Using a Random Forest Classifier to recognise translations of biomedical terms across languages

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 present a novel method to recognise semantic equivalents of biomedical terms in language pairs. We hypothesise that biomedical term are formed by semantically similar textual units across languages. Based on this hypothesis, we employ a Random Forest (RF) classifier that is able to automatically mine *higher order associations* between textual units of the source and target language when trained on a corpus of both positive and negative examples. We apply our method on two language pairs: one that uses the same character set and another with a different script, English-French and English-Chinese, respectively. We show that English-French pairs of terms are highly transliterated in contrast to the English-Chinese pairs. Nonetheless, our method performs robustly on both cases. We evaluate RF against a state-of-the-art alignment method, GIZA++, and we report a statistically significant improvement. Finally, we compare RF against Support Vector Machines and analyse our results.

1 Introduction

Given a term in a source language and term in a target language the task of this paper is to classify this pair as a translation or not. We investigate the performance of the proposed classifier by applying it on a balanced classification problem, i.e. our experimental datasets contain an equal number of positive and negative examples. The proposed classification model can be used as a component of a larger system that automatically compiles bilingual dictionaries of technical terms across languages. Bilingual dictionaries of terms are important resources for many *Natural Language Processing* (NLP) applications including *Statistical Machine Translation* (SMT) (Feng et al., 2004; Huang and Vogel, 2002; Wu et al., 2008), *Cross-Language Information Retrieval* (Ballesteros and Croft, 1997) and *Question Answering* systems (Al-Onaizan and Knight, 2002). Especially in the biomedical domain, manually creating and more importantly updating such resources is an expensive process, due to the vast amount of *neologisms*, i.e. newly introduced terms (Pustejovsky et al., 2001). The UMLS metathesaurus which is one the most popular *hub* of multilingual resources in the biomedical domain, contains technical terms in 21 languages that are linked together using a *concept* identifier. In Spanish, the second most popular language in UMLS, only 16.44% of the 7.6M English terms are covered while other languages fluctuate between 0.0052% (for Hebrew terms) to 3.26% (for Japanese terms). Hence, these lexica are far for complete and methods that *semi-automatically* (i.e., in a post-processing step, curators can manually remove erroneous dictionary entries) discover pairs of terms across languages are needed to enrich such multilingual resources. Our method can be applied to parallel, aligned corpora, where we expect approximately the same, balanced classification problem. However, in comparable corpora the search space of candidate alignments is of vast size, i.e., quadratic the the size of the input data. To cope with this heavily unbalanced classification problem, we would need to narrow down the number of negative instances before classification.

We hypothesise that there are *language independent* rules that apply to biomedical terms across many languages. Often the same or similar textual units (e.g., morphemes and suffixes) are concatenated to realise the same terms in different languages. For example, Table 1 illustrates how a morpheme expressing *pain* (*ache* in English) is used to realise the same terms in English, Chinese and French. The realisations of the term "head-

English Morpheme: **-ache**	Chinese Morpheme: 痛	French Morpheme: **-mal**
head-**ache**	头-**痛**	**mal** de tête
back-**ache**	腰-**痛**	**mal** au dos
ear-**ache**	耳朵-**痛**	**mal** d'oreille

Table 1: An example of English, Chinese and French terms consisting of the same morphemes

ache" is expected to consist of the units for "head" and "ache" regardless of the language of realisation. Hence, knowing the translations of "head" and "ache" allows the reconstruction "headache" in a target language.

In our method, we use a *Random Forest (RF)* classifier (Breiman, 2001) to learn the underlying rules according to which terms are being constructed across languages. An RF is an ensemble of Decision Trees voting for the most *popular* class. RF classifiers are popular in the biomedical domain for various tasks: classification of microarray data (Díaz-Uriarte and De Andres, 2006), compound classification in cheminformatics (Svetnik et al., 2003), classification of microRNA data (Jiang et al., 2007) and protein-protein interactions in Systems Biology (Chen and Liu, 2005). In NLP, RF classifiers have been used for: Language Modelling (Xu and Jelinek, 2004) and semantic parsing (Nielsen and Pradhan, 2004). To the best of the authors' knowledge, this is the first attempt to employ RF for identifying translation equivalents of biomedical terms.

We prefer RF over other traditional machine learning approaches such as *Support Vector Machines (SVMs)* for a number of reasons. Firstly, RF is able to automatically construct *correlation paths* from the feature space, i.e. decision rules that correspond to the translation rules that we intend to capture. Secondly, RF is considered one of the most accurate classifier available (Díaz-Uriarte and De Andres, 2006; Jiang et al., 2007). Finally, RF is reported to cope well with datasets where the number of features is larger than the number of observations (Díaz-Uriarte and De Andres, 2006). In our dataset, the number of features is almost four times more than that of the observations.

We represent pairs of terms using character gram features (i.e., *first order* features). Such shallow features have been proven effective in a number of NLP applications including: Named Entity Recognition (Klein et al., 2003), *Multilingual Named Entity Transliteration* (Klementiev and Roth, 2006; Freitag and Khadivi, 2007) and predicting authorship (Stamatatos, 2006). In addition, by selecting character n-grams instead of word n-grams, one avoids to segment words in Chinese which has been proven to be a challenging topic (Sproat and Emerson, 2003). We evaluate our proposed method on two datasets of biomedical terms (English-French and English-Chinese) that contain equal numbers of positive and negative instances. RF achieves higher classification performance than baseline methods. To boost SVM's performance further, we used a *second order* feature space to represent the data. It consists of pairs of character grams that co-occur in translation pairs. In the second order feature space, the performance of SVMs improved significantly.

The rest of the paper is structured as follows. In Section 2, we present previous approaches in identifying translation equivalents of terms or named entities. In Section 3, we define the classification problem, we formulate the RF classifier and we discuss the first and second order feature space that we use to represent pairs of terms. In Section 4, we show that RF achieves superior classification performance. In Section 5, we overview our method and we discuss how it can be used to compile large-scale bilingual dictionaries of terms from comparable corpora.

2 Related Work

In this section, we review previous approaches that exploit the internal structure of sequences to align terms or named entities across languages. (Klementiev and Roth, 2006; Freitag and Khadivi, 2007) use character gram features, similar to the feature space that we propose in this paper, to train discriminative, supervised models. Klementiev and Roth (2006) introduce a supervised *Perceptron* model for English and Russian named entities. They construct a character gram feature space as follows: firstly, they extract all distinct character grams from both source and target named entity. Then, they pair character grams of the source named entity with character grams of the corresponding target named entity into features. In or-

der to reduce the number of features, they link only those character grams whose position offsets in the source and target sequence differs by -1, 0 or 1. Freitag and Khadivi (2007) employ the same character gram feature space but they do not constraint the included character-grams to their relative position offsets in the source and target sequence. The *boolean* features are defined for every distinct character-grams observed in the data of length k or shorter. Using this feature space they train an *Averaged Perceptron* model, able to incorporate an arbitrary number of features in the input vectors, for English and Arabic named entities. The above character gram based methods mainly focused on *aligning* named entities of the general domain, i.e. person names, locations, organizations, etc., that are transliterated, i.e. present phonetic similarities, across languages.

SMT-based approaches built on top of existing SMT frameworks to identify translation pairs of terms (Tsunakawa et al., 2008; Wu et al., 2008). Tsunakawa et al. (2008), align terms between a source language L_s and a target language L_t using a pivot language L_p. They assume that two bilingual dictionaries exist: from L_s to L_p and from L_p to L_t. Then, they train *GIZA++* (Och and Ney, 2003) on both directions and they merge the resulting phrase tables into one table between L_s and L_t, using grow-diag-final heuristics (Koehn et al., 2007). Wu et al. (2008), use morphemes instead of words as translation units to train a phrase based SMT system for technical terms in English and Chinese. The use of shorter lexical fragments, e.g. lemmas, stems and suffixes, as translation units has reportedly reduced the *Out-Of-Vocabulary* problem (Virpioja et al., 2007; Popovic and Ney, 2004; Oflazer and El-Kahlout, 2007).

Hybrid methods exploit that a term or a named entity can be translated in various ways across languages (Shao and Ng, 2004; Feng et al., 2004; Lu and Zhao, 2006). For instance, person names are usually *translated by transliteration* (i.e., words exhibiting pronunciation similarities across languages, are likely to be mutual translations) while technical terms are likely to be *translated by meaning* (i.e., the same semantic units are used to generate the translation of the term in the target language). The resulting hybrid systems were reported to perform at least as well as existing SMT systems (Feng et al., 2004).

Lepage and Denoual (2005) presented an analogical learning machine translation system as part of the IWSLT task (Eck and Hori, 2005) that requires no training process and it is able to achieve state-of-the art performance. The core method of their system models relationships between sequences of characters, e.g., sentences, phrases or words, across languages using *proportional analogies*, i.e., $[a : b = c : d]$, "a is to b as c is to d", and is able to solve unknown *analogical equations*, i.e., $[x : y = z :?]$ (Lepage, 1998). Analogical learning has been proven effective in translating unseen words (Langlais and Patry, 2007). Furthermore, analogical learning is reported to achieve a better precision but a lower recall than a phrase-based machine translation system when translating medical terms (Langlais et al., 2009).

3 Methodology

Let $e^m = (e_1, \cdots, e_m)$ be an English term consisting of m translation units and $f^n = (f_1, \cdots, f_n)$ a French or Chinese term consisting of n units. As translation units, we consider character grams. We define a function $f : (e^m, f^n) \longrightarrow \{0, 1\}$:

$$f(e^m, f^n) = \begin{cases} 1, & \text{if } e^m \text{ translates into } f^n \\ 0, & \text{otherwise} \end{cases}$$

The function can be learned by training a *Random Forest (RF)* classifier[1]. Let N be the number of training instances, $|\Omega|$ the total number of features, i.e. the number of dimensions of the feature space, $|\tau|$ a predefined number of random decision trees and $|\phi|$ a predefined number of random features. An RF classifier is defined as a collection of fully grown decision tree classifiers, $\delta_i(X)$ (Breiman, 2001):

$$RF = \{\delta_1(X), \cdots, \delta_\tau(X)\},\ X = (e^m, ch^n) \tag{1}$$

A pair of terms is classified as a *translation* pair if the majority of the trees is voting for this class label. Let $I(\delta_i(X))$ be the vote of the i^{th} tree in the forest and $av_{j\in\{0,1\}}$ the average number of votes for class labels 0 (*translation*) and 1 (*non-translation*). The function f of τ decision trees can be written as the majority function:

$$\begin{aligned} f(e^m, ch^n) &= \text{Maj}\left(I(\delta_1(X)), \cdots, I(\delta_\tau(X))\right) \\ &= \left\lfloor \frac{1}{2} \frac{\sum_1^\tau I(\delta_i(X)) + 1/2(-1)^r}{\tau} \right\rfloor \end{aligned} \tag{2}$$

[1]The WEKA implementation (Hall et al., 2009) of RF was used for all experiments of this paper.

The majority function returns 1 if the majority of $I(\delta_i(X))$ is 1, or returns 0 if the majority of $I(\delta_i(X))$ is 0. Adding or subtracting $1/2$ controls whether a tie is resolved towards 1 or 0, respectively. In RF ties are resolved randomly. To represent this, the negative unit (-1) is raised to a randomly chosen positive integer $r \in \mathbb{N}^+$.

We tuned the RF classifier using 140 random trees and $|\phi| = \log_2 |\Omega| + 1$ features as suggested in Breiman (Breiman, 2001).

The RF mechanism that triggers term construction rules across languages lies in the decision trees. A RF grows a decision tree by selecting the most informative feature, i.e. corresponding to the lowest entropy, out of ϕ random features. For each selected feature, a node is created and this process is repeated for all ϕ random features of the unprunned decision trees. In other words, the process starts with the most informative feature and builds association rules between all random features. These are the construction rules that we are interested in. Figure 1 illustrates a path in one of the decision trees of an RF classifier taken from the experiments we conducted on the English-Chinese dataset. In only one of thousands of branches of the *forest*, the classifier is able to partially trigger the construction rule of *kinase*, a type of enzyme, between English and Chinese. The translation rule correctly associates the English n-grams *kin* and *as* with their Chinese translation 激酶. In addition, the translation rule contains both positive and negative associations between features. The English n-grams *ing* and *or* are negatively correlated with the term *kinase*.

3.1 Feature Engineering

Each pair of terms is represented as a feature vector of character n-grams. We further define two types of character n-gram features, namely *first order* and *second order*. First order character n-grams are boolean features that designate the occurrence of a corresponding character gram of predefined length in the input term. These features are monolingual, extracted separately from the source and target term. The RF classifier is shown to benefit from only monolingual features and achieves the best observed performance. In contrast, SVMs were shown not to perform well using the *first order* feature space because they cannot directly associate the source with the target character grams. To enhance the performance of SVMs, we constructed a *second order* feature space that contains associations between *first order* features. A *second order* feature is a tuple of a source and a target character gram that co-occur in one or more translation pairs. Table 2 illustrates an example. *Second order* character n-grams are multilingual features and are defined over true translation pairs. For this reason, we extract *second order* features from the training data only.

In all experiments, the features were sorted in decreasing order of frequency of occurrence. We trained a RF and two SVM classifiers, namely linear-SVM and RBF-SVM, using a gradually increasing number of features, always starting from the top of the list. SMT frameworks cannot be trained on an increasing number of features because each training instance needs to correspond to at least one known translation unit (i.e., first order features). Therefore, GIZA++ is trained on the complete set of translation units.

4 Experiments

In this section, we discuss the employed datasets of biomedical terms in English-French and English-Chinese and three baseline methods. We compare and discuss RF and SVMs trained on the *first order* and *second order* features. Finally, we report results of all classification methods evaluated on the same datasets.

4.1 Datasets

For our experiments, we used an online bilingual dictionary[2] for English-Chinese terms and the UMLS metathesaurus[3] for English-French terms. The former contains $31,700$ entries while the latter is a much larger dictionary containing $84,000$ entries. For training, we used the same number of instances for both language pairs (i.e., $21,000$ entries) in order not to bias the performance towards the larger English-French dataset. The remaining instances were used for testing (i.e., $10,7000$ and $63,000$ English-Chinese and English-French respectively). In the case where a source term corresponded to more that one target terms according to the seed dictionary, we randomly selected only one translation. Negative instances were created by randomly matching non-translation pairs of terms. Since we are dealing with a balanced clas-

[2] www2.chkd.cnki.net/kns50/

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

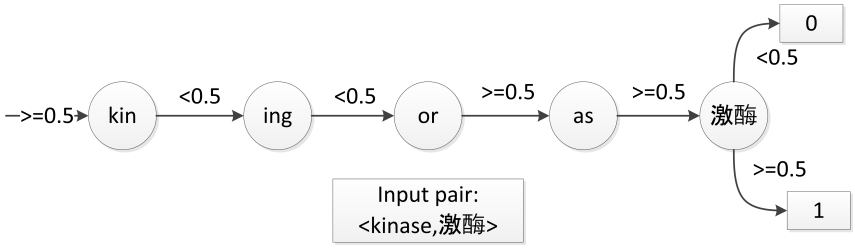


Figure 1: Example of a term construction rule as a branch in a decision tree.

Input pair of English-French terms : $(e_1, e_2, e_3, f_1, f_2, f_3)$		
English first order	French first order	Second order
$\phi_1(e_1, e_2)$	$\phi_1(f_1, f_2)$	$\phi_1(e_1e_2, f_1f_2)$, $\phi_1(e_1e_2, f_2f_3)$
$\phi_1(e_2, e_3)$	$\phi_1(f_2, f_3)$	$\phi_1(e_2e_3, f_1f_2)$, $\phi_1(e_2e_3, f_2f_3)$

Table 2: Example of first and second order features using a predefined n-gram size of 2.

sification problem, we created as many negative instances as the positive ones in all our datasets. In all experiments we performed a 3-fold cross-validation.

4.2 Baselines

We evaluated RF against three classification methods, namely SVMs, GIZA++ and a Levenshtein distance-based classifier.

SVMs coordinate a hyperplane in the hyperspace defined by the features to best separate the positive and negative instances, i.e. aligned from non-aligned pairs. In contrast to RF, SVMs do not support building association rules between features, i.e., translation units, which in our task seems to be a deficiency. SVMs produce one final association rule, i.e. the *classification boundary* which separates positive from negative examples. Its ability to distinguish aligned from non-aligned pair of terms depends on how separable the two clusters are. We evaluated several settings for the SVM classifier. Apart from the default linear kernel function, we applied a radial basis function, i.e. RBF-SVM. RBF-SVM uses the *kernel trick* to project the instances in a higher dimensional space to better separate the two clusters. While tuning the SVM's classification cost *C*, we observed optimal performance for a value of 100. Secondly, we seeded the association rules of translation units to the SVM classifier by creating a *second order* feature space, discussed in detail in section 3.1. We employed the *LIBSVM* implementation (Chang and Lin, 2011) of SVMs using both the linear and RBF kernels.

The second baseline method is GIZA++, an open source implementation of the 5 IBM-models (Brown et al., 1993). GIZA++ is traditionally trained on a bilingual, parallel corpus of aligned sentences and estimates the probability $P(s|t)$ of a source translation unit (typically a word), s, given a target unit t. To apply GIZA++ on our dataset, we consider the list of terms as parallel sentences. GIZA++, trained on a list of terms, estimates the alignment probability of English-Chinese and English-French textual units, i.e. character n-grams. Each entry i, j in the *translation table* is the probability $P(s_i|t_j)$, where s_i and t_j are the source and target character n-grams in row i and column j, respectively. Further details about training a SMT toolkit for aligning technical terms can be found in (Tsunakawa et al., 2008; Freitag and Khadivi, 2007; Wu et al., 2008). After training GIZA++ we estimate the posterior probability $P(cf^n|e^m)$ that a test, Chinese or French term $cf^n = \{cf_1, \cdots, cf_n\}$ is aligned with a given English term $e^m = \{e_1, \cdots, e_m\}$ as follows:

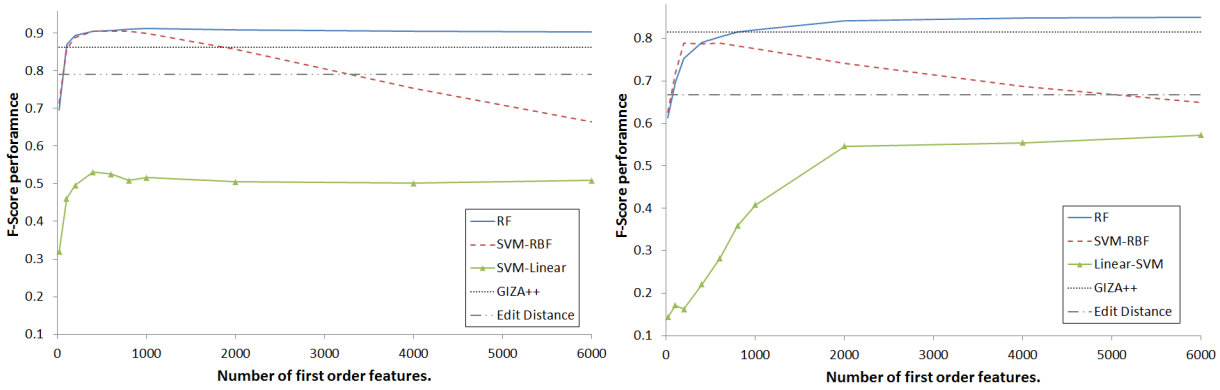
$$p(cf^n|e^m) = n^{-m} \sum_{i=1}^{n} \sum_{j=1}^{m} P(cf_i|e_j) \quad (3)$$

A threshold ξ was defined to classify a pair of terms into *translations* or *non-translations*:

$$f(e^m, cf^n) = \begin{cases} 1, & \text{if } p(cf^n|e^m) \geq \xi \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

We experimented with different values of ξ (*greedy search*) and we selected a value that maximizes classification performance.

In order to estimate how phonetically similar the two language pairs are, we employed a third base-



(a) English-French dataset

(b) English-Chinese dataset

Figure 2: F-Score of the RF and SVM, GIZA++ and Levenshtein distance-based classifier on the *first order* dataset

line method that uses the *Edit/Levenshtein distance* of pairs of terms to classify instances as translations or not. The Levenshtein distance is defined as the minimum edit operations, i.e., insertion, deletions and substitution, required to transform one sequence of characters to another. We cannot directly calculate the Levenshtein distance between English-Chinese pairs of terms since the two languages are using different scripts. Therefore, before we applied the Levenshtein distance-based classifier, we converted the Chinese terms to their *pinyin* form, i.e., Romanization system of Chinese characters. As with GIZA++, we selected a threshold ξ that maximizes the performance of the classifier.

4.3 Results

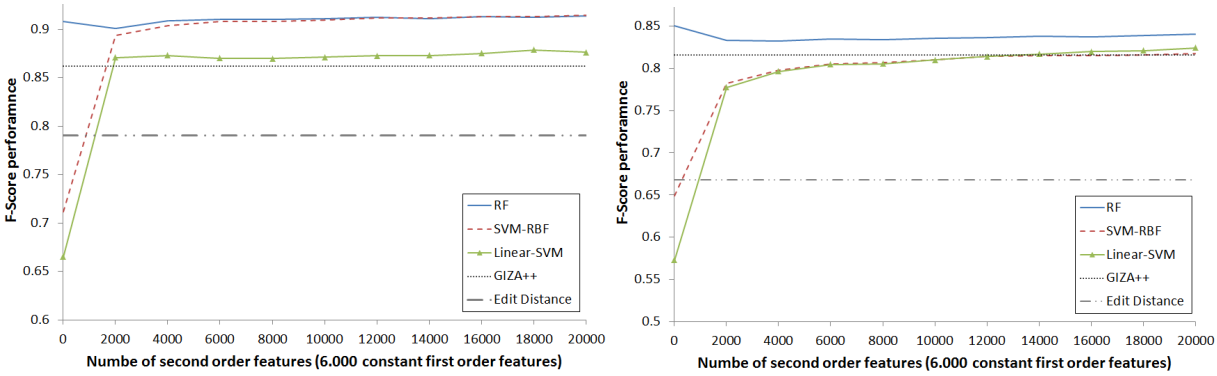
We hypothesise that a RF classifier is able to form association paths between first order features. We also have the theoretical intuition that SVM classifiers are not able to form such association paths. As a result, we expect limited performance on the first order feature set, because it does not contain any associations among character grams.

Figure 2 shows the F-Score achieved by RF, linear-SVM, RBF-SVM, GIZA++ and Levenshtein/Edit distance-based classifier on the English-French and English-Chinese datasets. RF and SVMs are trained on an increasing number of features. The behaviour of the classifiers is approximately the same in both datasets. Performance is greater on the English-French dataset since English is more similar to French than to Chinese.

We also observe that linear-SVM and RBF-SVM do not behave consistently. RBF-SVM's performance quickly climbs to a maximum and afterwards it declines while linear-SVM's performance is constantly increasing until it balances to a very high error rate, almost corresponding to random classification. The linear-SVM classifier performs poorly using *first order* features only, indicating that this feature space is *non-linearly* separable, i.e. there exists no hyperplane that separates *translation* from *non-translation* instances. Contrary, RBF-SVM is able to construct a higher dimensional space by applying the *kernel trick* so as to take full advantage of a small number of frequent and informative *first order* features. In this higher dimensional space of few but informative first order features, the RBF-SVM classifier coordinates a hyperplane that effectively separates positive from negative instances. However, increasing the number of features introduces noise that affects the performance.

The RF is able to profit from larger sets of *first order* features; thus, its performance is continuously increasing until it stabilises at $6,000$ features. The branches of the decision trees are shown to manage features correctly to construct most of the translation rules. Increasing the size of the feature space minimises the classification error, because more translation rules that generalize well on unseen data are constructed.

The bilingual dictionary that we use for our experiments contains heterogeneous biomedical terms of diverse semantic categories. For example, our data-set contains common medical terms such as *Intellectual Products* (e.g. *Pain Management, prise en charge de la douleur,* 控制疼痛) or complex biological concepts such as *Enzymes* (e.g. *homogentisate 1,2-dioxygenase,*



(a) English-French dataset

(b) English-Chinese dataset

Figure 3: F-Score of the RF and SVM, GIZA++ and Levenshtein distance-based classifier on the *second order* dataset

	English-French pairs			English-Chinese pairs		
	P	R	F_1	P	R	F_1
GIZA++	0.901	0.826	0.862	0.907	0.742	0.816
Levenshtein Distance	0.762	0.821	0.791	0.501	**0.990**	0.668
$SVM\text{-}RBF_{second\text{-}order}$	0.946	0.884	0.914	0.750	0.899	0.818
$Linear\text{-}SVM_{second\text{-}order}$	0.866	**0.887**	0.8763	0.765	0.893	0.824
$RF_{first\text{-}order}$	**0.962**	0.874	**0.916**	0.779	0.940	**0.851**

Table 3: Best observed performance of RF, SVM and GIZA++ and Levenshtein Distance

acide homogentisique-oxydase, 尿黑酸*1,2*-双氧酶). Therefore, we would expect poor performance of the supervised methods using only a small portion of the total set of *first order* features due to the high diversity of the terms. For example the morpheme *ache/ mal/* 痛 is more frequent in *Disease or Syndrome* named entities rather than *Enzyme* named entities. However, the results indicate that RF can generalize well on heterogeneous terms. Figure 2 shows that the RF classifier outperforms SMT based methods, using only 1000 features.

The Levenshtein distance-based classifier performs considerably better in the English-French dataset than in English-Chinese. In fact, its best performance for the English-Chinese dataset is achieved when classifying every pair of terms as a translation, i.e. 100% recall but 50% precision.

In a second experiment, we attempted to explore whether the performance of SVMs can be improved by providing cross-language association features. We employed the *second order* feature set discussed in subsection 3.1. We used a constant number of $6,000$ *first order* features, the number of features that achieved maximum F-Score for RF in the previous experiment. Besides these *first order* features, we added an increasing number of *second order* ones. Figure 3 shows the F-Score curves of the RF, linear-SVM, RBF-SVM, GIZA++ and Levenshtein distance using this feature space.

We observe that *second order* features improved the performance of both SVMs considerably. In contrast to the previous experiment, the two SVMs present consistent bevaviour. Interestingly, the performance of the RF slightly decreased when using a small number of *second order* features. A possible explanation of this behaviour is that the *second order* associative features added noise, since the RF had already formed the association rules from *first order* features. In addition, for m English and n Chinese or French *first order* features there were $m \times n$ possible combinations of *second order* features as explained in Subsection 3.1. Hence, there was a large number of *second order* features that we excluded from the training process. Consequently, decision tree branches were populated with incomplete association rules while the RF was able to form these associations automatically. Nevertheless, as more *second order* features were added, more association rules were explored and the RF performance in-

creased. Table 3 summarises the highest performance achieved by the RF, SVMs, GIZA++ and Levenshtein distance all trained and tested on the same dataset. The resulting performance of the RF compared with GIZA++ is statistically significant ($p < 0.0001$) in all experiments. Comparing the RF with the SVMs, we note that in the English-French dataset, the performance of the SVM-RBF is approximately the same with the performance of our proposed method. However, this comes with a cost. Firstly, SVMs can possibly achieve a comparable performance to the RF when using multilingual, second order features. In contrast, our experiments show that RF benefit from monolingual, first order features only. Secondly, SVMs need a large number of additional multilingual features, (6.000 second order features or more) to perform similarly to RF. As a consequence, the resulting models of the SVM classifiers are more complex. We measured the average time needed by the two classifiers to decide for a single pair of terms. The RF is approximately 30 times faster than SVMs (on average 0.010 and 0.292 seconds, respectively). Finally, in the English-Chinese dataset the RF performed significantly better than both SVMs.

5 Discussion And Future Work

In this paper, we presented a novel classification method that uses *Random Forest (RF)* to recognise translations of biomedical terms across languages. Our approach is based on the hypothesis that in many languages, there exist some rules for combining textual units, e.g. n-grams, to form biomedical terms. Based on this assumption, we defined a *first order* feature space of character grams and demonstrated that an RF classifier is able to discover such cross language translation rules for terms. We experimented with two diverse language pairs: English-French and English-Chinese. In the former case, pairs of terms exhibit high phonetic similarity while in the latter case they do not. Our results showed that the proposed method performs robustly in both cases and achieves a significantly better performance than GIZA++. We also evaluated *Support Vector Machines (SVM)* classifiers on the same *first order* feature space and showed that they fail to form translation rules in both language pairs, possibly because it cannot associate *first order* features with each other successfully. We attempted to boost the performance of the SVM classifier by adding association evidence of textual units to the features. We extracted *second order* features from the training data and we defined a new feature set consisting of both *first order* and *second order* features. In this feature space, the performance of the SVMs improved significantly.

In addition to this, we observe from the reported experiments that RF achieves a better F-Score performance than GIZA++ in all datasets. Nonetheless, GIZA++ presents a better precision (but lower recall) in one dataset, i.e., English/Chinese. Based on this observation we plan to investigate the performance of a hybrid system combining RF with MT approaches.

One trivial approach to apply the proposed method for compiling large-scale bilingual dictionaries of terms from comparable corpora would be to directly classify all possible pairs of terms into *translations* or *non-translations*. However, in comparable corpora, the size of the search space is quadratic to the input data. Therefore, the classification task is much more challenging since the distribution of positive and negative instances is highly skewed. To cope with the vast search space of comparable corpora, we plan to incorporate context-based approaches with the RF classification method. Context-based approaches, such as *distributional vector similarity* (Fung and McKeown, 1997; Rapp, 1995; Koehn and Knight, 2002; Haghighi et al., 2008), can be used to limit the number of candidate translations by filtering out pairs of terms with low contextual similarity.

Finally, the proposed method can be also used to *online* augment the *phrase table* of *Statistical Machine Translation (SMT)* in order to better handle the *Out-of-Vocabulary* problem i.e. inability to translate textual units that consist of one or more words and do not occur in the training data (Habash, 2008).

Acknowledgements

The work described in this paper is partially funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 318736 (OSSMETER).

References

Y. Al-Onaizan and K. Knight. 2002. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 400–408. Association for Computational Linguistics.

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

L. Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, and R.L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

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

X.W. Chen and M. Liu. 2005. Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics*, 21(24):4394–4400.

R. Díaz-Uriarte and S.A. De Andres. 2006. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3.

Matthias Eck and Chiori Hori. 2005. Overview of the iwslt 2005 evaluation campaign. In *Proc. of the International Workshop on Spoken Language Translation*, pages 1–22.

D. Feng, Y. Lv, and M. Zhou. 2004. A new approach for english-chinese named entity alignment. In *Empirical Methods in Natural Language Processing*, pages 372–379.

D. Freitag and S. Khadivi. 2007. A sequence alignment model based on the averaged perceptron. In *Conference on Empirical methods in Natural Language Processing*, pages 238–247.

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

N. Habash. 2008. Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 57–60. Association for Computational Linguistics.

A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein. 2008. Learning bilingual lexicons from monolingual corpora. *Proceedings of ACL-08: HLT*, pages 771–779.

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.

F. Huang and S. Vogel. 2002. Improved named entity translation and bilingual named entity extraction. In *International Conference on Multimodal Interaction*, pages 253–258. IEEE.

P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu. 2007. Mipred: classification of real and pseudo microrna precursors using random forest prediction model with combined features. *Nucleic acids research*, 35(suppl 2):W339–W344.

D. Klein, J. Smarr, H. Nguyen, and C.D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, pages 180–183. Association for Computational Linguistics.

A. Klementiev and D. Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 817–824. Association for Computational Linguistics.

P. Koehn and K. Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*, pages 9–16. Association for Computational Linguistics.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. 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.

Philippe Langlais and Alexandre Patry. 2007. Translating unknown words by analogical learning. In *Proceedings of EMNLP-CoNLL*, pages 877–886.

Philippe Langlais, François Yvon, and Pierre Zweigenbaum. 2009. Improvements in analogical learning: application to translating multi-terms of the medical domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 487–495. Association for Computational Linguistics.

Yves Lepage. 1998. Solving analogies on words: an algorithm. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 728–734. Association for Computational Linguistics.

M. Lu and J. Zhao. 2006. Multi-feature based chinese-english named entity extraction from comparable corpora. pages 131–141.

R.D. Nielsen and S. Pradhan. 2004. Mixing weak learners in semantic parsing. In *Empirical Methods in Natural Language Processing*.

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

K. Oflazer and I.D. El-Kahlout. 2007. Exploring different representational units in english-to-turkish statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 25–32. Association for Computational Linguistics.

Maja Popovic and Hermann Ney. 2004. Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. In *4th International Conference on Language Resources and Evaluation (LREC)*, pages 1585–1588, Lisbon,Portugal.

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

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.

L. Shao and H.T. Ng. 2004. Mining new word translations from comparable corpora. In *Proceedings of the 20th international conference on Computational Linguistics*, page 618. Association for Computational Linguistics.

R. Sproat and T. Emerson. 2003. The first international chinese word segmentation bakeoff. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 133–143. Association for Computational Linguistics.

Efstathios Stamatatos. 2006. Ensemble-based author identification using character n-grams. In *In Proc. of the 3rd Int. Workshop on Textbased Information Retrieval*, pages 41–46.

V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston. 2003. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958.

T. Tsunakawa, N. Okazaki, and J. Tsujii. 2008. Building bilingual lexicons using lexical translation probabilities via pivot languages. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may.

S. Virpioja, J.J. Väyrynen, M. Creutz, and M. Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. *Machine Translation Summit XI*, 2007:491–498.

X. Wu, N. Okazaki, T. Tsunakawa, and J. Tsujii. 2008. Improving English-to-Chinese Translation for Technical Terms Using Morphological Information. In *AMTA-2008. MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, pages 202–211, Waikiki, Hawai'i, October.

P. Xu and F. Jelinek. 2004. Random forests in language modeling. In *Empirical Methods in Natural Language Processing*, pages 325–332. Association for Computational Linguistics.