Italian[6], which contains text only. We used Python and the Gensim library to process the file[7]. The original number of documents was 1,454,884. We applied the following exclusion criteria: we removed pages whose title contains numbers only (as they mostly describe years and contain lists of events and references), lists (titles starting with "Lista d"), pages with summaries with less than 80 characters and articles and pages for which the article is less than 1.5 times longer than the lead.

We then preprocessed the text in the following way: from the summary, we removed the content of parentheses (as they often contain alternative names or names in a different language, which cannot be inferred from the article). For the article, we further excluded the following sections, which are not relevant for our task: Note (Footnotes), Bibliografia (References), Voci correlate (See also), Altri progetti (Other projects), Collegamenti esterni (External links), Galleria di Immagini (Images).

### A.3.3 Dataset statistics

Table A.1 shows some statistics on the dataset and compares WITS with the Italian split of WikiLingua (which we will refer to as IT-WikiLingua).

IT-WikiLingua contains documents from 17,673 WikiHow pages, but some of these pages describe more than one method related to the same topic. For example, the page "How to Reduce the Redness of Sunburn" contains several methods: "Healing and Concealing Sunburns", "Lessening Your Pain and Discomfort", and "Preventing a Sunburn". We

[6] `https://dumps.wikimedia.org/itwiki/latest/itwiki-latest-pages-articles.xml.bz2`

[7] `https://radimrehurek.com/gensim/scripts/segment_wiki.html`

| | WITS | | IT-Wikilingua | |
|---|---|---|---|---|
| | Summary | Source | Summary | Source |
| PER (avg) | 1.13 | 26.21 | 0.32 | 1.05 |
| LOC (avg) | 2.03 | 24.07 | 0.42 | 1.39 |
| ORG (avg) | 0.60 | 6.65 | 0.68 | 0.37 |
| MISC (avg) | 19.68 | 19.68 | 0.84 | 3.07 |
| All (avg) | 23.44 | 76.61 | 1.65 | 5.88 |

TABLE A.2: Named Entities in WITS and IT-WikiLingua.

consider distinct methods as separate documents, as they can be summarized in isolation. Notice that WITS is more than an order of magnitude larger than IT-Wikilingua.

We computed the number of tokens and the number of sentences through the spaCy it-core-news-lg[8] model. Compared to IT-WikiLingua, documents in WITS contain more tokens both in their summary and in their source (which is more than double in length), making the dataset particularly challenging. Note how the sentences are also more lengthy (thus complex) on average. For example, summaries in WITS contain on average less than 4 sentences, but more than 70 words; in contrast, IT-WikiLingua's summaries consist of more than 5 sentences but contain on average 44 tokens. Not surprisingly, WITS' compression ratio is larger than IT-WikiLingua's and very high in absolute value. Finally, we also notice that the dataset is very rich in named entities. Table A.2 reports the Named Entities as extracted with spaCy from WITS and IT-Wikilingua.

## A.4 Models performance

We tested some preliminary baseline methods on the dataset, reported in Table A.3. We evaluated the summaries using ROUGE [105].

We considered the following baselines:

**Lead-3** We extract the first three sentences from the source. Previous work has shown that this baseline is often hard to beat [157], especially in news summarization, which presents an "inverted pyramid" structure and tends to report the most important content at the start.

**TextRank [124]** TextRank is an unsupervised algorithm that extracts the most relevant sentences in the source. The algorithm constructs a graph with sentences as nodes and sentence similarity (in terms of shared vocabulary) as edges. The sentences are then ranked by using the PageRank [139] algorithm.

**LexRank [46]** LexRank works in a similar way as TextRank. However, instead of computing sentence similarity on normalized shared vocabulary, it uses the cosine similarity of their TF-IDF vectors.

**SumBasic [134]** SumBasic extracts sentences based on their word probabilities. Specifically, it scores each sentence as the mean of the probability of the words it contains (based on their frequency in the document). Iteratively, the sentence with the best

[8] `https://spacy.io/models/it`

score among the ones containing the most probable word is chosen. The probability of the words in the chosen sentence is then squared to limit redundancy.

We also performed some experiments with IT5 Small in our work. Here, we prefer to show results reported in [154] (which extensively experimented with WITS) as they also perform experiments with larger versions of the model.

We also reports some results obtained from BART-IT [94] for completeness.

| | R-1 | R-2 | R-L |
|---|---|---|---|
| Lead-3 | 24.76 | 5.54 | 16.54 |
| TextRank | 30.20 | 6.57 | 19.67 |
| LexRank | 26.90 | 5.91 | 17.52 |
| SumBasic | 20.60 | 4.80 | 14.01 |
| mT5 Small [154] | 34.7 | 20.0 | 31.6 |
| mT5 [154] | 34.8 | 20.0 | 31.5 |
| IT5 Small [154] | 33.7 | 19.1 | 30.6 |
| IT5 Base [154] | 36.9 | 21.7 | 33.3 |
| IT5 Large [154] | 33.5 | 19.1 | 30.1 |
| mBART [110] | 39.32 | 26.18 | 35.9 |
| BART-IT [94] | 42.32 | 28.83 | 38.84 |

TABLE A.3: ROUGE results on WITS. Results from T5 and BART-inspired models are from the related papers.

Results show that the Lead-3 baseline performance is low; this is likely due to the structure of Wikipedia, which contains several thematic sections without a general introduction outside the lead section. Extracting the first sentence(s) from each section would likely produce better results and could be investigated in future work.

In contrast, TextRank is the best non-neural baseline, with a ROUGE-2 score of 6.57; LexRank performs comparably. SumBasic metrics are even lower than those obtained with the Lead-3 baseline, suggesting that a purely frequency-based approach is insufficient given the dataset complexity.

Not surprisingly, neural models achieve the best results on the dataset. Among T5-based models, IT5 Base achieves the best performance. BART-IT is, to date, the best-performing model on the dataset.

## A.5 Conclusions

We have presented WITS, the first large-scale dataset for abstractive summarization in Italian. We have exploited Wikipedia's articles' structure to build a challenging, non-technical dataset with high-quality human-written abstracts. Given the lengthy source documents, the short summaries, and the short extractive fragments, the dataset calls for an abstractive approach. We have explored some standard non-neural extractive baselines and a neural abstractive baseline. Subsequent work has investigated further neural baselines for the dataset. The dataset can be easily extended by applying the procedure described in the paper to more languages, including low-resource ones, given the Wikipedia structure. We are confident that research in summarization in languages other than English will become more active in the near future and hope that WITS can be a valuable step in this direction.

Appendix B

# Appendix: US4575330 full text

# United States Patent [19]

**Hull**

[11] **Patent Number: 4,575,330**

[45] **Date of Patent: Mar. 11, 1986**

[54] **APPARATUS FOR PRODUCTION OF THREE-DIMENSIONAL OBJECTS BY STEREOLITHOGRAPHY**

[75] Inventor: **Charles W. Hull**, Arcadia, Calif.

[73] Assignee: **UVP, Inc.**, San Gabriel, Calif.

[21] Appl. No.: **638,905**

[22] Filed: **Aug. 8, 1984**

[51] **Int. Cl.$^4$** ...................... **B29D 11/00**; G03C 00/00

[52] **U.S. Cl.** ................................ **425/174.4**; 425/174; 425/162; 264/22; 430/269; 156/58; 365/119; 365/120

[58] **Field of Search** ...................... 425/162, 174, 174.4, 425/425; 264/22, 183, 40.1; 430/269; 156/38, 58, 275.5; 365/107, 119, 127

[56] **References Cited**

U.S. PATENT DOCUMENTS

| | | | |
|---|---|---|---|
| 2,708,617 | 5/1955 | Magat et al. | 264/183 X |
| 2,908,545 | 10/1959 | Teja | 264/22 X |
| 3,306,835 | 2/1967 | Magnus | 425/174.4 X |
| 3,635,625 | 1/1972 | Voss | 425/162 X |
| 3,775,036 | 11/1973 | Winning | 425/174.4 |
| 3,974,248 | 8/1976 | Atkinson | 425/162 X |
| 4,041,476 | 8/1977 | Swainson | 365/119 |
| 4,078,229 | 3/1978 | Swainson et al. | 365/107 |
| 4,081,276 | 3/1978 | Crivello | 430/269 |
| 4,238,840 | 12/1980 | Swainson | 365/119 |
| 4,252,514 | 2/1981 | Gates | 425/162 |
| 4,288,861 | 9/1981 | Swainson et al. | 365/127 |
| 4,292,015 | 9/1981 | Hritz | 425/162 X |
| 4,329,135 | 5/1982 | Beck | 425/174 |
| 4,333,165 | 6/1982 | Swainson et al. | 365/127 X |
| 4,374,077 | 2/1983 | Kerfeld | 264/22 |
| 4,466,080 | 8/1984 | Swainson et al. | 365/127 X |
| 4,471,470 | 9/1984 | Swainson et al. | 365/127 |

*Primary Examiner*—J. Howard Flint, Jr.

*Attorney, Agent, or Firm*—Fulwider, Patton, Rieber, Lee & Utecht

[57] **ABSTRACT**

A system for generating three-dimensional objects by creating a cross-sectional pattern of the object to be formed at a selected surface of a fluid medium capable of altering its physical state in response to appropriate synergistic stimulation by impinging radiation, particle bombardment or chemical reaction, successive adjacent laminae, representing corresponding successive adjacent cross-sections of the object, being automatically formed and integrated together to provide a step-wise laminar buildup of the desired object, whereby a three-dimensional object is formed and drawn from a substantially planar surface of the fluid medium during the forming process.

**47 Claims, 8 Drawing Figures**

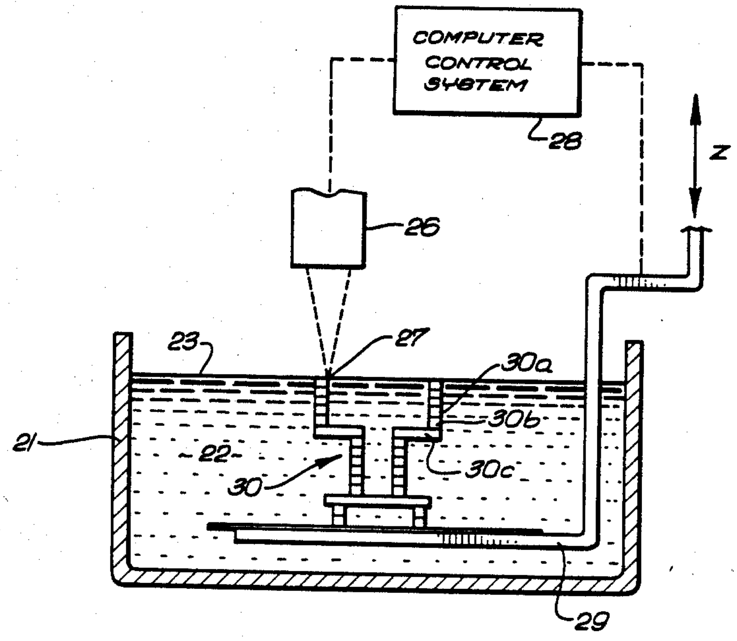


FIG. 1

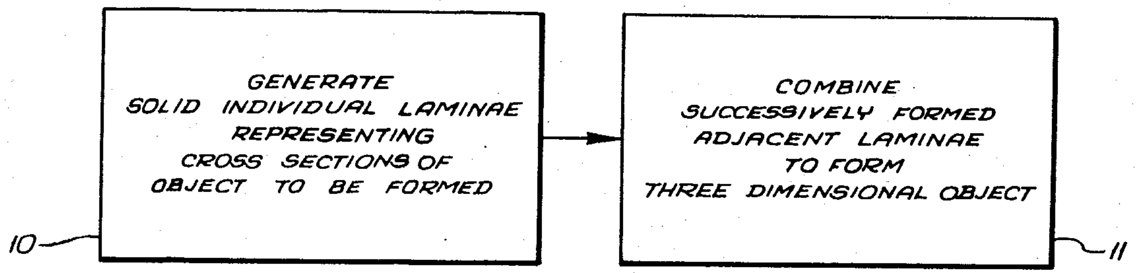

GENERATE SOLID INDIVIDUAL LAMINAE REPRESENTING CROSS SECTIONS OF OBJECT TO BE FORMED
10
COMBINE SUCCESSIVELY FORMED ADJACENT LAMINAE TO FORM THREE DIMENSIONAL OBJECT
11

FIG. 2

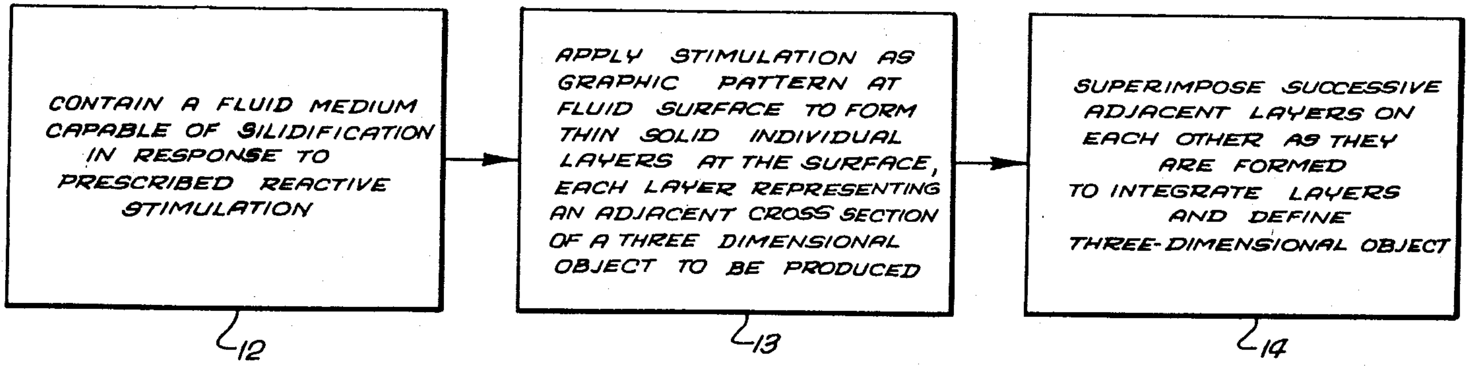

CONTAIN A FLUID MEDIUM CAPABLE OF SILIDIFICATION IN RESPONSE TO PRESCRIBED REACTIVE STIMULATION
12
APPLY STIMULATION AS GRAPHIC PATTERN AT FLUID SURFACE TO FORM THIN SOLID INDIVIDUAL LAYERS AT THE SURFACE, EACH LAYER REPRESENTING AN ADJACENT CROSS SECTION OF A THREE DIMENSIONAL OBJECT TO BE PRODUCED
13
SUPERIMPOSE SUCCESSIVE ADJACENT LAYERS ON EACH OTHER AS THEY ARE FORMED TO INTEGRATE LAYERS AND DEFINE THREE-DIMENSIONAL OBJECT
14

FIG. 3

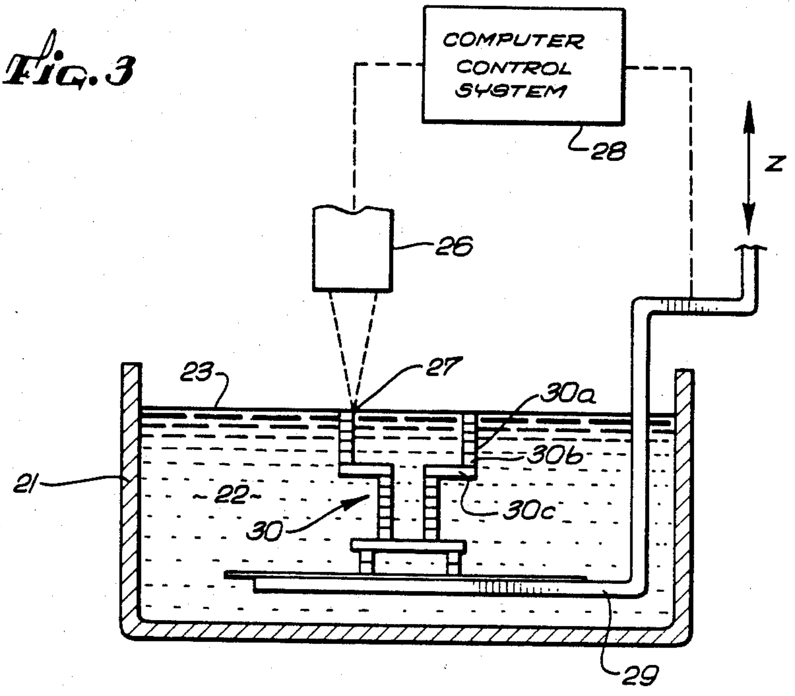

COMPUTER
CONTROL
SYSTEM
28
Z
26
23
27
30a
30b
21
22
30c
30
29

FIG. 4

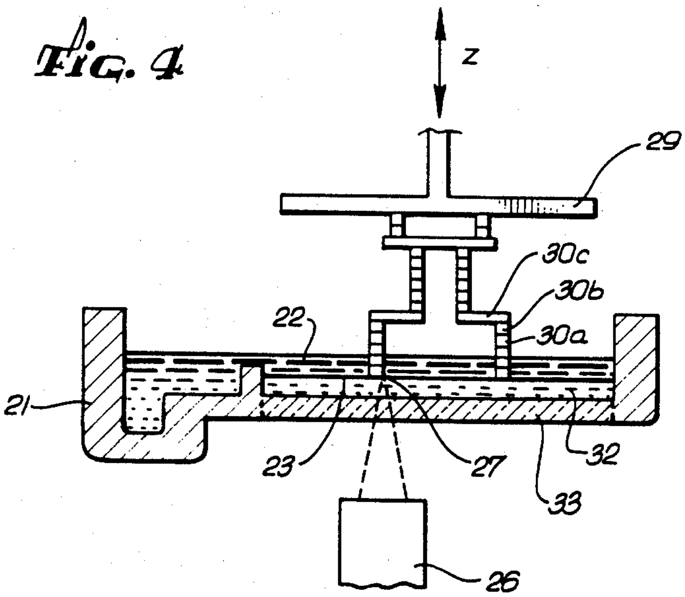

Z
29
30c
30b
22
30a
21
23
27
32
33
26

FIG. 5

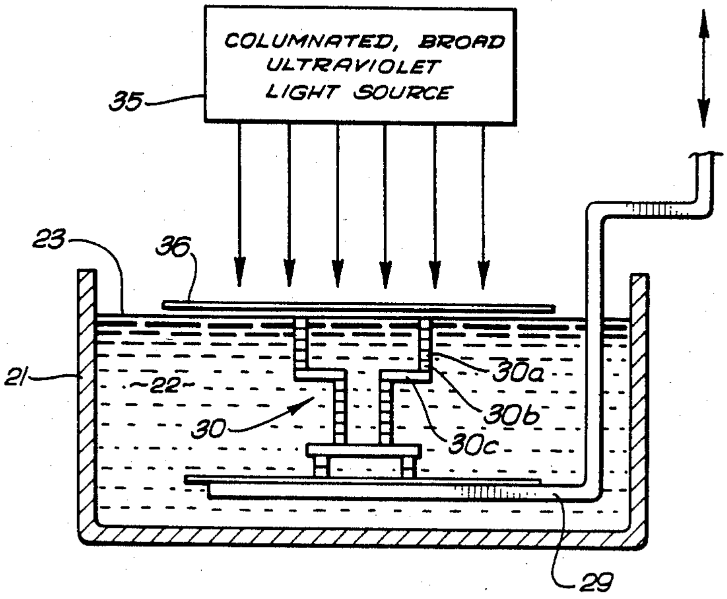

COLUMNATED, BROAD ULTRAVIOLET LIGHT SOURCE
35
23
36
21
~22~
30
30a
30b
30c
29

FIG. 6

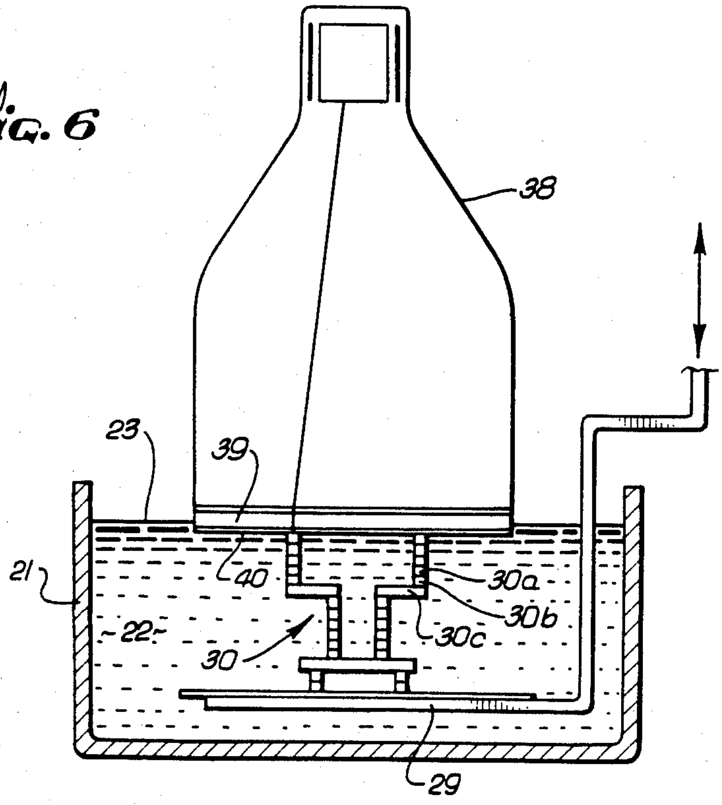

38
23
39
21
40
~22~
30
30a
30b
30c
29

FIG. 7

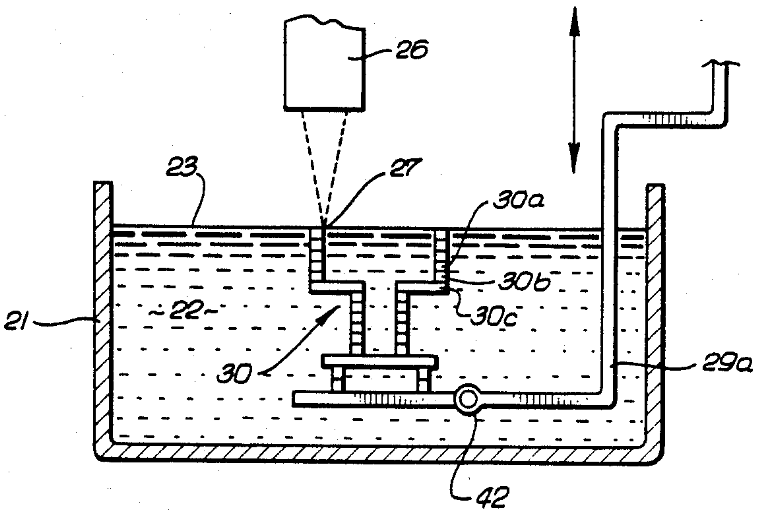
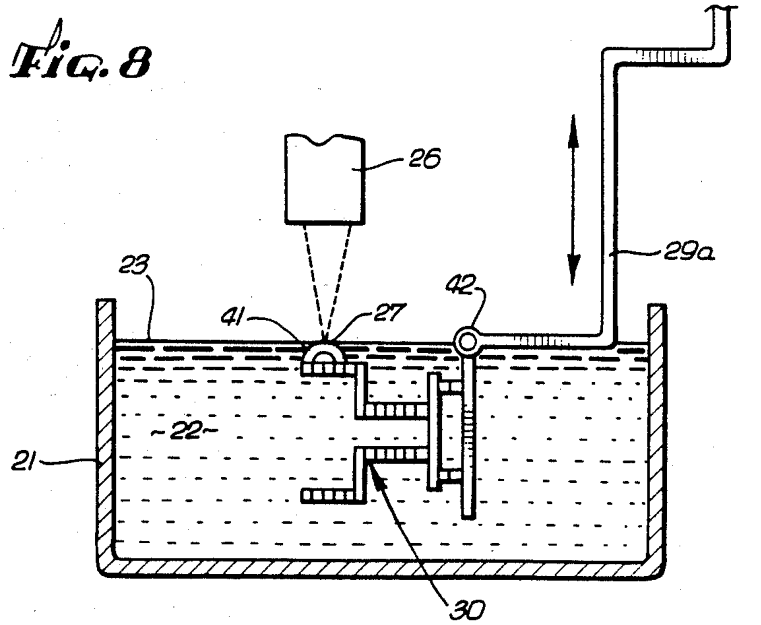


FIG. 8



# APPARATUS FOR PRODUCTION OF THREE-DIMENSIONAL OBJECTS BY STEREOLITHOGRAPHY

## BACKGROUND OF THE INVENTION

This invention relates generally to improvements in apparatus for forming three-dimensional objects from a fluid medium and, more particularly, to stereolithography involving the application of lithographic techniques to production of three-dimensional objects, whereby such objects can be formed rapidly, reliably, accurately and economically.

It is common practice in the production of plastic parts and the like to first design such a part and then painstakingly produce a prototype of the part, all involving considerable time, effort and expense. The design is then reviewed and, oftentimes, the laborious process is again and again repeated until the design has been optimized. After design optimization, the next step is production. Most production plastic parts are injection molded. Since the design time and tooling costs are very high, plastic parts are usually only practical in high volume production. While other processes are available for the production of plastic parts, including direct machine work, vacuum-forming and direct forming, such methods are typically only cost effective for short run production, and the parts produced are usually inferior in quality to molded parts.

In recent years, very sophisticated techniques have been developed for generating three-dimensional objects within a fluid medium which is selectively cured by beams of radiation brought to selective focus at prescribed intersection points within the three-dimensional volume of the fluid medium. Typical of such three-dimensional systems are those described in U.S. Pat. Nos. 4,041,476, 4,078,229, 4,238,840 and 4,288,861. All of these systems rely upon the buildup of synergistic energization at selected points deep within the fluid volume, to the exclusion of all other points in the fluid volume, using a variety of elaborate multibeam techniques. In this regard, the various approaches described in the prior art include the use of a pair of electromagnetic radiation beams directed to intersect at specified coordinates, wherein the various beams may be of the same or differing wavelengths, or where beams are used sequentially to intersect the same points rather than simultaneously, but in all cases only the beam intersection points are stimulated to sufficient energy levels to accomplish the necessary curing process for forming a three-dimensional object within the volume of the fluid medium. Unfortunately, however, such three-dimensional forming systems face a number of problems with regard to resolution and exposure control. The loss of radiation intensity and image forming resolution of the focused spots as the intersections move deeper into the fluid medium create rather obvious complex control situations. Absorption, diffusion, dispersion and defraction all contribute to the difficulties of working deep within the fluid medium on any economical and reliable basis.

Yet there continues to be a long existing need in the design and production arts for the capability of rapidly and reliably moving from the design stage to the prototype stage and to ultimate production, particularly moving directly from computer designs for such plastic parts to virtually immediate prototypes and the facility for large scale production on an economical and automatic basis.

Accordingly, those concerned with the development and production of three-dimensional plastic objects and the like have long recognized the desirability for further improvement in more rapid, reliable, economical and automatic means which would facilitate quickly moving from a design stage to the prototype stage and to production, while avoiding the complicated focusing, alignment and exposure problems of the prior art three dimensional production systems. The present invention clearly fulfills all of these needs.

## SUMMARY OF THE INVENTION

Briefly, and in general terms, the present invention provides a new and improved system for generating a three-dimensional object by forming successive, adjacent, cross-sectional laminae of that object at the surface of a fluid medium capable of altering its physical state in response to appropriate synergistic stimulation, the successive laminae being automatically integrated as they are formed to define the desired three-dimensional object.

In a presently preferred embodiment, by way of example and not necessarily by way of limitation, the present invention harnesses the principles of computer generated graphics in combination with stereolithography, i.e., the application of lithographic techniques to the production of three dimensional objects, to simultaneously execute computer aided design (CAD) and computer aided manufacturing (CAM) in producing three-dimensional objects directly from computer instructions. The invention can be applied for the purposes of sculpturing models and prototypes in a design phase of product development, or as a manufacturing system, or even as a pure art form.

"Stereolithography" is a method and apparatus for making solid objects by successively "printing" thin layers of a curable material, e.g., a UV curable material, one on top of the other. A programmed movable spot beam of UV light shining on a surface or layer of UV curable liquid is used to form a solid cross-section of the object at the surface of the liquid. The object is then moved, in a programmed manner, away from the liquid surface by the thickness of one layer, and the next cross-section is then formed and adhered to the immediately preceding layer defining the object. This process is continued until the entire object is formed.

Essentially all types of object forms can be created with the technique of the present invention. Complex forms are more easily created by using the functions of a computer to help generate the programmed commands and to then send the program signals to the stereolithographic object forming subsystem.

Of course, it will be appreciated that other forms of appropriate synergistic stimulation for a curable fluid medium, such as particle bombardment (electron beams and the like), chemical reactions by spraying materials through a mask or by ink jets, or impinging radiation other than ultraviolet light, may be used in the practice of the invention without departing from the spirit and scope of the invention.

By way of example, in the practice of the present invention, a body of a fluid medium capable of solidification in response to prescribed stimulation is first appropriately contained in any suitable vessel to define a designated working surface of the fluid medium at which successive cross-sectional laminae can be generated. Thereafter, an appropriate form of synergistic stimulation, such as a spot of UV light or the like, is applied as a graphic pattern at the specified working surface of the fluid medium to form thin, solid, individual layers at that surface, each layer representing an adjacent cross-section of the three-dimensional object to be produced. Superposition of successive adjacent layers on each other is automatically accomplished, as they are formed, to integrate the layers and define the desired three-dimensional object. In this regard, as the fluid medium cures and solid material forms as a thin lamina at the working surface, a suitable platform to which the first lamina is secured is moved away from the working surface in a programmed manner by any appropriate actuator, typically all under the control of a micro-computer of the like. In this way, the solid material that was initially formed at the working surface is moved away from that surface and new liquid flows into the working surface position. A portion of this new liquid is, in turn, converted to solid material by the programmed UV light spot to define a new lamina, and this new lamina adhesively connects to the material adjacent to it, i.e., the immediately preceding lamina. This process continues until the entire three-dimensional object has been formed. The formed object is then removed from the container and the apparatus is ready to produce another object, either identical to the first object or an entirely new object generated by a computer or the like.

The stereolithographic apparatus of the present invention has many advantages over currently used apparatus for producing plastic objects. The apparatus of the present invention avoids the need of producing design layouts and drawings, and of producing tooling drawings and tooling. The designer can work directly with the computer and a stereolithographic device, and when he is satisfied with the design as displayed on the output screen of the computer, he can fabricate a part for direct examination. If the design has to be modified, it can be easily done through the computer, and then another part can be made to verify that the change was correct. If the design calls for several parts with interacting design parameters, the method of the invention becomes even more useful because all of the part designs can be quickly changed and made again so that the total assembly can be made and examined, repeatedly if necessary.

After the design is complete, part production can begin immediately, so that the weeks and months between design and production are avoided. Ultimate production rates and parts costs should be similar to current injection molding costs for short run production, with even lower labor costs than those associated with injection molding. Injection molding is economical only when large numbers of identical parts are required. Stereolithography is useful for short run production because the need for tooling is eliminated and production set-up time is minimal. Likewise, design changes and custom parts are easily provided using the technique. Because of the ease of making parts, stereolithography can allow plastic parts to be used in many places where metal or other material parts are now used. Moreover, it allows plastic models of objects to be quickly and economically provided, prior to the decision to make more expensive metal or other material parts.

Hence, the stereolithographic apparatus of the present invention satisfies a long existing need for a CAD and CAM system capable of rapidly, reliably, accurately and economically designing and fabricating three-dimensional plastic parts and the like.

The above and other objects and advantages of this invention will be apparent from the following more detailed description when taken in conjunction with the accompanying drawings of illustrative embodiments.

## BRIEF DESCRIPTION OF THE DRAWINGS

FIG. **1** and FIG. **2** are flow charts illustrating the basic concepts employed in practicing the method of stereolithography of the present invention;

FIG. **3** is a combined block diagram, schematic and elevational sectional view of a presently preferred embodiment of a system for practicing the invention;

FIG. **4** is an elevational sectional view of a second embodiment of a stereolithography system for the practice of the invention;

FIG. **5** is an elevational sectional view, illustrating a third embodiment of the present invention;

FIG. **6** is an elevational sectional view illustrating still another embodiment of the present invention; and

FIGS. **7** and **8** are partial, elevational sectional views, illustrating a modification of the stereolithographic system of FIG. **3** to incorporate an elevator platform with multiple degrees of freedom.

## DESCRIPTION OF THE PREFERRED EMBODIMENT

Referring now to the drawings, FIGS. **1** and **2** are flow charts illustrating the basic system of the present invention for generating three-dimensional objects by means of stereolithography.

Many liquid state chemicals are known which can be induced to change to solid state polymer plastic by irradiation with ultraviolet light (UV) or other forms of synergistic stimulation such as electron beams, visible or invisible light, reactive chemicals applied by ink jet or via a suitable mask. UV curable chemicals are currently used as ink for high speed printing, in processes of coating of paper and other materials, as adhesives, and in other specialty areas.

Lithography is the art of reproducing graphic objects, using various techniques. Modern examples include photographic reproduction, xerography, and microlithography, as is used in the production of microelectronics. Computer generated graphics displayed on a plotter or a cathode ray tube are also forms of lithography, where the image is a picture of a computer coded object.

Computer aided design (CAD) and computer aided manufacturing (CAM) are techniques that apply the abilities of computers to the processes of designing and manufacturing. A typical example of CAD is in the area of electronic printed circuit design, where a computer and plotter draw the design of a printed circuit board, given the design parameters as computer data input. A typical example of CAM is a numerically controlled milling machine, where a computer and a milling machine produce metal parts, given the proper programming instructions. Both CAD and CAM are important and are rapidly growing technologies.

A prime object of the present invention is to harness the principles of computer generated graphics, combined with UV curable plastic and the like, to simultaneously execute CAD and CAM, and to produce three-dimensional objects directly from computer instructions. This invention, referred to as stereolithography, can be used to sculpture models and prototypes in a design phase of product development, or as a manufacturing device, or even as an art form.

Referring now to FIG. 1, the stereolithographic method is broadly outlined. Step 10 in FIG. 1 calls for the generation of individual solid laminae representing cross-sections of a three-dimensional object to be formed. Step 11, which inherently occurs if Step 10 is performed properly, combines the successively formed adjacent laminae to form the desired three-dimensional object which has been programmed into the system for selective curing. Hence, the stereolithographic system of the present invention generates three-dimensional objects by creating a cross-sectional pattern of the object to be formed at a selected surface of a fluid medium, e.g., a UV curable liquid or the like, capable of altering its physical state in response to appropriate synergistic stimulation such as impinging radiation, electron beam or other particle bombardment, or applied chemicals (as by ink jet or spraying over a mask adjacent the fluid surface), successive adjacent laminae, representing corresponding successive adjacent cross-sections of the object, being automatically formed and integrated together to provide a step-wise laminar or thin layer buildup of the object, whereby a three-dimensional object is formed and drawn from a substantially planar or sheet-like surface of the fluid medium during the forming process.

The aforedescribed technique is more specifically outlined in the flowchart of FIG. 2, wherein Step 12 calls for containing a fluid medium capable of solidification in response to prescribed reactive stimulation. Step 13 calls for application of that stimulation as a graphic pattern at a designated fluid surface to form thin, solid, individual layers at that surface, each layer representing an adjacent cross-section of a three-dimensional object to be produced. It is desirable to make each such layer as thin as possible during the practice of the invention in order to maximize resolution and the accurate reproduction of the three-dimensional object being formed. Hence, the ideal theoretical state would be an object produced only at the designated working surface of the fluid medium to provide an infinite number of laminae, each lamina having a cured depth of approximately only slightly more than zero thickness. Of course, in the practical application of the invention, each lamina will be a thin lamina, but thick enough to be adequately cohesive in forming the cross-section and adhering to the adjacent laminae defining other cross-sections of the object being formed.

Step 14 in FIG. 2 calls for superimposing successive adjacent layers or laminae on each other as they are formed, to integrate the various layers and define the desired three-dimensional object. In the normal practice of the invention, as the fluid medium cures and solid material forms to define one lamina, that lamina is moved away from the working surface of the fluid medium and the next lamina is formed in the new liquid which replaces the previously formed lamina, so that each successive lamina is superimposed and integral with (by virtue of the natural adhesive properties of the cured fluid medium) all of the other cross-sectional laminae. Hence, the process of producing such cross-sectional laminae is repeated over and over again until the entire three-dimensional object has been formed. The object is then removed and the system is ready to produce another object which may be identical to the previous object or may be an entirely new object formed by changing the program controlling the stereolithographic system.

FIGS. 3-8 of the drawings illustrate various apparatus suitable for implementing the stereolithographic methods illustrated and described by the flow charts of FIGS. 1 and 2.

As previously indicated, "Stereolithography" is a method and apparatus for making solid objects by successively "printing" thin layers of a curable material, e.g., a UV curable material, one on top of the other. A programmed movable spot beam of UV light shining on a surface or layer of UV curable liquid is used to form a solid cross-section of the object at the surface of the liquid. The object is then moved, in a programmed manner, away from the liquid surface by the thickness of one layer and the next cross-section is then formed and adhered to the immediately preceding layer defining the object. This process is continued until the entire object is formed.

Essentially all types of object forms can be created with the technique of the present invention. Complex forms are more easily created by using the functions of a computer to help generate the programmed commands and to then send the program signals to the stereolithographic object forming subsystem.

A presently preferred embodiment of the stereolithographic system is shown in elevational cross-section in FIG. 3. A container 21 is filled with a UV curable liquid 22 or the like, to provide a designated working surface 23. A programmable source of ultraviolet light 26 or the like produces a spot of ultraviolet light 27 in the plane of surface 23. The spot 27 is movable across the surface 23 by the motion of mirrors or other optical or mechanical elements (not shown) that are a part of light source 26. The position of the spot 27 on surface 23 is controlled by a computer or other programming device 28. A movable elevator platform 29 inside container 21 can be moved up and down selectively, the position of the platform being controlled by the computer 28. As the device operates, it produces a three-dimensional object 30 by step-wise buildup of integrated laminae such as 30*a*, 30*b*, 30*c*.

The surface of the UV curable liquid 22 is maintained at a constant level in the container 21, and the spot of UV light 27, or other suitable form of reactive stimulation, of sufficient intensity to cure the liquid and convert it to a solid material is moved across the working surface 23 in a programmed manner. As the liquid 22 cures and solid material forms, the elevator platform 29 that was initially just below surface 23 is moved down from the surface in a programmed manner by any suitable actuator. In this way, the solid material that was initially formed is taken below surface 23 and new liquid 22 flows across the surface 23. A portion of this new liquid is, in turn, converted to solid material by the programmed UV light spot 27, and the new material adhesively connects to the material below it. This process is continued until the entire three-dimensional object 30 is formed. The object 30 is then removed from the container 21, and the apparatus is ready to produce another object. Another object can then be produced, or some new object can be made by changing the program in the computer 28.

The curable liquid 22, e.g., UV curable liquid, must have several important properties. (A) It must cure fast enough with the available UV light source to allow practical object formation times. (B) It must be adhesive, so that successive layers will adhere to each other.

(C) Its viscosity must be low enough so that fresh liquid material will quickly flow across the surface when the elevator moves the object. (D) It should absorb UV so that the film formed will be reasonably thin. (E) It must be reasonably soluble in some solvent in the liquid state, and reasonably insoluble in that same solvent in the solid state, so that the object can be washed free of the UV cure liquid and partially cured liquid after the object has been formed. (F) It should be as non-toxic and non-irritating as possible.

The cured material must also have desirable properties once it is in the solid state. These properties depend on the application involved, as in the conventional use of other plastic materials. Such parameters as color, texture, strength, electrical properties, flammability, and flexibility are among the properties to be considered. In addition, the cost of the material will be important in many cases.

The UV curable material used in the presently preferred embodiment of a working stereolithograph (e.g., FIG. 3) is Potting Compound 363, a modified acrylate, made by Locktite Corporation of Newington, CT. A process to make a typical UV curable material is described in U.S. Pat. No. 4,100,141, entitled Stabilized Adhesive and Curing Compositions.

The light source **26** produces the spot **27** of UV light small enough to allow the desired object detail to be formed, and intense enough to cure the UV curable liquid being used quickly enough to be practical. The source **26** is arranged so it can be programmed to be turned off and on, and to move, such that the focused spot **27** moves across the surface **23** of the liquid **22**. Thus, as the spot **27** moves, it cures the liquid **22** into a solid, and "draws" a solid pattern on the surface in much the same way a chart recorder or plotter uses a pen to draw a pattern on paper.

The light source **26** for the presently preferred embodiment of a stereolithograph is made using a 350 watt mercury short arc lamp in a housing, with the light output of the housing focused on the end of a 1 mm diameter UV transmitting fiber optic bundle (not shown). The end of the bundle next to the lamp is water cooled, and there is an electronically controlled shutter blade between the lamp and the end of the bundle, which can turn the light through the bundle on and off. The bundle is 1 meter long, and the optical output is fitted into a lens tube that has a quartz to focus the UV to a spot. The light source **26** is capable of producing a spot somewhat less than 1 mm in diameter, with a long wave UV intensity of about 1 watt/cm2.

In the system of FIG. 3, means may be provided to keep the surface **23** at a constant level and to replenish this material after an object has been removed, so that the focus spot **27** will remain sharply in focus on a fixed focus plane, thus insuring maximum resolution in forming a thin layer along the working surface. In this regard, it is desired to shape the focal point to provide a region of high intensity right at the working surface **23**, rapidly diverging to low intensity and thereby limiting the depth of the curing process to provide the thinnest appropriate cross-sectional laminae for the object being formed. This is best accomplished by using a short focal length lens and bringing the source **26** as close as possible to the working surface, so that maximum divergence occurs in the cone of focus entering the fluid medium. The result is substantially enhanced resolution.

An H-P Model 9872 Digital Plotter (not shown) manufactured by Hewlett-Packard, of Palo Alto, Calif., is used to move the light source **26**. The lens tube is attached to the pen carriage of the plotter, and the plotter is driven by a computer **28** using normal graphic commands. The shutter is controlled by an H-P 3497A Data Acquisition/Control Unit, using computer commands.

Other physical forms of the light source **26** or its equivalent are feasible. Scanning could be done with optical scanners, and this would eliminate the fiber optic bundle and the digital plotter. A UV laser might ultimately be a better light source than a short arc lamp. The speed of the stereolithographic process is mainly limited by the intensity of the light source and the response of the UV curable liquid.

The elevator platform **29** is used to support and hold the object **30** being formed, and to move it up and down as required. Typically, after a layer is formed, the object **30** is moved beyond the level of the next layer to allow the liquid **22** to flow into the momentary void at surface **23** left where the solid was formed, and then it is moved back to the correct level for the next layer. The requirements for the elevator platform **29** are that it can be moved in a programmed fashion at appropriate speeds, with adequate precision, and that it is powerful enough to handle the weight of the object **30** being formed. In addition, a manual fine adjustment of the elevator platform position is useful during the set-up phase and when the object is being removed.

The elevator platform **29** for the embodiment of FIG. **3** is a platform attached to an analog plotter (not shown). This plotter is driven the H-P 3497A Data Acquisition/Control Unit with its internal digital to analog converter, under program control of the computer **28**.

The computer **28** in the stereolithographic system of the present invention has two basic functions. The first is to help the operator design the three-dimensional object in a way that it can be made. The second is to translate the design into commands that are appropriate for the other stereolithographic components, and to deliver these commands in a way so that the object is formed. In some applications, the object design will exist, and the only function of the computer will be to deliver the appropriate commands.

In an ideal situation, the operator will be able to design the object and view it three-dimensionally on the CRT screen of the computer **28**. When he is finished with the design, he will instruct the computer **28** to make the object, and the computer will issue the appropriate instructions to the stereolithographic components.

In a present working embodiment of the invention, the computer **28** is an H-P 9816, using a Basic Operating System. A typical program is shown in Appendix A. In this system, the operator programs using H-P Graphic Language, the command structure for the 3497A, plus the Basic Language commands. The operator also must set the appropriate exposure times and rates for the UV curable material. To operate the system an image of the object is created and a program is written to drive the stereolithograph to make that object.

The elevator platform **29** can be mechanical, pneumatic, hydraulic, or electrical and may also use optical or electronic feedback to precisely control its position. The elevator platform **29** is typically fabricated of either glass or aluminum, but any material to which the cured plastic material will adhere is suitable.

In some cases, the computer **28** becomes unnecessary and simpler dedicated programming devices can be used, particularly where only simply shaped objects are to be formed. Alternatively, the computer control system **28** can be simply executing instructions that were generated by another, more complex, computer. This might be the case where several stereolithography units are used to produce objects, and another device is used to initially design the objects to be formed.

A computer controlled pump (not shown) may be used to maintain a constant level of the liquid **22** at the working surface **23**. Appropriate level detection system and feedback networks, well known in the art, can be used to drive a fluid pump or a liquid displacement device, such as a solid rod (not shown) which is moved out of the fluid medium as the elevator platform is moved further into the fluid medium, to offset changes in fluid volume and maintain constant fluid level at the surface **23**. Alternatively, the source **26** can be moved relative to the sensed level **23** and automatically maintain sharp focus at the working surface **23**. All of these alternatives can be readily achieved by conventional software operating in conjunction with the computer control system **28**.

After the three-dimensional object **30** has been formed, the elevator platform **29** is raised and the object is removed from the platform. Typically, the object is then ultrasonically rinsed in a solvent, such as acetone, that dissolves the liquid state of the uncured fluid medium and not the cured solid state medium. The object **30** is then placed under an intense ultraviolet floodlight, typically a 200 watt per inch UV cure lamp, to complete the curing process.

In addition, there may be several containers **21** used in the practice of the invention, each container having a different type of curable material that can be automatically selected by the stereolithographic system. In this regard, the various materials might provide plastics of different colors, or have both insulating and conducting material available for the various layers of electronic products.

Referring now more particularly to the remaining drawings, in connection with various alternative embodiments of the invention, like reference numerals throughout the various figures of the drawings denote like or corresponding parts as those previously discussed in connection with the preferred embodiment of the invention shown in FIG. **3**.

As will be apparent from FIG. **4** of the drawings, there is shown an alternate configuration for a stereolithograph wherein the UV curable liquid **22** or the like floats on a heavier UV transparent liquid **32** which is non-miscible and non-wetting with the curable liquid **22**. By way of example, ethylene glycol or heavy water are suitable for the intermediate liquid layer **32**. In the system of FIG. **4**, the three-dimensional object **30** is pulled up from the liquid **22**, rather than down and further into the liquid medium, as shown in the system of FIG. **3**.

The UV light source **26** in FIG. **4** focuses the spot **27** at the interface between the liquid **22** and the non-miscible intermediate liquid layer **32**, the UV radiation passing through a suitable UV transparent window **33**, of quartz or the like, supported at the bottom of the container **21**. The curable liquid **22** is provided in a very thin layer over the non-miscible layer **32** and thereby has the advantage of limiting layer thickness directly, rather than relying solely upon adsorption and the like to limit the depth of curing, since ideally an ultrathin lamina is to be provided. Hence, the region of formation will be more sharply defined and some surfaces will be formed smoother with the system of FIG. **4** than with that of FIG. **3**. In addition, a smaller volume of UV curable liquid **22** is required, and the substitution of one curable material for another is easier.

The system of FIG. **5** is similar to that of FIG. **3**, but the movable UV light source **26** is eliminated and a collimated, broad UV light source **35** and suitable apertured mask **36** is substituted for the programmed source **26** and focused spot **27**. The apertured mask **36** is placed as close as possible to the working surface **23**, and collimated light from the UV source **35** passes through the mask **36** to expose the working surface **23**, thereby creating successive adjacent laminae, as in the embodiments of FIGS. **3** and **4**. However, the use of a fixed mask **36** provides three-dimensional objects with a constant cross-sectional shape. Whenever that cross-sectional shape is to be changed, a new mask **36** for that particular cross-sectional shape must be substituted and properly aligned. Of course, the masks can be automatically changed by providing a web of masks (not shown) which are successively moved into alignment with with the surface **23**.

FIG. **6** of the drawings again provides a stereolithographic system configuration similar to that previously described in connection with FIG. **3**. However, a cathode ray tube (CRT) **38**, fiber optic faceplate **39** and water (or other) release layer **40** are provided as a substitute for the light source **26** and focus spot **27**. Hence, the graphic image provided by a computer **28** to the CRT **38** produces the forming image upon the UV emitting phosphor face of the tube where it passes through the fiber optic layer **39** and release layer **40** to the working surface **23** of the fluid medium **22**. In all other respects, the system of FIG. **6** forms successive cross-sectional laminae defining the desired three-dimensional object to be formed, in exactly the same way as the embodiments of the invention previously discussed.

FIGS. **7** and **8** illustrate an embodiment of a stereolithographic system wherein the elevator platform **29** has additional degrees of freedom, so that different faces of the object **30** may be exposed for alternate methods of construction. Similarly, the stereolithography process may be utilized as an "add on" process where the elevator platform **29** will be used to pick up and locate another part for supplementary stereolithographic processing. In this regard, the systems shown in FIGS. **7** and **8** are identical to that of FIG. **3** with the exception of the elevator platform **29** which, in the systems of FIGS. **7** and **8** have a second degree of freedom via manual or automatically controlled rotation about a pivot pin or hinge member **42**. In this regard, FIG. **7** illustrates an adjustable elevator platform **29***a* in the conventional position, while FIG. **8** shows the platform **29***a* rotated 90° so that a supplementary, stereolithographically formed structure **41** can be selectively formed as an addition to one side of the three-dimensional object **30**.

A commercial stereolithography system will have additional components and subsystems besides those previously shown in connection with the schematically depicted systems of FIGS. **3–8**. For example, the commercial system would also have a frame and housing, and a control panel. It should have means to shield the operator from excess UV and visible light, and it may also have means to allow viewing of the object **30** while it is being formed. Commercial units will provide safety means for controlling ozone and noxious fumes, as well as conventional high voltage safety protection and interlocks. Such commercial units will also have means to effectively shield the sensitive electronics from electronic noise sources.

As previously mentioned, a number of other possible apparatus may be utilized to practice the stereolithographic method. For example, an electron source, a visible light source, or an x-ray source or other radiation source could be substituted for the UV light source 26, along with appropriate fluid media which are cured in response to these particular forms of reactive stimulation. For example, alphaoctadecylacrylic acid that has been slightly prepolymerized with UV light can be polymerized with an electron beam. Similarly, poly(2,3-dichloro-1-propyl acrylate) can be polymerized with an x-ray beam.

The stereolithographic method and apparatus has many advantages over currently used methods for producing plastic objects. The method avoids the need of producing design layouts and drawings, and of producing tooling drawings and tooling. The designer can work directly with the computer and a stereolithographic device, and when he is satisfied with the design as displayed on the output screen of the computer, he can fabricate a part for direct examination. If the design has to be modified, it can be easily done through the computer, and then another part can be made to verify that the change was correct. If the design calls for several parts with interacting design parameters, the method becomes even more useful because all of the part designs can be quickly changed and made again so that the total assembly can be made and examined, repeatedly if necessary.

After the design is complete, part production can begin immediately, so that the weeks and months between design and production are avoided. Ultimate production rates and parts costs should be similar to current injection molding costs for short run production, with even lower labor costs than those associated with injection molding. Injection molding is economical only when large numbers of identical parts are required. Stereolithography is useful for short run production because the need for tooling is eliminated and production set-up time is minimal. Likewise, design changes and custom parts are easily provided using the technique. Because of the ease of making parts, stereolithography can allow plastic parts to be used in many places where metal or other material parts are now used. Moreover, it allows plastic models of objects to be quickly and economically provided, prior to the decision to make more expensive metal or other material parts.

It will be apparent from the foregoing that, while a variety of stereolithographic systems have been disclosed for the practice of the present invention, they all have in common the concept of drawing upon a substantially two-dimensional surface and extracting a three-dimensional object from that surface.

The present invention satisfies a long existing need in the art for a CAD and CAM system capable of rapidly, reliably, accurately and economically designing and fabricating three-dimensional plastic parts and the like.

It will be apparent from the foregoing that, while particular forms of the invention have been illustrated and described, various modifications can be made without departing from the spirit and scope of the invention.

Accordingly, it is not intended that the invention be limited, except as by the appended claims.

I claim:

1. A system for producing a three-dimensional object from a fluid medium capable of solidification when subjected to prescribed synergistic stimulation, said system comprising:

means for drawing upon and forming successive cross-sectional laminae of said object at a two-dimensional interface; and

means for moving said cross-sections as they are formed and building up said object in step wise fashion, whereby a three-dimensional object is extracted from a substantially two-dimensional surface.

2. An improved system for producing a three-dimensional object from a fluid medium capable of solidification when subjected to prescribed synergistic stimulation, said system comprising:

a body of fluid medium capable of transforming its physical state in response to synergistic stimulation;

object support means immersed within said fluid medium for supporting a three-dimensional object to be formed;

translational means for selectively moving said object support means progressively away from a designated surface of said fluid medium; and

reaction means capable of altering the physical state of said fluid medium and operating in a prescribed pattern upon said designated surface of said fluid medium to provide a thin solid lamina at said surface representing a corresponding cross-sectional lamina of said three-dimensional object to be formed,

whereby successive adjacent laminae are provided to form said three-dimensional object on said object support means as said translational means moves said support means away from said designated surface.

3. A system as set forth in claim 2, and further including:

programmed control means for varying the graphic pattern of said reaction means operating upon said designated surface of said fluid medium.

4. A system as set forth in claim 2, wherein said reaction means includes:

a beam of impinging radiation.

5. A system as set forth in claim 2, wherein said reaction means includes:

an electron beam.

6. A system as set forth in claim 2, wherein said reaction means includes:

a beam of high energy particles.

7. A system as set forth in claim 2, wherein said reaction means includes:

a beam of light.

8. A system as set forth in claim 2, wherein said reaction means includes:

x-rays.

9. A system as set forth in claim 2, wherein said reaction means includes:

a beam of ultraviolet light.

10. A system as set forth in claim 2, wherein said reaction means includes:

a jet of a reactive chemical to induce solidification of said fluid medium.

11. A system as set forth in claim 2, wherein said reaction means includes:

a patterned mask overlying said designated surface for selectively applying a chemical to induce solidification of said fluid medium.

12. A system as set forth in claim 2, wherein said reaction means includes:

a patterned mask overlying said designated surface for selectively exposing said surface to synergistic stimulation.

13. A system as set forth in claim 2, wherein said reaction means includes:

a patterned mask overlying said designated surface for selectively exposing said surface to radiation.

14. A system as set forth in claim 2, wherein said translational means moves said object as it is formed away from said designated surface and further into said fluid medium.

15. A system as set forth in claim 2, wherein said translational means moves said object as it is formed away from said surface and out of said fluid medium.

16. A system as set forth in claim 2, wherein exposure to said reaction means at said designated surface is through a second non-reactive medium.

17. A system as set forth in claim 2, and further including:

a container for said fluid medium, wherein exposure of said designated surface to said reaction means is through the bottom of said container and a second non-reactive medium adjacent said designated surface.

18. A system as set forth in claim 17, wherein said second non-reactive medium is heavy water.

19. A system as set forth in claim 17, wherein said second non-reactive medium is ethylene glycol.

20. A system as set forth in claim 2, and further including:

rotational means, supplementing said translational means for altering the orientation of said object relative to said designated surface at which laminae are being formed.

21. A system as set forth in claim 2, wherein the level of said fluid medium locating said designated surface is variable.

22. A system as set forth in claim 2, wherein the level of said fluid medium locating said designated surface is maintained constant.

23. A system as set forth in claim 2, wherein said translational means has multiple degrees of freedom of movement.

24. A system as set forth in claim 4, wherein precise focus of said beam of impinging radiation upon said designated surface is maintained.

25. A system as set forth in claim 2, wherein said prescribed pattern is formed upon said designated surface by radiation emanating from the face of a cathode ray tube.

26. A system as set forth in claim 2, wherein said prescribed pattern is formed by light directly emanating from a phosphor image.

27. A system for directly producing a three-dimensional object as it is designed by a computer, comprising:

deriving graphic image output from said computer, said graphic image defining successive adjacent cross-sections of the three-dimensional object designed by said computer;

means for drawing upon and forming successive cross-sections, corresponding to said computer designed cross-sections of said object, at a two-dimensional interface; and

means for moving said cross-sections as they are formed and building up said object in a stepwise fashion, whereby the three-dimensional object designed by said computer is automatically extracted from a substantially two-dimensional surface.

28. An improved system for producing a three-dimensional object from a fluid medium capable of altering its physical state when subjected to prescribed radiation, said system comprising:

a body of fluid medium capable of altering its physical state;

means for forming said three-dimensional object from said fluid medium by irradiating a designated surface of said medium to provide integrated, successive surface laminae at said surface, said laminae together defining said three-dimensional object.

29. An improved system for producing a three-dimensional object from a fluid medium, said system comprising:

a body of fluid medium capable of altering its physical state in response to prescribed radiation;

a radiation source for impinging said prescribed radiation in a selected pattern upon a designated surface of said fluid medium to provide only at said surface a thin solid lamina representing a cross-sectional lamina of a three-dimensional object to be formed; and

means for combining successive adjacent laminae to form said three-dimensional object from said fluid medium.

30. A system as set forth in claim 29, wherein said radiation source includes:

a beam of impinging radiation.

31. A system as set forth in claim 29, wherein said radiation source includes:

an electron beam.

32. A system as set forth in claim 29, wherein said radiation source includes:

a beam of high energy particles.

33. A system as set forth in claim 29, wherein said radiation source includes:

a beam of light.

34. A system as set forth in claim 29, wherein said radiation source includes:

a beam of ultraviolet light.

35. A system as set forth in claim 29, wherein said radiation source includes:

x-rays.

36. A system as set forth in claim 29, wherein said radiation source and pattern includes:

a patterned mask overlying said designated surface for selectively exposing said surface to synergistic stimulation.

37. A system as set forth in claim 29, wherein said radiation source and pattern includes:

a patterned mask overlying said designated surface selectively exposing said surface to radiation.

38. A system as set forth in claim 29, wherein exposure to said prescribed radiation at said designated surface is through a second non-reactive medium.

39. A system as set forth in claim 29, and further including:

a container for said fluid medium, wherein exposure of said designated surface to said prescribed radiation is through the bottom of said container and a second non-reactive medium adjacent said designated surface.

**40**. A system as set forth in claim **39**, wherein said second non-reactive medium is heavy water.

**41**. A system as set forth in claim **39**, wherein said second non-reactive medium is ethylene glycol.

**42**. A system as set forth in claim **39**, wherein the level of said fluid medium locating said designated surface is maintained constant.

**43**. A system as set forth in claim **39**, wherein said translational means has multiple degrees of freedom of movement.

**44**. A system as set forth in claim **39**, wherein precise focus of said prescribed radiation upon said designated surface is maintained.

**45**. A system as set forth in claim **39**, wherein said selected pattern is formed upon said designted surface by radiation emanating from the face of a cathode ray tube.

**46**. A system as set forth in claim **39**, wherein said selected pattern is formed by light directly emanating from a phosphor image.

**47**. A system as set forth in claim **39**, and further including:

programmed control means for varying the pattern of said impinging radiation upon said designated surface of said fluid medium.

* * * * *

# REEXAMINATION CERTIFICATE (1177th)

# United States Patent [19]

**Hull**

[11] **B1 4,575,330**

[45] Certificate Issued **Dec. 19, 1989**

[54] **APPARATUS FOR PRODUCTION OF THREE-DIMENSIONAL OBJECTS BY STEREOLITHOGRAPHY**

[75] Inventor: **Charles W. Hull**, Arcadia, Calif.

[73] Assignee: **UVP, Inc.**

**Reexamination Request:**
No. 90/001,611, Sep. 28, 1988

**Reexamination Certificate for:**

| | |
|---|---|
| Patent No.: | **4,575,330** |
| Issued: | **Mar. 11, 1986** |
| Appl. No.: | **638,905** |
| Filed: | **Aug. 8, 1984** |

[51] **Int. Cl.⁴** .......................... **B29C 35/08**; G03C 9/08

[52] **U.S. Cl.** .................................... **425/174.4**; 156/58; 264/22; 365/119; 365/120; 425/162; 425/174

[58] **Field of Search** ...................... 425/162, 174, 174.4, 425/425; 264/22, 40.1, 25, 219, 183, 250, 255, 298, 308; 365/107, 119, 127, 106, 120; 156/58, 246, 273.3, 273.5, 275.5; 250/432 R, 433, 492.3, 492.1, 558; 522/910; 430/269, 296, 942; 342/179; 427/53.1, 54.1, 55

[56] **References Cited**

U.S. PATENT DOCUMENTS

| | | | |
|---|---|---|---|
| 2,381,234 | 8/1945 | Symmes | 430/294 |
| 2,525,532 | 10/1950 | Dreywood | 430/419 |
| 2,708,617 | 5/1955 | Magat et al. | 264/183 X |
| 2,775,758 | 12/1956 | Munz | 342/179 |
| 2,908,545 | 10/1959 | Teja | 264/22 X |
| 3,306,835 | 2/1967 | Magnus | 425/174.4 X |
| 3,428,503 | 2/1969 | Beckerle | 430/269 |
| 3,609,707 | 9/1971 | Lewis et al. | 365/119 |
| 3,635,625 | 1/1972 | Voss | 425/162 X |
| 3,723,120 | 3/1973 | Hummel | 522/910 X |
| 3,775,036 | 11/1973 | Winning | 425/174.4 |
| 3,866,052 | 2/1975 | Di Matteo | 250/558 |
| 3,932,923 | 1/1976 | Di Matteo | 29/407 |
| 3,974,248 | 8/1976 | Atkinson | 425/162 X |
| 4,041,476 | 8/1977 | Swainson | 365/119 |
| 4,078,229 | 3/1978 | Swainson et al. | 365/107 |
| 4,081,276 | 3/1978 | Crivello | 430/269 |
| 4,100,141 | 7/1978 | O'Sullivan | 526/301 |
| 4,238,840 | 12/1980 | Swainson | 365/119 |
| 4,247,508 | 1/1981 | Housholder | 264/219 |
| 4,252,514 | 2/1981 | Gates | 425/162 |
| 4,288,861 | 9/1981 | Swainson et al. | 365/127 |
| 4,292,015 | 9/1981 | Hritz | 425/162 X |
| 4,329,135 | 5/1982 | Beck | 425/174 |
| 4,333,165 | 6/1982 | Swainson et al. | 365/127 X |
| 4,374,077 | 2/1983 | Kerfield | 264/22 |
| 4,466,080 | 8/1984 | Swainson et al. | 365/127 X |
| 4,471,470 | 9/1984 | Swainson et al. | 365/127 |

FOREIGN PATENT DOCUMENTS

| | | |
|---|---|---|
| 1917294 | 4/1969 | Fed. Rep. of Germany . |
| 2553039 | 4/1985 | France . |
| 144478 | 11/1981 | Japan . |
| 58-211413 | 12/1983 | Japan . |
| 566795 | 1/1945 | United Kingdom . |
| 1556451 | 11/1979 | United Kingdom . |
| 2035602 | 6/1980 | United Kingdom . |

OTHER PUBLICATIONS

H. Kodama, "Automatic Method for Fabricating a Three-Dimensional Plastic Model with Photo-Hardening Polymer," Review of Scientific Instruments, vol 52, No. 11, Nov. 1981, pp. 1770–1773.

A. J. Herbert, "Solid Object Generation," Journal of Applied Photographic Engineering, 8(4), Aug. 1982, pp. 185–188.

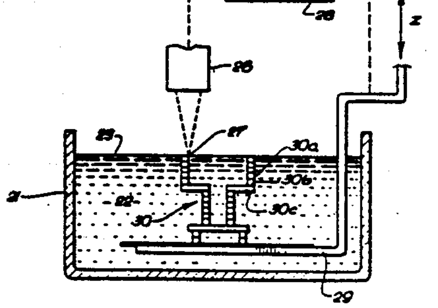
H. Kodama, "A Scheme for Three-Dimensional Display by Automatic Fabrication of Three-Dimensional Model," IECE, vol. J64–C, No. 4, Apr. 1981, pp. 237–241.

Request for Reexamination of U.S. Patent No. 4,575,330 filed Sep. 27, 1988, by E. I. Du Pont de Nemours & Company.

*Primary Examiner*—Richard L. Chiesa

[57] **ABSTRACT**

A system for generating three-dimensional objects by creating a cross-sectional pattern of the object to be formed at a selected surface of a fluid medium capable of altering its physical state in response to appropriate synergistic stimulation by impinging radiation, particle bombardment or chemical reaction, successive adjacent laminae, representing corresponding successive adjacent cross-sections of the object, being automatically formed and integrated together to provide a step-wise laminar buildup of the desired object, whereby a three-dimensional object is formed and drawn from a substantially planar surface of the fluid medium during the forming process.



## REEXAMINATION CERTIFICATE ISSUED UNDER 35 U.S.C. 307

THE PATENT IS HEREBY AMENDED AS INDICATED BELOW.

**Matter enclosed in heavy brackets [ ] appeared in the patent, but has been deleted and is no longer a part of the patent; matter printed in italics indicates additions made to the patent.**

AS A RESULT OF REEXAMINATION, IT HAS BEEN DETERMINED THAT:

Claims 1, 2 and 27–29 are determined to be patentable as amended.

Claims 3–26 and 30–47, dependent on an amended claim, are determined to be patentable.

1. A system for producing a three-dimensional object from a fluid medium capable of solidification when subjected to prescribed synergistic stimulation, said system comprising:

*a container holding said fluid medium, said medium being sufficiently absorptive of solidifying radiation to enable formation of an adequately cohesive layer of structure capable of being partially unsupported by any other layer of structure during formation;*

means for drawing upon and forming successive cross-sectional laminae of said object at a two-dimensional interface [and] *of said fluid medium defining a designated working surface, said laminae including a first cross-sectional layer of structure at said working surface;*

*means for automatically recoating over the entire said first cross-sectional layer of structure with a body of fluid and decreasing a substantial portion of said body of fluid in thickness from a fluid layer of excess fluid thickness to a successive fluid layer of less thickness in preparation for formation of a second cross-sectional layer of structure adhered to said first cross-sectional layer; and*

means for moving said [cross-sections] *cross-sectional layers* as they are formed and building up said object *from a plurality of successively adhered layers of structure* in step wise fashion, whereby a three-dimensional object is extracted from a substantially two-dimensional surface.

2. An improved system for producing a three-dimensional object from a fluid medium capable of solidification when subjected to prescribed synergistic stimulation, said system comprising:

a body of fluid medium capable of transforming its physical state in response to synergistic stimulation, *said fluid medium being sufficiently absorptive of said synergistic stimulation to enable formation of an adequately cohesive lamina of structure capable of being partially unsupported by any other lamina during formation;*

object support means immersed within said fluid medium for supporting a three-dimensional object to be formed;

translational means for selectively moving said object support means progressively away from a designated *working* surface of said fluid medium; [and]

reaction means capable of altering the physical state of said fluid medium and operating in a prescribed pattern upon said designated *working* surface of said fluid medium to provide a thin, solid, *first cross-sectional* lamina *of structure* at said *working* surface representing a corresponding cross-sectional lamina of said three-dimensional object to be formed[,]*; and*

*means for automatically recoating over the entire said first cross-sectional lamina of structure with a body of fluid and decreasing a substantial portion of said body of fluid in thickness from a fluid layer of excess fluid thickness to a successive fluid layer of less thickness in preparation for formation of a second cross-sectional lamina of structure adhered to said first cross-sectional lamina of structure,* whereby [successive] *a plurality of successively* adjacent *adhered* laminae are provided to form said three-dimensional object on said object support means as said translational means moves said support means away from said designated *working* surface.

27. A system for directly producing a three-dimensional object as it is designed by a computer, comprising:

deriving graphic image output from said computer, said graphic image *output* defining successive adjacent cross-sections of the three-dimensional object designed by said computer;

*a container holding a fluid medium capable of solidification, said fluid medium being sufficiently absorptive of solidifying radiation to enable formation of an adequately cohesive layer of structure capable of being partially unsupported by any other layer of structure during formation;*

means for drawing upon and forming successive *layer* cross-sections, corresponding to said computer designed cross-sections of said object, at a two-dimensional interface [and] *of said fluid medium defining a working surface, said layer cross-sections including a first cross-sectional layer of structure at said working surface;*

*means for automatically recoating over the entire said first cross-sectional layer of structure with a body of fluid and decreasing a substantial portion of said body of fluid in thickness from a fluid layer of excess fluid thickness to a successive fluid layer of less thickness in preparation for formation of a second cross-sectional layer of structure adhered to said first cross-sectional layer; and*

means for moving said [cross-sections] *cross-sectional layers* as they are formed and building up said object *from a plurality of successively adhered layers* in a stepwise fashion, whereby the three-dimensional object designed by said computer is automatically extracted from a substantially two-dimensional surface.

28. An improved system for producing a three-dimensional object from a fluid medium capable of altering its physical state when subjected to prescribed radiation, said system comprising:

a body of fluid medium capable of altering its physical state, *said fluid medium defining a designated working surface and being sufficiently absorptive of solidifying radiation to enable formation of an adequately cohesive layer of structure capable of being partially unsupported by any other layer of structure during formation;*

means for forming said three-dimensional object from said fluid medium by irradiating [a] *said* designated *working* surface of said *fluid* medium to provide integrated, successive surface laminae at said *working surface*, said laminae together defining said three-dimensional object [.], *said laminae including a first cross-sectional layer of structure at said working surface; and*

*means for automatically recoating over the entire said first cross-sectional layer of structure with a body of fluid and decreasing a substantial portion of said body of fluid in thickness from a fluid layer of excess fluid thickness to a successive fluid layer of less thickness in preparation for formation of a second cross-sectional layer of structure adhered to said first cross-sectional layer, whereby a plurality of successively adhered layers of structure form said three-dimensional object.*

29. An improved system for producing a three-dimensional object from a fluid medium, said system comprising:

a body of fluid medium capable of altering its physical state in response to prescribed radiation, *said fluid medium being sufficiently absorptive of said prescribed radiation to enable formation of an adequately cohesive lamina of structure capable of being partially unsupported by any other lamina of structure during formation;*

a radiation souce for impinging said prescribed radiation in a selected pattern upon a designated *working* surface of said fluid medium to provide only at said *working* surface a thin, solid, *first cross-sectional* lamina *of structure* representing a cross-sectional lamina of a three-dimensional object to be formed; [and]

*means for automatically recoating over the entire said first cross-sectional lamina of structure with a body of fluid and decreasing a substantial portion of said body of fluid in thickness from a fluid layer of excess fluid thickness to a successive fluid layer of less thickness in preparation for formation of a second cross-sectional lamina of structure adhered to said first cross-sectional lamina of structure; and*

means for combining [successive] *a plurality of successively adhered* adjacent laminae *of structure* to form said three-dimensional object from said fluid medium.

* * * * *

# Bibliography

[1] Armen Aghajanyan et al. "Muppet: Massive Multi-task Representations with Pre-Finetuning". In: *CoRR* abs/2101.11038 (2021). arXiv: 2101.11038. URL: https://arxiv.org/abs/2101.11038.

[2] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. "Data-Driven Sentence Simplification: Survey and Benchmark". In: *Computational Linguistics* 46.1 (Mar. 2020), pp. 135–187. ISSN: 0891-2017. DOI: 10.1162/coli_a_00370. eprint: https://direct.mit.edu/coli/article-pdf/46/1/135/1847760/coli_a_00370.pdf. URL: https://doi.org/10.1162/coli_a_00370.

[3] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. "The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification". In: *Computational Linguistics* 47.4 (Dec. 2021), pp. 861–889. ISSN: 0891-2017. DOI: 10.1162/coli_a_00418. eprint: https://direct.mit.edu/coli/article-pdf/47/4/861/1979827/coli_a_00418.pdf. URL: https://doi.org/10.1162/coli_a_00418.

[4] Fernando Alva-Manchego et al. "ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4668–4679. DOI: 10.18653/v1/2020.acl-main.424. URL: https://aclanthology.org/2020.acl-main.424.

[5] Fernando Alva-Manchego et al. "EASSE: Easier Automatic Sentence Simplification Evaluation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 49–54. DOI: 10.18653/v1/D19-3009. URL: https://aclanthology.org/D19-3009.

[6] Fernando Alva-Manchego et al. "Learning How to Simplify From Explicit Labeling of Complex-Simplified Text Pairs". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 295–305. URL: https://aclanthology.org/I17-1030.

[7] Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. "Exploring Transformer Text Generation for Medical Dataset Augmentation". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4699–4708. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.578.

[8] Linda Andersson, Mihai Lupu, and Allan Hanbury. "Domain Adaptation of General Natural Language Processing Tools for a Patent Claim Visualization System". In: *Multidisciplinary Information Retrieval*. Ed. by Mihai Lupu, Evangelos

Kanoulas, and Fernando Loizides. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 70–82. ISBN: 978-3-642-41057-4.

[9] Evlampios Apostolidis et al. "Summarizing Videos Using Concentrated Attention and Considering the Uniqueness and Diversity of the Video Frames". In: *Proceedings of the 2022 International Conference on Multimedia Retrieval*. ICMR '22. Newark, NJ, USA: Association for Computing Machinery, 2022, 407–415. ISBN: 9781450392389. DOI: 10.1145/3512527.3531404. URL: https://doi.org/10.1145/3512527.3531404.

[10] Mikel Artetxe and Holger Schwenk. "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 597–610. DOI: 10.1162/tacl_a_00288. URL: https://aclanthology.org/Q19-1038.

[11] Geert Asche. ""80% of technical information found only in patents"– Is there proof of this?" In: *World Patent Information* 48 (2017), pp. 16–28.

[12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1409.0473.

[13] Michele Banko, Vibhu O. Mittal, and Michael J. Witbrock. "Headline Generation Based on Statistical Translation". In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. ACL '00. Hong Kong: Association for Computational Linguistics, 2000, 318–325. DOI: 10.3115/1075218.1075259. URL: https://doi.org/10.3115/1075218.1075259.

[14] Eduard Barbu et al. "Language Technologies applied to Document Simplification for Helping Autistic People". In: *Expert Systems with Applications* 42 (July 2015), 5076–5086. DOI: 10.1016/j.eswa.2015.02.044.

[15] Leonard E. Baum and Ted Petrie. "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". In: *The Annals of Mathematical Statistics* 37.6 (1966), pp. 1554–1563. ISSN: 00034851. URL: http://www.jstor.org/stable/2238772 (visited on 09/23/2022).

[16] Iz Beltagy, Matthew E. Peters, and Arman Cohan. "Longformer: The Long-Document Transformer". In: *ArXiv* abs/2004.05150 (2020).

[17] Christian Bentz et al. "A Comparison Between Morphological Complexity Measures: Typological Data vs. Language Corpora". In: *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 142–153. URL: https://aclanthology.org/W16-4117.

[18] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[19] Piotr Bojanowski et al. "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146. DOI: 10.1162/tacl_a_00051. URL: https://aclanthology.org/Q17-1010.

[20] N. Bouayad-Agha et al. "Simplification of Patent Claim Sentences for their Paraphrasing and Summarization". In: *FLAIRS - PROCEEDINGS, International Florida Artificial Intelligence Research Society Conference, 22nd, International Florida Artificial Intelligence Research Society Conference*. Aaai Press; 2009, pp. 302–303. ISBN:

9781577354192, 1577354192. URL: https://www.tib.eu/de/suchen/id/BLCP%3ACN073481348.

[21] Nadjet Bouayad-Agha et al. "Improving the comprehension of legal documentation: The case of patent claims". In: *Proceedings of the International Conference on Artificial Intelligence and Law*. Jan. 2009, pp. 78–87. DOI: 10.1145/1568234.1568244.

[22] Fabienne Braune and Alexander Fraser. "Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora". In: *Coling 2010: Posters*. Beijing, China: Coling 2010 Organizing Committee, Aug. 2010, pp. 81–89. URL: https://aclanthology.org/C10-2010.

[23] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle et al. 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[24] Sören Brügmann et al. "Towards content-oriented patent document processing: Intelligent patent analysis and summarization". In: *World Patent Information* 40 (2015), pp. 30 –42. ISSN: 0172-2190. DOI: https://doi.org/10.1016/j.wpi.2014.10.003. URL: http://www.sciencedirect.com/science/article/pii/S0172219014001410.

[25] Alicia Burga et al. "The challenge of syntactic dependency parsing aNatural Language Processingtion for the patent domain". In: *ESSLLI-13 workshop on extrinsic parse improvement*. 2013.

[26] Ziqiang Cao et al. "Faithful to the Original: Fact Aware Neural Abstractive Summarization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018. URL: https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16121.

[27] Silvia Casola, Ivano Lauriola, and Alberto Lavelli. "Pre-trained transformers: an empirical comparison". In: *Machine Learning with Applications* 9 (2022), p. 100334. ISSN: 2666-8270. DOI: https://doi.org/10.1016/j.mlwa.2022.100334. URL: https://www.sciencedirect.com/science/article/pii/S2666827022000445.

[28] Silvia Casola and Alberto Lavelli. "FBK@SMM4H2020: RoBERTa for Detecting Medications on Twitter". In: *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. Barcelona, Spain (Online): Association for Computational Linguistics, Dec. 2020, pp. 101–103. URL: https://aclanthology.org/2020.smm4h-1.15.

[29] Silvia Casola and Alberto Lavelli. "Summarization, simplification, and generation: The case of patents". In: *Expert Systems with Applications* 205 (2022), p. 117627. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.117627. URL: https://www.sciencedirect.com/science/article/pii/S0957417422009356.

[30] Silvia Casola and Alberto Lavelli. "WITS: Wikipedia for Italian Text Summarization". In: *Italian Conference on Computational Linguistics*. 2021.

[31] Silvia Casola, Alberto Lavelli, and Horacio Saggion. "What's in a (dataset's) name? The case of BigPatent". In: *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 399–404. URL: https://aclanthology.org/2022.gem-1.34.

[32] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. "Evaluation of Text Generation: A Survey". In: *ArXiv* abs/2006.14799 (2020).

[33] Daniel Cer et al. "Universal Sentence Encoder for English". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 169–174. DOI: 10.18653/v1/D18-2029. URL: https://www.aclweb.org/anthology/D18-2029.

[34] Yen-Chun Chen and Mohit Bansal. "Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 675–686. DOI: 10.18653/v1/P18-1063. URL: https://www.aclweb.org/anthology/P18-1063.

[35] Yiran Chen et al. "CDEvalSumm: An Empirical Study of Cross-Dataset Evaluation for Neural Summarization Systems". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3679–3691. DOI: 10.18653/v1/2020.findings-emnlp.329. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.329.

[36] Sumit Chopra, Michael Auli, and Alexander M. Rush. "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 93–98. DOI: 10.18653/v1/N16-1012. URL: https://aclanthology.org/N16-1012.

[37] Joan Codina-Filbà et al. "Using genre-specific features for patent summaries". In: *Information Processing & Management* 53.1 (2017), pp. 151 –174. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2016.07.002. URL: http://www.sciencedirect.com/science/article/pii/S0306457316302825.

[38] Arman Cohan et al. "A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 615–621. DOI: 10.18653/v1/N18-2097. URL: https://www.aclweb.org/anthology/N18-2097.

[39] Will Coster and David Kauchak. "Learning to Simplify Sentences Using Wikipedia". In: *Proceedings of the Workshop on Monolingual Text-To-Text Generation*. Portland, Oregon: Association for Computational Linguistics, June 2011, pp. 1–9. URL: https://aclanthology.org/W11-1601.

[40] William Coster and David Kauchak. "Simple English Wikipedia: A New Text Simplification Task". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 665–669. URL: https://aclanthology.org/P11-2117.

[41] Rumen Dangovski et al. "Rotational Unit of Memory: A Novel Representation Unit for RNNs with Scalable Applications". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 121–138. DOI: 10.1162/tacl_a_00258. URL: https://aclanthology.org/Q19-1008.

[42] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association

for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[43] Jay DeYoung et al. "MS^2: Multi-Document Summarization of Medical Studies". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7494–7513. DOI: 10.18653/v1/2021.emnlp-main.594. URL: https://aclanthology.org/2021.emnlp-main.594.

[44] Oluwajana Dokun and Erbug Celebi. "Single-document summarization using latent semantic analysis". In: *International Journal of Scientific Research in Information Systems and Engineering (IJSRISE)* 1.2 (2015), pp. 57–64.

[45] Mahmoud El-Haj et al. "The Financial Narrative Summarisation Shared Task (FNS 2022)". In: *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*. Marseille, France: European Language Resources Association, June 2022, pp. 43–52. URL: https://aclanthology.org/2022.fnp-1.6.

[46] Günes Erkan and Dragomir R. Radev. "LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization". In: *J. Artif. Int. Res.* 22.1 (Dec. 2004), 457–479. ISSN: 1076-9757.

[47] Richard J. Evans. "Comparing methods for the syntactic simplification of sentences in information extraction". In: *Literary and Linguistic Computing* 26.4 (Aug. 2011), pp. 371–388. ISSN: 0268-1145. DOI: 10.1093/llc/fqr034. eprint: https://academic.oup.com/dsh/article-pdf/26/4/371/6197103/fqr034.pdf. URL: https://doi.org/10.1093/llc/fqr034.

[48] A. R. Fabbri et al. "Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation". In: *arXiv preprint arXiv:2010.12836* (Nov. 2020).

[49] Alexander Fabbri et al. "Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1074–1084. DOI: 10.18653/v1/P19-1102. URL: https://aclanthology.org/P19-1102.

[50] R. Feldman. "Plain Language Patents". In: vol. 17. 2008, p. 289.

[51] Gabriela Ferraro, Hanna Suominen, and Jaume Nualart. "Segmentation of patent claims for improving their readability". In: *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 66–73. DOI: 10.3115/v1/W14-1208. URL: https://www.aclweb.org/anthology/W14-1208.

[52] R. Flesch. *How to Write Plain English: A Book for Lawyers and Consumers*. Harper & Row, 1979.

[53] Philip Gage. "A new algorithm for data compression". In: *The C Users Journal archive* 12 (1994), pp. 23–38.

[54] Alexios Gidiotis and Grigorios Tsoumakas. "A Divide-and-Conquer Approach to the Summarization of Long Documents". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 3029–3040. DOI: 10.1109/TASLP.2020.3037401.

[55] K. Girthana and S. Swamynathan. "Query Oriented Extractive-Abstractive Summarization System (QEASS)". In: *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. CoDS-COMAD '19. Kolkata,

India: Association for Computing Machinery, 2019, 301–305. ISBN: 9781450362078. DOI: 10.1145/3297001.3297046. URL: https://doi.org/10.1145/3297001.3297046.

[56] K. Girthana and S. Swamynathan. “Query-Oriented Patent Document Summarization System (QPSS)”. In: *Soft Computing: Theories and Applications*. Ed. by Millie Pant et al. Singapore: Springer Singapore, 2020, pp. 237–246. ISBN: 978-981-15-0751-9.

[57] K. Girthana and S. Swamynathan. “Semantic Query-Based Patent Summarization System (SQPSS)”. In: *Advances in Data Science*. Ed. by Leman Akoglu et al. Singapore: Springer Singapore, 2019, pp. 169–179. ISBN: 978-981-13-3582-2.

[58] Tomas Goldsack et al. “Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 10589–10604. URL: https://aclanthology.org/2022.emnlp-main.724.

[59] Yihong Gong and Xin Liu. “Generic text summarization using relevance measure and latent semantic analysis”. In: *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2001.

[60] Max Grusky, Mor Naaman, and Yoav Artzi. “Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 708–719. DOI: 10.18653/v1/N18-1065. URL: https://aclanthology.org/N18-1065.

[61] Shuhao Gu et al. “Token-level Adaptive Training for Neural Machine Translation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1035–1046. DOI: 10.18653/v1/2020.emnlp-main.76. URL: https://aclanthology.org/2020.emnlp-main.76.

[62] Juncai Guo et al. “Modeling Hierarchical Syntax Structure with Triplet Position for Source Code Summarization”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 486–500. DOI: 10.18653/v1/2022.acl-long.37. URL: https://aclanthology.org/2022.acl-long.37.

[63] Mandy Guo et al. “LongT5: Efficient Text-To-Text Transformer for Long Sequences”. In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 724–736. URL: https://aclanthology.org/2022.findings-naacl.55.

[64] Junxian He et al. “CTRLsum: Towards Generic Controllable Text Summarization”. In: *arXiv preprint arXiv:2012.04281* (2020).

[65] Junxian He et al. “CTRLsum: Towards Generic Controllable Text Summarization”. In: *ArXiv* abs/2012.04281 (2020).

[66] Michael Heilman and Noah A Smith. *Question generation via overgenerating transformations and ranking*. Tech. rep. Carnegie-Mellon Univ Pittsburgh pa language technologies insT, 2009.

[67] Matthew Honnibal and Mark Johnson. “An Improved Non-monotonic Transition System for Dependency Parsing”. In: *Proceedings of the 2015 Conference on Empirical*

*Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 1373–1378. URL: https://aclweb.org/anthology/D/D15/D15-1162.

[68] Eduard Hovy and Chin-Yew Lin. "Automated Text Summarization and the SUMMARIST System". In: *Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998*. TIPSTER '98. Baltimore, Maryland: Association for Computational Linguistics, 1998, 197–214. DOI: 10.3115/1119089.1119121. URL: https://doi.org/10.3115/1119089.1119121.

[69] David M. Howcroft et al. "Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions". In: *Proceedings of the 13th International Conference on Natural Language Generation*. Dublin, Ireland: Association for Computational Linguistics, Dec. 2020, pp. 169–182. URL: https://aclanthology.org/2020.inlg-1.23.

[70] Si Huang et al. "An Extraction-Abstraction Hybrid Approach for Long Document Summarization". In: *2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC)*. 2019, pp. 1–6. DOI: 10.1109/BESC48373.2019.8962979.

[71] Weijing Huang et al. "Generating Reasonable Legal Text through the Combination of Language Modeling and Question Answering". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. Ed. by Christian Bessiere. Main track. International Joint Conferences on Artificial Intelligence Organization, July 2020, pp. 3687–3693. DOI: 10.24963/ijcai.2020/510. URL: https://doi.org/10.24963/ijcai.2020/510.

[72] William Hwang et al. "Aligning Sentences from Standard Wikipedia to Simple Wikipedia". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, 2015, pp. 211–217. DOI: 10.3115/v1/N15-1022. URL: https://aclanthology.org/N15-1022.

[73] "Indexing by Latent Semantic Analysis". In: *Journal of the Association for Information Science and Technology* 41.6 (1990), pp. 391–407. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[74] John Irvine and Ben R Martin. *Foresight in science: Picking the winners*. 338.06/I72f. 1984.

[75] Chuleerat Jaruskulchai and Canasai Kruengkrai. "A Practical Text Summarizer by Paragraph Extraction for Thai". In: *Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages - Volume 11*. AsianIR '03. Sapporo, Japan: Association for Computational Linguistics, 2003, 9–16. DOI: 10.3115/1118935.1118937. URL: https://doi.org/10.3115/1118935.1118937.

[76] Hongyan Jing and Kathleen R. McKeown. "Cut and Paste Based Text Summarization". In: *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*. NAACL 2000. Seattle, Washington: Association for Computational Linguistics, 2000, 178–185.

[77] Hongyan Jing and Kathleen R. McKeown. "The Decomposition of Human-Written Summary Sentences". In: *Proceedings of the 22nd Annual International ACM R Conference on Research and Development in Information Retrieval*. simplificationR '99. Berkeley, California, USA: Association for Computing Machinery, 1999, 129–136. ISBN: 1581130961. DOI: 10.1145/312624.312666. URL: https://doi.org/10.1145/312624.312666.

[78] James M. Joyce. "Kullback-Leibler Divergence". In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 720–722. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_327. URL: https://doi.org/10.1007/978-3-642-04898-2_327.

[79] Tomoyuki Kajiwara and Mamoru Komachi. "Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 1147–1158. URL: https://aclanthology.org/C16-1109.

[80] Tomoyuki Kajiwara and Mamoru Komachi. "Text Simplification without Simplified Corpora". In: *Journal of Natural Language Processing* 25 (Mar. 2018), pp. 223–249. DOI: 10.5715/jnlp.25.223.

[81] Dongyeop Kang et al. "A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1647–1661. DOI: 10.18653/v1/N18-1149. URL: https://aclanthology.org/N18-1149.

[82] Jeongwoo Kang, Achille Souili, and Denis Cavallucci. "Text Simplification of Patent Documents". In: *Automated Invention for Smart Industries*. Ed. by Denis Cavallucci, Roland De Guio, and Sebastian Koziołek. Cham: Springer International Publishing, 2018, pp. 225–237. ISBN: 978-3-030-02456-7.

[83] Neel Kanwal and Giuseppe Rizzo. "Attention-based clinical note summarization". In: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. 2022, pp. 813–820.

[84] David Kauchak. "Improving Text Simplification Language Modeling Using Unsimplified Text Data". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 1537–1546. URL: https://aclanthology.org/P13-1151.

[85] Xiaopeng Ke et al. "Towards Practical and Efficient Long Video Summary". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 1770–1774. DOI: 10.1109/ICASSP43922.2022.9746911.

[86] Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. "Abstractive Summarization of Reddit Posts with Multi-level Memory Networks". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 2519–2531. DOI: 10.18653/v1/N19-1260. URL: https://aclanthology.org/N19-1260.

[87] J. Peter Kincaid et al. "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel". In: 1975.

[88] V. Klema and A. Laub. "The singular value decomposition: Its computation and some applications". In: *IEEE Transactions on Automatic Control* 25.2 (1980), pp. 164–176. DOI: 10.1109/TAC.1980.1102314.

[89] Kevin Knight and Daniel Marcu. "Summarization beyond sentence extraction: A probabilistic approach to sentence compression". In: *Artificial Intelligence* 139.1

(2002), pp. 91–107. ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(02)00222-9. URL: https://www.sciencedirect.com/science/article/pii/S0004370202002229.

[90] Huan Yee Koh et al. "An Empirical Survey on Long Document Summarization: Datasets, Models and Metrics". In: *ACM Comput. Surv.* (2022). Just Accepted. ISSN: 0360-0300. DOI: 10.1145/3545176. URL: https://doi.org/10.1145/3545176.

[91] Anastassia Kornilova and Vladimir Eidelman. "BillSum: A Corpus for Automatic Summarization of US Legislation". In: *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 48–56. DOI: 10.18653/v1/D19-5406. URL: https://aclanthology.org/D19-5406.

[92] Mahnaz Koupaee and William Yang Wang. "WikiHow: A Large Scale Text Summarization Dataset". In: *ArXiv* abs/1810.09305 (2018).

[93] Wojciech Kryscinski et al. "Evaluating the Factual Consistency of Abstractive Text Summarization". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 9332–9346. DOI: 10.18653/v1/2020.emnlp-main.750. URL: https://aclanthology.org/2020.emnlp-main.750.

[94] Moreno La Quatra and Luca Cagliero. "BART-IT: An Efficient Sequence-to-Sequence Model for Italian Text Summarization". In: *Future Internet* 15.1 (2023). ISSN: 1999-5903. DOI: 10.3390/fi15010015. URL: https://www.mdpi.com/1999-5903/15/1/15.

[95] Faisal Ladhak et al. "WikiLingua: A New Benchmark Dataset for Multilingual Abstractive Summarization". In: *Findings of EMNLP, 2020*. 2020.

[96] Nicola Landro et al. "Two New Datasets for Italian-Language Abstractive Text Summarization". In: *Information* 13.5 (2022). ISSN: 2078-2489. DOI: 10.3390/info13050228. URL: https://www.mdpi.com/2078-2489/13/5/228.

[97] Hugo Larochelle and Yoshua Bengio. "Classification Using Discriminative Restricted Boltzmann Machines". In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. Helsinki, Finland: Association for Computing Machinery, 2008, 536–543. ISBN: 9781605582054. DOI: 10.1145/1390156.1390224. URL: https://doi.org/10.1145/1390156.1390224.

[98] Jieh-Sheng Lee. "Controlling Patent Text Generation by Structural Metadata". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2020, 3241–3244. ISBN: 9781450368599. URL: https://doi.org/10.1145/3340531.3418503.

[99] Jieh-Sheng Lee and Jieh Hsiang. "Patent claim generation by fine-tuning OpenAI GPT-2". In: *World Patent Information* 62 (2020), p. 101983. ISSN: 0172-2190. DOI: https://doi.org/10.1016/j.wpi.2020.101983. URL: http://www.sciencedirect.com/science/article/pii/S0172219019300766.

[100] Jieh-Sheng Lee and Jieh Hsiang. "PatentTransformer-1.5: Measuring Patent Claim Generation by Span Relevancy". In: *New Frontiers in Artificial Intelligence*. Ed. by Maki Sakamoto et al. Cham: Springer International Publishing, 2020, pp. 20–33. ISBN: 978-3-030-58790-1.

[101] Jieh-Sheng Lee and Jieh Hsiang. *Prior Art Search and Reranking for Generated Patent Text*. 2020. arXiv: 2009.09132 [cs.CL].

[102] Mike Lewis et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7871–7880. DOI: `10.18653/v1/2020.acl-main.703`. URL: `https://aclanthology.org/2020.acl-main.703`.

[103] Quentin Lhoest et al. “Datasets: A Community Library for Natural Language Processing”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 175–184. arXiv: `2109.02846 [cs.CL]`. URL: `https://aclanthology.org/2021.emnlp-demo.21`.

[104] Rensis Likert. *A technique for the measurement of attitudes / by Rensis Likert.* eng. Archives of psychology ; no. 140. New York: [s.n.], 1985 - 1932.

[105] Chin-Yew Lin. “ROUGE: a Package for Automatic Evaluation of Summaries”. In: *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*. 2004, pp. 74–81.

[106] Chin-Yew Lin and Franz Josef Och. “ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation”. In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland: COLING, 2004, pp. 501–507. URL: `https://aclanthology.org/C04-1072`.

[107] Harold A Linstone, Murray Turoff, et al. *The Delphi method*. Addison-Wesley Reading, MA, 1975.

[108] Peter J. Liu et al. “Generating Wikipedia by Summarizing Long Sequences”. In: *International Conference on Learning Representations*. 2018. URL: `https://openreview.net/forum?id=Hyg0vbWC-`.

[109] Yang Liu and Mirella Lapata. “Text Summarization with Pretrained Encoders”. In: *ArXiv* abs/1908.08345 (2019).

[110] Yinhan Liu et al. “Multilingual Denoising Pre-training for Neural Machine Translation”. In: *Trans. Assoc. Comput. Linguistics* 8 (2020), pp. 726–742. URL: `https://transacl.org/ojs/index.php/tacl/article/view/2107`.

[111] Elena Lloret, Laura Plaza, and Ahmet Aker. “The Challenging Task of Summary Evaluation: An Overview”. In: *Lang. Resour. Eval.* 52.1 (Mar. 2018), 101–148. ISSN: 1574-020X. DOI: `10.1007/s10579-017-9399-2`. URL: `https://doi.org/10.1007/s10579-017-9399-2`.

[112] Annie Louis and Ani Nenkova. “Automatically Evaluating Content Selection in Summarization without Human Models”. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, Aug. 2009, pp. 306–314. URL: `https://aclanthology.org/D09-1032`.

[113] Samuel Louvan, Silvia Casola, and Bernardo Magnini. “Investigating Continued Pretraining for Zero-Shot Cross-Lingual Spoken Language Understanding”. In: *Italian Conference on Computational Linguistics*. 2021.

[114] Xinyu Lu et al. “An Unsupervised Method for Building Sentence Simplification Corpora in Multiple Languages”. In: *EMNLP*. 2021.

[115] H. P. Luhn. “The Automatic Creation of Literature Abstracts”. In: *IBM Journal of Research and Development* 2.2 (1958), pp. 159–165. DOI: `10.1147/rd.22.0159`.

[116] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. "Readability Controllable Biomedical Document Summarization". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4667–4680. URL: `https://aclanthology.org/2022.findings-emnlp.343`.

[117] Potsawee Manakul and Mark Gales. "Long-Span Summarization via Local Attention and Content Selection". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 6026–6041. DOI: `10.18653/v1/2021.acl-long.470`. URL: `https://aclanthology.org/2021.acl-long.470`.

[118] Inderjeet Mani et al. "The TIPSTER SUMMAC Text Summarization Evaluation". In: *Ninth Conference of the European Chapter of the Association for Computational Linguistics*. Bergen, Norway: Association for Computational Linguistics, June 1999, pp. 77–85. URL: `https://aclanthology.org/E99-1011`.

[119] Ben R. Martin. "Foresight in science and technology". In: *Technology Analysis & Strategic Management* 7.2 (1995), pp. 139–168. DOI: `10.1080/09537329508524202`.

[120] Louis Martin et al. "Controllable Sentence Simplification". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4689–4698. ISBN: 979-10-95546-34-4. URL: `https://aclanthology.org/2020.lrec-1.577`.

[121] Louis Martin et al. "MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 1651–1664. URL: `https://aclanthology.org/2022.lrec-1.176`.

[122] C. Mastronardo and F. Tamburini. "Enhancing a Text Summarization System with ELMo". In: *CLiC-it*. 2019.

[123] Lorenzo De Mattei et al. "GePpeTto Carves Italian into a Language Model". In: *CLiC-it*. 2020.

[124] Rada Mihalcea and Paul Tarau. "TextRank: Bringing Order into Text". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 404–411. URL: `https://www.aclweb.org/anthology/W04-3252`.

[125] Ian Miles. "The development of technology foresight: A review". In: *Technological Forecasting and Social Change* 77.9 (2010). Strategic Foresight, pp. 1448–1456. ISSN: 0040-1625. DOI: `https://doi.org/10.1016/j.techfore.2010.07.016`. URL: `https://www.sciencedirect.com/science/article/pii/S0040162510001794`.

[126] Simon Mille and Leo Wanner. "Making Text Resources Accessible to the Reader: the Case of Patent Claims". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), May 2008. URL: `http://www.lrec-conf.org/proceedings/lrec2008/pdf/352_paper.pdf`.

[127] Simon Mille and Leo Wanner. "Multilingual summarization in practice: the case of patent claims". In: *Proceedings of the 12th Annual conference of the European Association for Machine Translation*. Hamburg, Germany: European Association for Machine Translation, 2008, pp. 120–129. URL: `https://www.aclweb.org/anthology/2008.eamt-1.18`.

[128] Mandar Mitra, Amit Singhal, and Chris Buckley. "Automatic Text Summarization by Paragraph Extraction". In: *Workshop On Intelligent Scalable Text Summarization*. 1997.

[129] Ramesh Nallapati et al. "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond". In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 280–290. DOI: 10.18653/v1/K16-1028. URL: https://aclanthology.org/K16-1028.

[130] Ramesh Nallapati et al. "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond". In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 280–290. DOI: 10.18653/v1/K16-1028. URL: https://aclanthology.org/K16-1028.

[131] Feng Nan et al. "Entity-level Factual Consistency of Abstractive Text Summarization". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 2727–2733. DOI: 10.18653/v1/2021.eacl-main.235. URL: https://aclanthology.org/2021.eacl-main.235.

[132] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 1797–1807. DOI: 10.18653/v1/D18-1206. URL: https://aclanthology.org/D18-1206.

[133] Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. "The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation". In: *ACM Trans. Speech Lang. Process.* 4.2 (2007), 4–es. ISSN: 1550-4875. DOI: 10.1145/1233912.1233913. URL: https://doi.org/10.1145/1233912.1233913.

[134] Ani Nenkova and Lucy Vanderwende. "The impact of frequency on summarization". In: *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005* 101 (2005).

[135] Sergiu Nisioi et al. "Exploring Neural Text Simplification Models". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 85–91. DOI: 10.18653/v1/P17-2014. URL: https://aclanthology.org/P17-2014.

[136] Ishmael Obonyo, Silvia Casola, and Horacio Saggion. "Exploring the limits of a base BART for multi-document summarization in the medical domain". In: *Proceedings of the Third Workshop on Scholarly Document Processing*. Gyeongju, Republic of Korea: Association for Computational Linguistics, Oct. 2022, pp. 193–198. URL: https://aclanthology.org/2022.sdp-1.23.

[137] Masayuki Okamoto, Zifei Shan, and Ryohei Orihara. "Applying Information Extraction for Patent Structure Analysis". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '17. Shinjuku, Tokyo, Japan: Association for Computing Machinery, 2017, 989–992. ISBN: 9781450350228. DOI: 10.1145/3077136.3080698. URL: https://doi.org/10.1145/3077136.3080698.

[138] Gustavo Paetzold and Lucia Specia. "Unsupervised lexical simplification for non-native speakers". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. 1. 2016.

[139] Lawrence Page et al. "The PageRank Citation Ranking : Bringing Order to the Web". In: *WWW 1999*. 1999.

[140] Kishore Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.

[141] Nikhil Pattisapu et al. "Leveraging Social Media for Medical Text Simplification". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '20. Virtual Event, China: Association for Computing Machinery, 2020, 851–860. ISBN: 9781450380164. DOI: 10.1145/3397271.3401105. URL: https://doi.org/10.1145/3397271.3401105.

[142] Eyal Peer et al. "Data quality of platforms and panels for online behavioral research". In: *Behavior Research Methods* 54.4 (2022), pp. 1643–1662. ISSN: 1554-3528. DOI: 10.3758/s13428-021-01694-3. URL: https://doi.org/10.3758/s13428-021-01694-3.

[143] Carlo Pietrobelli and Fernanda Puppato. "Technology foresight and industrial strategy". In: *Technological Forecasting and Social Change* 110 (2016), pp. 117–125. ISSN: 0040-1625. DOI: https://doi.org/10.1016/j.techfore.2015.10.021. URL: https://www.sciencedirect.com/science/article/pii/S0040162515003169.

[144] Jonathan Pilault et al. "On Extractive and Abstractive Neural Document Summarization with Transformer Language Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 9308–9319. DOI: 10.18653/v1/2020.emnlp-main.748. URL: https://www.aclweb.org/anthology/2020.emnlp-main.748.

[145] Basel Qenam et al. "Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: translation and evaluation". In: *Journal of medical Internet research* 19.12 (2017), e417.

[146] A. Radford et al. "Language Models are Unsupervised Multitask Learners". In: 2019.

[147] Alec Radford et al. *Improving language understanding by generative pre-training*. 2018.

[148] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[149] Peter Gordon Roetzel. "Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development". In: *Business Research* 12.2 (2019), pp. 479–522. ISSN: 2198-2627. DOI: 10.1007/s40685-018-0069-z. URL: https://doi.org/10.1007/s40685-018-0069-z.

[150] Alexander M. Rush, Sumit Chopra, and Jason Weston. "A Neural Attention Model for Abstractive Sentence Summarization". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 379–389. DOI: 10.18653/v1/D15-1044. URL: https://aclanthology.org/D15-1044.

[151] Horacio Saggion. "Learning Predicate Insertion Rules for Document Abstracting". In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 301–312. ISBN: 978-3-642-19437-5.

[152] Tetsuya Sakai and Karen Sparck-Jones. "Generic Summaries for Indexing in Information Retrieval". In: *Proceedings of the 24th Annual International ACM simplificationR Conference on Research and Development in Information Retrieval*. SIGIR '01. New Orleans, Louisiana, USA: Association for Computing Machinery, 2001, 190–198. ISBN: 1581133316. DOI: 10.1145/383952.383987. URL: https://doi.org/10.1145/383952.383987.

[153] Evan Sandhaus. "The New York Times annotated corpus". In: *Linguistic Data Consortium, Philadelphia* 6.12 (2008), e26752.

[154] Gabriele Sarti and Malvina Nissim. *IT5: Large-scale Text-to-text Pretraining for Italian Language Understanding and Generation*. 2022. arXiv: 2203.03759 [cs.CL].

[155] Carolina Scarton, Gustavo Paetzold, and Lucia Specia. "SimPA: A Sentence-Level Simplification Corpus for the Public Administration Domain". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. URL: https://aclanthology.org/L18-1685.

[156] Thomas Scialom et al. "MLSUM: The Multilingual Summarization Corpus". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 8051–8067. DOI: 10.18653/v1/2020.emnlp-main.647. URL: https://aclanthology.org/2020.emnlp-main.647.

[157] Abigail See, Peter J. Liu, and Christopher D. Manning. "Get To The Point: Summarization with Pointer-Generator Networks". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1073–1083. DOI: 10.18653/v1/P17-1099. URL: https://www.aclweb.org/anthology/P17-1099.

[158] Eva Sharma, Chen Li, and Lu Wang. "BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2204–2213. DOI: 10.18653/v1/P19-1212. URL: https://www.aclweb.org/anthology/P19-1212.

[159] Kim Cheng Sheang and Horacio Saggion. "Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer". In: *Proceedings of the 14th International Conference on Natural Language Generation*. Aberdeen, Scotland, UK: Association for Computational Linguistics, Aug. 2021, pp. 341–352. URL: https://aclanthology.org/2021.inlg-1.38.

[160] Svetlana Sheremetyeva. "Automatic Text Simplification For Handling Intellectual Property (The Case of Multiple Patent Claims)". In: *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, Aug. 2014, pp. 41–52. DOI: 10.3115/v1/W14-5605. URL: https://www.aclweb.org/anthology/W14-5605.

[161] Svetlana Sheremetyeva. "Natural Language Analysis of Patent Claims". In: *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing - Volume 20*. PATENT

'03. Sapporo, Japan: Association for Computational Linguistics, 2003, 66–73. DOI: 10.3115/1119303.1119311. URL: https://doi.org/10.3115/1119303.1119311.

[162] Ensheng Shi et al. "CAST: Enhancing Code Summarization with Hierarchical Splitting and Reconstruction of Abstract Syntax Trees". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4053–4062. DOI: 10.18653/v1/2021.emnlp-main.332. URL: https://aclanthology.org/2021.emnlp-main.332.

[163] Akihiro Shinmori and M. Okumura. "Aligning Patent Claims with Detailed Descriptions for Readability". In: *NII Testbeds and Community for Information Access Research*. 2004.

[164] Akihiro Shinmori et al. "Patent Claim Processing for Readability: Structure Analysis and Term Explanation". In: *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing - Volume 20*. PATENT '03. Sapporo, Japan: Association for Computational Linguistics, 2003, 56–65. DOI: 10.3115/1119303.1119310. URL: https://doi.org/10.3115/1119303.1119310.

[165] Akihiro Shinmori et al. "Rhetorical Structure Analysis of Japanese Patent Claims using Cue Phrases." In: *NII Testbeds and Community for Information Access Research*. 2002.

[166] Kai Shu et al. "Fact-Enhanced Synthetic News Generation". English. In: Conference on Artificial Intelligence, AAAI. AAAI press, 2020.

[167] Advaith Siddharthan. "A survey of research on text simplification". In: *International Journal of Applied Linguistics* 165.2 (2014), pp. 259–298. URL: http://oro.open.ac.uk/58886/.

[168] Sara Silveira and António Horta Branco. "Enhancing Multi-document Summaries with Sentence Simplification". In: 2012.

[169] Amanpreet Singh and Niranjan Balasubramanian. "Open4Business(O4B): An Open Access Dataset for Summarizing Business Documents". In: *ArXiv* abs/2011.07636 (2020).

[170] Matthew Snover et al. "A Study of Translation Edit Rate with Targeted Human Annotation". In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, 2006, pp. 223–231. URL: https://aclanthology.org/2006.amta-papers.25.

[171] Cinthia M. de Souza et al. "Using Summarization Techniques on Patent Database Through Computational Intelligence". In: *Progress in Artificial Intelligence*. Ed. by Paulo Moura Oliveira, Paulo Novais, and Luís Paulo Reis. Springer International Publishing, 2019, pp. 508–519. ISBN: 978-3-030-30244-3.

[172] Cinthia Mikaela de Souza, Magali R. G. Meireles, and P. Almeida. "A comparative study of abstractive and extractive summarization techniques to label subgroups on patent dataset". In: *Scientometrics* 126 (2021), pp. 135–156.

[173] Lucia Specia. "Translating from Complex to Simplified Sentences". In: *Computational Processing of the Portuguese Language*. Ed. by Thiago Alexandre Salgueiro Pardo et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 30–39. ISBN: 978-3-642-12320-7.

[174] Sanja Štajner et al. "Sentence Alignment Methods for Improving Text Simplification Systems". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for

Computational Linguistics, July 2017, pp. 97–102. DOI: 10.18653/v1/P17-2016. URL: https://aclanthology.org/P17-2016.

[175] Josef Steinberger and Karel Jezek. "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation". In: 2004.

[176] Elior Sulem, Omri Abend, and Ari Rappoport. "BLEU is Not Suitable for the Evaluation of Text Simplification". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 738–744. DOI: 10.18653/v1/D18-1081. URL: https://aclanthology.org/D18-1081.

[177] Elior Sulem, Omri Abend, and Ari Rappoport. "Semantic Structural Evaluation for Text Simplification". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 685–696. DOI: 10.18653/v1/N18-1063. URL: https://aclanthology.org/N18-1063.

[178] Hong Sun and Ming Zhou. "Joint Learning of a Dual SMT System for Paraphrase Generation". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 38–42. URL: https://aclanthology.org/P12-2008.

[179] Renliang Sun, Hanqi Jin, and Xiaojun Wan. "Document-Level Text Simplification: Dataset, Criteria and Baseline". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7997–8013. DOI: 10.18653/v1/2021.emnlp-main.630. URL: https://aclanthology.org/2021.emnlp-main.630.

[180] Hanna Suominen et al. "User Study for Measuring Linguistic Complexity and Its Reduction by Technology on a Patent Website". In: *Conference: 34 International Conference on Machine Learning*. ICML'17. Sydney, Australia, Feb. 2018.

[181] Sai Surya et al. "Unsupervised Neural Text Simplification". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2058–2068. DOI: 10.18653/v1/P19-1198. URL: https://aclanthology.org/P19-1198.

[182] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. "Sequence to Sequence Learning with Neural Networks". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, 3104–3112.

[183] Mirac Suzgun et al. *The Harvard USPTO Patent Dataset: A Large-Scale, Well-Structured, and Multi-Purpose Corpus of Patent Applications*. 2022. URL: https://arxiv.org/abs/2207.04043.

[184] A. Trappey, C. Trappey, and B. H. Kao. "Automated Patent Document Summarization for R&D Intellectual Property Management". In: *2006 10th International Conference on Computer Supported Cooperative Work in Design* (2006), pp. 1–6.

[185] Amy Trappey and Charles Trappey. "An R&D knowledge management method for patent document". In: *Industrial Management and Data Systems* 108 (Mar. 2008), pp. 245–257. DOI: 10.1108/02635570810847608.

[186] Amy Trappey, Charles Trappey, and Chun-Yi Wu. "Automatic patent document summarization for collaborative knowledge systems and services". In: *Journal of*

*Systems Science and Systems Engineering* 18 (Mar. 2009), pp. 71–94. DOI: 10.1007/s11518-009-5100-7.

[187] Amy J. C. Trappey, Charles V. Trappey, and Chun-Yi Wu. "A Semantic Based Approach for Automatic Patent Document Summarization". In: *Collaborative Product and Service Life Cycle Management for a Sustainable World*. Ed. by Richard Curran, Shuo-Yan Chou, and Amy Trappey. London: Springer London, 2008, pp. 485–494. ISBN: 978-1-84800-972-1.

[188] Amy J.C. Trappey et al. "Intelligent compilation of patent summaries using Machine Learning and Natural Language Processing techniques". In: *Advanced Engineering Informatics* 43 (2020), p. 101027. ISSN: 1474-0346. DOI: https://doi.org/10.1016/j.aei.2019.101027. URL: http://www.sciencedirect.com/science/article/pii/S1474034619306007.

[189] Yuen-Hsien Tseng, Chi-Jen Lin, and Yu-I Lin. "Text mining techniques for patent analysis". In: *Information Processing & Management* 43.5 (2007). Patent Processing, pp. 1216 –1247. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2006.11.011. URL: http://www.sciencedirect.com/science/article/pii/S0306457306002020.

[190] Yuen-Hsien Tseng et al. "Patent surrogate extraction and evaluation in the context of patent mapping". In: *Journal of Information Science* 33.6 (2007), pp. 718–736. DOI: 10.1177/0165551507077406. eprint: https://doi.org/10.1177/0165551507077406. URL: https://doi.org/10.1177/0165551507077406.

[191] Masao Utiyama and Hitoshi Isahara. "A Japanese-English patent parallel corpus". In: *MTSUMMIT*. 2007.

[192] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[193] S Verberne et al. "Quantifying the Challenges in Parsing Patent Claims". In: *Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval at ECIR 2010*. [Sl: sn]. 2010, pp. 14–21.

[194] David Vickrey and Daphne Koller. "Sentence Simplification for Semantic Role Labeling". In: *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, June 2008, pp. 344–352. URL: https://aclanthology.org/P08-1040.

[195] Alex Wang, Kyunghyun Cho, and Mike Lewis. "Asking and Answering Questions to Evaluate the Factual Consistency of Summaries". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5008–5020. DOI: 10.18653/v1/2020.acl-main.450. URL: https://www.aclweb.org/anthology/2020.acl-main.450.

[196] Jun Wang. "ESSumm: Extractive Speech Summarization from Untranscribed Meeting". In: *Interspeech 2022* (2022).

[197] Mu-Chun Wang, Zixuan Liu, and Sheng Wang. "Textomics: A Dataset for Genomics Data Summary Generation". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4878–4891. DOI: 10.18653/v1/2022.acl-long.335. URL: https://aclanthology.org/2022.acl-long.335.

[198] Leo Wanner et al. "Towards content-oriented patent document processing". In: *World Patent Information* 30 (Mar. 2008), pp. 21–33. DOI: 10.1016/j.wpi.2007.03.008.

[199] Katharina Wäschle and Stefan Riezler. "Analyzing Parallelism and Domain Similarities in the MAREC Patent Corpus". In: *Multidisciplinary Information Retrieval*. Ed. by Michail Salampasis and Birger Larsen. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 12–27. ISBN: 978-3-642-31274-8.

[200] Michael J. Witbrock and Vibhu O. Mittal. "Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries". In: *Proceedings of the 22nd Annual International ACM R Conference on Research and Development in Information Retrieval*. R '99. Berkeley, California, USA: Association for Computing Machinery, 1999, 315–316. ISBN: 1581130961. DOI: 10.1145/312624.312748. URL: https://doi.org/10.1145/312624.312748.

[201] Kristian Woodsend and Mirella Lapata. "WikiSimple: Automatic Simplification of Wikipedia Articles". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 25.1 (2011), pp. 927–932. DOI: 10.1609/aaai.v25i1.7967. URL: https://ojs.aaai.org/index.php/AAAI/article/view/7967.

[202] Sander Wubben, Antal van den Bosch, and Emiel Krahmer. "Sentence Simplification by Monolingual Machine Translation". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 1015–1024. URL: https://aclanthology.org/P12-1107.

[203] Katharina Wäschle and Stefan Riezler. *PatTR: Patent Translation Resource*. 2014. DOI: 10.11588/data/10002. URL: https://doi.org/10.11588/data/10002.

[204] Wei Xu, Chris Callison-Burch, and Courtney Napoles. "Problems in Current Text Simplification Research: New Data Can Help". In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 283–297. DOI: 10.1162/tacl_a_00139. URL: https://aclanthology.org/Q15-1021.

[205] Wei Xu et al. "Optimizing Statistical Machine Translation for Text Simplification". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 401–415. DOI: 10.1162/tacl_a_00107. URL: https://aclanthology.org/Q16-1029.

[206] Linting Xue et al. "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer". In: *NAACL*. 2021.

[207] Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. "Evaluating the Readability of Text Simplification Output for Readers with Cognitive Disabilities". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 293–299. URL: https://aclanthology.org/L16-1045.

[208] Tiezheng Yu et al. "Vision Guided Generative Pre-trained Language Models for Multimodal Abstractive Summarization". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3995–4007. DOI: 10.18653/v1/2021.emnlp-main.326. URL: https://aclanthology.org/2021.emnlp-main.326.

[209] Manzil Zaheer et al. "Big Bird: Transformers for Longer Sequences". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 17283–17297. URL: https://proceedings.neurips.cc/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

[210] David Zajic, Bonnie Dorr, and Richard Schwartz. "BBN/UMD at DUC-2004: Topiary". In: (May 2004).

[211] Sina Zarrieß, Henrik Voigt, and Simeon Schüz. "Decoding Methods in Neural Language Generation: A Survey". In: *Information* 12.9 (2021). ISSN: 2078-2489. DOI: `10.3390/info12090355`. URL: `https://www.mdpi.com/2078-2489/12/9/355`.

[212] Jingqing Zhang et al. "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 11328–11339. URL: `http://proceedings.mlr.press/v119/zhang20ae.html`.

[213] Tianyi Zhang et al. "BERTScore: Evaluating Text Generation with BERT". In: *International Conference on Learning Representations*. 2020. URL: `https://openreview.net/forum?id=SkeHuCVFDr`.

[214] Xingxing Zhang and Mirella Lapata. "Sentence Simplification with Deep Reinforcement Learning". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 584–594. DOI: `10.18653/v1/D17-1062`. URL: `https://aclanthology.org/D17-1062`.

[215] Yuan Zhang, Jason Baldridge, and Luheng He. "PAWS: Paraphrase Adversaries from Word Scrambling". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1298–1308. DOI: `10.18653/v1/N19-1131`. URL: `https://aclanthology.org/N19-1131`.

[216] Hao Zheng and Mirella Lapata. "Sentence Centrality Revisited for Unsupervised Summarization". In: *ArXiv* abs/1906.03508 (2019).

[217] Ming Zhong et al. "Extractive Summarization as Text Matching". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 6197–6208. DOI: `10.18653/v1/2020.acl-main.552`. URL: `https://aclanthology.org/2020.acl-main.552`.

[218] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. "A Monolingual Tree-Based Translation Model for Sentence Simplification". In: *Proceedings of the 23rd International Conference on Computational Linguistics*. COLING '10. Beijing, China: Association for Computational Linguistics, 2010, 1353–1361.