Extracting bilingual terminologies from comparable corpora

Ahmet Aker, Monica Paramita, Robert Gaizauskas
University of Sheffield
ahmet.aker, m.paramita, r.gaizauskas@sheffield.ac.uk

Abstract

In this paper we present a method for extracting bilingual terminologies from comparable corpora. In our approach we treat bilingual term extraction as a classification problem. For classification we use an SVM binary classifier and training data taken from the EUROVOC thesaurus. We test our approach on a held-out test set from EUROVOC and perform precision, recall and f-measure evaluations for 20 European language pairs. The performance of our classifier reaches the 100% precision level for many language pairs. We also perform manual evaluation on bilingual terms extracted from English-German term-tagged comparable corpora. The results of this manual evaluation showed 60-83% of the term pairs generated are exact translations and over 90% exact or partial translations.

1 Introduction

Bilingual terminologies are important for various applications of human language technologies, including cross-language information search and retrieval, statistical machine translation (SMT) in narrow domains and computer-aided assistance to human translators. Automatic construction of bilingual terminology mappings has been investigated in many earlier studies and various methods have been applied to this task. These methods may be distinguished by whether they work on parallel or comparable corpora, by whether they assume monolingual term recognition in source and target languages (what Moore (2003) calls symmetrical approaches) or only in the source (asymmetric approaches), and by the extent to which they rely on linguistic knowledge as opposed to simply statistical techniques.

We focus on techniques for bilingual term extraction from comparable corpora – collections of source-target language document pairs that are not direct translations but are topically related. We choose to focus on comparable corpora because for many less widely spoken languages and for technical domains where new terminology is constantly being introduced, parallel corpora are simply not available. Techniques that can exploit such corpora to deliver bilingual terminologies are of significant practical interest in these cases.

The rest of the paper is structured as follows. In Section 2 we outline our method. In Section 3 we review related work on bilingual term extraction. Section 4 describes feature extraction for term pair classification. In Section 5 we present the data used in our evaluations and discuss our results. Section 6 concludes the paper.

2 Method

The method we present below for bilingual term extraction is a symmetric approach, i.e. it assumes a method exists for monolingual term extraction in both source and target languages. We do not prescribe what a term must be. In particular we do not place any particular syntactic restrictions on what constitutes an allowable term, beyond the requirement that terms must be contiguous sequences of words in both source and target languages.

Our method works by first pairing each term extracted from a source language document S with each term extracted from a target language document T aligned with S in the comparable corpus. We then treat term alignment as a binary classification task, i.e. we extract features for each source-target language potential term pair and decide whether to classify the pair as a term equivalent or not. For classification purposes we use an SVM binary classifier. The training data for the classifier is derived from EUROVOC (Steinberger et al., 2002), a term thesaurus covering the activities of the EU and the European Parliament. We have run our approach on the 21 official EU languages covered by EUROVOC, constructing 20 language pairs with English as the source

language. Considering all these languages allows us to directly compare our method's performance on resource-rich (e.g. German, French, Spanish) and under-resourced languages (e.g. Latvian, Bulgarian, Estonian). We perform two different tests. First, we evaluate the performance of the classifier on a held-out term-pair list from EUROVOC using the standard measures of recall, precision and F-measure. We run this evaluation on all 20 language pairs. Secondly, we test the system's performance on obtaining bilingual terms from comparable corpora. This second test simulates the situation of using the term alignment system in a real world scenario. For this evaluation we collected English-German comparable corpora from Wikipedia, performed monolingual term tagging and ran our tool over the term tagged corpora to extract bilingual terms.

3 Related Work

Previous studies have investigated the extraction of bilingual terms from parallel and comparable corpora. For instance, Kupiec (1993) uses statistical techniques and extracts bilingual noun phrases from parallel corpora tagged with terms. Daille et al. (1994), Fan et al. (2009) and Okita et al. (2010) also apply statistical methods to extract terms/phrases from parallel corpora. In addition to statistical methods Daille et al. use word translation information between two words within the extracted terms as a further indicator of the correct alignment. More recently, Bouamor et al. (2012) use vector space models to align terms. The entries in the vectors are co-occurrence statistics between the terms computed over the entire corpus.

Bilingual term alignment methods that work on comparable corpora use essentially three sorts of information: (1) cognate information, typically estimated using some sort of transliteration similarity measure (2) context congruence, a measure of the extent to which the words that the source term co-occurs with have the same sort of distribution and co-occur with words with the same sort distribution as do those words that co-occur with the candidate term and (3) translation of component words in the term and/or in context words, where some limited dictionary exists. For example, in Rapp (1995), Fung and McKeown (1997), Morin et. al. (2007), Cao and Li (2002) and Ismail and Manandhar (2010) the context of text units is used to identify term mappings. Transliteration and cognate-based information is exploited in Al-Onaizan and Knight (2002), Knight and Graehl (1998), Udupa et. al. (2008) and Aswani and Gaizauskas (2010).

Very few approaches have treated term alignment as a classification problem suitable for machine learning (ML) techniques. So far as we are aware, only Cao and Li (2002), who treat only base noun phrase (NP) mapping, consider the problem this way. However, it naturally lends itself to being viewed as a classification task, assuming a symmetric approach, since the different information sources mentioned above can be treated as features and each source-target language potential term pairing can be treated as an instance to be fed to a binary classifier which decides whether to align them or not. Our work differs from that of Cao and Li (2002) in several ways. First they consider only terms consisting of noun-noun pairs. Secondly for a given source language term $\langle N_1, N_2 \rangle$, target language candidate terms are proposed by composing all translations (given by a bilingual dictionary) of N_1 into the target language with all translations of N_2. We remove both these restrictions. By considering all terms proposed by monolingual term extractors we consider terms that are syntactically much richer than noun-noun pairs. In addition, the term pairs we align are not constrained by an assumption that their component words must be translations of each other as found in a particular dictionary resource.

4 Feature extraction

To align or map source and target terms we use an SVM binary classifier (Joachims, 2002) with a linear kernel and the trade-off between training error and margin parameter $c = 10$. Within the classifier we use language dependent and independent features described in the following sections.

4.1 Dictionary based features

The dictionary based features are language dependent and are computed using bilingual dictionaries which are created with GIZA++ (Och and Ney, 2000; Och and Ney, 2003). The DGT-TM parallel data (Steinberger et al., 2012) was input to GIZA++ to obtain the dictionaries. Dictionary entries have the form $\langle s, t_i, p_i \rangle$, where s is a source word, t_i is the i-th translation of s in the dictionary and p_i is the probability that s is translated by t_i, the p_i's summing to 1 for each s in the dictionary. From the dictionaries we removed all entries with $p_i < 0.05$. In addition we also removed

every entry from the dictionary where the source word was less than four characters and the target word more than five characters in length and vice versa. This step is performed to try to eliminate translation pairs where a stop word is translated into a non-stop word. After performing these filtering steps we use the dictionaries to extract the following language dependent features:

- ***isFirstWordTranslated*** is a binary feature indicating whether the first word in the source term is a translation of the first word in the target term. To address the issue of compounding, e.g. for languages like German where what is a multi-word term in English may be expressed as a single compound word, we check whether the compound source term has an initial prefix that matches the translation of the first target word, provided that translation is at least 5 character in length.
- ***isLastWordTranslated*** is a binary feature indicating whether the last word in the source term is a translation of the last word in the target term. As with the previous feature in case of compound terms we check whether the source term ends with the translation of the target last word.
- ***percentageOfTranslatedWords*** returns the percentage of words in the source term which have their translations in the target term. To address compound terms we check for each source word translation whether it appears anywhere within the target term.
- ***percentageOfNotTranslatedWords*** returns the percentage of words of the source term which have no translations in the target term.
- ***longestTranslatedUnitInPercentage*** returns the ratio of the number of words within the longest contiguous sequence of source words which has a translation in the target term to the length of the source term, expressed as a percentage. For compound terms we proceed as with ***percentageOfTranslatedWords***.
- ***longestNotTranslatedUnitInPercentage*** returns the percentage of the number of words within the longest sequence of source words which have no translations in the target term.

These six features are direction-dependent and are computed in both directions, reversing which language is taken as the source and which as the target. We also compute another feature *averagePercentageOfTranslatedWords* which builds the average between the feature values of *percentageOfTranslatedWords* from source to target and target to source. Thus in total we have 13 dictionary based features. Note for non-compound terms if we compare two words for equality we do not perform string match but rather use the Levenshtein Distance (see Section 4.2) between the two words and treat them as equal if the Levenshtein Distance returns $>= 0.95$. This is performed to capture words with morphological differences. We set 0.95 experimentally.

4.2 Cognate based features

Dictionaries mostly fail to return translation entries for named entities (NEs) or specialized terminology. Because of this we also use cognate based methods to perform the mapping between source and target words or vice versa. Aker et al. (2012) have applied (1) Longest Common Subsequence Ratio, (2) Longest Common Substring Ratio, (3) Dice Similarity, (4) Needleman-Wunsch Distance and (5) Levenshtein Distance in order to extract parallel phrases from comparable corpora. We adopt these measures within our classifier. Each of them returns a score between 0 and 1.

- **Longest Common Subsequence Ratio (LCSR):** The *longest common subsequence* (LCS) measure measures the longest common non-consecutive sequence of characters between two strings. For instance, the words "*dollars*" and "*dolari*" share a sequence of 5 non-consecutive characters in the same ordering. We make use of dynamic programming (Cormen et al., 2001) to implement LCS, so that its computation is efficient and can be applied to a large number of possible term pairs quickly. We normalize relative to the length of the longest term:

$$LCSR(X,Y) = \frac{\text{len}[LCS(X,Y)]}{\max[\text{len}(X), \text{len}(Y)]}$$

  where LCS is the longest common subsequence between two strings and characters in this subsequence need not be contiguous. The shorthand *len* stands for *length*.
- **Longest Common Substring Ratio (LCSTR):** The *longest common substring* (LCST) measure is similar to the LCS measure, but measures the longest common

consecutive string of characters that two strings have in common. I.e. given two terms we need to find the longest character n-gram the terms share. The formula we use for the LCSTR measure is a ratio as in the previous measure:

$$LCSTR(X,Y) = \frac{\text{len}[LCST(X,Y)]}{\max[\text{len}(X), \text{len}(Y)]}$$

- **Dice Similarity:**

$$dice = \frac{2 * LCST}{\text{len}(X) + \text{len}(Y)}$$

- **Needlemann Wunsch Distance (NWD):**

$$NWD = \frac{LCST}{\min[\text{len}(X) + \text{len}(Y)]}$$

- **Levenshtein Distance (LD):** This method computes the minimum number of operations necessary to transform one string into another. The allowable operations are insertion, deletion, and substitution. Compared to the previous methods, which all return scores between 0 and 1, this method returns a score s that lies between 0 and n. The number n represents the maximum number of operations to convert an arbitrarily dissimilar string to a given string. To have a uniform score across all cognate methods we normalize s so that it lies between 0 and 1, subtracting from 1 to convert it from a distance measure to a similarty measure:

$$LD_{normalized} = 1 - \frac{LD}{\max[\text{len}(X), \text{len}(Y)]}$$

4.3 Cognate based features with term matching

The cognate methods assume that the source and target language strings being compared are drawn from the same character set and fail to capture the corresponding terms if this is not the case. For instance, the cognate methods are not directly applicable to the English-Bulgarian and English-Greek language pairs, as both the Bulgarian and Greek alphabets, which are Cyrillic-based, differ from the English Latin-based alphabet. However, the use of distinct alphabets is not the only problem when comparing source and target terms. Although most EU languages use the Latin alphabet, the occurrence of special characters and diacritics, as well spelling and phonetic variations, are further challenges which are faced by term or entity mapping methods, especially in determining the variants of the same mention of the entity (Snae, 2007; Karimi et al., 2011).[1] We address this problem by mapping a source term to the target language writing system or vice versa. For mapping we use simple character mappings between the writing systems, such as $\alpha \rightarrow a$, $\phi \rightarrow ph$, etc., from Greek to English. The rules allow one character on the lefthand side (source language) to map onto one or more characters on the righthand side (target language). We created our rules manually based on sound similarity between source and target language characters. We created mapping rules for 20 EU language pairs using primarily Wikipedia as a resource for describing phonetic mappings to English.

After mapping a term from source to target language we apply the cognate metrics described in 4.2 to the resulting mapped term and the original term in the other language. Since we perform both target to source and source to target mapping, the number of cognate feature scores on the mapped terms is 10 – 5 due to source to target mapping and 5 due to target to source mapping.

4.4 Combined features

We also combined dictionary and cognate based features. The combined features are as follows:

- ***isFirstWordCovered*** is a binary feature indicating whether the first word in the source term has a translation (i.e. has a translation entry in the dictionary regardless of the score) or transliteration (i.e. if one of the cognate metric scores is above 0.7[2]) in the target term. The threshold 0.7 for transliteration similarity is set experimentally using the training data. To do this we iteratively ran feature extraction, trained the classifier and recorded precision on the training data using a threshold value chosen from the interval $[0, 1]$ in steps of 0.1. We selected as final threshold value, the lowest value for which the precision score was the same as when the threshold value was set to 1.
- ***isLastWordCovered*** is similar to the previous feature one but indicates whether the last word in the source term has a translation or

[1] Assuming the terms are correctly spelled, otherwise the misspelling is another problem.

[2] Note that we use the cognate scores obtained on the character mapped terms.

transliteration in the target term. If this is the case, 1 is returned otherwise 0.

- ***percentageOfCoverage*** returns the percentage of source term words which have a translation or transliteration in the target term.
- ***percentageOfNonCoverage*** returns the percentage of source term words which have neither a translation nor transliteration in the target term.
- ***difBetweenCoverageAndNonCoverage*** returns the difference between the last two features.

Like the dictionary based features, these five features are direction-dependent and are computed in both directions – source to target and target to source, resulting in 10 combined features.

In total we have 38 features – 13 features based on dictionary translation as described in Section 4.1, 5 cognate related features as outlined in Section 4.2, 10 cognate related features derived from character mappings over terms as described in Section 4.3 and 10 combined features.

5 Experiments

5.1 Data Sources

In our experiments we use two different data resources: EUROVOC terms and comparable corpora collected from Wikipedia.

5.1.1 EUROVOC terms

EUROVOC is a term thesaurus covering the activities of the EU and the European Parliament in particular. It contains 6797 term entries in 24 different languages including 22 EU languages and Croatian and Serbian (Steinberger et al., 2002).

5.1.2 Comparable Corpora

We also built comparable corpora in the information technology (IT) and automotive domains by gathering documents from Wikipedia for the English-German language pair. First, we manually chose one seed document in English as a starting point for crawling in each domain[3]. We then identified all articles to which the seed document is linked and added them to the crawling queue. This process is performed recursively for each document in the queue. Since our aim is to build a comparable corpus, we only added English documents which have an inter-language link in Wikipedia to a German document. We set a maximum depth of 3 in the recursion to limit size of the crawling set, i.e. documents are crawled only if they are within 3 clicks of the seed documents. A score is then calculated to represent the importance of each document d_i in this domain:

$$score_{d_i} = \sum_{j=1}^{n} \frac{freq_{d_{ij}}}{depth_{d_j}}$$

where n is the total number of documents in the queue, $freq_{d_{ij}}$ is 1 if d_i is linked to d_j, or 0 otherwise, and $depth_{d_j}$ is the number of clicks between d_j and the seed document. After all documents in the queue were assigned a score, we gathered the top 1000 documents and used inter-language link information to extract the corresponding article in the target language.

We pre-processed each Wikipedia article by performing monolingual term tagging using TWSC (Pinnis et al., 2012). TWSC is a term extraction tool which identifies terms ranging from one to four tokens in length. First, it POS-tags each document. For German POS-tagging we use TreeTagger (Schmid, 1995). Next, it uses term grammar rules, in the form of sequences of POS tags or non-stop words, to identify candidate terms. Finally, it filters the candidate terms using various statistical measures, such as pointwise mutual information and TF*IDF.

5.2 Performance test of the classifier

To test the classifier's performance we evaluated it against a list of positive and negative examples of bilingual term pairs using the measures of precision, recall and F-measure. We used 21 EU official languages, including English, and paired each non-English language with English, leading to 20 language pairs.[4] In the evaluation we used 600 positive term pairs taken randomly from the EUROVOC term list. We also created around 1.3M negative term pairs by pairing a source term with 200 randomly chosen distinct target terms. We select such a large number to simulate the real application scenario where the classifier will be confronted with a huge number of negative cases

[3] http://en.wikipedia.org/wiki/Information_technology for IT and http://en.wikipedia.org/wiki/Automotive_industry for automotive domain.

[4] Note that we do not use the Maltese-English language pair, as for this pair we found that 5861 out of 6797 term pairs were identical, i.e. the English and the Maltese terms were the same. Excluding Maltese, the average number of identical terms between a non-English language and English in the EUROVOC data is 37.7 (out of a possible 6797).

Table 1: Wikipedia term pairs processed and judged as positive by the classifier.

	Processed	Positive
DE IT	11597K	3249
DE Automotive	12307K	1772

and a relatively small number of positive pairs. The 600 positive examples contain 200 single term pairs (i.e. single word on both sides), 200 term pairs with a single word on only one side (either source or target) and 200 term pairs with more than one word on each side. For training we took the remaining 6200 positive term pairs from EUROVOC and constructed another 6200 term pairs as negative examples, leading to total of 12400 term pairs. To construct the 6200 negative examples we used the 6200 terms on the source side and paired each source term with an incorrect target term. Note that we ensure that in both training and testing the set of negative and positive examples do not overlap. Furthermore, we performed data selection for each language pair separately. This means that the same pairs found in, e.g., English-German are not necessarily the same as in English-Italian. The reason for this is that the translation lengths, in number of words, vary between language pairs. For instance *adult education* is translated into *Erwachsenenbildung* in German and contains just a single word (although compound). The same term is translated into *istruzione degli adulti* in Italian and contains three words. For this reason we carry out the data preparation process separately for each language pair in order to obtain the three term pair sets consisting of term pairs with only a single word on each side, term pairs with a single word on just one side and term pairs with multiple words on both sides.

5.3 Manual evaluation

For this evaluation we used the Wikipedia comparable corpora collected for the English-German (EN-DE) language pair. For each pair of Wikipedia articles we used the terms tagged by TWSC and aligned each source term with every target term. This means if both source and target articles contain 100 terms then this leads to 10K term pairs. We extracted features for each pair of terms and ran the classifier to decide whether the pair is positive or negative. Table 1 shows the number of term pairs processed and the count of pairs classified as positive. Table 2 shows five positive term pairs extracted from the English-German comparable corpora for each of the IT and automotive domains. We manually assessed a subset of the positive examples. We asked human assessors to categorize each term pair into one of the following categories:

1. **Equivalence:** The terms are exact translations/transliterations of each other.
2. **Inclusion:** Not an exact translation/transliteration, but an exact translation/transliteration of one term is entirely contained within the term in the other language, e.g: "F1 car racing" vs "Autorennen (car racing)".
3. **Overlap:** Not category 1 or 2, but the terms share at least one translated/transliterated word, e.g: "hybrid electric vehicles" vs "hybride bauteile (hybrid components)".
4. **Unrelated:** No word in either term is a translation/transliteration of a word in the other.

In the evaluation we randomly selected 300 pairs for each domain and showed them to two German native speakers who were fluent in English. We asked the assessors to place each of the term pair into one of the categories 1 to 4.

5.4 Results and Discussion

5.4.1 Performance test of the classifier

The results of the classifier evaluation are shown in Table 3. The results show that the overall performance of the classifier is very good. In many cases the precision scores reach 100%. The lowest precision score is obtained for Lithuanian (LT) with 67%. For this language we performed an error analysis. In total there are 221 negative examples classified as positive. All these terms are multi-term, i.e. each term pair contains at least two words on each side. For the majority of the misclassified terms – 209 in total – 50% or more of the words on one side are either translations or cognates of words on the other side. Of these, 187 contained 50% or more translation due to cognate words – examples of such cases are *capital increase – kapitalo eksportas* or *Arab organisation – Arabu lyga* with the cognates *capital – kapitalo* and *Arab – Arabu* respectively. For the remainder, 50% or more of the words on one side are dictionary translations of words on the other side. In order to understand the reason why the classifier treats such cases as positive we examined the

Table 2: Example positive pairs for English-German.

IT	Automotive
chromatographic technique — chromatographie methode	distribution infrastructure — versorgungsinfrastruktur
electrolytic capacitor — elektrolytkondensatoren	ambient temperature — außenlufttemperatur
natural user interfaces — natürliche benutzerschnittstellen	higher cetane number — erhöhter cetanzahl
anode voltage — anodenspannung	fuel tank — kraftstoffpumpe
digital subscriber loop — digitaler teilnehmeranschluss	hydrogen powered vehicle — wasserstoff fahrzeug

Table 3: Classifier performance results on EUROVOC data (P stands for precision, R for recall and F for F-measure). Each language is paired with English. The test set contains 600 positive and 1359400 negative examples.

	ET	HU	NL	DA	SV	DE	LV	FI	PT	SL	FR	IT	LT	SK	CS	RO	PL	ES	EL	BG
P	1	1	.98	1	1	.98	1	1	.7	1	1	1	.67	.81	1	1	1	1	1	1
R	.67	.72	.82	.69	.81	.77	.78	.65	.82	.66	.66	.7	.77	.84	.72	.78	.69	.8	.78	.79
F	.80	.83	.89	.81	.89	.86	.87	.78	.75	.79	.79	.82	.71	.91	.83	.87	.81	.88	.87	.88

training data and found 467 positive pairs which had the same characteristics as the negative examples in the testing set classified. We removed these 467 entries from the training set and re-trained the classifier. The results with the new classifier are 99% precision, 68% recall and 80% F score.

In addition to Lithuanian, two further languages, Portuguese (PT) and Slovak (SK), also had substantially lower precision scores. For these languages we also removed positive entries falling into the same problem categories as the LT ones and trained new classifiers with the filtered training data. The precision results increased substantially for both PT and SK – 95% precision, 76% recall, 84% F score for PT and 94% precision, 72% recall, 81% F score for SK. The recall scores are lower than the precision scores, ranging from 65% to 84%. We have investigated the recall problem for FI, which has the lowest recall score at 65%. We observed that all the missing term pairs were not cognates. Thus, the only way these terms could be recognized as positive is if they are found in the GIZA++ dictionaries. However, due to data sparsity in these dictionaries this did not happen in these cases. For these term pairs either the source or target terms were not found in the dictionaries. For instance, for the term pair *offshoring — uudelleensijoittautuminen* the GIZA++ dictionary contains the entry *offshoring* but according to the dictionary it is not translated into *uudelleensijoittautuminen*, which is the matching term in EUROVOC.

5.4.2 Manual evaluation

The results of the manual evaluation are shown in Table 4. From the results we can see that both assessors judge above 80% of the IT domain terms as category 1 – the category containing equivalent term pairs. Only a small proportion of the term pairs are judged as belonging to category 4 (3–7%) – the category containing unrelated term pairs. For the automotive domain the proportion of equivalent term pairs varies between 60 and 66%. For unrelated term pairs this is below 10% for both assessors.

Table 4: Results of the EN-DE manual evaluation by two annotators. Numbers reported per category are percentages.

Domain	Ann.	1	2	3	4
IT	P1	81	6	6	7
	P2	83	7	7	3
Automotive	P1	66	12	16	6
	P2	60	15	16	9

We investigated the inter-annotator agreement. Across the four classes the percentage agreement was 83% for the automotive domain term pairs and 86% for the IT domain term pairs. The kappa statistic, κ, was .69 for the automotive domain pairs and .52 for the IT domain. We also considered two class agreement where we treated term pairs within categories 2 and 3 as belonging to category 4 (i.e. as "incorrect" translations). In this case, for the automotive domain the percentage agreement was 90% and $\kappa = 0.72$ and for the IT domain percentage agreement was 89% with $\kappa = 0.55$. The agreement in the automotive domain is higher than in the IT one although both judges were computer scientists. We analyzed the differences and found that they differ in cases where the German and the English term are both in English. One of the annotators treated such cases as correct translation, whereas the other did not.

We also checked to ensure our technique was not simply rediscovering our dictionaries. Since the GIZA++ dictionaries contain only single word–single word mappings, we examined the

newly aligned term pairs that consisted of one word on both source and target sides. Taking both the IT and automotive domains together, our algorithm proposed 5021 term pairs of which 2751 (55%) were word-word term pairs. 462 of these (i.e. 17% of the word-word term pairs or 9% of the overall set of aligned term pairs) were already in either the EN-DE or DE-EN GIZA++ dictionaries. Thus, of our newly extracted term pairs a relatively small proportion are rediscovered dictionary entries. We also checked our evaluation data to see what proportion of the assessed term pairs were already to be found in the GIZA++ dictionaries. A total of 600 term pairs were put in front of the judges of which 198 (33%) were word-word term pairs. Of these 15 (less than 8% of the word-word pairs and less then 3% of the overall assessed set of assessed term pairs) were word-word pairs already in the dictionaries. We conclude that our evaluation results are not unduly affected by assessing term pairs which were given to the algorithm.

Error analysis For both domains we performed an error analysis for the unrelated, i.e. category 4 term pairs. We found that in both domains the main source of errors is due to terms with different meanings but similar spellings such as the following example (1).

(1) *accelerator — decelerator*

For this example the cognate methods, e.g. the Levenshtein similarity measure, returns a score of 0.81. This problem could be addressed in different ways. First, it could be resolved by applying a very high threshold for the cognate methods. Any cognate score below that threshold could be regarded as zero – as we did for the combined features (cf. Section 4.4). However, setting a similarity threshold higher than 0.9 – to filter out cases as in (1) – will cause real cognates with greater variation in the spellings to be missed. This will, in particular, affect languages with a lot of inflection, such as Latvian. Another approach to address this problem would be to take the contextual or distributional properties of the terms into consideration. To achieve this, training data consisting of term pairs along with contextual information is required. However, such training data does not currently exist (i.e. resources like EUROVOC do not contain contextual information) and it would need to be collected as a first step towards applying this approach to the problem.

Partial Translation The assessors assigned 6 – 7% of the term pairs in the IT domain and 12 – 16% in the automotive domain to categories 2 and 3. In both categories the term pairs share translations or cognates.

Clearly, if humans such as professional translators are the end users of these terms, then it could be helpful for them to find some translation units within the terms. In category 2 this will be the entire translation of one term in the other such as the following examples.[5]

(2) *visible graphical interface — grafische benutzerschnittstelle*

(3) *modern turbocharger systems — moderne turbolader*

In example (3) the a translation of the German term is to be found entirely within in the English term but the English term has the additional word *visible*, a translation of which is not found in the German term. In example (4), again the translation of the German term is entirely found in the English term, but as in the previous example, one of the English words – *systems* – in this case, has no match within the German term. In category 3 there are only single word translation overlaps between the terms as shown in the following examples.

(4) *national standard language — niederländischen standardsprache*

(5) *thermoplastic material — thermoplastische elastomere*

In example (5) *standard language* is translated to *standardsprache* and in example (6) *thermoplastic* to *thermoplastische*. The other words within the terms are not translations of each other.

Another application of the extracted term pairs is to use them to enhance existing parallel corpora to train SMT systems. In this case, including the partially correct terms may introduce noise. This is especially the case for the terms within category 3. However, the usefulness of terms in both these scenarios requires further investigation, which we aim to do in future work.

[5] In our data it is always the case that the target term is entirely translated within the English one and the other way round.

6 Conclusion

In this paper we presented an approach to align terms identified by a monolingual term extractor in bilingual comparable corpora using a binary classifier. We trained the classifier using data from the EUROVOC thesaurus. Each candidate term pair was pre-processed to extract various features which are cognate-based or dictionary-based. We measured the performance of our classifier using Information Retrieval (IR) metrics and a manual evaluation. In the IR evaluation we tested the performance of the classifier on a held out test set taken from EUROVOC. We used 20 EU language pairs with English being always the source language. The performance of our classifier in this evaluation reached the 100% precision level for many language pairs. In the manual evaluation we had our algorithm extract pairs of terms from Wikipedia articles – articles forming comparable corpora in the IT and automotive domains – and asked native speakers to categorize a selection of the term pairs into categories reflecting the level of translation of the terms. In the manual evaluation we used the English-German language pair and showed that over 80% of the extracted term pairs were exact translations in the IT domain and over 60% in the automotive domain. For both domains over 90% of the extracted term pairs were either exact or partial translations.

We also performed an error analysis and highlighted problem cases, which we plan to address in future work. Exploring ways to add contextual or distributional features to our term representations is also an avenue for future work, though it clearly significantly complicates the approach, one of whose advantages is its simplicitiy. Furthermore, we aim to extend the existing dictionaries and possibly our training data with terms extracted from comparable corpora. Finally, we plan to investigate the usefulness of the terms in different application scenarios, including computer assisted translation and machine translation.

Acknowledgements

The research reported was funded by the TaaS project, European Union Seventh Framework Programme, grant agreement no. 296312. The authors would like to thank the manual annotators for their helpful contributions. We would also like to thank partners at Tilde SIA and at the University of Zagreb for supplying the TWSC term extraction tool, developed within the EU funded project ACCURAT.

References

A. Aker, Y. Feng, and R. Gaizauskas. 2012. Automatic bilingual phrase extraction from comparable corpora. In *24th International Conference on Computational Linguistics (COLING 2012), IIT Bombay, Mumbai, India, 2012*. Association for Computational Linguistics.

Y. Al-Onaizan and K. Knight. 2002. Machine transliteration of names in arabic text. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, pages 1–13. Association for Computational Linguistics.

N. Aswani and R. Gaizauskas. 2010. English-hindi transliteration using multiple similarity metrics. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valetta, Malta*.

D. Bouamor, N. Semmar, and P. Zweigenbaum. 2012. Identifying bilingual multi-word expressions for statistical machine translation. In LREC 2012, Eigth International Conference on Language Resources and Evaluation, pages 674-679, Istanbul, Turkey, 2012. ELRA.

Y. Cao and H. Li. 2002. Base noun phrase translation using web data and the em algorithm. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. 2001. *Introduction to Algorithms*. The MIT Press, 2nd revised edition, September.

B. Daille, É. Gaussier, and J.M. Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 515–521. Association for Computational Linguistics.

X. Fan, N. Shimizu, and H. Nakagawa. 2009. Automatic extraction of bilingual terms from a chinese-japanese parallel corpus. In *Proceedings of the 3rd International Universal Communication Symposium*, pages 41–45. ACM.

P. Fung and K. McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.

A. Ismail and S. Manandhar. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 481–489. Association for Computational Linguistics.

T. Joachims. 2002. *Learning to classify text using support vector machines: Methods, theory and algorithms*, volume 186. Kluwer Academic Publishers Norwell, MA, USA:.

S. Karimi, F. Scholer, and A. Turpin. 2011. Machine transliteration survey. *ACM Computing Surveys (CSUR)*, 43(3):17.

K. Knight and J. Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

J. Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 17–22. Association for Computational Linguistics.

R. Moore. 2003. Learning translations of named-entity phrases from parallel corpora. In *In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1, pages 259266*. Association for Computational Linguistics.

E. Morin, B. Daille, K. Takeuchi, and K. Kageura. 2007. Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 664–671, Prague, Czech Republic, June. Association for Computational Linguistics.

F. J. Och and H. Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics*, pages 1086–1090, Morristown, NJ, USA. Association for Computational Linguistics.

F. J. Och Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

T. Okita, A. Maldonado Guerra, Y. Graham, and A. Way. 2010. Multi-word expression-sensitive word alignment. Association for Computational Linguistics.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, and Tatiana Gornostay. 2012. Term extraction, tagging, and mapping tools for under-resourced languages. In *Proc. of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June*, pages 20–21.

R. Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322. Association for Computational Linguistics.

Helmut Schmid. 1995. Treetagger— a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, page 43.

C. Snae. 2007. A comparison and analysis of name matching algorithms. *International Journal of Applied Science. Engineering and Technology*, 4(1):252–257.

R. Steinberger, B. Pouliquen, and J. Hagman. 2002. Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. *Computational Linguistics and Intelligent Text Processing*, pages 101–121.

R. Steinberger, A. Eisele, S. Klocek, S. Pilos, and P. Schlter. 2012. Dgt-tm: A freely available translation memory in 22 languages. In *Proceedings of LREC*, pages 454–459.

R. Udupa, K. Saravanan, A. Kumaran, and J. Jagarlamudi. 2008. Mining named entity transliteration equivalents from comparable corpora. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1423–1424. ACM.