

Automatic Term Extraction from Newspaper Corpora: Making the Most of Specificity and Common Features

Patrick Drouin, Jean-Benoît Morel, Marie-Claude L’Homme

Observatoire de linguistique Sens-Texte (OLST)

Université de Montréal

C.P. 6128, succ. Centre-ville

Montréal (Québec) H3C 3J7 CANADA

{patrick.drouin, jean-benoit.morel, mc.lhomme}@umontreal.ca

Abstract

The first step of any terminological work is to setup a reliable, specialized corpus composed of documents written by specialists and then to apply automatic term extraction (ATE) methods to this corpus in order to retrieve a first list of potential terms. In this paper, the experiment we describe differs from this usual process. The corpus used for this study was built from newspaper articles retrieved from the Web using a short list of keywords. The general intuition on which this research is based is that ATE based corpus comparison techniques can be used to capture both similarities and dissimilarities between corpora. The former are exploited through a termhood measure and the latter through word embeddings. Our initial results were validated manually and show that combining a traditional ATE method that focuses on dissimilarities between corpora to newer methods that exploit similarities (more specifically distributional features of candidates) leads to promising results.

Keywords: terminology, automatic term extraction, unspecialized corpora

1. Introduction

The first step of any terminological work is to setup a reliable, specialized corpus composed of documents written by specialists. It is usually assumed that only domain-specific corpora compiled according to criteria defined by terminologists can represent good starting points for terminological description. This is especially true when relying on automatic term extraction (ATE) tools as the quality of the output is in direct relation to the quality of the input.

However, these "ideal" requirements are not always met in certain fields of knowledge. This is the case of the domain explored in this work, i.e. problematic behavior in the workplace. Its terminology can be disseminated in various forms of textual genres, including unspecialized corpora.

Extracting terminology from unspecialized corpora raises new challenges for ATE since most tools and methodologies are built around the assumption that the corpora being processed are specialized. Tools and methodologies thus tend to target features specific to this type of corpora. One efficient strategy for spotting domain-specific terms consists in comparing the behavior of the lexicon of a specialized corpus (an analysis corpus, AC) to the behavior of the lexicon in a general language corpus (a reference corpus, RC), thus exploiting the difference between text genres.

Such a technique has proved efficient for extracting relevant and interesting term candidates. One question remains however: Can we expect this comparison method to yield similar results when comparing corpora that belong to the same genre or when using an analysis corpus that is unspecialized? We believe that, although still useful, the method would need to be complemented with further processing.

This paper presents an automatic term extraction experiment carried out on a newspaper corpus that contains texts that address directly or indirectly the topic of discrimination. We first explore the results of a hybrid corpus comparison ATE experiment and propose new techniques in order

to increase the precision of the results obtained. We believe that the proposed approach is useful to tackle ATE from unspecialized corpora and that the underlying ideas can be used for ATE in other situations.

2. The task

For the project described in this paper, we have been working with a private company (Valital¹) whose core business is the real-time online analysis of job candidates behavior and the automated confirmation of their work experience. Their process digs into various sources of information with the aim of defining a textual profile for different kinds of misconduct in the workplace. Among these sources, are newspaper articles dealing with problematic behavior (e.g. violence, discrimination), but most articles do not concern the workplace as such. One of the tasks assigned to our team was to capture the terminological profile for each of these behaviors. This terminological profile was to be implemented in an ontology at a later stage.

From a terminological standpoint, newspaper articles are "atypical" textual sources since they are targeted at the general public. Even if these articles were automatically filtered according to the topic they address based on a short list of keywords, they may or may not concern the workplace as such. In other words, articles can report on a discrimination situation, but this situation could have taken place anywhere. The challenge in this case was to be able to locate relevant terms in an unspecialized corpus.

Our task involved an additional less typical aspect. The terminology related to misconduct includes various types of terms such as verbs (e.g. *discriminate*), adjectives (e.g. *discriminatory*) or single-word predicative nouns (e.g. *discrimination*). The term extraction method needed to be able to identify single-word terms and terms that belong to different parts of speech.

¹<https://www.valital.com>

3. Related Work

Different methods were devised to identify terminology and such methods are now well-established and used for different applications (Indurkha and Damerou, 2010). Automatic term extraction (ATE) methods are usually categorized as linguistic, statistical or hybrid. The first techniques rely on linguistic descriptions (grammars, dictionaries, surface patterns), while statistical methods rely on information like frequency and co-occurrence, etc. In the last 20 years, most tools use both statistical and linguistic information and fall into the hybrid category. The tools try to evaluate how interesting items extracted are for terminologists, leading to various methods for calculating *termhood* (Kageura and Umno, 1996). Among the three traditional categories, hybrid methods were evaluated as those that led to better results (Macken et al., 2013). But in the last few years, the research field of ATE has undergone profound changes. Progress in machine learning and more specifically in deep learning has led to methodologies which cannot be easily described using the three traditional categories (Rigouts-Terryn et al., 2020). In this work, we will explore a traditional hybrid method that compares corpora and combine it with more recent techniques such as word embeddings (Mikolov et al., 2013; Amjadian et al., 2016; Kucza et al., 2018; Qasemizadeh and Handschuh, 2014; Pollak et al., 2019). Our work is similar to (Hätty et al., 2019) as far as the method is concerned. However, our aim is to identify terms in unspecialized corpora. Given this, we cannot only target changes in meaning or reduction of number of attested meanings in a specialized corpus when compared to a general one. We take the opposite approach and attempt to spot potential meaning similarities to remove candidates that would be very similar regardless of the corpora.

An efficient method for ATE consists of comparing a domain-specific corpus (an analysis corpus, AC) to a general one (a reference corpus, RC) and computing a specificity score for lemmas. For instance, a corpus of English texts dealing with the topic of climate change can be compared to a general balanced corpus such as the British National Corpus. This method was implemented in TermoStat described in (Drouin, 2003). It was evaluated for the extraction of single-word terms with satisfactory results (Lemay et al., 2005) and supports multiple languages². The concept of “specificity” aims to capture the potential of term candidates to behave like terms (termhood). In most cases, termhood is linked to a higher than expected frequency in a specialized corpus based on a theoretical frequency computed from a general corpus. Various statistical measures can be used to compute specificity.

When comparing corpora of different genres, terms ranking high retrieved from the AC usually correspond to terms. When the analysis corpus is less specialized (even if its content is topic-specific), it is to be expected that the strong opposition between corpora is lost. We can no longer focus on the single assumption that there is a high level of divergence in the way words behave in the AC and the RC as

²<http://termostat.ling.umontreal.ca>, (Drouin, 2020)

in (Hätty et al., 2019). This work addresses this problem and suggests a method that could still make the most of the terminological content of the AC even if it belongs to a text genre that is the same or very similar to that of the RC.

4. Hypotheses

In this paper, we are dealing with corpora that belong to the same genre even though one of the corpora covers a broader spectrum of topics. Our hypotheses are:

- A traditional approach to ATE based on frequency comparison can still be used to locate relevant terminology. In other words, **the dissimilarity between the topics of the two corpora can still be exploited by an automatic term extraction method (Hypothesis 1)**.
- However, given the fact that textual genres are quite similar, it is likely that a number of tokens will need to be filtered (probably more than usual). One strategy consists in using some of the features shared by both corpora to further refine term extraction. We can exploit the fact that some words have a similar behavior in the two corpora and use this feature to filter out the results obtained by simple corpus comparison. This method is likely to increase precision. However, in order to capture this behavior, we need to go beyond frequency measures and model semantic features in some way, e.g. using distributional information and word embeddings. Thus, **the similarities between the corpora are also useful and can be exploited with distributional analysis and word embeddings (Hypothesis 2)**.

The main idea behind (1) is that, since our AC is limited to one topic, specificity can be used to retrieve term candidates (TC). In contrast, since both the AC and the RC are comparable from a text genre point of view, in (2) we want to capture the fact that some items that might be retrieved by the specificity carry *meanings* that do not contrast sharply with the ones they convey in general language corpora. In order to do so, we will compare word embeddings built from our AC and freely available prebuilt embeddings. This comparison will be used to filter out the results obtained based on (1).

5. Methodology

The overall process is illustrated in Figure 1. The regular approach to term extraction when comparing corpora is represented by the stages in light blue. The analysis and reference corpora are preprocessed; term extraction is performed using specificity scores; finally, term candidates are ranked according to the score they obtained. We are adding a layer (steps in light yellow) designed to compare word embeddings in order to re-rank the output produced by steps in blue.

5.1. Corpora

Basic preprocessing is applied to both the AC and the RC. All files from the corpora are tokenized, then tagged and lemmatized using TreeTagger (Schmid, 1994). The TreeTagger format is used as a common input for subsequent tasks.

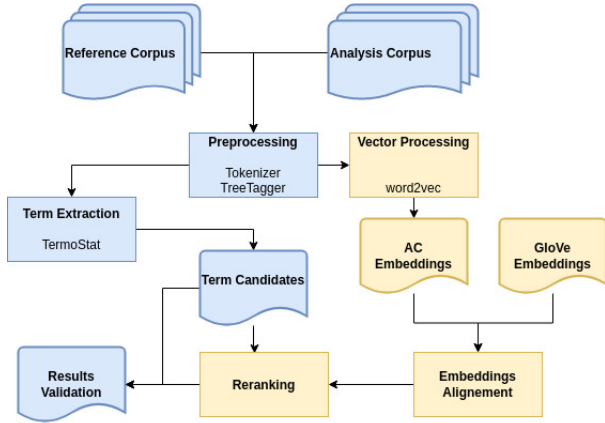


Figure 1: Overview of the process

5.1.1. Analysis Corpus

The corpus that was built by our partner comprises several text documents dealing with unwanted behavior from potential employees: *Addiction*, *Discrimination*, *Fraud*, *Harassment*, and *Violence*. It is important to mention that all files in the corpus were retrieved automatically from the web based on a short list of keywords related to each of these topics. All files come from online Canadian English newspapers and have been preprocessed to remove HTML markup. Since the crawling process was keyword based, the various corpora are noisy and thus do not lend themselves easily to standard term extraction process. In this work, we will focus solely on the Discrimination corpus as work on this topic is more advanced than for the other topics. The corpus contains 1,541,987 tokens.

5.1.2. Reference Corpus

The reference corpus used was built from subsets of two large corpora: the British National Corpus (BNC) (Consortium, 2007) and the American National Corpus (ANC) (Reppen et al., 2005). We extracted 4M tokens from each of these corpora in order to compile our 8M tokens reference corpus. In both cases, only newspaper texts were retrieved.

5.2. Term Extraction

The extraction process was limited to single-word lexical items including nouns, verbs and adjectives, since, as was mentioned above, important concepts in this field can be expressed with terms that belong to different parts of speech. TermoStat computes a *Specificity* score to represent how far the frequency in the specialized corpus deviates from a theoretical frequency. Its calculation relies on an approximation of the binomial distribution using standard normal distribution. In order to do so, a measure proposed by Lafon (1980) is used.

Using values from Table 1, specificity can be calculated as follows:

$$\log P(X=b) = \log(a+b)! + \log(N-(a+b))! + \log(b+d)! + \log(N-(b+d))! - \log N! - \log b! - \log((a+b)-b)! - \log((b+d)-b)! - \log(N-(a+b)-(b+d)+b)!$$

This measure was tested in previous studies (Lemay et al., 2005; Drouin and Langlais, 2006; Drouin, 2006; Drouin

	Reference Corpus	Specialized Corpus	Total
Freq. term	a	b	a+b
Freq. of other words	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

Table 1: Contingency table of frequencies

and Doll, 2008; Drouin et al., 2018) and leads to excellent results for the extraction of both single-word and multi-word terms. Specificity can be used to spot items that are both over- and under-represented in a corpus. In the case of terminology, a domain- and genre-oriented lexicon, we are solely interested in positive specificities which highlight items that are over-represented in the AC.

Since the specificity scores cannot be represented on a pre-defined scale, for the current experiment, we expressed them on a scale ranging from 0 to 1 where the max specificity score is mapped to 1. This mapping, which does not impact the overall distribution of scores, leads to a less granular representation of the scores and a more flexible set of scores to assess. The specificity score is used to test hypothesis 1.

5.3. Embeddings

5.3.1. Computed Word Embeddings

To build embeddings for our AC, we used the word2vec (Mikolov et al., 2013) implementation included in Gensim (Řehůřek and Sojka, 2010). We used default values for the skipgram algorithm with a window of 5 words, a minimum frequency threshold of 5 and 300 dimensions for the vectors.

5.3.2. Pre-trained Word Embeddings

To compare the behavior of tokens in a large unspecialized language corpus, we used the pre-trained word GloVe embeddings (Pennington et al., 2014). More specifically, we used the Common Crawl embeddings built from 42B tokens with a 1.9M vocabulary (uncased) and. The embeddings' vectors have 300 dimensions.

5.3.3. Alignment of Word Embeddings

Since our embeddings and the GloVe embeddings are built from different corpora and we want to be able to compare the vectors for words in both of them, the embeddings must be aligned. In order to do this, we used the technique proposed by (Hamilton et al., 2016) based on the code provided by Tyler Angert and available from his GitHub³. Such an approach is been used in (Hätty et al., 2019) to compare vectors between corpora for term extraction. During the alignment process, only the shared vocabulary⁴ between embeddings is kept.

³<https://gist.github.com/tangert/106822a0f56f8308db3f1d77be2c7942>

⁴By *shared vocabulary*, we mean words that are common to both embeddings.

5.4. DistSpecificity

Words with similar behaviors in a large unspecialized corpus (Glove embeddings in our case) and our AC (our corpus built embeddings) are assumed to carry the same *meanings* based on the distributional features/patterns captured by the embeddings. From this idea we can use a simple cosine distance to compare vectors. Similar vectors will lead to cosine distance closer to 0 and dissimilar vectors to values closer to 1. We represent the distance using a score called GloveDist.

What is of interest to us is to lower the Specificity score for TCs whose distributional behavior is the same in both corpora. The rationale behind this strategy is that even though the Specificity score seems to indicate that TCs are valid terms, their overall *meaning* is the same. We thus factor this information in a new score called DistSpecificity which is used to test our hypothesis 2.

$$\text{DistSpecificity} = \text{GloveDist} * \text{Specificity}$$

Using this score, the Specificity score of a very specific TC that has almost the same distributional behavior in both corpora will be closer to 0 (since GloveDist will tend towards 0). On the other hand, a dissimilar behavior in both corpora will not impact Specificity as such (since GloveDist will have a value closer to 1).

5.5. Validation

All results were manually validated by a terminologist who has been involved in the project from the start. For the purpose of the current experiment, we are mainly interested in the potential of our score to rank valid terms at the top of the list of term candidates. Our manual validation was limited to the first 250 TCs retrieved using each of our three scores (Specificity, GloveDist and DistSpecificity) ranked from the highest to the lowest value. We thus validated a total of 750 TCs. As can be seen in Table 3, some TCs could appear in two or three lists.

The criteria used for the validation of TC were the following:

1. Terms must appear in contexts that are *meaningful* according to our task;
2. Terms must appear in at least 10 knowledge-rich contexts (KRC) (Meyer, 2001) related to *discrimination*;
3. TCs can also be considered terms if they hold syntagmatic or paradigmatic relations (e.g., as synonymy, antonymy or argumental) with already validated terms. (L’Homme, 2020).

What we define as a *meaningful* context (Criteria 1) is a context in which a misconduct is described. Even though some TCs could appear in an important number of contexts, we selected to base our study on KRCs only (Criteria 2). This methodological decision makes our validation process more challenging but our results more interesting.

The following sentence is a good example of a KRC for TCs such as *race* or *religion*: *In New York State, we have no tolerance for discrimination on the basis of race, religion, sex, national origin, disability, sexual orientation or perceived sexual orientation.* KRCs provide insights on how

TCs can be linked to each other in a specific domain. In this KRC, it shows us how *race* and *religion* can be linked to *discrimination* (also a TC) in our domain.

In addition to meeting the above-mentioned criteria, some TCs were also validated according to Criteria 3. For example, *anti-discriminatory* was labelled as a term on the basis of being an antonym for *discriminatory*; *woman* on the basis of being an argument of verbs such as *discriminate* or predicative nouns such as *discrimination*. Both TCs meet the other criteria as well.

The validation process was challenging due to the fact that often TCs did not convey a very technical meaning in the AC, i.e. a meaning that one could easily distinguish from general usage. Our approach was to consider TCs that were relevant according to the topic of discrimination and this “relevance” was constantly refined as we skimmed through the list of candidates.

TCs that met the previous criteria were labelled as *Term*; TCs that did not meet these criteria as *Non-Term*; and TCs that we had doubts about as *Uncertain* (see Tables 5 to 7).

6. Results and Discussion

As can be seen from the precision values in Table 2, ATE on unspecialized corpora is not a trivial process. We provide two precision measures for each score. Precision¹ is obtained by dividing the total number of valid TCs by 250 (the total in our lists) while Precision² corresponds to the number of valid TCs evaluated on the set of TC that we could validate (ignoring the TCs classified as Uncertain from the calculation). Values obtained by both measures are quite low, but not to the point of making the ATE extraction useless. Recall was not evaluated for this experiment since we do not have a gold standard that can be used and a manual evaluation of recall on newspaper corpora does not serve a larger purpose for the time being. The main issue with a task like the one we describe in this paper is still reaching acceptable precision values.

Score	Specificity	GloveDist	DistSpecificity
Term	145	106	135
Non-Term	87	128	97
Uncertain	18	16	18
Total	250	250	250
Precision ¹	0.58	0.42	0.54
Precision ²	0.63	0.45	0.58

Table 2: Precision values for all 3 scores

Since we are more interested in the potential of each score to rank the valid information at the top of the list presented to the terminologist, we can evaluate precision at each position in the TC lists. This information is provided in Figure 2 which shows the precision values obtained by the three scores (Specificity, GloveDist and DistSpecificity) over the whole list. For these scores, entries identified as Uncertain were considered as errors, we are thus using the Precision¹. We can easily see that GloveDist does not perform as expected. This means that using solely distributional information from a large unspecialized corpus as captured by GloVe embeddings and comparing them to our local vectors

is not sufficient in itself. The distance between the vectors does not allow us to distinguish Terms from Non-Terms. Specificity presents a somewhat stable curve which means that valid TCs are distributed evenly along the list of 250 TCs. These results show that Specificity remains an interesting score to identify potential terms in unspecialized corpora by comparing them to larger unspecialized corpus. On the other hand, Figure 2 shows that it is not the best score to maximize valid TCs at the top of list. As mentioned earlier, our DistSpecificity score combines both Specificity and GloveDist, the idea being to lower the importance of TCs that have a high Specificity but a similar behavior in both our corpora and the corpus used to build the GloVe embeddings. Figure 2 shows that this seems to be the case as precision values for DistSpecificity are higher for an important part of the list of TCs (until we reach candidate 165).

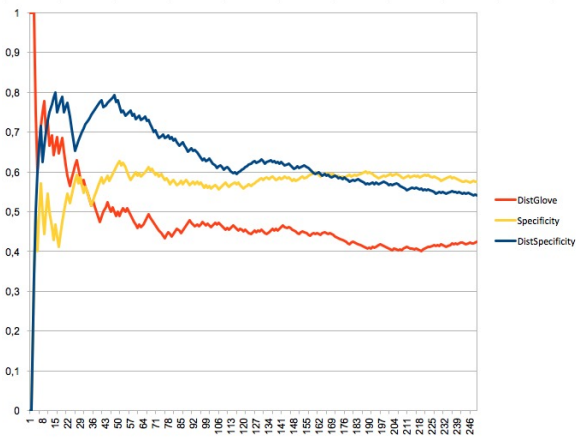


Figure 2: Overall precision of the scores

Table 3 details the contributions of each score. It shows that they share 60 common terms while bringing unique contributions to the overall list of TCs. However, Specificity locates more valid terms than GloveDist.

	Specificity	DistSpecificity
Common	60	60
Unique	84	75
Total	144	135

Table 3: Overall Contribution of Scores for Valid Terms

	Specificity	DistSpecificity
Common	50	50
Unique	47	47
Total	97	97

Table 4: Contribution of Scores for Valid Terms < 165

Nearly a third (30%) of the top 165 candidates are common to both scores, the top 15 can be seen in Table 5. One can clearly see by looking at the *Non-Terms* that the nature of the corpus had an impact on the results. For example, items

Spec	Status	DistSpecificity	Status
ms	Non-Term	ms	Non-Term
law	Term	read	Non-Term
woman	Term	employee	Term
newsletter	Non-Term	white	Term
lawsuit	Term	state	Term
story	Non-Term	law	Term
court	Term	file	Term
employee	Term	hide	Non-Term
subscribe	Non-Term	court	Term
photo	Non-Term	bill	Term
read	Non-Term	case	Term
state	Term	lawyer	Term
case	Term	justice	Term
lawyer	Term	complaint	Term
plaintiff	Term	religion	Term

Table 5: Top common TCs

such as *ms*, *newsletter*, *subscribe*, *hide* can be attributed to the fact that the corpus was built from Web pages. The results at the top of the list for DistSpecificity are much better and contain terms relevant to the task at hand.

Spec	Status	DistSpecificity	Status
discrimination	Term	dismissal	Term
gender	Term	argument	Term
percent	Non-Term	argue	Term
update	Non-Term	politics	Term
advertisement	Non-Term	contend	Term
transgender	Term	person	uncertain
discriminate	Term	epithet	Term
right	Term	man	uncertain
emails	Non-Term	retaliate	Term
program	Non-Term	advertiser	Non-Term
verify	Non-Term	caste	Term
robot	Non-Term	city	Non-Term
minority	Term	engage	Term
sex	Term	request	Non-Term
disability	Term	resign	Term
hire	Term	asylum	Non-Term
racism	Term	noose	Term
ruling	Term	dissent	Non-Term
view	Non-Term	analyze	Non-Term
neighborhood	uncertain	officer	Non-Term

Table 6: Top unique TCs

Table 6 shows some of the unique contributions of the scores. Once again in this context we can observe the influence of the nature of the corpus on the TCs retained: *advertisement*, *robot*, *view*, *request*, *verify*, etc.. Such TCs were again more present in the first TCs proposed by the Specificity score which means that DistSpecificity was, to some extent, successful in re-ranking them.

Table 7 contains the TCs that were most positively affected by the re-ranking. Although some results can be explained by the content of the documents that make up the corpus

Term	Status	Delta
flag	Non-Term	+184
buraku	Term	+180
enact	Term	+179
try	Non-Term	+159
be	Non-Term	+155
harass	Term	+149
department	Non-Term	+146
defendant	Term	+144
university	Non-Term	+143
accuse	Term	+142
gap	Term	+135
legislation	Term	+132
allege	Term	+123
ethnicity	Term	+123
remark	Term	+122

Table 7: Top positive re-ranking of Specificity by Dist-Specificity

(*university, department*), some are quite puzzling (*try, be*) and need to be investigated further. Since the AC is made of newspaper articles, academics who study the phenomenon of *discrimination* are often quoted and it explains the strong presence of the former in our corpus. However, it does not explain why their distributional features are so different in the two corpora. This will also be subjected to further investigation.

Term	Status	Delta
advocate	Term	-59
orientation	Term	-63
bias	Term	-71
bar	Term	-73
hate	Term	-73
subscribe	Non-Term	-75
promotion	Non-Term	-76
segregation	Term	-78
newsletter	Non-Term	-78
behavior	Term	-87
administration	Uncertain	-114
lawsuit	Term	-138
housing	Non-Term	-149
suit	Term	-165
click	Non-Term	-189

Table 8: Top negative re-ranking of Specificity by Dist-Specificity

At the other end of the spectrum are the results contained in Table 8 which include the TCs that have been negatively re-ranked by DistSpecificity. As we mentioned earlier, the good news is that this score is able to capture the fact that some TCs that are more closely related to the Web than the subject matter of the corpus and lower their termhood lowered (*click, promotion, etc.*). Nonetheless, some valid terms are being affected quite strongly while they should not be (*advocate, orientation, bias, bar*). In some cases

(*bar, hate, orientation*), it seems that polysemy can be a factor affecting the quality of the results.

7. Future Work

All experiment results were evaluated by a single terminologist and limited to the first 250 TCs provided by each score. Working with a larger sample and a team of validators would allow us to test inter-annotator agreement over a larger sample.

For the current task we limited our investigation to single-word TCs. We believe our findings could be applied to multiword TCs in order to see if we can corroborate the results obtained here. In order to do so, we would need to conduct an experiment using word embeddings that can capture distributional information from multiword TCs. Relying on more recent (and more complex) embeddings algorithms would also help to capture contexts in which TCs are used and perhaps mitigate the effects of polysemy observed in our results.

An interesting extension of the method presented here would be to apply it other genres of unspecialized corpora such as texts retrieved from social media. Some social platforms such as Twitter and Reddit host communities of specialists. These specialists exchange knowledge in informal settings and the terms carrying this knowledge should be described.

Provided that our results can be replicated in larger settings, integrating our method into the compilation process of terminological resources and into our term extraction tool would be beneficial.

8. Conclusion

In this paper we proposed a method for extracting terminology from unspecialized corpora. Our first hypothesis was that traditional corpus comparison techniques could be used in such a task in order to capture the dissimilarity between the topics of the two corpora. We have verified that this is the case and that the results of such a technique could still be used in terminology work although they are noisy. Our second hypothesis was that the similarities between the corpora are also useful and can be exploited with distributional analysis and word embeddings. To test our second hypothesis we devised a new way to re-rank TCs provided by a classic corpus comparison method in an effort to compare distributional features of TCs in our unspecialized corpus to those observed in a general langue corpus. Using this technique leads to very good results, as far as we could tell from the first part of a list of candidate terms. For terminologists, this method would allow them to focus on more relevant terms.

9. Acknowledgements

The authors would like to thank Valital for working with our team on this project and for the corpora they provided as well as IVADO (Institut de valorisation des données)⁵ for making this possible through its financial support.

⁵<https://ivado.ca/>

10. Bibliographical References

- Amjadian, E., Inkpen, D., Paribakht, T., and Faez, F. (2016). Local-Global Vectors to Improve Unigram Terminology Extraction. In *Proceedings of the 5th International Workshop on Computational Terminology*, pages 2–11, Osaka, Japan.
- Drouin, P. and Doll, F. (2008). Quantifying termhood through corpus comparison. In *Terminology and Knowledge Engineering (TKE-2008)*, pages 191–206, Copenhagen, Denmark, Août. Copenhagen Business School, Copenhagen, Copenhagen Business School, Copenhagen.
- Drouin, P. and Langlais, P. (2006). Évaluation du potentiel terminologique de candidats termes. In *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles. (JADT 2006)*, pages 389–400, Besançon, France.
- Drouin, P., L'Homme, M.-C., and Robichaud, B. (2018). Lexical profiling of environmental corpora. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3419–3425, Paris, France, May. European Language Resources Association (ELRA).
- Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.
- Drouin, P. (2006). Termhood experiments: quantifying the relevance of candidate terms. *Modern Approaches to Terminological Theories and Applications*, 36:375–391.
- Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proc. Assoc. Comput. Ling. (ACL)*.
- Hätty, A., Schlechtweg, D., and Schulte im Walde, S. (2019). SUREl: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 1–8, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Indurkha, N. and Damerau, F. (2010). *Handbook of Natural Language Processing, Second Edition*. Chapman & Hall/CRC machine learning & pattern recognition series. CRC Press.
- Kageura, K. and Umino, B. (1996). Methods of automatic term recognition. *Terminology*, 3(2):259–289.
- Kucza, M., Niehues, J., Zenkel, T., Waibel, A., and Stüker, S. (2018). Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks. In *Interspeech 2018*, pages 2072–2076, Hyderabad, India, September. ISCA.
- Lemay, C., L'Homme, M.-C., and Drouin, P. (2005). Two methods for extracting specific single-word terms from specialized corpora: Experimentation and evaluation. *International Journal of Corpus Linguistics*, 10(2):227–255.
- L'Homme, M.-C. (2020). *Lexical Semantics for Terminology: An introduction*. John Benjamins.
- Macken, L., Lefever, E., and Hoste, V. (2013). TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment. *Terminology*, 19(1):1–30.
- Meyer, I. (2001). Extracting knowledge-rich contexts for terminography. In Didier Bourigault, et al., editors, *Recent advances in computational terminology*, page 279. John Benjamins.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Pollak, S., Repar, A., Martinc, M., and Podpečan, V. (2019). Karst Exploration: Extracting Terms and Definitions from Karst Domain Corpus. In *Proceedings of eLex 2019*, pages 934–956, Sintra, Portugal.
- Qasemizadeh, B. and Handschuh, S. (2014). Investigating Context Parameters in Technology Term Recognition. In *Proceedings of SADAATL 2014*, pages 1–10, Dublin, Ireland.
- Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. <http://is.muni.cz/publication/884893/en>.
- Rigouts-Terryn, A., Drouin, P., Hoste, V., and Lefever, E. (2020). Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. In *Proceedings of Computerm 2020*.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

11. Language Resource References

- BNC Consortium. (2007). *British National Corpus, version 3 BNC XML edition*. British National Corpus Consortium, ISLRN 143-765-223-127-3.
- Drouin, Patrick. (2020). *TermoStat 3.0*. <http://termostat.ling.umontreal.ca>.
- Reppen, Randi and Ide, Nancy and Suderman, Keith. (2005). *American National Corpus (ANC) Second Release*. Linguistic Data Consortium, ISLRN 797-978-576-065-6.