Analysing the Impact of Supervised Machine Learning on Automatic Term Extraction: HAMLET vs TermoStat

Ayla Rigouts Terryn*, Patrick Drouin**, Veronique Hoste* and Els Lefever*

*LT³ Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Gent; name.surname@ugent.be
**Observatoire de Linguistique Sens-Texte, Université de Montréal
Succ. Centre-ville, Montréal, QC, H3C 3J7; patrick.drouin@umontreal.ca

Abstract

Traditional approaches to automatic term extraction do not rely on machine learning (ML) and select the top n ranked candidate terms or candidate terms above a certain predefined cut-off point, based on a limited number of linguistic and statistical clues. However, supervised ML approaches are gaining interest. Relatively little is known about the impact of these supervised methodologies; evaluations are often limited to precision, and sometimes recall and f1-scores, without information about the nature of the extracted candidate terms. Therefore, the current paper presents a detailed and elaborate analysis and comparison of a traditional, state-of-the-art system (TermoStat) and a new, supervised ML approach (HAMLET), using the results obtained for the same, manually annotated, Dutch corpus about dressage.

1 Introduction

Automatic term extraction (ATE), also known as automatic term recognition (ATR), has long been an established task within the field of natural language processing. It can be used both in its own right, to automatically obtain a list of candidate terms (cts) from a specialised corpus, or as a preprocessing step for other tasks, such as machine translation (Wolf et al., 2011). The traditional method for ATE is a hybrid approach, combining both linguistic and statistical information. In a first step, linguistic preprocessing is performed and a preliminary list of cts is produced based on part-of-speech (POS) patterns. Next, statistical metrics are applied to measure termhood (to what degree a term is related to the domain) and unithood for multi-word terms (whether the individual tokens combine to form a lexical unit) (Kageura and Umino, 1996). These metrics are used to sort the cts based on their likelihood to be actual terms. To filter the list, one can either determine a cut-off value or select the top n or top n percent of terms. As a final step, manual validation is required.

This has been a standard methodology for some time (Daille, 1994) and is still used by state-of-the-art systems such as TermoStat (Drouin, 2003) and TExSIS (Macken et al., 2013). However, the problem with these methodologies is determining the cut-off point (Lopes and Vieira, 2015) and combining multiple features (e.g., separate measures for termhood and unithood). It has become clear that multiple evidence (i.e. combining multiple features) is highly beneficial for ATE (Dobrov and Loukachevitch, 2011; Loukachevitch, 2012). Supervised machine learning (ML) methodologies are now being used in answer to these problems. By automatically learning an optimal combination of features and cut-off points, many features can be efficiently combined.

One of the biggest hurdles for the progress of ATE technologies has been the data acquisition bottleneck, both for evaluation and now also as training data. Manually annotating terms is a slow and arduous task, with notoriously low inter-annotator agreement due to the ambiguous nature of terms. This lack of agreement on the basic characteristics of terms is also reflected in the different methodologies of various ATE research, e.g., min./max. length and frequency, POS patterns and degree of specialisation. As a result, the supervised methodologies that have been developed are extremely difficult to compare (both to each other and to non-ML systems) and qualitative analyses that go beyond calculating precision (how many of the extracted cts are true terms), recall (how many of the true terms are extracted) and f1-scores (weighted average of precision and recall) are rare.

The construction of a diverse and extensive dataset for ATE (Rigouts Terryn et al., 2019) pro-

Proceedings of Recent Advances in Natural Language Processing, pages 1012–1021,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_117

vided an opportunity to (1) develop a supervised ML approach for ATE (HAMLET) and (2) perform a detailed evaluation of this system compared to a traditional tool without ML: TermoStat (Drouin, 2003). These specific systems were chosen because they both allow extraction of single- and multi-word terms (swts and mwts) and are not restricted to only nouns and noun phrases, but instead also allow verbs, adjectives and adverbs to be extracted. Moreover, their methodology is similar, so the research can focus on one main difference: the fact that HAMLET uses supervised ML to combine different features, rather than relying on manually set filters and thresholds like TermoStat. This is important to better understand the impact of the methodology. Are the same terms found with both methodologies? Do they make similar mistakes? Is it possible to see the impact of the training data? The analysis is performed by a terminologist, in her native language (Dutch) and on a subject for which she is a domain specialist (equitation - dressage).

2 Related Research

Some of the original supervised approaches to ATE start appearing in the early 2000s. Vivaldi and Rodríguez (2001) claim to be the first to combine different methodologies for term extraction into a single system. Based on two manually annotated Spanish corpora in the medical domain, four different strategies are combined. The first strategy is to use EuroWordNet (EWN) (Vossen, 1998) to determine whether a word belongs to the medical domain. Next, Greek and Latin word forms are detected. Context is analysed as well, focusing on prime term candidates, i.e. those that are validated with EWN as medical terms. Finally, three unithood measures help to find relevant multi-word terms. Combining these four techniques leads to better results than using any one of them separately. The system is only tested on the Spanish medical domain; performance may vary significantly depending on EWN coverage of the corpus and relevance of the Latin and Greek words. Later research does test on multiple domains, for instance, an evolutionary algorithm based on the optimisation of the Receiver Operating Characteristics curve for the extraction of mwts (Azé et al., 2005), tested on the domains of biology and HR; or a system for both swts and mwts (Yuan et al., 2017), elaborately evaluated with different algorithms, using undersampling to obtain more balanced data, and cross-domain training/testing on four domains.

In 2016, neural network word embeddings are applied to ATE for the first time (Amjadian et al., 2016), first as a filter on an existing tool (TermoStat), later on also as a full ATE pipeline (Amjadian et al., 2018). The success of multiple features for ATE has been proven repeatedly (Dobrov and Loukachevitch, 2011; Loukachevitch, 2012; Nokel, Michael et al., 2012) and aside from the original binary classification approach of cts, sequence labelling approaches are also gaining interest (Judea et al., 2014; Kucza et al., 2018). Additionally, There has been an increased interest in more nuanced term labels (Ljubei et al., 2019; Hätty and Schulte im Walde, 2018), even though binary classification is still the norm.

Unsupervised and semi-supervised approaches are starting to appear as well, which is interesting considering the time and effort associated with constructing good gold standard data. Judea, Schütze and Brügmann (2014) use the specific layout of patents to generate training data. Cts are extracted based on their POS pattern and filtered with an elaborate stopword list. When these cts were preceded by a figure reference in patents, 95% of them were true terms. Since these terms could be identified with high precision, they were used as training data to detect other terms without figure references. Another strategy is fault-tolerant learning, which has been used for Chinese ATE (Yang et al., 2011). Two sets of seed terms are extracted from the same, unlabelled dataset, with two different termhood metrics methods. By comparing the results of the two classifiers and retraining on only the best results (for n iterations), a system can be trained without any labelled training data. Human annotation is only used for evaluation, where an approximation of precision is calculated by randomly sampling and annotating 10% of the extracted cts. Patry and Langlais (2005) take an unusual approach regarding the difficulty of obtaining data and ask users to provide an annotated corpus. This added effort on the part of the user would be rewarded in the form of a customised tool, considering the user's own definition of the ambiguous concept of a term. They also cite two of the most common problems for ATE: the lack of a common benchmark for evaluation and the difficulty extracting hapax terms, especially con-

sidering that these make up 75% of the terms in their test corpus.

Despite the increasing research interest, research on the impact of ML approaches on ATE is limited (Amjadian et al., 2018; Nokel, Michael et al., 2012). Comparative evaluations are highly problematic for several reasons. First, established benchmarks such as the GENIA corpus (Kim et al., 2003) and the ACL RD-TEC (Qasemizadeh and Schumann, 2016) are rare and often only available in a single language and domain. Second, reported evaluation scores (usually precision, recall and f1-score) differ greatly depending on the strictness of the evaluation (e.g., whether or not partial matches are approved). Third, the difficulty of the task varies considerably depending on the ct selection. For instance, limiting POS patterns and frequency thresholds can result in a more balanced data set and narrower search space. Finally, results are rarely discussed beyond reporting the scores, which may result in a distorted image, given the ambiguous nature of terms, as will be discussed further on. Therefore, while researchers regularly mention the suspected impact of methodology, term definitions, language and domain, little is known about how these factors influence the actual results. The research presented in this paper presents an elaborate and qualitative evaluation and comparison of two tools and will focus on the difference between a supervised ML approach and a traditional approach.

3 Data and Tools

3.1 Data

The dataset is described in detail in (Rigouts Terryn et al., 2019). The Dutch corpus on dressage was chosen as the evaluation corpus. The annotation scheme is based on lexicon-specificity (whether a term belongs to general language or only the vocabulary of experts) and domain-specificity (how relevant the term is to the given domain). Terms are annotated with three different labels: Specific Terms (which are both domain-specific and lexicon-specific), Common Terms (which are domain-specific but not lexicon-specific) and Out-Of-Domain (OOD) Terms (which are not domain specific but are lexicon-specific). Named Entities are annotated as well. In this corpus of around 55k tokens (64 documents), this resulted in 1326 different manual annotations (excluding Split Terms).

3.2 TermoStat

TermoStat is a hybrid term extractor developed by Drouin (2003) which is still continuously updated. It is currently available in French, English, Spanish, Italian, and Portuguese, with beta versions for German, Catalan, Korean, Chinese and Dutch. It is customisable in the sense that users can choose to extract swts, mwts, or both and can also select which POS (nouns, adjectives, adverbs and/or verbs) should be extracted. TermoStat selects cts based on their POS pattern and filters and sorts these cts with the Specificity score, a measure that takes into account the relative frequency of a ct in the specialised corpus, compared to that in a general reference corpus to calculate termhood.

3.3 HAMLET

HAMLET stands for Hybrid Adaptable Machine Learning approach to Extract Terminology and is a supervised methodology for ATE based on the data described in (Rigouts Terryn et al., 2019). HAMLET's architecture is inspired by traditional hybrid systems such as TermoStat. First, cts are extracted based on their POS pattern. However, rather than a predefined list, the patterns are obtained from the annotated training corpus. Since there were no restrictions on which POS could be annotated, this results in an extensive list. Moreover, incorrect patterns due to POS-tagging errors are included as well. This may result in a lot of noise but could also increase recall if similar tagging mistakes are made on terms in the test corpus.

Next, a series of features are calculated for each ct. There are six different feature groups: morphological/shape (e.g., term length, capitalisation, special characters), frequency (e.g., relative frequencies in specialised corpus, newspaper corpus and Wikipedia corpus), statistical (e.g., various termhood and unithood measures), related cts (e.g., information about terms with same lemma or normalised form), linguistic (e.g., POS pattern) and corpus features (e.g., domain of corpus of origin). There are 152 distinct features in total. In contrast to most other term extractors, no restrictions are placed on term length or frequency.

This information is fed to a binary decision tree classifier in Scikit-learn (Pedregosa et al., 2011). Hyperparameter optimisation with grid search is performed in 5 folds on the training data. All values are scaled to a value between 0 and 1. For the experiment discussed in the current contribu-

tion, HAMLET was trained on the Dutch corpora about heart failure and wind energy and tested on the Dutch corpus about dressage. Irrelevant features (with the same value for all instances) are discarded, leaving 136 features in this case. The data is highly imbalanced, with fewer than 10% positive instances (similar distribution in train and test sets). While other algorithms were able to reach better scores (e.g., a random forest classifier obtained an f1-score of 61% on the same dataset), only the decision tree model is discussed, because it offers both decent performance and is easy to interpret. Future research will devote more attention to the differences between algorithms for this task.

4 Experiments and Comparisons

4.1 Candidate Terms and Part-of-Speech Patterns

The gold standard data (test data) contains 1326 unique annotations: 985 Specific Terms, 190 Common Terms, 45 OOD Terms and 106 Named Entities. For this experiment, HAMLET was trained to find all annotation types, which is the configuration that lead to the best results for TermoStat. However, HAMLET could also be trained on specific combinations of these labels to customise the results for different applications. Out of the 1326 annotations which were considered true terms, only two could not be found because the annotations were made below token-level and were therefore never selected as a ct by either extractor: *promotie* (promotion, i.e. moving to a higher level of competition) and *k* (one of the letters indicating a certain position in the riding arena). Another portion could not be found due to their POS pattern. This is always a problem for non-ML extractors, since it is nearly impossible to manually define all possible patterns, especially considering that POS taggers can make mistakes. However, the supervised system has similar troubles. HAMLET's preprocessing can only select terms for which the POS pattern occurred in the training corpora (the two Dutch corpora on heart failure and wind energy). In this case, there are 216 different patterns in the training data, but the test corpus still contains terms with 63 patterns that are not in the training data. This illustrates how domain-specific terminology can be. Dressage terminology contains many terms that start with a preposition. For instance, there are 85 annotations of the preposition+determiner+noun pattern, e.g., *aan het been* (responsive to a rider's leg aids). Patterns including verbs are common in dressage as well, e.g., *vierkant halthouden* (stopping the horse so all four hoofs form a rectangle). Due to the absence of such patterns in the training data, 104 terms were not extracted by HAMLET, while 11 of these were found by TermoStat.

Across all 3 languages and 4 domains in the complete dataset, a total of 1345 distinct POS patterns are identified (419 in Dutch in all four domains), meaning that these types of errors are greatly reduced when HAMLET is trained on a larger portion of the data, though that also leads to more noise. This emphasises the importance of diverse datasets to train robust term extractors and to evaluate extractors in multiple domains.

4.2 Decision Tree

The decision tree (of depth 8) that was created based on the training data of Dutch corpora on heart failure and wind energy uses 64 out of the 152 distinct features. All feature categories are represented, except corpus features. In other experiments involving more domains and languages, corpus features are regularly used, but in this setting, with only two different domains in the training data, they did not appear to be informative. Statistical features are used most often (66 nodes, using 17 distinct features), followed by linguistic features (35 nodes, 16 features), related ct features (28 nodes, 9 features), morphological/shape features (25 nodes, 9 features), and frequency features (16 nodes, 11 features).

The most discriminating feature (first node in the decision tree) is Vintar's termhood score (Vintar, 2010), calculated for the original, unlemmatised ct, compared to a reference corpus of newspaper articles. This is also the feature that, following domain consensus, is used most often (10 times and 8 times, respectively). The most frequently used features in the other categories are: number of characters (morphological/shape feature used 7 times), number of cts that contain the current ct (related feature used 6 times), the presence of either a preposition or a noun (linguistic features, both used 4 times). The frequency features are all used 0-2 times and none stand out. A possible explanation for the comparative irrelevance of frequency features, is that frequency is most informative when already incorporated into termhood or unithood measures and that many fre-

quency features are strongly correlated.

A feature indicating presence in a list of stopwords is not used, even though lists of stopwords are generally very useful for ATE. This may be related to the limited list used for Dutch (414 tokens) or the way it is currently implemented (only complete matches are counted). This analysis shows how the statistical termhood and unithood features are indeed most useful for ATE, but that there are many other informative features as well, in a range of different categories.

4.3 Precision, Recall and F1-scores

HAMLET extracts 1352 cts with a precision of 55.03%, a recall of 56.11% and an f1-score of 55.56%. TermoStat extracts many more cts (4671) and has a much lower precision of only 18.18% but a higher recall at 64.03%, resulting in an f1-score of 28.31%. This is where the supervised ML component becomes immediately apparent: HAMLET is trained to optimise f1-score, whereas the cut-off point for TermoStat had to be set manually based on a limited set of experiments. Figures 1 and 2 show the precision, recall and f1-score curves for HAMLET and TermoStat. In this case, HAMLET did not print the predicted label of the classifier, but the predicted probability of label 1, i.e. the predicted probability that the ct is a true term. In Figure 1, only terms with a probability higher than 50% (up until rank 1352) were labelled as terms by HAMLET. However, for the sake of comparison with TermoStat, the graph was calculated supposing that all 4671 highest ranked cts were predicted as terms. As can be seen in the graph, the decision boundary is very close to the highest possible f1-score. According to this ranking, that would have been 57.05%, if HAMLET had extracted the highest ranked 1619 cts instead of the first 1352. The TermoStat results in Figure 2 show a different trend. Here, the ideal cut-off point would have been after the 1307th highest ranked term (Specificity of 16.06), which would have resulted in an f1-score of 42.61%. Instead, 3362 more terms were extracted, causing a large drop in f1-score.

Another notable peculiarity in these curves is that the TermoStat curves are smoother and follow a more predictable pattern: precision starts high and decreases gradually, recall increases but starts to slowly flatten out. The recall curve for HAMLET follows this pattern and even reaches over 80% at rank 4671, where TermoStat's recall is still only at 64%. However, HAMLET's precision curve is far from smooth in the beginning, with the highest precision only around rank 285. These fluctuations are due to two factors. First, precision curves are very susceptible to small changes at the start, when it is calculated for few examples. Second, surprisingly, HAMLET's predicted probability that a ct is a true term does not always correspond with the reality. For instance, 13 cts were given a 100% probability and only 7 of these were actual terms. So, while the predicted true term probability for these cts was 100%, the actual precision was only 54%. One of the false positives should have been in the gold standard and was missed by the annotators. Two were parts of terms and the remaining three were very common words: *bovenstaande* (above), *moet* (has to), and *werd* (became). Further research is needed to explain this behaviour and compare results with other algorithms and corpora.

4.4 Term Labels

While the extractors only performed binary classification, the gold standard does contain more detailed labels (Specific Terms, Common Terms, OOD Terms and Named Entities, see section 3.1). It was already established that TermoStat extracts many more terms, resulting in a lower precision but also a higher recall. An additional analysis can show whether both tools extract the same term types based on the more fine-grained labels. Regarding these labels, two hypotheses were formulated. First, we expect HAMLET to be better than TermoStat at extracting Named Entities and maybe also OOD Terms, since these were included in the training data, while TermoStat's Specificity score is designed mostly to detect domain-specific terms, i.e. Specific and Common Terms. TermoStat may still extract Named Entities and OOD Terms, since they share many characteristics with the other two categories, but the hypothesis is that it will extract comparatively fewer than HAMLET. This hypothesis was partly confirmed by the results. Even though HAMLET extracts fewer terms in total, it extracts more Named Entities than TermoStat (63 versus 43) and a larger percentage of all HAMLET's extractions are Named Entities (5% versus 1%). For OOD Terms the hypothesis could not be confirmed. since the difference was too small. This may be due, at least

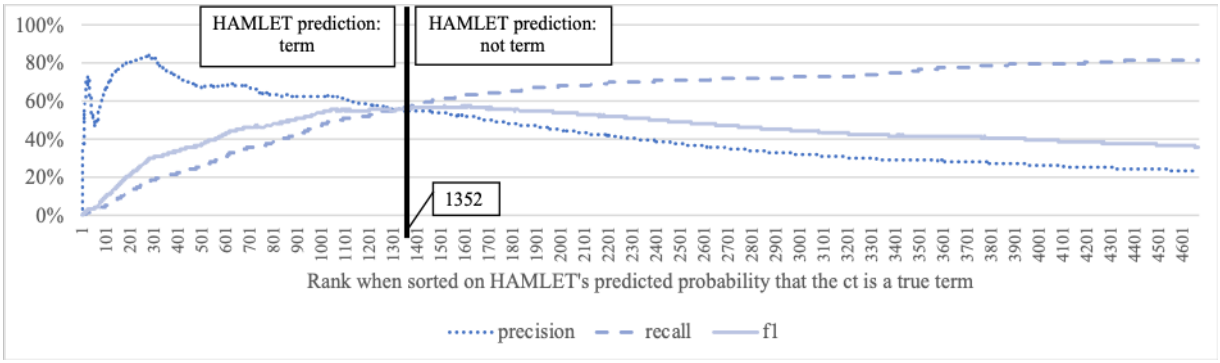

Figure 1: Precision, recall and f1-curves of HAMLET, including the 4671 highest ranked cts based on predicted probability that the ct is a true term; black line indicates boundary between cts that were predicted to be terms and those that were predicted not to be terms.

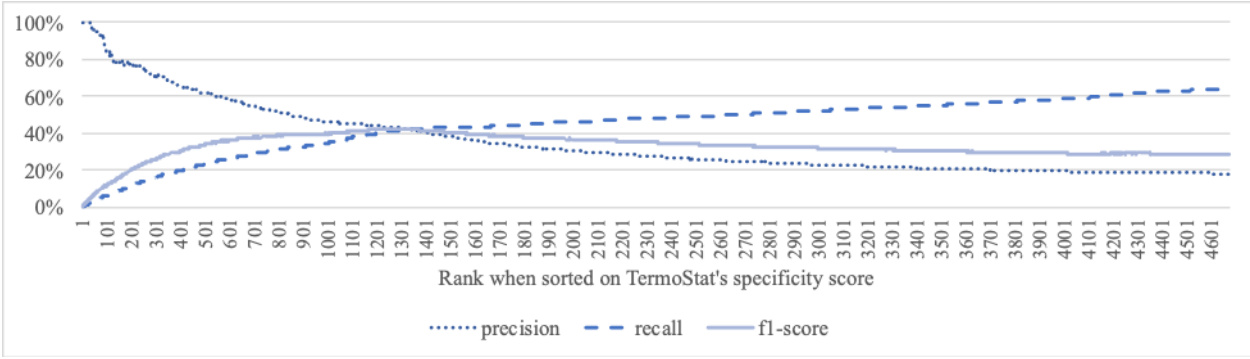

Figure 2: Precision, recall and f1-score curves of TermoStat for all 4669 extracted terms, ranked based on specificity score

in part, to the annotation. Since the corpus subject was dressage (a subdomain of equitation), rather than equitation as a whole, many terms that are specific to other branches of equitation were annotated as OOD Terms. These are nearly all terms related to other equitation disciplines, such as *gymkhana* (same in English) or *voltige* (equestrian vaulting). Had the annotation been slightly less strict about the domain-specificity, at least 27 of the 34 OOD Term annotations would have been Specific Terms. This illustrates how a subjective decision about whether or not to include a certain group of terms, can have a large impact on the results.

The second hypothesis concerns Specific Terms: we expect HAMLET to outperform TermoStat for Specific Terms. TermoStat relies heavily on a single termhood measure, which means it has the typical drawback of being very sensitive to frequency, leading to low recall on rare terms. HAMLET combines many more features, which may mean that it is less sensitive to frequency. This is important for Specific Terms, since they are often rare. The average relative frequency of Specific Terms versus Common Terms in the domain-specific corpus, calculated by HAMLET is 0.0001268 versus 0.0003642 (similar for document frequency). Again, the hypothesis could only partially be confirmed. HAMLET extracts fewer Specific terms than TermoStat (540 versus 626), though this is similar when considering the comparative difference in total number of extracted terms. However, HAMLET does extract more hapax terms (291 versus 241 by TermoStat), despite extracting fewer terms in total, confirming the part of the hypothesis about HAMLET's improved ability to extract rare terms.

4.5 Agreement between HAMLET and TermoStat

The agreement between HAMLET and TermoStat is very low, with a Cohen's Kappa score of only 0.162. Part of the disagreement is due to the much higher number of non-terms extracted by TermoStat, but even agreement on true terms is only

slightly more elevated (0.28). These numbers indicate that the two tools have different strengths and weaknesses. In previous sections, two main strengths of the supervised approach were already discussed: it is better at optimising for f1-score and it is better at extracting rare terms. The variety of features also visibly results in other improvements. For instance, there is a feature indicating the presence of a dash at the end of a ct. HAMLET has incorporated this feature into the decision tree and extracts only 3 wrong cts that begin or end with a dash. This is not included in TermoStat's preprocessing, resulting in 43 wrong extractions. This does not always explain the results, as illustrated by the fact that the feature indicating the presence of digits in a ct is never used, but HAMLET still correctly extracts 10 out of 21 gold standard terms with digits, whereas TermoStat does not recognise any. Another notable result is that TermoStat extracts 90 cts that begin or end with an article (compared to only 5 such error extracted by HAMLET). These types of mistakes were actually expected from HAMLET, rather than TermoStat, since HAMLET selects cts based on a list of POS patterns that is not manually validated and includes wrong patterns. The fact that only TermoStat makes this error, indicates that the former may have learnt to exclude such cts, while the latter may include wrong patterns, due to its susceptibility to human error. Another possibility is that the POS tagger used by TermoStat is less accurate, resulting in more such errors.

There are also disadvantages to the supervised method, specifically due to the differences between training and test data. For instance, single letters, indicating certain positions in a dressage arena, can be terms. This is not be the case in most other domains, so a supervised system may learn rules that obstruct the extraction of single-character terms. HAMLET only extracts 3 out of 10 single-character terms in the gold standard, while TermoStat extracts 6. This is an illustration of how domain-dependent term characteristics can be and how this could impact supervised systems. Furthermore, HAMLET's lower sensitivity to frequency is not only an advantage but can also backfire. A few seemingly obvious terms with very high frequencies are not extracted, e.g., *hulpen* (aids) and *hand* (meaning both literally hand, but also the direction the horse is going in the arena). Even *paarden* (horses) received a 0% probability of being a term by HAMLET. A final category of terms both extractors struggle with, are those that are also part of general language and only become terms in this context. An example is *pijp*, which usually means pipe, but, in the context of dressage, refers to a part of a horse's leg. At least half of the terms that were not found by either tool concern terms that are also part of general language.

	H = 1	**H = 0**	**SUM**
TS = 1	852	3819	4671
TS = 0	500	9360	9860
SUM	1352	13179	14530

Table 1: Agreement between TermoStat (TS) and HAMLET (H); =0.162

The described differences illustrate various strengths and weaknesses of both approaches and inspires a few suggestions for improvement. TermoStat's approach could benefit from more elaborate preprocessing (e.g., removing cts ending in a dash) and an evaluation of the POS patterns. The supervised approach is clearly influenced by the domain-dependence of term characteristics and could benefit from in-domain training data or training data in more domains. The two approaches are at least partly complementary and a combination of the output results in a recall of 77.45%, which is high, considering the strictness of the evaluation and the gold standard.

4.6 Agreement Between Tools and Gold Standard

Even though the gold standard was rigorously annotated, there is always the possibility of human error and the ambiguous nature of terms, which means that these annotations are not the only possible correct annotations. Therefore, it is worth looking at the ATE results in more detail. Are there any terms that should have, or could have been annotated among the false positives? Or the opposite: terms which could or should not have been annotated among the false negatives? Are the mistakes made by the tools understandable or undeniably wrong? In an attempt to answer these questions, HAMLET's results were analysed in more detail.

Only a single annotation was found to be undeniably wrong: *veel* (many) was mistakenly annotated as a term. However, there were 76 others which were labelled: should (not) have been

annotated, including terms such as *uitzwaaien* (wrong positioning of the horse's hindquarters, mostly during a turn), and *verruim* (specific way of lengthening the horse's stride; other forms of this verb were annotated correctly). Looking at the 608 false positives, at least 217 of them could have been terms, which would increase precision to over 70%. It implies that there is at least some logic in the errors and that, overall, HAMLET does appear to have learnt informative general characteristics of terms. However, this analysis should also be interpreted as a cautionary tale regarding ATE evaluation. When evaluating a list of already extracted cts, annotators are biased to evaluate favourably. Therefore, results compared to a predetermined gold standard may tend to be worse than results based on the annotation of the ATE output. Any comparisons between such results should be interpreted with due caution.

Nevertheless, it is encouraging to see logical patterns in the ATE results. For instance, many cts were extracted related to body parts (of both horse and rider). These were not always consistently annotated but are still logical terms in the field of dressage, which is a sport where the positioning of horse and rider are crucial. At least 171 cts extracted by HAMLET were related to body parts or bodily functions (e.g., *schuimproductie*, the production of foam in the horse's mouth). They were actually extracted by HAMLET more consistently than they were labelled by the human annotator. Also encouraging was the fact that, despite only very limited information about term variation, HAMLET often makes the same decision for related terms, such as terms with different full forms sharing the same lemma. Still, TermoStat's strategy of grouping terms with the same lemma is more effective and should be considered as an option to improve HAMLET.

One last item to mention here is that HAMLET is still susceptible to classic ATE errors, such as wrongly extracting parts of terms, combinations of different terms, or very frequent terms in combination with a non-term. For instance, 35 false positives contain the word *paard* or *paarden* (horse(s)), but in combinations that are not terms, e.g., *paard gaat* (horse goes), *paard niet* (horse not), and *paard symmetrisch* (horse symmetrical). These are typical errors because such combinations are much more frequent in the domain-specific corpus than in reference corpora, so they get high termhood values. Even though HAMLET still makes these mistakes, there is a marked improvement compared to TermoStat, which relies more heavily on termhood statistics. For instance, TermoStat wrongly extracts 320 cts that contain *paard(en)*, compared to only 35 for HAMLET. This further illustrates the positive effect of multiple features to limit frequency-related errors.

5 Conclusions and Future Research

The research described in this paper presents an elaborate evaluation of a supervised ML approach to automatic term extraction (HAMLET), compared to a traditional system without training data (TermoStat). As expected, the supervised system obtains higher f1-scores by combining features with various types of information and optimising f1-score. A closer look at the results confirms that the system has clearly learnt informative general characteristics of terms. It is less reliant on frequency, leading to fewer mistakes on rare terms or frequent non-terms. However, the supervised system also has a distinct weakness, namely its domain-dependence, since it was trained on out-of-domain data. This emphasises the need for annotated data, though there are also indications that very little training data could suffice (Amjadian et al., 2018). Nevertheless, annotated data remains critical for a nuanced evaluation.

Current versions of HAMLET can already obtain an average f1-score of 53%, using cross-validation on all domains and languages combined. Preliminary results already show the impact of factors such as algorithm, language, domain, term definition, and in-domain training data, with f1-scores of up to 66% depending on the combination. Precision and recall are not always as balanced as for the presented use-case, and results vary greatly per corpus. Future research will concentrate on further exploring the robustness of HAMLET, with more contrasting results for different configurations and data. Aside from the binary classifier, a sequence labelling approach, which is further removed from the original methodology, will also be explored and will provide further material for comparison.

6 Acknowledgements

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

References

Ehsan Amjadian, Diana Inkpen, T.Sima Paribakht, and Farahnaz Faez. 2016. Local-Global Vectors to Improve Unigram Terminology Extraction. In *Proceedings of the 5th International Workshop on Computational Terminology*. Osaka, Japan, pages 2–11.

Ehsan Amjadian, Diana Zaiu Inkpen, T. Sima Paribakht, and Farahnaz Faez. 2018. Distributed specificity for automatic terminology extraction. *Terminology* 24(1):23–40. https://doi.org/10.1075/term.00012.amj.

Jérôme Azé, 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*. Brest, France, pages 209–2019. ArXiv: cs/0512050. http://arxiv.org/abs/cs/0512050.

Béatrice Daille. 1994. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In J. Klavans and P. Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Aproaches to Language*, MIT Press, Massachusetts, pages 49–66.

Boris Dobrov and Natalia Loukachevitch. 2011. Multiple evidence for term extraction in broad domains. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Hissar, Bulgaria, pages 710–715.

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

Anna Hätty 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)*. Sante Fe, New Mexico, USA, pages 62–73.

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

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

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics* 19(1):180–182.

Maren Kucza, Jan Niehues, Thomas Zenkel, Alex Waibel, and Sebastian Stüker. 2018. Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks. In *Interspeech 2018*. ISCA, Hyderabad, India, pages 2072–2076. https://doi.org/10.21437/Interspeech.2018-2017.

Nikola Ljubei, Darja Fier, and Toma Erjavec. 2019. KAS-term: Extracting Slovene Terms from Doctoral Theses via Supervised Machine Learning. *arXiv:1906.02053 [cs]* ArXiv: 1906.02053. http://arxiv.org/abs/1906.02053.

Lucelene Lopes and Renata Vieira. 2015. Evaluation of cutoff policies for term extraction. *Journal of the Brazilian Computer Society* 21(1). https://doi.org/10.1186/s13173-015-0025-0.

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

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

Nokel, Michael, Bolshakova, E.i., and Loukachevitch, Natalia. 2012. Combining multiple features for single-word term extraction. In *Proceedings of Dialog 2012*. pages 490–501.

Alexandre Patry and Philippe Langlais. 2005. Corpus-Based Terminology Extraction. In *Terminology and Content Development - Proceedings of the 7th International Conference on Terminology and Knowledge Engineering*. Copenhagen, Denmark, pages 313–321.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. 2011. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON* (12):2825–2830.

Behrang Qasemizadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods. In *Proceedings of LREC 2016*. ELRA, Portoro, Slovenia, pages 1862–1868.

Ayla Rigouts Terryn, Véronique Hoste, Joost Buysschaert, Robert Vander Stichele, Elise Van Campen, and Els Lefever. 2019. Validating multilingual hybrid automatic term extraction for search engine optimisation: the use case of EBM-GUIDELINES. *Argentinian Journal of Applied Linguistics* 7(1):93–108.

Jorge Vivaldi and Horacio Rodríguez. 2001. Improving term extraction by combining different techniques. *Terminology* 7(1):31–48. https://doi.org/10.1075/term.7.1.04viv.

Piek Vossen, editor. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-94-017-1491-4.

Petra Wolf, Ulrike Bernardini, Christian Federmann, and Hunsicker Sabine. 2011. From statistical term extraction to hybrid machine translation. In Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste, editors, *Proceedings of the 15th Conference of the European Association for Machine Translation*. Leuven, Belgium, pages 225–232.

Yuhang Yang, Hao Yu, Yao Meng, Yingliang Lu, and Yingju Xia. 2011. Fault-Tolerant Learning for Term Extraction. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*. Sendai, Japan, pages 321–330.

Yu Yuan, Jie Gao, and Yue Zhang. 2017. Supervised Learning for Robust Term Extraction. In *The proceedings of 2017 International Conference on Asian Language Processing (IALP)*. IEEE.