Proceedings of the TERM21 Workshop , pages 26–29
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

Terminology extraction using co-occurrence patterns as predictors of semantic relevance

Rogelio Nazar, David Lindemann
Instituto de Literatura y Ciencias del Lenguaje, Pontificia Universidad Católica de Valparaíso
Faculty of Arts, UPV/EHU University of the Basque Country
rogelio.nazar@pucv.cl, david.lindemann@ehu.eus

Abstract

We propose a method for automatic term extraction based on a statistical measure that ranks term candidates according to their semantic relevance to a specialised domain. As a measure of relevance we use term co-occurrence, defined as the repeated instantiation of two terms in the same sentences, in indifferent order and at variable distances. In this way, term candidates are ranked higher if they show a tendency to co-occur with a selected group of other units, as opposed to those showing more uniform distributions. No external resources are needed for the application of the method, but performance improves when provided with a pre-existing term list. We present results of the application of this method to a Spanish-English Linguistics corpus, and the evaluation compares favourably with a standard method based on reference corpora.

Keywords: terminology extraction, co-occurrence patterns, semantic relevance

1. Introduction

In this short paper, we present a methodological proposal for automatic terminology extraction (ATE), which forms part of a larger terminology software project, currently in development, aimed at the automation of different tasks of glossary creation. Here we explain therefore only the task of creating the list of entries for the glossary by means of term extraction from a specialised corpus. With this goal in mind, we experimented with the application of a co-occurring measure, which we used as a means to operationalise a key concept of the ATE problem such as semantic relevance.

Using word co-occurrence as indicator of semantic relevance is something that has been tried in the past for different terminology related applications (Nazar et al., 2007; Wartena et al., 2010). An earlier attempt to use this type of measures in an ATE system was Termout[1] (Nazar, 2016), which proved effective as a method to extract terms from a single document but too computationally expensive to analyse a full corpus, making it impractical in environments like web applications.

On this occasion, we further explore these co-occurrence measures and present a significant improvement. This new method is simple, computationally efficient and scalable: after a classical workflow involving the filtering of single and multi-word units based on syntactic patterns, the central idea is to promote candidates that show a particular profile of co-occurrence, i.e., a tendency to appear with a selected number of other lexical units in the same sentences. This is regardless of the order of appearance of the terms, as well as their relative distance, as in the case of the terms *signifier* and *signified* in the field of Linguistics. We observe that when a candidate has a persistent group of 'friends', it usually is a specialised term, as opposed to those showing more uniform distributions.

The co-occurrence information is obtained from the same specialised corpus, and for this reason, a minimum corpus size is required (ca. 5 million tokens). Apart from a POS-tagger, no external resources are needed. But if a list of terms of the domain is already available, then it can be used to improve performance by identifying its members among the co-occurring words of a given candidate.

We present results of the application of the method to a Spanish-English linguistic corpus, in which evaluation figures compare favourably with a standard method based on reference corpora. More data is available on the project's website[2].

2. Related Work

The field of terminology has always been intrinsically related with that of computational linguistics because of the variety of natural language processing tools and methods that can be applied to at least partially automatise the terminology workflow and the process of dictionary creation (Sager, 1990). ATE, however, was consolidated as a particular field of research after Kageura and Umino's survey (1996), where the authors defined the task of separating the terms from the rest of the vocabulary of a specialised corpus. They also presented the main approaches (i.e., based on statistical or on linguistic knowledge) and explained the procedure for evaluation, which continues to be the standard today.

Different methods have been proposed in the span of several decades, but no consensus has yet been reached concerning which one is preferable, since different methods show a better performance than others, depending on the use case. The lack of a standard evaluation dataset is one of the main difficulties for evalu-

[1] http://www.termout.org

[2] http://www.tecling.com/cgi-bin/termout/ling

ating ATE methods (Astrakhantsev, 2017; Zhang et al., 2017).

Overall, certain tendencies can appreciated in the history of this field. Earlier methods began to explore statistics of term distribution. The work of Spärk-Jones (1972) in Information Retrieval is often credited as a trailblazer in ATE, as she proposed an algorithm to promote term candidates that show concentrated frequency in fewer documents of a corpus. An earlier study by Juilland and Chang-Rodriguez (1964) also deserves mention, as they too were looking at how lexical units are distributed in a corpus in order to separate terms from general vocabulary.

Later models, in the eighties and nineties, involved a greater degree of linguistic sophistication, with the application of morphosyntactic patterns for the correct segmentation of multiword units (Justeson and Katz, 1995). They observed that multiword terms most often occur as certain types of noun phrases (e.g., noun, adjective-noun, noun-preposition-noun).

In parallel, with the rise of Corpus Linguistics in the British lexicographic tradition (Sinclair, 1991), the concept of 'keyness' or 'keywordness' began to develop, according to which lexical units are weighted using large reference corpora of non-specialised discourse. Keywords are defined as those that occur relatively more often in the domain-specific target corpus than would be expected in comparison with a reference corpus that represent general or every-day language. Functions to extract keywords were then offered by classical corpus linguistics software such as Wordsmith Tools (Scott, 1997) or AntConc (Anthony, 2005). Later term extraction systems were also inspired by this approach, such as Termostat (Drouin, 2003), and Sketch Engine (Kilgarriff et al., 2014), and others using similar notions such as term 'weirdness' (Ahmad et al., 1994).

By the turn of the century, surveys show a progressive hybridisation of methodologies, involving both statistical and linguistic data (Cabré et al., 2001). More recently, however, a new tendency seems to be gaining ground, one that takes into account contextual features and distributional semantics. TerMine (Frantzi et al., 2000) is an earlier example of ATE method that uses some form of contextual features. Its 'C-/NC-value' combines statistical measures and distributional information. The common statistical measure is improved in the sense that it adjusts frequency values of single or multiword terms that also occur as part of longer multiword terms (C measure), while information about words that tend to appear next to term candidates is also taken into account (NC measure).

However, it is in more recent approaches where the semantic component is most evident. Some researchers are introducing semantic relatedness of term candidates as a measure in addition to a combination of methods based on statistics (frequencies) and linguistics (lexico-syntactic patterns, distributional information). For instance, 'KeyConceptsRelatedness' (Astrakhantsev, 2014) is the semantic relatedness of candidates to already validated domain terms, where semantic relatedness is computed according to a word embedding model trained on Wikipedia text. Similar work relies on lexico-semantic knowledge represented in semantic networks and ontologies, as shown in the survey by Maynard et al. (2008). In this line, Zhang et al. (2017) propose a generic method for enhancing ATE results, using a small set of validated seed terms to compute the distributional similarity against term candidates.

Our present proposal can be considered similar to this later trend, as it uses co-occurrence to operationalise semantic relevance.

3. Method

As usual in ATE projects, this method begins with the selection of a language and a domain of interest. As we were already embarked in a project to develop a large Spanish-English Linguistics glossary, we decided to test our method with a linguistics corpus. To this end, we used all the articles published in the last 25 years by *Revista Signos*[3], an open-access linguistics journal that accepts papers in both languages. This constitutes a corpus of 602 papers with a total of approximately 6.5 million tokens.

We developed a pipeline to download the papers and convert them from their original HTML format to plain text. As usual in some academic journals, the papers have bilingual titles, abstracts and keywords. They also often mix reference titles mainly in both of these languages. In this paper we set up the ATE task to be applied on monolingual corpora. At a later stage, we will exploit the fact that it is a pseudo-parallel corpus in order to align the extracted terms, but as we said, we leave those details for a future paper. For the present stage, we opted to separate the corpus in both languages and apply the method one language at a time. This separation is done automatically with Linguini[4], a Perl script that detects the main language of every text in a corpus and then deletes any fragments in other languages found inside each text. This is relevant in our use case, since also the text bodies frequently contain e.g. quotes and examples in another language.

As is normal in this type of workflows, the next step in the pre-processing the corpus consists of the application of a POS-tagger. In our case, we used UDPipe (Straka and Straková, 2017) because of the quality of its lemmatisation and POS-tagging. It also offers full syntactic parsing, and some authors have suggested the use of this type of parsers in order to better segment multiword terminology (Judea et al., 2014). However, we opted for a more conservative approach, and ignored the syntactic annotation. Instead, we defined a list of morphosyntactic patterns typical of multiword terminology, such as noun-noun or adjective-noun (e.g., *corpus linguistics, specific language im-*

[3]http://www.revistasignos.cl

[4]http://www.tecling.com/linguini

pairment) or constructions with certain propositions (e.g., in Spanish, *lingüística de corpus*). This is undoubtedly an oversimplification of the problem because morphosyntactic patterns found in multiword terminology can be extremely diverse, and this will have to be addressed in future work.

The previous step results in a first unrefined list of term candidates. Next, the algorithm extracts the contexts of occurrence of each candidate in the specialised corpus. The intuition is that genuine terms of the domain will show a particular profile of co-occurrence, as indicative of how informative they are. This can be seen, for instance, in Figure 1, which depicts this type of analysis for the case of the term *second language acquisition*. In this case, we can see a characteristic shape of the co-occurrence frequency curve, showing that there is a limited number of vocabulary units that appear with a significant frequency in the same sentences. One can notice, among the most frequent co-occurring units, some words and parts of terms and proper names that are semantically related to the candidate (e.g. *learning, feedback, corrective*).

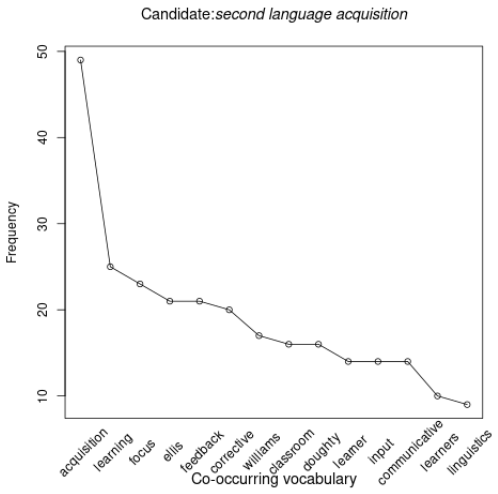

Figure 1: Co-occurrence profile of candidate *second language acquisition* in the corpus.

In order to account for this phenomenon as a predictor of terminology, we developed a co-occurrence measure (1) that will promote a candidate based on its co-occurrence frequency curve.

$$I(x) = \frac{\log_2 \sum_{i=1}^{n} R_{x,i}}{\log_2 |m(x)|} \quad (1)$$

Here, x represents some term candidate; R_x is the set of co-occurring words; $m(x)$ is the set of contexts of occurrence of x and $R_{x,i}$ is the frequency of occurrence of a word in the ith position of the n most frequent words in those contexts. The parameter n is arbitrary, and we set it to 20 in our experiments. Larger values would imply longer processing times.

Another arbitrary parameter would be a threshold k, used by a binary function ATE(x) (2) if one needs to accept or reject each candidate. Alternatively, one can rank all candidates in a list according to (1).

$$ATE(x) = \begin{cases} 1 & I(x) > k \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

As a final note for the explanation of the methodology, we add that the sensitivity of the term detection can be amplified with the use of a pre-existent list of terminological units. If a user can provide a large list of terms as examples (ca. 2000), then the algorithm can calculate the intersection between such list and the vocabulary co-occurring with a candidate. Of course, this is then used to promote a candidate as a relevant term, but it can also be used to narrow down the selection according to the interest of the researcher, e.g. to extract term candidates for the enrichment of a vocabulary of the domain of Lexicography rather than Linguistics.

4. Results and Evaluation

After processing the corpus, the algorithm first obtained a list of approximately 46,000 different noun phrases with term-like morphosyntactic patterns, and then ranked them according to the co-occurrence measure. We only considered as a result the best 4,000 candidates of the list, and we conducted a manual evaluation of the first and the last 500 rank positions. The error rates obtained were 23% and 48%, respectively, with 84% inter-coder agreement.

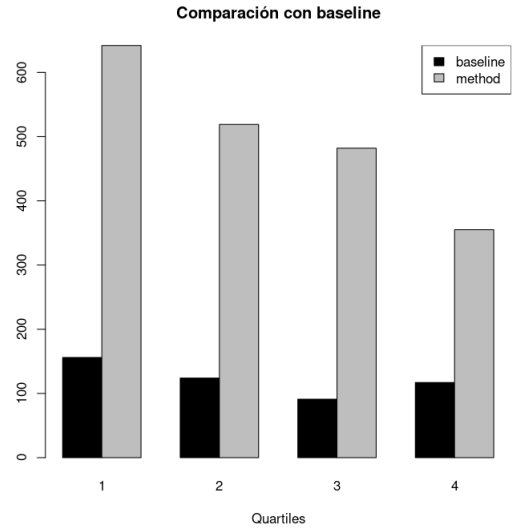

Figure 2: Contrast between method and baseline by the intersection of terms with Gold Standard.

Also, as a baseline we used Sketch Engine's term extraction function (Kilgarriff et al., 2014), as it represents a classical approach based on reference corpora (cf. Section 2). We submitted the same corpus,

and again considered only the best 4,000 term candidates. To automatise the comparison, we used as gold-standard a manually curated list of circa 3,500 linguistics terms. Figure 2 shows the comparison with the baseline in the number of matches with the gold standard. The first quartile corresponds to the best 1000 candidates. As can be seen, the matching is significantly higher than the baseline in each quartile, and then it decreases non-randomly, meaning that the ranking is effective.

5. Conclusions

In this paper we proposed an ATE method and described its results on a Spanish-English linguistics corpus. The method is relatively simple, it is computationally efficient and the evaluation shows promising results.

In future work we will be describing subsequent steps to further improve the quality of results. We already mentioned some of these steps, like a better segmentation of multiword terms. But we also discovered other simple strategies which have a significant impact, such as promoting candidates that appear in bibliographic references. We are also working on how to automatise other operations such as filling in fields of a terminological database, such as equivalences in another language, morphological categories, inflected forms, related terms, definitions, and others.

6. Bibliographical References

Ahmad, K., Davies, A., Fulford, H., and Rogers, M. (1994). What is a term?: The semi-automatic extraction of terms from text. In Mary Snell-Hornby, et al., editors, *Benjamins Translation Library*, volume 2, page 267. John Benjamins, Amsterdam.

Anthony, L. (2005). Antconc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *IPCC 2005. Proceedings. International Professional Communication Conference*, pages 729–737.

Astrakhantsev, N. (2014). Automatic term acquisition from domain-specific text collection by using Wikipedia. *Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS)*, 26(4):7–20.

Astrakhantsev, N. (2017). ATR4S: toolkit with state-of-the-art automatic terms recognition methods in Scala. *Language Resources and Evaluation*, 52(3):853–872.

Cabré, M. T., Estopà, R., and Vivaldi, J. (2001). Automatic term detection: A review of current systems. In Didier Bourigault, et al., editors, *Natural Language Processing*, volume 2, pages 53–87.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Judea, A., Schütze, H., and Bruegmann, S. (2014). Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 290–300, Dublin, Ireland, August.

Juilland, A. and Chang-Rodriguez, E. (1964). *Frequency Dictionary of Spanish Words*. De Gruyter.

Justeson, J. S. and Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Kageura, K. and Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, 3(1):259–289.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The sketch engine: Ten years on. *Lexicography*, 1(1):7–36.

Maynard, D., Li, Y., and Peters, W. (2008). NLP Techniques for Term Extraction and Ontology Population. In Paul Buitelaar et al., editors, *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 107–127. IOS Press.

Nazar, R., Vivaldi, J., and Wanner, L. (2007). Towards quantitative concept analysis. *Procesamiento del Lenguaje Natural*, 39:139–46.

Nazar, R. (2016). Distributional analysis applied to terminology extraction: First results in the domain of psychiatry in Spanish. *Terminology*, 22(2):141–170.

Sager, J. C. (1990). *A Practical Course in Terminology Processing*. John Benjamins, Amsterdam.

Scott, M. (1997). The right word in the right place: Key word associates in two languages. *AAA: Arbeiten Aus Anglistik Und Amerikanistik*, 22(2):235–248.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada.

Wartena, C., Brussee, R., and Slakhorst, W. (2010). Keyword Extraction Using Word Co-occurrence. In *2010 Workshops on Database and Expert Systems Applications*, pages 54–58, August.

Zhang, Z., Gao, J., and Ciravegna, F. (2017). SemRe-Rank: Improving Automatic Term Extraction By Incorporating Semantic Relatedness With Personalised PageRank.