Terminology Extraction for and from Communications in Multi-disciplinary Domains

Xiubo Zhang*, Raffaella Panizzon†, Maria Teresa Musacchio†, Khurshid Ahmad*

*Trinity College Dublin
Dublin, Republic of Ireland
zhangx6, khurshid.ahmad@scss.tcd.ie

†University of Padova
Padova, Italy
raffaella.panizzon, mt.musacchio@unipd.it

Abstract

Terminology extraction generally refers to methods and systems for identifying term candidates in a uni-disciplinary and uni-lingual environment such as engineering, medical, physical and geological sciences, or administration, business and leisure. However, as human enterprises get more and more complex, it has become increasingly important for teams in one discipline to collaborate with others from not only a non-cognate discipline but also speaking a different language. Disaster mitigation and recovery, and conflict resolution are amongst the areas where there is a requirement to use standardised multilingual terminology for communication. This paper presents a feasibility study conducted to build terminology (and ontology) in the domain of disaster management and is part of the broader work conducted for the EU project Slándáil (FP7 607691). We have evaluated CiCui (for Chinese name 词萃, which translates to *words gathered*), a corpus-based text analytic system that combine frequency, collocation and linguistic analyses to extract candidates terminologies from corpora comprised of domain texts from diverse sources. CiCui was assessed against four terminology extraction systems and the initial results show that it has an above average precision in extracting terms.

Keywords: terminology extraction, software evaluation, multilingual communication

1. Terminology and Ontology of Social Media Streams

The existence of a term relies on textual evidence, i.e. the statistically significant occurrence of a term in a number of randomly sampled texts within a domain (Ahmad et al., 1994; Ahmad, 2001). Term extraction can be exploited to meet specific needs such as glossary compilation, translation, information retrieval (IR), ontology or conceptual map generation among others. The extraction of terms and of the information associated with them (i.e. definitions, synonyms, related concepts, etc.) from domain-specific corpora has received considerable attention, thus encouraging the study and development of a variety of automatic or semi-automatic extraction methods and tools – especially in fast growing disciplines such as biotechnology or computer science – and the broadening, update or harmonisation of existing termbanks and glossaries. A number of different systems are currently available both as freeware and proprietary software for a number of purposes such as research in linguistics, improvement of professional translators' performance or social media trend monitoring. Term extraction methods also facilitate the extraction of candidate ontologies (Ahmad and Gillam, 2005). The rise of Internet-based communications, and especially social media, has expedited the development of multidisciplinary terminology through the availability of large volumes of specialist texts in a variety of domains (Ahmad et al., 2006). Term extraction methods have been used in one of the traditional foci of social media analytics – film reviews and viewer sentiment (Manek et al., 2016). Term extraction and ontological mapping have made considerable progress since pharmaceutical companies have started to use social media to monitor reports of adverse reactions to drugs, also called pharmacovigilance — to do this the precise terminology of the domain needs to match a layperson's (patient administered a drug) use of the term. Terminology extraction techniques have been used in the pharmacovigilance domain with some success and are based on statistical machine learning techniques (Nikfarjam et al., 2015). Terminology extraction has been put to use in the surveillance of automotive component failure as reported by the 'buzz' in social media (Abrahams et al., 2013).

Multidisciplinary subjects, especially disaster management that involves a large number of agencies with different objectives but focussed on disaster mitigation and recovery, are characterised by terminology that is essentially a federated collection of terms from different constituent domains. A terminology collection in a multidisciplinary domain has to be carefully prepared and terms need to be elaborated fully. Emergency management and disaster relief organisations have developed and maintained terminology concerning natural hazards. The Slándáil project, whose goal is to ethically improve the use of social media in enhancing the response of disaster related agencies, has surveyed existing terminology resources, which usually take the form of glossaries containing entries with a term and its definition. These usually take the form of glossaries containing entries with a term and its definition. FEMA has an online glossary[1], Australia's EMA also has an emergency management glossary[2], while New Zealand's Ministry of Civil Defence and Emergency Management has local emergency management plans with short glossaries[3]. Similarly, Germany's BBK has a glossary[4] in

[1] https://www.fema.gov/rules-tools/glossary-terms

[2] https://www.ag.gov.au/EmergencyManagement/Tools-and-resources/Publications/Documents/Manual-series/manual-3-australian-emergency-glossary.pdf

[3] http://www.gdc.govt.nz/assets/Files/Civil-Defence/Glossary-Abbreviations-2009.pdf

[4] http://www.bbk.bund.de/SharedDocs/Downloads/BBK/DE/Publikationen/Praxis_Bevoelkerungsschutz/Band_8_Praxis_BS_BBK_

German and Italy's Protezione Civile has an official, concise one[5] on its webpage, while more extended glossaries are available in the webpages of voluntary organizations more or less closely associated with the national one[6]. While these glossaries reflect the need of emergency management organizations to communicate their terminology in a concise or more extended form, there is little if any information on how they were developed. Glossaries in PDF format are difficult to search and update. There are also some bilingual glossaries – Canada's English-French[7], Italy's Italian-English – both with definitions – and Germany's German-English, which provides only English equivalents. A trilingual glossary was found in Italy's South Tirol (Zivilschutz Glossary[8]). International organizations include some emergency management and natural hazards terms in their glossaries – cf. EIONET and IATE – while UNISDR developed a glossary of 53 terms which was extended to 80 in 2015 [9]. UNISDR also offers some information about terminology development[10], as it clearly states that terms were identified in a corpus of 35,000 documents and then validated by a group of experts. There seems to be still ample scope for automating term extraction, populating term entries and establishing conceptual relations through ontologies. Table 1 provides further details about each glossary, in particular how term entries are structured and how many terms are included as well as whether terms and other relevant information are linked to each other (hypertextuality), whether the glossary can be searched and navigated (interactivity) and if their format can be easily updated.

2. Sublanguages of Specialist Domains and Social Media

The use of language by humans in every domain of enterprise shows that some words and some linguistic structures are used more frequently than others. A sublanguage is "a specialized language or jargon associated with a specific group or context"[11] where words describing key objects/events/ideas and key activities are used almost exclusively for key objects and activities: 'bank' in financial transactions is not the same 'bank' as used in river engineering, and the activity where we rely or bank on others is confused by someone going to 'bank' money in a bank. The notion of sublanguages was propounded by Zellig Harris, a pioneering figure in modern linguistics, who has looked at the language used by mathematicians and biochemists and noted subtle differences (Schwartz et al., 2013) in language use between the two domains and between the general everyday language and the sublanguages in the two domains (Harris, 1991). Since Harris and others, there has been much work on the translation of sublanguage texts (Grishman and Kittredge, 2014).

There is also an equally important sublanguage that is shaped by the medium used – we had telegraphese in the 19th-20th century due to telegraphy technology, and the 140 character language, complete with shriek symbols and @ signs, i.e. Twitter, for the 21st century. It has been suggested that this is "what people say in social media to find distinctive words, phrases, and topics as functions of known attributes of people such as gender, age, location, or psychological characteristics" (Schwartz et al., 2013). A combined study of the sublanguage of a specialist domain and that of the (micro) blogging and social networking has been deployed to understand how patients are reacting to diseases like breast cancer (Elhadad et al., 2014). Sublanguage studies have been used in retrieving and analysing 'hazard related' posts on social media networks (Bolea, 2015).

3. Terminology Extraction Method

The terminology used in a sublanguage plays a crucial role in the characterisation of the conceptual composition of the corpus. Such information provides insights into the key issues and concerns in the domain of emergency management. The CiCui system implements a machine learning based automatic terminology extraction procedure. The system first extracts preliminary term candidates (TCs) by matching preprocessed text against predefined linguistic patterns; it then further refines the resultant TCs using statistical classifiers trained on previously labelled data. The workflow of the CiCui system is summarised in Figure 1.

3.1. Extracting Preliminary Term Candidates

The system first treats the input documents with natural language processing techniques; running texts are tokenised and tagged with part-of-speech information using the Stanford CoreNLP package (Manning et al., 2014). We employed the TokenRegex facility in Stanford CoreNLP to extract preliminary term candidates (TCs); TokenRegex allows matching word sequences using regular expressions specified at a token level (instead of at a character level as in normal regular expressions). The linguistic pattern used in our method is: `[word:/[a-zA-Z-]+/; tag:/NN|NNS|JJ/]+? [word:/[a-zA-Z-]+/; tag:/NN|NNS/]+`. The pattern matches noun sequences optionally modified with adjectives; all words in the term must consist of only letters and hyphens. Word sequences that match the above pattern are kept as preliminary TCs. Frequencies, document frequencies, tf-idf scores, and weirdness scores are computed for each word in the vocabulary of the corpus. The weirdness score for a certain word is a keywordness measure defined as the ratio between the word's relative frequency in a domain corpus and its relative frequency in a reference general corpus; in this case, the frequencies of

Glossar.pdf

[5]http://www.protezionecivile.gov.it/jcms/it/glossario.wp

[6]http://www.proingpa.it/wp-content/uploads/2011/10/Glossario-protezione-civile-rev1.pdf

[7]https://www.sdc.gov.on.ca/sites/mgcs-onterm/Documents/Glossaries/EMOGlossaryEN-FR.htm

[8]http://www.provinz.bz.it/zivilschutz/service/veroeffentlichungen.asp

[9]http://www.unisdr.org/we/inform/terminology

[10]http://www.preventionweb.net/files/45462_backgoundpaperonterminologyaugust20.pdf

[11]http://www.oxforddictionaries.com/definition/english/sublanguage

Name	Lang.	Term Entry	# of Terms	Hypertextual-ity	Interactivity	Updatable Format
(FEMA) Glossary of Terms – 2015	ENG	term (and acronym), definition	154	×	✓ (navigation from one letter to the other)	✓ (web page)
(EMA) Australian Emergency Management Glossary – 2016	ENG	term (and acronym), abbreviation, definition(s), synonyms, related terms, references	ca. 1100	×	×	× (PDF)
(NZ MCDEM) Glossary/Abbreviations – 2009	ENG	term, abbreviation, definition, references	66	×	×	× (PDF)
BBK-Glossar: Ausgewählte zentrale Begriffe des Bevölkerungsschutzes (glossary of selected key concepts of emergency management) – 2011	GER	term (and acronym), abbreviation, definition, synonyms, related terms, notes, references	176	×	×	× (PDF)
(Protezione Civile) Glossario (glossary) – n.d.	ITA	term (and acronym), definition, synonyms, related terms	278	✓ (some related terms are hyper-linked)	✓ (navigation from one letter to the other)	✓ (web page)
(Palermo Engineers' Association for Civil Protection and Emergency Management) Nuovo Glossario di Protezione Civile (new civil protection glossary) – 2012	ITA	term (and acronym), abbreviation, definition, related terms, references	459	× (only one hyperlink to an external source)	×	× (PDF)
(Emergency Management Ontario) English-French Emergency Management Glossary Of Terms – 2011	ENG FR	term (and acronym), definition, synonyms, references	123	×	×	✓ (web page)
(Bolzano Province) Civil Protection Glossary – 2013	ITA GER ENG	term(and acronym), synonyms	357	×	×	× (PDF)
(UNISDR) Terminology on Disaster Risk Reduction – 2009	ENG	term (and acronym), definition, synonyms, related terms, notes, references	53	×	×	× (PDF)
(UNISDR) Proposed Updated Terminology on Disaster Risk Reduction: A Technical Review – 2015	ENG	term (and acronym), definition, synonyms, related terms, notes, references	80	×	×	× (PDF)

Table 1: Features of glossaries from emergency management and disaster relief organisations

words in the COCA corpus (Davies, 2008) were used to derive the reference relative frequencies.

3.2. Refining Term Candidates with Statistical Methods

3.2.1. Model Training

The preliminary TCs extracted then undergo additional filtering using statistical classifiers. These classifiers were trained using a set of labelled terms extracted from another disaster management corpus separately prepared. The training corpus contained 2015 announcements and documents published by various disaster management agencies around the globe[12], totalling 418,513 words. From the list of 11,721 preliminary TCs which were sorted descendingly according to their tf-idf scores, five batches of TCs each containing 500 entries were sampled evenly every 2000 items; the sampled TCs were manually evaluated by two human raters and divided into three categories: 'Green' (G), 'Amber' (A), and 'Red' (R). 'Red' indicates that the TC does not constitute a valid term; 'Amber' indicates the TC is not identified as a validate term as a whole but may contain one; and the 'Green' category signifies a valid term. The distribution of the manually labelled categories is shown in Table 2. The distribution of the classes in the training set appears unbalanced; as a result, the baseline accuracy for classifications on the dataset will be 65%, which is achieved when all TCs are labelled with 'Red'.

Features For training the classifiers, the following features were used:

[12]The sources include FEMA, NASA Earth Observatory Natural Hazards, CDC Emergency Preparedness and Response, GDACS, and ReliefWeb.

Batch	R	A	G
1-501	164	205	132
2001-2501	294	186	21
4001-4501	350	144	7
6001-6501	422	75	4
8001-8501	364	136	1
Total	1594	746	165
Percentage (%)	64	30	6

Table 2: RAG distribution in training set grouped by batch

- The frequency, document frequency, and tf-idf of a TC: these statistics characterise the usage of a preliminary TC in the corpus as a whole. *TC frequency* is the number of times the TC was encountered in the text; *TC document frequency* counts the number of documents in which the TC ocurred; *TC tf-idf* is the tf-idf score calculated from TC frequency (tf) and TC document frequency df:

  $$tf\text{-}idf = tf \times \log \frac{|D|}{df}$$

  where $|D|$ is the total number of documents in the corpus.
- The mean, standard deviation, maximum, and minimum of the frequencies, document frequencies, tf-idf scores, and weirdness scores of the constituent words for each TC: these features summarise the characteristics of the individual words in each TC.
- The length of a TC: unusually long word sequences can be a result of misclassification of part-of-speech or other issues such as malformed sentences.

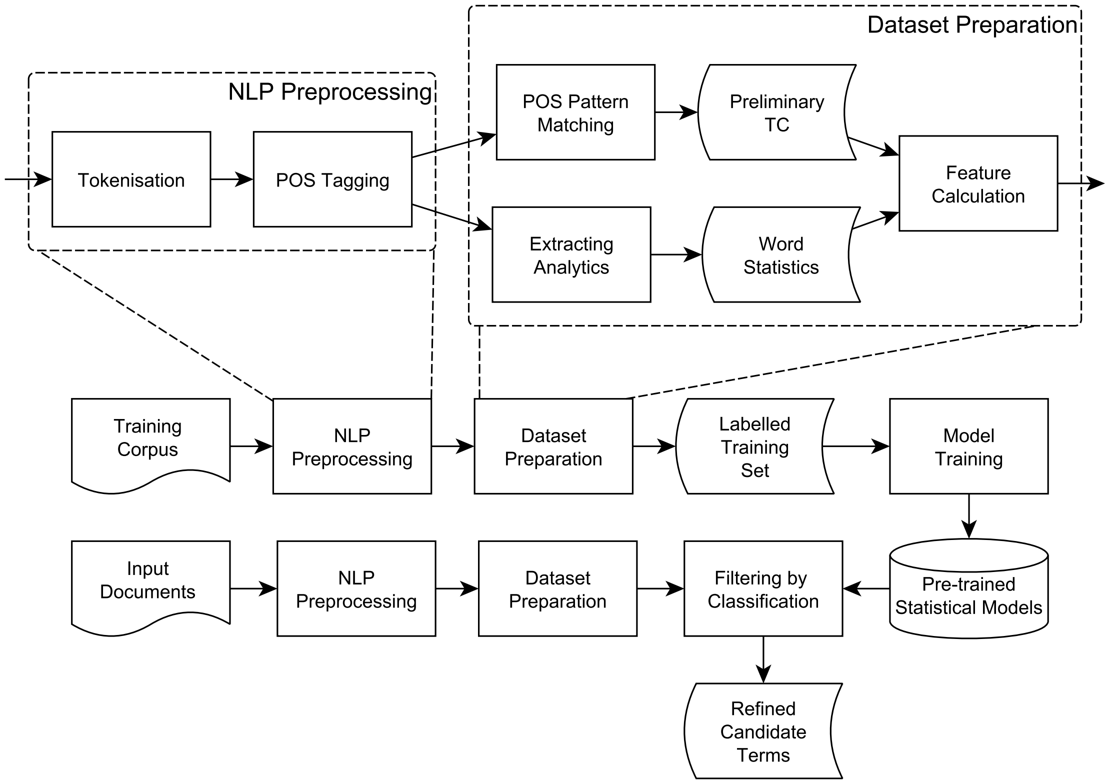

Figure 1: CiCui's Term Extraction Workflow

- The proportion of nouns in the TC: TCs comprised of mostly nouns are more likely to be valid terms.
- A binary feature indicating whether or not the first word in a TC is an adjective, and also the weirdness score of the first word of a TC.

All numerical features except for the proportion of nouns and the binary indicator were first logarithmically transformed and then normalised between 0 and 1 before further processing.

Model Performances We cross-validated a number of different classifiers on the training set using implementations provided by a data mining platform called KNIME (Berthold et al., 2007). The results of the cross-validation are tabulated in Table 3. Among the classifiers tested, random forest, neural network, SVM, and radial basis function network performed significantly better than the baseline, with the best being the random forest, which conferred a 7.47% increase in accuracy compared to the baseline.

3.3. Classification of Preliminary Term Candidates

Features for preliminary TCs extracted from new documents were prepared in the same way during the training. Based on the results from the training session, we classified new preliminary TCs by consulting an ensemble of three classifiers: a random forest, a SVM, and a multilayer neural network, all trained using the set-up described in Section 3.2.1. Each of the three classifiers votes for the preliminary TCs independently. A preliminary TC is kept if and only if it satisfies any of the following two criteria: (i) it received no 'R' label and at least one 'G' label; (ii) it received exactly one 'R' vote and two 'G' votes.

4. Term Extraction Evaluation

In this section, we present our evaluations of four commercial and open-source terminology extraction systems by looking at their extraction methods (statistical, linguistic or hybrid). The performances of these systems on a test corpus were evaluated and compared to the results returned by CiCui on the same test corpus. The performance scores of each are analysed and individual features are discussed. A summary of these systems is presented in Table 4.

Synchroterm is a Canadian-based statistical term extractor from Terminotix designed for professional translators and terminologists needing to create and manage translation memories. Though created to work with parallel texts (two languages), it can also extract terminology from unilingual documents. The testing has been conducted by selecting compound nouns from 2 to 8 elements without any stop list (though the option is available). The software allows a user to import lists of terms and expressions to be ignored during an extraction, and to create and modify a list of deleted items that are then ignored during all further extractions.

TaaS has been created within the EU project Accurat and uses a hybrid method, thus combining linguistic analysis (part of speech tagging, morpho-syntactic patterns, etc.) enriched by statistical features (e.g., frequency score). It supports all EU languages and Russian.

TermoStat Web 3.0 is a term extractor that uses both linguistic and statistical methods taking the potential terms'

Classifier	mean acc.	s.d.	mean diff.	t-statistic	p-value
Random Forest (100 Decision Trees)	71.47	2.02	7.47	11.71	< 0.01
Neural Network (2 layers, 10 nodes per layer)	70.27	2.35	6.27	8.43	< 0.01
SVM (linear kernel)	69.82	2.88	5.82	6.40	< 0.01
Radial Basis Function Network (Weka 3.7)	69.66	3.43	5.66	5.21	< 0.01
Logistic Regression	65.93	4.93	1.93	1.24	0.25
Multinomial Naive Bayes (Weka 3.7)	65.53	1.57	1.53	3.08	0.01
Naive Bayes	63.77	3.69	-0.23	-0.20	0.85
Fuzzy Rule Learner	63.00	2.67	-1.00	-1.18	0.27
Ordinal Logistic Regression (R, `MASS` package)	60.83	8.22	-3.17	-1.22	0.25

Table 3: Each classifier listed in the table was trained and tested with 10-fold cross-validation. The *mean acc.* and the *s.d.* column show the mean and the standard deviation of accuracies from the 10 classifications respectively. The *mean diff.*, *t-statistic*, and *p-value* columns show the result from a one-sample t-test (degree-of-freedom $= 9$) with the null hypothesis being that the average accuracy of the 10 classifications does not differ from the baseline (i.e. 64%).

structures and relative frequencies into account in the corpus analysis. It compares the specialised corpus provided by users with an in-built reference corpus for each of the languages it can process (French, English, Italian, Spanish and Portuguese). Users can choose to analyse mono and/or polylexical units. The English reference corpus has approximately $8{,}000{,}000$ words; half of it consists of news articles from the daily *The Gazette* published between March and May 1989 while the other half is taken from the British National Corpus (BNC). The corpus submitted is first tagged using TreeTagger and then, using regular expressions, simple or compound words are matched with predefined syntactic matrices.
Vocabgrabber is a software system that analyses text and generates lists of words and their use in context. No indication is provided as to the method applied to select and rank TCs though it is likely to be based on statistical frequency which can be sorted by subject (geography, people, social studies, etc.); ordered alphabetically, by relevance or "familiarity", i.e. frequency in general language.

Terminology Extraction System	Year	Target	Method
Synchroterm	2014	Translators, Terminologists	statistical
TaaS	2016	Companies	hybrid
TermoStat	2010	Linguists	hybrid
Vocabgrabber	2016	General public	statistical
CiCui	2014	Linguists, Companies	hybrid

Table 4: Summary of Terminology Extraction Systems

A test corpus consisting of $326{,}319$ words and comprising texts from FEMA fact sheets, handbooks from different emergency management agencies and news items extracted from LexisNexis (keywords: weather, emergency, disaster, Sandy, hurricane, superstorm, storm) has been used to test the systems described. In some cases only part of the corpus was analysed due to restrictions applied by systems. The results provided by each software system have been manually evaluated by assigning RAG labels to the top 100 TCs produced by each, and then compared with CiCui's performance.
Because an evaluation of recall requires to know all the terms present in the corpus in advance, only precision scores were calculated. The formula to calculate precision is:

$$P = \frac{A}{A+C} \times 100\%$$

where A is the number of accepted terms (i.e. 'Green') and C is the number of discarded results ('Amber' and 'Red'). The four extractors presented above have been tested on the same corpus, though in the case of TaaS and Vocabgrabber restrictions were applied, so that only part of the corpus was analysed. For the former $100{,}000$ tokens were processed, while for the latter a threshold was set at $200{,}000$ characters. Below are the results for precision and manual (RAG) evaluation calculated on the top 100 CTs returned by each system. Although 'Green' terms and precision express the same measure, i.e. the percentage of validated terms, both have been included in the table below for the sake of completeness.
The evaluation presented above allows for the comparison of the automatic term extraction system (CiCui) with currently available software performing similar tasks. It has been observed that term extractors have been designed with different users in mind (professional translators, businesses, linguists or general users), which strongly influenced the quality of their output. It also shows that automatic term extraction can greatly benefit from the adoption of machine learning techniques.

5. Conclusions

Successful disaster mitigation and recovery would not be feasible without the collaboration of experts from a variety of domains, who are bound to use more or less overlapping terminology. Therefore, all those involved in emergency management can greatly benefit from the standardisation of multidisciplinary and multilingual terminology. The evaluation presented above allows for the comparison of the Slándáil automatic term extraction system (CiCui) with four currently available terminology extraction tools (two commercial and two open-source) performing similar tasks. It has been observed that term extractors have been designed with different users in mind (professional translators, businesses, linguists or general users), which strongly influenced the quality of their output. The comparison was carried out by calculating their performance on precision scores and by manually evaluating the top 100 TCs ranked by frequency.

System	# of TCs extracted	Precision	RAG evaluation		
			Red (%)	Amber (%)	Green (%)
Synchroterm	14791	56%	30	14	56
TaaS	345	42%	48	10	42
TermoStat	4082	64%	20	16	64
Vocabgrabber	2501	14%	78	14	14
CiCui	602	77%	6	17	77
Average			36.4	14.2	50.6
Standard Deviation			27.8	2.7	24.1

Table 5: RAG evaluation between terminology extraction systems

To this end, a test corpus was created and used to trial all five tools. Results highlighted substantial differences among them, with hybrid systems generally performing better in terms of precision and in the number of potentially invalid TCs. CiCui's term recognition showed above average results for precision and segmentation, although extensive work is being carried out to achieve further improvement.

6. Acknowledgements

The research leading to these results has received funding from the European community's Seventh Framework Programme under grant agreement No. 607691 (SLANDAIL).

7. Bibliographical References

Abrahams, A. S., Jiao, J., Fan, W., Wang, G. A., and Zhang, Z. (2013). What's buzzing in the blizzard of buzz? automotive component isolation in social media postings. *Decision Support Systems*, 55(4):871 – 882. 1. Social Media Research and Applications 2. Theory and Applications of Social Networks.

Ahmad, K. and Gillam, L., (2005). *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2005, Agia Napa, Cyprus, October 31 - November 4, 2005, Proceedings Part II*, chapter Automatic Ontology Extraction from Unstructured Texts, pages 1330–1346. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ahmad, K., Davies, A., Fulford, H., and Rogers, M. (1994). What is a term? the semi-automatic extraction of terms from text. *Translation studies: an interdiscipline*, 267:278.

Ahmad, K., Cheng, D., and Almas, Y. (2006). Multi-lingual sentiment analysis of financial news streams. In *Proc. of the 1st Intl. Conf. on Grid in Finance*.

Ahmad, K. (2001). The role of specialist terminology in artificial intelligence and knowledge acquisition. *Handbook of Terminology Management: Application-oriented terminology management*, 2:809.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., and Wiswedel, B. (2007). KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer.

Bolea, S. C. (2015). Vocabulary, synonyms and sentiments of hazard-related posts on social networks. In *Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on*, pages 1–6. IEEE.

Elhadad, N., Zhang, S., Driscoll, P., and Brody, S. (2014). Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions. In *AMIA Annual Symposium Proceedings*, volume 2014, page 516. American Medical Informatics Association.

Grishman, R. and Kittredge, R. (2014). *Analyzing language in restricted domains: sublanguage description and processing*. Psychology Press.

Harris, Z. (1991). Theory of language and information: a mathematical approach.

Manek, A. S., Shenoy, P. D., Mohan, M. C., and Venugopal, K. (2016). Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and svm classifier. *World Wide Web*, pages 1–20.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., and Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

8. Language Resource References

Davies, Mark. (2008-). *The Corpus of Contemporary American English: 520 million words, 1990-present*. Availalbe online at `http://corpus.byu.edu/coca/`.