Isolating Terminology Layers in Complex Linguistic Environments: a Study About Waste Management (Short Paper)

Nicola Cirillo[1]

[1]*Università degli Studi di Salerno*

Abstract

Automatic term extraction aims at extracting terminological units from specialized corpora in order to assist terminolographers to develop lexicographic resources. In this paper, we introduce Domain Concept Relatedness, a novel term extraction technique meant to isolate the terminology of a given subject field. In order to evaluate our technique, we apply it to the extraction of waste management terms from a new Italian corpus about waste management legislation. We test it against Sketch Engine and the contrastive approach showing that our technique effectively extracts multi-word terms belonging to a given subject field but still fails to extract single-word terms.

Keywords

Terminology Extraction, Waste Management, Domain-knowledge, Specialized Languages

1. Introduction and related work

Automatic Term Extraction (ATE) is a natural language processing task. Its focus is the extraction of terms from specialized corpora. Typically, an ATE tool extracts all the terms that occur in a corpus disregarding their domain. While this approach is totally reasonable in most cases, it would be beneficial to distinguish terms related to different subject fields in some contexts. This demand has already been pointed out by [1] and [2]. They state that when ATE is applied to legislative documents, it becomes crucial to differentiate terms that belong to the regulated sector from legal terms but unfortunately only a few ATE methods address this task.

It is too simplistic to assume that all the terms contained in a specialized corpus belong to the same domain. In practice, many specialized languages include the terminology of multiple subject fields. For instance, the terminology of institutional languages comes from different domains of knowledge, namely law, administration, economy, and finance, plus all the technical terms of the regulated sectors [3]. In addition, each field can be divided into more specific sub-fields that can, in turn, be separated into smaller sub-fields, leading to a complex hierarchical model. We will refer to terms of a given subject field as a *Terminology Layer*, from now on TLR, regardless of its position in the hierarchy. Thus, the terminology of a specialized language has a hierarchical structure. Top-level layers contain terms related to broader domains, while

2nd International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2023, June 29–30, 2023, Lisbon, Portugal

nicirillo@unisa.it (N. Cirillo)

0000-0002-2107-1313 (N. Cirillo)

lower-level layers incorporate more topic-specific terms [4]. For example, a corpus of legislative documents about waste management contains at least three layers: the *law* TLR, the *waste management* TLR, and the *waste management law* TLR (which is an intersection of the first two). The first TLR contains terms such as *competent authority*, *member state*, and *regulation*, the second includes *incineration plant*, *separate collection*, and *landfill*, and the third comprises *competent authority of dispatch*, *country of destination*, and *list of wastes*.

We believe that ATE research has mostly overlooked the complex stratification held by terms. To date, research mainly focuses on the extraction of all the terms with ever-increasing accuracy, tackling ATE through several strategies. Standard methods couple linguistic approaches to simple corpus statistics [5, 6, 7, 8]. Other methods involve the comparison between a *focus corpus* (i.e. a specialized corpus) and a *reference corpus* (i.e. a general language corpus) [9, 10, 11]. It is also worth mentioning that deep learning approaches are becoming increasingly popular and obtain notable results [12, 13, 14]. Even though most approaches do not address the isolation of TLRs, there are at least two exceptions that we are aware of, namely the *contrastive approach* [15, 2] and the approach proposed by [4]. The idea behind the contrastive approach is that the distribution of a term in the focus corpus is insufficient to determine its domain-specificity. Instead, it is also necessary to examine its distribution across a *contrastive corpus* (i.e. a corpus containing documents about different domains). [4] attempted to isolate the TLR of a vast domain such as the environment. In particular, they test two measures. *Specificity*, a measure that compares the distribution of terms in the focus corpus with their distribution in a reference corpus and *Inverse Document Frequency*. They expect the terminology of the general environmental lexicon to obtain a low Inverse Document Frequency because it is evenly distributed throughout the document collection.

The remainder of this paper is organized as follows. In section 2 we describe Domain Concept Relatedness, a novel ATE technique meant to isolate the terminology of a given subject field. Section 3 contains an overview of the corpora that we created to evaluate our technique and the results of the evaluation. Finally, In section 4 we draw conclusions and propose future research directions. In summary, the main contributions of this paper are:

- We organize the terminology of a subject field into Terminology Layers.
- We propose a novel ATE technique to isolate a specific Terminology Layer.
- We create a focus corpus about waste management legislation and a contrastive corpus about legislation of other sectors.
- We test our technique against existing ATE techniques.

2. Proposed Approach

We propose *Domain Concept Relatedness* (DCR), an ATE technique capable of isolating TLRs. It is based on *Key Concept Relatedness* (KCR) [16] and addresses its limitations. DCR involves four phases: *candidate extraction*, *generation of concept embeddings*, *key concept extraction*, and *relatedness computation*. The code and the resources used for the evaluation are publicly available on GitHub[1].

[1] https://github.com/nicolaCirillo/termdomain

2.1. Key concept relatedness

KCR [16] is an ATE technique based on the assumption that terms of a given domain must be semantically related to already-known *key concepts* from that domain. To measure semantic relatedness, KCR relies on cosine similarities between *concepts embeddings*. In the latest version of KCR [16], concept embeddings are generated from hyperlinks of Wikipedia pages. In particular, the Wikipedia dump is preprocessed by removing markups and by replacing each hyperlink with a special token, and then the embeddings are generated via *word2vec* [17]. Next, to obtain the list of *key concepts*, KCR employs a keyword extraction algorithm, namely a simplified version of KP-Miner [18]. It automatically extracts key concepts from each document in the corpus. Finally, for each candidate term t for which a concept embedding exists, the algorithm computes the semantic relatedness to a subset of k key concepts (selected using the k Nearest Neighbour algorithm, with only positive examples). Below is the formula to compute *relatedness*.

$$relatedness(t, C_d) = \frac{1}{k} \sum_{i=1}^{k} cos(v_t, v_i) \tag{1}$$

where C_d is a set of key concepts sorted by cosine similarity to the candidate term t, k is a parameter from kNN while v_t and v_i are the word vectors of the term t and the key concept i, respectively. The relatedness of candidates that do not have a concept embedding is set to 0.

Overall, KCR is a promising technique. According to the evaluation done by [16], KCR is the best-performing technique on the FAO dataset. Furthermore, when it comes to document-level ATE, KCR clearly outperforms other techniques [19]. Nonetheless, it has two main limitations. The first limitation is the fact that it is unsuited to treat domains with low coverage by Wikipedia [16]. The second limitation has to do with the selection of the key concepts. Selecting the key concepts via automatic keyword extraction has the advantage of making KCR fully unsupervised, nevertheless, it does not ensure the adherence of all the key concepts to the investigated domain. preventing the possibility of using KCR to isolate a TLR.

2.2. Domain Concept Relatedness

In the candidate extraction step, DCR employs syntactic patterns based upon *RegexpParser* from the *nltk* Python library[2]. Since our main contribution is the scoring mechanism, we decided to keep the candidate extraction fairly simple. Thus, we extracted only nouns and noun phrases.

To rank candidates, we employed a modified version of KCR. Our aim is to make it extract only those terms that belong to the investigated domain, ignoring other terms that may occur in the focus corpus, however, KCR is not intended to isolate a given TLR. Worse still, it is not suitable for domains with low coverage by Wikipedia. To address the first issue, we made DCR semi-supervised. In particular, the set of key concepts is taken from an existing thesaurus[3] (or provided by the user). This modification ensures that all the key concepts belong to the investigated domain, thus enabling DCR to isolate a given TLR. In order to overcome the Wikipedia coverage problem, DCR employs a different technique to produce *concept embeddings*.

[2] https://www.nltk.org/

[3] in the evaluation (section 3), key concepts are taken from the following glossary: https://www.zerosprechi.eu/index.php/glossario

Corpus	Subjects	Documents	Sentences	Tokens
focus	waste management	**148**	**15,463**	**630,456**
contrastive	transport	60	8,239	331,061
	health	28	3,271	130,708
	safety	15	1,959	71,005
	agriculture	56	1,807	68,859
	TOT	**159**	**15,276**	**601,633**
Contrastive	transport, health, safety, and agriculture	159	15,276	601,633

Table 1
Structure of the corpora.

While KCR produces them from the Wikipedia dump, DCR produces concept embeddings directly from the focus corpus by employing the *Alacarte* [20] algorithm. The advantage of this technique is two-fold: it eliminates the Wikipedia coverage problem and ensures that each candidate has its vector representation. The convenience of Alacarte over traditional models is its ability to induce embeddings on the fly. Moreover, Alacarte generally produces more robust representations than traditional models when dealing with rare words and this is a key advantage in term extraction. To learn the transformation matrix, Alacarte needs only a set of pre-trained embeddings and the corpus used to induce them. For this purpose, we trained a *word2vec* model on the Paisà corpus [21]. Once the embeddings are generated, relatedness scores are computed as in equation 1 with k set to 5.

3. Evaluation

To test our novel technique, we compared it to two already existing ATE techniques. The first one is the term extraction tool of Sketch Engine, it compares the frequency of terms in the focus corpus with their frequency in a reference corpus. The second one is the contrastive approach [2], a technique that has already been employed to extract the terminology of the regulated sector from legislative documents. Due to the lack of gold-labelled corpora, we constructed a corpus of EU directives and regulations about waste management and applied the selected techniques to it. Then, the list of extracted terms was evaluated by three different annotators. This methodology has the drawback of assessing only *precision* (i.e. the correctness of the extracted items) but not *recall* (i.e the fraction of terms that have been extracted). Therefore, we computed precision at k (P@k) and average precision (AP). P@k is simply the number of correct terms out of the top k extracted items while AP indicates how highly are correct terms ranked.

3.1. Corpora

For our experiment, we constructed two corpora (ca. 600,000 tokens each), both composed of EU directives and regulations. The *focus* corpus is about the waste management sector, while the *contrastive* corpus (required to test the contrastive approach) covers four different subject matters (agriculture and fisheries, public health, safety at work, and transport). To build the focus corpus, we selected all the directives and regulations that are about a Eurovoc concept

related to the waste domain [22] (i.e. environmental policy, waste, and waste management) or one of its hyponyms. On the other hand, directives and regulations that compose the contrastive corpus concern different subject matters that, in our opinion, show enough variety to let the contrastive approach capture terms that characterize directives and regulations in general.

3.2. Dataset

We produced the evaluation dataset by selecting the first 200 single-word terms and the first 200 multi-word terms extracted from the focus corpus by each tool (1039 unique items in total). Then, these items were judged by three annotators, namely the author and two linguistics students. We provided each annotator with annotation guidelines and with the focus corpus. Prior to giving judgments, annotators had to check the usage of each item in the focus corpus. Each annotator provided two judgments. Firstly, they decided which items are terms. Secondly, they specified the domain to which terms belong. We defined five domains: legislation in general (LAW); waste management legislation (WASTE LAW); waste management (WASTE); topics related to waste management (WASTE REL); other domains (OTHER). Overall, the inter-annotator agreement concerning term identification is moderate (Fleiss k 0.54) and the agreement on domains is slightly lower (Fleiss k 0.47). Nonetheless, these figures are consistent with the literature. Agreement tends to be quite low due to the lack of clear boundaries between terminology and general language [23]. In the final version of the dataset, an item is considered a valid term only if at least two annotators judged it as such. The same holds for domain tags, if at least two annotators assigned an item to the same domain, that tag was kept in the final dataset. Otherwise, the main annotator (the author of this paper) decided which tag to keep.

3.3. Results

We ran three evaluations. In the first one, all terms are considered to be correctly extracted regardless of the domain. In the second one, only the terms belonging to the WASTE, WASTE LAW, and WASTE REL domains (waste terms) are considered correctly extracted, and in the last one, we consider to be correctly extracted only the terms belonging to the LAW domain (legal terms). Results are shown in Table 2. By looking at the P@200, it is clear that Sketch Engine is not capable of discriminating waste terms from legal ones since they have been extracted in similar percentages. Conversely, the contrastive approach and DCR are better suited for isolating waste terminology, notably the latter, for which the percentage of extracted legal terms is almost negligible. In addition, DCR also obtained the highest AP in the extraction of multi-word terms regarding waste while keeping a P@200 similar to the contrastive approach. Thus they extract almost the same number of terms but DCR tends to rank them higher. The same does not hold in the single-word scenario, where DCR shows lower scores than the contrastive approach and Sketch Engine. Overall, DCR works well with multi-word terms. It is able to isolate the terminology of the waste domain better than Sketch Engine and the contrastive approach while performing poorly on single-word terms.

Term type	Tool	All terms		Waste terms		Legal terms	
		P@200	AP	P@200	AP	P@200	AP
single-word	Sketch Engine	0.40	0.52	0.23	0.33	**0.18**	**0.23**
	contrastive approach	**0.47**	**0.55**	**0.31**	**0.37**	0.17	0.21
	DCR	0.20	0.28	0.17	0.29	0.03	0.02
multi-word	Sketch Engine	0.49	0.53	0.29	0.36	**0.21**	**0.20**
	contrastive approach	**0.54**	**0.67**	**0.40**	0.54	0.15	0.15
	DCR	0.43	0.62	0.37	**0.62**	0.06	0.05

Table 2
Results of the evaluation.

4. Conclusions and perspectives

The ability to isolate the terms of a specific subject field is an important feature of ATE tools that has been underlined in the context of legislative documents, where it is crucial to separate the terminology of the regulated sector from legal terminology. To this end, we propose Domain Concept Relatedness (DCR), a novel ATE technique to isolate terminology layers. The main difference between DCR and other existing ATE techniques with the same goal is that it is not based on corpus statistics but on word embeddings. Due to the lack of gold-labelled data, we create a corpus of EU directives and regulations about waste management and use it to test DCR. Our evaluation suggests that DCR is effective at extracting multi-word terms belonging to a given subject field but fails to extract single-word terms.

Furthermore, it must be noted that DCR has many parameters that will be optimized to increase its effectiveness, notably the set of word embeddings employed. It would also be interesting to combine DCR with traditional ATE techniques based on corpus statistics by using it as a re-ranking technique.

References

[1] A. Lenci, S. Montemagni, V. Pirrelli, G. Venturi, Ontology learning from italian legal texts, in: Law, Ontologies and the Semantic Web, IOS Press, 2009, pp. 75–94.

[2] F. Bonin, F. Dell'Orletta, S. Montemagni, G. Venturi, A contrastive approach to multi-word extraction from domain-specific corpora, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/553_Paper.pdf.

[3] D. Vellutino, L'italiano istituzionale per la comunicazione pubblica, il Mulino, 2018.

[4] P. Drouin, M.-C. L'Homme, B. Robichaud, Lexical profiling of environmental corpora, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[5] J. S. Justeson, S. M. Katz, Technical terminology: some linguistic properties and an algorithm for identification in text, Natural language engineering 1 (1995) 9–27.

[6] D. A. Evans, R. G. Lefferts, Clarit - trec experiments, Information Processing & Management 31 (1995) 385–395.
[7] K. Church, W. Gale, Inverse document frequency (idf): A measure of deviations from poisson, in: Natural language processing using very large corpora, Springer, 1999, pp. 283–295.
[8] R. Navigli, P. Velardi, Learning domain ontologies from document warehouses and dedicated web sites, Computational Linguistics 30 (2004) 151–179.
[9] K. Ahmad, L. Gillam, L. Tostevin, et al., University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder), in: TREC, 1999, pp. 1–8.
[10] Y. Park, R. J. Byrd, B. Boguraev, Automatic glossary extraction: Beyond terminology identification., in: COLING, volume 10, 2002, pp. 1072228–1072370.
[11] K. Meijer, F. Frasincar, F. Hogenboom, A semantic approach for extracting domain taxonomies from text, Decision Support Systems 62 (2014). doi:10.1016/j.dss.2014.03.006.
[12] M. Kucza, J. Niehues, T. Zenkel, A. Waibel, S. Stüker, Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks, volume 2018-September, International Speech Communication Association, 2018, pp. 2072–2076. doi:10.21437/Interspeech.2018-2017.
[13] A. Hazem, M. Bouhandi, F. Boudin, B. Daille, Termeval 2020: Taln-ls2n system for automatic term extraction, in: 6th International Workshop on Computational Terminology (COMPUTERM 2020), 2020.
[14] S. H. Manjunath, J. P. McCrae, Encoder-attention-based automatic term recognition (ea-atr), in: 3rd Conference on Language, Data and Knowledge (LDK 2021), Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.
[15] R. Basili, A. Moschitti, P. M. Teresa, F. M. Zanzotto, A contrastive approach to term extraction, 2001. URL: https://www.researchgate.net/publication/283854408.
[16] N. Astrakhantsev, Atr4s: toolkit with state-of-the-art automatic terms recognition methods in scala, Language Resources and Evaluation 52 (2018) 853–872.
[17] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 1st International Conference on Learning Representations, {ICLR} 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings (2013).
[18] S. R. El-Beltagy, A. Rafea, Kp-miner: Participation in semeval-2, Association for Computational Linguistics, 2010, pp. 190–193. URL: https://aclanthology.org/S10-1041.
[19] A. Šajatović, M. Buljan, J. Šnajder, B. D. Bašić, Evaluating automatic term extraction methods on individual documents, Association for Computational Linguistics, 2019, pp. 149–154. URL: https://aclanthology.org/W19-5118. doi:10.18653/v1/W19-5118.
[20] M. Khodak, N. Saunshi, Y. Liang, T. Ma, B. Stewart, S. Arora, A la carte embedding: Cheap but effective induction of semantic feature vectors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 12–22. URL: https://aclanthology.org/P18-1002. doi:10.18653/v1/P18-1002.
[21] V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'orletta, H. Dittmann, A. Lenci, V. Pirrelli, The paisa' corpus of italian web texts, Association for Computational

Linguistics, 2014, pp. 36–43.
[22] D. Vellutino, R. Maslias, F. Rossi, Verso l'interoperabilità semantica di iate. studio preliminare per il dominio "gestione dei rifiuti urbani", Terminologie specialistiche e diffusione dei saperi (2016) 1–240.
[23] A. Rigouts Terryn, V. Hoste, E. Lefever, In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora, Language Resources and Evaluation 54 (2020) 385–418.