languages

MDPI

Article

Isolating Terminology Layers in Complex Linguistic Environments: A Study about Waste Management

Nicola Cirillo

Department of Political and Communication Science (DiSPC), University of Salerno, 84084 Fisciano, SA, Italy; nicirillo@unisa.it

Abstract: Automatic term extraction aims at extracting terminological units from specialized corpora to assist terminographers in developing glossaries, thesauri, and termbases. Unfortunately, traditional methods often overlook the complex relation between terminologies of different subject fields that co-occur in a single specialized corpus. This study illustrates Domain Concept Relatedness, a novel term extraction technique meant to isolate the terminology of a given subject field. We test our technique against the term extraction tool of Sketch Engine and the contrastive approach by applying them to the extraction of waste management terms from a new Italian corpus about waste management legislation. The results show that Domain Concept Relatedness effectively extracts multi-word terms belonging to a given subject field but still fails to extract single-word terms.

Keywords: automatic term extraction; waste management terminology; domain knowledge; specialized languages

check for updates

Citation: Cirillo, Nicola. 2024. Isolating Terminology Layers in Complex Linguistic Environments: A Study about Waste Management. *Languages* 9: 68. https://doi.org/10.3390/languages9030068

Academic Editor: Jeanine Treffers-Daller

Received: 8 December 2023
Revised: 11 February 2024
Accepted: 12 February 2024
Published: 20 February 2024

1. Introduction

Automatic term extraction (ATE) is a natural language processing task. Its focus is the extraction of terms from specialized corpora. A typical ATE tool extracts all of the terms occurring in a corpus, disregarding their subject fields. While this approach is reasonable in most cases, it would be beneficial to distinguish terms related to different subject fields in some contexts. Lenci et al. (2009) and Bonin et al. (2010) stated that when ATE is applied to legislative documents, it becomes crucial to differentiate terms that belong to the regulated sector from legal terms. Unfortunately, only a few ATE methods address this task.

Moreover, from a theoretical perspective, it is too simplistic to assume that all of the terms in a specialized corpus belong to the same subject field. 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 of the technical terms of the regulated sectors (Vellutino 2018). In addition, each subject field can be divided into more specific sub-fields, which can, in turn, be separated into even smaller sub-fields, leading to a complex hierarchical model. Top-level layers contain terms related to broader domains, while lower-level layers incorporate more topic-specific terms. For example, Drouin et al. (2018) divided the terminology of the environment into two layers, the general environmental lexicon and the topic-specific lexicon. The former includes terms that are used in all environmental texts, regardless of the topic (e.g., *biologist*, *ecosystem*, *green*), while the latter is specific to a given topic, such as renewable energy (e.g., *photovoltaic*), waste management (e.g., *compostable*), climate change (e.g., *global warming*), etc. We will refer to the terminology of a given subject field as a terminology layer—from now on, TLR—regardless of its position in the hierarchy.

To further exemplify the concept of a TLR, let us analyze an excerpt from *Directive 2006/21/EC of the European Parliament and of the Council of 15 March 2006 on the management of waste from extractive industries and amending Directive 2004/35/EC*.

> "The *[1] competent authority* shall, prior to the commencement of any operations involving the accumulation or deposit of *[2] extractive waste* in a *[3] waste facility*, require a *[4] financial guarantee* (e.g., in the form of a financial deposit, including industry-sponsored mutual guarantee funds) or equivalent, in accordance with procedures to be decided by the *[5] Member States* [...]"

The terms [1], [3], and [4] are easy to classify. They belong to the law, waste management, and finance TLRs, respectively. In spite of being a legal term (such as the term [1]), the term [5] is specific to European law. Thus, it belongs to the European law TLR, which descends from the law TLR. Being composed of two subterms, the term [2] is more complex. Namely, the noun *waste* is related to waste management, while the adjective *extractive* is related to mining industry; therefore, the term [2] belongs to the waste management of mining industry TLR, which descends from both the waste management and the mining industry TLRs. Figure 1 summarizes this classification scheme.

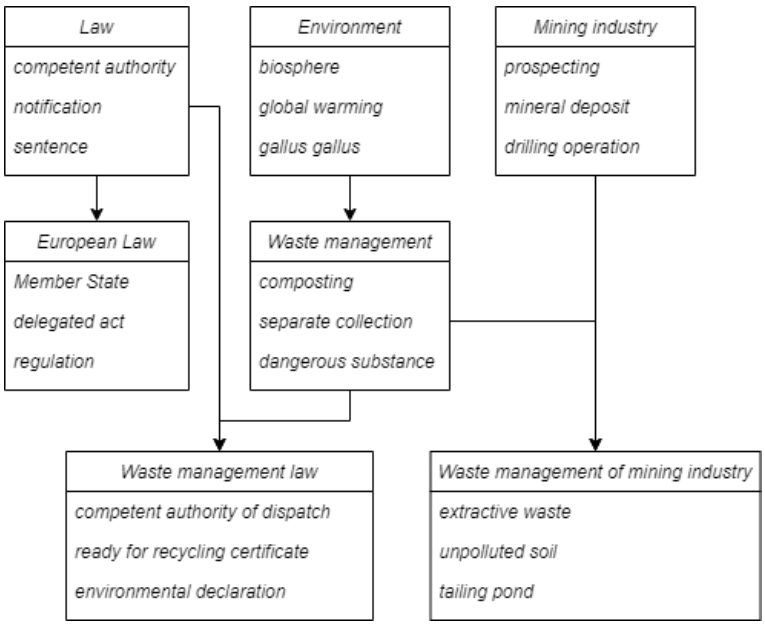

Figure 1. Terminology layers.

In this study, we propose and test a novel ATE technique: Domain Concept Relatedness—DCR. Based on Key Concept Relatedness (KCR) (Astrakhantsev 2018), DCR is one of the few ATE techniques exploiting word embeddings. It isolates a given TLR by extracting only those terms that are semantically related to already-known terms of a glossary or thesaurus.

In summary, the main contributions of this study are the following:

- We organize the terminology of a specialized corpus into TLRs.
- We propose a novel ATE technique for isolating a specific TLR.
- 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.

The remainder of this paper is organized as follows. Section 2 gives an overview of previous literature about ATE. In Section 3, we describe and evaluate DCR, a novel ATE technique meant to isolate the terminology of a given subject field. Section 4 shows the evaluation results. Finally, in Section 5, we discuss the results, draw conclusions, and propose future research directions.

2. Related Work

Most ATE algorithms share a similar architecture (Astrakhantsev et al. 2015). They employ morpho-syntactic patterns (Meyers et al. 2018; Pais and Ion 2020) or sentence

chunking (Oliver and Vàzquez 2020; Rose et al. 2010) to extract a list of candidates (i.e., words and phrases). Then, candidates are ranked according to their distribution in a focus corpus (i.e., a specialized corpus), which is sometimes compared to a reference corpus (i.e., general language corpus) (Kilgarriff 2009; Meijer et al. 2014; Park et al. 2002). When a gold-labeled dataset is available, it is possible to train a supervised classifier to combine multiple statistical measures (Patry and Langlais 2005). Some approaches mentioned above are implemented in the JATE 2.0 (Zhang et al. 2016) and the ATR4S (Astrakhantsev 2018) toolkits.

A more distinctive approach takes advantage of word embeddings. KCR (Astrakhantsev 2018) exploits the semantic relationships between candidates and already-known concepts of the target domain. Since DCR is a modified version of KCR, the latter is further described in Section 3.1. Pais and Ion (2020) also exploited word embeddings; they tested a clustering algorithm, DBSCAN, which outperformed statistical approaches.

Recently, deep learning methods have become widely popular in ATE, as in other natural language processing tasks. They do not need the explicit computation of metrics but, rather, rely on internal representations of words and sentences. Kucza et al. (2018) proposed an LSTM model that performed sequence labeling with a BILOU schema (Beginning, Inside, Last, Outside, Unit). Hazem et al. (2020) and Manjunath and McCrae (2021) employed a BERT model with notable results. However, despite being quite effective, deep learning methods have a relevant drawback: They learn from gold-labeled datasets, which are available only for a few subject fields.

The approaches to ATE discussed above do not address the isolation of TLRs. Nevertheless, there are at least two approaches that specifically address this ATE subtask, namely, the contrastive approach (Basili et al. 2001; Bonin et al. 2010) and the approach proposed by Drouin et al. (2018). The idea behind the contrastive approach is that the distribution of a term in a 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). Therefore, Basili et al. (2001) proposed two contrastive metrics to evaluate single-word and multi-word terms. The contrastive weight for a single-word term is equal to the product of its frequency in the focus corpus and its inverse word frequency across all corpora. On the other hand, the probability of multi-word terms is harder to estimate because of their sparsity. Thus, the weight of a multi-word term is obtained by multiplying the contrastive weight of its syntactical head by the frequency of the whole term in the focus corpus. Furthermore, Bonin et al. (2010) weighed both single-word and multi-word terms with the same metric through a function suitable for low-frequency events.

Drouin et al. (2018) attempted to isolate the TLR of a vast domain, such as the environment, to capture the general environmental lexicon and to separate it from the lexica of more specific topics (i.e., renewable energy, waste management, climate change, and endangered species). In particular, they tested 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 expected that the general environmental lexicon would obtain a low inverse document frequency, since it was evenly distributed throughout the document collection.

3. Materials and Methods

We propose Domain Concept Relatedness (DCR), an ATE technique capable of isolating TLRs. It is based on KCR (Astrakhantsev 2018) and addresses its limitations. The code and the material described in this section are publicly available on GitHub.[1]

3.1. Key Concept Relatedness

KCR (Astrakhantsev 2018) 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, it relies on cosine similarity between concept

embeddings. In its latest version (Astrakhantsev 2018), concept embeddings are generated from hyperlinks of Wikipedia pages. In particular, a Wikipedia dump is preprocessed by removing markups and replacing each hyperlink with a special token. Then, the embeddings are generated via word2vec (Mikolov et al. 2013). Next, to obtain the key concepts, KCR employs a keyword extraction algorithm, namely, a simplified version of KP-Miner (El-Beltagy and Rafea 2010). 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 Neighbor algorithm (kNN) with only positive examples). Below is the formula for computing relatedness.

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

where C_d is a set of key concepts sorted by cosine similarity to the candidate term t, k is a parameter from kNN, and 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 0.

Overall, KCR is a promising technique. According to the evaluation performed by Astrakhantsev (2018), KCR is the best-performing technique on the FAO dataset. Furthermore, with regard to document-level ATE, KCR outperforms other techniques (Šajatović et al. 2019). Nonetheless, it has two main limitations. Firstly, it is unsuited for treating domains with low coverage by Wikipedia (Astrakhantsev 2018). Secondly, the selection of key concepts via keyword extraction does not ensure their adherence to the investigated domain, thus preventing the possibility of using KCR to isolate TLRs.

3.2. Domain Concept Relatedness

DCR involves three phases: candidate extraction, generation of concept embeddings, and relatedness computation. 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, DCR extracts only nouns and noun phrases.

DCR aims to extract only 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 was taken from an existing thesaurus (or provided by the user). This modification ensured that all key concepts belonged to the investigated domain, thus enabling DCR to isolate a given TLR. Moreover, to overcome the problem of Wikipedia coverage, DCR employed a different technique for producing concept embeddings. While KCR produces them from a Wikipedia dump, DCR produces concept embeddings directly from the focus corpus by employing the Alacarte algorithm (Khodak et al. 2018). The advantage of this technique is two-fold: It solves the problem of Wikipedia coverage 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. In addition, Alacarte generally produces more robust representations than those of traditional models when dealing with rare words (and this is a key advantage in term extraction). Alacarte needs a set of pre-trained embeddings and the corpus used to induce them to learn a transformation matrix. For this purpose, we trained a word2vec model on the Paisà corpus (Lyding et al. 2014). Once the embeddings were generated, relatedness scores are computed as in Equation (1) with $k = 5$.

3.3. Corpora

To test DCR, we constructed two corpora (ca. 600,000 tokens each). They were composed of EU directives and regulations. The focus corpus was about the waste management sector. Conversely, the contrastive corpus, which was required to test the contrastive ap-

proach (Bonin et al. 2010), covered four different subject matters (agriculture and fisheries, public health, safety at work, and transport). The structure of the corpora is summarized in Table 1. To build the focus corpus, we selected all of the directives and regulations that were about a Eurovoc concept related to the waste management domain (Vellutino et al. 2016) (i.e., environmental policy, waste, and waste management) or one of its hyponyms. On the other hand, the directives and regulations that composed the contrastive corpus concerned different subject matters that, in our opinion, showed enough variety to let the contrastive approach capture terms that characterize directives and regulations in general.

Table 1. Structure of the corpora.

Corpus	Subjects	Documents	Sentences	Tokens
focus	waste management	148	15,463	630,456
contrastive	transport	60	8239	331,061
	health	28	3271	130,708
	safety	15	1959	71,005
	agriculture	56	1807	68,859
	TOT	159	15,276	601,633

3.4. Dataset

The evaluation dataset was obtained by selecting the first 200 single-word terms and the first 200 multi-word terms extracted from the focus corpus through DCR, Sketch Engine (Kilgarriff 2009), and the contrastive approach (Bonin et al. 2010) (1039 unique items in total). Three annotators evaluated the items: the author and two students of institutional languages with a fair knowledge of the domain. We provided each annotator with the list of unique items[3] (arranged in random order), annotation guidelines, and the focus corpus. The annotators had to check the item usage in the focus corpus and provide two judgments. Firstly, they decided if an item was a term. Then, they specified the domain to which it belonged. We defined five domains: legislation in general (LAW), waste management legislation (WASTE LAW), waste management (WASTE), topics related to waste management (WASTE REL), and other domains (OTHER). Overall, the inter-annotator agreement concerning term identification was moderate (Fleiss k 0.54), and the agreement on domains was slightly lower (Fleiss k 0.47). Nonetheless, these figures were consistent with the literature; agreement tends to be quite low due to the lack of clear boundaries between terminology and general language (Rigouts Terryn et al. 2020). In the final version of the dataset, an item was a valid term if at least two annotators judged it as such. The same held 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 domain tag to keep.

3.5. Evaluation

To test DCR, we compared it to two already existing ATE techniques. The first one was the term extraction tool of Sketch Engine (Kilgarriff 2009). It compares the frequency of terms in the focus corpus with their frequency in a reference corpus[4] and is not meant to isolate a terminology layer. Therefore, it should provide a clue about the distribution of different TLRs in a corpus. Moreover, we preferred it among other similar techniques because it is widely popular among terminologists. The second one was the contrastive approach (Bonin et al. 2010), a technique that has already been employed to extract the terminology of the regulated sector from legislative documents.

Due to the lack of gold-labeled corpora, we evaluated the list of items extracted by each tool. This method had the drawback of assessing only precision (i.e., the correctness of the extracted items) but not recall (i.e., the fraction of terms extracted). To measure precision, we computed two metrics: precision at k (P@k) and average precision (AP). P@k is simply the number of correct terms out of the top k items extracted by the tool. The more

extracted items are correct, the higher P@k is. We set k to 200. Contrary to P@k, AP also accounts for the ranking of items. Thus, the more correct terms appear at the top of the list, the higher the AP is. AP was computed with the following formula (2):

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \quad (2)$$

where P_n and R_n are the precision and recall at the nth threshold. We computed AP considering only the first 200 items extracted by each tool.

We ran three evaluations. In the first one, all terms were considered to be correct, regardless of the domain. In the second one, only the terms belonging to the WASTE, WASTE LAW, and WASTE REL domains (waste terms) were considered to be correct, and in the last one, we considered to be correct only the terms belonging to the LAW domain (legal terms). In all three evaluations, the key concepts provided to DCR were taken from the online glossary of Zerosprechi,[5] an information website that aims to provide high-quality environmental information. It contains 58 terms and represents an example of an easy-to-find online glossary.

4. Results

The results of the evaluation described in Section 3.5 are shown in Table 2. By looking at the results of Sketch Engine, it emerged that the corpus contained waste and legal terms in similar percentages. Conversely, the contrastive approach and DCR were better suited for isolating waste terminology—notably the latter, for which the ratio of extracted legal terms was almost negligible. In addition, DCR obtained the highest AP in the extraction of multi-word terms of the waste TLR while keeping a P@200 similar to that of the contrastive approach. Thus, they extracted almost the same numbers of terms, but DCR tended to rank them higher. The same did not hold in the single-word scenario, where DCR showed lower scores than those of the contrastive approach and Sketch Engine. Overall, DCR worked well with multi-word terms. It was able to isolate the terminology of the waste TLR better than Sketch Engine and the contrastive approach while performing poorly on single-word terms.

Table 2. Results of the evaluation.

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

5. Discussion

To assess the differences among Sketch Engine, DCR, and the contrastive approach, we qualitatively analyzed the results. Examining single-word terms, it emerged that DCR did not extract any acronyms, while Sketch Engine did (e.g., *CER* 'European waste catalogue', *BAT* 'best available techniques', *ERU* 'emission reduction unit'). The cause probably lay in the candidate extraction step of DCR; acronyms were not tagged as nouns by the part-of-speech tagger. Furthermore, we observed that DCR extracted certain common words (e.g., *operazione* 'operation', *processo* 'process', and *trattamento* 'processing) that were not terms per se but might convey a terminological meaning when employed as anaphoric

references, replacing multi-word terms (e.g., *operazione di smaltimento* 'disposal operation', *processo di riciclaggio* 'recycling process', and *trattamento dei rifiuti* 'waste processing'). Other differences arose from multi-word terms, with DCR extracting more low-frequency terms compared to the other two algorithms. For example, *incenerimento mediante ossidazione* 'oxidative combustion' (two occurrences) and *stoccaggio in loco* 'storage of waste on-site' (one occurrence). Moreover, while both Sketch Engine and the contrastive approach extracted 19 terms related to waste management legislation (e.g., *dichiarazione ambientale* 'environmental declaration' and *documento di movimento* 'movement document'), DCR extracted only 6.

Among other algorithms proposed to tackle ATE, DCR is one of the few that employs static word embeddings with satisfactory results. This approach bridges the gap between simple frequency-based methods and advanced deep learning techniques. Moreover, the advantages of DCR emerge when the task is the isolation of a TLR in a specialized corpus. In this scenario, DCR is less demanding than the contrastive approach. For instance, despite providing good results, the contrastive approach (Basili et al. 2001; Bonin et al. 2010) requires a contrastive corpus with the terminology that the user wants to exclude. Unfortunately, such a corpus may not be readily available. On the other hand, DCR only requires a list of already-known terms that are easier to find. Worse still, to the best of our knowledge, TLR isolation has never been addressed via deep learning. Probably, a deep learning algorithm may be easily trained on the classification of terms according to a given set of subject fields. However, users may also be interested in the terminology of ad hoc topics that the model has never seen during training.

Because DCR is specifically made to isolate TLRs, its performance is not directly comparable to that of state-of-the-art ATE algorithms. Nonetheless, we illustrate the results of the TermEval 2020 shared task (Rigouts Terryn et al. 2020) in Table 3 to provide an interpretive framework for our evaluation. Five teams participated in this shared task. Two of them, namely, TALN-LS2N (Hazem et al. 2020) and NLPLabUQAM, employed a deep learning approach. The NYU and e-Terminology teams (Oliver and Vàzquez 2020) opted for a traditional statistical approach. Finally, the RACAI team (Pais and Ion 2020) mixed statistical approaches and word embeddings.

Table 3. Results of TermEval 2020. Source: Rigouts Terryn et al. (2020).

Track	Rank	Team	Scores incl. Named Entities		
			Precision	Recall	f1-Score
English	1	TALN-LS2N	34.8	70.9	46.7
	2	RACAI	42.4	40.3	41.3
	3	NYU	43.5	23.6	30.6
	4	e-Terminology	34.4	14.2	20.1
	5	NLPLab_UQAM	21.4	15.6	18.1
French	1	TALN-LS2N	45.2	51.5	48.1
	2	e-Terminology	36.3	13.5	19.7
	3	NLPLab_UQAM	16.1	11.2	13.2
Dutch	1	NLPLab_UQAM	18.9	18.6	18.7
	2	e-Terminology	29.0	9.6	14.4

From TermEval 2020, it emerged that ATE is far from being a solved problem. The precision scores never reached 0.5. Also, the recall was generally around 0.5, except for the best system, TALN-L2SN, which obtained a recall of 0.71 in the English track (but a precision of 0.35). In this scenario, the precision obtained by DCR on the extraction of multi-word terms is promising, especially considering that DCR requires little supervision. However, it still needs to be refined to treat single-word terms effectively.

In summary, DCR is one of the few embedding-based approaches to ATE that provides a convenient and flexible solution to TLR isolation. Moreover, this work represents the

starting point for developing datasets that take into account the complex hierarchical organization of terms contained in specialized corpora.

In the future, we plan to integrate DCR with existing large termbases such as EuroVoc and IATE. Furthermore, DCR has some 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 statistical ATE techniques by using it for the re-ranking of candidates.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: https://github.com/nicolaCirillo/termdomain (accessed on 15 February 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AP	Average Precision
ATE	Automatic Term Extraction
BERT	Bidirectional Encoder Representations from Transformers
BILOU	Beginning, Inside, Last, Outside, Unit
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCR	Domain Concept Relatedness
KCR	Key Concept Relatedness
kNN	k Nearest Neighbors
LSTM	Long Short-Term Memory
P@k	Precision at k
TLR	Terminology Layer

Notes

1 https://github.com/nicolaCirillo/termdomain (accessed on 15 February 2024).
2 https://www.nltk.org/ (accessed on 15 February 2024).
3 Annotators were unaware of which tool extracted an item.
4 The reference corpus chosen in the evaluation was itTenTen.
5 https://www.zerosprechi.eu/index.php/glossario (accessed on 15 February 2024).

References

Astrakhantsev, Nikita. 2018. Atr4s: Toolkit with state-of-the-art automatic terms recognition methods in scala. *Language Resources and Evaluation* 52: 853–72. [CrossRef]

Astrakhantsev, Nikita A., Denis G. Fedorenko, and Denis Yu Turdakov. 2015. Methods for automatic term recognition in domain-specific text collections: A survey. *Programming and Computer Software* 41: 336–49. [CrossRef]

Basili, Roberto, Alessandro Moschitti, Pazienza Maria Teresa, and Fabio Massimo Zanzotto. 2001. A Contrastive Approach to Term Extraction. Paper presented at Terminology and Artificial Intelligence Conference (TIA 2001), Nancy, France, May 3–4.

Bonin, Francesca, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2010. A contrastive approach to multi-word extraction from domain-specific corpora. Paper presented at Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, May 17–23. Marseille: European Language Resources Association (ELRA).

Drouin, Patrick, Marie-Claude L'Homme, and Benoît Robichaud. 2018. Lexical profiling of environmental corpora. Paper presented at Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 7–12.

El-Beltagy, Samhaa R., and Ahmed Rafea. 2010. Kp-Miner: Participation in Semeval-2. Paper presented at 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, July 15–16. Kerrville: Association for Computational Linguistics, pp. 190–93.

Hazem, Amir, Mérieme Bouhandi, Florian Boudin, and Béatrice Daille. 2020. Termeval 2020: Taln-ls2n system for automatic term extraction. Paper presented at 6th International Workshop on Computational Terminology (COMPUTERM 2020), Marseille, France, May 11–16.

Khodak, Mikhail, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. 2018. A la carte embedding: Cheap but effective induction of semantic feature vectors. Paper presented at 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, July 15–20. Kerrville: Association for Computational Linguistics, pp. 12–22. [CrossRef]

Kilgarriff, Adam. 2009. Simple maths for keywords. Paper presented at 5th Corpus Linguistic Conference (CL2009), Liverpool, UK, July 20–23.

Kucza, Maren, Jan Niehues, Thomas Zenkel, Alex Waibel, and Sebastian Stüker. 2018. *Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks*. New York: Curran Associates, pp. 2072–76. Volume 2018-September. [CrossRef]

Lenci, Alessandro, Simonetta Montemagni, Vito Pirrelli, and Giulia Venturi. 2009. Ontology learning from italian legal texts. In *Law, Ontologies and the Semantic Web*. Amsterdam: IOS Press, pp. 75–94.

Lyding, Verena, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The Paisa' Corpus of Italian Web Texts. Paper presented at 9th Web as Corpus Workshop (WaC-9), Gothenburg, Sweden, April 26. Kerrville: Association for Computational Linguistics, pp. 36–43.

Manjunath, Sampritha H., and John P. McCrae. 2021. Encoder-attention-based automatic term recognition (ea-atr). Paper presented at 3rd Conference on Language, Data and Knowledge (LDK 2021), Zaragoza, Spain, September 1–3. Wadern: Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Meijer, Kevin, Flavius Frasincar, and Frederik Hogenboom. 2014. A semantic approach for extracting domain taxonomies from text. *Decision Support Systems* 62: 78–93. [CrossRef]

Meyers, Adam L., Yifan He, Zachary Glass, John Ortega, Shasha Liao, Angus Grieve-Smith, Ralph Grishman, and Olga Babko-Malaya. 2018. The termolator: Terminology recognition based on chunking, statistical and search-based scores. *Frontiers in Research Metrics and Analytics* 3: 19. [CrossRef]

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Paper presented at 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, AZ, USA, May 2–4. Workshop Track Proceedings.

Oliver, Antoni, and Mercè Vàzquez. 2020. Termeval 2020: Using tsr filtering method to improve automatic term extraction. In *Proceedings of the 6th International Workshop on Computational Terminology*. Marseille: European Language Resources Association, pp. 106–13.

Pais, Vasile, and Radu Ion. 2020. TermEval 2020: RACAI's automatic term extraction system. In *Proceedings of the 6th International Workshop on Computational Terminology*. Marseille: European Language Resources Association, pp. 101–5.

Park, Youngja, Roy J. Byrd, and Branimir Boguraev. 2002. Automatic glossary extraction: Beyond terminology identification. *COLING* 10: 1072228–370.

Patry, Alexandre, and Philippe Langlais. 2005. Corpus-based terminology extraction. In *Terminology and Content Development—Proceedings of 7th International Conference on Terminology and Knowledge Engineering*. Copenhagen: Litera.

Rigouts Terryn, Ayla, Veronique Hoste, Patrick Drouin, and Els Lefever. 2020. TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset. In *Proceedings of the 6th International Workshop on Computational Terminology*. Edited by Béatrice Daille, Kyo Kageura and Ayla Rigouts Terryn. Marseille: European Language Resources Association, pp. 85–94.

Rigouts Terryn, Ayla, Véronique Hoste, and Els Lefever. 2020. In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation* 54: 385–418. [CrossRef]

Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory* 1: 1–20.

Šajatović, Antonio, Maja Buljan, Jan Šnajder, and Bojana Dalbelo Bašić. 2019. Evaluating Automatic Term Extraction Methods on Individual Documents. Paper presented at Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), Florence, Italy, August 2. Kerrville: Association for Computational Linguistics, pp. 149–54. [CrossRef]

Vellutino, Daniela. 2018. *L'italiano istituzionale per la comunicazione pubblica*. Bologna : il Mulino.

Vellutino, Daniela, Rodolfo Maslias, and Francesco Rossi. 2016. Verso l'interoperabilità semantica di iate. studio preliminare per il dominio "gestione dei rifiuti urbani". In *Terminologie specialistiche e diffusione dei saperi*. Milano: EDUCatt—Ente per il Diritto allo studio universitario dell'Università Cattolica, pp. 1–240.

Zhang, Ziqi, Jie Gao, and Fabio Ciravegna. 2016. Jate2. 0: Java automatic term extraction with apache solr. Paper presented at Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, May 23–28. Marseille: European Language Resources Association (ELRA).