Multilingual Term Extraction from Comparable Corpora: Informativeness of Monolingual Term Extraction Features

Kim Steyaert and Ayla Rigouts Terryn

LT^3 Language and Translation Technology Team
Ghent University, Groot-Brittanniëlaan 45, 9000 Gent;
ayla.rigoutsterryn@ugent.be; kim.steyaert@gmail.com

Abstract

Most research on bilingual automatic term extraction (ATE) from comparable corpora focuses on both components of the task separately, i.e. monolingual automatic term extraction and finding equivalent pairs cross-lingually. The latter usually relies on context vectors and is notoriously inaccurate for infrequent terms. The aim of this pilot study is to investigate whether using information gathered for the former might be beneficial for the cross-lingual linking as well, thereby illustrating the potential of a more holistic approach to ATE from comparable corpora with re-use of information across the components. To test this hypothesis, an existing dataset was expanded, which covers three languages and four domains. A supervised binary classifier is shown to achieve robust performance, with stable results across languages and domains.

1 Introduction

Bilingual automatic term extraction (ATE) from comparable corpora aims to identify equivalent term pairs cross-lingually in monolingual corpora that are similar in terms of size, topic and style. Strongly related to bilingual lexicon induction (BLI), the main difference is that ATE from comparable corpora focuses on terminology (domain-specific, specialised vocabulary), rather than general language. Despite the difficulty of finding cross-lingual equivalents in unaligned text, comparable corpora have a substantial added value over parallel corpora, since they are much easier to create and, therefore, less expensive. This is especially relevant for low-resourced languages and specialised domains and has made both ATE and BLI popular research topics over the past years.

The most successful strategies for finding cross-lingual equivalents rely on the distributional hypothesis or compositionality (see related research). The hypothesis of this project is that information from the monolingual term extraction phase (e.g., frequency, termhood and unithood statistics, part-of-speech (POS)) could be re-used as additional clues for finding equivalents cross-lingually. While it is not expected that these features alone will suffice to find cross-lingual equivalents, they might provide complementary information using features that have already been calculated for the monolingual ATE and could also help counter disadvantages of current approaches, such as the dependence on huge corpora. This pilot study was set up to test the potential informativeness of features from monolingual ATE to recognise term pairs cross-lingually in comparable corpora. First, an existing dataset for ATE (Rigouts Terryn et al., 2019b) was expanded with more cross-lingual annotations. Subsequently, a supervised binary classifier was constructed using only the features designed for monolingual term extraction. Further analyses of the classifier and the features illustrate how this information might complement the more established features.

2 Related Research

ATE from comparable corpora and BLI have received much research interest and certain trends have emerged. The distributional hypothesis appears to be the most popular approach for finding cross-lingual equivalents. This hypothesis states that equivalent lexical units will appear in similar contexts. The contexts of potential equivalents are compared by using some form of word vector representations. This can be done through the so-called standard approach (or a variation thereof).

In this case, context vectors are created for an n-word window around the lexical unit, subsequently, these vectors are normalised (e.g., using mutual information) and a bilingual seed lexicon is used to project between the source and target language. Once it is possible to map between the two vector spaces, a similarity measure (e.g., cosine similarity) can be calculated to measure the context similarity (Liu et al., 2018).

A second approach based on the distributional hypothesis consists in using neural networks to obtain word embeddings. An example is presented in (Hazem and Morin, 2018), where embeddings from the specialised corpus are combined with those from a larger, general corpus. Apart from word embeddings, they also experiment with character n-gram embeddings, which take into account the internal structure of words. This is another popular strategy, especially for morphologically-rich languages and the medical domain (Heyman et al., 2018; Hakami and Bollegala, 2017; Bollegala et al., 2015; Kontonatsios et al., 2013; Hazem and Morin, 2018). Character n-grams have repeatedly been shown to outperform word embeddings, or at least to be useful in combination with them. Since the previously described methods are mainly applied to single-word terms, there is another strategy specifically for multi-word terms, which is the compositionality approach, whereby each part of a multi-word term is translated separately to map it to potential equivalents. Such methods are highly reliant on bilingual dictionaries. Other common features are string similarity measures (Pinnis et al., 2019), Wikipedia-based features (Jakubina and Langlais, 2016, 2018), and temporal clues, burstiness and frequency (Irvine and Callison-Burch, 2013).

Some of the most commonly cited problems with current methodologies for ATE from comparable corpora are that they are dependent on very large resources for the context vectors, which is a big disadvantage for a task that has the specific goal of making bilingual lexicon building less reliant on expensive resources. While the increasing availability of large-scale, multilingual pre-trained language models (e.g., BERT (Devlin et al., 2019)) can be very helpful for BLI in general, it is less well-suited for multilingual ATE, since terminology is both less frequent and more domain-dependent than general language. Therefore, the specific characteristics of terms may not necessarily be captured well in these general language models, especially for those terms that also occur in general language but acquire a different meaning as a term in a specialised domain. A second, related disadvantage of most current approaches for ATE from comparable corpora is that they score badly for infrequent terms, even when ”infrequent” is broadly interpreted as having a frequency of up to 25 (Jakubina and Langlais, 2018). Other disadvantages are reliance on existing resources, such as bilingual seed lexicons or Wikipedia, separate methodologies for single- and multi-word terms (Liu et al., 2018) and, in the case of string similarity, dependence on similarities between source and target language. A final remark in this regard, is that it is very difficult to compare reported results. This is partly because of differences in methodology (e.g., entire ATE from comparable corpora pipeline or only classifying term pairs, focus on single- or multi-word terms, etc.) and evaluation measures (precision@rank, mean average precision and f1-score being the most common). Another important reason is the ambiguous nature of the task: determining whether two terms are equivalent is by no means straightforward. This can range from technical questions such as whether terms with an almost identical meaning, but from a different word class are considered correct, to more theoretical problems regarding the nature of equivalence (Le Serrec, 2012).

3 Data

For previous research on monolingual ATE (Rigouts Terryn et al., 2019b), three comparable corpora had been created in the domains of dressage, wind energy and heart failure, as well as one parallel corpus in the domain of corruption. All four corpora were constructed with English, French and Dutch texts. Around 50k tokens were manually annotated per domain/language, resulting in over 100k annotations of single- and multi-word terms and Named Entities (NEs). Cross-lingual annotations had already been added for the complete corpus on heart failure. A similar methodology was adopted to annotate cross-lingual equivalents in the other domains as well, although the annotations are less elaborate than those for the corpus on heart failure, which includes annotations of synonyms, abbreviations, alternative spellings, lemmas, hypernyms, hyponyms and other strongly related terms, in addition to cross-lingual equivalents. The annotations that were added for the other domains only concern cross-lingual equivalents. Moreover, the added annotations do not cover the entire corpora (as the original ones did), but they are sufficient to allow various cross-domain comparisons. The annotation work resulted in a total of over 3.5k validated term pairs per language pair (see Table 1). With the current methodology (see section 5), the order of the languages (English-French, English-Dutch, and French-Dutch) is irrelevant, so it could be reversed without influencing either the numbers or the results.

Domain	**EN-FR**	**EN-NL**	**FR-NL**
corruption	358	401	397
dressage	402	407	525
heart failure	2362	2467	2611
wind energy	425	598	389
Total	**3547**	**3873**	**3892**

Table 1: Number of positively validated term pairs per corpus

While the ultimate goal is to develop an entire pipeline for ATE from comparable corpora, from monolingual term extraction to bilingual term linking, the aim of the current pilot study was to test the potential of re-using information from the former for the latter. For this purpose, the previously mentioned annotations were transformed into datasets with positive (equivalent) and negative (non-equivalent) term pairs, which could be used as input for a binary classifier. All positive term pairs were manually annotated as valid equivalents and random sampling was used for negative examples, a methodology adopted in previous research as well (Kontonatsios et al., 2013; Hakami and Bollegala, 2017). For the sake of comparison, the methodology of Kontonatsios et al. (2013) was followed in other respects as well, for instance by starting with a balanced data set, with 50% positive and 50% negative instances. However, since this is not realistic in an actual pipeline for multilingual ATE from comparable corpora, imbalanced datasets were created as well with only 20% and 5% positive instances. The number of positives always remains the same (see Table 1), only the number of negatives varies according to these percentages.

A final note on the data is that the specialised corpora that were used are extremely small (50k tokens per language/domain) compared to the ones used in similar research (rarely under 1M tokens). Some of the features do refer to frequencies in large reference corpora (see section 4), but due to the specialised nature of the corpora and the fact that multi-word terms and NEs are included, many of the terms (single-word, multi-word and NE) have very low frequencies. For instance, out of the 3873 valid term pairs in the English-Dutch corpus, 1125 of the English source terms and 1340 Dutch target terms appear only once in the specialised corpus, and 1242 of the English terms and 2154 of the Dutch terms do not appear in any of the reference corpora. Considering that in similar research, terms appearing fewer than 25 times are considered to be infrequent (Jakubina and Langlais, 2016, 2018), it is interesting to see whether a decent performance can be obtained on such infrequent terms.

4 Monolingual ATE Features

The monolingual ATE features are based on the HAMLET tool (Rigouts Terryn et al., 2019a) and can be divided into 5 groups: *shape*, *frequency*, *statistics*, *related terms*, and *linguistics*. The number of features in each group and the description of these features can be found in Table 2. Most of these features have already been used for monolingual ATE, though most approaches are limited to a small number of these features. The reference corpora are Wikipedia dumps and news corpora in the respective languages, all limited to 10M tokens. For English, the News on Web corpus was used (Davies, 2017), for French the Gigaword corpus (Graff et al., 2011) and for Dutch the newspaper section of OpenSONAR (Oostdijk et al., 2013). The linguistic preprocessing was performed with the LeTs Preprocess toolkit (van de Kauter et al., 2013) and the part-of-speech (POS) tag sets of the three different languages were all mapped to a single set of 23 tags, so the same tags could be used across the languages. Preliminary experiments determined that the best way to encode the POS-patterns, was to have 3 vectors for all 23 individual tags: one for the tag of the first token of the term, one for the last and one for the frequency of all tags in the term. In the case of single-word terms, these would all be the same, but it was still an efficient way to encode the POS pattern for terms of varying lengths, without either losing too much

Feature group	#	Features
Shape	20	term length (in tokens or characters), capitalisation, presence of special characters
Frequency	12	relative frequency and document frequency of original term or lemmatised term in domain-specific corpus, newspaper reference corpus and Wikipedia corpus
Statistics	25	various termhood and unithood measures, calculated both for the original term and the lemmatised form (Vintar's termhood measure (Vintar, 2010)), C-value, TF-IDF, log-likelihood ratio, domain consensus, domain specificity, weirdness, basic, combo basic (more information about measures in (Astrakhantsev et al., 2015); measures that require a general reference corpus are calculated twice: once for the newspaper reference corpus, once for the Wikipedia reference corpus
Related Terms	12	count, combined frequency and average domain specificity of related terms, i.e. terms with the same lemma or normalised form and terms that are part of or contain the term in question
Linguistics	75	presence in stopword list, tag by automatic named entity recognition and POS, encoded as 3 one-hot vectors for the POS of the first token, POS of the last token and the frequency of all POS tags in the term

Table 2: Feature groups of the monolingual ATE with the number (#) of features in each group and a description of the features in that group

information or creating a disproportionate amount of POS-related features. There are no restrictions on term length, frequency or part-of-speech.

5 Experiments

5.1 Classifier and Features

By interpreting ATE from comparable corpora as a supervised binary classification task, we aim to test the usefulness of the monolingual ATE features for bilingual linking. Precision, Recall and f1-scores were calculated for each experiment. All experiments were performed with Python's scikit-learn package. Hyperparameter optimisation was performed through grid search and to counter the effect of random variations, the results of each experiment are averaged over 5 trials. Experiments were performed with either 5-fold cross-validation (within all domains of a single language pair or within one domain and language pair) or with a separate train and test set (test on one domain in one language pair and train on the three others). Preliminary experiments showed that the Random Forrest Classifier (RFC) and Multi-Layer Perceptron (MLP) outperformed the Decision Tree Classifier and the Logistic Regression Classifier. Since the RFC was more efficient than the MLP, had been used in previous research (Kontonatsios et al., 2013) and had a more stable performance, all further experiments were performed with the RFC. Positive instances (valid equivalents) were labelled as '1' and negatives (wrong equivalents) as '0'. The hyperparameter search space remained unchanged throughout the project ('min_samples_leaf': [5, 10], 'min_samples_split': [2, 10, 20], 'n_estimators': [150] and standard settings for all other hyperparameters), with the exception of 'class_weight', which varied from ['balanced', 0: 1, 1: 1.5, 0: 1, 1: 2, 0: 1, 1: 2.5] for the balanced dataset, to ['balanced', 0: 1, 1: 2, 0: 1, 1: 3, 0: 1, 1: 4, 0: 1, 1: 5, 0: 1, 1: 6] for the dataset with 20% positives and ['balanced', 0: 1, 1: 8, 0: 1, 1: 10, 0: 1, 1: 12, 0: 1, 1: 15] for the dataset with 5% positives.

As stated in section 4, the features are the ones used for monolingual ATE. There were two different setups to combine the features. In the first (CONCAT), the monolingual features of source and target term were simply concatenated, without any additional transformations. For the second (ABSDIF), the absolute difference was taken for all respective features. The features regarding the terms' POS pattern were analysed in more detail, since it was assumed that these features could potentially be very informative. Since there was no restriction on term length or POS pattern, the list of possible patterns across all languages is very long (200+ unique patterns). Therefore, as explained in the previous section, for the monolingual ATE, instead of a one-hot vector for all possible patterns, three (much shorter) vectors were used for all tags: one for the frequency of each tag in the pattern, one for the tag of the first token and one for that of the last token. While some information is lost this way, its compactness and ability to generalise was proven with good results for monolingual ATE. However, since POS pattern might be even more important for the bilingual linking, both approaches were tested and compared. Preliminary experiments showed better results (gain of 0.05 in f1-score) for the compact representations. Consequently, all further experiments were performed with this version of the features. Since only limited performance was expected from these features, it was decided to also test their compatibility with a string similarity feature (Levenshtein ratio), which seems intuitively more directly useful for the detection of equivalents in related languages, such as the ones used in this project. Using the python-Levenshtein package[1], Levenshtein ratio was calculated between all source and target terms. Before training the models, all features that showed no variance in the training data were removed. Generally, this affected some of the POS-features and special character features. Finally, the remaining features were scaled to [-1,1].

All these methodological difference lead to many different configurations: separate train/test sets versus 5-fold cross-validation, CONCAT versus ABSDIF features, with and without Levenshtein ratio, balanced dataset (50/50) versus slightly imbalanced dataset (80/20) versus very imbalanced dataset (95/5), and also three language pairs and four domains. Various experiments will be described in more detail in the following sections, but it can already be stated that the results were surprisingly good. The best obtained f1-score with Levenshtein features was 0.970 (precision 0.957 and recall 0.984). This was on a balanced dataset for the domain of corruption, French to Dutch, with ABSDIF features and separate train and test sets. The standard deviation of the f1-scores over the 5 trials was 0.002, indicating a rather stable performance. The best f1-score without Levenshtein features was still 0.939 (precision 0.911 and recall 0.970), with a standard deviation of 0.015. This was for the balanced data in the domain of dressage, French to Dutch, with CONCAT features and 5-fold cross-validation of only in-domain data. For comparison, the best reported state-of-the-art f1-score with a similar setup (supervised binary classifier with balanced data and 3-fold cross-validation) and of character n-grams features for English-French is 0.916 (Kontonatsios et al., 2013). Considering the nature of our features and the amount of infrequent terms in the data, our results compare much more favourably than expected and are a promising indication that features from monolingual ATE are relevant enough to be re-used for cross-lingual linking for ATE from comparable corpora.

5.2 Impact of Domain and Training Data

While domain can have a substantial effect on performance of monolingual ATE (Fedorenko et al., 2013; Conrado et al., 2013), performance across domains for our experiments with the cross-lingual linking of term equivalents appears to be largely domain-independent. For instance, f1-scores for experiments on the balanced datasets, using 5-fold cross-validation and averaged over experiments with different features are extremely similar: 0.928 (corruption), 0.927 (dressage), 0.928 (heart failure), and 0.930 (wind energy). Scores for more imbalanced datasets and with different features are comparably similar. This is somewhat surprising, considering that terms do have different characteristics in different domains (Rigouts Terryn et al., 2018, 2019b), that corruption is actually a parallel corpus and that there is much more data available for the domain of heart failure. The fact that corruption is a parallel, rather than a comparable corpus, should make it easier to find equivalents, but that fact may be compensated by the difficulty of the domain, since it was reported to be the most difficult to annotate (both monolingually and cross-lingually). Nevertheless, despite the similar results in this case, some of the highest obtained f1-scores were still obtained in the domain of corruption. As for the much larger size of the heart failure dataset: this may not affect the cross-validation experiments, but for the experiments with separate train and test sets, which use only training data from the other domains, heart failure does have a lower f1-score (averaged over all experiments with separate test set) than the other domains: 0.688 versus 0.811, 0.806, and

[1] https://pypi.org/project/python-Levenshtein/

0.800 in corruption, dressage, and wind energy respectively.

domain	p	r	f1
corruption	0.881	0.936	0.907
dressage	0.911	0.955	0.932
heart failure	0.875	0.953	0.912
wind energy	0.908	0.954	0.930
Average	**0.894**	**0.950**	**0.920**

Table 3: Precision (p), recall (r) and f1-scores (f1) per domain, averaged over all language pairs, on balanced datasets, without Levenshtein features, with concatenated features, using 5-fold cross-validation (and in-domain training data)

domain	p	r	f1
corruption	0.903	0.827	0.887
dressage	0.903	0.889	0.896
heart failure	0.829	0.868	0.848
wind energy	0.894	0.869	0.881
Average	**0.882**	**0.874**	**0.878**

Table 4: Precision (p), recall (r) and f1-scores (f1) per domain, averaged over all language pairs, on balanced datasets, without Levenshtein features, with concatenated features, using separate train and test sets (without in-domain training data)

While performance is stable across domains, training data does have an impact. Experiments with separate test sets (and only out-of-domain training data) perform worse than cross-validation experiments (with in-domain training data). Tables 3 and 4 show the results with the same experimental setup (balanced datasets, without Levenshtein features, with concatenated features) with cross-validation versus separate train and test sets. It is worth noting that, with different experimental configurations, the conclusions remain the same: with cross-validation, there is little to no difference in performance between domains, whereas separate train and test data results in slightly lower f1-scores for heart failure, the domain for which less training data is available. Moreover, performance is better for the former. In conclusion, while this methodology seems to work equally well for different domains, the presence of in-domain training data is important, and the amount of training data could also influence the scores.

lng.	**without Lev.**			**with Lev.**		
pair	**p**	**r**	**f1**	**p**	**r**	**f1**
en-fr	0.81	0.89	0.85	0.91	0.94	0.92
en-nl	0.78	0.90	0.83	0.91	0.93	0.92
fr-nl	0.79	0.91	0.85	0.94	0.96	0.95
Av.	**0.79**	**0.90**	**0.84**	**0.92**	**0.94**	**0.93**

Table 5: Precision (p), recall (r) and f1-scores (f1) per language pair, evaluated with 5-fold cross-validation on all domains of a language pair combined, evaluated on slightly imbalanced datasets (20% positives), with concatenated features, with and without Levenshtein features

5.3 Impact of Features and Language Pair

The impact of CONCAT versus ABSDIF features is minimal, with a slight advantage for CONCAT features (average difference in f1-score of 0.04). This is not surprising, since both contain almost the same information, and it indicates that the model is able to generalise well from concatenated features without any explicit link between equivalent features of source and target language terms. Still, a little information is lost by taking the absolute difference, so for future research it could be worth investigating other ways of combining the features. Since CONCAT features work best, the following experiments will all use these, unless stated otherwise.

The Levenshtein feature does have a large impact, as expected. Table 5 compares the results of two experiments with the same settings, with and without Levenshtein ratio as a feature. Since the difference in performance is more pronounced for imbalanced datasets (though it is noticeable as well on the balanced data), the reported results are for the dataset with only 20% positive instances. As can be seen, the models that include Levenshtein features achieve higher f1-scores, more specifically by increasing precision. This is true for all language pairs and also holds with other experimental settings. The only notable difference in this regard is between language pairs: including the Levenshtein feature has a bigger impact on the French-Dutch language pairs than on the others, which is somewhat unexpected, since the other language pairs seem more related (historically, English and French have influenced each other a lot and English and Dutch are both Germanic languages). No immediate explanation has been found to explain this phenomenon, especially since it is present in all domains and almost all configurations of the experiment. It is also reflected in the feature importance of Levenshtein ratio (see section 5.5). Apart from the Levenshtein feature, results for all language pairs are comparable for all settings.

5.4 Data Balance

As has already become clear, performance with balanced data is surprisingly good. However, in an actual pipeline for multilingual ATE from comparable corpora, this is not realistic, so the stability of the performance for imbalanced data was tested as well. Table 6 reports precision, recall and f1-scores for balanced (50/50), slightly imbalanced (80/20) and very imbalanced (95/5) data, using cross-validation to test on all domains of a language pair combined, and without Levenshtein feature. It should be noted that these scores are even higher when including Levenshtein ratio (f1-score for highly imbalanced data with Levenshtein is on average 0.902 with these settings).

Balanced data (50/50)			
	Precision	**Recall**	**f1-score**
en-fr	0.882	0.951	0.915
en-nl	0.885	0.953	0.917
fr-nl	0.889	0.957	0.921
Imbalanced data A (80/20)			
	Precision	**Recall**	**f1-score**
en-fr	0.810	0.888	0.847
en-nl	0.776	0.897	0.832
fr-nl	0.795	0.907	0.847
Imbalanced data B (95/5)			
	Precision	**Recall**	**f1-score**
en-fr	0.796	0.806	0.801
en-nl	0.755	0.813	0.783
fr-nl	0.753	0.741	0.794

Table 6: Precision (p), recall (r) and f1-scores (f1) per language pair evaluated with 5-fold cross-validation on all domains of a language pair combined, without Levenshtein features and for three differently balanced datasets

The first thing that can be seen in these tables, is that performance remains relatively high, despite the imbalance in the datasets. This will, of course, be partly due to the RFC's 'class_weight' parameter, but it is still promising, especially given the nature of the features. In all cases, recall is favoured over precision, even though precision never drops below 0.741. Conclusions are similar for all domains and with different experimental setups.

5.5 Feature Importance

To analyse the importance the model attributed to the various features, we looked at the models for the balanced dataset, created with 5-fold cross-validation on all domains combined per language pair. Conclusions across the language pairs are very similar, except that, when included, the Levenshtein feature gets a much higher importance for the French-Dutch language pair. Naturally, this feature is important in all models, but even more so for this language combination. For instance, in the models with ABSDIF features, the Levenshtein gets an importance of 20.8% for English-French, 24.5% for English-Dutch and 30.9% for French-Dutch. As mentioned in section 5.3, we have not yet been able to explain this difference satisfactorily. Since the models for all language pairs are similar in all other respects, the rest of the discussion will focus on a single language pair (English-French) as an example.

group	**feature**	**imp.**
SIM	Levenshtein	21%
STAT	Combo Basic	3.6%
LING	freq. of determiner POS tag	3.4%
SHAPE	nr. of tokens	2.9%
STAT	domain specificity of lemmatised form vs. Wikipedia	2.9%
STAT	domain specificity of original form vs. Wikipedia	2.7%
STAT	Vintar's termhood measure of original form vs. newspaper corpus	2.6%
LING	freq. of preposition POS tag	2.4%
LING	preposition as first POS tag	2.3%
SHAPE	nr. of characters	2.2%
REL	average domain specificity of all terms that contain the current term	2.1%

Table 7: Top ranked features with their feature groups and their attributed importance for the balanced en-fr models, created with 5-fold cross-validation on all domains combined, including Levenshtein features, with ABSDIF features

Table 7 shows all features that were attributed an importance of over 2% in a model with ABSDIF features. These results are for a model with Levenshtein features, but the ranking of the features remains similar without this feature. As can be seen, features from almost all feature groups are included (see also section 4): the string similarity feature (Levenshtein) (SIM), statistical (STAT) features, linguistic (LING) features, morphological/shape features (SHAPE), and related terms features (REL). The highest ranked frequency feature is not far behind in the ranking, in 18th place: relative frequency of the lemmatised form in the Wikipedia corpus (1.4%). Logically, features that show the least variance are also least important (e.g., features about rare special characters or rare first/last POS tags). Still, many features from many different groups are used.

Results with CONCAT features are more difficult to interpret, because the features from source and target term are separate. When included, Levenshtein ratio remains most important, but the other results differ. Strangely, the highest ranked features are all about the target language term; the first source term feature is only ranked 25th. Another difference with the ABSDIF models, is that, apart from the Levenshtein feature, the 15 highest ranked features are all statistical (12) or about related terms (3).

5.6 Error Analysis

To get a more in-depth idea of the performance, a limited error analysis was performed on one of the models. The results of an RFC model were analysed in English-Dutch, tested on the domain of dressage and trained on all other domains in the same language pair. This experiment used CONCAT features, including Levenshtein ratio and was performed on a balanced dataset. The f1-score for this particular run was 0.952 (precision 0.932 and recall 0.973). Out of 814 instances, there were 396 true positives, 378 true negatives, 11 false negatives and 29 false positives. Out of 11 false negatives, 4 contained numbers in either source or target language, which were written in full in the other language (e.g., *three-loop serpentine* and *slangenvolte met 3 bogen*). If the model has learnt to look at the presence of a number (shape feature), it is not surprising that equivalents where only one term contains a number are wrongly classified, even though a few other examples were correctly recognised despite this difficulty. Of the others, 5 concern either a source or target term that can be interpreted differently depending on the POS-tag, so the term pair may only be truly equivalent in some contexts (e.g., the English term *hoofs* and its Dutch equivalent *hoeven*, which can mean either *hoofs*, but also, more commonly, *ought to*). The remaining two concern pairs with no string similarity, and also different length: *equestrianism* and *equitation* as equivalents for *hippische sport* (some discussion is possible about the exact equivalence in this case). The false positives can be similarly explained. Only two are due to a coincidentally high Levenshtein ratio. Among the true positives, it is clear that even formally very different term pairs (e.g., *half-pass* and *appuyement*, or *inside hind leg* and *binnenachterbeen*) and infrequent terms can be correctly recognised with this methodology.

6 Conclusions and Future Research

The goal of this pilot study was to investigate whether features used for monolingual ATE could also be used to detect cross-lingual equivalents in comparable corpora. For this purpose, an existing dataset was expanded and these data were used to build binary classifiers in various experiments, testing the impact of certain features, domains, language pairs and the distribution of the dataset. Considering the models use none of the traditional features for this task and that the corpora were small and, therefore, contained many infrequent terms, the results were very promising and even outperformed some of the state-of-the-art approaches. Future research will have to determine whether these conclusions hold up in a complete pipeline for multilingual ATE from comparable corpora and whether and how they can best be combined with more typical features, e.g., distributional linking.

7 Acknowledgements

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

References

Nikita Astrakhantsev, 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):336–349. https://doi.org/10.1134/S036176881506002X.

Danushka Bollegala, Georgios Kontonatsios, and Sophia Ananiadou. 2015. A Cross-Lingual Similarity Measure for Detecting Biomedical Term Translations. *PLOS ONE* 10(6). https://doi.org/10.1371/journal.pone.0126196.

Merley da Silva Conrado, 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 NAACL HLT 2013 Student Research Workshop*. ACL, Atlanta, GA, USA, pages 16–23.

Mark Davies. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* ArXiv: 1810.04805. http://arxiv.org/abs/1810.04805.

Denis Fedorenko, 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*. Kazan, Russia, volume 26, pages 15–23.

David Graff, Ângelo Mendonça, and Denise DiPersio. 2011. French Gigaword Third Edition LDC2011t10. Technical report, Linguistic Data Consortium, Philadelphia, USA.

H. Hakami and D. Bollegala. 2017. A classification approach for detecting cross-lingual biomedical term translations. *Natural Language Engineering* 23(1):31–51. https://doi.org/10.1017/S1351324915000431.

Amir Hazem and Emmanuel Morin. 2018. Leveraging Meta-Embeddings for Bilingual Lexicon Extraction from Specialized Comparable Corpora. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA, pages 937–949.

Geert Heyman, Ivan Vulić, and Marie-Francine Moens. 2018. A deep learning approach to bilingual lexicon induction in the biomedical domain. *BMC Bioinformatics* 19:259. https://doi.org/10.1186/s12859-018-2245-8.

Ann Irvine and Chris Callison-Burch. 2013. Supervised Bilingual Lexicon Induction with Multiple Monolingual Signals. In *Proceedings of NAACL-HLT*. ACL, Atlanta, GA, USA, pages 518–523.

Laurent Jakubina and Philippe Langlais. 2016. A Comparison of Methods for Identifying the Translation of Words in a Comparable Corpus: Recipes and Limits. *Computación y Sistemas* 20(3):449–458. https://doi.org/10.13053/cys-20-3-2465.

Laurent Jakubina and Philippe Langlais. 2018. Reranking Candidate Lists for Improved Lexical Induction. In Ebrahim Bagheri and Jackie C.K. Cheung, editors, *Advances in Artificial Intelligence. Canadian AI 2018*, Springer International Publishing, Cham, volume 10832 of *Lecture Notes in Computer Science*, pages 121–132. https://doi.org/10.1007/978-3-319-89656-4$_1$0.

Georgios Kontonatsios, Ioannis Korkontzelos, Sophia Ananiadou, and Junichi Tsujii. 2013. Using a Random Forest Classifier to recognise translations of biomedical terms across languages. In *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*. Association for Computational Linguistics, Sofia, Bulgaria, pages 95–104.

Anaïch Le Serrec. 2012. *Analyse comparative de l'équivalence terminologique en corpus parallèle et en corpus comparable : application au domaine du changement climatique*. Doctor of Philosophy, Université de Montréal.

Jingshu Liu, Emmanuel Morin, and Peña Saldarriaga. 2018. Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA, pages 2855–2866.

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 219–247. https://doi.org/10.1007/978-3-642-30910-6$_1$3.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, Tatjana Gornostaja, Špela Vintar, and Darja Fišer. 2019. Extracting Data from Comparable Corpora. In Inguna Skadiņa, Robert Gaizauskas, Bogdan Babych, Nikola Ljubešić, Dan Tufiş, and Andrejs Vasiļjevs, editors, *Using Comparable Corpora for Under-Resourced Areas of Machine Translation*, Springer International Publishing, Cham, pages 89–139. https://doi.org/10.1007/978-3-319-99004-0$_4$.

Ayla Rigouts Terryn, Patrick Drouin, Véronique Hoste, and Els Lefever. 2019a. Analysing the Impact of Supervised Machine Learning on Automatic Term Extraction: HAMLET vs TermoStat. In *Proceedings of RANLP 2019*. Varna, Bulgaria.

Ayla Rigouts Terryn, 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 LREC 2018*. ELRA, Miyazaki, Japan.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2019b. In No Uncertain Terms: A

Dataset for Monolingual and Multilingual Automatic Term Extraction from Comparable Corpora. *Language Resources and Evaluation* pages 1–34. https://doi.org/https://doi.org/10.1007/s10579-019-09453-9.

Marian van de Kauter, 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* 3:103–120.

Spela Vintar. 2010. Bilingual Term Recognition Revisited. *Terminology* 16(2):141–158.