The Recent Advances in Automatic Term Extraction: A survey

HANH THI HONG TRAN, Jozef Stefan Institute, Slovenia; University of La Rochelle, France, Slovenia

MATEJ MARTINC, Jozef Stefan Institute, Slovenia, Slovenia

JAYA CAPORUSSO, Jozef Stefan Institute, Slovenia, Slovenia

ANTOINE DOUCET, University of La Rochelle, France, France

SENJA POLLAK, Jozef Stefan Institute, Slovenia, Slovenia

Automatic term extraction (ATE) is a Natural Language Processing (NLP) task that eases the effort of manually identifying terms from domain-specific corpora by providing a list of candidate terms. As units of knowledge in a specific field of expertise, extracted terms are not only beneficial for several terminographical tasks, but also support and improve several complex downstream tasks, e.g., information retrieval, machine translation, topic detection, and sentiment analysis. ATE systems, along with annotated datasets, have been studied and developed widely for decades, but recently we observed a surge in novel neural systems for the task at hand. Despite a large amount of new research on ATE, systematic survey studies covering novel neural approaches are lacking. We present a comprehensive survey of deep learning-based approaches to ATE, with a focus on Transformer-based neural models. The study also offers a comparison between these systems and previous ATE approaches, which were based on feature engineering and non-neural supervised learning algorithms.

1 INTRODUCTION

Terms are textual expressions that denote concepts in a specific field of expertise. They are beneficial for several terminographical tasks performed by linguists, e.g., construction of term dictionaries [60]. Moreover, terms can also improve several complex downstream natural language processing (NLP) tasks, e.g., information retrieval [62] and machine translation [106]. Manual term extraction is labor- and time-demanding. Therefore, with the new advances in machine learning and the recent introduction of novel neural architectures, we observed a push towards automatizing the task and reducing the time and effort needed to identify terms in domain-specific corpora.

ATE systems usually employ a two-step procedure: (1) extracting a list of candidate terms, and (2) determining which candidate terms are correct using supervised or unsupervised techniques. Early ATE systems were based on handcrafted rules and substantial feature engineering effort, which relied on linguistic knowledge and distinctive linguistic aspects

of terms. They employed several NLP tools, e.g., tokenization, lemmatization, stemming, chunking, and POS tagging. Next, several other studies opted for a statistical approach based on various measures, such as termhood [101], unithood [23], or C-value [36]. These strategies are also employed in some current hybrid systems, which rely on combining linguistic and statistical information.

Inspired by the success of deep neural networks on several NLP tasks, a variety of neural approaches were proposed for ATE over the past few years, the most popular being the approaches that rely on Transformer-based language models [100]. Such models are appealing due to two major reasons: (1) they can deduce features themselves, which dismisses the need for hand-crafted feature engineering; and (2) they typically do not require domain-specific resources and are thus more domain-independent. Overall, several studies used neural architectures for ATE, by either leveraging embedding representations as the input for the classifiers [1, 37, 56, 109] or by employing pretrained language models followed by fine-tuning in the recent NLP transfer learning paradigm [15, 44, 87, 95, 98].

The recent advances in supervised machine learning and deep learning approaches have become more popular than the unsupervised ones in the context of term extraction. However, while the traditional approaches had been comprehensively addressed in previous surveys [20, 51], no systematic summary of the improvements made in supervised term extraction systems using machine learning or deep learning has been proposed yet. Therefore, our paper aims to provide a comprehensive research trajectory in the supervised term extraction task, including the following aspects.

(1) We survey the resources for term extraction of the past three decades. All the well-annotated corpora are taken into consideration if the two following requirements are satisfied: (1) they are publicly available, and (2) a description paper explaining their construction for the term extraction was published;

(2) We propose a systematic review of deep learning-based approaches to the term extraction task (with a specific focus on the Transformer-based models) and compare them with previous approaches based on feature engineering and shallow supervised learning algorithms. Our comparison analysis highlights the improvements achieved by neural networks and shows that incorporating some of the insights learned from past work on feature engineering-based ATE systems can yield further improvements;

(3) We present all the metrics used in term extraction and categorize them based on either indirect (i.e., through a downstream application) or direct evaluation methodology (i.e., whether the evaluation is human judges, dictionary-based, or gold standard based) and the scope of results (i.e., whether we evaluate the entire results, parts of results, or top-k best results);

(4) Incorporating the lessons learned from the existing techniques, we point out the current challenges in correctly capturing multi-word and nested terms, and in improving the robustness of the term extractor. Based on the current challenges, we propose some suggestions for further exploration in the term extraction research.

The rest of the survey is organized as follows. Section 2 presents the related surveys of approaches to ATE written in the last few decades. In Section 3, we introduce our approach to identifying the articles for the survey. Section 4 presents a comparative study of different monolingual and multilingual corpora that can be used for the training and evaluation of ATE systems. This is followed by a systematic description of different ATE approaches in Section 5, different evaluation metrics used in ATE in Section 6, and a comparative evaluation of different systems in Section 7. Based on the current works and the trending technology, we discuss the challenges we are still facing and the possible future directions in Section 8. Finally, we conclude our work in Section 9.

2 PREVIOUS SURVEYS AND COMPARATIVE STUDIES

The first comprehensive survey of approaches to solving ATE was written by [51]. It provided a systematic overview of the two major trends in the principles and methods of automatic term recognition and introduced a distinction between linguistic and statistical approaches. [80] argued that all contemporary algorithms employed linguistic methods as a filtering step. A decade later, [20] presented a survey of different approaches in automatic term extraction, which can be divided into three sub-categories: statistical, linguistics, and hybrid (i.e., statistical and linguistic). Furthermore, they made a brief review of different corpora in different domains that have been used in term extraction, but only for Brazil Portuguese. Regarding monolingual rich-resourced languages such as English, [5] presented a survey of existing notions of a term and its linguistic features. They formulated the definition of automatic term extraction, analyzed available approaches, and proposed a general ATE pipeline consisting of four consecutive steps: preprocessing, term candidate collecting, term candidate scoring, and term candidate ranking.

Besides these initial survey studies, the systematic comparison of various ATE systems on the same corpora was also conducted in the scope of several shared tasks and workshops. NTCIR [50] is an evaluation-based project for information retrieval and term extraction carried out between 1998 and 1999. The term recognition part of this project tackled three subtasks, including the initial term extraction, keyword extraction, and role analysis tasks. However, this study suffered from a limited number of participants and the absence of previous evaluation initiatives for computational terminology. Later, CoRReCT [31] introduced an interesting data set with two proposed systems to detect the candidate terms (i.e., FASTR and SYRETE) and a new protocol to evaluate term detection, the so-called controlled indexing. A few years later, the Campagne d'Evaluation des Systèmes d'Acquisition des Ressources Terminologiques (CESART) [30], presented term extraction as one of three subtasks with a more user-oriented evaluation. This project proposed an interesting new protocol for term extraction, including a gold standard list of terms and a corresponding domain-specific acquisition corpus. This setting is considered as one of the better baseline approaches toward term annotation in the latest term extraction studies.

The TermEval 2020 shared task on monolingual automatic term extraction, organized as part of the CompuTerm workshop [87], presented one of the first opportunities to systematically study and compare various term extraction systems on comparable corpora in English, French, and Dutch from four domains (Corruption, Wind, Equitation, Heart Failure). The participating systems ranged from traditional approaches based on linguistic and statistical features to systems based on shallow machine learning and neural networks. While the workshop was an important step forward in terms of systematic comparison among different methodologies, there is still room for improvement as the open-source code is not available for most participating systems, hindering their replicability.

As of yet, to the best of our knowledge, there is no comprehensive survey of modern neural network ATE systems, nor is there a survey that compares systems based on extensive feature engineering and neural systems in both multi-lingual and multi-domain settings. Recently proposed approaches, which in most cases rely on transferring knowledge from one domain to another or from one language to another, have not been thoroughly evaluated and a comparison between these systems and other ATE systems is still lacking.

3 METHODOLOGY

We searched mainly on the five websites, including Google[1], Google Scholar[2], Semantic Scholar[3], ACL Anthology[4], and Terminology[5] journals to identify suitable articles for this survey. The search progress was taken from October to December 2022. Our query terms covered two distinct ATE resources: available corpora and neural systems. Therefore, we used 80 different queries to collect the papers that describe either methods/systems (16) or corpora (64), which are the combinations of different keywords as illustrated in Fig. 1 and are listed as examples below:

- Methods/Systems: *automatic term extraction, automatic term recognition, automatic term detection, automatic term retrieval, automated term extraction, automated term recognition, automated term detection, automated term retrieval, automatic terminology extraction, automatic terminology recognition, automatic terminology detection, automatic terminology retrieval, automated terminology extraction, automated terminology recognition, automated terminology detection, automated terminology retrieval,...*
- Corpora: We combined the above keywords with each of the four following keywords: *data, dataset, corpus, corpora.* For example, *automatic term extraction data, automatic term extraction dataset, automatic term extraction corpus, automatic term extraction corpora*, to mention a few.
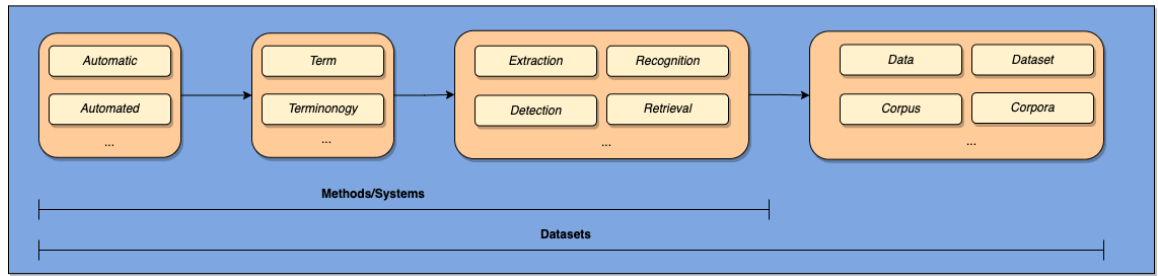

Fig. 1. The combination of keywords to search for term datasets and systems.

We sorted the resulting papers obtained by each query according to the citation counts. After collecting based on queries, we filtered out all the papers that have less than five citations if the papers were released before 2022. After this filtering step, we continued to select only the papers that either introduced a neural architecture for automatic term extraction or described a novel ATE dataset to investigate in our survey. There were no limits in the languages and the domains of the dataset that the papers aimed to solve the tasks. When it came to articles presenting neural architecture, we included them in the survey only if they represented the first article to introduce it; otherwise, we traced citations back until we found the original paper that introduced the architecture. We followed the same citation-tracing approach when searching for ATE systems that employ extensive feature engineering for performance comparison. In total, 329 articles were collected, 139 articles were reviewed after filtering, and 112 articles were finally selected from the point of view of the dataset, code, and tool availability after applying the citation-tracing mechanism for the survey.

---

[1]https://www.google.com/
[2]https://scholar.google.com/
[3]https://www.semanticscholar.org/
[4]https://aclanthology.org/
[5]https://benjamins.com/catalog/term

4 ATE DATASETS

In this section, we cover linguistic resources for term extraction developed in the past three decades. More specifically, we present and analyze monolingual and multilingual domain-specific gold standard corpora that can be used for the training and evaluation phase of ATE systems. We only consider resources that are publicly available and for which a description paper explaining their construction has been published.

4.1 Monolingual corpora

Table 1 lists manually annotated monolingual corpora that: (1) have been used for monolingual ATE evaluation, (2) contain more than 100 terms, (3) have annotation or gold standards, and (4) are publicly available at a specific URL or easily to reconstruct based on the data description. Note that the years on the Table refer to when the corpora were released and that the statistics are up-to-date with the latest release. We attach the link of each corpus in the footnote for facilitating access.

GENIA [53] is one of the best-known biomedical English corpora that was used for evaluating monolingual term extraction in many studies [33, 55, 56, 76, 107, 111]. The version 3.0 of GENIA was a collection of 2000 abstracts in the articles extracted from MEDLINE database. 93,293 out of all 436,967 words were annotated as biological terms. This corpus was the first publicly available corpus with linguistically rich annotations that include sentence boundaries, term boundaries, term classifications, semi-structured coordinated clauses, recovered ellipsis in terms, etc.

Another example of a bio-text-mining dataset is the Colorado Richly Annotated Full Text Corpus (CRAFT) [7]. CRAFT is a collection of full-text, open-access biomedical journal articles, each of which is a member of the PubMed Central Open Access Subset[6]. The latest public release (version 2.0) includes around 100,000 ontologies/terms annotated from 67 out of the 97 articles in seven different categories: Chemical Entities of Biological Interest, Cell Ontology, Entrez Gene, three sub-ontologies of Gene Ontology, NCBI Taxonomy, Protein Ontology, and Sequence Ontology. Further details are described in CRAFT website[7]. However, both GENIA and CRAFT consisted of applying ontologies on text, and not of labeling terms as an open task.

In the context of the Gene Ontology project[8], the Gene Ontology corpus [17] provides structured controlled vocabularies of molecular and cellular biology that are freely available for community use in the annotation of genes, gene products, and sequences. Between 2005 and 2022, a total of 7,694,564 annotations was made, with 43,329 valid terms, 4.024 obsoleted terms, and 2,438 merged terms. These terms were organized in a hierarchical manner under three sub-ontologies, including Molecular Function, Biological Process, and Cellular Component. Further statistical details about different versions from the beginning of 2018 up to October 2022 can be accessed via the Gene Ontology website[9]. A few approaches to term extraction have been tested on this corpus, for example, [24, 112].

The ACL Reference Dataset for Terminology Extraction and Classification (ACL RD-TEC) is a dataset for evaluating the term extraction and classification from literature in the domain of computational linguistics. Derived from the ACL Anthology Reference Corpus, ACL RD-TEC consists of two versions: ACL RD-TEC v1.0 and ACL RD-TEC v2.0. The 1.0 version [85] was released in 2014, providing 82,000 manually annotated candidate terms of three types: valid, invalid, or technology terms. Successive to the ACL RD-TEC v1.0, the ACL RD-TEC v2.0 [84] embraces 300 unique abstracts from articles in the ACL Anthology Reference Corpus, published between 1978 and 2006. The annotated terms are classified

---

[6]https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[7]https://bionlp-corpora.sourceforge.net/CRAFT/
[8]http://www.geneontology.org/
[9]http://geneontology.org/stats.html

into one of the seven categories: Method, Tool, Language resource, Language resource product, Model, Measures and measurements, and Other. Both versions are experimented with and evaluated by several approaches, such as linguistic and statistical ones [104], machine learning ones [42, 72, 104, 107], or even deep learning ones [39, 56, 58].

Regarding the Brazilian Portuguese language, the number of available Brazilian Portuguese corpora has increased significantly in different domains, including and excluding well-constructed gold standards. In our survey, we focus only on the accessible corpora that have well-annotated labels or gold standards. For example, the Ceramic Coating corpus [93] collected texts (448,352 words in total) from 164 papers in the Industrial Ceramics Magazine. Each paper contains 4 to 8 pages on average (approximately 4,000 words per paper). In the domain of pediatrics, JPED [18] includes 283 texts from the Pediatrics Journal (Jornal de Pediatria) and a gold standard of 1,534 bigrams and 2,647 trigrams. Another popular corpus is ECO [108]. This available corpus, constructed in the context of the BLOC-Eco project, contains 390 Portuguese documents pertaining to the domain of Ecology. In the context of Nanoscience and Nanotechnology, the N&N [13, 26] corpus includes 1,057 texts, divisible into 5 categories: scientific, divulgation-scientific, informative, technical-administrative, and others. EaD [92] contains 347 texts extracted from the Internet in the field of Distance Education. Further information about all the Brazilian Portuguese-related corpora is described in [20]'s work.

Table 1. List of monolingual term datasets.

Datasets	Years	Domain(s)	Language(s)	No of docs	No of words	No of terms
GENIA[10]	2003	Biomedicine	English	193	436,967	93,293
GO[11]	2005	Gene	English	-	-	43,329
CRAFT v2.0[12]	2016	Biomedicine	English	67	> 560,000	≈ 100,000
ACL RD-TEC v1.0 [13]	2014	Computational linguistics	English	10,922	36,729,513	82,000
ACL RD-TEC v2.0[14]	2016	Computational linguistics	English	300	33,216	6,818
JPED [15]	2005	Pediatrics	Portuguese	283	-	-
ECO [16]	2005	Ecology	Portuguese	390	-	322
N&N [17]	2008	Nanoscience Nanotechnology	Portuguese	1,057	-	1,794
EaD [18]	2010	Distance education	Portuguese	347	-	118
Irish Wiki[19]	2019	Education	Irish	11	5,178	864
Hindu Wiki[20]	2022	Education	Hindu	71	11,960	953
RSDO5 [21]	2021	Ph.D. theses, Scientific book, Journal articles, Graduate textbooks	Slovene	12	257,029	37,985

For what concerns less-resourced languages, RSDO5[22] is one of the academic Slovenian corpora with well-documented annotation guidelines. It consists of 12 texts, collected between 2000 and 2019, from the fields of biomechanics, linguistics, chemistry, and veterinary science. Over 250,000 words and almost 38,000 manually annotated terms were included. Apart from the manually annotated terms, the corpus was also annotated using the Universal Dependencies annotations (e.g., tokenization, sentence segmentation, lemmatization, morphological features, and dependency syntax). The corpus can serve as training sets for supervised term extraction and for term extraction tool benchmarking. Recently, while [73] applied term extraction on 11 Wikipedia Irish documents of 5,178 words and 864 extracted terms, [8] introduced a collection of 71 Wikipedia Hindu documents comprising 11,960 words and 953 manually annotated terms.

The monolingual datasets differ considerably, as there is a lot of variation in both the annotation and evaluation. Diverse approaches have been applied to each different aspect, including annotation type (candidate term list or just source text, etc.), annotation scheme (binary or multi-label, etc.), annotation guidelines (term length, POS patterns, whether or not to add named entities, etc.). The inter-annotator agreement (IAA), a measure of how well multiple annotators can make the same annotation decision for a certain category, varies as well, and there is often a lack of meta-information to better understand the corpus. Another problem is that each monolingual dataset covers a different domain and usually covers only a limited amount of terms within that specific domain. Consequently, it is extremely difficult to make any comprehensive comparison among corpora or combine multiple datasets into an extensive corpus.

4.2 Multilingual corpora

With the development of multilingual tasks [25, 58], datasets that cover multiple languages and domains are becoming increasingly important for the training and evaluation of multilingual and cross-lingual models. Table 2 presents the available manually annotated corpora used in multilingual ATE training and/or evaluation that cover more than one language.

The Terminology Extraction, Translation Tools and Comparable Corpora (TTC) project [22, 38] was one of the first domain-specific corpora made purposefully to enable training and evaluation of multilingual term extraction models. The corpus contains manually annotated texts from two domains (Wind energy and Mobile technology) in seven languages (Chinese, English, French, German, Latvian, Russian, and Spanish). The summary for each domain can be found on this website[23].

Also, other corpora focused on specific domains in multiple languages. The BitterCorpus [4] was released in 2014 as a manually-annotated collection of parallel English-Italian documents with 874 domain-specific bilingual terms in the Information Technology (IT) domain. The documents are extracted from two corpora: the GNOME and the KDE data collections. The summary for these two corpora and the BitterCorpus itself can be found on this website[29].

TermFrame v1.0 [83, 103] is a specialized corpus of karstology literature in Slovene, Croatian, and English. The terms were extracted from 24 English and 60 Slovene documents by comparing the domain corpus to the reference corpus, and with additional processing steps (e.g., filtering based on nested terms, stopwords, and fuzzy matching). The summary for each domain can be found on this website[30].

KAS-biterm [66] is a collection of 700 bilingual Ph.D. theses (40 million tokens) in the KAS corpus about Slovene academic writing of three domains: Chemistry, Computer Science, and Political Science. The sentences in the corpus were filtered by noSketch Engine to keep only those with a high chance of containing the term in the original language

---

[22]https://www.clarin.si/repository/xmlui/handle/11356/1400
[23]http://www.ttc-project.eu/
[29]https://mt4cat.fbk.eu/benchmarks/bittercorpus
[30]http://termframe.ff.uni-lj.si/

Table 2. A list of multilingual term datasets.

Datasets	Years	Domain(s)	Language(s)	No of docs	No of words	No of terms	No of entities
TTC[24]	2012	Wind energy, Mobile technology	Chinese	270	238,299	293	-
			English	209	51,699	269	-
			French	-	-	-	-
			German	-	-	-	-
			Latvian	-	-	-	-
			Russian	-	-	-	-
			Spanish	-	-	-	-
BitterCorpus[25]	2014	Information Technology	English	56	-	874	-
			Italian	56	-	874	-
TermFrame v1.0[26]	2019	Karstology	Slovene	60	987,801	2,946	341
			Croatian	43	969,735	-	-
			English	54	1,968,509	4,397	156
KAS-biterm[27]	2019	Academic writing	Slovene	176	-	1,895	-
			English	176	-	2,138	-
			others	176	-	168	-
ACTER v1.5[28]	2020	Corruption, Dressage, Heart failure, Wind energy	English	248	222,281	5,534	1,333
			French	309	240,207	4,941	789
			Dutch	259	218,989	5,454	884

and its translation into Slovene. The terms were labeled using three labels (terms, partial terms, abbreviations) in three categories of languages (Slovene, English, or other languages). The summary for each domain can be found on this website[31].

Annotated Corpora for Term Extraction Research (ACTER) dataset [87] was the first meticulously annotated corpus that covered multiple languages and domains and that offered available up-to-date documentation and transparent annotation guidelines. Texts from four different domains (Corruption, Dressage, Heart failure, and Wind energy) and three languages (English, French, and Dutch) were manually annotated. The texts for a specific domain are comparable across languages, meaning that texts from the same domain but across different languages cover the same subject. While most texts cannot be aligned as parallel translations, the texts from the corruption domain are sentence-aligned. Furthermore, ACTER solves the debate about whether or not to consider named entities as terms by providing two different versions of manual annotations: one containing only terms, and the other containing both terms and named entities. As of yet, these characteristics make ACTER the most useful resource for benchmarking and method comparison in both monolingual and multilingual ATE. The summary for each domain can be found on this website[32].

While we investigated different available corpora, we figured out that most of the original papers fully describe the primary version of the dataset. Some papers included both data and method descriptions but rarely documented data versions of the dataset and whether an entire dataset was used or only a part of it. As a result, given the same corpora, it is hard to make a thorough comparison among methods given they use different data versions, especially when the latest version has a significant update compared to the previous one.

---

[31]https://www.clarin.si/repository/xmlui/handle/11356/1199
[32]https://github.com/AylaRT/ACTER

5 SUPERVISED ATE SYSTEMS

In this section, we offer an extensive overview of recent ATE systems, which tend to rely either on shallow machine learning techniques or deep neural networks. In most cases, traditional approaches to term extraction do not rely on machine learning. Instead, they tend to extract candidate terms above a certain threshold, derived from several linguistic and statistical features. Traditional approaches are not covered in this survey, since they have been covered in previous surveys (e.g., [51] and [5]). Instead, in this section, starting from the advent of machine learning up until now, we present machine learning-based (neural and non-neural) ATE systems, that have not been systematically described and compared in any recent survey study.

5.1 Machine learning approaches

Despite several variations in features and models, most supervised machine learning approaches to ATE comply with the traditional approach, following three principal steps: (1) preprocessing, (2) feature engineering, and (3) term extracting classifier. In the preprocessing step, some operations are performed on input texts (e.g., sentence segmentation, word segmentation, and POS tagging) to make the input text ready for further steps. In the feature engineering step, we describe the candidate terms through several different features (see examples of different types of features in Figure 2). In the final steps, the features are fed into the machine learning classifiers so that the classifiers can learn from the training set and use that knowledge to make predictions on new unseen texts.

Due to the relatively low accuracy, these first machine-learning approaches were mostly used to complement approaches based on hand-crafted rules. This, along with the success of traditional systems (e.g., TermoStat [29]), which relied on several linguistic and statistical features, lead to the idea of combining different types of information. That is to say, multiple linguistic and statistical termhood indicators were fed as features to a variety of machine learning algorithms. While several NLP tools (e.g., tokenization, lemmatization, stemming, chunking, POS tagging, and full syntactic parsing) are employed in this approach to obtain linguistic profiles of term candidates, numerous statistical measures are also applied to this approach, including the termhood [101], unithood [23], C-value [36]. Regarding machine learning algorithms, the most popular algorithms used for ATE include AdaBoost [12], ROGER evolutionary algorithm [6], RIPPER rule induction [35], CRF++ [48], K-nearest neighbors [86], Logistic Regression [11, 27, 32, 68, 77], Decision Trees [52], and Support Vector Machines [65].

An example of a machine learning approach employing extensive feature engineering and several classifiers is given in [16]. In this study, the authors proposed to select statistical and linguistic features and feed them into different machine learning classifiers (e.g., JRip, Naive Bayes, J48, or SMO from WEKA). Several post-processing steps were carried out (e.g., word variation standardization, POS annotation, and stopword and punctuation removal) before the selection of the final candidate terms.

[107] instead proposed a machine learning method with common features (e.g., term frequency, c-value, weirdness) extracted from the token n-grams (n=1,2,3,4,5) excluding all the stopwords. The selected features were fed into several different classifiers, namely Random Forest, Linear Support Vector Machine, Multinomial Naive Bayes classifier, Logistic Regression, and SGD classifiers.

With the advent of the newly annotated ACTER corpora, one of the TALN-LS2N approaches [44] used the combination of different meaningful information—such as linguistic, stylistic, statistic, and distributional descriptors—to generate the feature vectors. It then made use of the XGBoost model to learn several classifiers, which were weighted according to their performance and aggregated iteratively.
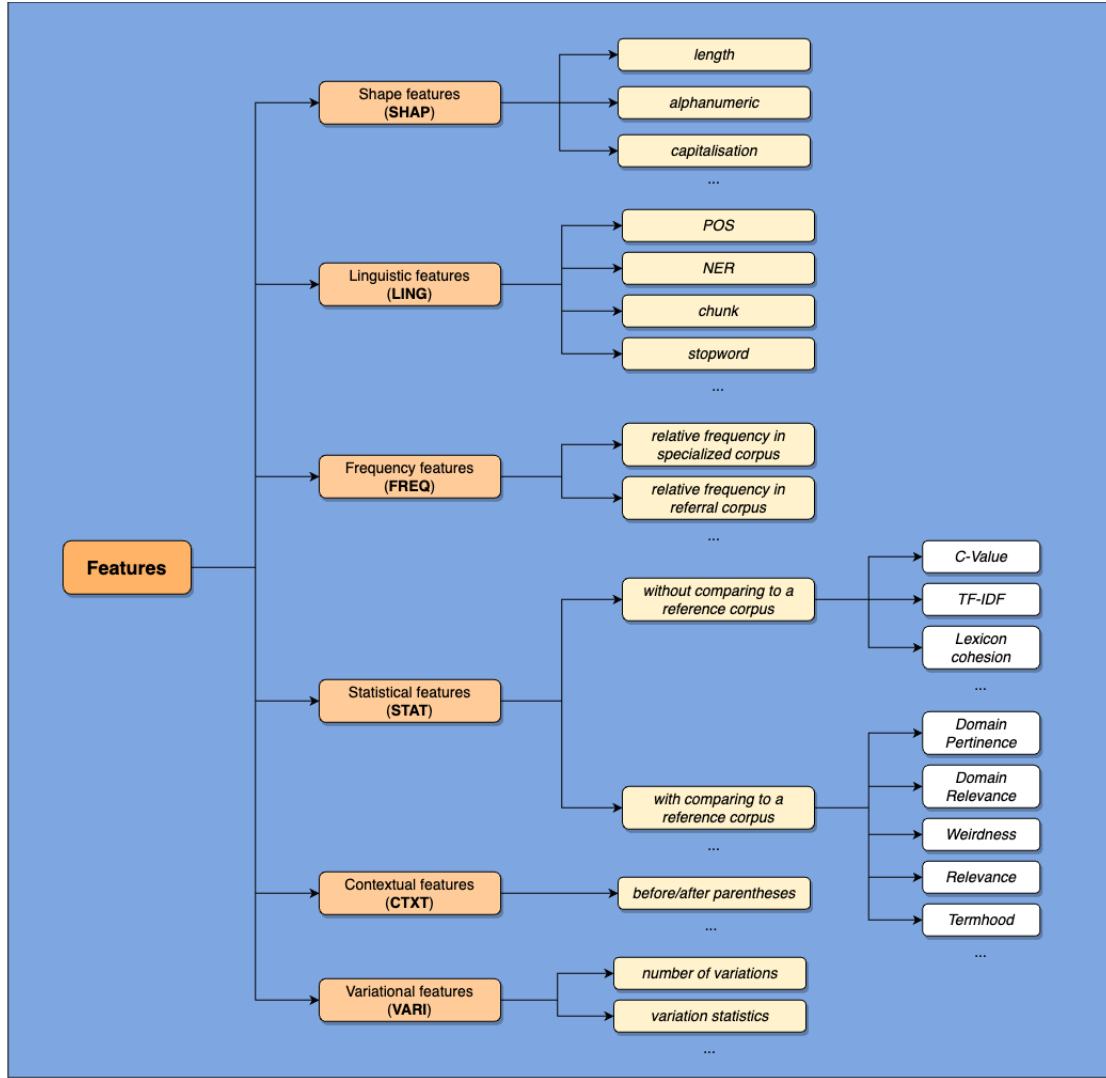
Fig. 2. Feature group and subgroup for machine learning models based on [88]'s work.

The HAMLET approach [88] included a relatively wide range of supervised machine learning methods (e.g., Decision Tree, Random Forest, Multi-layer Perceptron, and Logistic Regression) and relied on a total of 152 features from six different feature groups: morphological, frequency-based, statistical, relational and linguistic, and corpus-based. The Binary Random Forest classifier performed the best out of all tested classifiers.

Furthermore, [78] proposed a new approach that combines probabilistic topic modeling (PTM) and non-negative matrix factorization (NMF). They compared five different NMF algorithms and four different NMF initializations, and found optimal combinations of NMF to compare with the extraction baseline (e.g., TF-IDF, RAKE, YAKE, and TextRank).

5.2 Deep learning approaches

The use of neural networks, especially language models, to solve ATE, has been gaining a lot of traction recently. Their application is performed either to just represent the information in the text with word embeddings or to apply a deep architecture as an end-to-end classifier. In this section, we describe studies that employed any of these two approaches.

5.2.1 Embeddings. The embeddings for term extraction are often pretrained on a large general corpus, and then potentially fine-tuned on specific corpora during classification. The most popular general non-contextual word embeddings that were used in several ATE studies [2, 56, 59, 111] are GloVe embeddings[33] [81]. The second most popular option is constituted by domain-specific Word2Vec embeddings [75], either employing the CBOW or the skip-gram architecture [2, 10, 104] and FastText [94].

The first study to successfully apply word embeddings for the ATE task was the one by [1]. The embeddings, pretrained on a general corpus, were initially used in a filtering step on top of an existing rule-based tool TermoStat [29], and later as a standalone term extraction tool [2]. Meanwhile, SemRe-Rank [109] made use of Word2vec embeddings to create semantic relatedness graphs of words, thus computing their "semantic importance" scores. These scores were then used to revise the base scores of term candidates computed by another ATE algorithm.

Several papers also researched domain-specific embedding representations for ATE usage in a specific field, for example, as [10]'s research in medicine. Meanwhile, some studies attempted to combine domain-specific and general embeddings. [2] explored the idea of concatenating general and domain-specific embeddings as an input (in English) for a Multi-Layer Perceptron (MLP) from the 1M+ token mathematics corpus. Similarly, [43] tested two new (neural) approaches to exploit general vs. domain-specific comparisons. They used a simple neural network model with pre-computed Ge-embedding information as input, and a multi-channel model computing the comparison internally. These embedding concatenation techniques demonstrated that any strategy using both types of embeddings performed better than the strategies only using either the general or the domain-specific ones.

Considering the importance of general knowledge encoded in the pretrained language model and contextual information for the downstream task, the employment of several contextual embeddings for the ATE task was proposed. This is the case, for example, of the (stack) bi-directional Flair embeddings [94] with a neural, character-based language model from the FlairNLP framework[34] for multiple languages that incorporate subword information, and newly adopted Transformer-based (monolingual and multilingual) embeddings, some of which were already tested in term extraction (e.g. BERT embeddings [3, 94] and its variants), or the stacked Flair + BERT embeddings [94].

5.2.2 ATE Classifiers. The use of deep neural networks in ATE is not only limited to the generation of embedding representations. Neural architectures are also used as end-to-end term extractors. Based on the methodology, we can divide neural extractors into three main types, as demonstrated in Figure 3: (1) sequence classifiers; (2) token classifiers; and (3) sequence-to-sequence generation models, which are used for several generation tasks (e.g., summarization, translation, and headline generation).

When deep neural models were first adopted for ATE, ATE was considered as a sequence or binary classification task. The resulting approaches relied on the positive (is-a-term) and negative (not-a-term) samples, which are generated from all possible n-grams of a fixed length of a given sentence. TALN-LS2N [44] was the first team to propose this, using BERT-based language models in the recent ATE challenge TermEval 2020. Their approach outperformed all classification methods based on feature engineering. They tested different versions of BERT (RoBERTa for the English

---

[33]https://nlp.stanford.edu/projects/glove/
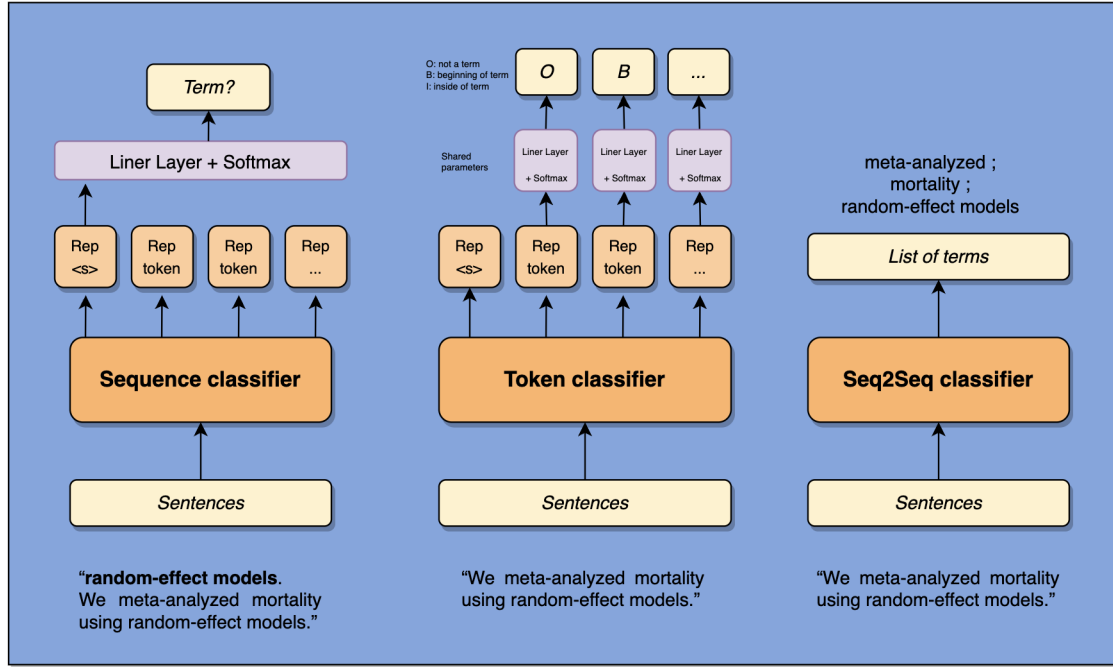[34]https://github.com/flairNLP/flair

Fig. 3. An example of how three types of neural ATE classifiers work (from left to right: Sequence classifier (e.g., XLMR); Token classifier (e.g., BERT); Seq-2-seq (e.g., mBART).

corpus, CamemBERT for the French corpus) as binary classifiers for term prediction. The input consisted of a sentence and a selected n-gram within the sentence. If the n-gram was a term, the input was labeled as a positive training example, otherwise as a negative example. Using the same approach and training datasets, [58] improved the performance of the ATE binary classifier by employing a multilingual version of RoBERTa, the so-called XLM-RoBERTa (or XLM-R, for short) and compared it with other neural approaches. Although this approach proved to have better predictive power in comparison with other machine-learning-based approaches, generating all possible n-grams contained in each sentence across all documents in order to create the training set is computationally expansive and space-demanding. For this reason, later studies decided to shift from this approach toward sequence labeling or token classification.

The latest trend is the use of language models and sequential methods where the candidate terms are detected in their original contexts, usually by classifying each token in the text sequence as (part of) a term or not. Thus, instead of having a label for the whole text sequence, the token classifier assigns a label for each token. This sequence labeling approach was first tested by [48], who proposed a non-neural conditional random field (CRF) classifier to investigate the utility of different types of features. [104] introduced co-training—a weakly supervised bootstrapping paradigm that helps in stabilizing classifier's predictions—using a long short-term memory (LSTM) network and a convolutional neural network (CNN). Alternatively, [56] applied a token classifier by taking advantage of recurrent neural networks (RNN). [40] proposed the CNN-BiLSTM-CRF model to minimize the influence of different word segmentation results on Chinese term extraction. Another attempt at deep learning based sequential labeling for ATE task is presented in [3]. This study experimented on several models, such as LSTM, BiLSTM, LSTM-CRF, and BiLSTM-CRF, with the FastText

and BERT embeddings, to obtain a high evaluation score for ATE in the cybersecurity domain from the less-resourced Lithuanian language.

Regarding more recent competitions, NLPLab UQAM [59], the team which won in the use of TermEval's Dutch corpus, also modeled the task as token classification by applying a bidirectional LSTM (biLSTM) with pretrained GloVe word embeddings. The monolingual version of D-Terminer [89] and [45] took advantage of the Flair framework to propose a feature-based strategy that consists in using BERT embeddings as input features for a biLSTM with a CRF layer. Given the same dataset, [94] made a comparison between the neural sequence-labeling approach and the traditional feature-based approach, where they proposed the evaluation of three-word embeddings (i.e., FastText, Flair, and BERT) in monolingual and multilingual settings.

With the advent of Transformer-based models, several pretrained language models have been applied as token classifiers to boost the extraction performance [45, 58]. Above all, XLM-R [14] was applied and mentioned in several publications and is now considered a benchmark for several languages [98]. Since this approach operates without upfront n-gram generation and handles each sentence as a single example that needs to be labeled, it is considerably more time efficient than the previous sequence classification approach.

Despite the widespread use of Transformer-based sequence-to-sequence generation models for NLP, the adaption of self-supervised pre-training approaches for ATE has just recently begun to gain traction. [58] was the first to employ the sequence generation model mBART [64]) to transform the input sentences to sequences of comma-separated terms. This approach was inspired by the neural machine translation-based ontology learning proposed by [82]. While promising, the performance and applicability of this approach require additional testing and remain an unsolved issue.

5.3 Multilingual learning

Recently, a new research direction in ATE aims to conduct term extraction in a multilingual setting, testing the hypothesis that multilingual information can help during the monolingual extraction phase.

Multilingual term extraction can be performed on parallel or comparable corpora. According to [74]'s definition, a comparable corpus is composed of monolingual data collected from different languages using the same sampling techniques. Meanwhile, a parallel corpus is one that is composed of source texts and their translations in one or more different languages. While the comparable corpora are not translations, they do share some vocabulary (e.g., cognates) due to the fact that they cover the same topic. This can be exploited to extract equivalent terms found in parallel corpora, usually by leveraging term alignment approaches.

Regarding parallel corpora, [34] proposed to divide the methods into two groups: "align-extract" and "extract-align", depending on whether monolingual candidate terms are extracted first, or whether the alignment is performed first. Several tools have been proposed (e.g., the TExSIS tool [71]) for bilingual automatic term extraction from parallel corpora. The TExSIS tool extracts word alignments with GIZA++ [79] and applies the rule-based chunking [70] with alignment progress and additional statistical (termhood) filtering to extract the candidate terms. Besides GIZA++, some of the most popular methods for the alignment [47] are Moses phrases tables [54], Awesome Align neural word alignment [28], and ASTrED aligned syntactic tree edit distance [99], to mention a few.

Despite the significant efforts that have been made to create multilingual corpora that cover more than one domain, parallel multilingual datasets have been rarely created for less-resourced specialized domains. This is due to the fact that the creation of these datasets tends to be error-prone, labor-intensive, and time-expensive, and was also criticized in the past for subjectiveness. Instead, most of the multilingual corpora are comparable, and multilingual language models are commonly used for term extraction on these corpora. An example of this kind of approach is HAMLET [88].

[88] proposed six different settings, in which they used 5-fold cross-validation to train and test on a single corpus (one domain, one language), single language (all domains, one language), or all corpora (all domains, all languages), with or without POS patterns. [98] leveraged the multilingual setup by fine-tuning the model using training datasets from all the languages in the ACTER corpora (English, French, and Dutch) and the Slovenian RSDO5 corpus, and then applying the model to the test sets of all languages. By doing so, they examined whether adding more data from other languages to the training set that matches the target language in the testing set improves the predictive performance of the model.

5.4 Cross-domain and cross-lingual transfer learning

Current research on term extraction deals with two major issues: (1) the data collection for a specific field of expertise or domain is often non-existent or small, and (2) the available well-annotated corpora are scarce, especially for lesser-known languages (e.g., Slavic languages). To address the former issue, [63] proposed to combine specialized and general domain word embeddings to increase the training corpora. However, this technique is hard to apply for most ATE tasks, since it still requires a "reasonable" amount of domain-specific manually annotated data, which is scarce (i.e., the second issue). To solve both problems, cross-domain and cross-lingual approaches have been recently introduced.

5.4.1 Cross-domain learning. Zero-shot and few-shot learning recently gained in popularity as advanced approaches to solve the scarcity of available annotated corpora. They rely on transferring knowledge the model obtains from a well-resourced domain/language with plenty of annotated data to a less-resourced domain/language with scarce or no data. This way, cross-domain learning can also be used as a way of checking the generalization capabilities (and therefore usefulness) of the model, i.e., how successfully the knowledge that the model obtains on one domain can be applied to the arbitrary new unseen domains.

[45] was the first to propose to fine-tune several Transformer-based language models and to test them in a cross-domain scenario. In all the scenarios, two domains were used for training, one for validation, and one for testing. The authors proposed two evaluation settings: 1) fixing the training data sets while varying BERT models to test different model versions, and 2) fixing the models while varying the training data sets to test the success of different cross-domain transfers. The results suggested that while cross-domain learning for ATE is possible, the models are still domain sensitive, and a careful choice of training datasets is necessary to achieve competitive performance.

[58] used the same cross-domain setting with the same corpora, and experimented with three different Transformer-based term extraction models operating on sentence level: a language model for token classification; one for sequence classification; and innovative use of sequence to sequence (seq2seq), which learns to reduce sentences to terms. [98] tested the same cross-domain setting on a less-resourced language, Slovenian (RSDO5 corpus), using a multilingual model. They showcased a generalization of cross-domain transfer learning also on Slovenian. Finally, [97] employed several Transformer-based multilingual and monolingual model ensembles in a cross-domain setting on both the ACTER and RSDO5 dataset, further improving the results of both previous studies and confirming the success of the cross-domain transfer.

5.4.2 Cross-lingual learning. Similarly to cross-domain learning, cross-lingual learning is based on transferring knowledge obtained by the model from annotated data in a high-resourced language to the input data in a less-resourced language. We examine how well the ATE model performs without the language-specific training corpus and how good the knowledge transfer between different languages is.

[45] tested several multilingual fine-tuned Transformer language models in a zero-shot cross-lingual setting. They did so by fine-tuning the model on the ATE task in one or several languages, and by testing them on a new target language,

on which the model was not fine-tuned. The results indicate a successful transfer, which led to achieving competitive results on the ACTER dataset and opening a new promising direction in ATE. [58, 97] further extended this work with a new benchmark language model, namely XLMR [14], testing its performance in a cross-lingual setting and using the sequence labeling approach. The results re-confirmed the potential of transferring knowledge from rich-resourced to less-resourced languages in the field of ATE since this technique allows us to beat the previous state-of-the-art methods on the ACTER dataset.

6 ATE EVALUATION METRICS

ATE models usually provide a list of candidate terms from the given domain-specific corpus as the final output, and it is important to define the correct methods and metrics to evaluate the quality of this output. There are several variations among the evaluation methods, including both intrinsic and extrinsic mechanisms. The extrinsic methods assess the quality of the extracting system by measuring the improvement in the performance of another system or application that uses the results of term extraction as described in [102]. Meanwhile, the intrinsic ones measure the quality of term extractor by evaluating some intrinsic properties, which is independent of their intended planned use. We can classify the direct evaluation approaches into two main aspects: the evaluation methodology and the scope of the results [110] as shown in Figure 4. While the evaluation methodology focuses on whether the evaluation is performed by human judges, or whether it is a dictionary-based or gold standard based, the scope of results answers the question of whether we evaluate the entire results, parts of the results, or the top-k results. Due to this variety in evaluation types, the performance of different approaches is often not directly comparable.

Regarding the evaluation methods, in the initial studies about ATE, having a human judge was one of the first approaches to evaluate how well the candidate terms were extracted. [49] evaluated their system by asking a domain expert (i.e., a terminologist) to judge whether the extracted candidates were domain-specific terms.

In order to reduce the human effort, the dictionary-based evaluation was proposed either in the form of a list or as a dictionary of the pre-defined criteria to map with the extracted terms. [51] took advantage of statistical features, e.g., termhood and unithood, for evaluation. [61] applied five pre-evaluation criteria from the basic design evaluation to present how the results are shown in the candidate lists. [71] identified the terms with linguistic preprocessing steps and matched the candidate terms to a pre-defined dictionary of POS patterns.

Another popular approach is based on reference term lists, or the so-called gold standard corpora, which can be an adaptation of a pre-existing list, a sample (seed terms), or the list of all the terms in the corpus. The pre-existing term list [27, 105] is not collected directly from the training corpora but is an already existing and community-wide terminology. Thus, the term extraction approaches using the corpora with this gold standard often evaluate how many candidate terms were actual terms but failed to measure how many terms in the text were correctly extracted. Meanwhile, the approach of considering a sample as the gold standard [9, 67] took advantage of a web crawler to collect the domain-specific texts (e.g., TTC project). The crawler took a list of domain-specific words, the so-called seed terms, as input, and as outputs, the texts found on the Web in the domain of interest. The seed terms are usually term representative of the domain for which we want to retrieve the web documents. To resolve the existing issues of the two mentioned types, the list of all the terms in the corpus [53, 87] was used as a gold standard where the list was annotated directly from the corpora. This became a benchmark for manual annotation in term extraction until now. If this gold standard corpus is used, the final evaluation score is usually obtained by calculating Precision (i.e., how many candidate terms were actual terms), Recall (i.e., how many terms in the text were correctly extracted), and F1-score (i.e.,
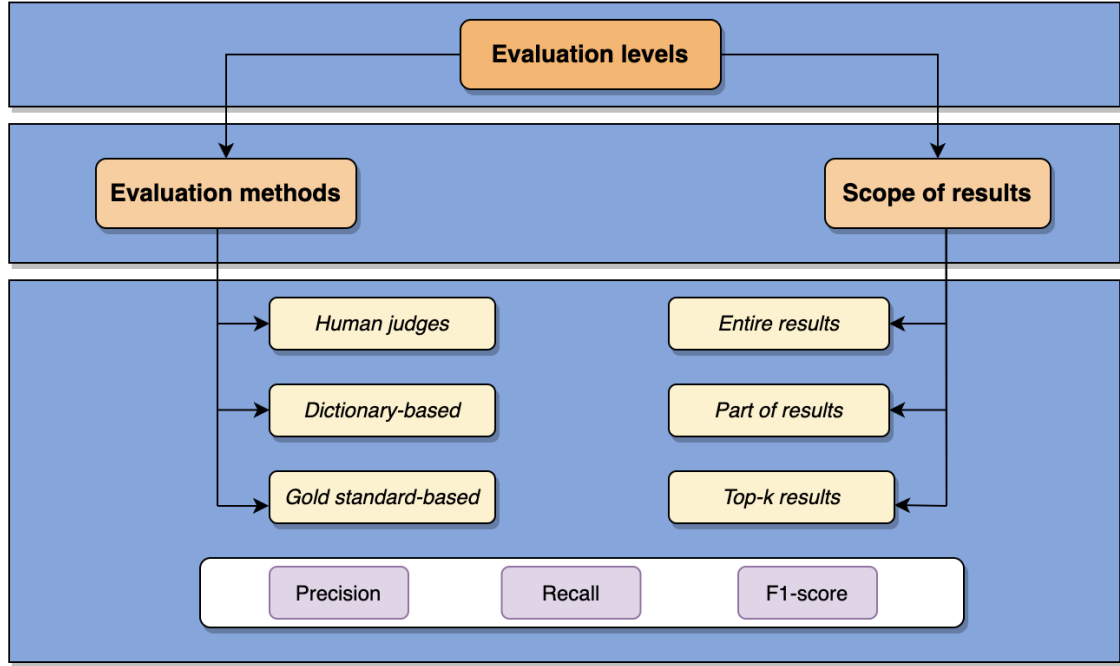
Fig. 4. The overview of different evaluation metrics in ATE task.

the harmony mean of Precision and Recall) for the obtained list of term candidates. We will specify these metrics when discussing the evaluation from different scopes of results.

Finally, besides the performance of the ATE system in terms of the evaluation metrics described above, some claim that other factors, such as the consistency of the predicted candidate terms, are also important. For example, [90] proposed that the measures of a model's quality should also concern reliability, efficiency, maintainability, usability, and portability.

Regarding the scope of results, the most common ATE evaluation approach is to compare the candidate term lists obtained by the system against the list of terms extracted by human annotators (i.e., the so-called gold standard corpus [14, 53, 87, 88, 97]). This is done by calculating Precision (i.e., how many candidate terms are actually terms, see Equation 1), Recall (i.e., how many terms in the text were correctly extracted, see Equation 2), and F1-score (i.e., the harmonic mean between Precision and Recall, see Equation 3).

$$Precision = \frac{Number\ of\ correctly\ terms}{All\ the\ candidate\ terms} = \frac{True\ Positives}{True\ Positives + False\ Positive} \tag{1}$$

$$Recall = \frac{Number\ of\ correctly\ terms}{All\ terms\ in\ the\ corpus} = \frac{True\ Positives}{True\ Positives + False\ Negative} \tag{2}$$

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Several ATE systems use the same evaluation approach as the Keyword Extraction task, i.e., they choose top-k best candidate terms for evaluation (top-100 candidate terms [46, 57], top-300 candidate terms [101], top-500 candidate

terms [21]). [71] and [111] evaluated the best-k candidate terms by using a variable k. Most of these studies employed Precision for evaluation of the best-k terms [29, 41, 71, 91, 111].

One notable insight which emerged from our literature review is that the number of studies using just Precision is higher than the number of studies that evaluate the system according to all three evaluation criteria (i.e., Precision, Recall, and F1-score). This is due to the relative lack of available gold standard data in the past. However, thanks to the more recent effort of constructing and manually annotating domain-specific datasets for ATE, most of the current systems employed on contemporary benchmark datasets (e.g., ACTER) are evaluated according to all three metrics [87, 88].

7 COMPARATIVE RESULTS

Since term extraction methods vary greatly in regard to definition, corpora, domains, languages, and evaluation metrics, a comparative evaluation of term extraction methods is hardly achievable. Nevertheless, in this section, we propose to measure the performance of several ATE systems by comparing the candidate term list extracted on the whole test set level with the manually annotated gold standard of each domain using a strictly matching F1-score. All approaches are applied to the newest and the most popular up-to-date ATE corpus, namely the ACTER dataset. We chose this dataset for evaluation because it is the most systematically annotated corpus that covers multiple languages and domains, and that also contains available up-to-date documentation and transparent annotation guidelines. The results of the evaluation are presented in Table 3, including both the approaches evaluated on the ACTER dataset in the related work (non-neural models and some deep learning models) and our own experiments on the same dataset.

Note that deep learning-based approaches are divided into three different categories based on how they took advantage of languages in the training corpus. While monolingual deep learning models use a single language for training and predict on the new unseen data from the same language, multilingual deep learning models are fine-tuned on multiple languages at once (e.g., on a combination of English, French, and Dutch data in the ACTER corpus) and generate a prediction on the new unseen data from all the languages. Finally, cross-lingual deep learning models are trained on one (or more) languages and generate predictions on unseen data from a new language that the model did not encounter during the training.

TALN-LS2N, RACAI, NYU, e-Terminology, and NLPLab_UQAM were all competitors in the TermEval 2020 shared task on the ACTER dataset [87]. The TALN-LS2N approach is the winning solution on the English and French datasets, whereas NLPLab_UQAM is the winning team on the Dutch corpus. While RACAI, NYU, and e-Terminology rely on feature engineering and statistical approaches, TALN-LS2N and NLPLab UQAM applied neural network-based models. After the competition, several new approaches were developed. For example, HAMLET is a novel supervised method inspired by traditional hybrid systems such as TermoStat. Furthermore, [78] combined probabilistic topic modeling (PTM) and non-negative matrix factorization (NMF), and the optimal combinations of NMF outperformed four baseline extraction methods (e.g., TF-IDF, RAKE, YAKE, and TextRank). All the described approaches were implemented on the ACTER dataset of version 1.2, which was first released by the shared task competition. However, none of them had considered the problem as a token-level sequence labeling task and had taken advantage of the Transformer-based models, until the novel approach from [58, 98] and [98].

[58] proposed a comprehensive comparison between three Transformer-based ATE models operating on the sentence level: token classification (sequence labeling task), sequence classification, and sequence-to-sequence generation. This proved three points. First, the superiority of token classification in terms of performance, which, therefore, affirmed itself as the state-of-the-art (SOTA) with regard to the ACTER dataset. Second, the multilingual models perform better

Table 3. The F1-score evaluation of benchmark approaches on the Heart Failure test set from the ACTER corpus in three languages (English (EN), French (FR), and Dutch (NL)) and two types of annotation, with named entities (NER) and without named entities (ANN). The best approach for each category is highlighted in bold.

Methods	ACTER					
	ANN			NES		
Non-neural models	EN	FR	NL	EN	FR	NL
RACAI [87]	39.3	-	-	41.3	-	-
e-Terminology [87]	21.4	20.6	15.3	20.1	19.7	14.4
NYU [87]	31.5	-	-	30.6	-	-
HAMLET [88]	**54.2**	**60.2**	**66.1**	**55.4**	**60.8**	**66.0**
Feature-based XGBoost [45]	-	-	-	33.6	50.9	34.1
NMF [78]	33.5	30.9	30.1	33.7	30.7	30.3
Monolingual deep learning models						
TALN-LS2N [44]	45.0	45.9	-	46.7	48.1	-
NLPLab UQAM [59]	17.8	12.9	18.6	18.1	13.2	18.7
XLM-R sequence Classifier [58]	-	-	-	45.2	46.0	48.5
XLM-R Token Classifier [58]	-	-	-	58.3	52.9	**69.6**
mBART NMT [58]	-	-	-	53.2	55.9	65.2
Vanilla-biLSTM-CRF [45]	-	-	-	8.17	6.53	7.50
BERT [45]	-	-	-	48.2	-	-
CamemBERT [45]	-	-	-	-	51.1	-
BERT (NER) [45]	-	-	-	37.4	-	51.0
CamemBERT (NER) [45]	-	-	-	-	51.1	-
BERT-biLSTM-CRF [45]	-	-	-	29.5	25.6	27.4
mBERT [45]	-	-	-	45.8	47.2	-
ALBERT [96]	49.9	-	-	54.5	-	-
DistilBERT [96]	46.5	-	-	53.9	-	-
ELECTRA [96]	51.2	-	-	52.2	-	-
RoBERTa [96]	54.3	-	-	59.1	-	-
XLNet [96]	55.2	-	-	**57.8**	-	-
FlauBERTa [96]	-	38.9	-	-	51.4	-
Dutch BERT [96]	-	-	65.6	-	-	66.8
robBERTa [96]	-	-	48.2	-	-	62.2
robBERTa-v2 [96]	-	-	48.3	-	-	63.4
mInfoXLM [96]	**56.1**	50.4	**68.5**	57.6	58.0	68.6
mBERT [96]	43.0	48.4	66.3	55.0	**57.4**	67.8
mDistilBERT [96]	49.8	**51.9**	65.3	55.6	55.8	67.8
Multilingual deep learning models						
XLM-R Sequence Classifier [58]	-	-	-	46.0	46.7	56.0
XLM-R Token Classifier [58]	-	-	-	**56.2**	55.3	**67.8**
mBART NMT [58]	-	-	-	55.3	**57.6**	64.9
mBERT [45]	-	-	-	45.4	44.9	51.0
Cross-lingual deep learning models						
XLM-R Sequence Classifier [58]	-	-	-	44.7	48.1	58.0
XLM-R Token Classifier [58]	-	-	-	**58.3**	57.6	**69.8**
mBART NMT [58]	-	-	-	55.2	57.4	59.6
XLM-R Token Classifier [95]	56.0	59.9	68.5	58.0	**61.1**	68.3

than the monolingual ones. Third, zero-shot learning shows good potential in extracting the candidate terms in new unseen languages and domains. [96] extended this research with an empirical evaluation of several monolingual and multilingual Transformer-based language models, including both masked (e.g., BERT and its variants) and autoregressive (e.g., XLNet) models, on the cross-domain ATE tasks with additional ensembling two best classifiers to optimize the Recall. Their approaches were applied to the ACTER dataset of version 1.5. However, there were not many changes to actual annotations, but major updates to how the annotations were presented.

Most deep learning based approaches also proved to be very competitive and surpassed the performance of most non-neural methods by a large margin on all languages and annotation types. [58] and [98] also showcased the potential of cross-lingual learning when considering ATE tasks as a token classification problem. These approaches currently represent the new state-of-the-art methods regarding most languages in the ACTER corpus.

8 CHALLENGES AND FUTURE DIRECTIONS

With the development of deep neural networks and the advent of recent Transformer-based models, deep learning-based ATE benefited significantly from the advances of pretrained embeddings in language models. We now have the opportunity to re-examine the term extraction task, the relative challenges, and the potential future directions, without having to involve complicated feature engineering.

Considering the scarcity of well-annotated and well-documented corpora, especially for less-resourced languages, the main objective in the field is to develop a robust approach for candidate terms detection that will be able to extract meaningful candidate terms from text sequences from a variety of domains and languages. This comes with the challenges addressed in the following subsections.

8.1 Nested terms

Nested terms refer to words or phrases that are both terms themselves and parts of longer terms. This means that a shorter term may appear within a larger term and vice versa. For example, in ACTER's English Corruption dataset, the word *"confiscation"* is a standalone term that also appears as a part of other longer terms such as *"confiscation of corruption proceeds"*, *"confiscation of criminal assets"*, and *"confiscation of the proceeds of crime"*. This makes the labeling harder, as the classifier needs to infer from the context whether a specific term is part of a longer term. This leads to the most common problem of current systems, which is that instead of predicting the whole multi-word term, they predict only a shorter gold standard term nested in the multi-word gold standard term. Vice versa, the model sometimes also generates incorrect predictions containing the correct nested terms.

8.2 Generalisability

In the past decade, we have witnessed a growing data production involving different forms of data, including raw documents. However, building domain-specific corpora with manually annotated terms from the available raw texts is often error-prone, labor-intensive, time-consuming, and subjective[35]. As a consequence, annotated datasets, on which the models can be trained, are often scarce, limited, and usually do not cover many languages and domains. Therefore, it is necessary for the model to have the ability to adapt and extract terms from new languages and domains, which were not present in its training dataset. For this reason, one of the future challenges is to build models that will be able to transfer the knowledge between different domains and languages efficiently and to extract language- and

---

[35]Subjectiveness emerges due to two major reasons: the low agreement among the annotators and the lack of consensus about an annotation protocol.

domain-specific terminology even either without the language- or domain-specific training corpus, or with only a few available training examples.

With the recent advances in language models and the huge demand for real-world applications, we expect ATE to receive more attention from researchers in the future. Based on the comparative studies in this survey, we list some proposals for further directions that the research on term extraction could take.

8.3 Ensembling

Deep neural networks based on pretrained language models have a strong predictive power on downstream tasks (e.g., term extraction), and can be fed raw data directly, without any feature engineering required. However, linguistic and statistical measures of termhoodness can still contribute to the research on ATE, by being added to neural models as additional features with the aim of boosting performance. The integration of models with different combinations of non-contextual features, contextual features from language models, and linguistics and statistical measures (e.g., termhood, unithood, C-value) should be investigated. Furthermore, [96] presented the potential of ensembling two best classifiers by simply performing the union operation on the candidate term lists of the best two classifiers. Further ensembling techniques should be employed to enhance the performance in Recall and by extension in F1-score.

8.4 Scalability

Despite showing competitive performance in extracting good candidate terms, large pretrained neural language models are computationally expensive. Furthermore, new research is going in the direction of exponentially increasing the parameters of the model and the training data size in order to improve performance. While this has been shown to be a promising research direction, these models are usually too expensive for being used in production or in academia. Therefore, we believe that going in the opposite direction, i.e., making neural approaches more scalable, is also promising. We moreover propose to develop approaches with good trade-offs between performance and scalability. Model compression and pruning techniques are also good options to reduce the space and computation time required for model learning, without compromising the overall performance.

8.5 Prompting

Prompt-based methods have been successfully applied to sentence-level few-shot learning tasks such as in NER tasks [19]. Instead of replacing the language model head with the classifier head, prompt-based approaches induce the prediction from the language model by constructing a template. By re-using the masked language model objective, prompt-based approaches alleviate the problem of pre-training and fine-tuning discrepancy (caused by their different training objectives), as well as benefit the few-shot performance. Furthermore, to avoid the time-consuming issues of searching for appropriate templates, a template-free approach [69], which re-used the language model head and leads the model to predict label words through an Entity-oriented objective, was also introduced. However, few efforts have been made to apply these methods to the token-level learning tasks and ATE in specific. Therefore, we believe that there is still room for improvement in this direction.

9 CONCLUSION

Our survey summarizes the recent advances in the automatic term extraction task, covering both classic machine learning models based on feature engineering and novel neural network models that have yielded several important insights, especially with the advent of pretrained language models. We first surveyed different resources and presented a

systematic list of well-annotated monolingual and multilingual corpora for term extraction from the past three decades. Then, we presented the first systematic summary of deep learning-based approaches and made a comparison of their performance in supervised term extraction to machine learning based approaches. Furthermore, we indicated all the metrics used in term extraction and categorized them based on the evaluation methodology and the scope of results. Incorporating the lessons learned from the existing techniques, we pointed out the current challenges and proposed some suggestions for further exploration in the term extraction research.

The main findings of this survey are that neural models generally outperform machine learning models based on feature engineering by a large margin and that Transformer-based cross-domain and cross-lingual models tend to perform really well, establishing new state-of-the-art methods on the benchmark ACTER dataset. When three types of Transformer-based ATE models were implemented to operate on the sentence level, the superiority of token classification in terms of performance affirmed itself even in cross-domain or cross-lingual in comparison with the monolingual setting. Despite monolingual language models (e.g, SloBERTa) on less-resourced languages (e.g., Slovenian) captured the information about terms than multilingual ones (e.g., multilingual BERT) if named entities are not included, multilingual ones outperform the others for extracting terms with entities and adding more data from other languages to the train set that matches the target language improves the predictive performance of the model. The zero-shot learning shows good potential in extracting the candidate terms in new unseen languages and domains.

ACKNOWLEDGMENTS

REFERENCES

[1] Ehsan Amjadian, Diana Inkpen, T Sima Paribakht, and Farahnaz Faez. 2016. Local-Global Vectors to Improve Unigram Terminology Extraction. *Computerm 2016* (2016), 2.

[2] Ehsan Amjadian, Diana Inkpen, T Sima Paribakht, and Farahnaz Faez. 2018. Distributed specificity for automatic terminology extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 24, 1 (2018), 23–40.

[3] UTKA Andrius. 2020. Automatic extraction of lithuanian cybersecurity terms using deep learning approaches. In *Human Language Technologies–The Baltic Perspective: Proceedings of the Ninth International Conference Baltic HLT 2020*, Vol. 328. IOS Press, 39.

[4] Mihael Arcan, Marco Turchi, Sara Topelli, and Paul Buitelaar. 2014. Enhancing statistical machine translation with bilingual terminology in a cat environment. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*. 54–68.

[5] Nikita A Astrakhantsev, Denis G Fedorenko, and D Yu Turdakov. 2015. Methods for automatic term recognition in domain-specific text collections: A survey. *Programming and Computer Software* 41, 6 (2015), 336–349.

[6] Jérôme Azé, Mathieu Roche, Yves Kodratoff, and Michele Sebag. 2005. Preference learning in terminology extraction: A ROC-based approach. *arXiv preprint cs/0512050* (2005).

[7] Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. 2012. Concept annotation in the CRAFT corpus. *BMC bioinformatics* 13, 1 (2012), 1–20.

[8] Shubhanker Banerjee, Bharathi Raja Chakravarthi, and John Philip McCrae. 2022. A Dataset for Term Extraction in Hindi. In *Proceedings of the Workshop on Terminology in the 21st century: many faces, many places*. 19–25.

[9] Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web.. In *LREC*. Citeseer, 1313–1316.

[10] Matthias Bay, Daniel Bruneß, Miriam Herold, Christian Schulze, Michael Guckert, and Mirjam Minor. 2021. Term extraction from medical documents using word embeddings. In *2020 6th IEEE Congress on Information Science and Technology (CiSt)*. IEEE, 328–333.

[11] Elena Bolshakova, Natalia Loukachevitch, and Michael Nokel. 2013. Topic models can improve domain term extraction. In *European Conference on Information Retrieval*. Springer, 684–687.

[12] M Teresa Cabré Castellví, Rosa Estopa Bagot, and Jordi Vivaldi Palatresi. 2001. Automatic term detection: A review of current systems. *Recent advances in computational terminology* 2 (2001), 53–88.

[13] JS Coleti, DF Mattos, and GMB Almeida. 2009. Primeiro dicionário de nanociência e nanotecnologia em língua portuguesa. *II Encontro Acadêmico de Letras (EALE)* (2009), 1–10.

[14] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[15] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*.

[16] Merley Conrado, Thiago Pardo, and Solange Rezende. 2013. A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*. Association for Computational Linguistics, Atlanta, Georgia, 16–23. https://aclanthology.org/N13-2003

[17] Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* 32, suppl_1 (2004), D258–D261.

[18] Robert James Coulthard et al. 2005. The application of corpus methodology to translation: the JPED parallel corpus and the Pediatrics comparable corpus. (2005).

[19] Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-Based Named Entity Recognition Using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 1835–1845.

[20] Merley da Silva Conrado, Ariani Di Felippo, Thiago Alexandre Salgueiro Pardo, and Solange Oliveira Rezende. 2014. A survey of automatic term extraction for Brazilian Portuguese. *Journal of the Brazilian Computer Society* 20, 1 (2014), 1–28.

[21] Béatrice Daille. 1994. Study and implementation of combined techniques for automatic extraction of terminology. In *The balancing act: Combining symbolic and statistical approaches to language*.

[22] Béatrice Daille. 2012. Building bilingual terminologies from comparable corpora: The TTC TermSuite. In *5th Workshop on Building and Using Comparable Corpora with special topic" Language Resources for Machine Translation in Less-Resourced Languages and Domains", co-located with LREC 2012*.

[23] Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

[24] Nikolai Daraselia, Anton Yuryev, Sergei Egorov, Ilya Mazo, and Iaroslav Ispolatov. 2007. Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks. *BMC bioinformatics* 8, 1 (2007), 1–17.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[26] Ariani Di. 2008. Compilação de corpus em língua portuguesa na área de nanociência/nanotecnologia: problemas e soluções. *Avanços da linguística de Corpus no Brasil* (2008), 167.

[27] Boris V Dobrov and Natalia Loukachevitch. 2011. Multiple evidence for term extraction in broad domains. In *Proceedings of the international conference recent advances in natural language processing 2011*. 710–715.

[28] Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231* (2021).

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

[30] Widad Mustafa El Hadi, Ismail Timimi, Marianne Dabbadie, Khalid Choukri, Olivier Hamon, and Yun-Chuang Chiao. 2006. Terminological resources acquisition tools: Toward a user-oriented evaluation model. *System* 1, C2 (2006), C3.

[31] Chantal Enguehard. 2003. Correct: Démarche coopérative pour l'évaluation de systemes de reconnaissance de termes. In *Actes de la 10ème conférence sur le Traitement Automatique des Langues Naturelles. Posters*. 339–346.

[32] Denis Fedorenko, N Astrakhantsev, and D Turdakov. 2014. Automatic recognition of domain-specific terms: an experimental evaluation. *Proceedings of the Institute for System Programming* 26, 4 (2014), 55–72.

[33] Evelina Fedorenko, John Duncan, and Nancy Kanwisher. 2013. Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences* 110, 41 (2013), 16616–16621.

[34] Jody Foo. 2012. *Computational terminology: Exploring bilingual and monolingual term extraction*. Ph. D. Dissertation. Linköping University Electronic Press.

[35] Jody Foo and Magnus Merkel. 2010. Using machine learning to perform automatic term recognition. In *LREC 2010 Workshop on Methods for automatic acquisition of Language Resources and their evaluation methods, 23 May 2010, Valletta, Malta*. European Language Resources Association, 49–54.

[36] Katerina T Frantzi, Sophia Ananiadou, and Junichi Tsujii. 1998. The c-value/nc-value method of automatic recognition for multi-word terms. In *International conference on theory and practice of digital libraries*. Springer, 585–604.

[37] Yuze Gao and Yu Yuan. 2019. Feature-less end-to-end nested term extraction. In *CCF international conference on natural language processing and Chinese computing*. Springer, 607–616.

[38] Tatiana Gornostay, Anita Gojun, Marion Weller, Ulrich Heid, Emmanuel Morin, Beatrice Daille, Helena Blancafort, Serge Sharoff, and Claude Méchoulam. 2012. Terminology Extraction, Translation Tools and Comparable Corpora: TTC concept, midterm progress and achieved results. In *LREC 2012 workshop on creating cross-language resources for disconnected languages and styles (CREDISLAS)*. 4–pages.

[39] Dagmar Gromann, Lennart Wachowiak, Christian Lang, and Barbara Heinisch. 2021. Multilingual Extraction of Terminological Concept Systems. *Deep Learning and Neural Approaches for Linguistic Data* (2021), 5.

[40] Xiaowei Han, Lizhen Xu, and Feng Qiao. 2018. CNN-BiLSTM-CRF model for term extraction in Chinese corpus. In *International Conference on Web Information Systems and Applications*. Springer, 267–274.

[41] Rejwanul Haque, Sergio Penkale, and Andy Way. 2018. TermFinder: log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction. *Language Resources and Evaluation* 52, 2 (2018), 365–400.

[42] Anna Hätty, Michael Dorna, and Sabine Schulte im Walde. 2017. Evaluating the reliability and interaction of recursively used feature classes for terminology extraction. In *Proceedings of the student research workshop at the 15th conference of the European chapter of the association for computational linguistics*. 113–121.

[43] Anna Hätty, Dominik Schlechtweg, Michael Dorna, and Sabine Schulte im Walde. 2020. Predicting degrees of technicality in automatic terminology extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2883–2889.

[44] Amir Hazem, Mérieme Bouhandi, Florian Boudin, and Béatrice Daille. 2020. Termeval 2020: Taln-ls2n system for automatic term extraction. In *6th International Workshop on Computational Terminology (COMPUTERM 2020)*.

[45] Amir Hazem, Mérieme Bouhandi, Florian Boudin, and Béatrice Daille. 2022. Cross-lingual and Cross-domain Transfer Learning for Automatic Term Extraction from Low Resource Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 648–662.

[46] Masamichi Ideue, Kazuhide Yamamoto, Masao Utiyama, and Eiichiro Sumita. 2011. A comparison of unsupervised bilingual term extraction methods using phrase-tables. In *Proceedings of Machine Translation Summit XIII: Papers*.

[47] Tanja Ivanović, Ranka Stanković, Branislava Šandrih Todorović, and Cvetana Krstev. 2022. Corpus-based bilingual terminology extraction in the power engineering domain. *Terminology* (2022).

[48] Alex Judea, Hinrich Schütze, and Sören Brügmann. 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*. 290–300.

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

[50] Kyo Kageura, Takahiro Fukushima, Noriko Kando, Manabu Okumura, Satoshi Sekine, Kazuko Kuriyama, Koichi Takeuchi, Masaharu Yoshioka, Teruo Koyama, and Hitoshi Isahara. 2000. Ir/ie/summarisation evaluation projects in japan. In *LREC2000 Workshop on Using Evaluation within HLT Programs*. 19–22.

[51] Kyo Kageura and Bin Umino. 1996. Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 3, 2 (1996), 259–289.

[52] Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Evaluation of classification algorithms and features for collocation extraction in croatian. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. 657–662.

[53] J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics* 19, suppl_1 (2003), i180–i182.

[54] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*. 177–180.

[55] Michael Krauthammer and Goran Nenadic. 2004. Term identification in the biomedical literature. *Journal of biomedical informatics* 37, 6 (2004), 512–526.

[56] Maren Kucza, 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.. In *Interspeech*. 2072–2076.

[57] Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *31st Annual Meeting of the Association for Computational Linguistics*. 17–22.

[58] Christian Lang, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann. 2021. Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 3607–3620.

[59] Ngoc Tan Le and Fatiha Sadat. 2021. Multilingual Automatic Term Extraction in Low-Resource Domains. In *The International FLAIRS Conference Proceedings*, Vol. 34.

[60] Annaïch Le Serrec, Marie-Claude L'Homme, Patrick Drouin, and Olivier Kraif. 2010. Automating the compilation of specialized dictionaries: Use and analysis of term extraction and lexical alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 16, 1 (2010), 77–106.

[61] Marie-Claude L'Homme, Loubna Benali, Claudine Bertrand, and Patricia Lauduique. 1996. Definition of an evaluation grid for term-extraction software. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 3, 2 (1996), 291–312.

[62] Yang Lingpeng, Ji Donghong, Zhou Guodong, and Nie Yu. 2005. Improving retrieval effectiveness by using key terms in top retrieved documents. In *European Conference on Information Retrieval*. Springer, 169–184.

[63] Jingshu Liu, Emmanuel Morin, and Sebastián Peña Saldarriaga. 2018. Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms. In *Proceedings of the 27th International Conference on Computational Linguistics*. 2855–2866.

[64] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics* 8 (2020), 726–742.

[65] Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. 2018. KAS-term and KAS-biterm: Datasets and baselines for monolingual and bilingual terminology extraction from academic writing. *Digital Humanities* 7 (2018).

[66] Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. Kas-term: Extracting slovene terms from doctoral theses via supervised machine learning. In *International Conference on Text, Speech, and Dialogue*. Springer, 115–126.

[67] Elizaveta Loginova, Anita Gojun, Helena Blancafort, Marie Guégan, Tatiana Gornostay, and Ulrich Heid. 2012. Reference lists for the evaluation of term extraction tools. In *Proceedings of the Terminology and Knowledge Engineering Conference (TKE'2012)*. Citeseer.

[68] Natalia Loukachevitch. 2012. Automatic term recognition needs multiple evidence. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. 2401–2407.

[69] Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Qi Zhang, and Xuanjing Huang. 2021. Template-free prompt tuning for few-shot NER. *arXiv preprint arXiv:2109.13532* (2021).

[70] Lieve Macken and Walter Daelemans. 2010. A Chunk-Driven Bootstrapping Approach to Extracting Translation Patterns. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 394–405.

[71] Lieve Macken, Els Lefever, and Veronique Hoste. 2013. TExSIS: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 19, 1 (2013), 1–30.

[72] Alfredo Maldonado and David Lewis. 2016. Self-tuning ongoing terminology extraction retrained on terminology validation decisions. In *Proceedings of The 12th International Conference on Terminology and Knowledge Engineering*. 91–100.

[73] John Philip McCrae and Adrian Doyle. 2019. Adapting Term Recognition to an Under-Resourced Language: The Case of Irish. In *Proceedings of the Celtic Language Technology Workshop*. 48–57.

[74] Tony McEnery, Richard Xiao, and Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

[75] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. 746–751.

[76] Goran Nenadić and Sophia Ananiadou. 2006. Mining semantically related terms from biomedical literature. *ACM Transactions on Asian Language Information Processing (TALIP)* 5, 1 (2006), 22–43.

[77] Michael Nokel, EI Bolshakova, and Natalia Loukachevitch. 2012. Combining multiple features for single-word term extraction. *Proceedings of Dialog* (2012), 490–501.

[78] Aliya Nugumanova, Darkhan Akhmed-Zaki, Madina Mansurova, Yerzhan Baiburin, and Almasbek Maulit. 2022. NMF-based approach to automatic term extraction. *Expert Systems with Applications* 199 (2022), 117179.

[79] Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics* 29, 1 (2003), 19–51.

[80] Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2005. Terminology extraction: an analysis of linguistic and statistical approaches. In *Knowledge mining*. Springer, 255–279.

[81] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[82] Giulio Petrucci, Marco Rospocher, and Chiara Ghidini. 2018. Expressive ontology learning as neural machine translation. *Journal of Web Semantics* 52 (2018), 66–82.

[83] Senja Pollak, Andraž Repar, Matej Martinc, and Vid Podpečan. 2019. Karst exploration: extracting terms and definitions from karst domain corpus. *Proceedings of eLex* 2019 (2019), 934–956.

[84] Behrang Qasemi Zadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 1862–1868.

[85] Behrang QasemiZadeh and Siegfried Handschuh. 2014. The ACL RD-TEC: a dataset for benchmarking terminology extraction and classification in computational linguistics. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*. 52–63.

[86] Behrang Qasemizadeh and Siegfried Handschuh. 2014. Evaluation of technology term recognition with random indexing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*.

[87] Ayla Rigouts Terryn, Véronique 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 *6th International Workshop on Computational Terminology (COMPUTERM 2020)*. European Language Resources Association (ELRA), 85–94.

[88] Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2021. HAMLET: Hybrid Adaptable Machine Learning approach to Extract Terminology. *Terminology* 27, 2 (2021), 254–293.

[89] Ayla Rigouts Terryn, Veronique Hoste, and Els Lefever. 2022. D-Terminer: Online Demo for Monolingual and Bilingual Automatic Term Extraction. In *Proceedings of the TERM21 Workshop*. Language Resources and Evaluation Conference (LREC 2022), 33–40.

[90] Véronique Anne Sauron. 2002. Tearing out the terms: evaluating terms extractors. In *Proceedings of Translating and the Computer 24*.

[91] Francesco Sclano and Paola Velardi. 2007. Termextractor: a web application to learn the shared terminology of emergent web communities. In *Enterprise Interoperability II*. Springer, 287–290.

[92] JWC Souza and AD Felippo. 2010. *Um exercício em lingüística de Corpus no âmbito do Projeto TermiNet*. Technical Report. Tech. Rep. NILC-TR-10-08, Instituto de Ciências Matemáticas e de Computação . . . .

[93] Maria Fernanda Teline et al. 2004. Avaliação de métodos para a extração Automática de terminologia de textos em português. *ICMC-USP (Dissertação de Mestrado), São Carlos, SP* (2004).

[94] Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2022. Tagging terms in text: A supervised sequential labelling approach to automatic term extraction. *Terminology: international journal of theoretical and applied issues in specialized communication* 28, 1 (2022), 157–189.

[95] Hanh Thi Hong Tran, Matej Martinc, Antoine Doucet, and Senja Pollak. 2022. Can Cross-Domain Term Extraction Benefit from Cross-lingual Transfer?. In *International Conference on Discovery Science*. Springer, 363–378.

[96] Hanh Thi Hong Tran, Matej Martinc, Andraz Pelicon, Antoine Doucet, and Senja Pollak. 2022. Ensembling Transformers for Cross-domain Automatic Term Extraction. In *International Conference on Asian Digital Libraries*. Springer, 90–100.

[97] Thi Hong Hanh Tran, Matej Martinc, Antoine Doucet, and Senja Pollak. 2022. Ensembling Transformers for Cross-domain Automatic Term Extraction. (2022).

[98] Thi Hong Hanh Tran, Matej Martinc, Andraz Repar, Antoine Doucet, and Senja Pollak. 2022. A Transformer-based Sequence-labeling Approach to the Slovenian Cross-domain Automatic Term Extraction. (2022).

[99] Bram Vanroy, Orphée De Clercq, Arda Tezcan, Joke Daems, and Lieve Macken. 2021. Metrics of syntactic equivalence to assess translation difficulty. In *Explorations in empirical translation process research*. Springer, 259–294.

[100] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[101] Spela Vintar. 2010. Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 16, 2 (2010), 141–158.

[102] Jorge Vivaldi and Horacio Rodríguez. 2007. Evaluation of terms and term extraction systems: A practical approach. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 13, 2 (2007), 225–248.

[103] Katarina Vrtovec, Špela Vintar, Amanda Saksida, and Uroš Stepišnik. 2019. TermFrame: Knowledge Frames in Karstology. *of TOTh2019* (2019), 109–126.

[104] Rui Wang, Wei Liu, and Chris McDonald. 2016. Featureless domain-specific term extraction with minimal labelled data. In *Proceedings of the Australasian Language Technology Association Workshop 2016*. 103–112.

[105] Joachim Wermter and Udo Hahn. 2005. Massive biomedical term discovery. In *International Conference on Discovery Science*. Springer, 281–293.

[106] Petra Wolf, Ulrike Bernardi, Christian Federmann, and Sabine Hunsicker. 2011. From statistical term extraction to hybrid machine translation. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*.

[107] Yu Yuan, Jie Gao, and Yue Zhang. 2017. Supervised learning for robust term extraction. In *2017 International Conference on Asian Language Processing (IALP)*. IEEE, 302–305.

[108] Claudia Zavaglia, Leandro Henrique Mendonça de Oliveira, Maria das Graças Volpe Nunes, Maria Fernanda Teline, Sandra Maria Aluisio, et al. 2005. Avaliação de métodos de extração automática de termos para a construção de ontologias. (2005).

[109] Ziqi Zhang, Jie Gao, and Fabio Ciravegna. 2018. Semre-rank: Improving automatic term extraction by incorporating semantic relatedness with personalised pagerank. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12, 5 (2018), 1–41.

[110] Ziqi Zhang, José Iria, Christopher Brewster, and Fabio Ciravegna. 2008. A comparative evaluation of term recognition algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

[111] Ziqi Zhang, Johann Petrak, and Diana Maynard. 2018. Adapted textrank for term extraction: A generic method of improving automatic term extraction algorithms. *Procedia Computer Science* 137 (2018), 102–108.

[112] Bin Zheng, David C McLean, and Xinghua Lu. 2006. Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC bioinformatics* 7, 1 (2006), 1–10.