Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020), pages 85–94
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset

Ayla Rigouts Terryn*, Veronique Hoste*, Patrick Drouin, Els Lefever***
(*) LT^3 Language and Translation Technology Team, Ghent University;
(**) Observatoire de Linguistique Sens-Texte, Université de Montréal;
(*) firstname.lastname@ugent.be, (**) patrick.drouin@umontreal.ca

Abstract

The TermEval 2020 shared task provided a platform for researchers to work on automatic term extraction (ATE) with the same dataset: the Annotated Corpora for Term Extraction Research (ACTER). The dataset covers three languages (English, French, and Dutch) and four domains, of which the domain of *heart failure* was kept as a held-out test set on which final f1-scores were calculated. The aim was to provide a large, transparent, qualitatively annotated, and diverse dataset to the ATE research community, with the goal of promoting comparative research and thus identifying strengths and weaknesses of various state-of-the-art methodologies. The results show a lot of variation between different systems and illustrate how some methodologies reach higher precision or recall, how different systems extract different types of terms, how some are exceptionally good at finding rare terms, or are less impacted by term length. The current contribution offers an overview of the shared task with a comparative evaluation, which complements the individual papers by all participants.

Keywords: ATE, automatic term extraction,terminology

1. Introduction

Automatic Term Extraction (ATE) can be defined as the automated process of identifying terminology from a corpus of specialised texts. Despite receiving plenty of research attention, it remains a challenging task, not in the least because terms are so difficult to define. Terms are typically described as ”lexical items that represent concepts of a domain” (Kageura and Marshman, 2019), but such definitions leave room for many questions about the fundamental nature of terms. Since ATE is supposed to automatically identify terms from specialised text, the absence of a consensus about the basic characteristics of terms is problematic. The disagreement covers both practical aspects, such as term length and part-of-speech (POS) pattern, and theoretical considerations about the difference between words (or collocations/phrases) and terms. This poses great difficulties for many aspects of ATE, from data collection, to extraction methodology, to evaluation.

Data collection, i.e. creating domain-specific corpora in which terms have been annotated, is time- and effort-consuming. When manual term annotation is involved, inter-annotator agreement is notoriously low and there is no consensus about an annotation protocol (Estopà, 2001). This leads to a scarcity in available resources. Moreover, it means that the few available datasets are difficult to combine and compare, and often cover only a single language and domain. While the manual annotation bottleneck has often been circumvented by starting from existing resources, such as ontologies or terminological databases, specialised dictionaries, or book indexes, such strategies do not have the same advantages as manual annotation and will rarely cover all terms in an entire corpus.

This is linked to the evaluation of ATE, for which the accepted metrics are precision (how many of the extracted terms are correct), recall (how many of the terms in the text have correctly been extracted), and f1-score (harmonic mean of the two). To calculate recall (and, therefore, also f1-score), it is necessary to know all true terms in a text. Since manual annotation is such an expensive operation, and relatively few resources are currently available, evaluation is often limited to either a single resource, or the calculation of precision.

The ATE methodology itself, most notably the types of terms a system is designed to find, is impacted as well. Some of the most fundamental differences are term length (in number of tokens), term POS-pattern (sometimes only nouns and noun phrases, sometimes adjectives, adverbs, and verbs are included), and minimum term frequency. Differences which are more difficult to quantify are, for instance, how specialised or domain-specific a lexical unit needs to be before it is considered a term. These three aspects are closely related, since different systems and evaluation methods will be suited for different datasets. This combination of difficulties creates a hurdle for clear, comparative research.

All of this can slow down the advance of ATE, especially now that (supervised) machine learning techniques are becoming more popular for the task. The TermEval shared task on ATE, using the ACTER Annotated Corpora for Term Extraction Research, was designed to lower these hurdles. The ACTER dataset contains specialised corpora in three languages (English, French, and Dutch), and four domains (corruption, dressage (equitation), heart failure, and wind energy), which have been meticulously, manually annotated according to transparent guidelines. Both the texts and the annotations have been made freely available. The current version of the dataset presents the annotations as unstructured lists of all unique annotated terms (one term and its label per line), rather than providing the span of each occurrence of annotated terms in their context (which may be provided in future releases). The shared task brought together researchers to work on ATE with the same data and evaluation setup. It allowed a detailed comparison of different methodologies. Standard evaluation methods (precision, recall, f1-score) were used for the basic evaluation and ranking; these are elaborated with more detailed evaluations as presented both in the current overview paper and in participants' contributions.

The following sections start with a brief overview of current datasets and methodologies for ATE. In section 3, the ACTER dataset is described in some detail. The fourth section contains an overview of the shared task itself and the results. The final section is dedicated to a discussion and the conclusions.

2. Related Research

2.1. Manually Annotated Gold Standards for ATE

Two of the most commonly used annotated datasets are GENIA (Kim et al., 2003), and the ACL RD-TEC 2.0 (Qasemizadeh and Schumann, 2016), both of which are in English. GENIA is a collection of 2000 abstracts from the MEDLINE database in the domain of bio-medicine, specifically "transcription factors in human blood cells". Over 400k tokens were annotated by two domain experts to obtain 93,293 term annotations. The ACL-RD-TEC 2.0 contains 300 annotated abstracts from the ACL Anthology Reference Corpus. Again, two experts performed the annotation of 33k tokens, which resulted in 6818 term annotations. They claim three main advantages over GENIA: first, the domain (computational linguistics) means that ATE researchers will have a better understanding of the material. Second, the ACL RD-TEC corpus covers three decades, which allows some research of the evolution of terms. Third and finally, the annotation is more transparent, with freely available annotation guidelines and the possibility to download the annotations of both experts separately.

There are other examples as well, such as the CRAFT corpus, another English corpus in the biomedical domain (99,907 annotations over 560k tokens) (Bada et al., 2012), an English automotive corpus (28,656 annotations over 224,159 tokens) (Bernier-Colborne, 2012; Bernier-Colborne and Drouin, 2014), a diachronical English corpus on mechanical engineering (+10k annotations over 140k words) (Schumann and Fischer, 2016), the TermITH French corpus on language sciences (14,544 unique validated terms found over 397,695 words) (TermITH, 2014; Billami et al., 2014), a small German corpus on DIY, cooking, hunting and chess which focused on inter-annotator agreement between laypeople (912 annotations on which at least 5 out of 7 annotators agreed, over 3075 words) (Hätty and Schulte im Walde, 2018b) and, within the framework of the TTC project (Loginova et al., 2012), lists of 107-159 annotated terms in corpora in seven languages and two domains (wind energy and mobile technology). While this is a non-exhaustive list, it illustrates an important and logical trend: either the created gold standard is quite large, with over 10k annotations, or it covers multiple languages and/or domains.

While this is not necessarily problematic, the annotation guidelines for all of these corpora differ, and, therefore, the annotations themselves as well. That does create difficulties, since comparing ATE performance on multiple corpora will not necessarily reflect differences in performance between domains or languages, but may also show the contrast between the different annotation styles. The differences can be quite substantial, e.g. in GENIA and ACL RD-TEC, nested annotations are not allowed, in CRAFT they are only allowed under certain conditions, while in the TermITH project they are allowed in most cases. Moreover, it is important to note that the annotations of both the TermITH project and the TTC project are based on the manual annotation of ATE results, rather than manual annotations in the unprocessed text. A final remark is that some corpora have been annotated with multiple term labels or have even been annotated according to large taxonomies, while others don't make any distinctions beyond terms. As will be discussed in more detail in section 3, the ACTER dataset has been specifically designed to deal with some of the issues addressed here.

2.2. ATE

Traditionally, three types of ATE methodologies are identified: linguistic (relying on linguistic information, such as POS-patterns and chunking), statistical (using frequencies, often compared to a reference corpus, to calculate termhood and unithood (Kageura and Umino, 1996)), and hybrid methods (which combine the two). It has been established for some time that hybrid methods appear to outperform the other two (Macken et al., 2013). These methods typically select candidate terms based on their POS-pattern and rank these candidate terms using the statistical metrics, thus combining the advantages of both techniques. A particular difficulty is defining the cut-off threshold for the term candidates, which can be defined as the top-n terms, the top-n percentage of terms, or all terms above a certain threshold score. Manually predicting the ideal cut-off point is extremely difficult and can result in a skew towards either precision or recall, which can be detrimental to the final f1-score (Rigouts Terryn et al., 2019a).

While this typology of linguistic, statistical, and hybrid systems is sometimes still used today, in recent years, the advance of machine learning techniques has made such a simple classification of ATE methodologies more complicated (Gao and Yuan, 2019). Methodologies have become so diverse that they are no longer easily captured in such a limited number of clearly delineated categories. For instance, apart from the distinction between statistical and linguistic systems, one could also distinguish between rule-based methods and machine learning methods. However, rather than a simple binary distinction, there is quite a range of options: methods that rely on a single statistical score (Drouin, 2003; Kosa et al., 2020), systems that combine a limited number of features with a voting algorithm (Fedorenko et al., 2013; Vivaldi and Rodríguez, 2001), an evolutionary algorithm that optimises the ROC-curve (Azé et al., 2005), rule-induction (Foo and Merkel, 2010), support-vector models (Ramisch et al., 2010), logistic regression (Bolshakova et al., 2013; Judea et al., 2014), basic neural networks (Hätty and Schulte im Walde, 2018a), recursive neural networks (Kucza et al., 2018), siamese neural networks (Shah et al., 2019), and convolutional neural networks (Wang et al., 2016). Within the machine learning systems, there are vast differences between supervised, semi-supervised, and unsupervised systems, as well as the distinction between sequence labelling approaches and systems that start from a limited list of unique term candidates. Splitting systems by their features is perhaps even more difficult, since research has moved far beyond using simple linguistic and statistical features. Some examples include the use of topic modelling (Šajatović et al., 2019; Bolshakova et al., 2013), queries on search engines, Wikipedia, or other external resources (Kessler et al., 2019; Vivaldi and Rodríguez, 2001), and word embeddings (Amjadian et al., 2016; Kucza et al., 2018; Qasemizadeh and Handschuh, 2014; Pollak et al., 2019). Some methods are even called "featureless" (Gao and Yuan, 2019; Wang et al., 2016).

There are many more ways in which ATE systems can vary. Some can already be deduced from the ways in which the datasets are annotated, such as support for nested terms. Another very fundamental difference is the frequency cut-off: many ATE systems only extract terms which appear above a certain frequency threshold in the corpora. This threshold is extremely variable, with some systems that do not have any threshold, others that only extract candidate terms which appear 15 times or more (Pollak et al., 2019), and still others where only the top-n most frequent terms are extracted (Loukachevitch, 2012). Term length is similarly variable, with systems that don't place any restrictions, others that extract only single-word terms, only multi-word terms, or those that extract all terms between 1 and n tokens (with n ranging from 2 to 15), where n is sometimes determined by the restrictions of a system, sometimes experimentally set to an optimal value, and at other times directly determined by the maximum term length in a gold standard. There are many other possible differences, such as POS patterns, which will not be discussed in any detail here. More information regarding both datasets for ATE and different ATE methodologies can be found in Rigouts Terryn et al. (2019b).

With such a great variety of methodologies, comparative research is essential to identify the strengths and weaknesses of the respective strategies. However, as discussed, appropriate datasets are scarce and often limited. This means that ATE systems are regularly scored solely on precision (or some variation thereof), since recall and f1-score cannot be calculated without knowing all true terms in a corpus. Considering the expense of data annotation, the extra effort required is rarely feasible. The strictness of the evaluation varies as well, such as determining how specialised a term candidate needs to be for it to be considered a true term, and validating only full matches or also partial ones. Moreover, scores for sequence labelling approaches are difficult to compare to scores for approaches that provide ranked lists of unique terms. There is even disagreement on the required expertise for annotators: do they need to be domain experts or terminologists? This disparity does not only make comparisons between systems highly problematic, it also means that many systems are evaluated on only a single domain (and language).

3. ACTER Annotated Corpora for Term Extraction Research

ACTER is a collection of domain-specific corpora in which terms have been manually annotated. It covers three languages (English, French, and Dutch) and four domains (corruption, dressage (equitation), heart failure, and wind energy). It has been created in light of some of the perceived difficulties that have been mentioned. A previous version (which did not yet bear the ACTER acronym) has already been elaborately described (Rigouts Terryn et al., 2019b), so we refer the interested reader to this work for more detailed information. However, the current version of the dataset has been substantially updated since then, to be even more consistent. All previous annotations have been double-checked, inconsistent annotations were automatically found and manually edited when necessary, and, with this shared task, a first version has been made publicly available. Therefore, the remainder of this section will focus on the up-to-date statistics of version 1.2 of the ACTER dataset (version 1.0 was the first to appear online for the shared task). The annotation guidelines have been updated as well and are freely available[1]. Discontinuous terms (e.g. in ellipses) have been annotated, but are not yet included in ACTER 1.2, and neither are the cross-lingual annotations in the domain of heart failure. The changes made between ACTER versions are indicated in detail in the included README.md file and the biggest difference between version 1.0 and 1.2 (besides some 120 removed or added annotations) is the inclusion of the label of each annotation.

The dataset contains trilingual comparable corpora in all domains: the corpora in the same domain are similar in terms of subject, style, and length for each language, but they are not translations (and, therefore, cannot be aligned). Additionally, for the domain of corruption, there is a trilingual parallel corpus of aligned translations. For each language and domain, around 50k tokens have been manually annotated (in the case of corruption, the annotations have only been made in the parallel corpus, so the comparable corpus on corruption is completely unannotated). In all domains except heart failure, the complete corpora are larger than only the annotated parts, and unannotated texts are included (separately) as well. The texts are all plain text files and the sources have been included in the downloadable version. The annotations have been performed in the BRAT annotation tool (Stenetorp et al., 2011), but they are currently provided as flat lists with one term per line. The annotations have all been performed by a single annotator with experience in the field of terminology and ATE and fluent in all three languages. However, she is not a domain-expert, except in the domain of dressage. Multiple semi-automatic checks have been performed to ensure the best possible annotation quality and inter-annotator agreement studies were performed and published (Rigouts Terryn et al., 2019b) to further validate the dataset. Furthermore, the elaborate guidelines helped the annotator to make consistent decisions and make the entire process more transparent. Nevertheless, term annotation remains an ambiguous

[1] http://hdl.handle.net/1854/LU-8503113

bioprosthetic valve replacement	Specific_Term
biopsies	Common_Term
biopsy	Common_Term
biosynthetic enzymes	Specific_Term
bisoprolol	Specific_Term
bisphosphonates	Specific_Term

Table 1: Sample of one of the gold standard term lists in the ACTER 1.2 dataset to illustrate the format

and subjective task. We do not claim that ours is the only possible interpretation and, therefore, when using ACTER for ATE evaluation purposes, always recommend checking the output for a more nuanced evaluation (e.g. Rigouts Terryn et al. (2019a)).

While ATE for TermEval has been perceived as a binary task (term or not), the original annotations included four different labels. There are three term labels, for which terms are defined by their degree of domain-specificity (are they relevant to the domain) and lexicon-specificity (are they known only by experts, or by laypersons as well). The three term labels defined this way are: Specific Terms (which are both domain- and lexicon-specific), Common Terms (domain-specific, not lexicon-specific), and Out-of-Domain (OOD) Terms (not domain-specific, lexicon-specific). In the domain of heart failure, for instance, *ejection fraction* might be a Specific Term: laypersons generally do not know what it means, and it is strongly related to the domain of heart failure, since it is an indication of the volume of blood the heart pumps on each contraction. *Heart* is an example of a Common Term: it is clearly domain-specific to heart failure and you do not need to be an expert to have a basic idea of what a heart is. An example of an OOD term might be *p-value*, which is lexicon-specific since you need some knowledge of statistics to know the term, but it is not domain-specific to heart failure. In addition to these three term labels, Named Entities (proper names of persons, organisations, etc.) were annotated as well, as they share a few characteristics with terms: they will appear more often in texts with a relevant subject (e.g. brand names of medicine in the field of heart failure) and, like multi-word terms, have a high degree of unithood (internal cohesion). Labelling these does not mean we consider them to be terms, but it offers more options for the evaluation and training based on the dataset.

Since TermEval was set up as a binary task, all three term labels were combined and considered as true terms. There were two separate datasets regarding the Named Entities: one including both terms and Named Entities, one with only terms. All participating systems were evaluated on both datasets. Moreover, while the evaluation for the ranking of the participating systems was based only on these two binary interpretations, the four labels were made available afterwards for a more detailed evaluation of the results. The gold standard lists of terms were ordered alphabetically, so with no relation to their labels or degree of termhood. Table 1 shows a sample of such a gold standard list, with one unique term per line followed by its label.

Tables 2 and 3 provide more details on ACTER 1.2. Table 2 shows the number of documents and words per corpus, both in the entire corpus and only the annotated part of the corpus. Table 3 provides details on the number of annotations per corpus, counting either all annotations or all unique annotations. In total, 119,455 term and Named Entity annotations have been made over 596,058 words, resulting in 19,002 unique annotations. As can be seen, the number of annotations within a domain is usually somewhat similar for all languages (since the corpora are comparable), with larger differences between the domains. Version 1.2 of ACTER only provides a list of all unique lowercased terms (and Named Entities) per corpus. The aim is to release future versions with all in-text annotation spans, where every occurrence of each term is annotated, so that it can be used for sequence-labelling approaches as well. It is important to note that, since the annotation process was completely manual, each occurrence of a term was evaluated separately. When a lexical unit was only considered a term in some contexts, it was only annotated in those specific contexts. For instance, the word *sensitivity* can be used in general language, where it will not be annotated, but also as a synonym of *recall* (true positive rate), in which case it was annotated as a term.

Additional characteristics to bear in mind about these annotations are that nested annotations are allowed (as long as the nested part can be used as a term on its own), and that there were no restrictions on term length, term frequency, or term POS-pattern (apart from the condition that terms had to contain a content word). If a lexical unit was used as a term in the text, it was annotated, even if it was not the best or most frequently used term for a certain concept. The reasoning behind this strategy was that one of the most important applications of ATE is to be able to keep up with fast-evolving terminology in increasingly more specialised domains. If only well-established, frequent terms are annotated, the rare and/or new terms will be ignored, even though these could be particularly interesting for ATE. While these qualities were all chosen to best reflect the desired applications for ATE, they do result in a particularly difficult dataset for ATE, so f1-scores for ATE systems tested on ACTER are expected to be rather modest in comparison to some other datasets.

4. TermEval Shared Task on ATE

4.1. Setup

The aim of the TermEval shared task was to provide a platform for researchers to work on the same task, with the same data, so that different methodologies for ATE can easily be compared and current strengths and weaknesses of ATE can be identified. During the training phase, participants all received the ACTER dataset as described in the previous section, with all domains apart from *heart failure*. The latter is provided during the final phase as the test set on which the scores are calculated. As described in the previous section, ACTER 1.2 consists of flat lists of unique terms per corpus, with one term per line. Since this first version of the shared task aims to focus on ATE in general, rather than term variation, all terms are lowercased, and only identical lowercased terms are merged in a single entry, without lemmatisation. Even when terms acquire

				# Words	
Type	**Domain**	**Language**	**# Texts**	**in entire corpus**	**in annotated part of corpus**
Parallel	Corruption	en	24	176,314	45,234
		fr	24	196,327	50,429
		nl	24	184,541	47,305
Comparable	Corruption	en	44	468,711	-
		fr	31	475,244	-
		nl	49	470,242	-
	Dressage	en	89	102,654	51,470
		fr	125	109,572	53,316
		nl	125	103,851	50,882
	Heart failure	en	190	45,788	45,788
		fr	215	46,751	46,751
		nl	175	47,888	47,888
	Wind Energy	en	38	314,618	51,911
		fr	12	314,681	56,363
		nl	29	308,742	49,582
		TOTAL	**3,365,924**	**1194**	**596,058**

Table 2: Number of documents and words in the entire corpus vs. the annotated part of each corpus in ACTER 1.2

		# Annotations			
Domain	**Language**	**Terms (all)**	**Terms (unique)**	**NEs (all)**	**NEs (unique)**
Corruption	en	6,385	927	2,373	247
	fr	5,930	982	2,186	235
	nl	5,163	1,047	2,334	248
Dressage	en	10,889	1,155	970	420
	fr	9,397	963	467	220
	nl	11,207	1,395	295	151
Heart failure	en	14,011	2,361	526	224
	fr	10,801	2,276	319	147
	nl	10,219	2,077	433	180
Wind Energy	en	9,478	1,091	1,429	443
	fr	8,524	773	439	195
	nl	5,044	940	636	305
	TOTAL	**107,048**	**15,987**	**12,407**	**3,015**

Table 3: Number of annotations (counting all annotations separately or all unique annotations) of terms and Named Entities (NEs), per corpus in ACTER 1.2

a different meaning through different capitalisation options or POS patterns, they only count as a single annotation in this version. For example, the English corpus on dressage contains the term *bent* (verb – past tense of *to bend*), but also *Bent* (proper noun – person name). While both capitalisation and POS differ, and *bent* is not the lemmatised form, there is only one entry: *bent* (lowercased) in the gold standard (other full forms of the verb *to bend* have separate entries, if they are present and annotated in the corpus). We do not discount the importance of ATE systems that handle term variation, but a choice was made to focus on the core task for the first edition of the task.

There are three different tracks (one per language) and participants could enter in one or multiple tracks. When participants submitted their final results on the test data (as a flat list of unique lowercased terms, like the training data), f1-scores were calculated twice: once compared to the gold standard with only terms, once compared to the gold standard with both terms and Named Entities. These double scores did not influence the final ranking based on f1-scores. The dataset has been used for more detailed evaluations as well (see section 4.3) and participants were encouraged to report scores on the training domains in their own papers as well.

4.2. Participants

Five teams participated in the shared task: TALN-LS2N (Hazem et al., 2020), RACAI (Pais and Ion, 2020), e-Terminology (Oliver and Vàzquez, 2020), NLPLab_UQAM (no system description paper), and NYU (no system description paper but based on previous work in Meyers et al. (2018)). NYU and RACAI participated only in the English track, TALN-LS2N participated in both the English and French tracks, and e-Terminology and NLPLab_UQAM participated in all tracks. We refer to their own system description papers for more details, but will provide a short summary of each of their methodologies.

Team **NYU** has applied an updated version of Termolator (Meyers et al., 2018). Candidate terms are selected based on "terminological chunking and abbreviations". The terminological chunking focuses, among others, on nominalisations, out-of-vocabulary words, and technical adjectives (based on suffixes) to find terms. Constructions where full forms are followed by their abbreviations are also taken into account. Next, three distributional metrics (e.g. TFIDF) are combined with equal weights and a "well-formedness score" is calculated, using mainly linguistic and morphological information. Additionally, a relevance score is based on the results of an online search engine. The final selection of candidate terms is made based on the product of these three metrics. Due to the timing of the shared task, Termolator was not specifically tuned to the ACTER dataset.

Team **e-Terminology** uses the TSR (Token Slot Recognition) technique, implemented in TBXTools (Oliver and Vazquez, 2015; Vàzquez and Oliver, 2018). For Dutch, the statistical version of TBXTools is employed, for English and French the linguistic version is used. Stopwords are filtered out and all candidate terms that appear below a frequency threshold of two. As a terminological reference for each language (required for the TSR technique), the IATE database for 12-Law was chosen.

Team **RACAI** uses a combination of statistical approaches, such as an improved TextRank (Zhang et al., 2018), TFIDF, clustering, and termhood features. Algorithms were adapted where possible to make use of pre-trained word embeddings and the result was generated using several voting and combinatorial approaches. Special attention is also paid to the detection of nested terms.

Team **TALN-LS2N** uses BERT as a binary classification model for ATE. The model's input is represented as the concatenation of a sentence and a selected n-gram within the sentence. If the n-gram is a term, the input is labelled as positive training example. If not, a corresponding negative example is generated.

Team **NLPLab_UQAM** applied a bidirectional LSTM. Pre-trained GloVe word embedding were used to train a neural network-based model on the training corpora.

4.3. Results

Precision, recall, and f1-scores were calculated both including and excluding Named Entities, for each team in all tracks. The scores and resulting ranking are presented in Table 3. As can be seen, TALN-LS2N's system outperforms all others in the English and French tracks. NLPLab_UQAM's system outperforms e-Terminology for the Dutch track (though their respective rankings for English and Dutch are reversed). Scores with and without Named Entities are usually very similar (average difference of one percentage point), with e-Terminology and NYU scoring slightly better when Named Entities are excluded, and the others scoring better when they are included. On average, precision is higher than recall, especially when Named Entities are included. However, there is much variation. For instance, TALN-LS2N's English system obtains 36-40% more recall than precision (the difference is only 6-9% for their French system). Comparatively, e-Terminology obtains 20% higher precision than recall on average and NLPLab_UQAM obtains more balanced precision and recall scores. The number of extracted term candidates varies greatly as well, from 744 (e-Terminology in Dutch), to 5267 (TALN-LS2N in English). Therefore, even though TALN-LS2N achieves the highest f1-scores thanks to great recall in English, their system also produces most noise, with 3435 false positives (including Named Entities). The average number of extracted candidate terms (2038) is not too different from the average number of terms in the gold standard (2422 incl. Named Entities, 1720 without). Looking at performance of systems in multiple tracks, there does not appear to be one language that is inherently easier or more difficult. TALN-LS2N's best performance is reached for French, e-Terminology's for English, and NLPLab_UQAM's for Dutch.

As with many other task within natural language processing, the methodology based on the BERT transformer model appears to outperform other approaches. However, the large gap between precision and recall for the English model, which is much smaller for the French model, may be an indication of an often-cited downside of deep learning models: their unpredictability. For ATE, predictability is cited as at least as important as f1-scores: "for ATE to be usable, its results should be consistent, predictable and transparent" (Kageura and Marshman, 2019). Additionally, it appears that neural networks and word embeddings do not always work for this task, as demonstrated by the fact that, for English and French, NLPLab_UQAM's bidirectional LSTM approach with GLOVE embeddings is ranked last, below non-neural approaches such as NYU's.

Apart from the ranking based on f1-scores, three different aspects of the results are analysed in more detail: composition of the output, recall of terms with different frequencies, and recall of terms with different lengths. Figure 1 shows the first of these, illustrating the composition of the gold standard regarding the four annotation labels, versus the true positives from each team. The results are averaged over all languages, as the differences between the languages were small. False positives were not included, since these can be deduced from the precision scores. The graphs are relative, so they do not represent the absolute number of annotations per type, only the proportions. The order of the teams is the order of their ranks for the English track. A first observation is that all teams seem to extract at least some Named Entities, except for e-Terminology. This may be partly due to their low recall, but since they did not extract a single Named Entity in any of the languages, it does appear that their system is most focused on terms. While the differences are never extreme, the various systems do show some variation in this respect. For instance, the two lowest ranked systems can be seen to extract relatively more Common Terms. This may be an indication that they are sensitive to frequency, as many of the Specific Terms are rarer (e.g., e-Terminology employs a frequency threshold of two). Conversely, NYU's system appears to excel at extracting these Specific Terms and also extracts relatively few Named Entities. The output of two top-scoring teams has a very similar composition to the gold standard, which

Track	Rank	Team	Scores incl. NEs			Scores excl. NEs		
			precision	recall	f1-score	precision	recall	f1-score
English	1	TALN-LS2N	34.8	70.9	46.7	32.6	72.7	45.0
	2	RACAI	42.4	40.3	41.3	38.6	40.1	39.3
	3	NYU	43.5	23.6	30.6	42.2	25.1	31.5
	4	e-Terminology	34.4	14.2	20.1	34.4	15.5	21.4
	5	NLPLab_UQAM	21.4	15.6	18.1	20.1	16.0	17.8
French	1	TALN-LS2N	45.2	51.5	48.1	41.9	50.9	45.9
	2	e-Terminology	36.3	13.5	19.7	36.3	14.4	20.6
	3	NLPLab_UQAM	16.1	11.2	13.2	15.1	11.2	12.9
Dutch	1	NLPLab_UQAM	18.9	18.6	18.7	18.1	19.3	18.6
	2	e-Terminology	29.0	9.6	14.4	29.0	10.4	15.3

Table 4: Scores (as percentages) and rank for all teams per track

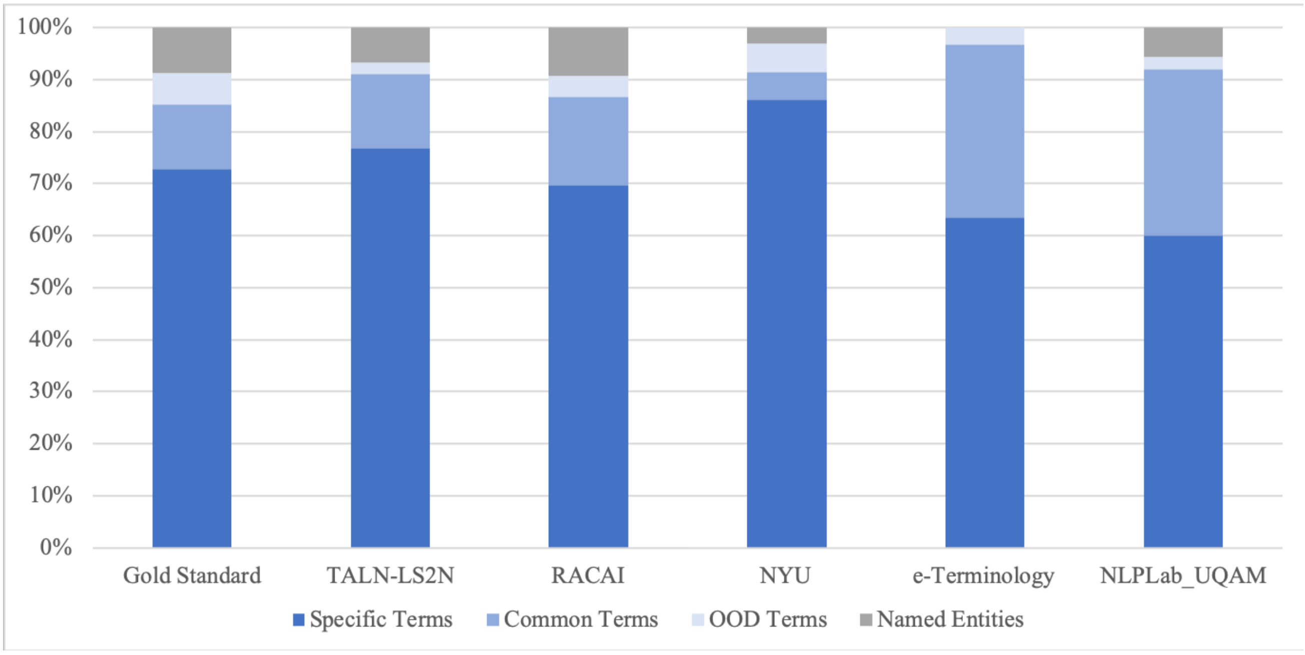

Figure 1: Proportion of Specific, Common, and OOD Terms, and Named Entities in the gold standard versus the true positives extracted by each team (averaged over all languages if teams participated in multiple tracks).

may be part of the explanation for their high scores, and, in the case of TALN-LS2N's system, may be related to their reliance on the training data.

A preference for Common Terms or Specific Terms can already give an indication of the system performance for rare terms, but we can also look directly at the recall of terms for various frequencies, as shown in Figure 2. Here, the recall of all systems for various term frequencies is shown for the English track. Results for the other languages were similar, so will not be discussed separately. The dataset actually contains many hapax terms (which appear only once). In English, when Named Entities are included, there are 1121 (43%) hapax terms, 398 (15%) terms that appear twice, 220 (9%) terms that appear three times, 232 (9%) terms with a frequency between 4 and 5, 259 (10%) terms with a frequency between 5 and 10, 199 (8%) terms with a frequency between 10 and 25, and only 156 (6%) terms that appear more than 25 times. In line with previous findings on the difficulties of ATE, recall is lowest for hapax terms for all systems, and increases as frequency increases. Of course, e-Terminology has 0% recall for hapax terms due to the frequency cut-off, but the other systems also have difficulties. Notably, TALN-LS2N's system obtains a surprisingly stable recall for various frequencies and a very high recall of 64% for hapax terms. This is likely a consequence of the fact that they use none of the traditional statistical (frequency-related) metrics for ATE. Recall is almost always highest for the most frequent terms, though when looking at these frequent terms in more detail, recall appears to drop again for the most extreme cases (terms appearing over 100 times; not represented separately in Figure 2), presumably because these are more difficult to distinguish from common general language words.

The final analysis concerns term length. Similarly to the analysis for frequency, Figure 3 presents recall for different term lengths per team, using the English data, including Named Entities, as a reference. The majority of gold standard terms are single-word terms (swts) (1170, or 45%),

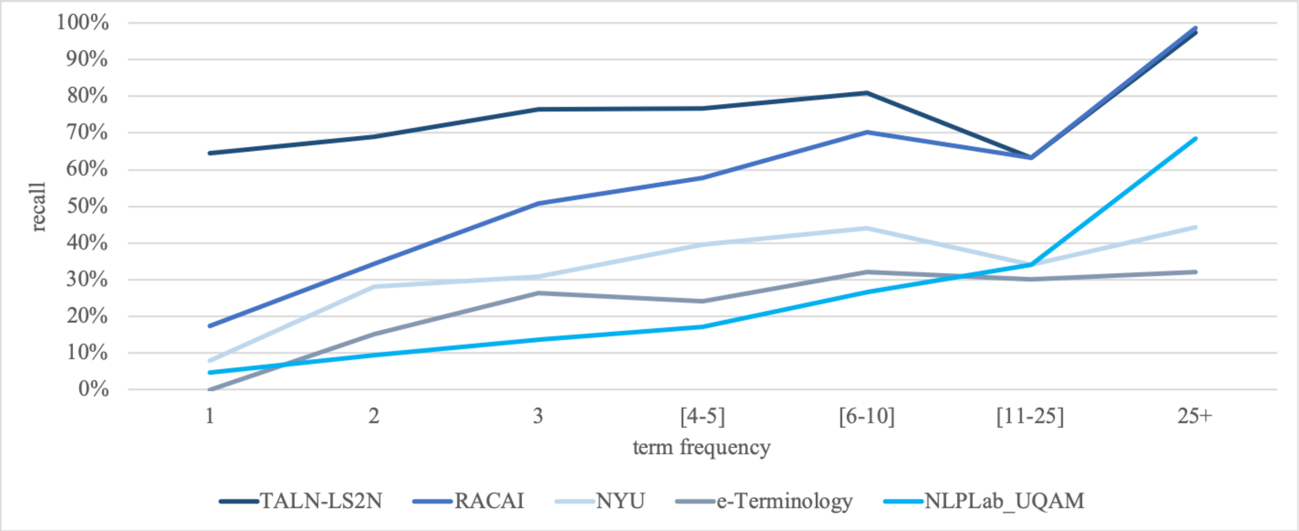

Figure 2: Recall for terms with various frequencies per team in English, including Named Entities

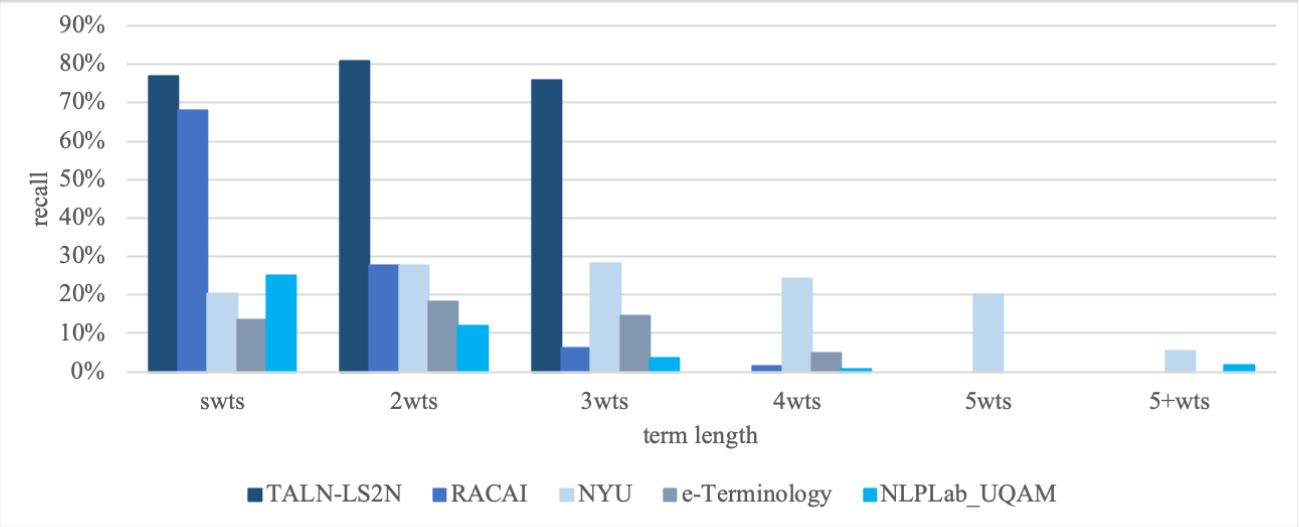

Figure 3: Recall per term length (single-word terms (swts) to terms with over 5 tokens (5+wts) for each team in English, including Named Entities

with frequencies decreasing as term length increases (800 or 31% 2-word terms (2wts), 376 or 15% 3wts, 144 or 6% 4wts, 40 or 2% 5wts, and 55 or 2% terms that are longer than 5 tokens. As can be seen in Figure 3, two out of five teams (RACAI and NLPLab_UQAM) have lower recall for 2wts than for swts, and, overall, recall decreases for terms with more than 3 tokens. TALN-LS2N extracts no terms beyond a length of 3 tokens at all, though this is different for their French system, where recall decreases more gradually with term length. NYU's system has a surprisingly stable performance for different term lengths, especially compared to TALN-LS2N and RACAI.

5. Discussion and Conclusions

Five different teams submitted their results for the TermEval shared task on ATE, based on the ACTER dataset. With the domains of corruption, dressage, and wind energy from the dataset as training data or simply as reference material, the teams either used (and adapted) their existing systems or developed a new methodology for ATE. The domain of heart failure was used as the test set, with three different tracks for English, French and Dutch. The teams were all ranked based on the f1-score they obtained on the test data, with additional evaluations of the types of terms they extracted and recall for different term frequencies and term lengths.

The results show quite a large variation between all methodologies. The highest scores were obtained by a deep learning methodology using BERT as a binary classification model. The second best system does not rely on deep learning and combines pre-trained word embeddings with more classical features for ATE, such as statistical termhood measures. Such results show how there is still a lot of potential for deep learning techniques in the field of ATE, highlighting also the importance of large datasets like ACTER. However, it also illustrates that more traditional methodologies can still lead to state-of-the-art results as well, especially when updated with features like word embeddings.

The more detailed analyses also revealed how the composition of the output of the different systems varies, e.g., including or excluding more Named Entities, and focusing on either the most domain-specific and specialised terms (Specific Terms) or also on more general terms (Common Terms). This is a clear indication of how different applications for ATE may require different methodologies. For instance, translators may be more interested in a system that extracts mostly Specific Terms, since Common Terms may already be part of their general vocabulary.

Checking recall for terms with different frequencies and terms with different lengths confirmed two often-cited weaknesses of ATE: low-frequency terms and long terms are more difficult to extract. However, in each case, there were some systems for which the performance was more stable and less impacted by these factors. The winning deep learning approach achieves a high recall even for hapax terms (64%) and one of the rule-based systems maintains a more or less stable recall for terms up to a length of five tokens.

With these results, we conclude that there remains a lot of room for improvement in the field of ATE, both by trying the latest deep learning methodologies which have been successfully used in other natural language processing tasks, and by updating and combining more traditional methodologies with state-of-the-art features and algorithms. Taking into account the unpredictability of many machine learning approaches and the considerable variety between the potential outputs, as demonstrated in this shared task, it is essential for ATE to be evaluated beyond precision, recall, and f1-scores. To further encourage and facilitate both supervised machine learning approaches and high-quality evaluations on diverse data, the complete ACTER dataset has been made freely available online (Rigouts Terryn, Ayla and Drouin, Patrick and Hoste, Véronique and Lefever, Els, 2020).

6. Bibliographical References

Amjadian, E., Inkpen, D., Paribakht, T., and Faez, F. (2016). Local-Global Vectors to Improve Unigram Terminology Extraction. In *Proceedings of the 5th International Workshop on Computational Terminology*, pages 2–11, Osaka, Japan.

Azé, J., Roche, M., Kodratoff, Y., and Sebag, M. (2005). Preference Learning in Terminology Extraction: A ROC-based approach. In *Proceeedings of Applied Stochastic Models and Data Analysis*, pages 209–2019, Brest, France. arXiv: cs/0512050.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K. B., Verspoor, K., Blake, J. A., and Hunter, L. E. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13:161–180.

Bernier-Colborne, G. and Drouin, P. (2014). Creating a test corpus for term extractors through term annotation. *Terminology*, 20(1):50–73.

Bernier-Colborne, G. (2012). Defining a Gold Standard for the Evaluation of Term Extractors. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey. ELRA.

Billami, M., Camacho-Collados, J., Jacquey, E., and Kister, L. (2014). Annotation sémantique et validation terminologique en texte intégral en SHS. In *Proceedings of TALN 2014*, pages 363–376, Marseille, France.

Bolshakova, E., Loukachevitch, N., and Nokel, M. (2013). Topic Models Can Improve Domain Term Extraction. In David Hutchison, et al., editors, *Advances in Information Retrieval*, volume 7814, pages 684–687. Springer Berlin Heidelberg, Berlin, Heidelberg.

Drouin, P. (2003). Term Extraction Using Non-Technical Corpora as a Point of Leverage. *Terminology*, 9(1):99–115.

Estopà, R. (2001). Les unités de signification spécialisées élargissant l'objet du travail en terminologie. *Terminology*, 7(2):217–237.

Fedorenko, D., Astrakhantsev, N., and Turdakov, D. (2013). Automatic recognition of domain-specific terms: an experimental evaluation. In *Proceedings of the Ninth Spring Researcher's Colloquium on Database and Information Systems*, volume 26, pages 15–23, Kazan, Russia.

Foo, J. and Merkel, M. (2010). Using machine learning to perform automatic term recognition. In *Proceedings of the LREC 2010 Workshop on Methods for automatic acquisition of Language Resources and their evaluation methods*, pages 49–54, Valetta, Malta. ELRA.

Gao, Y. and Yuan, Y. (2019). Feature-Less End-to-End Nested Term Extraction. *arXiv:1908.05426 [cs, stat]*, August. arXiv: 1908.05426.

Hätty, A. and Schulte im Walde, S. (2018a). Fine-Grained Termhood Prediction for German Compound Terms Using Neural Networks. In *Proceedings of the Joint Workshop on,Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 62–73, Sante Fe, New Mexico, USA.

Hätty, A. and Schulte im Walde, S. (2018b). A Laypeople Study on Terminology Identification across Domains and Task Definitions. In *Proceedings of NAACL-HLT 2018*, pages 321–326, New Orleans, USA. ACL.

Hazem, A., Bouhandi, M., Boudin, F., and Daille, B. (2020). Termeval 2020: Taln-ls2n system for automatic term extraction. In *Proceedings of CompuTerm 2020.*

Judea, A., Schütze, H., and Brügmann, S. (2014). Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical Papers*, pages 290–300, Dublin, Ireland.

Kageura, K. and Marshman, E. (2019). Terminology Extraction and Management. In O'Hagan, Minako, editor, *The Routledge Handbook of Translation and Technology.*

Kageura, K. and Umino, B. (1996). Methods of automatic term recognition. *Terminology*, 3(2):259–289.

Kessler, R., Béchet, N., and Berio, G. (2019). Extraction of terminology in the field of construction. In *Proceedings of the First International Conference on Digital Data Processing (DDP)*, pages 22–26, London, UK. IEEE.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GE-

NIA corpus - a semantically annotated corpus for biotextmining. *Bioinformatics*, 19(1):180–182.

Kosa, V., Chaves-Fraga, D., Dobrovolskyi, H., and Ermolayev, V. (2020). Optimized Term Extraction Method Based on Computing Merged Partial C-Values. In *Information and Communication Technologies in Education, Research, and Industrial Applications. ICTERI 2019*, volume 1175 of *Communications in Computer and INformation Science*, pages 24–49. Springer International Publishing, Cham.

Kucza, M., Niehues, J., Zenkel, T., Waibel, A., and Stüker, S. (2018). Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks. In *Interspeech 2018*, pages 2072–2076, Hyderabad, India, September. ISCA.

Loginova, E., Gojun, A., Blancafort, H., Guégan, M., Gornostay, T., and Heid, U. (2012). Reference Lists for the Evaluation of Term Extraction Tools. In *Proceedings of the 10th International Congress on Terminology and Knowledge Engineering*, Madrid, Spain. ACL.

Loukachevitch, N. (2012). Automatic Term Recognition Needs Multiple Evidence. In *Proceedings of LREC 2012*, pages 2401–2407, Istanbul, Turkey. ELRA.

Macken, L., Lefever, E., and Hoste, V. (2013). TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment. *Terminology*, 19(1):1–30.

Meyers, A. L., He, Y., Glass, Z., Ortega, J., Liao, S., Grieve-Smith, A., Grishman, R., and Babko-Malaya, O. (2018). The Termolator: Terminology Recognition Based on Chunking, Statistical and Search-Based Scores. *Frontiers in Research Metrics and Analytics*, 3.

Oliver, A. and Vazquez, M. (2015). TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction. In *Proceedings of Recent Advances in Natural Language Processing*, pages 473–479, Hissar, Bulgaria.

Oliver, A. and Vàzquez, M. (2020). Termeval 2020: Using tsr filtering method to improve automatic term extraction. In *Proceedings of CompuTerm 2020*.

Pais, V. and Ion, R. (2020). Termeval 2020: Racai's automatic term extraction system. In *Proceedings of CompuTerm 2020*.

Pollak, S., Repar, A., Martinc, M., and Podpečan, V. (2019). Karst Exploration: Extracting Terms and Definitions from Karst Domain Corpus. In *Proceedings of eLex 2019*, pages 934–956, Sintra, Portugal.

Qasemizadeh, B. and Handschuh, S. (2014). Investigating Context Parameters in Technology Term Recognition. In *Proceedings of SADAATL 2014*, pages 1–10, Dublin, Ireland.

Qasemizadeh, B. and Schumann, A.-K. (2016). The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods. In *Proceedings of LREC 2016*, pages 1862–1868, Portorož, Slovenia. ELRA.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010). Multiword Expressions in the wild? The mwetoolkit comes in handy. In *Coling 2010: Demonstration Volume*, pages 57–60, Beijing, China.

Rigouts Terryn, A., Drouin, P., Hoste, V., and Lefever, E. (2019a). Analysing the Impact of Supervised Machine Learning on Automatic Term Extraction: HAMLET vs TermoStat. In *Proceedings of RANLP 2019*, Varna, Bulgaria.

Rigouts Terryn, A., Hoste, V., and Lefever, E. (2019b). In No Uncertain Terms: A Dataset for Monolingual and Multilingual Automatic Term Extraction from Comparable Corpora. *Language Resources and Evaluation*, pages 1–34.

Schumann, A.-K. and Fischer, S. (2016). Compasses, Magnets, Water Microscopes. In *Proceedings of LREC 2016*, pages 3578–3584, Portorož, Slovenia. ELRA.

Shah, S., Sarath, S., and Shreedhar, R. (2019). Similarity Driven Unsupervised Learning for Materials Science Terminology Extraction. *Computación y Sistemas*, 23(3):1005–1013.

Stenetorp, P., Topić, G., Pyysalo, S., Ohta, T., Kim, J.-D., and Tsujii, J. (2011). BioNLP Shared Task 2011: Supporting Resources. In *Proceedings of BioNLP Shared Task 2011 Workshop*.

TermITH, P. (2014). Annotation sémantique et terminologique avec la plateforme SMARTIES.

Vivaldi, J. and Rodríguez, H. (2001). Improving term extraction by combining different techniques. *Terminology*, 7(1):31–48, December.

Vàzquez, M. and Oliver, A. (2018). Improving term candidates selection using terminological tokens. *Terminology*, 24(1):122–147, May.

Wang, R., Liu, W., and McDonald, C. (2016). Featureless Domain-Specic Term Extraction with Minimal Labelled Data. In *Proceedings of Australasian Language Technology Association Workshop*, pages 103–112, Melbourne, Australia.

Zhang, Z., Petrak, J., and Maynard, D. (2018). Adapted TextRank for Term Extraction: A Generic Method of Improving Automatic Term Extraction Algorithms. *ACM Transactions on Knowledge Discovery from Data*, 12(5):1–7.

Šajatović, A., Buljan, M., Šnajder, J., and Bašić, B. D. (2019). Evaluating Automatic Term Extraction Methods on Individual Documents. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 149–154, Florence, Italy. ACL.

7. Language Resource References

Rigouts Terryn, Ayla and Drouin, Patrick and Hoste, Véronique and Lefever, Els. (2020). *Annotated Corpora for Term Extraction Research (ACTER)*. Ghent University, 1.2.