Supervised Feature-based Classification Approach to Bilingual Lexicon Induction from Specialised Comparable Corpora

Ayla Rigouts Terryn
KU Leuven, Kulak
Centre for Computational Linguistics
ayla.rigoutsterryn@kuleuven.be

Abstract

This study, submitted to the BUCC2023 shared task on bilingual term alignment in comparable specialised corpora, introduces a supervised, feature-based classification approach. The approach employs both static cross-lingual embeddings and contextual multilingual embeddings, combined with surface-level indicators such as Levenshtein distance and term length, as well as linguistic information. Results exhibit improved performance over previous methodologies, illustrating the merit of integrating diverse features. However, the error analysis also reveals remaining challenges.

1 Introduction

The current contribution represents a submission to the BUCC2023 shared task on bilingual term alignment in comparable specialised corpora[1], specifically for the English-French language pair. The task can alternatively be phrased as bilingual lexicon induction (BLI) for terminology. It holds significant potential: it can benefit end-users with ad hoc bilingual terminology construction from relatively easily available comparable corpora, and offers researchers a probing task to assess the cross-lingual lexico-semantic knowledge of language models.

This complex task encompasses many current challenges in natural language processing. First, there are the challenges related to the data. With parallel corpora, identifying an equivalent term in the aligned sentence is, if not simple, at least a task with limited possible answers. With comparable corpora, the task becomes exponentially harder. There is no straightforward place in the corpus to start looking for equivalents, and no guarantee that there will be a valid cross-lingual equivalent for each term. This makes it difficult both to construct a gold standard dataset and to automate the task. For the shared task, the former issue was addressed by creating comparable corpora based on parallel corpora (Adjali et al., 2022b). Moreover, the shared task starts from a predefined list of candidate terms, so the focus is only on the cross-lingual alignment, and not term identification. Besides the data-related challenges, there are conceptual challenges. Terminological equivalence must be defined (Should terms and meanings be considered in context? How close does the meaning have to be, for a term to be considered valid? Do equivalents need to have the same syntactic function or can, e.g., an adjective be a valid equivalent for a noun?). This issue is circumvented in the shared task because the dataset was created based on parallel data, where the equivalence can be defined in context. As will be seen in the error analysis, this also means there are remaining questions as to the equivalence of, for instance, false positives. A final challenge concerns the choice of lexical items, in this case: single- and multi-word terms. Popular embedding-based approaches still struggle with accurate representations for multi-words. Including multi-words alongside single-words, with pairs of different lengths, forces participants to develop a methodology that handles both. For instance, the French equivalent for *train station* is *gare*, and for *database* it is *base de données*. Additionally, terminology is typically not as common as general vocabulary, so methodologies need to be more robust for smaller datasets and lower frequencies.

This paper starts with information on the shared task dataset and setup, and a section on related research. Next, the methodology is described, followed by the results and a brief error analysis, before summarising the findings and looking ahead in the conclusion.

[1] 2023 Building and Using Comparable Corpora shared task website: https://comparable.limsi.fr/bucc2023/bucc2023-task.html

Proceedings of the First ConTenNTS Workshop and the 16th BUCC workshop, pages 59–68
Varna, Sep 7, 2023
https://doi.org/10.26615/978-954-452-090-8_008

2 Dataset and Task

This year's shared task uses an identical setup and dataset to that of last year, so detailed information on the dataset and shared task rules can be found in last year's overview paper (Adjali et al., 2022a). As this was the only submission to this year's shared task, there is no separate overview paper this year, but additional information can be found on the website (see footnote 1). Shared task participants received a comparable corpus in source and target language, as well as lists of terms in source and target language. For the English-French language pair, a gold standard list of equivalents was provided as training data. Thus, the focus of this task is on the cross-lingual alignments. Not all terms in the lists of source and target language terms are present in the cross-lingual gold standard, and some terms have multiple correct equivalents.

Number of:	**training**	**test**
tokens in src corp.	19,358,505	4,464,919
tokens in tgt corp.	21,378,916	14,158,415
GS term pairs	2,519	1,970
src terms	3,132	1,270
tgt terms	2,984	9,712
src terms not in txt	17	0
tgt terms not in txt	30	9

Table 1: Number of tokens and terms in source (src) and target language (tgt) parts of the BUCC2023 dataset (tokenisation with spaCy); GS=Gold Standard, txt=text

Looking at the sizes of the datasets (see Table 1), a few things stand out. First, the corpora are quite large, with a slightly larger training corpus than test corpus. Though the source and target language parts of the training corpora are very similar in size, this is not the case for the test data, where the target language part is over three times as large. A second observation is that, for both train and test data, more terms are provided in the target language. However, this difference is once again much larger in the test corpus, with over seven times as many target language terms as source language terms. Third, as indicated, not all terms are included in the gold standard list of pairs. For instance, in the training data, around 80% of all source and target terms occur in the list of gold standard term pairs. One final difference between train and test data stands out: the number of gold standard term pairs in relation to the number of terms in each language. The number of gold standard term pairs is significantly lower than the number of terms in each language in the training data, whereas for the test data, there are more gold standard term pairs than source language terms. All of these differences between training and test data will influence the performance of any supervised system trained on this dataset.

All occurrences of all terms were identified in the lowercased and tokenised corpora. Most, but not all terms were found. In the training data, terms that were not found did not appear in the gold standard list of term pairs. However, in the test data some terms among the gold standard term pairs were not found in the corpus. Therefore, with this methodology, these terms could not be found by the system either. This was only the case for four pairs in the test data. There is a relatively even distribution between single- and multi-word terms in all parts of the corpora: 41% and 61% single-word terms in source and target training data; and 48% and 44% in source and target test data respectively).

The corpora contain texts from many different domains, and they are not very specialised. There are many very general terms (e.g., *water bottle*, *slow*, *remarks*, *young adults*), and much fewer specialised terms (e.g., *sovereignty*, *probiotic*, *legal person*). In both train and test data, there are many instances that would not conventionally be called terms e.g., *whosoever*, *very long time*, *necessarily*, *mere*, *friendly atmosphere*, etc. This is due to the automatic creation of the dataset, based on automatic term extraction with TermSuite (Cram and Daille, 2016). This is not to say that TermSuite's performance is bad, but there will inevitably be errors. Moreover, TermSuite is meant to work well on domain-specific specialised corpora and, while the BUCC corpora are somewhat specialised, they cover many different domains.

A ranked list of term pairs for the test data had to be submitted, with the most confidently predicted pair at the top. Up to five submissions were allowed per team. This list was evaluated through uninterpolated average precision (AP), with an evaluation script provided on GitHub[2].

3 Related Research

Last year, two teams submitted three runs each to this shared task (Adjali et al., 2022a). Team Jozef Stefan Institute (JSI) (Repar et al., 2022) trained an SVM binary classifier (Joachims, 2002), using features based on both the shared task resources,

[2]https://github.com/PierreZweigenbaum/bucc2022

and external (freely available) resources. They originally experimented with cross-lingual embeddings and sentence transformers, but chose a feature-based approach instead, due to unsatisfying results. This approach was based on previous work (Repar et al., 2021), where they incorporated "the cosine similarity values of the crosslingual and sentence transformer models into features of the machine learning model" (Repar et al., 2022, p. 63). They use four types of features. The "cognate-based features" take into account the specific differences between language. For instance, words ending in *-ology* in English are likely to end in *-ologie* in French. Their "dictionary-based features" rely on GIZA++ word alignment (Och and Ney, 2003). The "embedding-based features" use cosine similarity scores from cross-lingually aligned embeddings and five language models. The final group of "combined features" combines parts of all three other groups.

Team CUNI (Požár et al., 2022) submitted three different systems: one with cross-lingually aligned static embeddings, one with contextual multilingual embeddings, and one with unsupervised phrase-based machine translation. For the former, they trained FastText embeddings (Bojanowski et al., 2016) for both languages and aligned them cross-lingually using the MUSE tool (Conneau et al., 2018). For the contextual embeddings they worked with multilingual BERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019). Finally, they used the Monoses tool (Artetxe et al., 2019) to train an unsupervised phrase-based machine translation model on the provided comparable corpora. They also submitted a combined approach using both the cross-lingually aligned embeddings and phrase-based machine translation.

On the test set, the CUNI team obtained the highest uninterpolated average precision score (0.2816) with their combined system, closely followed by two of the submissions of JSI (0.2685 and 0.2674). Team JSI concluded that "careful feature engineering could still produce better results than more novel deep learning approaches", though they admit their system is "quite resource intensive" (Repar et al., 2022, p. 64). Team CUNI concluded that they were able to get the highest mean average precision (MAP) on the train set with the XLM-model, fine-tuned on the task dataset. The task organisers noticed that 10.7% of the gold standard term pairs were not found by any of the six submitted systems. A recurring issue was when multiple equivalents were present in the gold standard data, and the systems did not find all options. Multi-word terms were also found to be more difficult (Adjali et al., 2022a).

Of course, there is other related research outside of the shared task, though rarely including multi-words. Generally, it is interesting to see experiments where the information from language models is supplemented with additional (linguistic) information. Researchers argue that "there is still room in the NLP toolbox for methods that utilise discrete, symbolic linguistic knowledge; in fact, the two paradigms can be successfully combined for an amplified effect" (Majewska et al., 2022). Specifically for BLI, there is also a call for more rigour on the definition of the task and the used datasets. Laville et al. (2022) address the challenges related to evaluating BLI. Focusing on the popular and valuable MUSE dataset (Conneau et al., 2018), they identify several issues: there is an overrepresentation of proper nouns, of graphically similar (or identical) word pairs, and of high frequency words. A similar argument is made in the work of Kementchedjhieva et al (2019), who, additionally, talk about the gaps in the gold standard datasets. Some of these issues are notably less present in the BUCC shared task dataset because it focuses on terminology, making it more interesting and challenging. Nevertheless, the gold standard data is still automatically generated, so any research requires thorough evaluation that goes beyond simple scores to identify system strengths and weaknesses.

4 Experimental Setup

The methodology of this work is partly inspired by last year's submissions by team JSI (Repar et al., 2022): it is also a feature-based classifier that combines different types of features, including ones based on embeddings. The provided training data was used to train a supervised, binary classifier. Besides the data provided by the shared task, the methodology also relies on pretrained embeddings (no embeddings were trained or fine-tuned on the corpora from the shared task). Additionally, two of the submitted systems were trained on a combination of the provided training data and a supplementary dataset: the Annotated Corpora for Term Extraction Research (ACTER) (Rigouts Terryn et al., 2020), specifically using the cross-lingual annota-

tions in the domain of heart failure as described in Rigouts Terryn et al. (2018). Contrary to the shared task dataset, ACTER contains manual annotations, both for term identification and cross-lingual term alignment. It contains a mix of general and very specific terms, and the corpus is much smaller than that of the shared task (± 60k tokens per language). The English-French part of the dataset contains 2455 term pairs. The monolingual annotations of this dataset are publicly available[3], but the cross-lingual annotations require further validation before being released. Therefore, the methodology does not rely heavily on this dataset, except to test the impact of different training data.

4.1 Preprocessing

The first step in the methodology is the linguistic preprocessing of the corpora, including tokenisation, part-of-speech tagging, lemmatisation, and named entity recognition. This was performed using the English and French NLP pipelines of spaCy (version 3.5.4, *en_core_web_lg* and *fr_core_new_lg*) (Honnibal and Montani, 2017). Once the corpora have been preprocessed, all terms in the term lists are tokenised and mapped to the preprocessed corpora. All data is lowercased, but otherwise only exact matches are included. As discussed, not all terms were found in the corpus (see Table 1), and those that were not were excluded from this step onwards. Next, features were calculated for each possible term pair. With 3,115 English and 2,954 French terms remaining in the dataset, this meant 9,201,701 possible term pairs, with only 2519 (0.027%) positive (equivalent) instances. While some basic filters were applied afterwards to reduce this size, calculating all features and training remain computationally intense.

4.2 Features

cross-lingually aligned static embeddings (1) The strategy for the alignment of the static embeddings was based on previous research (Singh et al., 2022) on the improvement of domain-specific cross-lingual embeddings. For the monolingual embeddings, the same setup is used as in the previous study: FastText (Bojanowski et al., 2016), pretrained on the Common Crawl corpus and Wikipedia. "These models were trained using the Continuous Bag of Words (CBOW) model with position weights, a dimensionality of 300, character n-grams of length 5, a window of size 5, and 10 negative samples" (Singh et al., 2022, p. 128). The monolingual embeddings were aligned using VecMap (Artetxe et al., 2018). As shown in the study by Singh et al., cross-lingually aligned embeddings rely heavily on a relevant seed lexicon for the alignment. In their study, the lexicon was automatically constructed based on Wikipedia titles and the cross-lingual Wikipedia links. For the shared task, this approach could not easily be used, because the data is not limited to a single domain. Nevertheless, it was felt that including more specialised vocabulary in the seed dictionary could be beneficial. For the seed lexicon in the current study, the MUSE dataset (Conneau et al., 2018) was taken as a starting point. Though the quality of the English-French MUSE dataset was found to be high for the most frequent words, it was still manually amended (no automatic filtering was performed). This mainly meant removing some named entities to balance out their overrepresentation, focusing on those named entities that would be much more commonly used in English than in French, such as the names of the US states. A few errors were also removed. Starting from the MUSE list of 113,286 word pairs, 10,021 of the most common pairs were maintained. Additionally, 600 medical single-word medical term pairs from the work of Singh et al. (2022) were included. Finally, almost 1500 more single-word terms were manually added from diverse domains, based on the following online resources: Dictionnaire de l'Académie Nationale de Médicine[4] (379 term pairs), Anglais Pratique[5] (819 term pairs, including chemical elements and biological terms), and Lexique anglais-français d'écologie numérique et de statistique[6] (Legendre and Legendre, 1999) (285 terms from statistics). This resulted in a seed lexicon of 12,104 word pairs in total.

In the training dataset, there were no out-of-vocabulary terms; in the test dataset there were seven in English and thirteen in French (a possible indication that the test data is slightly more specialised than the training data). For multi-word terms, the token embeddings were combined using mean pooling, and only if at least half of the tokens were in-vocabulary. This was to avoid cases where embeddings only existed for the common

[3] https://github.com/AylaRT/ACTER

[4] http://dictionnaire.academie-medecine.fr/

[5] https://anglais-pratique.fr/

[6] http://www.numericalecology.com/lex/index.html

parts of the multi-word term, and not for the more meaningful part(s). This could be especially important in French, where multi-word terms are regularly connected by prepositions or articles (e.g., *environmental protection* in French is *protection de l'environnement*). FastText cosine similarity score was included as a single feature. In the case of out-of-vocabulary terms, average BERT cosine similarity (see next section) was taken instead.

multilingual contextual embeddings (5 or 3) The contextual embeddings of choice were pretrained multilingual BERT embeddings (Devlin et al., 2019), accessed through Hugging Face Transformers (Wolf et al., 2020-07-13). Again, the mean of the token embeddings is used for terms that contain multiple tokens. Five contexts were selected per term, evenly divided over the corpus. This strategy was meant to increase the possibility of finding the term in different informative contexts, without increasing the computational load too much by getting embeddings for all occurrences of all terms. For each term pair, five cosine similarity scores were calculated between the five embeddings for source and target terms. The official submissions to the shared task use these five features. However, it was then observed that including five cosine similarity scores from randomly selected contexts might not be ideal, as there is no telling which of the five will be more informative. Therefore, for subsequent experiments, the five original features were turned into three more interpreted ones: minimum, mean, and maximum cosine similarity scores (out of the original five).

edit distance (1) For the English-French dataset, edit distance could clearly be a relevant feature for many (though certainly not all) term pairs. Only Levenshtein distance (Levenshtein, 1966) was included as a feature, but more advanced implementations, e.g., like the cognate-based features of Repar et al. (2022), might be considered in the future.

frequency (6) Relative frequencies of the source and target terms in the term pair were included as well, with both the relative frequencies of the full forms and the lemmas. Additionally, the difference between the relative frequencies for full forms and lemmas was included as well, resulting in six frequency-related features. These will be more relevant for more comparable corpora, and less so for corpora that are more different in each language.

length (8) The length of source and target terms, measured in tokens and in characters, was included as well, alongside features with the difference (length source term minus length target term) and ratio (length source term divided by sum of length source term and length target term) between these lengths. This results in eight length-related features: four counting tokens, four with characters.

linguistic information (26) The most commonly assigned (out of five contexts) part-of-speech pattern (single tag in case of single-word terms) and named entity recognition label was obtained for each term. These were turned into numeric features in several ways. For the part-of-speech patterns, the five potentially most informative tags were selected: adjective, adverb, noun, proper noun, and verb. For all of these, the numbers of tokens with that tag in source and target terms were added as features, as well as the difference and ratio between the counts for source and target terms. This means that, for each of the five selected tags, four features were calculated (number of tokens with tag in source term, number of tokes with tag in target term, difference between these two, and ratio between these two), adding up to twenty part-of-speech features. Three additional part-of-speech features were added: (1) whether or not the pattern is identical for source and target terms, (2) whether or not the tags (regardless of their order) are identical for source and target terms, and (3) how many tags only occur in either source or target term. Finally, three more named entity recognition features are added: the average number of tokens of the source and target terms tagged as a named entity (across five contexts), and the difference between these averages for source and target terms. In total, there are 26 linguistic features.

4.3 Filtering

The resulting term pairs with features were filtered, e.g., removing any pairs with a FastText cosine similarity below 0.1, an average BERT cosine similarity below 0.1, or a very large difference in length (e.g., over 30 characters). The filters were intentionally set very broadly, so that no positive equivalents were removed from the training data. This means a very large number remains for training and classification (8,391,279). These filters could be set more strictly without losing (much) accuracy in the training data. Even so, 19 equivalent term pairs were removed from the test data with the broad

submission	Training data	Scoring	# predictions	AP	P	R	F1
1	BUCC-train + ACTER	f1_weighted	790	.30	.82	.33	.47
2	BUCC-train + ACTER	roc	1606	**.42**	.60	.49	.54
3	BUCC-train	f1_weighted	785	.30	.82	.32	.46
4	BUCC-train	roc	1205	.39	.71	.43	.54

Table 2: Details of the submitted systems, including the training data and scoring metric used for optimisation, as well as official results in terms of uninterpolated average precision (AP), precision (P), recall (R), and F1-score (F1)

filter, further illustrating the differences between the datasets.

4.4 Classifier

The experiments were performed in Scikit-learn (Pedregosa et al., 2011) with the Random Forest Classifier (Ho, 1995). This choice was motivated by its relative efficiency, interpretability, and the options to get probability scores for each prediction and estimate the importance of each feature. All features were scaled using the *StandardScaler*. Limited hyperparameter optimisation was used for the systems submitted to the shared task for the hyperparameters *min_sample_leaf*, *min_sample_split*, and *n_estimators*. For the remaining experiments in this contribution, no more optimisation was used and hyperparameters were set to: *class_weight='balanced', min_samples_leaf=5, min_samples_split=5, n_estimators=500*. Optimisation was either based on weighted f1-score (f1_w), or on Area Under the Receiver Operating Characteristic Curve (roc).

4.5 Data for Experiments

Four systems were officially submitted with the settings detailed in Table 2, and 47 features. These systems were trained on either the provided training data, or a combination of that training data with the ACTER dataset. Predictions were sorted based on the predicted probability of equivalence. Only positively predicted pairs were included (predicted probability of equivalence at least 50%), but this threshold could easily be adapted to favour either precision or recall.

Further experiments were performed after the official submissions. These used the three adapted features for cosine similarity from contextual embeddings (min, mean, and max cosine distance based on five contexts) and no hyperparameter optimisation. A first batch of experiments used just the BUCC training dataset, which was split into a separate train and test set. This was done by splitting the gold standard into 80% training pairs and 20% test pairs, and then splitting off the term pairs with features based on whether the source term was in the test set. The final batch of experiments used the same settings on the test data, which was made available by the organisers.

5 Results

The official results for the shared task can be found in Table 2. Though there were no other participants for a comparison this year, there is a considerable improvement over last year's top score of 0.28 AP. The best results were obtained by a system trained on a combination of the BUCC and the ACTER datasets, and optimised for roc. The addition of the ACTER dataset did not appear to have a big influence on the scores, but optimising for roc clearly worked better than optimising for f1_weighted. Precision scores are much higher than recall in all submitted systems, and many equivalent pairs could still be found below the threshold of 50% predicted confidence of equivalence, meaning that scores might be further improved by lowering the threshold.

As described, further experiments were performed to analyse the system and results in more detail. The experiments focused on the impact of: the scoring used for optimisation, the features, and the threshold value (i.e., the minimum predicted probability score for equivalence). Originally, this threshold was always set at 50% (only pairs the system actually predicted as equivalent), but since it was observed that uninterpolated average precision could be further improved by lowering this threshold, scores were also calculated at a cut-off point of 25%. For each experiment, uninterpolated average precision (AP) is reported as defined by shared task, as well as precision (P), recall (R), and F1-score (F1). Additionally, F1-score of the true label in the classification task (F1_true) is included, and the number of predicted equivalent pairs above the threshold (#),

Concerning the **features**, experiments were per-

train data	score	features	F1_ true	threshold@50% #	AP	P	R	F1	threshold@25% #	AP	P	R	F1
experiments on train data (80/20-split)													
BUCC	f1_w	all	.80	618	.82	.72	.88	.79	946	.86	.50	.94	.65
BUCC	roc	all	.81	612	.82	.72	.88	.79	950	**.87**	.50	.94	.65
BUCC	f1_w	cos	.71	647	.70	.62	.80	.70	958	.74	.45	.86	.59
BUCC	roc	cos	.70	655	.70	.61	.80	.69	956	.73	.45	.85	.59
BUCC	f1_w	cos&lev	.77	622	.77	.69	.85	.76	867	.80	.52	.89	.66
BUCC	roc	cos&lev	.77	624	.78	.69	.85	.76	875	.81	.52	.90	.66
BUCC	f1_w	limited	.83	599	.83	.75	.89	**.82**	864	.86	.54	.93	.69
BUCC	roc	limited	.83	598	**.84**	.75	.89	**.82**	846	.86	.55	.93	.69
experiments on test data													
BUCC+ ACTER	roc	all	.52	1177	.36	.69	.41	.52	2610	**.46**	.45	.60	.51
BUCC	roc	all	.52	1127	.36	.71	.41	.52	2355	**.46**	.47	.56	.51
BUCC	roc	limited	.52	1142	**.37**	.71	.41	.52	2065	.45	.52	.54	**.53**
BUCC	roc	cos	.39	1009	.24	.58	.30	.39	2132	.29	.37	.40	.39

Table 3: Results of further experiments on training data (80/20-split) and on test data

formed with: all described features (**all**: 45 features), only the cosine similarity features and Levenshtein distance (**cos&lev**; 5 features), only the cosine similarity features (**cos**: 4 features), or a limited set of features, including cos & lev, the difference in frequency, the four combined length features, the three part-of-speech features that are not about specific tags, and the difference in the average number of tokens recognised as named entities (**limited**: 14 features). The latter was meant to reduce some of the redundant information in the features, as there were many with both separate values for source and target terms, as well as a feature combining that information.

The results of these additional experiments can be seen in Table 3. The minor difference in setup for experiments with the test data as compared to the submitted runs (different features for contextual embeddings and no hyperparameter optimisation) results in slightly different, but still similar, scores for otherwise comparable experiments.

The first observation about the results in Table 3 is that all scores are much higher for experiments on a train/test-split of the training data, than for experiments trained on the training data and evaluated on the test data. While some deviation is to be expected, as discussed, there are significant differences between training and test datasets. Where AP scores were up to 0.87 for the training experiments, the highest score obtained on the test data is significantly lower at 0.46. A similar drop is seen for the F1-scores. For the experiments with threshold 50%, recall is only half of what it was for the training experiments. And though it is increased with a lower threshold, it is nowhere near the very high recall of 0.94 for the first experiments. Similar differences with results were reported last year. Despite the lower scores compared to the experiments on the training data, the top score of 0.46 AP is much higher than the best score of .28 submitted to the shared task last year.

The next observation is that a lower threshold results in (much) higher scores for AP. For the experiments on the training data, this improvement in AP is due to an increase in recall (up to .94), but the drop in precision results in a lower F1-score. For the experiments on the test data, AP is also highest with the lower threshold thanks to an improved recall, but in this case, the F1-scores are not much affected. Lowering the threshold has a higher impact on the number of predictions for the test data experiments. For the training data experiments, only 245 to 338 more pairs are extracted (+39% to 55%), whereas for the experiments on the test data, lowering the threshold results in up to 1433 more pairs, i.e., an increase of up to 122%. Being able to easily adjust this threshold depending on the requirements of the experiment is a considerable advantage.

Conversely to the results of the officially submitted runs, scoring used for optimisation has only a very minor impact, so, for the experiments on the

feature	importance
cos. sim. FastText	0.373
max. cos. sim. BERT	0.229
mean cos. sim. BERT	0.152
min. cos. sim. BERT	0.104
same POS tags	0.042
Levenshtein distance	0.034
difference in POS	0.015
length tokens ratio	0.013
same POS pattern	0.011
length chars ratio	0.007
length tokens difference	0.006
length chars difference	0.005
frequency difference	0.005
named entity rec. difference	0.002

Table 4: Importance of features limited features

test data, all training was optimised with roc. Predictably, the features do influence results. Clearly and unsurprisingly, the cosine features are most important, and results are not bad with just those features. The addition of Levenshtein is, predictably, an advantage for the English-French language pair as well. Interestingly, the other features also add relevant information, though the individual features are much less important. The system with more limited features appears to efficiently capture the relevant information.

Feature importance scores of a system trained with limited features and optimised with roc are shown in table 4. FastText cosine similarity score is the most important feature by some margin, followed by the three BERT cosine features which, together, are even more important. None of the other features are very important by themselves. Interestingly, the feature indicating whether source and target terms have the same part-of-speech tags (regardless of order) is more important than Levenshtein distance. In conclusion, these experiments show very promising results, especially for systems where training and test data are very similar.

6 Error Analysis

The output of the system trained on the shared task training data and tested on the test data (including all features) was analysed in more detail. Among the most confidently predicted pairs, there is a good mix of single- and multi-word terms, so not all multi-word term pairs were difficult to predict correctly. At the top of this list, there are a lot of pairs with a low Levenshtein distance, though not exclusively. For instance, at rank 7 there is the pair *typical recipe* and *recette typique*, and at rank 31 *disabled children* and *enfants handicapés*. The first false positive is found at rank 74, where *economic difficulties* is aligned with *problèmes économiques* (literally *economic problems*). While a more literal equivalent is available, this pair could certainly be considered equivalent in many contexts. This is seen for many of the highly ranked false positives: they either could be equivalent in certain contexts, or they should have been considered equivalents in the first place, e.g., *strategic game* and *jeu stratégique*, and *direct taxes* and *taxes directes*. Out of 1127 ranked equivalents, there were 325 false positives and 58 of those could be considered equivalent in many contexts, with an additional 33 deemed strongly related or potentially equivalent in some contexts. While these results require a more thorough analysis (with inter-annotator agreement), these numbers are an indication of the importance of a nuanced definition of equivalence and a thorough error analysis.

Naturally, some terms are also clearly misaligned. One of the, probably less serious, common misalignments is between terms with a different number. For instance, the singular *tumor* is aligned with the plural *tumeurs*, and the reverse is done for *wine bottles* and *bouteille de vin*. There are also a few false positives due to different parts-of-speech, for instance, *infected* was matched to *infection*. However, this only occurred eight times, so the part-of-speech features may have already prevented some of these mismatches. Multi-word terms with relatively general words were also found to be difficult. The term *access control system* was linked to 16 different French terms with a probability of at least 25%. A couple of other categories of terms that cause multiple false positives are: numbers, family relations, and colours. For instance, *eighth* is most confidently correctly aligned with *huitième*, but then also (with much lower probability) to *dix-septième* (*seventeenth*). Similarly, *aunt* is correctly matched to *tante* with a 94% probability, but then also to *oncle* (*uncle*) (91%), *mère* (*mother*) (86%), *père* (*father*) (71%), *neveu* (*nephew*) (69%), and so on. Similar issues are found for colours. Sometimes cultural differences play a role, for instance when *pound* is wrongly, but understandably, matched to *kilo*. While performance on multi-word terms was not especially

bad, the rather simplistic approach of averaging embeddings has clear downsides. This can be seen in misalignments where word order plays a role, e.g., *wine bottles* is misaligned to *vin en bouteille* (*botteled wine*), and *product safety* to *produits de sécurité* (*safety products*).

7 Conclusion

This contribution to the BUCC2023 shared task on bilingual term alignment in comparable specialised corpora presents a supervised approach with a feature-based classifier that combines features from embeddings with other information, including edit distance and linguistic characteristics. Results are promising and the system outperforms those from last year's submissions. Though the efficient random forest classifier is used, preparing the experiments is, admittedly, computationally expensive, since all source language terms are matched with all target language terms, and contextual features are calculated for each pair. However, it also provides interesting insights, for instance showing the relative importance of the various features. The error analysis illustrates various challenges, both in terms of the dataset and in terms of system weaknesses. Future research is planned to look into rich datasets for BLI from specialised corpora, to facilitate more thorough work on this task. Further experiments will include more features and compare different embeddings, as well as experiments with different types of classifiers. A more elaborate error analysis and the inclusion of more language pairs could further improve our understanding of the cross-lingual knowledge captured (or not) by both static and contextual embeddings.

References

Omar Adjali, Emmanuel Morin, Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. 2022a. Overview of the 2022 BUCC Shared Task- Bilingual Term Alignment in Comparable Specialized Corpora. In *Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022) @LREC2022*, pages 67–76, Marseille, France. European Language Resources Association.

Omar Adjali, Emmanuel Morin, and Pierre Zweigenbaum. 2022b. Building Comparable Corpora for Assessing Multi-Word Term Alignment. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 3103–3112, Marseille, France. European Language Resources Association.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A Robust Self-Learning Method for Fully Unsupervised Cross-Lingual Mappings of Word Embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An Effective Approach to Unsupervised Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, pages 7059–7069, Vancouver, Canada.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data.

Damien Cram and Beatrice Daille. 2016. TermSuite: Terminology Extraction with Term Variant Detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Berlin, Germany. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282. IEEE.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Thorsten Joachims. 2002. *Learning to Classify Text Using Support Vector Machines*. Springer US, Boston, MA.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in Evaluation: Misleading Benchmarks for Bilingual Dictionary Induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3334–3339, Hong Kong, China. Association for Computational Linguistics.

Martin Laville, Emmanuel Morin, and Philippe Langlais. 2022. About Evaluating Bilingual Lexicon Induction. In *Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022) @LREC2022*, pages 8–14, Marseille, France. European Language Resources Association.

Pierre Legendre and Louis Legendre. 1999. Lexique anglais-français d'écologie numérique et de statistique — English-French vocabulary of numerical ecology and statistics.

V.I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. 10(8):707–710.

Olga Majewska, Ivan Vulić, and Anna Korhonen. 2022. *Linguistically Guided Multilingual NLP: Current Approaches, Challenges, and Future Perspectives*, 1 edition, pages 163–188. CRC Press.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. 29(1):19–51.

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. (12):2825–2830.

Borek Požár, Klara Tauchmanová, Kristyna Neumannová, Ivana Kvapilíková, and Ondřej Bojar. 2022. CUNI Submission to the BUCC 2022 Shared Task on Bilingual Term Alignment. In *Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022) @LREC2022*, pages 43–49, Marseille, France. European Language Resources Association.

Andraz Repar, Senja Pollak, Matej Ulčar, and Boshko Koloski. 2022. Fusion of Linguistic, Neural and Sentence-Transformer Features for Improved Term Alignment. In *Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022) @LREC2022*, pages 61–66, Marseille, France. European Language Resources Association.

Andraž Repar, Matej Martinc, Matej Ulčar, and Senja Pollak. 2021. Word-embedding Based Bilingual Terminology Alignment. In *Proceedings Electronic Lexicography in the 21st Century (eLex 2021) Post-editing Lexicography*, page 98.

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 the Eleventh International Conference on Language Resources and Evaluation (LREC)*, pages 1803–1808, Miyazaki, Japan. European Language Resources Association.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2020. In No Uncertain Terms: A Dataset for Monolingual and Multilingual Automatic Term Extraction from Comparable Corpora. 54(2):385–418.

Pranaydeep Singh, Ayla Rigouts Terryn, and Els Lefever. 2022. Improving Domain-specific Cross-lingual Embeddings with Automatically Generated Bilingual Dictionaries. 12:124–140.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020-07-13. HuggingFace's Transformers: State-of-the-art Natural Language Processing.