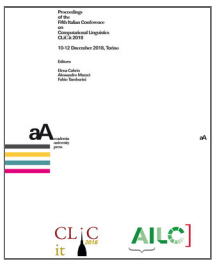

Elena Cabrio, Alessandro Mazzei and Fabio Tamburini (dir.)

Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018
10-12 December 2018, Torino

Accademia University Press

A Linguistic Failure Analysis of Classification of Medical Publications: A Study on Stemming vs Lemmatization

Giorgio Maria Di Nunzio and Federica Vezzani

DOI: 10.4000/books.aaccademia.3327
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2018
Published on OpenEdition Books: 8 April 2019
Serie: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9788831978682

A Linguistic Failure Analysis of Classification of Medical Publications: A Study on Stemming vs Lemmatization

Giorgio Maria Di Nunzio
Dept. of Information Engineering
University of Padua, Italy
dinunzio@dei.unipd.it

Federica Vezzani
Dept. of Languages and Literary Studies
University of Padua, Italy
federica.vezzani@phd.unipd.it

Abstract

English. Technology-Assisted Review (TAR) systems are essential to minimize the effort of the user during the search and retrieval of relevant documents for a specific information need. In this paper, we present a failure analysis based on terminological and linguistic aspects of a TAR system for systematic medical reviews. In particular, we analyze the results of the worst performing topics in terms of recall using the dataset of the CLEF 2017 eHealth task on TAR in Empirical Medicine.

Italiano. I sistemi TAR (Technology-Assisted Review) sono fondamentali per ridurre al minimo lo sforzo dell'utente che intende ricercare e recuperare i documenti rilevanti per uno specifico bisogno informativo. In questo articolo, presentiamo una *failure analysis* basata su aspetti terminologici e linguistici di un sistema TAR per le revisioni sistematiche in campo medico. In particolare, analizziamo i topic per i quali abbiamo ottenuto dei risultati peggiori in termini di recall utilizzando il dataset di *CLEF 2017 eHealth task on TAR in Empirical Medicine*.

1 Introduction

The Cross Language Evaluation Forum (CLEF) (Goeuriot et al., 2017) Lab on eHealth has proposed a task on Technology-Assisted Review (TAR) in Empirical Medicine since 2017. This task focuses on the problem of systematic reviews in the medical domain, that is the retrieval of all the documents presenting some evidence regarding a certain medical topic. This kind of problem is also known as total recall (or total sensitivity) problem since the main goal of the search is to find possibly all the relevant documents for a specific topic.

In this paper, we present a failure analysis based on terminological and linguistic aspects of the system presented by (Di Nunzio, 2018) on the CLEF 2017 TAR dataset. This system uses a continuous active learning approach (Di Nunzio et al., 2017) together with a variable threshold based on the geometry of the two-dimensional space of documents (Di Nunzio, 2014). Moreover, the system performs an automatic estimation of the number of documents that need to be read in order to declare the review complete.

In particular, 1) we analyze the results of those topics for which the retrieval system does not achieve a perfect recall; 2) based on this analysis, we perform new experiments to compare the results achieved with the use of either a stemmer or a lemmatizer. This paper is organized as follows: in Section 1.1, we give a brief summary of the use of stemmers and lemmatizers in Information Retrieval; in Section 3, we describe the failure analysis carried out on the CLEF 2017 TAR dataset and the results of the new experiments comparing the use of stemmers vs lemmatizers. In Section 4, we give our conclusions.

1.1 Stemming and Lemmatization

Stemming and lemmatization play an important role in order to increase the recall capabilities of an information retrieval system (Kanis and Skorkovská, 2010; Kettunen et al., 2005). The basic principle of both techniques is to group similar words which have either the same root or the same canonical citation form (Balakrishnan and Lloyd-Yemoh, 2014). Stemming algorithms remove suffixes as well as inflections, so that word variants can be conflated into their respective stems. If we consider the words *amusing* and *amusement*, the stem will be *amus*. On the other hand, lemmatization uses vocabularies and morphological anal-

yses to remove the inflectional endings of a word and to convert it in its dictionary form. Considering the example below, the lemma for *amusing* and *amused* will be *amuse*. Stemmers and lemmatizers differ in the way they are built and trained. Statistical stemmers are important components for text search over languages and can be trained even with few linguistic resources (Silvello et al., 2018). Lemmatizers can be generic, like the one in the Stanford coreNLP package (Manning et al., 2014), or optimized for a specific domain, like BioLemmatizer which incorporates several published lexical resources in the biomedical domain (Liu et al., 2012).

2 System

The system we used in this paper is based on a Technologically Assisted Review (TAR) system which uses a two-dimensional representation of probabilities of a document d being relevant $\mathcal{R}$, or non-relevant, $\mathcal{NR}$ respectively $P(d|\mathcal{R})$ and $P(d|\mathcal{NR})$ (Di Nunzio, 2018).

This system uses an alternative interpretation of the BM25 weighting schema (Robertson and Zaragoza, 2009) by splitting the weight of a document in two parts (Di Nunzio, 2014):

$$P(d|\mathcal{R}) = \sum_{w_i \in d} w_i^{BM25,\mathcal{R}}(tf) \quad (1)$$

$$P(d|\mathcal{NR}) = \sum_{w_i \in d} w_i^{BM25,\mathcal{NR}}(tf) \quad (2)$$

The system uses a bag-of-words approach on the words w_i (either stemmed or lemmatized) that appear in the document and an explicit relevance feedback approach to continuously update the probability of the terms in order to select the next document to show to the user.

In addition, for each topic the system uses a query expansion approach with two variants per topic in order to find alternative and valid terms for the retrieval of relevant documents. Our approach for the query reformulation is based on a linguistic analysis performed by means of the model of terminological record designed in (Vezzani et al., 2018) for the study of medical language and this method allows the formulation of two different query variants. The first is a list of key-words resulting from a systematic semic analysis (Rastier, 1987) consisting in the decomposition of the meaning of technical terms (that is the lexematic or morphological unit) into minimum unit of meaning that cannot be further segmented. The second is a human-readable reformulation using validly attested synonyms and orthographic alternatives as variants of the medical terms provided in the original query. The following examples show our query reformulations given the initial query provided with the CLEF 2017 TAR dataset:

Table 1: CLEF 2017 TAR topics selected for the linguistic failure analysis.

topic ID	# docs shown	# relevant	# missed
CD009579	4000	138	1
CD010339	3000	114	6
CD010653	3320	45	2
CD010783	3004	30	2
CD011145	4360	202	8

- Initial query: *Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain*;
- First variant: *Sensitivity, specificity, test, tests, diagnosis, examination, physical, straight leg raising, slump, radicular, radiculopathy, pain, inflammation, compression, compress, spinal nerve, spine, cervical, root, roots, sciatica, vertebrae, lumbago, LBP, lumbar, low, back, sacral, disc, discs, disk, disks, herniation, hernia, herniated, intervertebral*;
- Second variant: *Sensitivity and specificity of physical tests for the diagnosis of nerve irritation caused by damage to the discs between the vertebrae in patients presenting LBP (lumbago)*.

Given a set of documents, the stopping strategy of the system is based on an initial subset (percent p) of documents that will be read and a maximum number of documents (threshold t) that an expert is willing to judge.

3 Experiments

The dataset provided by the TAR in Empirical Medicine Task at CLEF 2017[1] is based on 50 systematic reviews (or topics) conducted by Cochrane experts on Diagnostic Test Accuracy (DTA). For each topic, the set of PubMed Document Identifiers (PIDs) returned by running the

[1] https://goo.gl/jyNALo

query proposed by the physicians in MEDLINE as well as the relevance judgements are made available (Kanoulas et al., 2017). The aim of the task is to retrieve all the documents that have been judged as relevant by the physicians. The results achieved by the participating teams to this task showed that it is possible to get very close to a perfect recall; however, there are some topics for which most of the systems did not retrieve all the possible relevant documents, unless an unfeasible amount of documents is read by the user.

In this paper, i) we present a linguistic and terminological failure analysis of such topics and, based on this analysis, ii) the results of a new set of experiments that compare the use of either a stemmer or a lemmatizer in order to evaluate a possible improvement in the performance in terms of recall. As a baseline for our analyses, we used the source code provided by (Di Nunzio, 2018). The two parameters of the system — the percentage p of initial training documents that the physician has to read, and the maximum number of documents t a physician is willing to read — were set to $p = 500$ and $t = 100, 500, 1000$.

3.1 Linguistic Failure Analysis

In order to select the most difficult topics for the failure analysis, we run the retrieval system with parameters $p = 50\%$ and threshold $t = 1000$ and selected those topics for which the system could not retrieve all the relevant documents, five in total, shown in Table 1. In order to find out why the system did not retrieve all the relevant documents for these topics, we focused on linguistic and terminological aspects both of technical terms in the original query and of the abstracts of missing relevant documents.

We started by reading the abstract of all 19 missing relevant documents and manually selecting technical terms, defined as as all the terms that are strictly related to the conceptual and practical factors of a given discipline or activity (Vezzani et al., 2018), in this case the medical discipline. Then, we compared these terms with those previously identified in the two query variants encoded in the retrieval system. From this comparison, we noticed that most of the relevant terms extracted from the abstracts were not present in the previous two reformulation (a minimum of 0 and a maximum of 8 terms in common), so that some relevant documents in which such terms were present have not been retrieved. By focusing on the morphological point of view, we have been able to categorize such techincal terms in: 1) acronyms; 2) pairs of terms, in particular noun-adjective; 3) triad of terms, in particular noun-adjective-noun.

The category of acronyms is not an unexpected outcome. Medical language is caracterized by an high level of abbreviations and acronyms (Rouleau, 2003) and, in order to retrieve those missing relevant documents, we should have considered all the orthographic variants of a technical term as well as its acronym or expansion according to the case.

Regarding the second and the third category, that is the pairs noun-adjective (e.g.: bile/biliary, pancreas/pancreatic, schizophrenia/schizophrenetic) and the triad of terms noun-adjective-noun (e.g.: psychiatry/psychiatric/psychiatrist), we noticed some problems related to the stemming process. The analysis carried out allowed us to identify numerous cases of understemming, as for example the case of *psychiatry* stemmed as *psychiatri*, *psychiatric* stemmed as *psychiatr* and *psychiatrist* stemmed as *psychiatrist*, all of them belonging to the same conceptual group. The fact that the stemmer recognizes these three words as different suggests us that the conflation of the inflected forms of a lemma in the query expansion procedure may help to retrieve the missed relevant documents.

3.2 Stemming vs Lemmatization

For the reasons explained in the previous section, we decided to perform a new set of experiments on these "difficult" topics to study whether a lemmatization approach can improve the recall compared to the stemming approach. We used the standard algorithms implemented in the two R packages SnowballC[2] and Textstem.[3] Both implements the Porter stemmer (Porter, 1997), while the second uses the TreeTagger algorithm (Schmid, 1999) to select the lemma of a word. To make a fair comparison for the stemming vs lemmatization part of the analysis, in our experiments we did not use any of the two query variants. By reproducing the results presented in (Di Nunzio, 2018), we discovered an issue in the original source code concerning the stemming phase. The R package *tm* for text mining[4] calls the stemming function of the Snow-

[2] https://goo.gl/n3WexD
[3] https://goo.gl/hCLGP8
[4] https://goo.gl/wp859o

ballC with the "english" language instead of the default "porter" stemmer. This caused a substantial difference in the terms produced for the index and those stemmed during the query analysis. For this reason, all our results are significantly higher compared to those presented by (Di Nunzio, 2018) which makes this approach more effective than the original work.

We studied the performance in terms of recall, and precision at 100, 500, and 1000 documents read (p@100, P@500, and P@1000 respectively) for different values of the threshold t. In Table 2, we report in the first column of each value of t the performance of the original experiment compared to our results (only recall is available from (Di Nunzio, 2018)). If we observe the performances on the whole set of test queries, there is no substantial difference between stemming and lemmatization. There is some improvement in terms of recall when threshold $t = 100$, however 85% of recall is usually considered a 'low' score in total recall tasks. Table 3 compares the number of relevant documents missed by the stemming and lemmatization approaches on the difficult topics. The differences between the original experiments and these new experiments are minimal apart from topic CD010339 for which the absence of the two query reformulations led to a worse performance.

4 Final Remarks and Future Work

In this work, we have presented a linguistic failure analysis in the context of medical systematic reviews. The analysis showed that, for those topics where the system does not retrieve all the relevant information, the main issues are related to abbreviations and pairs noun-adjective and the triad of terms noun-adjective-noun. We performed a new set of experiments to see whether lemmatization could improve over stemming but the results were not conclusive. The issues remain the same since the type of relation noun-adjective or noun-adjective-noun, cannot be resolved by a lemmatizer. For this reason, we are currently studying an approach that conflates morphosyntactic variants of medical terms into the same lemma (or 'conceptual sphere') by means of medical terminological records (Vezzani et al., 2018) and the use of the Medical Subject Headings (MesH) dictionary. [5] In this way, we expect that the system will automatically identify all the related forms (such as all the derivative nouns, adjectives or adverbs) of a lemma in order to include them in the retrieval process of potentially relevant documents.

[5] https://meshb.nlm.nih.gov/search

Acknowledgments

The authors would like to thank Sara Bosi and Fiorenza Germana Grilli, students of the Master Degree in Modern Languages for International Communication and Cooperation of the Department of Linguistics and Literary Study of the University of Padua, who helped us in the linguistic failure analysis phase.

References

Vimala Balakrishnan and Ethel Lloyd-Yemoh. 2014. Stemming and lemmatization: A comparison of retrieval performances. *Lecture Notes on Software Engineering*, 2(3):262 – 267.

Giorgio Maria Di Nunzio, Federica Beghini, Federica Vezzani, and Geneviève Henrot. 2017. An Interactive Two-Dimensional Approach to Query Aspects Rewriting in Systematic Reviews. IMS Unipd At CLEF eHealth Task 2. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*

Giorgio Maria Di Nunzio. 2014. A New Decision to Take for Cost-Sensitive Naïve Bayes Classifiers. *Information Processing & Management*, 50(5):653 – 674.

Giorgio Maria Di Nunzio. 2018. A study of an automatic stopping strategy for technologically assisted medical reviews. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 672–677.

Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, João Palotti, and Guido Zuccon, 2017. *CLEF 2017 eHealth Evaluation Lab Overview*, pages 291–303. Springer International Publishing, Cham.

Jakub Kanis and Lucie Skorkovská. 2010. Comparison of different lemmatization approaches through the means of information retrieval performance. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, pages 93–100, Berlin, Heidelberg. Springer Berlin Heidelberg.

Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker, editors. 2017. *CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview. In Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017.*, CEUR Workshop Proceedings. CEUR-WS.org.

Table 2: Performance of stemming vs lemmatization for different values of t

	t = 100			t = 500			t = 1000		
	(Di Nunzio, 2018)	stem	lemma	(Di Nunzio, 2018)	stem	lemma	(Di Nunzio, 2018)	stem	lemma
recall	.645	.854	.875	.940	.976	.969	.988	.992	.992
P@100	-	.194	.208	-	.194	.194	-	.194	.208
P@500	-	.113	.108	-	.098	.976	-	.098	.099
P@1000	-	.100	.096	-	.070	.070	-	.071	.071

Table 3: Number of relevant documents missed by the original experiment (see Table 1), the stemming approach (original experiment corrected), and the lemmatization approach.

topic ID	# original	# stem	# lemma
CD009579	1	1	1
CD010339	6	15	16
CD010653	2	1	1
CD010783	2	1	1
CD011145	8	7	9

Kimmo Kettunen, Tuomas Kunttu, and Kalervo Järvelin. 2005. To stem or lemmatize a highly inflectional language in a probabilistic ir environment? *Journal of Documentation*, 61(4):476–496.

Haibin Liu, Tom Christiansen, William A. Baumgartner, and Karin Verspoor. 2012. Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics*, 3(1):3, Apr.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Martin F. Porter. 1997. An algorithm for suffix stripping. In Karen Sparck Jones and Peter Willett, editors, *Readings in Information Retrieval*, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

François Rastier. 1987. *Sémantique interprétative*. Formes sémiotiques. Presses universitaires de France.

Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Maurice Rouleau. 2003. La terminologie médicale et ses problèmes. Tribuna, Vol. IV, n. 12.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural Language Processing Using Very Large Corpora*, pages 13–25. Springer Netherlands, Dordrecht.

Gianmaria Silvello, Riccardo Bucco, Giulio Busato, Giacomo Fornari, Andrea Langeli, Alberto Purpura, Giacomo Rocco, Alessandro Tezza, and Maristella Agosti. 2018. Statistical stemmers: A reproducibility study. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 385–397.

Federica Vezzani, Giorgio Maria Di Nunzio, and Geneviève Henrot. 2018. TriMED: A Multilingual Terminological Database. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).