 ISBN: 978-989-95618-4-7

Using SVMs for Filtering Translation Tables for Parallel Corpora Alignment

K. M. Kavitha, Luís Gomes, and Gabriel Pereira Lopes

Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
Quinta da Torre. 2829–516 Caparica, Portugal
k.mahesh@fct.unl.pt luismsgomes@gmail.com
gpl@fct.unl.pt

Abstract. Translation Lexicons are known to improve the quality of parallel corpora alignment at sub-sentence granularity, the quality of newly extracted translations, and as a consequence, Machine Translation and cross language information retrieval. Bilingual pairs (entries) that are part of such translation lexicons should be correct if they are to contribute positively to the improvement of application's quality. This paper proposes and focuses on a method for classifying bilingual entries that were automatically extracted from aligned parallel corpora as correct or incorrect, by using a Support Vector Machine based classifier. Experimental results demonstrate that the classification approach enabled a Micro f-measure higher than 85% for language pair English-Portuguese.

Keywords: Translation equivalents, Translation Lexicon, Translation tables, Bilingual translation pairs, Phrase table filtering, Classification, Support Vector Machine, SVM

1 Introduction

An expression in one language having the same meaning as (or usable in a similar context to) an expression from another language may be referred to as a translation equivalent. Not all automatically extracted translation equivalents should make their way into a bilingual translation lexicon as 'appropriate entries'. For instance, consider the extracted term-pair '*declaration on the* ⇔ *declaração relativa a a*'. The term '*declaration on the*' ends with determiner '*the*' which depends on a noun or noun phrase that is not present. So it makes no sense for the determiner to appear in that position in that entry, otherwise, other entries should also occur, one for each form of Portuguese definite article, '*o*', '*a*', '*os*', '*as*', and many others for taking account all possible determiners that might appear there. Including such terms pairs as good candidates, results in an artificially huge lexicon. However, it would be allowable to include '*declaration on* ⇔ *declaração relativa a*' as it includes agreement information despite the fact that '*on*' may occur in the lexicon having possible translations as '*sobre*', '*relativa a*', '*relativo a*', '*relativos a*' and many others. Again, consider another

example of incorrectly extracted term-pairs '*capacity* ⇔ *capacidade de produção*' and '*commission of the European communities* ⇔ *comissão*'. In the first bilingual pair, the Portuguese word '*produção*' does not have a translation in its English counterpart, while in the second example, '*European communities*' doesn't have an equivalent translation in Portuguese. And so, such term-pairs are questionable candidates to be considered as appropriate entries in a translation lexicon[1].

A common approach for acquiring such a lexicon is based on aligning texts that are translations of each other (parallel texts) [1], [9], [13]. The mainstream strategy for aligning parallel texts [13] is to apply a fully unsupervised machine learning[2] algorithm to learn the parameters (including alignment) of statistical translation models [4], [19]. Naturally, this fully unsupervised learning strategy produces alignment errors, which propagate into the bilingual lexicons extracted from the alignment.

A different strategy is to use a bilingual lexicon to align parallel texts [9] and then extract new[3] term-pairs from the aligned texts [1]. Afterwards, the extracted term-pairs are manually verified and the correct ones are added to the bilingual lexicon, marked as accepted. Incorrect ones are also added to the lexicon marked as rejected. It was this strategy that enabled the construction of the bilingual English-Portuguese translation lexicon with accepted and rejected entries, that was used in this work to train the SVM based classifier. Iterating over these three steps (parallel text alignment, extraction of new translation pairs and their validation) improves the alignment quality [9] and enriches the lexicon, without the risk of decreasing its quality (because of manual validation).

This supervised strategy presents twofold advantages of allowing an improved alignment precision while reducing uncertainty [4], which in turn enables a more accurate extraction of new term-pairs. The verification step is crucial for keeping alignment and extraction errors from being fed back into subsequent alignment and extraction iterations, which would lead the system to degenerate.

In this paper, we propose an automatic classifier that classifies extracted term-pairs as *correct* or *incorrect*, based on a Support Vector Machine (SVM) trained upon a set of manually classified entries. This classification phase, prior to validation, improves validation productivity.

In conventional statistical machine translation systems all phrase pairs that are considerably consistent with the word alignment are extracted and compiled into a phrase table along with their associated probabilities [17] [19]. Such *com-*

[1] A translation lexicon can be simply thought of as a dictionary which contains a term (taken as a single word - any contiguous sequence of characters delimited by white space-, a phrase - contiguous sequence of words-, or a pattern of words or phrases) in the first language cross-listed with the corresponding word, phrase or pattern in the second language such that they share the same meaning or are usable in equivalent contexts

[2] the Expectation Maximization algorithm (EM) to be more specific

[3] by *new* we mean that they were not in the bilingual lexicon that was used for aligning the parallel texts

[4] uncertainty is reduced because a fraction of the aligned phrases is part of the lexicon and thus known to be correct translations

pletely automated training models involve no human supervision and the selection of appropriate translation pairs is only done during the translation process by the decoder. However, studies point out that productivity of the whole translation process can be enhanced by incorporating the human correction activities within the translation process itself [2], thus emphasizing the benefits of supervision. Such an *interactive phrase-based SMT system* is seen to contribute towards a reduced search space unlike the conventional unsupervised models. Again, it is important to note that many of the translations in phrase tables produced are either wrong or will never be used in any translation [12].

The decision on whether or not to incorporate the extracted pair of translation candidates into the bilingual lexicon as an appropriate entry requires judgment. Relying on the evaluation to be done manually, demands that the evaluator has a good knowledge of the languages being dealt with, is time-consuming and thus expensive. As an alternative, prior to human evaluation, an attempt to automatically classify the extracted translation equivalents [1] [13] based on a machine learning approach is proposed in this paper. The experiments presented here are intended to facilitate the validation process by providing the human validator with newly extracted translation pairs automatically classified as correct or incorrect with high precision. Thus the human validation effort becomes lighter and the validation productivity dramatically improves, and may attain 1,000 validated entries per hour per validator, thereby contributing to significantly decrease the time consumed on manual validation. It should be stressed that we are not advocating that just this evaluated bilingual translation lexicon is used for translation. It would certainly decrease translation quality. Our focus is on the improvement of alignment precision and the subsequent extraction accuracy on each cycle of iteration.

1.1 Related Work

The approaches for enhancing the lexicon quality might be viewed as a *filtering process* that discards spurious entries from the lexicon or as a *learning process* that identifies lexicon entries as being correct or incorrect based on examples. Our proposal falls down in the latter approach, wherein, each pair of automatically extracted translation equivalent is classified into one of the pre-defined *accepted* or *rejected* categories.

Filtering Approaches Filter-based approach for enhancing statistical translation models by inducing N-best translation lexicons with non-statistical sources of information is introduced in [18]. A cascade of non-statistical filters is used based on particular knowledge sources such as part of speech information, machine-readable bilingual dictionaries (MRBDs), cognate and word alignment heuristics to remove inappropriate pairs from consideration. The effectiveness of each of the cognate, part of speech, MRBD and word alignment filters is discussed to respectively depend on the particular pair of languages under consideration, the availability of part of speech taggers for both languages, the extent to which the

vocabulary of the MRBD intersects with the vocabulary of the training text, and model of typical word alignments between the pair of languages in question [18].

The use of automatic evaluation filter for discarding the most unlikely translation candidates extracted from parallel corpora may be found in [22]. Several approaches are discussed, namely, the length based filter (using the length difference ratio), similarity filter (based on the comparisons of similarity scores between the most likely translation and alternative candidates), frequency based filters (using absolute and co-occurrence frequencies) and subset filter (for discarding a translation candidate completely included within another candidate). Also, the possibility of combining these filters so as to have separate approaches for identification of most likely translations and for comparing alternative translations with the most likely candidate is stated. In our experiments, the similarity and frequency based features have been used as baseline.

Aires et al. [1] discuss two frequency based scoring functions to filter bad entries extracted from aligned parallel corpora. The scoring functions are developed mainly using source, target and matching frequencies of translation equivalents and are based on the observation that most of the wrong translations revealed considerable differences between those properties. The scoring functions are evaluated for a set of thresholds and the f-measure results obtained varied from 70-82% for correct entries while it varied from 43-60% for incorrect entries.

As far as the state-of-the-art of enhancing quality of phrase tables are concerned, [7] highlight the significance of association scores between phrase-pairs in parallel corpora and utilize them as feature functions to enhance the phrase translation model. Other features used for the same purpose are, the tf-idf term weights for choosing phrase pairs containing infrequent words [25], word-based co-occurrence scores for re-ranking n-best list of translations [6], significance testing of phrase pair co-occurrence with chosen threshold for removal of unlikely translation pairs [12] and the statistical independence measure namely Noise, for filtering phrase tables in Statistical Machine Translation System [23].

Approaches based on Support Vector Machines (SVM) SVM, introduced by [24] is a learning machine based on the Structural Risk Minimization principle and mapping of input vectors into high-dimensional feature space. Adequate feature identification that appropriately represent the knowledge implicit in data is fundamental to enable good learning. SVM had been successfully used for translation related tasks such as learning translation model for extracting word sequence correspondences (phrase translations) [20] and automatic annotation of cognate pairs [3].

The only work we know employing SVMs for selecting appropriate entries into a dictionary from aligned expressions is presented in [15]. This work targets at complex proper noun phrases of the English-Japanese pair defined as proper noun phrases with prepositional phrases and/or co-ordinated phrases. They use SVM for constructing the selection model by taking as features, the common and the different parts between a current translation and a new trans-

lation. Morphemes, part of speech, semantic markers obtained by consulting EDR concept dictionary, and upper-level semantic markers are used as means for representing linguistic information and the features are generated by applying UNIX command 'diff' to the two translations represented in the above mentioned forms and an evaluation of their effect on selection performances is studied. Comparative studies depicted in the paper show that representation by morphemes provided the best f-measure of 0.803. In the experiments reported in this paper, we introduce the concept of *translation mis-coverage* of bilingual translation entries that may compare with the common and difference feature proposed by [15]. Translation mis-coverage is learnt considering both the source and target sides of the bilingual pair [8].

We propose the pre-validation phase of automatically extracted lexicon entries as a classification task, using a supervised learning technique to learn a classifier based on Support Vector Machines (SVM), that assigns the extracted term pairs to one of the predefined good or bad classes. The experiments performed consider language pairs English-Portuguese (EN-PT).

This paper is organized as follows. Section 2 introduces classification as a technique for selecting the bilingual pairs. Section 3 discusses the experiments carried out. The evaluation of the results obtained by the classifier trained using different data sets and features are discussed in Section 4. Finally, in section 5, we conclude with the most relevant results obtained and elaborate a bit over the future work.

2 Validation as a Classification Task

2.1 Data set

We considered classification of translation candidates that were extracted from aligned parallel corpora [5] [21] for language pairs EN-PT. Data consisted of a set of bilingual pairs representing terms in first language and its equivalent in second language, collected from an existing multi-lingual lexicon whose entries are manually tagged as being accepted (positive examples), rejected (negative examples) or left unverified. As was discussed earlier, the alignment and extractions follow the procedures as those proposed in [9] and [1] respectively.

2.2 Features

The features used in the learning process include base properties of translation equivalents, viz., the frequency of term X in first language(F_X), frequency of term Y in second language (F_Y) and matching (or co-occurrence) frequency (F_{XY}), all of which are estimated from the aligned parallel corpus. Two terms are said to co-occur if they are found in segments that have been aligned with each other according to the method proposed in [9]. Features derived using these frequencies, such as, the Dice coefficient of frequencies, the ratio of co-occurrence

[5] JRC-Acquis multilingual parallel corpus, sentence-aligned (22 languages)

frequency to source term frequency, ratio of the co-occurrence frequency to target term frequency, minimum to maximum frequency ratio are used as features in the baseline experiments. Orthogonal similarity features based on Levenshtein edit distance [16], longest common subsequence (LCS) [18], longest common prefix (LCP) [14] and length ratio are quantified as measurable characteristics and used as feature values to identify cognates. Each of these features are normalized by length of the longest terms in the bilingual pair under consideration and are respectively calculated using the equations listed below, where *Len* denotes the length, *MinLen* and *MaxLen*, the minimum and maximum length and *EditDist*, the edit distance used.

$$EditSim = 1 - EditDist(X,Y)/MaxLen(X,Y) \tag{1}$$

$$LCSR = Len(LCS(X,Y))/MaxLen(X,Y) \tag{2}$$

$$LCPR = Len(LCP(X,Y))/MaxLen(X,Y) \tag{3}$$

$$LENR = MinLen(X,Y)/MaxLen(X,Y) \tag{4}$$

Determiners and Co-ordinated Conjunctions Binary-valued features discriminating translation pairs ending with a determiner are used as additional features. Terms ending with determiners such as, '*a*', '*the*', '*certain*' etc., in EN and '*os*', '*uma*' in PT may not be considered as adequate candidates in the lexicon as was discussed in section 1. As '*a*' in PT can be a determiner or a preposition, in order to discriminate between them, including preposition prior to the determiner enabled greater precision on what we were willing to sign down. This knowledge was incorporated as a new feature that represents whether or not, EN terms end with words such as '*a*', '*an*', '*some*', '*one*', '*certain*', '*other*', '*those*', '*and*' etc., and PT terms end with '*o*', '*os*', '*as*', '*uma*', '*uns*', '*umas*', '*este*', '*esta*', '*estes*', '*estas*', '*algum*', '*alguma*', '*alguns*', '*algumas*', '*por a*', '*de a*', '*a a*', '*após a*', '*com a*', '*até a*', '*contra a*', '*desde a*', '*perante a*', '*em a*', '*outro*', '*outra*', '*outros*', '*outras*', '*aqueles*', '*aquela*', '*aquelas*', '*e*'. Co-ordinated conjunctions such as '*and* $\Leftrightarrow$ *e*', were included in each set of specific endings. For a term in each language, the feature value is set to 1 if the term ends with any of the patterns from the corresponding pattern list and to 0 otherwise.

Translation Mis-coverage As was discussed in section 1, the lexicon consisted of bilingual pairs wherein a term in one language differed from their counterparts (translations) by the *number of content words*. This asymmetry with respect to the content words in one language relative to other language is indicated as two additional features. Missing counterparts (translations) in second language for sub-expression in first language (and vice-versa) render an impression of such bilingual entries being bad candidate for translation pairs. This clue was used to

indicate that sub-expressions in one language may or may not have equivalents in the other language. These features indicating translation mis-coverage for terms in the bilingual pair are used with an intuition to correctly identify examples belonging to bad class. The features specify the *existence of translation gaps* or *missing translation segments* (or *mis-coverage* as it was named in [8]) in each of the first and second language terms, where a gap characterizes a sub-expression of the term in one language for which there is no known translation equivalent in the term of the other language. If both the terms in bilingual pair have full coverage, in the sense that, term in one language has a translation in another language taken *in its entirety* or *in constituent sub-expressions*[6] and vice-versa, then the feature value corresponding to each term in bilingual pair is assumed to have a value 0. There is no mis-coverage. But, if one of the terms in the bilingual pair doesn't have a translation then we represent this with the feature value for corresponding term set to 1. There is some mis-coverage. The procedure for identifying translation gaps follows Aho-corasick set-matching algorithm [11] that checks if the terms in the key-word tree (constructed from the bilingual training data separately for EN and PT terms) occur as sub-expressions in the bilingual pair to be validated and if they occur are accepted translations.

Term_{EN}	Term_{PT}	Gap_{EN}	Gap_{PT}
preliminary runs	ensaios preliminares	0	0
traditions and systems	tradições e os sistemas	0	0
vehicles crossing austria	que atravessam a áustria	1	0
training *schemes*	formação	1	0
violence	violência *doméstica*	0	1
union	*disposição de a* união	0	1
watertight *compartment*	*compartimento* estanque	1 (0)	1 (0)
rendering	*revestimento aplicado sobre isolante*	0.5	0.5
recollections	*recordações*	0.5	0.5
accrued	*vencidas*	0.5 (0)	0.5 (0)
accrued	vencido	0	0
accrued	vencida	0	0
watertight	estanque	0	0
compartments	compartimentos	0	0

Table 1. Example of features indicating translation coverage for *EN-PT*

Table 1 illustrates the feature values[7] representing coverage for EN-PT bilingual pairs. The first two examples with gap_{EN}, gap_{PT} set to 0 illustrate correctly translated bilingual pairs having complete coverage. However, the bilingual can-

[6] We look for words translating as multi-words, or multiwords translating as multi-words

[7] Sub-expressions/expressions in italics indicate segments for which translations are missing in other language. Values in parenthesis represent feature values after re-processing. The last 4 translation pairs represent positive training examples

didate '*vehicles crossing austria ⇔ que atravessam a áustria*' is an example of incorrect bilingual entry. We may see that, no translation exists for the English sub-expression '*vehicles*' in Portuguese. Hence we indicate this missing translation for '*vehicles*' by gap_{EN} set to 1. However, looking for coverage from the right-hand side term, we have gap_{PT} set to 0 as '*que atravessam ⇔ crossing*' and '*a áustria ⇔ austria*'. This asymmetry can be seen in the following three examples '*training schemes ⇔ formação*' (where '*schemes*' has no translation in Portuguese counterpart), '*violence ⇔ violência doméstica*' (where '*doméstica*' has no translations in the English counterpart) and '*union ⇔ disposição de a união*' (where '*disposição*' has no translation counterpart in the English side) accordingly setting the feature values with respect to PT sub-expressions and EN sub-expressions.

While looking for coverage, we ignore translations for words of shorter length such as *de, a* in PT. In other words, we neglect missing translations for sub-expression that are not content words, provided they are encapsulated between content words that have translations. For example in the translation pair '*attendance allowance ⇔ subsídio de assistência*', to be validated, if it was learnt that '*attendance ⇔ assistência*' and '*allowance ⇔ subsídio*', we consider the translation candidate to have a full coverage and hence gap_{EN} and gap_{PT} are both set to 0.

If both gap_{EN} and gap_{PT} are set to 1, we check if the sub-expressions indicating missing translations in English and Portuguese parts are *possible translations*. This is achieved by using the stemmed versions of the lexicon for accepted English and Portuguese terms. For instance, in the pair '*watertight compartment ⇔ compartimento estanque*' gap_{EN} and gap_{PT} are set to 1, as the translation '*compartment ⇔ compartimento*' does not appear in the lexicon of accepted pairs used for training. However, as the pair '*compartments ⇔ compartimentos*' exists in the training data as an accepted entry and as their roots appear as longest prefix for '*compartment*' and '*compartimento*', the feature values are reset to 0. In general, the values for gap_{EN} and gap_{PT} are reset to 0 if the stemmed versions of the accepted term pairs appear as prefixes of the sub-expression pairs indicating missing translations. If no match is found, or if at least one word is left out without a translation, the original values for gap_{EN} and gap_{PT} are retained.

To deal with situations where the expressions on either sides are not covered by the lexicon, we set the features gap_{EN} and gap_{PT} to 0.5, which is a neutral value reflecting our lack of support for deciding to accept or to reject that pair. All such pairs are also subjected to further processing to select from among them, those entries that might represent correct translations as for example with the pair '*accrued ⇔ vencidas*'. As explained in the previous paragraph, the stemmed training data is used. Additionally, orthogonal similarities of such expressions are taken into consideration in deciding the correctness using a similarity measure based on the Edit distance between words under consideration. The similarity between words (SpSim) is computed as in equation 1, but discounting the characteristic spelling differences that were learnt previously [10]. Examples of such

spelling differences include (ph — f) and (on — ão) found in English-Portuguese cognates as for example (phase — fase) and (photon — fotão) .

3 Experiments

Experiments were carried out by varying the size of the training data set. SVM based tool namely LIBSVM [5] was used to learn the classifier, which tries to find the hyperplane that separates the training examples with the largest margin. Data was scaled in range [0 1]. In the experiments discussed, the radial basis function (RBF) kernel, with parameters (g, C) shown in table 3 was used. The values presented for g and C reflect the best cross-validation rate.

About 90% of the term-pairs labeled as accepted are used as positive examples and 90% of the term-pairs labeled as rejected form the negative examples in the training data set. The remaining 10% of term pairs belonging to each class constitute the test set.

Data Set	Positive examples	Negative examples
Training	134,448	125,659
Test	14,939	13,962

Table 2. Overview of the Training and Test data set

Five different training data sets (containing 10,000, 25,000, 50,000, 100,000 positive and negative examples each and with the entire training set) (see results in table 3) were constructed from the training data presented in table 2. Training was performed by randomly considering equal number of examples belonging to positive and negative classes and with *entire data* in the training data set (unequal number of positive and negative examples) presented in table 2.

4 Discussion

The classifier results were evaluated with Precision (P), Recall (R) and Accuracy for accepted (Acc) and rejected (Rej) translation pairs, which were computed as given below:

$\text{Precision}_{Acc} = t_p/(t_p + f_p)$
$\text{Precision}_{Rej} = t_n/(t_n + f_n)$
$\text{Recall}_{Acc} = t_p/(t_p + f_n)$
$\text{Recall}_{Rej} = t_n/(t_n + f_p)$
$\text{Accuracy} = (t_p + t_n)/(t_p + f_p + t_n + f_n)$

where, t_p is the number of terms correctly classified as *accepted*, t_n is the number of terms correctly classified as *rejected*, f_p is the number of *incorrect* terms misclassified as *accepted* and f_n is the number of *correct* terms misclassified as *rejected*. The precision, recall and accuracy of the classification for each of the classes over various data sets are as shown in figure 1.

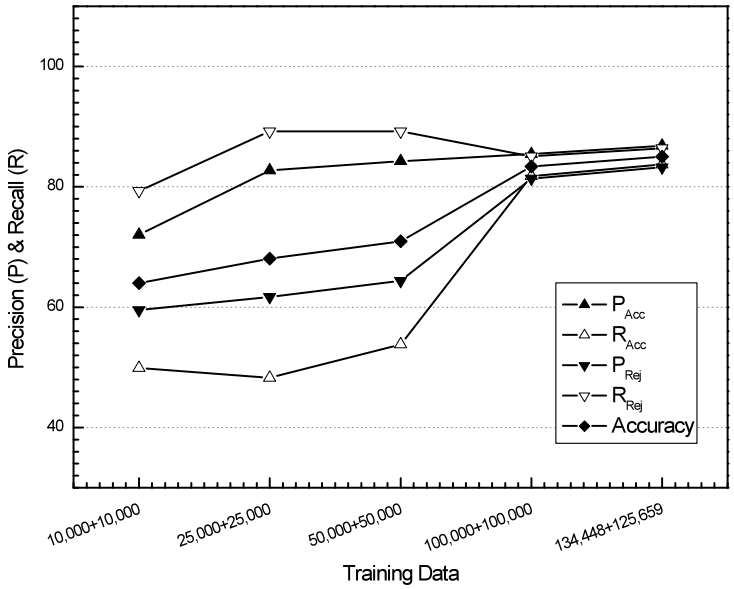

Fig. 1. Precision and Recall for different classes

Also, in order to assess the global performance over both classes, Micro-average Recall (μ_R), Micro-average Precision (μ_P), and Micro-average f-measure (μ_F) are used, and calculated as given below.

μ_P = (Precision$_{Acc}$ + Precision$_{Rej}$) / 2

μ_R = (Recall$_{Acc}$ + Recall$_{Rej}$) / 2

μ_F = 2 * μ_P * μ_R / (μ_P + μ_R)

Table 3 shows the μ_P, μ_R and μ_F obtained in classifying bilingual pairs together with the chosen kernel function and corresponding parameter values for different training sets. The classification approach discussed above, enabled a f-measure of 85.06% which is higher compared to the results attained with the scoring functions used in [1].

Training Data	Kernel = RBF	Type = C-SVC	μ_P	μ_R	μ_F	Accuracy
(Positive + Negative)	Gamma (g)	Cost (C)				
10,000 + 10,000	.125	32	65.80	64.52	65.15	64.02
25,000 + 25,000	.125	32	72.22	68.74	70.44	68.05
50,000 + 50,000	.5	32	74.32	71.54	72.90	70.94
100,000 + 100,000	.5	32	83.45	82.39	82.92	83.39
134,448 + 125,659	.5	32	85.04	85.08	85.06	85.03

Table 3. Performance results for different training data sets

The accuracy of estimated classifier for predicting classes using various features are presented in table 4 for the EN-PT pair trained with all positive and negative examples not in the test set. The information indicating term pairs ending in determiners and co-ordinated conjunctions proved beneficial in discarding unproductive bilingual pairs that would otherwise contribute to a huge lexicon. A rough alignment based method looking for translation coverage from either sides of the bilingual pair provided significant improvement in discriminating the classes. The underlying notion was to utilize the available knowledge about highly reliable translation pairs in deciding if newly extracted bilingual pairs are correct. Using the stemmed lexicon of accepted bilingual pairs (stemmed positive examples) in further processing the segments representing translation gaps, a remarkable overall improvement (almost 10%) was observed in the classification results.

Features	μ_P	μ_R	μ_F	Accuracy
Orthogonal Similarity + Frequency (Baseline)	54.13	53.95	54.04	54.30
Baseline + Determiners + Co-ordinated Conjunctions	67.10	66.73	66.91	66.96
Baseline + Determiners + Co-ordinated Conjunctions + Translation Coverage	75.47	74.93	75.19	75.16
Baseline + Determiners + Co-ordinated Conjunctions + Translation Coverage + Reprocessing of gaps	85.04	85.08	85.06	85.03

Table 4. Performance of classifier on EN-PT bilingual pairs for different features over the entire training data set

5 Conclusion

In this paper, we have introduced the classification process as a methodology for classifying bilingual translation pairs extracted from aligned parallel corpora. The motivation for the work reported in this paper is the fact that automatically extracted translation equivalents after human validation are used for iteratively aligning, extracting and validating new translation pairs. Moreover, evaluation of extracted translation equivalents depends heavily on the human evaluator, and hence incorporation of an automated filter for appropriate and inappropriate translation pairs prior to human evaluation tremendously augments the productivity of this work, attaining 1,000 entries validated per hour per validator, thereby saving the time involved and progressively improving alignment and extraction quality, and hence contributing to improve translation quality. Examples of manually validated adequate and inadequate translation entries also augmented. As seen in the results shown, the larger the training set the quality

of trained classifier improved. For the language pair EN-PT, the accuracy in classifying automatically extracted bilingual pairs is seen to be over 85%.

Our technique for deciding on appropriate entries in a translation lexicon using Support Vector Machine based classifiers considers as features the source term, target term and co-occurring frequencies as a measure to validate the bilingual pairs. Additionally, looking for translation coverage in the translation pairs and for term pairs ending with determiners and co-ordinated conjunctions, the features reflected an overall improvement in classification results by improving the precision for both classes. Further, by re-processing the missing translation segments using stemmed training data the micro-average f-measure improved by approximately 10% when compared to the mere identification of translation gaps.

In future, we intend to extend this technique of machine learning, to classify the translation equivalents extracted from distant language pairs such as English-Hindi. The use of suffix-based features will be examined. Experiments using lexicon extraction and our validation philosophy using Moses [13] by training it using a large parallel corpora and validated bilingual lexicons need to be done. Comparison with the usual approach using just Moses and parallel corpora aligned at sentence level are to be reported.

Acknowledgements K. M. Kavitha and Luís Gomes gratefully acknowledge the Research Fellowship by FCT/MCTES with Reference nos., SFRH/BD/64371/2009 and SFRH/BD/65059/2009, respectively. The authors would like to acknowledge VIP Access project (Ref. PTDC/PLP/72142/2006) and ISTRION project (Ref. PTDC/EIA-EIA/114521/2009) funded by FCT/MCTES that provided other means for the research carried out.

References

1. J. Aires, G. P. Lopes, and L. Gomes. Phrase translation extraction from aligned parallel corpora using suffix arrays and related structures. *Progress in Artificial Intelligence*, pages 587–597, 2009.
2. S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. Lagarda, H. Ney, J. Tomás, E. Vidal, et al. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28, 2009.
3. S. Bergsma and G. Kondrak. Alignment-based discriminative string similarity. In *Annual meeting-Association for Computational Linguistics*, volume 45, page 656, 2007.
4. P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, and R.L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.
5. Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
6. B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico. The ITC-irst SMT system for IWSLT-2005. *Proceeding of IWSLT*, pages 98–104, 2005.
7. B. Chen, G. Foster, and R. Kuhn. Phrase translation model enhanced with association based features. *Proceedings of MT-Summit XII*, 2009.

8. J. Costa, L. Gomes, G.P Lopes, and L.M.S. Russo. Managing and querying a bilingual lexicon with suffix trees. In *EPIA 2011*. APPIA, Portuguese Association for Artificial Intelligence, 2011. To be published in this volume.
9. L. Gomes. Parallel texts alignment. In *New Trends in Artificial Intelligence, 14th Portuguese Conference in Artificial Intelligence, EPIA 2009*, Aveiro, October 2009.
10. L. Gomes and G. P. Lopes. Measuring spelling similarity for cognate identification. In *Progress in Artificial Intelligence*, Lecture Notes in Computer Science. Springer-Verlag, October 2011.
11. D. Gusfield. *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge Univ Pr, 1997. pages 52–61.
12. J.H. Johnson, J. Martin, G. Foster, and R. Kuhn. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2007.
13. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 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, 2007.
14. G. Kondrak. Cognates and word alignment in bitexts. In *Proceedings of the 10th Machine Translation Summit*, pages 305–312, 2005.
15. T. Kutsumi, T. Yoshimi, K. Kotani, I. Sata, and H. Isahara. Selection of entries for a bilingual dictionary from aligned translation equivalents using support vector machines. In *Proceedings of Pacific Association for Computational Linguistics*, volume 2005, 2005.
16. V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710, 1966.
17. A. Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49, 2008.
18. I.D. Melamed. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 184–198. Boston, MA, 1995.
19. F.J. Och and H. Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, 2004.
20. K. Sato and H. Saito. Extracting Word Sequence Correspondences Based on Support Vector Machines. *Journal of Natural Language Processing*, 10(4):109–124, 2003.
21. R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142–2147, 2006.
22. J. Tiedemann. Extraction of translation equivalents from parallel corpora. In *Proceedings of the 11th Nordic conference on computational linguistics*, pages 120–128, 1998.
23. N. Tomeh, N. Cancedda, and M. Dymetman. Complexity-based phrase-table filtering for statistical machine translation. 2009.
24. V. Vapnik. The Nature of Statistical Learning Theory. *Data Mining and Knowledge Discovery*, pages 1–47, 2000.
25. B. Zhao et al. Phrase pair rescoring with term weightings for statistical machine translation. 2004.