Using a Random Forest Classifier to Compile Bilingual Dictionaries of Technical Terms 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

We describe a machine learning approach, a *Random Forest* (RF) classifier, that is used to automatically compile bilingual dictionaries of technical terms from comparable corpora. We evaluate the RF classifier against a popular term alignment method, namely *context vectors*, and we report an improvement of the translation accuracy. As an application, we use the automatically extracted dictionary in combination with a trained Statistical Machine Translation (SMT) system to more accurately translate *unknown* terms. The dictionary extraction method described in this paper is freely available [1].

1 Background

Bilingual dictionaries of technical terms are important resources for many *Natural Language Processing* (NLP) tasks including *Statistical Machine Translation* (SMT) (Och and Ney, 2003) and *Cross-Language Information Retrieval* (Ballesteros and Croft, 1997). However, manually creating and updating such resources is an expensive process. In addition to this, new terms are constantly emerging. Especially in the biomedical domain, which is the focus of this work, there is a vast number of *neologisms*, i.e., newly coined terms, (Pustejovsky et al., 2001).

Early work on bilingual lexicon extraction focused on clean, parallel corpora providing satisfactory results (Melamed, 1997; Kay and Röscheisen, 1993). However, parallel corpora are expensive to construct and for some domains and language pairs are scarce resources. For these reasons, the focus has shifted to comparable corpora that are more readily available, more up-to-date, larger and cheaper to construct than parallel data. Comparable corpora are collections of monolingual documents in a source and target language that share the same topic, domain and/or documents are from the same period, genre and so forth.

Existing methods for bilingual lexicon extraction from comparable corpora are mainly based on the same principle. They hypothesise that a word and its translation tend to appear in similar lexical context (Fung and Yee, 1998; Rapp, 1999; Morin et al., 2007; Chiao and Zweigenbaum, 2002). Context vector methods are reported to achieve robust performance on terms that occur frequently in the corpus. Chiao and Zweigenbaum (2002) achieved a performance of 94% accuracy on the top 20 candidates when translating high frequency, medical terms (frequency of 100 or more). In contrast, Morin and Daille (2010) reported an accuracy of 21% for multi-word terms occurring 20 times or less, noting that translating rare terms is a challenging problem for context vectors.

Kontonatsios et al. (2013) introduced an RF classifier that is able to automatically learn association rules of textual units between a source and target language. However, they applied their method only on artificially constructed datasets containing an equal number of positive and negative instances. In the case of comparable corpora, the datasets are highly unbalanced (given n, m source and target terms respectively, we need to classify $n \times m$ instances). In this work, we incorporate the classification margin into the RF model, to allow the method to cope with the skewed distribution of positive and negative instances that occurs in comparable corpora.

Our proposed method ranks candidate translations using the classification margin and suggests as the best translation the candidate with the *maximum margin*. We evaluate our method on an

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

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 111–116,
Gothenburg, Sweden, April 26-30 2014.

English-Spanish comparable corpus of Wikipedia articles that are related to the medical sub-domain of "breast cancer". Furthermore, we show that dictionaries extracted from comparable corpora can be used to dynamically augment an SMT system in order to better translate *Out-of-Vocabulary (OOV)* terms.

2 Methodology

A pair of terms in a source and target language is represented as a feature vector where each dimension corresponds to a unique character n-gram. The value of each dimension is 0 or 1 and designates the occurrence of the corresponding n-gram in the input terms. The feature vectors that we use contain $2q$ dimensions where the first q dimensions correspond to the n-gram features extracted from the source terms and the last q dimensions to those from the target terms. In the reported experiments, we use the 600 (300 source and 300 target) most frequently occurring n-grams.

The underlying mechanism that allows the RF method to learn character gram mappings between terms of a source and target language is the decision trees. A node in the decision tree is a unique character n-gram. The nodes are linked through the branches of the trees and therefore the two sub-spaces of q source and q target character grams are combined. Each decision tree in the forest is constructed as follows: every node is split by considering $|\phi|$ random n-gram features of the initial feature set Ω, and a decision tree is fully grown. This process is repeated $|\tau|$ times and constructs $|\tau|$ decision trees. We tuned the RF classifier using 140 random trees where we observed a plateau in the classification performance. Furthermore, we set the number of random features using $|\phi| = \log_2 |\Omega| + 1$ as suggested by Breiman (2001).

The classification margin that we use to rank the candidate translations is calculated by simply subtracting the average number of trees predicting that the input terms are not translations from the average number of decision trees predicting that the terms are mutual translations. A larger classification margin means that more decision trees in the forest classify an instance as a translation pair.

For training an RF model, we use a bilingual dictionary of technical terms. When the dictionary lists more than one translation for an English term, we randomly select only one. Negative instances are created by randomly matching non-translation pairs of terms. We used an equal number of positive and negative instances for training the model. Starting from $20,000$ translation pairs we generated a training dataset of $40,000$ positive and negative instances.

2.1 Baseline method

The context projection method was first proposed by (Fung and Yee, 1998; Rapp, 1999) and since then different variations have been suggested (Chiao and Zweigenbaum, 2002; Morin et al., 2007; Andrade et al., 2010; Morin and Prochasson, 2011). Our implementation more closely follows the context vector method introduced by (Morin and Prochasson, 2011).

As a preprocessing step, stop words are removed using an online list [2] and lemmatisation is performed using TreeTagger (Schmid, 1994) on both the English and Spanish part of the comparable corpus. Afterwards, the method proceeds in three steps. Firstly, for each source and target term of the comparable corpus, i.e., i, we collect all lexical units that: (a) occur within a window of 3 words around i (a seven-word window) and (b) are listed in the seed bilingual dictionary. The lexical units that satisfy the above two conditions are the dimensions of the context vectors. Each dimension has a value that indicates the correlation between the context lexical unit and the term i. In our approach, we use the log-likelihood ratio. In the second step, the seed dictionary is used to translate the lexical units of the Spanish context vectors. In this way the Spanish and English vectors become comparable. When several translations are listed in the seed dictionary, we consider all of them. In the third step, we compute the *context similarity*, i.e., distance metric, between the vector of an English term to be translated with every projected, Spanish context vector. For this we use the cosine similarity.

3 Experiments

In this section, we evaluate the two dictionary extraction methods, namely context vectors and RF, on a comparable corpus of Wikipedia articles.

For the evaluation metric, we use the top-k translation accuracy [3] and the mean reciprocal

[2] http://members.unine.ch/jacques.savoy/clef/index.html

[3] the percentage of English terms whose top k candidates contain a correct translation

rank (MRR) [4] as in previous approaches (Chiao and Zweigenbaum, 2002; Chiao and Zweigenbaum, 2002; Morin and Prochasson, 2011; Morin et al., 2007; Tamura et al., 2012). As a reference list, we use the UMLS metathesaurus[5]. In addition to this, considering that in several cases the dictionary extraction methods retrieved synonymous translations that do not appear in the reference list, we manually inspected the answers. Finally, unlike previous approaches (Chiao and Zweigenbaum, 2002), we do not restrict the test list only to those English terms whose Spanish translations are known to occur in the target corpus. In such cases, the performance of dictionary extraction methods have been shown to achieve a lower performance (Tamura et al., 2012).

3.1 Data

We constructed a comparable corpus of Wikipedia articles. For this, we used Wikipedia's search engine [6] and submitted the queries "breast cancer" and "cáncer de mama" for English and Spanish respectively. From the returned list of Wikipedia pages, we used the $1,000$ top articles for both languages.

The test list contains $1,200$ English single-word terms that were extracted by considering all nouns that occur more than 10 but not more than 200 times and are listed in UMLS. For the Spanish part of the corpus, we considered all nouns as candidate translations ($32,347$ in total).

3.2 Results

Table 1 shows the top-k translation accuracy and the MRR of RF and context vectors.

	Acc_1	Acc_{10}	Acc_{20}	MRR
RF	0.41	0.57	0.59	0.47
Cont. Vectors	0.1	0.21	0.26	0.11

Table 1: top-k translation accuracy and MRR of RF and context vectors on $1,200$ English terms

We observe that the proposed RF method achieves a considerably better top-k translation accuracy and MRR than the baseline method. Moreover, we segmented the $1,200$ test terms into 7 frequency ranges [7], from high-frequency to rare terms. Figure 1 shows the translation accuracy at top 20 candidates for the two methods. We note

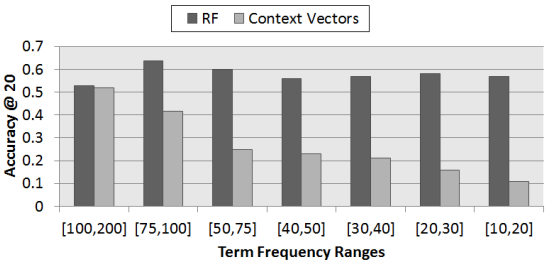

Figure 1: Translation accuracy of top 20 candidates on different frequency ranges

that for high frequency terms, i.e. [100,200] range, the performance achieved by the two methods is similar (53% and 52% for the RF and context vectors respectively). However, for lower frequency terms, the translation accuracy of the context vectors continuously declines. This confirms that context vectors do not behave robustly for rare terms (Morin and Daille, 2010). In contrast, the RF slightly fluctuates over different frequency ranges and presents approximately the same translation accuracy for both frequent and rare terms.

4 Application

As an application of our method, we use the previously extracted dictionaries to on-line augment the phrase table of an SMT system and observe the translation performance on test sentences that contain OOV terms. For the translation probabilities in the phrase table, we use the distance metric given by the dictionary extraction methods i.e., classification margin and cosine similarity of RF and context vectors respectively, normalised by the uniform probability (if a source term has m candidate translations, we normalise the distance metric by dividing by m as in (Wu et al., 2008) .

4.1 Data and tools

We construct a parallel, sentence-aligned corpus from the biomedical domain, following the process described in (Wu et al., 2011; Yepes et al., 2013). The parallel corpus comprises of article titles indexed by PubMed in both English and Spanish. We collect 120K parallel sentences for train-

[4] $MRR = \frac{1}{|Q|}\sum_{i=1}^{Q}\frac{1}{rank_i}$ where $|Q|$ is the number of English terms for which we are extracting translations and $rank_i$ is the position of the first correct translation from returned list of candidates

[5] nlm.nih.gov/research/umls

[6] http://en.wikipedia.org/wiki/Help:Searching

[7] each frequency range contains 100 randomly sampled terms

ing the SMT and 1K sentences for evaluation. The test sentences contain $1,200$ terms that do not appear in the training parallel corpus. These terms occur in the Wikipedia comparable corpus. Hence, the previously extracted dictionaries list a possible translation. Using the PubMed parallel corpus, we train Moses (Koehn et al., 2007), a phrase-based SMT system.

4.2 Results

We evaluated the translation performance of the SMT that uses the dictionary extracted by the RF against the following baselines: (i) Moses using only the training parallel data (Moses), (ii) Moses using the dictionary extracted by context vectors (Moses+context vector). The evaluation metric is BLEU (Papineni et al., 2002).

Table 2 shows the BLEU score achieved by the SMT systems when we append the top-k translations to the phrase table.

	BLEU on top-k translations		
	1	10	20
Moses	24.22	24.22	24.22
Moses+ RF	**25.32**	24.626	24.42
Moses+ Context Vectors	23.88	23.69	23.74

Table 2: Translation performance when adding top-k translations to the phrase table

We observe that the best performance is achieved by the RF when we add the top 1 translation with a total gain of 1.1 BLEU points over the baseline system. In contrast, context vectors decreased the translation performance of the SMT system. This indicates that the dictionary extracted by the context vectors is too noisy and as a result the translation performance dropped. Furthermore, it is noted that the augmented SMT systems achieve the highest performance for the top 1 translation while for k greater than 1, the translation performance decreases. This behaviour is expected since the target language model was trained only on the training Spanish sentences of the parallel corpus. Hence, the target language model does not have a prior knowledge of the OOV translations and as a result it cannot choose the correct translation among k candidates.

To further investigate the effect of the language model on the translation performance of the augmented SMT systems, we conducted an oracle experiment. In this ideal setting, we assume a strong language model, that is trained on both training and test Spanish sentences of the parallel corpus, in order to assign a higher probability to a correct translation if it exists in the deployed dictionary. As we observe in Table 3, a strong language model can more accurately select the correct translation among top-k candidates. The dictionary extracted by the RF improved the translation performance by 2.5 BLEU points for the top-10 candidates and context vectors by 0.45 for the top-20 candidates.

	BLEU on top-k translations		
	1	10	20
Moses	28.85	28.85	28.85
Moses+ RF	30.98	**31.35**	31.2
Moses+ Context Vectors	28.18	29.17	29.3

Table 3: Translation performance when adding top-k translations to the phrase table. SMT systems use a language model trained on training and test Spanish sentences of the parallel corpus.

5 Discussion

In this paper, we presented an RF classifier that is used to extract bilingual dictionaries of technical terms from comparable corpora. We evaluated our method on a comparable corpus of Wikipedia articles. The experimental results showed that our proposed method performs robustly when translating both frequent and rare terms.

As an application, we used the automatically extracted dictionary to augment the phrase table of an SMT system. The results demonstrated an improvement of the overall translation performance.

As future work, we plan to integrate the RF classifier with context vectors. Intuitively, the two methods are complementary considering that the RF exploits the internal structure of terms while context vectors use the surrounding lexical context. Therefore, it will be interesting to investigate how we can incorporate the two feature spaces in a machine learner.

6 Acknowledgements

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*, COLING '10, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

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

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

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.

Martin Kay and Martin Röscheisen. 1993. Text-translation alignment. *computational Linguistics*, 19(1):121–142.

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.

Georgios Kontonatsios, Ioannis Korkontzelos, Jun'ichi Tsujii, and Sophia Ananiadou. 2013. Using random forest to recognise translation equivalents of biomedical terms across languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 95–104. Association for Computational Linguistics, August.

I. Dan Melamed. 1997. A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 305–312. Association for Computational Linguistics.

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. 2011. 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.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. 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.

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.

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.

Hua Wu, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 993–1000. Association for Computational Linguistics.

Cuijun Wu, Fei Xia, Louise Deleger, and Imre Solti. 2011. Statistical machine translation for biomedical text: are we there yet? In *AMIA Annual Symposium Proceedings*, volume 2011, page 1290. American Medical Informatics Association.

Antonio Jimeno Yepes, Élise Prieur-Gaston, and Aurélie Névéol. 2013. Combining medline and publisher data to create parallel corpora for the automatic translation of biomedical text. *BMC bioinformatics*, 14(1):146.