HAMLET

Hybrid Adaptable Machine Learning approach to Extract Terminology

Ayla Rigouts Terryn, Véronique Hoste & Els Lefever
LT Language and Translation Technology Team

Automatic term extraction (ATE) is an important task within natural language processing, both separately, and as a preprocessing step for other tasks. In recent years, research has moved far beyond the traditional hybrid approach where candidate terms are extracted based on part-of-speech patterns and filtered and sorted with statistical termhood and unithood measures. While there has been an explosion of different types of features and algorithms, including machine learning methodologies, some of the fundamental problems remain unsolved, such as the ambiguous nature of the concept "term". This has been a hurdle in the creation of data for ATE, meaning that datasets for both training and testing are scarce, and system evaluations are often limited and rarely cover multiple languages and domains. The ACTER Annotated Corpora for Term Extraction Research contain manual term annotations in four domains and three languages and have been used to investigate a supervised machine learning approach for ATE, using a binary random forest classifier with multiple types of features. The resulting system (HAMLET Hybrid Adaptable Machine Learning approach to Extract Terminology) provides detailed insights into its strengths and weaknesses. It highlights a certain unpredictability as an important drawback of machine learning methodologies, but also shows how the system appears to have learnt a robust definition of terms, producing results that are state-of-the-art, and contain few errors that are not (part of) terms in any way. Both the amount and the relevance of the training data have a substantial effect on results, and by varying the training data, it appears to be possible to adapt the system to various desired outputs, e.g., different types of terms. While certain issues remain difficult – such as the extraction of rare terms and multiword terms – this study shows how supervised machine learning is a promising methodology for ATE.

Keywords: terminology, automatic term extraction, comparable corpora, named entities

https://doi.org/10.1075/term.20017.rig | Published online: 20 August 2021
Terminology 27:2 (2021), pp. 290–329. ISSN 0929-9971 | E-ISSN 1569-9994

1. Introduction

Automatic term extraction (ATE) attempts to automatically discover terms in collections of domain-specific texts, where terms can be described as the specialised vocabulary of that domain. Since manual term extraction is a time- and effort-consuming task, ATE has been a popular field of research within natural language processing (NLP). Extracted terms can be used for a multitude of applications, such as ontology construction and enrichment (Durán-Muñoz 2019), machine translation (Wolf et al. 2011), and sentiment analysis (Mayorov et al. 2015). Despite the importance of terms and the amount of research about ATE, identifying terms remains a difficult and largely subjective task. Three often-cited and strongly related hurdles are the lack of a clear distinction between terms and general language, the varying characteristics of terms for different domains, languages, and applications, and the time- and effort-consuming nature of manual term extraction for the creation of gold standard data. Consequently, annotated datasets are scarce, often limited, and rarely cover multiple languages and domains. Combining multiple datasets is problematic since each dataset has a different annotation protocol. However, the rise of machine learning methodologies means that good datasets are becoming increasingly important, both as training, and as evaluation data. The ACTER dataset (Annotated Corpora for Term Extraction Research) (Rigouts Terryn et al., 2019; Rigouts Terryn et al., 2020) was created with these problems in mind and covers multiple languages and domains. Manual annotations are made with four different labels, and both the dataset itself and the annotation guidelines are freely available online. These characteristics make ACTER an especially useful resource for research into machine learning methodologies for ATE, as demonstrated in the current project.

Based on ACTER, HAMLET has been developed, a Hybrid Adaptable Machine Learning approach to Extract Terminology. This system is hybrid, in the sense that it follows the traditional hybrid methodology for ATE and combines both linguistic and statistical clues to detect terminology. A list of unique candidate terms (CTs) is extracted based on part-of-speech (POS) patterns and this is list is further filtered and sorted based on other information, including statistical features, such as termhood and unithood measures (Kageura and Umino 1996). In contrast to the traditional hybrid approach, the HAMLET methodology is adapted to fit a supervised machine learning perspective. Instead of a single or very limited list of statistical features with manually set thresholds, dozens of features of various kinds are calculated and automatically combined using supervised machine learning. The aim of this research was not to build the ATE system with the highest possible f1-scores, but rather to explore the potential of supervised machine learning for ATE in more detail, since "to what extent the cur-

rent machine learning approach can deal with issues in terminology extraction is yet to be seen" (Kageura and Marshman 2019). Therefore, this contribution will focus on the strengths and weaknesses of the methodology and the impact of the various components. How important is domain-specific training data? Can training data from different languages be combined and what is the role of the volume of training data? How does the methodology perform on the different term types and how do terms and Named Entities relate in this context? Which are the most common errors and, finally, how do the different features contribute to the results?

Section 2 starts with an overview of the state-of-the-art of ATE and is followed by Section 3 on the ACTER dataset. Section 4 is dedicated to the methodology and the experiments. It discusses the experimental setup, the results per corpus and the impact of various factors on these results. Section 5 presents a more detailed analysis and discusses the impact of the different types of annotations and features. The paper concludes with a discussion and ideas for future research.

2. Related research

Traditionally, ATE methodologies have been categorised into three different types: linguistic, statistical, and hybrid. Linguistic systems rely on information from the linguistic preprocessing of texts, using POS tagging and, occasionally, more advanced syntactic chunking or parsing. Statistical systems depend on frequencies to calculate termhood and unithood measures (Kageura and Umino 1996), often using a general language reference corpus for comparison. Hybrid systems combine both strategies by selecting CTs using the linguistic method and sorting and filtering this list with statistical measures. These hybrid systems have long set the tone for ATE research and still obtain state-of-the-art performance, provided the language is well-resourced enough that reference corpora and POS tagging are available. Many current state-of-the-art systems still maintain some variation of this methodology (Drouin 2003; Kessler, Béchet, and Berio 2019; Kosa et al. 2020; Macken, Lefever, and Hoste 2013; Šajatović et al. 2019).

However, in recent years, research into ATE has outgrown this linguistic/statistical/hybrid typology. Gao and Yuan (2019) propose a typology of five, calling the three original categories "rule-based", "statistical", and "hybrid", and adding "machine-learning based" and "deep-learning based". While it is true that the original typology is due for an update, their suggestion may not be ideal, in the sense that "deep learning" is technically a type of "machine learning", which is considered the opposite of "rule-based". Moreover, it combines two characteristics into a single methodology (the type of features, e.g., linguistic vs. statistical,

and the type of algorithm, i.e., rule-based vs. machine learning). Furthermore, the variation in methodologies has increased to such an extent, that the fundamental differences can no longer be captured in 3 or 5 categories. Nevertheless, it can still be valuable to provide a theoretical framework to identify different methodologies. Consequently, we suggest a different approach to describe ATE methodologies in a clear and comprehensive way, while still leaving room for all possible variation in such a productive field of research.

To categorise methodologies for ATE, four fundamental aspects were identified in which current methodologies for ATE differ. Rather than trying to fit all methodologies into a single categorisation, we propose defining the methodologies according to the following four aspects, which are explained in more detail below:

1. Candidate term selection
2. Algorithm
3. Features
4. Term variation

Candidate term selection (1) refers to the preprocessing step of ATE where it is decided which lexical units are to be considered as potential terms. As mentioned, in the traditional, hybrid methodology, this would be done based on a predefined list of POS patterns, as with the TermoStat (Drouin 2003) and TExSIS (Macken, Lefever, and Hoste 2013) systems. Rather than starting from a predefined list of POS patterns, the POS patterns can also be derived from training data, as was done in the work of Patry and Langlais (2005), who trained a POS-based language model to determine appropriate POS patterns for terms. Another strategy is looking at n-grams (any sequence of n tokens), regardless of POS. This approach was tested, among others, by Wang et al. (2016). A third approach is sequence labelling, where, rather than extracting a flat list of CTs, each token in the text is considered sequentially and in relation to the surrounding tokens, for instance with IOB-labelling (where each token is either *Inside*, or *Outside* of a term, or the *Beginning* of a term) or some variation thereof (Kucza et al. 2018; McCrae and Doyle 2019). Most approaches fit into one of these three categories relatively easily, though there may be some exceptions. For instance, Gao and Yuan (2019) use a sequential approach with deep learning and, rather than traditional IOB-labelling, they work with all possible term spans in each sentence, with spans up to a maximum term length *k*, where k must be smaller than or equal to the sentence length. This allows the detection of nested terms with a sequential approach. This could be marked as a hybrid of the second and third categories proposed for the CT selection aspect: a sequence labelling approach with n-grams.

The **algorithm (2)** can easily be split into rule-based and machine learning methodologies. However, especially in the case of machine learning methodologies, many more distinctions can be made, e.g., supervised vs. semi-supervised vs. unsupervised. An example of an unsupervised deep learning approach is the work of Shah et al. (2019), who use statistical termhood and unithood features to find the most likely CTs, and then find similar terms through a siamese neural network with word embeddings. We refer to their work for more information about supervised vs. unsupervised methodologies for ATE. Machine learning algorithms can, of course, also be divided according to the learner. Many different kinds have already been used for ATE, e.g., logistic regression (Bolshakova, Loukachevitch, and Nokel 2013; Fedorenko, Astrakhantsev, and Turdakov 2013), the ROGER evolutionary algorithm (Azé et al. 2005), rule induction with RIPPER (Foo and Merkel 2010), CRF++ (Judea, Schütze, and Brügmann 2014), decision trees (Karan, Snajder, and Dalbelo Basic, Bojana 2012), support-vector machines (Ljubešić, Erjavec, and Fišer 2018), AdaBoost (Patry and Langlais 2005), k-nearest neighbours (Qasemizadeh and Handschuh 2014), and many types of neural networks (Amjadian et al. 2018; Hätty and Schulte im Walde 2018; Kucza et al. 2018; Shah, Sarath, and Shreedhar 2019; Wang, Liu, and McDonald 2016).

For the third aspect, **features (3)**, we do not attempt an exhaustive classification, since the variety and creativity of features that are invented to detect terms is too great. However, we do propose a number of categories for some of the most common types of features, specifying that methodologies may combine any number of these types of features. The first two categories have already been mentioned: linguistic (POS patterns, parsing, stopwords, etc.) and statistical (termhood and unithood measures). Related to the former are morphological or shape-related features, (e.g., length, capitalisation, presence of special characters, Greek or Latin forms etc.) and related to the latter are frequency features (frequencies that have not yet been transformed into statistical measures). Another large category is reserved for features based on external resources, such as existing terminologies and ontologies, Wikipedia, or internet searches. For instance, Vivaldi and Rodríguez (2001) rely on the lexical database EuroWordNet, Loukachevitch (2012) uses both features based on an internet search, and features based on a domain-specific thesaurus, and Ramisch et al. (2010) use the results of internet search engines as well. The next type of features are those based on topic modelling, as in the works of Šajatović et al. (2019) and Loukachevitch and Nokel (2013). Two less commonly used features are those based on language models, like measuring perplexity (Foo 2009), and features related to the layout and position of the term. Such features have been used for unsupervised training data generation (Judea, Schütze, and Brügmann 2014) or for related tasks such as indexing (Koutropoulou and Efstratios 2019). A hypothetical reason for the relative absence

of such potentially informative features, like the occurrence of CTs in bold or italics, or in the titles of texts, may be the fact that most systems work with plain text files, in which such information is not readily available. Another category can be reserved for features relating to context. For instance, the proximity of a CT to other highly scored CTs (Vivaldi, Màrquez, and Rodríguez 2001). The final category is devoted to features that use word- or character-embeddings, which are becoming ever more prevalent. Recently, embeddings are used in both feature-based and so-called "featureless" methodologies (Gao and Yuan 2019; Wang, Liu, and McDonald 2016). In the TermFrame project (Pollak et al. 2019) FastText embeddings trained on the small, domain-specific corpus are used to extend the list of CTs obtained through a traditional, hybrid approach.

The final fundamental aspect in which ATE methodologies differ, is in how they handle **term variation (4)**. Many systems currently do not go beyond providing the user with a flat list of (lowercased) unique CTs. A first step towards handling term variation is to perform lemmatisation or stemming (e.g., Conrado et al., 2013), i.e. combining different full forms of the same term. Related to this, one can distinguish formally identical terms with a different POS, e.g., having separate entries for *type* as a noun and as a verb. Handling term variation can go much further as well, e.g., with automatic abbreviation/acronym detection (Meyers et al. 2018) or automatic detection of syntactic term variation (Ville-Ometz, Royauté, and Zasadzinski 2007).

Apart from these four fundamental aspects, there are many other ways in which methodologies for ATE can differ. For instance, the limitations placed on the types of terms a system aims to extract can have a large impact: minimum frequency, minimum or maximum term length, limited POS patterns, inclusion or not of nested terms, etc. Other examples are the amount of preprocessing (e.g., removal of single-character terms, punctuation, special characters, etc.), specialisation to a single language and/or domain, binary or multiclass approaches, etc. It is beyond the scope of this paper to discuss each of these differences in detail, but imperative to consider that such differences have an enormous impact on the results and, therefore, also on the evaluation of ATE.

3. ACTER Annotated Corpora for Term Extraction Research

The ACTER dataset consists of manually annotated corpora in four specialised domains (corruption, dressage (equitation), heart failure, and wind energy), and three languages (English, French, and Dutch). It contains annotations with four different labels (Specific Terms, Common Terms, Out-of-Domain Terms, and Named Entities). The dataset has been introduced in previous publications,

(Rigouts Terryn et al., 2018; Rigouts Terryn, et al., 2020), and served as the basis for the TermEval 2020 shared task on ATE (Rigouts Terryn, et al., 2020). It is publicly available with up-to-date documentation as a Github repository,[1] and is accompanied by annotation guidelines.[2] For the current project, version 1.4 was used.

Around 60k tokens have been manually annotated per language & domain. All annotations were made in the texts, not using any preprocessing or filtering of CTs. Each occurrence of a term was annotated separately. Moreover, there were few limitations on what might be considered a term: no minimum frequency, no minimum or maximum length, no restrictions based on POS patterns (apart from that it had to contain a content word), and all nested terms were annotated as well. This makes it a challenging gold standard for ATE, with many hapax terms (that occur only once) and few ways to quickly reduce the search space (since using a frequency threshold, or length or POS restrictions will immediately limit recall). However, these characteristics also make it a more realistic gold standard.

Aside from the distinction between terms and Named Entities, there were three different term labels: Specific Terms, Common Terms, and Out-of-Domain Terms. These three categories of terms are defined according to their *domain-specificity*, i.e., how typical is the term for the relevant domain, and *lexicon-specificity*, i.e., how specialised does one have to be to know the term (or is it part of general language?). For instance, in the domain of heart failure, a term like *ejection fraction* is both domain-specific (strongly related to heart failure, as it refers to the percentage of blood that leaves the heart during each contraction) and lexicon-specific (laypersons will generally not know the term, only medical professionals will). Therefore, *ejection fraction* is a Specific Term. An example of a Common Term for that domain is *heart*, which is domain-specific (relevant to the domain of heart failure), but not lexicon-specific (it is assumed that every layperson has a basic knowledge of the concept). Conversely, *p-value* would be the opposite: not domain-specific (it is more of a statistic term, rather than a medical term), and lexicon-specific (some specialisation in statistics is required to understand the concept). Other researchers have used similar intuitions to define or categorise terms. For instance, Meyers et al. (2018) employ the model of the naïve adult, using Homer Simpson to decide whether a lexical unit might be considered specialised enough to be a term (if Homer Simpson would know it, it is not a term), in combination with the Juvenile Fiction subcorpus of COCA. Drouin et al. (2018) and Hätty et al. (2017) distinguish between different types of terminology as well.

1. https://github.com/AylaRT/ACTER
2. http://hdl.handle.net/1854/LU-8503113

The distinction between the four labels allows researchers to, at least partially, tailor the definition of terms to different possible applications. For instance, ATE used for ontology construction would ideally exclude Out-of-Domain terms which are not relevant to the ontology, whereas ATE for translators might be most useful for Specific Terms, which are not part of the translator's general vocabulary. Since HAMLET requires a binary distinction between terms and non-terms, we merge different annotation categories, as was also done in the KAS-term project (Ljubešić, Fišer, and Erjavec 2019). For the experiments, unless mentioned otherwise, all annotations are combined, so that Specific, Common and, Out-of-Domain Terms, and Named entities are all considered positives. The ±60k tokens of text annotated per language & domain, resulted in 119,450 annotations over 719,338 tokens.

4. Methodology and experiments

4.1 Experimental setup

4.1.1 *Preprocessing and CT selection based on POS*

All corpora are linguistically preprocessed using LeTs Preprocess (van de Kauter et al. 2013), which includes tokenisation, lemmatisation, POS tagging, chunking, and Named Entity Recognition (NER). To allow multilingual models and fair cross-lingual experiments, a single set of POS tags was required. The original sets used by LeTs are language-dependent: the English, French and Dutch tag sets were from Penn Treebank, TreeTagger and CGN (Corpus Gesproken Nederlands) respectively. Universal Dependencies (UD) (Petrov, Das, and McDonald 2012) were used as a basis, and a mapping was already available for the English and Dutch tagsets. By comparing these existing mappings, one could also be derived for the French tags. However, since the original LeTs tagsets were all more fine-grained, a few tags were added, which resulted in a final set of 26 tags. Starting from this fine-grained POS tagset, a more coarse-grained simple POS set was created counting 8 tags. Both sets are shown in Table 1.

Table 1. Standard and simple POS tagsets

Standard POS	Simple POS	Description
ABR	X	Abbreviation
ADJ	ADJ	Adjective
ADP	FUNC	Adposition (prepositions etc.)
ADV	ADV	Adverb
CONJ	FUNC	Conjunction
DET	FUNC	Determiner
FW	X	Foreign word
INTJ	X	Interjection
NOUN	NN	Noun
NUM	X	Numeral
PART	FUNC	Particle
PNO	FUNC	Pronoun (other)
PNPR	FUNC	Pronoun (personal)
PNPS	FUNC	Pronoun (possessive)
PROPN	FUNC	Proper noun
PUNCT	PUNCT	Punctuation (general)
PUNQ	PUNCT	Punctuation (quotation marks)
PUNB	PUNCT	Punctuation (parentheses)
SYM	X	Symbol
VB	VB	Verb (infinitive)
VBG	VB	Verb (present participle or gerund)
VBN	VB	Verb (past participle)
VBPA	VB	Verb (past tense)
VBPR	VB	Verb (present tense or imperative)
VBX	VB	Verb (other)
EOS	EOS	End Of Sentence

Since HAMLET is based on traditional ATE and aims to extract a list of all unique terms rather than all occurrences of each term, the next step was to decide how to combine terms. To make an informed decision, scenarios were tested with six types of variants. To avoid an overly fine-grained system which could be overly sensitive to small tagging errors, the simple POS tagset (rather than the more fine-

grained standard set) was used. The different variant forms, illustrated with the example term *Co-morbidities*, are:

1. Token_noPOS (token with original casing, without POS pattern), e.g., *Co-morbidities*
2. token_noPOS (lowercased token, without POS pattern), e.g., *co-morbidities*
3. Token_POS (token with original casing, with POS pattern), e.g., *Co-morbidities(NN)*
4. token_POS (lowercased token, with POS pattern), e.g., *co-morbidities(NN)*
5. lemma_POS (lowercased lemma, with POS pattern), e.g., *co-morbidity(NN)*
6. normalised_noPOS (normalised[3] token, without POS pattern), e.g., *comorbidities*

For each annotation, all six variant forms were constructed, and we calculated how consistently each form was annotated, i.e., out of all occurrences of that variant form in the corpus, how often it was annotated. While a small margin of inconsistency will always remain to account for both human error in the manually annotated data, and terms which may only be valid terms in some contexts, the goal was still to limit this inconsistency. On the other hand, since low-frequency terms are notoriously difficult for ATE, variants that are able to capture more annotations under a single entry (e.g., combining identical terms with different capitalisation) could also be beneficial, since they reduce the total number of unique terms (in that variant), and the number of very rare terms. Table 2 shows the average consistency and total number of gold standard terms per variation. As can be seen, there are three variants that lead to over 90% consistent annotations on average: Token_POS, token_POS, and lemma_POS. Since the token_POS variant considerably reduces the number of different annotations compared to Token_POS, while at the same time scoring very high on consistency, this is the variant that will be used for all experiments. Nevertheless, the same methodology can be applied with the other variants as well.

After linguistically preprocessing all texts, mapping all POS tags to a shared, language-independent set, and deciding to work with the token_POS variant, a preliminary list of unique CTs was extracted. HAMLET follows the traditional hybrid method for ATE and selects a list of CTs based on POS patterns. However, contrary to the traditional methods, these patterns are not manually defined, but extracted from the training data. This means that no restrictions had to be predefined with respect to term length, frequency, or POS type, but that all of this information would be derived from the training data. Since POS patterns were

3. Normalised in this case meant converting all tokens to only [a-z][0–9] characters, unless this meant no characters were left, in which case *UNK* was used as a placeholder.

Table 2. Total number of (unique) annotations per variant and how consistent annotations are when using this variant; for each annotated term/Named Entity as the specified variant, how many of the occurrences of that string in the text are annotated

Variant	# Annotations	% Occurrences that are annotated (average)
Token_noPOS	21,270	82.1%
token_noPOS	18,801	79.5%
Token_POS	22,776	92.9%
token_POS	**20,459**	**91.8%**
lemma_POS	18,375	90.2%
normalised_noPOS	18,611	79.0%

derived from the automatically tagged data, this means wrongly tagged patterns will be included. While this may lead to more noise among the CTs, it could also benefit recall when similar tagging errors are made in the test data. In English, 436 unique POS patterns were found; in French 353, and in Dutch 283. When excluding Named Entities, the numbers are slightly lower at 331 (en), 277 (fr), and 2020 (nl). Out of all term POS patterns, 61–70% contain at least one noun.

This selection of CTs based on POS patterns aims for high recall, since these are the CTs for which features will be calculated, to be processed by the machine learning classifier. The goal is not to "lose" (m)any real terms before these features can be calculated, so they can be used to make a more informed decision about which CTs to discard. However, this comes at a cost to precision. When the POS patterns of the test corpus itself are included, precision ranges between 3.2% and 8.4% and recall is perfect. When POS patterns from the test corpus are excluded, and only patterns from the other domains are used, precision is similar between 3.1% and 11.2%, but, since some POS patterns only occur in a single domain, recall is no longer perfect and ranges between 91.1% and 98.4% (95% on average). While the loss in recall is limited, this stresses the impact of both volume and relevance of training data.

4.1.2 *Features*

For each of the extracted CTs, 177 features were calculated. Most of these were based on the typical information used for hybrid ATE, such as termhood and unithood. Since previous research has repeatedly shown that no single feature appears to work best for all cases (Loukachevitch 2012), we investigated different feature combinations. Some of the features have not (often) been used for ATE and were based on findings during the annotation process. For instance, especially in the medical domain, terms often occur both in full, and as an abbreviated version.

In such cases, they are introduced in the full form, followed by the abbreviation between parentheses, e.g., *heart failure (HF)*. Therefore, features were added to indicate whether a CT occurs in the vicinity of parentheses. All features were divided into 6 groups and 18 subgroups. A summary per subgroup is given below and a complete overview is included in the Appendix.

A few of the features rely not only on the domain-specific corpora, but also on general language reference corpora. Two separate types of reference corpora were used per language: a Wikipedia reference corpus, based on Wikipedia dumps, and a newspaper reference corpus.[4] All reference corpora were limited to 10M tokens and artificially split into 5000 documents. Features that make use of reference corpora are always calculated twice, i.e., once for each type of reference corpus. Non-numeric features are converted to (one-hot) vectors. A non-trivial task was finding a way to encode the POS pattern into informative features, without having to add a separate feature for each of the 300+ possible patterns. Based on preliminary experiments, we decided to work with three vector representations: two one-hot vectors for all POS tags (not patterns) to represent the POS of the first token and the last token and one frequency vector for the tags of all tokens of the CT. For instance, a term like heart failure (noun+noun) would get three vectors representing all POS tags, with zeros in all places except for the first noun in the first vector (1), the last noun in the second vector (1), and the sum of all nouns in the last one (2).

Before training, all statistical features (including those in the variational features), were scaled using scikit-learn's (Pedregosa et al. 2011) RobustScaler, which is more robust towards outliers. Features without any variance were automatically removed, which mostly concerned the POS-related features, since not all POS tags can occur in first/last position. Out of 177 possible features, 150–160 usually remained (depending on the setup).

4.1.3 *Algorithm, evaluation, and optimisation*

Evaluation and optimisation of the models was based on f1-scores (harmonic mean of precision and recall). In this context, precision is defined as the percentage of true terms among all extracted CTs (number of true positives, divided by number of extracted terms), and recall as the percentage of all true terms that have been extracted (number of true positives divided by the number of gold standard terms). Evaluation is strict, in the sense that only exact matches are counted as correct. Relatively low scores were expected due to the inherent difficulty of

4. The newspaper reference corpora per language were: the English News on Web corpus (Davies 2017), the French Gigaword corpus (Graff, Mendonça, and DiPersio 2011), and the news-related subcorpora of the Dutch openSONAR (Oostdijk et al. 2013)

Table 3. Description of features per group and subgroup

1. Shape features (SHAP)	
length	number of characters & number of tokens
alphanumeric	whether the CT is alphabetic, numeric, alphanumeric, etc. & the number of digits and non-alphabetic characters
capitalisation	out of all occurrences of the CT, how often (%) is it all lowercase, all uppercase, title case, etc.
2. Linguistic features (LING)	
first POS	POS tag of the first token of the CT (simple & standard POS)
last POS	POS tag of the last token of the CT (simple & standard POS)
freq. POS	how frequently each POS tag (simple & standard) occurs within the CT (simple & standard POS)
NER	whether the CT was tagged (completely, partially, etc.) as a Named Entity during preprocessing
chunk	which chunk tag(s) were assigned to the CT in preprocessing
stopword	whether the CT contains a stopword or is a stopword*
3. Frequency features (FREQ)	
spec. freq.	relative (document) frequency in the specialised corpus
ref. freq.	relative (document) frequency in the news and Wikipedia reference corpora
4. Statistical features (STAT)	
stats without ref.	metrics to calculate termhood/unithood without comparing to a reference corpus: C-Value (Barrón-Cedeño et al. 2009), TF-IDF (Astrakhantsev, Fedorenko, and Turdakov 2015), Lexical Cohesion and Basic (Bordea, Buitelaar, and Polajnar 2013)
stats with ref. (basic)	metrics to calculate termhood/unithood by comparing frequencies to a reference corpus: Domain Pertinence (Meijer, Frasincar, and Hogenboom 2014), Domain Relevance (Bordea, Buitelaar, and Polajnar 2013), Weirdness (Astrakhantsev, Fedorenko, and Turdakov 2015), Relevance (Peñas, Verdejo, and Gonzalo 2001), Log-Likelihood Ratio (Macken, Lefever, and Hoste 2013)
stats with ref. (advanced)	similar to the basic termhood/unithood measures, but these measures do not just use the frequencies of the entire CT in the reference corpora, but also those of all separate tokens that make up the CT: Vintar's termhood measure (Vintar 2010), Domain Specificity (Kozakov et al. 2004)
5. Contextual features (CTXT)	
parentheses	CT occurs between parentheses or right before/after parentheses

Table 3. *(continued)*

6. Variational features (VARI)	
var. numbers	number of different variations for the CT for each variant type (types explained in Section 4.1.1) & the combined relative frequency in the domain specific corpus of the CT in each variation per variant
var. stats	sum of the domain specificity and Vintar termhood scores of all different variations for the CT for each variant type

* The ISO stopwords were used for all languages: https://github.com/stopwords-iso

the ACTER dataset (no minimum or maximum term length, no minimum frequency, no limitations on POS patterns, inclusion of nested terms). Preliminary experiments were performed to choose the best algorithm for this task. Since there is no way to predict the best algorithm for a specific task and dataset beforehand (no free lunch theorem (Wolpert 1996)), a relatively wide range of classifiers was tested. With scikit-learn, the decision tree classifier (DTC), random forest classifier (RFC), multi-layer perceptron (MLP), and logistic regression (LOGREG) were compared, all allowing hyperparameter optimisation. In these preliminary experiments, the best average f1-scores were obtained with the RFC, followed by DT (−5,6 percentage points), LOGREG (−19.4), and MLP (−20.5). All experiments reported in the current contribution were, therefore, performed with scikit-learn's RFC. Hyperparameter optimisation was performed through grid search with five folds. Whenever k-fold cross-validation was used, five folds were used, with nested hyperparameter optimisation.

4.2 Results per corpus

With the basic methodology outlined (variant token_POS, CTs based on standard POS patterns, RFC with 177 features), five experimental setups were defined. The first, basic setup, is trained on three out of the four domains in a single language and tested on the held-out test corpus (domain) in that same language (e.g., training on English corruption, dressage, and wind energy corpora, testing on English heart failure corpus). This was deemed the most realistic real-world setup, since in-domain test data is rarely available. However, to get a better idea of the impact of the training data, four additional setups were defined. As can be seen in the summary in Table 4, setups 1 and 2 use a separate, held-out test corpus and only train on data from other corpora. In setup 2, corpora from other languages are included, so domain-specific training data is included, but only in different languages from the test corpus. Setups 3, 4, and 5 do not have a held-out test corpus but use 5-fold cross-validation to train and test on a single corpus (setup 3), all

domains in a single language (setup 4), or all corpora in all languages (setup 5). The CT extraction based on POS patterns (see 4.1.1) always excludes the POS patterns from the test corpus for a fair evaluation, except for setups 4 and 5, which are evaluated simultaneously on multiple corpora, so that it is impossible to exclude all POS patterns. Only the language-specific POS patterns from the training corpora are used for CT extraction.

Table 4. Overview of experimental setups

Setup	Train/test	Lang.	Training	Corpus-specific POS patterns
1	held-out test	one	3 other domains in same language	excluded
2	held-out test	all	all other corpora in all languages	excluded
3	5-fold cv	one	single corpus (1 domain, 1 language)	excluded
4	5-fold cv	one	single language (all domains, 1 lang.)	included
5	5-fold cv	all	all corpora (all domains, all lang.)	included

Tables 5 and 6 contain the results for each of the described experimental setups per corpus. Table 5 shows the results for models trained and evaluated on terms (Specific, Common, and OOD Terms) and Named Entities, while Table 6 excludes Named Entities. So, both tables show results on with the same methodology, with the exact same data, except that in the first table, Named Entities are considered positive instances, and in the second they are not. All results are averaged over three trials, and the average standard deviation between trials is only 0.8%. On average, f1-scores for the models excluding Named Entities are 3.7 percentage points lower, with the biggest difference for the domains of corruption and dressage, especially in English. The cross-validation experiments are less influenced by the Named Entities than the experiments with a held-out test set. Since basic Named Entity Recognition is already included in the preprocessing, and it is generally considered an easier task, since Named Entities have clearer characteristics, it is not surprising that performance is higher when Named Entities are excluded. Results, both with and without Named Entities, are state-of-the-art. Compared to the TermEval shared task (Rigouts Terryn, et al., 2020), which was based on the same dataset, they are similar to the best-performing system using a deep neural network with BERT models (Hazem et al. 2020).

Table 5. Precision (p), recall (r), and f1-scores (f1) in percentages per language and domain and per setup (see Table 6), trained and evaluated to extract all terms and Named Entities

Corpus		Setup 1			Setup 2			Setup 3			Setup 4			Setup 5		
Language	**Domain**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**
English	corp	33.6	50.2	40.2	37.6	50.1	43.0	38.8	50.5	43.8	48.7	58.0	52.9	50.2	60.5	54.9
	equi	54.2	52.5	53.3	57.6	58.3	58.0	54.9	66.4	60.1						
	htfl	55.6	37.0	44.4	44.7	56.1	49.8	47.1	67.3	55.4						
	wind	36.7	57.1	44.7	37.7	54.7	44.6	44.8	57.1	50.1						
French	corp	35.8	37.2	36.5	38.9	37.1	37.9	38.5	42.9	40.4	48.6	55.6	51.8			
	equi	54.7	46.7	50.3	58.2	47.6	52.3	55.9	56.5	56.1						
	htfl	60.0	50.2	54.7	56.6	55.9	56.3	51.4	74.3	60.8						
	wind	26.7	48.4	34.4	31.4	48.9	38.2	39.2	44.5	41.7						
Dutch	corp	37.2	51	43	41.4	52.8	46.4	41.0	56.7	47.4	51.6	69.4	59.2			
	equi	71.6	56.8	63.3	71.9	59.4	65.1	61.5	77.1	68.4						
	htfl	62.9	41.4	49.9	55.7	70.6	62.2	55.8	80.9	66.0						
	wind	32.9	76.1	45.9	33.5	75.9	46.4	43.3	65.4	52.1						
averages		**46.8**	**50.4**	**46.7**	**47.1**	**55.6**	**50.0**	**47.7**	**61.6**	**53.5**	**49.6**	**61.0**	**54.6**	**50.2**	**60.5**	**54.9**

Table 6. Precision (p), recall (r), and f1-scores (f1) in percentages per language and domain and per setup (see Table 5), trained and evaluated to extract all terms, without Named Entities

Corpus		Setup 1			Setup 2			Setup 3			Setup 4			Setup 5		
Language	**Domain**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**
English	corp	28.2	42.3	33.8	32.9	39.3	35.8	35.1	41.6	38.0	45.8	53.3	49.2	48.9	56.2	52.3
	equi	45.7	41.7	43.6	57.3	42.2	48.6	53.5	54.6	54.0						
	htfl	48.9	31.2	38.1	46.6	49.4	48.0	45.6	66.9	54.2						
	wind	32.8	47.1	38.7	39.9	40.2	40.1	43.9	49.5	46.4						
French	corp	29.2	28.4	28.8	30.3	30.1	30.2	32.1	36.3	34.0	47.9	51.2	49.4			
	equi	48.4	41.7	44.8	51.7	41.4	46.0	51.9	49.3	50.5						
	htfl	59.6	46.3	52.1	54.5	55.2	54.9	50.7	74.3	60.2						
	wind	31.1	34.2	32.5	31.7	38.2	34.6	36.8	37.8	37.3						
Dutch	corp	32.0	41.6	36.1	36.1	49.8	41.8	36.8	53.3	43.4	51.1	66.7	57.9			
	equi	72.2	53.8	61.6	74.8	51.7	61.1	60.8	74.3	66.8						
	htfl	67.4	35.2	46.2	55.9	66.0	60.5	55.5	81.8	66.1						
	wind	29.2	71.7	41.5	28.3	74.9	41.0	40.2	59.9	48.1						
averages		**43.7**	**42.9**	**41.5**	**45.0**	**48.2**	**45.2**	**45.2**	**56.6**	**49.9**	**48.3**	**57.1**	**52.2**	**48.9**	**56.2**	**52.3**

A lot of conclusions can be drawn from the results in Tables 5 and 6. First of all, given the mentioned difficulty of the task, the scores are promising with f1-scores up to 68.4%. Nevertheless, there is a lot of variation, and the lowest f1-score is only 28.8%, so further analysis is required. While, on average, precision and recall are similar and the top scores are very good (74.8% for precision, 81.8% for recall), the balance between the two varies greatly. The most extreme differences are seen for the Dutch corpus on wind energy, where the recall is up to 46.6 percentage points higher than precision. The first setup appears to be most sensitive to these differences, which might be another indication of the importance of domain-specific data. This unpredictability is an important issue for real-world applications, since, even if machine learning approaches get higher f1-scores than rule-based approaches, "for ATE to be usable, its results should be consistent, predictable and transparent" (Kageura and Marshman 2019).

There are notable differences between the results of the different setups. Setups 4 and 5, using cross-validation on all corpora, or all corpora in the same language, perform best on average (better even than cross-validation on a single corpus or cross-validation on all corpora combined). Setup 1, which is both the most realistic, but also the strictest, achieves the lowest f1-scores. The models in setup 2 are similarly evaluated on a held-out test corpus, but, as opposed to those in setup 1, they have access to training data in the other languages, including domain-specific data. This combination of more training data and domain-specific training data, even if it is in other languages than the test corpus, appears to give the models in setup 2 a slight advantage over setup 1. The importance of volume of training data might also explain why results for cross-validation setup 3 are lower than the other cross-validation setups, despite being trained exclusively on the same domain and language. The potential of including data in other languages is not entirely clear, since it appears to help in setup 2 versus setup 1, and potentially also in setup 5 versus 4.

Language appears to have an undeniable impact on terminology. The results for Dutch are noticeably higher than those for the other languages, and French scores lowest. The most probable explanation is that Dutch compounding rules make ATE slightly easier. In Dutch, nominal terms are often long compound nouns, rather than multi-word terms, as in English and French, e.g., *ejectiefractie*, compared to *ejection fraction* and *fraction d'éjection*. Therefore, in Dutch, there are more single-word terms (which are easier to detect, see further). Domain has a notable impact on the results as well. Both dressage and heart failure obtain relatively high f1-scores compared to wind energy and corruption. The average f1-scores per domain (setups 1–3) including Named Entities are 42.1%, 58.5%, 55.5%, and 44.2% for corruption, dressage, heart failure, and wind energy respectively. This aligns with how the annotation process was perceived, as both heart

failure and dressage were found to be easiest to annotate, and corruption most challenging. Consequently, it is difficult to distinguish between cause and effect: terminology might be objectively more difficult to identify for certain domains, influencing both human annotators and ATE, or the data may have been annotated better for the domains which were perceived as easier to annotate, making it more suitable for ATE. Likely, both have some effect. In conclusion, the results reported in Tables 5 & 6 show that performance is promising, but not always predictable, and variable per corpus. There is a considerable impact of training data, language, and domain. Overall, even models tested on a completely unseen language and domain obtain robust performance, so it can be concluded that the HAMLET methodology is able to generalise relatively well across corpora, and is, therefore, a viable strategy for ATE, even without domain-specific training data

5. Analysis and discussion

5.1 Error analysis

Despite promising scores, there is still a lot of room for improvement. A more detailed error analysis of a previous version of the system, including a comparison to a non-machine learning tool, has already been presented in a previous pilot study (Rigouts Terryn et al. 2019). Without going into the same amount of detail, this section will provide an error analysis with some of the remaining challenges. To avoid an overly complex examination per corpus and per setup, the results of setup 5 will be used for this purpose, since this is a single model trained and evaluated on all data combined through cross-validation. Since the goal is to look beyond precision, recall, and f1-score at the actual output, the analysis will focus on one run (one trial) of the system (the results in the previous tables were averaged over three trails). Also, since the focus is on the extraction of terms, rather than Named Entities, one of the models of Table 6 (trained and evaluated exclusively on terms) will serve as the basis for the analysis. While this means only a single experiment is discussed in detail in this section, most of the conclusions were found to hold for the other experiments as well. For this particular run of setup 5, precision, recall, and f1-score were 48.9%, 56.3%, and 52.3% respectively. There were 17,400 terms in the gold standard, and 320,063 CTs were extracted based on POS patterns. Of these, 20,043 were classified as terms, resulting in 9,795 true positives, 10,248 false positives, 7,605 false negatives, and 292,415 true negatives.

As predicted, f1-scores are higher for single-word terms (58.6%) than for multi-word terms (41.8%). Recall for single-word terms is especially good at 69.1%, while only 39.2% of all multi-word terms are found. Single-word terms are gener-

ally considered easier to extract since only termhood needs to be calculated, not unithood (which only needs to be measured for multi-word terms, to test whether the separate words form a cohesive unit), so this was expected. Similarly, it was unsurprising to find that hapax terms, which only occur once in the entire corpus, are more difficult to find than terms that occur at least twice: recall of hapax terms is 45.1% versus 66.3% for more frequent terms. While some features are included that do not rely on frequency at all, these alone are not very efficient at detecting terms (see Section 5.3). Nevertheless, as seen in previous research (Rigouts Terryn et al. 2019), this approach still performs better for rare terms than the traditional hybrid method. The impact of a minimum frequency for CTs is further illustrated by the fact that even increasing the minimum frequency to 2, leads to a gain in f1-scores of 6.5 percentage points for setup 5 (when also evaluated on CTs that occur at least twice). However, since there are many hapax terms, the evaluation compared to the complete gold standard that still includes hapax terms drops by 11.3 percentage points. This is important to remember when comparing different ATE systems, as most do work with a minimum frequency threshold.

Recall of the different term types is 60.8% (Specific Terms), 48.3% (Common Terms), and 46.2% (OOD Terms). Since Specific Terms are those that are likely to be most important for most projects (these are the terms in the strictest sense of the word: both domain- and language-specific), it is promising to see a relatively high recall there, especially considering that, among Specific Terms, frequencies are often low. Lower recall with Common Terms may be due to the fact that they are not language-specific and are easier to confuse with general vocabulary. OOD Terms, conversely, are language-specific but not domain-specific, so they are also not the typical target for term extraction features. Even though the system was trained to find only terms, some Named Entities were included in the output as well, but compared to systems trained to find Named Entities, recall for this category was low at only 11.1%. To illustrate: an identical system trained to find both terms and Named Entities has a recall of 69.1% for the latter. This implies that, while some confusion between terms and Named Entities remains possible due to sometimes similar characteristics, in general, the system adapts well to the training data.

To get a more detailed idea of the results, the predicted CTs were sorted based on the predicted probability that they would be true terms. Since setup 5 combines all corpora, the corpus of origin was displayed for each CT as well. A sample of the results can be seen in Table 7, where only the 25 most highly ranked English terms can be seen. Only 3 of these (in grey) are not in the gold standard. The predicted probability appears to be an effective way to sort results: the average probability for true positives is 77.2%, compared to 67.4% for false positives, 26.6% for false negatives, and 6.4% for true negatives. However, some non-terms can still be predicted

as terms with high probability scores, and the average probability for false positives is only 10 percentage points lower than for true positives. The difference between true negatives and false negatives is slightly higher, with an average predicted probability of 26.6% for the latter. A sample of the results can be seen in Table 7, where only the 25 most highly ranked English terms can be seen. Both the predicted probability score and the rank are shown in this table. Since setup 5 combines all corpora, the corpus of origin was displayed for each CT as well. Only three of the extracted CTs are not in the gold standard; they have been marked in grey.

Table 7. 25 highest ranked English CTs of model trained to extract only terms (no Named Entities), for setup 5, marking false positives in grey and showing the CT (as variant token_POS), domain of origin, predicted probability, and when sorted by probability

Candidate Terms (grey = false positives)	Domain	Predicted Probability	Rank
Strides(NN)	dres	99.5%	23
stride(NN)	dres	99.4%	39
carvedilol(NN)	htfl	99.3%	46
spironolactone(NN)	htfl	99.3%	49
canter(NN)	dres	99.3%	51
metoprolol(NN)	htfl	99.1%	75
impulsion(NN)	dres	99.0%	94
forehand(NN)	dres	98.8%	105
travers(NN)	dres	98.8%	108
ivabradine(NN)	htfl	98.7%	124
rein(NN)	dres	98.6%	142
angiotensin(NN)	htfl	98.6%	148
kccq(NN)	htfl	98.5%	161
pesade(NN)	dres	98.5%	165
ballotade(NN)	dres	98.5%	168
forelegs(NN)	dres	98.5%	171
airfoils(NN)	wind	98.4%	183
elastance(NN)	htfl	98.4%	184
sitagliptin(NN)	htfl	98.4%	186
etidronate(NN)	htfl	98.4%	192
elektrine(NN)	wind	98.3%	197
halts(NN)	dres	98.3%	198
gaits(NN)	dres	98.3%	203
resynchronization(NN)	htfl	98.3%	204
gait(NN)	dres	98.2%	209

A first observation concerning these most highly ranked terms is the presence of many Dutch terms (Dutch terms not displayed in table), and terms from the domains of dressage and heart failure, which are also the language and domains that tend to obtain the highest f1-scores. In the list of the 25 most highly ranked English terms in Table 7, it is remarkable to see only single nouns. When including the other languages as well, the most highly ranked non-noun appears at rank 81, and it is a POS tagging mistake (noun tagged as adverb): *binnenachterbeen* (nl), meaning *inside hind leg* in the domain of dressage. To find the first multi-word term, we must go down even further in the list, to rank 147: *peptides natriurétiques* (fr), (*natriuretic peptides* in English). While it has already been confirmed that recall, in general, is lower for rare terms, this does not mean that all infrequent terms are hard to extract. The second most highly ranked term of all, *gedragenheid* (nl), a term in the domain of dressage, only appears twice in the entire corpus.

Among highly ranking false positives there are a few Named Entities, e.g., *KCCQ,* which is an abbreviation of *Kansas City Cardiomyopathy Questionnaire*. This was one of the instances which caused some hesitation during the annotation process as well since it could reasonably be considered both a term and a Named Entity. It is interesting to find that such cases which were considered ambiguous during the annotation process, are also problematic for the ATE. Other notable false positives are those that possibly should have been annotated, but were not, e.g., *elastance,* which was only annotated when combined with adjectives (e.g. *ventricular elastance*), but not by itself. Again, such cases regularly caused doubts during the annotation process, as it is often far from clear where to draw the boundary between terms, parts of terms, and general language. Of course, not all false positives are so ambiguous that they might be considered correct after all. For instance, *electrine* in the corpus on wind energy is the second token in a Named Entity *Lietuvos Elektrine* that simply happens to occur often in the corpus, but which is clearly not a term. Nevertheless, it is encouraging to see that, especially among highly ranked CTs, many of the false positives are understandable mistakes, i.e., they resemble the types of disagreements that might also occur between human annotators.

Among the false negatives there are, as expected, many multi-word terms. There also appear to be more non-nominal terms, e.g., *dynamic, covariate, blacklisting, diseased, hydroelastic*. Some of these are due to tagging errors, e.g., *clinician* is tagged as an adjective instead of a noun. Most terms contain at least one noun, which is reflected in the output, where only 33 of the 1000 most highly ranked CTs do not contain a noun, so the system seems to have learnt that such CTs are less likely to be valid terms. However, it also means that terms that do not contain a noun but are still valid terms, are less likely to be extracted. Out of the 1000 most

highly ranked false negatives (so gold standard terms that were not detected), 187 do not contain any nouns. Among the false negatives, there are also many terms that are common in general language, and are, therefore, more difficult to detect. Sometimes this concerns Common Terms which are domain-specific but also very common in general language, e.g., breeze, downwind, political; in other cases, they are ambiguous Specific Terms which use general language words that only acquire a more specialised meaning in the context of a domain, e.g., yield and collection, which are Specific Terms in the domain of dressage.

To conclude, despite remaining difficulties, like multi-word terms, non-noun terms, and rare terms, the system can extract over half of all terms in a very difficult setting, with a reasonable precision. The fact that many of the errors of the system are "understandable" is a promising indication that the system has been able to learn a robust definition of terminology.

5.2 Impact of annotation types

To find out how adaptable the methodology is to different configurations of the four labels, the binary approach was maintained, but the definition of what was considered a valid term changed to include or exclude various labels (both in training and evaluation). For instance, in the experiment of the first row of Table 8, all annotated instances of all labels are considered as positives (1), while all non-annotated instances are considered negatives (0), whereas, in the second row, only term labels are considered as positives and annotated Named Entities, as well as non-annotated instances, are negatives. The results of these experiments, using setup 5, can be found in Table 8. As can be seen, f1-score is highest when all categories are included. Leaving out Named Entities leads to a drop in both precision and recall. Trying to extract only Specific Terms leads to the lowest performance. This was expected, since, by excluding all other annotations as positives, the dataset of CTs becomes even more imbalanced, making it even more difficult to correctly identify the few valid terms correctly. Since there are very few OOD Terms, they do not have a big impact on the results.

5.3 Impact of Features

5.3.1 *Feature group selection*

Since HAMLET combines so many features of different types, it is important to investigate the role these features play in the eventual models. As a first experiment, we trained models according to setup 1, i.e., the strictest setup, and tested performance when including and excluding various feature groups. Alle exper-

Table 8. Scores (as percentages) of the HAMLET classifier for setup 5, training and evaluating on different configurations of the labels (those in grey are excluded)

Specific Terms	Common Terms	OOD Terms	Named Entities	p	r	f1
Specific Terms	Common Terms	OOD Terms	Named Entities	50.2	60.5	54.9
Specific Terms	Common Terms	OOD Terms		48.9	56.2	52.3
Specific Terms	Common Terms			48.9	56.2	52.3
Specific Terms				45.4	50.1	47.6
Specific Terms			Named Entities	47.8	55.8	51.5
Specific Terms	Common Terms		Named Entities	50.0	59.9	54.5

iments were performed twice: once including both terms and Named Entities, once including only terms. Table 9 shows the results with all scores (average scores over all corpora), sorted based on the average f1-score without Named Entities. The first, and perhaps most remarkable observation is that the highest scoring model excludes statistical features, even though these features have long been some of the most important features in ATE research. However, this conclusion needs to be nuanced, since the variational features include statistical metrics as well, calculated for different variants of the CTs. This is illustrated by the fact that leaving out either statistical or variational features has little impact but leaving out both leads to a much bigger drop in the f1-scores.

In most cases, leaving out one feature group has only a limited impact, except for linguistic features, in which case performance suddenly drops to only 33.0%, showing the importance of these features. Using only linguistic and statistical features, as in typical hybrid ATE methodologies, leads to moderate performance. When models are trained with only a single group of features, shape features are most informative, followed by linguistic features. Contextual features are least informative when used by themselves, which is hardly surprising, since they are currently very limited (only features relating to parentheses). In conclusion, investigating the impact of the feature groups based on this limited feature group selection leads to some surprising results. In the next section, this is investigated in more detail.

5.3.2 *Feature importance*

With scikit-learn's RFC, it is possible to see the importance assigned to each feature. Because it was assumed different types of features might be important for different corpora, we looked at those from experimental setup 3, where cross-validation is used within a single corpus (so all features are learnt from the same corpus). To avoid overcomplicating the conclusions with Named Entities,

Table 9. Average f1-scores (as percentages) for setup 1, trained & evaluated on terms (incl. & excl. Named Entities), including, and excluding various features groups (see Table 4 for setups, Table 3 for features)

Included feature groups						f1-scores	
Shape	**Ling.**	**Freq.**	**Stat.**	**Context**	**Variants**	**Incl. NEs**	**Excl. NEs**
SHAP	LING	FREQ		CTXT	VARI	46.2	41.8
SHAP	LING		STAT	CTXT	VARI	45.2	40.5
SHAP	LING	FREQ	STAT	CTXT	VARI	45.5	40.4
SHAP	LING	FREQ	STAT	CTXT		45.1	40.3
SHAP	LING	FREQ	STAT		VARI	45.5	40.2
	LING	FREQ	STAT	CTXT	VARI	44.4	39.2
	LING		STAT			42.8	38.0
SHAP	LING	FREQ		CTXT		41.3	37.7
SHAP		FREQ	STAT	CTXT	VARI	37.9	33.0
SHAP						24.0	28.7
	LING					32.2	28.3
					VARI	28.3	24.3
			STAT			24.3	23.1
		FREQ				22.6	22.1
				CTXT		18.3	16.4

for these experiments the models were trained only on terms. The heat map of the results per feature subgroup is displayed in Table 10. For all corpora, variant scores are very important, as well as advanced statistical measures that use the reference corpora. Both of these feature subgroups contain almost the same metrics, but the former subgroup calculates them for different variants of the CT. By comparison, the other statistical scores that use reference corpora are assigned a surprisingly low importance, especially considering that even simple frequency features get higher importance scores. Another rather universally informative group concerns stop words, which is logical, since stop words are not automatically filtered out, but can still be identified through these features. The differences between the corpora are relatively limited. One of the only clear patterns is that Dutch models place a higher importance on length features, which is unsurprising given the language's compounding rules (long compound SWTs without spaces, which can be very informative for terms).

Table 10. Heat map of assigned importance per feature subgroup, per corpus, for model trained with setup 3 on terms, without Named Entities; green=higher, red=lower importance

	corp			equi			htfl			wind		
	en	fr	nl	en	fr	nl	en	fr	nl	en	fr	nl
CTXT: brackets												
FREQ: ref. freq.												
FREQ: spec. freq.												
LING: NER												
LING: chunk												
LING: first POS												
LING: freq POS												
LING: last POS												
LING: stopword												
SHAP: alphanumeric												
SHAP: capitalisation												
SHAP: length												
STAT: with ref. (advanced)												
STAT: with ref. (basic)												
STAT: without ref.												
VARI: var. numbers												
VARI: var. scores												

Since the differences between corpora appear relatively small, the analyses of all separate features will take a more general approach using setup 5. With this setup, 160 features were maintained (the others were discarded due to a lack of variance). For more information on the features, see the Appendix. Table 11 shows the 30 most highly ranked features with the assigned importance scores. Interestingly, the four most important features are all Vintar's termhood score (Vintar 2010) compared to the newspaper reference corpus, for different variants of the CT (the variant used for all other features is token_POS, i.e. lowercased token followed by simple POS). The same metric for this standard variant only occurs in eighth place as a statistical feature. Out of the 160 features, only 28 are assigned an importance of over 1%; 62 features score below 0.1%. The most highly ranked non-variational feature is the one indicating there are no stop words in a term (5th place). The first shape feature occurs in 18th place (number of tokens). The first POS-related features are in 22nd and 23rd place and indicate how many function words and adpositions the CT contains. One frequency feature makes it in the top 30, in 28th place: the relative document frequency in the specialised corpus.

Since there are so few contextual features, which are only relevant in very specific contexts, the first one only occurs in 71st place (CT is followed by an open parenthesis). Most of the bottom-ranked features are related to POS, which is understandable since there are so many of them and most will apply only to a minority of CTs. It is surprising that the most highly ranked POS features concern function words and adpositions, rather than nouns. The first noun-related features only occur in 36th to 38th place.

Analysing the features shows that the relation between them can be quite complicated and that many different types contribute towards the final results. Termhood and unithood calculations remain invaluable, and the more advanced calculations appear to more informative than the simple ones like TF-IDF.

6. Conclusions and future research

In conclusion, in recent years, ATE has evolved beyond the traditional rule-based methods of extracting CTs based on POS patterns and filtering them with term hood and unithood measures. This explosion of new methods means that a new theoretical framework is required to describe each method systematically and with enough detail, since the simple distinction between linguistic, statistical, and hybrid methods no longer suffices. In this contribution, we propose moving away from a simple categorisation, and describing ATE in terms of at least four aspects: CT selection, algorithm, feature types, and term variation. Moreover, this evolution has emphasised the need for large, diverse, and reliable datasets. For this project, the freely available ACTER dataset was selected.

Based on this dataset, the HAMLET Hybrid Adaptable Machine Learning approach to Extract Terminology was developed. The diversity of the data allowed the development and evaluation of a robust supervised machine learning approach. This system was elaborately tested to evaluate the impact of training data, language, domain, types of terms and Named Entities. A simple Random Forest Classifier yielded f1-scores up 68%, which is promising when considering the difficulty of the dataset and the strictness of the evaluation (many low-frequency terms, no restriction on POS, no maximum length, only count of full matches). While the methodology is robust and can be used on an unseen domain, results vary widely depending on language, domain, available training data, and type of annotation. Domain-specific training data can considerably improve results, but the amount of training data plays a role as well, and even data from a different language can be helpful. The methodology can be adapted to focus on different types of terms and/or Named Entities and works especially well for the most specialised (Specific) terms. Some of the same difficulties as in tra-

Table 11. 30 highest ranked features according to assigned importance (as percentages) for setup 5, model excluding Named Entities

Rank	Group	Feature Name (see appendix for explanations)	Imp.
1	VARI	variant(normalised_noPOS)_sum_termhood_vintar_news	5.2
2	VARI	variant(token_noPOS)_sum_termhood_vintar_news	5.2
3	VARI	variant(Token_noPOS)_sum_termhood_vintar_news	4.6
4	VARI	variant(lemma_POS)_sum_termhood_vintar_news	4.6
5	LING	stopword_none	4.2
6	VARI	variant(lemma_POS)_sum_domain_specificity_wiki	3.8
7	VARI	variant(Token_POS)_sum_termhood_vintar_news	3.5
8	STAT	vintar_news	3.4
9	LING	stopword_partial	3.2
10	VARI	variant(normalised_noPOS)_sum_domain_specificity_wiki	3.2
11	STAT	vintar_wiki	3.1
12	VARI	variant(Token_noPOS)_sum_domain_specificity_wiki	3.0
13	VARI	variant(Token_POS)_sum_domain_specificity_wiki	2.8
14	VARI	variant(token_noPOS)_sum_domain_specificity_wiki	2.7
15	STAT	domain_specificity_wiki	2.5
16	STAT	domain_specificity_news	2.4
17	VARI	variant(token_noPOS)_rel_freq_in_spec_corp	1.5
18	SHAP	nr_tokens	1.5
19	VARI	variant(normalised_noPOS)_rel_freq_in_spec_corp	1.5
20	VARI	variant(lemma_POS)_rel_freq_in_spec_corp	1.4
21	STAT	basic	1.4
22	LING	POS_simple_freq_FUNC	1.4
23	LING	POS_standard_freq_ADP	1.3
24	SHAP	nr_characters	1.3
25	SHAP	nr_non_letters	1.3
26	VARI	variant(Token_noPOS)_rel_freq_in_spec_corp	1.2
27	STAT	llr_news	1.2
28	FREQ	freq(doc)_in_specialised_corpus	1.1
29	STAT	relevance_wiki	0.9
30	STAT	relevance_news	0.9

ditional ATE remain, such as rare terms, multi-word terms, and non-noun terms, the large number of varied features do help towards a robust performance.

This project has inspired many ideas for future research. First and foremost, the goal is to consider context by developing a sequential labelling approach. The hypothesis is that contextual features might be complementary with the current features and, thus, lead to a better performance, especially for those types of terms which are currently still problematic. Second, a sequential approach would create the opportunity of experimenting with Recurrent Neural Networks, so that these can be compared to a feature-rich approach. Finally, the issue of term variation has so far not been addressed in any detail and could be both an effective and necessary improvement to the methodologies.

Funding

This research has been carried out as part of a PhD fellowship on the EXTRACT project, funded by the FWO-Research Foundation – Flanders.

References

Amjadian, Ehsan, Diana Zaiu 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): 23–40. https://doi.org/10.1075/term.00012.amj

Astrakhantsev, Nikita, D. Fedorenko, and D. Yu. Turdakov. 2015. "Methods for Automatic Term Recognition in Domain-Specific Text Collections: A Survey." *Programming and Computer Software* 41 (6): 336–49. https://doi.org/10.1134/S036176881506002X

Azé, Jérôme, Mathieu Roche, Yves Kodratoff, and Michèle Sebag. 2005. "Preference Learning in Terminology Extraction: A ROC-Based Approach." In *Proceeedings of Applied Stochastic Models and Data Analysis*, 209–2019. Brest, France. http://arxiv.org/abs/cs/0512050

Barrón-Cedeño, Alberto, Gerardo Sierra, Patrick Drouin, and Sophia Ananiadou. 2009. "An Improved Automatic Term Recognition Method for Spanish." In *Computational Linguistics and Intelligent Text Processing*, edited by Alexander Gelbukh, 125–36. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-00382-0_10

Bolshakova, Elena, Natalia Loukachevitch, and Michael Nokel. 2013. "Topic Models Can Improve Domain Term Extraction." In *Advances in Information Retrieval*, edited by Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz, 7814:684–87. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-36973-5_60

Bordea, Georgeta, Paul Buitelaar, and Tamara Polajnar. 2013. "Domain-Independent Term Extraction Through Domain Modelling." In *Proceedings of the 10th International Conference for Terminology and Artificial Intelligence (TIA)*, 61–68. Paris, France.

Conrado, Merley da Silva, Thiago A. Salgueiro Pardo, and Solange Oliveira Rezende. 2013. "A Machine Learning Approach to Automatic Term Extraction Using a Rich Feature Set." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 16–23. Atlanta, GA, USA: Association for Computational Linguistics.

Davies, Mark. 2017. "The New 4.3 Billion Word NOW Corpus, with 4--5 Million Words of Data Added Every Day." In *Proceedings of the 9th International Corpus Linguistics Conference. Birmingham*. Birmingham, UK. https://www.english-corpora.org/now

Drouin, Patrick. 2003. "Term Extraction Using Non-Technical Corpora as a Point of Leverage." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 9 (1): 99–115. https://doi.org/10.1075/term.9.1.06dro

Drouin, Patrick, Marie-Claude L'Homme, and Benoit Robichaud. 2018. "Lexical Profiling of Environmental Corpora." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, 3419–25. Miyazaki, Japan: European Language Resources Association.

Durán-Muñoz, Isabel. 2019. "Methodological Proposal to Build a Corpus-Based Ontology in Terminology." *Lingue e Linguaggi*. https://doi.org/10.1285/i22390359v29p581

Fedorenko, Denis, Nikita Astrakhantsev, and Denis Turdakov. 2013. "Automatic Recognition of Domain-Specific Terms: An Experimental Evaluation." In *Proceedings of the Ninth Spring Researcher's Colloquium on Database and Information Systems*, 26:15–23. Kazan, Russia.

Foo, Jody. 2009. "Term Extraction Using Machine Learning." *Linköping University, LINKÖPING*, 1–8.

Foo, Jody, and Magnus Merkel. 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*, 49–54. Valetta, Malta: European Language Resources Association.

Gao, Yuze, and Yu Yuan. 2019. "Feature-Less End-to-End Nested Term Extraction." *ArXiv:1908.05426 [Cs, Stat]*, August. http://arxiv.org/abs/1908.05426. https://doi.org/10.1007/978-3-030-32236-6_55

Graff, David, Ângelo Mendonça, and Denise DiPersio. 2011. "French Gigaword Third Edition LDC2011T10." Philadelphia, USA: Linguistic Data Consortium.

Hätty, Anna, and Sabine Schulte im Walde. 2018. "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)*, 62–73. Sante Fe, New Mexico, USA: Association for Computational Linguistics.

Hätty, Anna, Simon Tannert, and Ulrich Heid. 2017. "Creating a Gold Standard Corpus for Terminological Annotation from Online Forum Data." In *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017)*. Montpellier, France: Association for Computational Linguistics.

Hazem, Amir, Mérieme Bouhandi, Florian Boudin, and Béatrice Daille. 2020. "TermEval 2020: TALN-LS2N System for Automatic Term Extraction." In *Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020)*, 95–100. Marseille, France: European Language Resources Association.

Judea, Alex, 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. Dublin, Ireland: Dublin City University and Association for Computational Linguistics.

Kageura, Kyo, and Elizabeth Marshman. 2019. "Terminology Extraction and Management." In *The Routledge Handbook of Translation and Technology*, edited by O'Hagan, Minako. https://doi.org/10.4324/9781315311258-4

Kageura, Kyo, and Bin Umino. 1996. "Methods of Automatic Term Recognition." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 3 (2): 259–89. https://doi.org/10.1075/term.3.2.03kag

Karan, Mladen, Jan Snajder, and Dalbelo Basic, Bojana. 2012. "Evaluation of Classification Algorithms and Features for Collocation Extraction in Croatian." In *Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, 657–62. Istanbul, Turkey: European Language Resources Association.

Kauter, Marian van de, Geert Coorman, Els Lefever, Bart Desmet, Lieve Macken, and Véronique Hoste. 2013. "LeTs Preprocess: The Multilingual LT3 Linguistic Preprocessing Toolkit." *Computational Linguistics in the Netherlands Journal* 3: 103–20.

Kessler, Rémy, Nicolas Béchet, and Giuseppe Berio. 2019. "Extraction of Terminology in the Field of Construction." In *Proceedings of the First International Conference on Digital Data Processing (DDP)*, 22–26. London, UK: IEEE Computer Society. https://doi.org/10.1109/DDP.2019.00015

Kosa, Victoria, David Chaves-Fraga, Hennadii Dobrovolskyi, and Vadim Ermolayev. 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*, 1175:24–49. Communications in Computer and Information Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-39459-2_2

Koutropoulou, Theoni, and Efstratios Efstratios. 2019. "TMG-BoBI: Generating Back-of-the-Book Indexes with the Text-to-Matrix-Generator." In *Proceedings of the 10th International Conference on Information, Intelligence, Systems and Applications, IISA 2019*, 1–8. Patras, Greece. https://doi.org/10.1109/IISA.2019.8900745

Kozakov, L., Y. Park, T. Fin, Y. Drissi, Y. Doganata, and T. Cofino. 2004. "Glossary Extraction and Utilization in the Information Search and Delivery System for IBM Technical Support." *IBM Systems Journal* 43 (3): 546–63. https://doi.org/10.1147/sj.433.0546

Kucza, Maren, Jan Niehues, Thomas Zenkel, Alex Waibel, and Sebastian Stüker. 2018. "Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks." In *Proceedings of Interspeech 2018, the 19th Annual Conference of the International Speech Communication Association*, 2072–76. Hyderabad, India: International Speech Communication Association. https://doi.org/10.21437/Interspeech.2018-2017

Ljubešić, Nikola, 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.

Ljubešić, Nikola, Darja Fišer, and Tomaž Erjavec. 2019. "KAS-Term: Extracting Slovene Terms from Doctoral Theses via Supervised Machine Learning." In *Text, Speech, and Dialogue. TSD 2019*. Vol. 11697. Lecture Notes in Computer Science. Springer. http://arxiv.org/abs/1906.02053. https://doi.org/10.1007/978-3-030-27947-9_10

Loukachevitch, Natalia. 2012. "Automatic Term Recognition Needs Multiple Evidence." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, 2401–7. Istanbul, Turkey: European Language Resources Association.

Loukachevitch, Natalia, and Michael Nokel. 2013. "An Experimental Study of Term Extraction for Real Information-Retrieval Thesauri." In *Proceedings 10th International Conference on Terminology and Artificial Intelligence TIA 2013*, 69–76. Paris, France.

Macken, Lieve, Els Lefever, and Véronique 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): 1–30. https://doi.org/10.1075/term.19.1.01mac

Mayorov, V., I. Andrianov, Nikita Astrakhantsev, Avanesov, V., Kozlov, I., and Turdakov, D. 2015. "A High Precision Method for Aspect Extraction in Russian." In *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue."* Vol. 2. Moscow, Russia.

McCrae, John P., 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. Dublin, Ireland.

Meijer, Kevin, Flavius Frasincar, and Frederik Hogenboom. 2014. "A Semantic Approach for Extracting Domain Taxonomies from Text." *Decision Support Systems* 62 (June): 78–93. https://doi.org/10.1016/j.dss.2014.03.006

Meyers, Adam L., Yifan He, Zachary Glass, John Ortega, Shasha Liao, Angus Grieve-Smith, Ralph Grishman, and Olga Babko-Malaya. 2018. "The Termolator: Terminology Recognition Based on Chunking, Statistical and Search-Based Scores." *Frontiers in Research Metrics and Analytics* 3 (June). https://doi.org/10.3389/frma.2018.00019

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. "The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch." In *Essential Speech and Language Technology for Dutch*, edited by Peter Spyns and Jan Odijk, 219–47. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30910-6_13

Patry, Alexandre, and Philippe Langlais. 2005. "Corpus-Based Terminology Extraction." In *Terminology and Content Development – Proceedings of the 7th International Conference on Terminology and Knowledge Engineering*, 313–21. Copenhagen, Denmark.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Machine Learning in Python*, no. 12: 2825–30.

Peñas, Anselmo, Felisa Verdejo, and Julio Gonzalo. 2001. "Corpus-Based Terminology Extraction Applied to Information Access." In *Proceedings of Corpus Linguistics*, 9. Lancaster, UK.

Petrov, Slav, Dipanjan Das, and Ryan McDonald. 2012. "A Universal Part-of-Speech Tagset." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2089–96. Istanbul, Turkey: European Language Resources Association.

Pollak, Senja, Andraž Repar, Matej Martinc, and Vid Podpečan. 2019. "Karst Exploration: Extracting Terms and Definitions from Karst Domain Corpus." In *Proceedings of ELex 2019*, 934–56. Sintra, Portugal.

Qasemizadeh, Behrang, and Siegfried Handschuh. 2014. "Evaluation of Technology Term Recognition with Random Indexing." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 4027–32. Reykjavik, Iceland: European Language Resources Association.

Ramisch, Carlos, Aline Villavicencio, and Christian Boitet. 2010. "Mwetoolkit: A Framework for Multiword Expression Identification." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 662–69. Valetta, Malta: European Language Resources Association.

Rigouts Terryn, Ayla, Patrick Drouin, Véronique Hoste, and Els Lefever. 2019. "Analysing the Impact of Supervised Machine Learning on Automatic Term Extraction: HAMLET vs TermoStat." In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 1012–21. Varna, Bulgaria. https://doi.org/10.26615/978-954-452-056-4_117

Rigouts Terryn, Ayla, 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 *Proceedings of the LREC 2020 6th International Workshop on Computational Terminology (COMPUTERM 2020)*, 85–94. Marseille, France: European Language Resources Association.

Rigouts Terryn, Ayla, Véronique Hoste, and Els Lefever. 2018. "A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, 1803–8. Miyazaki, Japan: European Language Resources Association.

Rigouts Terryn, Ayla, Véronique Hoste, and Els Lefever. 2020. "In No Uncertain Terms: A Dataset for Monolingual and Multilingual Automatic Term Extraction from Comparable Corpora." *Language Resources and Evaluation* 54 (2): 385–418. https://doi.org/10.1007/s10579-019-09453-9

Šajatović, Antonio, Maja Buljan, Jan Šnajder, and Bojana Dalbelo Bašić. 2019. "Evaluating Automatic Term Extraction Methods on Individual Documents." In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 149–54. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5118

Shah, Sapan, S. Sarath, and Reddy Shreedhar. 2019. "Similarity Driven Unsupervised Learning for Materials Science Terminology Extraction." *Computación y Sistemas* 23 (3): 1005–13. https://doi.org/10.13053/cys-23-3-3266

Ville-Ometz, Fabienne, Jean Royauté, and Alain Zasadzinski. 2007. "Enhancing in Automatic Recognition and Extraction of Term Variants with Linguistic Features." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 13 (1): 35–59. https://doi.org/10.1075/term.13.1.03vil

Vintar, Spela. 2010. "Bilingual Term Recognition Revisited." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 16 (2): 141–58. https://doi.org/10.1075/term.16.2.01vin

Vivaldi, Jorge, Luís Màrquez, and Horacio Rodríguez. 2001. "Improving Term Extraction by System Combination Using Boosting." In *Proceedings of the 12th European Conference on Machine Learning (ECML 2001)*, edited by Luc Raedt and Peter Flach, 2167:515–26. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44795-4_44

Vivaldi, Jorge, and Horacio Rodríguez. 2001. "Improving Term Extraction by Combining Different Techniques." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 7 (1): 31–48. https://doi.org/10.1075/term.7.1.04viv

Wang, Rui, Wei Liu, and Chris McDonald. 2016. "Featureless Domain-Specific Term Extraction with Minimal Labelled Data." In *Proceedings of Australasian Language Technology Association Workshop*, 103–12. Melbourne, Australia.

Wolf, Petra, Ulrike Bernardini, Christian Federmann, and Hunsicker Sabine. 2011. "From Statistical Term Extraction to Hybrid Machine Translation." In *Proceedings of the 15th Conference of the European Association for Machine Translation*, edited by Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste, 225–32. Leuven, Belgium: European Association for Machine Translation.

Wolpert, David H. 1996. "The Lack of a Priori Distinctions between Learning Algorithms." *Neural Computation* 8 (7): 1341–90. https://doi.org/10.1162/neco.1996.8.7.1341

Appendix. Features

For more explanation about all features, including the motivation behind them and the references for the statistical features, see Section 4.1.2.

Grp.	Subgrp.	Name & Description	Type	#
SHAP	length	*nr_characters* number of characters (incl. space)	integer (cont.)	1
		nr_tokens number of tokens	integer (cont.)	
	alpha-numeric	*is_alpha* contains only [a-zA-z] and space	integer (binary)	1
		is_alpha_with_dash_or_apostrophe is_alpha + dashes and/or apostrophes	integer (binary)	1
		is_alphanum is_alpha + [0–9]	integer (binary)	1
		is_numeric contains only [0–9] and space	integer (binary)	1

Grp.	Subgrp.	Name & Description	Type	#
		is_non_alphanum contains characters other than [a-zA-Z], [0–9], space, dash, apostrophe	integer (binary)	1
		nr_digits number of [0–9] characters	integer (cont.)	1
		nr_non_letters number of characters other than [a-zA-Z] and space	integer (cont.)	1
	capitalisation	*caps_all_lower_prob* how often the CT occurs completely lowercased	float (perc.)	1
		caps_all_upper_prob how often the CT occurs completely uppercased	float (perc.)	1
		caps_title_case_prob how often the CT occurs completely title-cased	float (perc.)	1
		caps_mwt_first_upper_prob for multi-word CTs: how often the CT occurs with only first letter capitalised	float (perc.)	1
		caps_mixed_case_prob how often the CT occurs with different capitalisation (anything not in other features)	float (perc.)	1
LING	first POS	*POS_simple_first_[POS tag]* simple POS tag of first token as one-hot vector with one feature for each of the simple POS tags	integer (binary)	8
		POS_standard_first_[POS tag] standard POS tag of first token as one-hot vector with one feature for each of the standard POS tags	integer (binary)	26
	last POS	*POS_simple_last_[POS tag]* simple POS tag of last token as one-hot vector with one feature for each of the simple POS tags	integer (binary)	8
		POS_standard_last_[POS tag] standard POS tag of last token as one-hot vector with one feature for each of the standard POS tags	integer (binary)	26
	freq. POS	*POS_simple_freq_[POS tag]*	integer (cont.)	8

Grp.	Subgrp.	Name & Description	Type	#
		frequency with which simple POS tag occurs in CT		
		POS_standard_freq_[POS tag]	integer (cont.)	26
		frequency with which standard POS tag occurs in CT		
	NER	*NER_completely_tagged*	int (binary)	1
		completely tagged as Named Entity LeTs Preprocess		
		NER_not_tagged	int (binary)	1
		not at all tagged as Named Entity LeTs Preprocess		
		NER_partially_tagged	int (binary)	1
		partially tagged as Named Entity by LeTs Preprocess		
	CHUNK	*chunk_contains_ADVP*	int (binary)	1
		LeTs Preprocess chunking assigned ADVP tag to one or more of the CT's tokens		
		chunk_contains_AP	int (binary)	1
		LeTs Preprocess chunking assigned AP tag to one or more of the CT's tokens		
		chunk_contains_NP	int (binary)	1
		LeTs Preprocess chunking assigned NP tag to one or more of the CT's tokens		
		chunk_contains_PP	int (binary)	1
		LeTs Preprocess chunking assigned PP tag to one or more of the CT's tokens		
		chunk_contains_VP	int (binary)	1
		LeTs Preprocess chunking assigned VP tag to one or more of the CT's tokens		
		chunk_contains_O	int (binary)	1
		LeTs Preprocess chunking assigned O (outside) tag to one or more of the CT's tokens		
		chunk_ends_with_I	int (binary)	1
		LeTs Preprocess chunking assigned I (inside) tag to the final token of the CT		
		chunk_starts_with_B	int (binary)	1

Grp.	Subgrp.	Name & Description	Type	#
		LeTs Preprocess chunking assigned B (beginning) tag to the first of the CT's tokens		
	Stopwords	*stopword_completely* CT is completely composed of stopwords	int (binary)	1
		stopword_none CT does not contain any stopwords	int (binary)	1
		stopword_partial CT contains stopword(s), but also other tokens	int (binary)	1
FREQ	spec. freq.	*freq_in_specialised_corpus* relative frequency in specialised corpus	float [0–1]	1
		freq(doc)_in_specialised_corpus relative document frequency in specialised corpus	float [0–1]	1
	ref. freq.	*freq_in_reference_corpus_news* relative frequency in news reference corpus	float [0–1]	1
		freq(doc)_in_reference_corpus_news relative document frequency in news reference corpus	float [0–1]	1
		freq_in_reference_corpus_wiki relative frequency in Wikipedia reference corpus	float [0–1]	1
		freq(doc)_in_reference_corpus_wiki relative document frequency in Wikipedia reference corpus	float [0–1]	1
STAT	stats without ref.	*tfidf* TF-IDF scores of CT in specialised corpus	float [0–1]	1
		cvalue C-value score of CT in specialised corpus	float [0–1]	1
		basic Basic score of CT in specialised corpus	float [0–1]	1
		lexical_cohesion Lexical Cohesion of CT in specialised corpus	float [0–1]	1
	stats with ref. (basic)	*domain_pertinence_news/wiki*	float [0–1]	2

Grp.	Subgrp.	Name & Description	Type	#
		Domain Pertinence scores compared to news and Wikipedia reference corpora		
		domain_relevance_news/wiki	float	2
		Domain Relevance scores compared to news and Wikipedia reference corpora	[0–1]	
		weirdness_news/wiki	float	2
		Weirdness scores compared to news and Wikipedia reference corpora	[0–1]	
		relevance_news/wiki	float	2
		Relevance scores compared to news and Wikipedia reference corpora	[0–1]	
		llr_news/wiki	float	2
		Log-Likelihood Ratio scores compared to news and Wikipedia reference corpora	[0–1]	
	stats with ref. (advanced)	*domain_specificity_news/wiki*	float	2
		Domain Specificity scores compared to news and Wikipedia reference corpora	[0–1]	
		vintar_news/wiki	float	2
		Vintar's termhood scores compared to news and Wikipedia reference corpora	[0–1]	
CTXT	parentheses	*parentheses_ct_between*	int	1
		CT occurs between parentheses in specialised corpus	(binary)	
		parentheses_ct_open_paranthesis	int	1
		CT occurs followed by open parenth. in spec. corpus	(binary)	
		parentheses_open_ paranthesis _ct	int	1
		CT occurs after open parenth. in spec. corpus	(binary)	
		parentheses_ct_closing_paranthesis	int	1
		CT occurs followed by closing parenth. in spec. corp.	(binary)	
VARI	var. numbers	*variant(X)_nr_possible_variants*	int (cont.)	5

Grp.	Subgrp.	Name & Description	Type	#
		for the CT in its current variant, how many different variations there are for the CT as the other 5 variants		
		variant(X)_rel_freq_in_spec_corp combined relative frequency of all possible variations of the CT as the other 5 variants	float [0–1]	5
	var. stats	*variant(X)_sum_termhood_vintar_news* sum of Vintar's termhood score compared to the news reference corpus for all possible variations of the CT as the other 5 variants	float [0–1]	5
		variant(X)_sum_domain_specificity_wiki sum of Domain Specificity score compared to the Wikipedia reference corpus for all possible variations of the CT as the other 5 variants	float [0–1]	5

Address for correspondence

Ayla Rigouts Terryn
Department of Translation, Interpreting and Communication
Universiteit Gent
Groot-Brittanniëlaan 45
9000 Gent
Belgium
ayla.rigoutsterryn@ugent.be
https://orcid.org/0000-0002-9936-9849

Co-author information

Véronique Hoste
Department of Translation, Interpreting and Communication
Universiteit Gent
veronique.hoste@ugent.be

Els Lefever
Department of Translation, Interpreting and Communication
Universiteit Gent
els.lefever@ugent.be

Publication history

Date received: 6 April 2020
Date accepted: 23 November 2020
Published online: 20 August 2021