Combining String and Context Similarity for Bilingual Term Alignment from Comparable Corpora

Georgios Kontonatsios[1,2] **Ioannis Korkontzelos**[1,2] **Jun'ichi Tsujii**[3] **Sophia Ananiadou**[1,2]
National Centre for Text Mining, University of Manchester, Manchester, UK[1]
School of Computer Science, University of Manchester, Manchester, UK[2]
Microsoft Research Asia, Beijing, China[3]
{gkontonatsios,ikorkontzelos,sananiadou}@cs.man.ac.uk
jtsujii@microsoft.com

Abstract

Automatically compiling bilingual dictionaries of technical terms from comparable corpora is a challenging problem, yet with many potential applications. In this paper, we exploit two independent observations about term translations: (a) terms are often formed by corresponding sub-lexical units across languages and (b) a term and its translation tend to appear in similar lexical context. Based on the first observation, we develop a new character n-gram compositional method, a logistic regression classifier, for learning a string similarity measure of term translations. According to the second observation, we use an existing context-based approach. For evaluation, we investigate the performance of compositional and context-based methods on: (a) similar and unrelated languages, (b) corpora of different degree of comparability and (c) the translation of frequent and rare terms. Finally, we combine the two translation clues, namely string and contextual similarity, in a linear model and we show substantial improvements over the two translation signals.

1 Introduction

Bilingual dictionaries of technical terms are resources useful for various tasks, such as computer-aided human translation (Dagan and Church, 1994; Fung and McKeown, 1997), *Statistical Machine Translation* (Och and Ney, 2003) and *Cross-Language Information Retrieval* (Ballesteros and Croft, 1997). In the last two decades, researchers have focused on automatically compiling bilingual term dictionaries either from *parallel* (Smadja et al., 1996; Van der Eijk, 1993) or *comparable* corpora (Rapp, 1999; Fung and Yee, 1998). While parallel corpora contain the same sentences in two languages, comparable corpora consist of bilingual pieces of text that share some features, only, such as topic, domain, or time period. Comparable corpora can be constructed more easily than parallel corpora. Freely available, up-to-date, on-line resources (e.g., Wikipedia) can be employed.

In this paper, we exploit two different sources of information to extract bilingual terminology from comparable corpora: the *compositional* and the *contextual clue*. The *compositional clue* is the hypothesis that the representations of a term in any pair of languages tend to consist of corresponding lexical or sub-lexical units, e.g., prefixes, suffices and morphemes. In order to capture associations of textual units across languages, we investigate three different character n-gram approaches, namely a *Random Forest* (RF) classifier (Kontonatsios et al., 2014), *Support Vector Machines* with an RBF kernel (SVM-RBF) and a *Logistic Regression* (LogReg) classifier. Whilst the previous approaches take as an input monolingual features and then try to find cross-lingual mappings, our proposed method (LogReg classifier) considers multilingual features, i.e., tuples of co-occurring n-grams.

The *contextual* clue is the hypothesis that mutual translations of a term tend to occur in similar lexical context. Context-based approaches are unsupervised methods that compare the context distributions of a source and a target term. A bilingual seed dictionary is used to map context vector dimensions of two languages. Li and Gaussier (2010) suggested that the seed dictionary can be used to estimate the degree of comparability of a bilingual corpus. Given a seed dictionary, the *corpus comparability* is the expectation of finding for each word of the source corpus, its translation in the target part of the corpus. The performance of context-based methods has been shown to depend on the frequency of terms to be translated and the

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1701–1712,
October 25-29, 2014, Doha, Qatar.

corpus comparability. In this work, we use an existing distributional semantics approach to locate term translations.

Furthermore, we hypothesise that the compositional and contextual clue are orthogonal, since the former considers the internal structure of terms while the latter exploits the surrounding lexical context. Based on the above hypothesis, we combine the two translation clues in a linear model.

For experimentation, we construct comparable corpora for four language pairs (English-Spanish, English-French, English-Greek and English-Japanese) of the biomedical domain.

We choose this domain because a large proportion of the medical terms tends to compositionally translate across languages (Lovis et al., 1997; Namer and Baud, 2007). Additionally, given the vast amount of newly introduced terms (neologisms) in the medical domain (Pustejovsky et al., 2001), term alignment methods are needed in order to automatically update existing resources.

We investigate the following aspects of term alignment: (a) the performance of compositional methods on closely related and on distant languages, (b) the performance of context vectors and compositional methods when translating frequent or rare terms, (c) the degree to which the corpus comparability affects the performance of context-based and compositional methods (d) the improvements that we can achieve when we combine the compositional and context clue.

Our experiments show that the performance of compositional methods largely depends on the distance between the two languages. The performance of the context-based approach is greatly affected by corpus-specific parameters (the frequency of occurrence of the terms to be translated and the degree of corpora comparability). It is also shown that the combination of compositional and contextual methods performs better than each of the clues, separately. Combined systems can be deployed in application environments with different language pairs, comparable corpora and seeds dictionaries.

The LogReg, dictionary extraction method described in this paper is freely available [1].

[1] http://personalpages.manchester.ac.uk/postgrad/georgios.kontonatsios/Software/LogReg-TermAlign.tar.gz

2 Related Work

Context-based methods (Fung and Yee, 1998; Rapp, 1999) adapt the *Distributional Hypothesis* (Harris, 1954), i.e., words that occur in similar lexical context tend to have the same meaning, in a multilingual environment. They represent the context of each term t as a context vector, usually following the *bag-of-words* model. Each dimension of the vector corresponds to a context word occurring within a predefined window, while the corresponding value is computed by a correlation metric, e.g., Log-Likelihood Ratio (Morin et al., 2007; Chiao and Zweigenbaum, 2002) or Point-wise Mutual Information (Andrade et al., 2010). A general bilingual dictionary is then used to translate/project the target context vectors into the source language. As a result, the source and target context vectors become directly comparable. In a final step, candidate translations are being ranked according to a distance metric, e.g., cosine similarity (Tamura et al., 2012) or Jaccard index (Zanzotto et al., 2010; Apidianaki et al., 2012).

Whilst context-based methods have become a common practise for bilingual dictionary extraction from comparable corpora, nonetheless, their performance is subject to various factors, one of which is the quality of the comparable corpus. Li and Gaussier (2010) introduced the corpus comparability metric and showed that it is related to the performance of context vectors. The higher the corpus comparability is, the higher the performance of context vectors is. Furthermore, context vector approaches are sensitive to the frequency of terms. For frequent terms, distributional semantics methods exhibit robust performance since the corresponding context is more informative. Chiao and Zweigenbaum (2002) reported an accuracy of 91% for the top 20 candidates when translating terms that occur 100 times or more. However, the performance of context vectors drastically decreases for lower frequency terms (Kontonatsios et al., 2014; Morin and Daille, 2010).

Our work is more closely related to a second class of term alignment methods that exploits the internal structure of terms between a source and a target language. Compositional translation algorithms are based on the *principal of compositionality* (Keenan and Faltz, 1985), which claims that the translation of the whole is a function of the translation of its parts. Lexical (Morin and Daille, 2010; Daille, 2012; Robitaille et al., 2006;

Tanaka, 2002) and sub-lexical (Delpech et al., 2012) compositional algorithms are knowledge-rich approaches that proceed in two steps, namely generation and selection. In the generation step, an input source term is segmented into basic translation units: words (lexical compositional methods) or morphemes (sub-lexical methods). Then a pre-compiled, seed dictionary of words or morphemes is used to translate the components of the source term. Finally, a permutation function generates candidate translations using the list of the translated segments. In the selection step, candidate translations are ranked according to their frequency (Morin and Daille, 2010; Robitaille et al., 2006) or their context similarity with the source term (Tanaka, 2002). The performance of the compositional translation algorithms is bound to the coverage of the seed dictionary (Daille, 2012). Delpech et al. (2012) noted that 30% of untranslated terms were due to the low coverage of the seed dictionary.

Kontonatsios et al. (2014) introduced a Random Forest (RF) classifier that learns correspondences of character n-grams between a source and target language. Unlike lexical and sub-lexical compositional methods, a RF classifier does not require a bilingual dictionary of translation units. The model is able to automatically build correlation paths between source and target sub-lexical segments that best discriminate translation from non-translation pairs. However, being a supervised method, it still requires a seed bilingual dictionary of technical terms for training. The RF classifier was previously applied on an English-Spanish comparable corpus and it was shown to significantly outperform context-based approaches.

3 Methods

In this section we describe the character n-gram models, the context vector method and the hybrid system. The lexicon induction task is formalised as a two-class classification problem. Given a pair of terms in a source and a target language, the output is a prediction of whether the terms are mutual translations are not. Furthermore, each term alignment method implements a ranking function that calculates a similarity score between a source and a target term. The methods rank target terms according to the similarity score and select the top N ranked terms as candidate translations. The ranking functions will be discussed in the following subsections.

3.1 Character n-gram models

Let s be a source term containing p character n-grams (s=$\{s_1, s_2, ..., s_p\}$ $s_i \in S$, $\forall i \in [1, p]$) and t a target term of q n-grams (t=$\{t_1, t_2, ..., t_q\}$ $t_i \in T$, $\forall i \in [1, q]$). We extract character n-grams by considering any contiguous, non-linguistically motivated sequence of characters that occurs within a window size of $[2-5]$ [2]) for English, French and Greek. For Japanese, unigrams are included (window size of $[1-5]$ because Japanese terms often contain Kanji (Chinese) characters.

Given the two lists of source and target n-grams, our objective is to find an underlying relationship between S and T that best discriminates translation from non-translation pairs. The RF classifier was previously shown to exhibit such behaviour (Kontonatsios et al., 2014). An RF classifier (Breiman, 2001) is a collection of decision trees voting for the most popular class. For a pair of source and target terms $\langle s, t \rangle$, the RF method creates feature vectors of a fixed size $2r$, i.e., *first order* feature space. The first r features are extracted from the source term, while the last r features from the target term. Each feature has a boolean value (0 or 1) that designates the presence/absence of the corresponding n-gram in the input instance.

The ability of the RF to detect latent associations between S and T relies on the decision trees. The internal nodes of a decision tree represent the n-gram features that are linked together in the tree-hierarchy. Each leaf node of the trees is labelled as *translation* or *non-translation* indicating whether the parent path of n-gram features is positively or negatively associated. The classification margin that we use to rank the candidate translations is given by a margin function (Breiman, 2001):

$$mg(X, Y) = av(I(x) = 1) - av(I(x)) = 0) \quad (1)$$

where x is an instance $\langle s, t \rangle$, $y \in Y = \{0, 1\}$ the class label, $I(\cdot) : (s, t) \longrightarrow \{0, 1\}$ is the indicator function of a decision tree and $av(I(\cdot))$ the average number of trees voting for the same class label. In our experiments, we used the same settings as the ones reported in Kontonatsios et al. (2014).

[2] we have experiments with larger and narrower window sizes but this setting resulted in better translation accuracy

We used 140 decision trees and $\log_2 |2q| + 1$ random features. For training an RF model, we used the WEKA platform (Hall et al., 2009).

The second class of machine learning algorithms that we investigate is Support Vector Machines (SVMs). The simplest version of SVMs is a linear classifier (linear-SVM) that tries to place a hyperplane, a decision boundary, that separates translation from non-translation instances. A linear-SVM is a feature agnostic method since the model only exploits the position of the vectors in the hyperspace to achieve class separation (Hastie et al., 2009).

The first order feature representation used with the RF classifier does not model associations between S and T. Hence, intuitively, a first order feature space is not linearly separable, i.e., there exists no decision boundary that divides the data points into translations and non-translations. [3]. To solve non-linear classification problems, SVMs employ non-linear kernels. A kernel function projects input instances into a higher dimensional space to discover non-linear associations between the initial features. In this new, projected feature space, the SVM attempts to define a separating plane. For training an non-linear SVM on the first order feature space, we used the LIBSVM package (Chang and Lin, 2011) with a *radial basis function* (RBF) kernel. For ranking candidate translations, we used the decision value given by LIBSVM which represents the distance between an instance and the hyperplane. To translate a source term, the method ranks candidate translations by decision value and suggests as best translation the candidate with the maximum distance (*maximum margin*).

While the first order models try to find cross-lingual mappings between monolingual features, our proposed method follows a different approach. It models cross-lingual links between the source and target character n-grams and uses them as *second order* features to train a linear classifier. A second order feature is a tuple of n-grams in S and T, respectively, that co-occur in a training, translation instance. Second order feature values are boolean. Given a translation instance $\langle s, t \rangle$ of p source and q target n-grams, there are $p \times q$ second order features. For dimensionality reduction, we consider as second order features the most frequent out of all possible first order feature combinations, only. Experiments indicate that a large number of features needs to be considered to achieve robust performance. To cope with the high dimensional second order space, we use LIBLINEAR (Fan et al., 2008), which is designed to solve large-scale, linear classifications problems. LIBLINEAR implements two linear classification algorithms: LogReg and linear-SVM. Both models solve the same optimisation problem, i.e., determine the optimal separating plane, but they adopt different loss functions. Since LIBLINEAR does not support decision value estimations for the linear-SVM, we only experimented with LogReg. Similarly to SVM-RBF, LogReg ranks candidate translations by classification margin.

3.2 Context vectors

We follow a standard approach to calculate context similarity of source and target terms (Rapp, 1999; Morin and Daille, 2010; Morin and Prochasson, 2011a; Delpech et al., 2012). Context vectors of candidate terms in the source and target language are populated after normalising each bilingual corpus, separately. Normalisation consists of stop-word filtering, tokenisation, lemmatisation and Part-of-Speech (PoS) tagging. For English, Spanish and French we used the TreeTagger (Schmid, 1994) while for Greek we used the ILSP toolkit (Papageorgiou et al., 2000). The Japanese corpus was segmented and PoS-tagged using Juman (Kurohashi and Kawahara, 2005).

In succession, monolingual context vectors are compiled by considering all lexical units that occur within a window of 3 words before or after a term (a seven-word window). Only lexical units (seeds) that occur in a bilingual dictionary are retained The values in context vectors are Log-Likelihood Ratio associations (Dunning, 1993) of the term and a seed lexical unit occurring in it. In a second step, we use the translations in the seed dictionary to map target context vectors into the source vector space. If there are several translations for a term, they are all considered with equal weights. Finally, candidate translations are ranked in descending order of the cosine of the angle between the mapped target vectors and the source

[3] We applied a linear-SVM with the first order feature representation on the four comparable corpora for English-French, English-Spanish, English-Greek and English-Japanese. In all cases, the best accuracies achieved were close to zero. Additionally, the ranked list of candidate translations was the same for all source terms. Hence, we can empirically suggest that the linear-SVM cannot exploit a first order feature space.

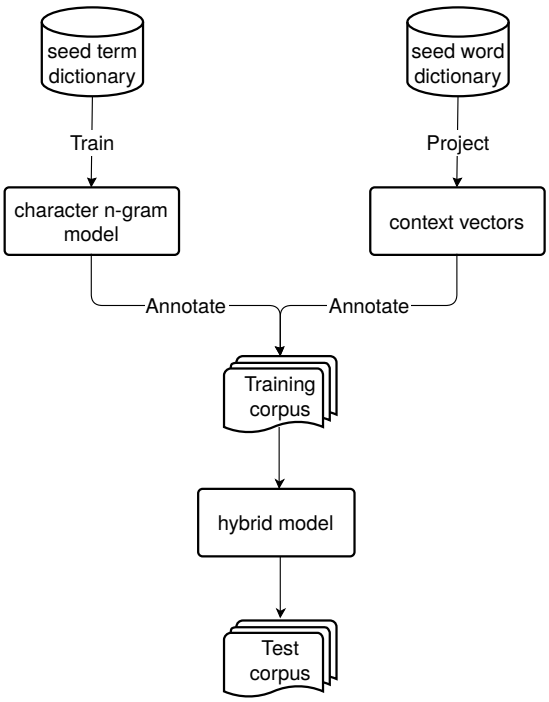


Figure 1: Architecture of the hybrid term alignment system.

vector.

3.3 Hybrid term alignment system

Figure 1 illustrates a block diagram of our term alignment system. We use two bilingual seed dictionaries: (a) a dictionary of term translation pairs to train the n-gram models and (b) a dictionary of word-to-word correspondences to translate target context vectors. The n-gram and context vector methods are used separately to score term pairs. The n-gram model computes the value of the *compositional clue* while the context vector estimates the score of the *contextual clue*. The hybrid model combines both methods by using the corresponding scores as features to train a linear classifier. For this, we used a linear-SVM of the LIBSVM package with default values for all parameters.

4 Data

Following previous research (Prochasson and Fung, 2011; Irvine and Callison-Burch, 2013; Klementiev et al., 2012), we construct comparable biomedical corpora using Wikipedia as a freely available resource.

Starting with a list of 4K biomedical English terms (query-terms), we collected 4K English Wikipedia articles, by matching query-terms to the topic signatures of articles. Then, we followed the Wikipedia interlingual links to retrieve thematically related articles in each target language. Since not all English articles contain links for all four target languages (Spanish, French, Greek and Japanese), we used a different list of query-terms for each language pair. Corpora were randomly divided into training and testing parts. For training we used 3K documents and for testing the remaining 1K. Table 1 shows the size of corpora in terms of numbers of source (SW) and target words (TW).

4.1 Seed dictionaries

As shown in Figure 1, the term alignment methods require two seed bilingual dictionaries: a term and a word dictionary. The character n-gram models rely on a bilingual term dictionary to learn associations of n-grams that appear often in technical terms. The dictionary may contain both single-word and multi-word terms. For English-Spanish and English-French we used UMLS (Bodenreider, 2004) while for English-Japanese we used an electronic dictionary of medical terms (Denshika and Kenkyukai, 1991).

An English-Greek biomedical dictionary was not available at the time of conducting these experiments, thus we automatically compiled a dictionary from a parallel corpus. For this, we trained a standard Statistical Machine Translation system (Koehn et al., 2007) on *EMEA* (Tiedemann, 2009), a biomedical parallel corpus containing sentence-aligned documents from the European Medicines Agency. Then, we extracted all English-Greek pairs for which: (a) the English sequence was listed in UMLS and (b) the translation probability was equal or higher to 0.7.

The sizes of the seed term dictionaries vary significantly, e.g., 500K entries for English-French but only 20K entries for English-Greek. However, the character n-gram models require a relatively small portion of the corresponding dictionary to converge. In the reported experiments, we used 10K translation pairs as positive, training instances. In addition, we generated an equal number of *pseudo-negative* instances by randomly matching non-translation terms.

Morin and Prochasson (2011b) showed that the translation accuracy of context vectors is higher when using bilingual dictionaries that contain both general language entries and technical terms rather than general or domain-specific dictionaries, sep-

	Training corpus		Test Corpus	
	# SW	# TW	# SW	# TW
en-fr	4.8M	2.2M	1.9M	1.1M
en-es	4.9M	2.5M	1.8M	0.9M
en-el	10.2M	2.4M	3.3M	1.3M
en-jpn	5.3M	2.4M	2.3M	1.2M

Table 1: Statistics of the English-French (en-fr), Engish-Spanish (en-es), English-Greek (en-el) and English-Japanese (en-jpn) Wikipedia comparable corpora. SW: source words, TW: target words

	Corpus Comparability	Seed words in dictionary
en-fr	0.71	66K
en-es	0.75	40K
en-el	0.68	22K
en-jpn	0.49	57K

Table 2: Corpus comparability and number of features of the seed word dictionaries

arately. In a mixed dictionary, lexical units are either single-word technical terms, such as "disease" and "patient", or general language words, such as "occur" and "high". Note that we have already compiled a seed term dictionary for each pair of languages. Following the suggestion of Morin and Prochasson (2011b), we attempt to enrich the seed term dictionaries with general language entries. For this, we extracted bilingual word dictionaries for English-Spanish, English-French and English-Greek by applying GIZA++ (Och and Ney, 2003) on the EMEA corpus. We then concatenated the word with the term dictionaries to obtain enhanced seeds for the three language pairs. For English-Japanese, we only used the term dictionary to translate the target context vectors.

Once the word dictionaries have been compiled, we compute the *corpus comparability* measure. Li and Gaussier (2010) define corpus comparability as the percentage of words that can be translated bi-directionally, given a seed dictionary.

Table 2 shows corpus comparability scores of the four corpora accompanied with the number of English, single words in the seed dictionaries. It can be observed that seed dictionary sizes are not necessarily proportional to the corresponding corpus comparability scores. As expected, for English-Japanese, corpus comparability is low because the dictionary contains single-word terms, only. The English-Spanish dictionary is smaller than the English-French but achieved higher corpus comparability, i.e., a higher percentage of words can be bi-directionally translated using the corresponding seed dictionary. A possible explanation is that the comparable corpora were constructed using different lists of query-terms. Hence, the query-terms used for English-Spanish retrieved a more coherent corpus. The resulting values of corpus comparability indicate that the context vectors will perform the best for English-Spanish while for English-Japanese the performance is expected to be substantially lower.

4.2 Training and evaluation datasets

For evaluation, we construct a test dataset of single-word terms, in particular nouns or adjectives. The dataset contains $1K$ terms that occur more frequently than 20 but not more than 200 times and are listed in the English part of the UMLS. In order to extract candidate translations, we considered all nouns or adjectives that occur at least 5 times in the target part of the corpus. Furthermore, we do not constraint the evaluation datasets only to those terms whose corresponding translation occurs in the corpus.

The hybrid model that combines the compositional and context clue, is based on a two-feature model. Therefore, the model converges using only a few hundred instances. For training a hybrid model, we used $1K$ translation instances that occurred in the training comparable corpora. Similarly, to the character n-gram models, *pseudo-negative* instances were generated by randomly coupling non-translation terms. The ratio of positive to negative instances is $1:1$.

5 Experiments

In this section, we present three experiments conducted to evaluate the *character n-gram*, *context vector* and *hybrid* methods. Firstly, we examine the performance of the n-gram models on closely related language pairs (English-French, English-Spanish), on a distant language pair (English-Greek) and on an unrelated language pair (English-Japanese). English and Greek are not unrelated because they are members of the same language family, but also not closely related because they use different scripts. Secondly,

we compare the character n-gram methods against context vectors when translating frequent or rare terms and on comparable corpora of similar language pairs (English-French, English-Spanish) but of different corpus comparability scores. Thirdly, we evaluate the hybrid method on all four comparable corpora and investigate the improvement margin of combining the *contextual* with the *compositional* clue.

As evaluation metrics, we adopt the top-N translation accuracy, following most previous approaches (Rapp, 1999; Chiao and Zweigenbaum, 2002; Morin et al., 2007; Tamura et al., 2012). The top-N translation accuracy is defined as the percentage of source terms for which a given method has output the correct translation among the top N candidate translations.

5.1 Character n-gram models

In the first experiment, we investigate the performance of the character n-gram models considering an increasing number of features. The features were sorted in order of decreasing frequency of occurrence. Starting from the top of the list, more features were incrementally added and translation accuracy was recorded.

Figure 2 shows the top-20 translation accuracy of single-word terms on an increasing number of first and second order features. With regards to the first order models (Subfigure 2a), the Random Forest (RF) classifier outperforms our baseline method (SVM-RBF) for all four language pairs. The largest margin between RF and SVM-RBF can be observed for the English-Greek dataset while for closely related language pairs, i.e., English-French and English-Spanish, the margin is smaller. Furthermore, it can be noted that using only a small number of first order features, $1K$ features (500 for the source and 500 for the target language, both n-gram models reach a stable performance.

In contrast to the first order models, the LogReg classifier requires a large number of second order features to achieve a robust performance (Subfigure 2b). Starting from $100K$ features, the translation accuracy continuously increases. The best performance is observed for a total number of $4M$ second order features when considering the English-French, English-Spanish and English-Greek datasets. For English-Japanese, the best performance is achieved for $2M$ features. Beyond this point, translation accuracy decreases slightly.

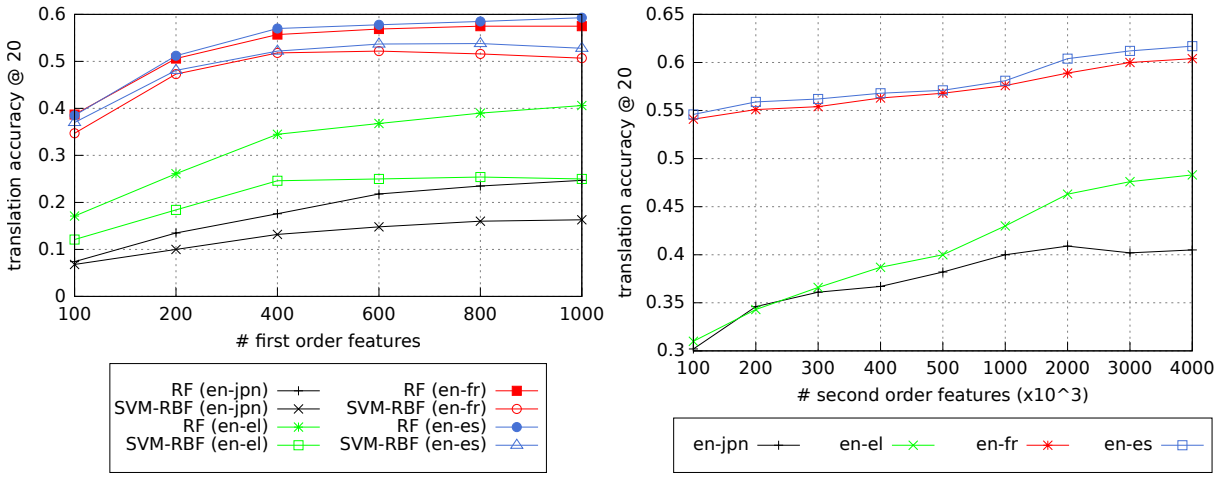
After feature selection is performed, we directly compare all the character n-gram models. Table 3 summarises performance achieved by the LogReg, RF and SVM-RBF models. It can be noted that LogReg and RF performed similarly for closely related languages (no statistically significant differences were observed) while both methods outperformed the SVM-RBF. However, for English-Greek and English-Japanese, LogReg achieved a statistically significant improvement over the translation accuracy of RF and SVM-RBF. LogReg outperformed RF by 7% for English-Greek, while for English-Japanese the improvement was 10% and 17% percent for top-1 and top-20 accuracy, respectively. Finally, it can be observed that the more distant the language pair is, the lower the performance.

5.2 N-gram methods and context vectors

In this experiment, we compare the n-gram methods against context vectors with regards to two parameters: (a) the frequency of source terms to be translated and (b) corpus comparability. English-French and English-Spanish are similar language pairs but the corresponding corpora are of different corpus comparability scores. To investigate how performance is affected by term occurrence frequency, we compiled an additional test dataset of $1K$ rare English terms in the frequency range $[10, 20]$. Our intuition is, that character n-gram methods will perform similarly for all settings since character n-grams are corpus independent features.

We compare (a) the character n-gram models (LogReg, RF and SVM-RBF) with (b) the context vector method (context) and (c) an upper bound. The latter represents the percentage of source terms for which a reference translation actually occurs in the target corpus. Hence, the upper bound is the maximum performance achievable according to the reference evaluation.

Figure 3a shows the top-20 translation accuracy for high and medium frequency terms, within the frequency range $[20, 200]$. Context vectors achieved a robust performance of 52% and 45% for English-Spanish and English-French, respectively. The difference in corpus comparability can explain this 7% margin between these performances. As shown in Table 2, the corpus comparability scores for English-Spanish and English-



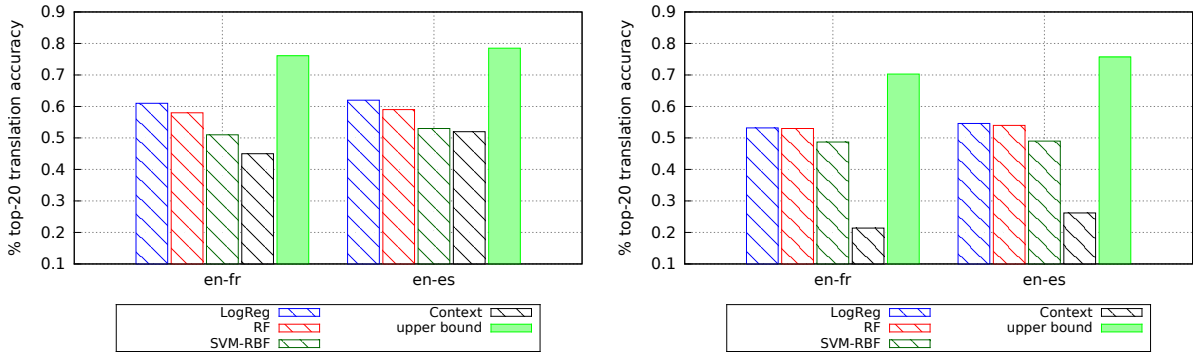
(a) First order n-gram models

(b) Second order n-gram model

Figure 2: Top-20 translation accuracy of models trained on (a) first and (b) second order features

	English-French		English-Spanish		English-Greek		English-Japanese	
	acc@1	acc@20	acc@1	acc@20	acc@1	acc@20	acc@1	acc@20
LogReg	0.45	**0.61**	0.42	**0.62**	**0.3**	**0.48**	**0.25**	**0.41**
RF	**0.47**	0.58	**0.43**	0.59	0.23	0.41	0.15	0.24
SVM-RBF	0.38	0.51	0.33	0.53	0.1	0.25	0.06	0.16

Table 3: Top-1 (acc@1) and top-20 (acc@20) translation accuracy of LogReg, RF and SVM-RBF



(a) Test terms with frequency $[20, 200]$

(b) Test terms with frequency $[10, 20]$

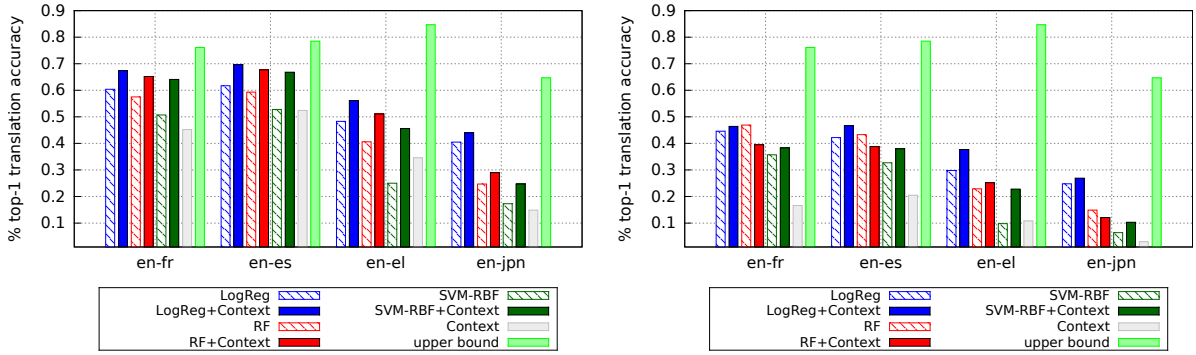
Figure 3: Top-20 translation accuracy of terms in the frequency range of $[10, 200]$ and $[10, 20]$

French are 0.75 and 0.71, respectively. In contrast to context vectors, the character n-gram methods performed comparably.

A second factor that affects the performance of context vectors, is the frequency of the terms to be translated. The translation of rare terms has been shown to be a challenging case for context vectors. For example, Morin and Daille (2010) reported low accuracy (21% for the top-20 candidates) of context vectors for terms occurring 20 times or less. In our experiments, Figure 3b illustrates accuracies achieved for less frequent terms ($[10, 20]$). The performance of context vectors is significantly lower, 26% for English-Spanish and 21% for English-French. Furthermore, the translation accuracy of the n-gram methods decreases slightly ($\sim$ 5% to 8%). This can be explained by the decrease of the upper bound for lower frequency terms ($\sim$ 3% to 6%).

5.3 Combining internal and contextual similarity

We have hypothesised that the *compositional* and *contextual* clue are orthogonal, i.e., they convey



(a) Top-20 accuracy (acc@20)

(b) Top-1 accuracy (acc@1)

Figure 4: Overall performance. Top-20 and top-1 translation accuracy

different and possibly complimentary information. To investigate this intuition, we evaluate the hybrid model on all four comparable corpora, for term occurrence frequencies in $[20, 200]$.

Figure 4a illustrates top-20 translation accuracy scores for (a) the character n-gram models, (b) the context vector method and (c) the hybrid models, i.e., LogReg+Context, RF+Context, SVM-RBF+Context. We observe that the combination of the *compositional* and *contextual* clue improved the performance of all methods. The hybrid model largely improved the performance of the SVM-RBF ($\sim$ 14% to 20%). With regards to the combined signals the translation accuracy of LogReg and RF increased by $\sim$ 4% for the English-Japanese corpus and $\sim$ 8% for all other corpora.

For the top 1 candidate translation, we observe in Figure 4 smaller improvements achieved by the hybrid model in comparison to the top-20 accuracy. Interestingly, the RF classifier performed slightly better on its own for English-French, English-Spanish and English-Japanese. This indicates that the hybrid method ranks more correct translations in the top 20 candidates but it does not always assign the best score to the correct answer.

6 Discusion and Future work

In this paper, we investigated a compositional and a context-based approach useful for compiling bilingual dictionaries of terms automatically from comparable corpora. Compositional translation methods exploit the internal structure of terms across languages while context-based approaches investigate the surrounding lexical context.

We proposed a character n-gram compositional method, i.e., a *Logistic Regression* classifier, which uses a multilingual representation, i.e., source and target terms. Experimental evidence showed that the LogReg classifier significantly outperformed the baseline methods on distant languages. For closely related languages, LogReg performed comparably to an existing n-gram method based on a *Random Forest* classifier.

Furthermore, we compared the n-gram models against a context-based approach under different corpus-specific parameters: (a) corpus comparability, which is relevant to the seed dictionary, and (b) the occurrence frequency of the terms to be translated. It was shown that the performance of n-gram methods was not affected by different parameter settings. Only small fluctuations were observed, since the n-gram methods are based on corpus-independent features, only. In contrast, the context-based method was affected by corpus comparability scores. The corresponding translation accuracy declined significantly for rare terms.

Finally, we hypothesised that the n-gram and context-based methods provide complimentary information. To test this hypothesis, we developed a hybrid method that combines compositional and contextual similarity scores as features in a linear classifier. The hybrid model achieved significantly better top-20 translation accuracy than the two methods separately but minor improvements were observed in terms of top-1 accuracy.

As future work, we plan to improve the quality of the extracted dictionary further by exploiting additional translation signals. For example, previous works (Schafer and Yarowsky, 2002; Klementiev et al., 2012) have reported that the *temporal* and *topic* similarity are clues that indicate translation equivalence. It would be interesting to investigate the contribution of different clues for various

experimental parameters, e.g., domain, distance of languages, types of comparable corpora.

Acknowledgements

The authors would like to thank Dr. Danushka Bollegala for providing feedback on this paper and the three anonymous reviewers for their useful comments and suggestions. This work was funded by the European Community's Seventh Framework Program (FP7/2007-2013) [grant number 318736 (OSSMETER)].

References

Daniel Andrade, Tetsuya Nasukawa, and Jun'ichi Tsujii. 2010. Robust measurement and comparison of context similarity for finding translation pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Marianna Apidianaki, Nikola Ljubešic, and Darja Fišer. 2012. Disambiguating vectors for bilingual lexicon extraction from comparable corpora. In *Eighth Language Technologies Conference*, pages 10–15.

Lisa Ballesteros and W.Bruce Croft. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *ACM SIGIR Forum*, volume 31, pages 84–91. ACM.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45:5–32.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. In *International Conference on Computational Linguistics*.

Ido Dagan and Ken Church. 1994. Termight: Identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied natural language processing*, pages 34–40. Association for Computational Linguistics.

Emmanuel Morin Bīeatrice Daille. 2012. Revising the compositional method for terminology acquisition from comparable corpora. *COLING 2012*, 1810.

Estelle Delpech, Béatrice Daille, Emmanuel Morin, and Claire Lemaire. 2012. Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. In *COLING*, pages 745–762.

Igakuyo Denshika and Jisho Kenkyukai. 1991. 250,000 medical term dictionary (in japanese). Nichigai Associates, Inc.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Pascale Fung and Kathleen McKeown. 1997. A technical word-and term-translation aid using noisy parallel corpora across language groups. *Machine Translation*, 12(1-2):53–87.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 414–420. Association for Computational Linguistics.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Z.S. Harris. 1954. Distributional structure. *Word*.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. 2009. *The elements of statistical learning*, volume 2. Springer.

Ann Irvine and Chris Callison-Burch. 2013. Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of NAACL-HLT*, pages 518–523.

Edward L Keenan and Leonard M Faltz. 1985. *Boolean semantics for natural language*, volume 23. Springer.

Alexandre Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 130–140. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*,

pages 177–180. Association for Computational Linguistics.

G. Kontonatsios, I. Korkontzelos, J. Tsujii, and S. Ananiadou. 2014. Using a random forest classifier to compile bilingual dictionaries of technical terms from comparable corpora. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 111–116. Association for Computational Linguistics.

Sadao Kurohashi and Daisuke Kawahara. 2005. Japanese morphological analysis system juman version 5.1 manual.

Bo Li and Eric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 644–652. Association for Computational Linguistics.

Christian Lovis, R Baud, PA Michel, JR Scherrer, and AM Rassinoux. 1997. Building medical dictionaries for patient encoding systems: A methodology. In *Artificial Intelligence in Medicine*, pages 373–380. Springer.

Emmanuel Morin and Béatrice Daille. 2010. Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, 44(1-2):79–95.

Emmanuel Morin and Emmanuel Prochasson. 2011a. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 27–34. Association for Computational Linguistics.

Emmanuel Morin and Emmanuel Prochasson. 2011b. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 27–34, Portland, Oregon, June. Association for Computational Linguistics.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo 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.

Fiammetta Namer and Robert Baud. 2007. Defining and relating biomedical terms: towards a cross-language morphosemantics-based system. *International Journal of Medical Informatics*, 76(2):226–233.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Harris Papageorgiou, Prokopis Prokopidis, Voula Giouli, and Stelios Piperidis. 2000. A unified pos tagging architecture and its application to greek. In *Proceedings of the 2nd Language Resources and Evaluation Conference*, pages 1455–1462, Athens, June. European Language Resources Association.

Emmanuel Prochasson and Pascale Fung. 2011. Rare word translation extraction from aligned comparable documents. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1327–1335. Association for Computational Linguistics.

James Pustejovsky, Jose Castano, Brent Cochran, Maciej Kotecki, and Michael Morrell. 2001. Automatic extraction of acronym-meaning pairs from medline databases. *Studies in health technology and informatics*, (1):371–375.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics.

Xavier Robitaille, Yasuhiro Sasaki, Masatsugu Tonoike, Satoshi Sato, and Takehito Utsuro. 2006. Compiling french-japanese terminologies from the web. In *EACL*.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, volume 12, pages 44–49. Manchester, UK.

Frank Smadja, Kathleen R McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, 22(1):1–38.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 24–36. Association for Computational Linguistics.

Takaaki Tanaka. 2002. Measuring the similarity between compound nouns in different languages using non-parallel corpora. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Jörg Tiedemann. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248.

Pim Van der Eijk. 1993. Automating the acquisition of bilingual terminology. In *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, pages 113–119. Association for Computational Linguistics.

Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1263–1271, Stroudsburg, PA, USA. Association for Computational Linguistics.