Using Bilingual Segments in Generating Word-to-word Translations

K. M. Kavitha[1,3] **Luís Gomes**[1,2] **José Gabriel Pereira Lopes**[1,2]
[1]NOVA Laboratory for Computer Science and Informatics (NOVA LINCS)
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
2829-516 Caparica, Portugal.
luismsgomes@gmail.com gpl@fct.unl.pt
[2]ISTRION BOX-Translation & Revision, Lda., Parkurbis, Covilhã 6200-865 Portugal.
[3] Department of Computer Applications, St Joseph Engineering College
Vamanjoor, Mangaluru, 575 028, India.
kavitham@sjec.ac.in

Abstract

We defend that bilingual lexicons automatically extracted from parallel corpora, whose entries have been meanwhile validated by linguists and classified as correct or incorrect, should constitute a specific parallel corpora. And, in this paper, we propose to use word-to-word translations to learn morph-units (comprising of bilingual stems and suffixes) from those bilingual lexicons for two language pairs L1-L2 and L1-L3 to induce a bilingual lexicon for the language pair L2-L3, apart from also learning morph-units for this other language pair. The applicability of bilingual morph-units in L1-L2 and L1-L3 is examined from the perspective of pivot-based lexicon induction for language pair L2-L3 with L1 as bridge. While the lexicon is derived by transitivity, the correspondences are identified based on previously learnt bilingual stems and suffixes rather than surface translation forms. The induced pairs are validated using a binary classifier trained on morphological and similarity-based features using an existing, automatically acquired, manually validated bilingual translation lexicon for language pair L2-L3. In this paper, we discuss the use of English (EN)-French (FR) and English (EN)-Portuguese (PT) lexicon of word-to-word translations in generating word-to-word translations for the language pair FR-PT with EN as pivot language. Generated translations are filtered out first using an SVM-based FR-PT classifier and then are manually validated.

1 Introduction

Translation lexicon coverage is one of the crucial factors influencing effective Machine Translation. To fill in the gap corresponding to certain missing translation pairs and/or to overcome the difficulties in acquiring translation lexicons for under-resourced language pairs, one can combine the already available bilingual knowledge bases using a common language referred to as pivot and hence automatically expand the translation coverage. Although the pivoted approach to lexicon induction is not new, the novelty of our approach lies in the use of bilingual morph-units rather than the surface translation forms.

We depart from 3 bilingual lexicons (EN-PT, EN-FR and FR-PT) that were automatically acquired from aligned parallel corpora using various extraction techniques (Brown et al., 1993; Lardilleux and Lepage, 2009; Aires et al., 2009; Gomes and Lopes, 2011), whose entries were classified as correct or incorrect by linguists making use of a bilingual concordancer (Costa et al., 2015). These lexicons will hence and along the paper be named as *validated bilingual lexicons*. To be specific, we discuss the use of EN-FR and EN-PT validated bilingual lexicons in inducing bilingual pairs for FR-PT with EN as the pivot language.

The task of bilingual lexicon augmentation is approached in two phases involving pivoted induction and binary classification for subsequent validation of the induced pairs, followed by manual validation. One of the concerns with pivoted translation induction is the generation of wrong translations primarily due to polysemy and ambiguous words. Hence, prior to human validation, for selecting the induced translations in FR-PT, an automatic filter in the form of an SVM-based binary classifier trained on the validated FR-PT bilingual lexicon is used. For every pair of newly induced morph-units/words in the first phase, the next phase involves determining whether the two are translations of each other or not.

Proceedings of the Sixth Workshop on Hybrid Approaches to Translation,
pages 61–71, Osaka, Japan, December 11, 2016.

2 Related Work

While the idea of using pivot language(s) for deriving bilingual lexicon is not new, the approaches differ with respect to the resources employed (Ács, 2014) (Wushouer et al., 2014b), languages dealt (Saralegi et al., 2012) (Wushouer et al., 2013) and the post-processing operations involved in selecting unambiguous and correct translations (Tanaka and Umemura, 1994) (Kaji and Erdenebat, 2008) (Shezaf and Rappoport, 2010). Earliest reported work on pivoted dictionary induction is credited to Tanaka et al. (Tanaka and Umemura, 1994), who proposed the Inverse consultation (IC) approach for pruning wrong translation candidates. Paik et al. (Paik et al., 2004) argued on the importance of directionality in automating the dictionary building process. His experiments are based on one-time inverse consultation method earlier proposed by Tanaka et al. (Tanaka and Umemura, 1994), the overlapping constraint method for improved equivalent pair extraction rate and the POS-based sorting of newly linked pairs to avoid polysemous entries.

In an exclusive analysis of the techniques used to filter wrong translation candidates induced by pivoting, Saralegi et al. (Saralegi et al., 2011) explored two of the common choices, namely the Inverse consultation (IC) method (Tanaka and Umemura, 1994) and the Distributional Similarity measure (DS) (Kaji and Erdenebat, 2008). An outcome of their analysis is that, IC relies on large number of lexical variants in the dictionaries for each sense in the pivot language. Further, given that DS identifies as translations those words exhibiting similar distributions or contexts across two corpora of different languages, it is learnt that, richer context representations and the translation quality of contexts contribute to its improved performance. Union and linear combination of IC and DS outperforms each of these measures taken individually. The authors (Saralegi et al., 2012) thereon discuss the applicability of these methods in building a Basque-Chinese dictionary via English. In a heuristic based approach, Wushouer et al. (Wushouer et al., 2013) explores the use of probability, semantics and spelling similarity heuristics for inducing one-to-one mapping dictionary of Uyghur and Kazakh languages from Chinese-Uyghur and Chinese-Kazakh dictionaries.

In a different study, transitive lexicon induction is centred on multilingual lexical databases such as, lexicon of language-specific word variants, lexemes and collocations, with the validation of new pairs achieved through parallel corpus consultation (Nerima and Wehrli, 2008). Another research (Ács, 2014) on augmenting existing dictionaries in multiple languages, relies on Wiktionary for exploring the links between translations. While exploiting the fact that pairs found via several pivot languages are more precise than those found via one (Tanaka and Umemura, 1994), Ács (Ács, 2014) proposes to extend IC (Tanaka and Umemura, 1994) from using single pivot up to 53 pivots (Ács, 2014). Addressing the task as an optimisation problem, Wushouer et al. (Wushouer et al., 2014a; Wushouer et al., 2014b), proposed extended constraint optimisation model, formalised on Integer Linear Programming for pivot-based dictionary induction of closely related languages by employing multiple dictionaries.

In each of the afore-mentioned approaches, new correspondences are induced by exploiting surface translation forms in two language pairs with one or more language(s) as bridge. In contrast, our approach deviates from transitive induction scheme discussed above with respect to the resources employed in learning correspondences. A specific distinction in our approach is that the resources used for pivoting consist of bilingual morph-units learnt using the bilingual learning method (Karimbi Mahesh et al., 2014a), unlike the traditional surface translation forms. To be specific, for each language pair the knowledge base employed in the experiments consists of bilingual stems, bilingual suffixes as explained in Section 3 and illustrated in Table 1.

3 Background - Bilingual Segments as Knowledge Base

Fundamental to the pivoted induction strategy are the bilingual resources comprising of *bilingual stems* and *bilingual suffixes* learnt from validated bilingual lexicons for the language pairs EN-FR and EN-PT extracted from the aligned parallel corpora[1] using various extraction techniques (Brown et al., 1993; Lardilleux and Lepage, 2009; Gomes, 2009; Aires et al., 2009; Gomes and Lopes, 2011). The methods proposed by Brown et al. (Brown et al., 1993) and Lardilleux and Lepage (Lardilleux and Lepage,

[1] DGT-TM - https://open-data.europa.eu/en/data/dataset/dgt-translation-memory
Europarl - http://www.statmt.org/europarl/
OPUS (EUconst, EMEA) - http://opus.lingfil.uu.se/

2009) were employed for an initial extraction as they do not require a priori validated lexicons. The former is based on corpus-wide frequency counts and provides an alignment for every word in the corpus, while the latter is based on random sub-corpus sampling, improving precision for some words but being omissive with respect to others. The alignment method proposed by Gomes (Gomes, 2009) projects the validated bilingual lexicons into the parallel corpus, aligning known expressions, and leaving the remainder words unaligned. The extraction method proposed by Aires et al. (Aires et al., 2009) uses these alignments as anchors to infer alignments of neighbouring unaligned words, based on co-occurrence statistics. Finally, the method proposed by Gomes and Lopes (Gomes and Lopes, 2011) combines these co-occurrence statistics with a spelling similarity score, SpSim, which is trained to recognize cognate words by learning regular spelling differences from previously validated bilingual cognates such as [ph]arm[a]c[y]↔[f]arm[á]c[ia] (EN-PT).

Induction of bilingual stems and suffixes follows the bilingual learning approach (Karimbi Mahesh et al., 2014a) applied on the bilingual lexicon of word-to-word translations for each of the language pairs EN-PT and EN-FR. The approach being purely suffixation based induces bilingual stems, suffixes and bilingual suffix replacement rules that allow one translation form to be obtained from the other (by identifying clusters of bilingual suffixes that associate with a set of induced bilingual stems). The bilingual stems and suffixes learnt, when productively combined, enable new translations to be suggested. Collectively, these bilingual stems and suffixes are referred to as *bilingual morph-units* and are fundamental to the pivoted translation suggestion task elaborated in the forthcoming sections. A bilingual stem conflates various inflected surface forms of a translation. The bilingual suffixes represent morphological extensions for the bilingual stems. The approach is illustrated below for the language pair EN-FR.

1. Decompose each bilingual pair in the lexicon as bilingual stems and bilingual suffixes by pairing similar translations.
   Example: Split pair of translations ‘ensured’ ⇔ ‘assuré’ and ‘ensuring’ ⇔ ‘assurer’ into bilingual stem (‘ensur’ ⇔ ‘assur’) with bilingual morphological extensions (‘ed’, ‘é’) and (‘ing’, ‘er’).

2. Group all the bilingual suffixes that associate with each of the bilingual stem identified in Step 1. Hence identify the bilingual suffix transformations (replacement rules). Each such grouping indicates the possibility of obtaining one surface form from another.
   Example:
   (‘ensure’, ‘assure’) : (‘’, ‘r’) (‘d’, ‘ée’) (‘d’, ‘és’) (‘d’, ‘ées’) (‘d’, ‘é’)
   (‘ensur’, ‘assur’) : (‘e’, ‘er’) (‘ed’, ‘é’) (‘ed’, ‘ée’) (‘ing’, ‘er’) (‘ed’, ‘és’) (‘ed’, ‘ées’)
   represent randomly selected groupings learnt from inflected translation forms ‘ensured’ ⇔ ‘assuré’, ‘ensuring’ ⇔ ‘assurer’ and so forth.

3. Eliminate redundant groups by retaining those bilingual stems that share higher number of transformations.
   Example: Among the two examples in the step 2, the second group (‘ensur’, ‘assur’) : (‘e’, ‘er’) (‘ed’, ‘é’) (‘ed’, ‘ée’) (‘ing’, ‘er’) (‘ed’, ‘és’) (‘ed’, ‘ées’) is retained.

4. Generalise the bilingual suffix replacement rules by looking for other bilingual stems sharing identical transformations. In other words, this involves identification of bilingual suffix clusters (set of bilingual stems sharing same bilingual suffix transformations).
   Example: (‘increas’, ‘augment’): (‘e’, ‘er’) (‘ed’, ‘é’) (‘ed’, ‘ée’) (‘ing’, ‘er’) (‘ed’, ‘és’) (‘ed’, ‘ées’) represents another grouping, where the bilingual stem (‘increas’, ‘augment’) shares same bilingual morphological extensions as the bilingual stem (‘ensur’, ‘assur’) and hence both bilingual stems belong to the same cluster.

   The partition approach provided in the clustering tool kit CLUTO[2] was used to identify the clusters of bilingual suffixes.

4 Approach Outline

The proposed approach works in two phases. Given the list of bilingual stems learnt from validated bilingual lexicons for the language pairs EN-FR and EN-PT as briefed in the Section 3, we derive a lexicon

[2] http://glaros.dtc.umn.edu/gkhome/views/cluto

of bilingual stems for the language pair FR-PT by inducing transitive correspondences between bilingual stems of EN-FR and EN-PT with the common language EN. Having determined the bilingual stems for the language pair FR-PT as mentioned, the associated morphological extensions in the form of bilingual suffixes are gathered for each newly induced bilingual stem based on the transitive correspondences between bilingual suffixes for EN-FR and EN-PT. We impose the constraint that, the bilingual suffixes representing transitive correspondences should occur substantial number of times in the reference set of bilingual suffixes learnt from FR-PT validated lexicon. Newly induced correspondences require validation, as not all of the generated translations are correct. Thus post-generation, prior to manual validation, we classify the generated pairs into one of the pre-defined correct or incorrect classes.

Table 1: Known stem, suffix correspondences for EN-PT and EN-FR (rows 1, 2) and the associated transitive correspondences learnt for FR-PT (row 3)

Language Pair	Bilingual Stems	Bilingual Suffixes	
EN-FR	(‘deliver’, ‘délivr’)	(‘’, ‘er’)	(‘ed’, ‘é’), (‘ed’, ‘és’)
EN-PT	(‘deliver’, ‘emit’)	(‘’, ‘ir’)	(‘ed’, ‘ido’), (‘ed’, ‘iu’)
FR-PT	(‘délivr’, ‘emit’)	(‘er’, ‘ir’)	(‘é’, ‘ido’), (‘é’, ‘iu’), (‘és, ‘ido’), (‘és’, ‘iu’)

Table 1 instantiates the use of EN-FR and EN-PT bilingual morph-units in learning new correspondences for FR-PT, with EN as the pivot language. In the second column of the table, the second and the third rows respectively show the known bilingual stems for each of the language pairs EN-FR and EN-PT. Similarly, the following columns in the second and third rows show the known bilingual suffixes attached to the corresponding bilingual stems shown in column 2. The last row shows the newly induced stem pairs for the target language pair FR-PT and their associated morphological extensions (suffix pairs) obtained by transitivity.

4.1 Pivoting Stem and Suffix Correspondences

First, using the list of bilingual stems for two language pairs L1-L2 and L1-L3 represented as relational tables, we perform a relational natural join over common stem in language L1[3].

Algorithm 1 Translation Generation as Pivoting and Classification

```
procedure PIVOT_BILINGUALMORPHS
    $A_{L1-L2}$, $A_{L1-L3}$ ← lexicon of bilingual stems for L1-L2, L1-L3
    $S_{L2-L3}$ ← bilingual suffix list learnt from validated lexicon for L2-L3
    Join relational tables for $A_{L1-L2}$ and $A_{L1-L3}$ on stems of the common language L1
    for each stem pair $(a_{i_{L1}}, a_{i_{L2}}) \epsilon A_{L1-L2}$ and $(a_{i_{L1}}, a_{i_{L3}}) \epsilon A_{L1-L3}$ do
        if suffix pair $(s_{i_{L1}}, s_{i_{L2}}) \epsilon$ bilingual suffix list associated with $(a_{i_{L1}}, a_{i_{L2}})$ &&
            suffix pair $(s_{i_{L1}}, s_{i_{L3}}) \epsilon$ bilingual suffix list associated with $(a_{i_{L1}}, a_{i_{L3}})$ then
            append $(s_{i_{L2}}, s_{i_{L3}})$ to the suffix list associated with $(a_{i_{L2}}, a_{i_{L3}})$ iff
            $(s_{i_{L2}}, s_{i_{L3}}) \epsilon S_{L2-L3}$ && $occurrence\_frequeny(s_{i_{L2}}, s_{i_{L3}}) \geq 3$.
        end if
    end for
end procedure
```

Let A_{L1-L2} and A_{L1-L3} be the lexicons consisting of bilingual stems for the language pairs L1-L2 and L1-L3 respectively. Further, let S_{L2-L3} be the list of bilingual suffixes learnt from validated lexicon for L2-L3. This list of bilingual suffixes is obtained by applying the bilingual learning approach (Karimbi Mahesh et al., 2014a) on the validated bilingual lexicon for L2-L3. The list serves in identifying valid bilingual suffixes from the set of candidate bilingual suffixes (for L2-L3) induced following

[3] Alternatively, we may perform search and replace operation on the bilingual stem file for L1-L3 using a two-column table (consisting of stems in L1 as first column and their corresponding translations in L2 as the second column) for L1-L2.

transitive correspondences between the bilingual suffixes for L1-L2 and L1-L3 over common suffix in L1 (L1 is the pivot language).

Initially, we perform a natural join on the relational tables for the bilingual lexicons A_{L1-L2} and A_{L1-L3} over common stems of the pivot language L1. Consequently, we obtain a lexicon of candidate bilingual stems for the language pairs L2-L3.

After the candidate bilingual stems are determined for the language pair L2-L3 as specified, the associated bilingual suffixes are predicted for each induced bilingual stem in L2-L3 based on the transitive correspondences between bilingual suffixes for L1-L2 and L1-L3 over common suffix in L1 (as enumerated in the steps 3 through 9 of the Algorithm 1). However, this results in an exhaustive list of candidate bilingual suffix correspondences, for each candidate bilingual stem induced in the previous step. Hence, the selection of valid correspondences from this initial list of candidate bilingual suffix correspondences is done in consultation with S_{L2-L3}, the list of known bilingual suffixes for L2-L3, i.e., valid correspondences between suffixes in L2 and L3 are determined based on their occurrence frequencies in S_{L2-L3}. Candidate bilingual suffixes (following transitive correspondences between suffixes in L1-L2 and L1-L3) with occurrence frequency less than 3 as observed in the bilingual suffix list for L2-L3 are discarded. Setting the occurrence frequency threshold below this value leads to over-generation of surface translation forms dropping the translation generation precision below 60%.

To illustrate the above outlined procedure, consider the examples in Table 1. The last row in the table represents the newly induced bilingual correspondences for FR-PT following the transitive correspondences between bilingual stems and suffixes in EN-FR (second row) and EN-PT (third row). For example, (‘délivr’, ‘emit’) represents the new bilingual stem induced following transitive correspondences between the known bilingual stems (‘deliver’, ‘délivr’) in EN-FR and (‘deliver’, ‘emit’) in EN-PT. The candidate bilingual suffixes that associate with (‘délivr’, ‘emit’) are (‘er’, ‘ir’) following the transitive correspondences between bilingual suffixes (‘’, ‘er’) in EN-FR and (‘’, ‘ir’) in EN-PT and similarly, (‘é’, ‘ido’), (‘é’, ‘iu’), (‘és’, ‘ido’) and (‘és’, ‘iu’) following correspondences between (‘ed’, ‘é’) in EN-FR and (‘ed’, ‘ido’), (‘ed’, ‘iu’) in EN-PT and between (‘ed’, ‘és’) in EN-FR and (‘ed’, ‘ido’), (‘ed’, ‘iu’) in EN-PT. Looking for the occurrence frequencies of each of these correspondences in the FR-PT bilingual suffix list, S_{FR-PT}, learnt from validated FR-PT lexicon, we choose to either retain or discard the associated suffixes.

4.2 Generation of Surface forms

Surface translation forms can be interpreted as the concatenation of newly induced bilingual stems and their associated suffixes. For instance, simple concatenation of the bilingual stem (‘délivr’, ‘emit’) with associated bilingual suffix (‘er’, ‘ir’) yields the surface form (‘délivrer’, ‘emitir’).

4.3 Validation as Binary Classification

We evaluate the newly induced FR-PT pairs by using the validated FR-PT bilingual lexicon for supervised learning and combining varied features derived from that lexicon. We train a SVM-based binary classifier that assigns each of the induced bilingual pairs into one of the previously defined correct or incorrect classes. Features (Karimbi Mahesh et al., 2014b) characterising correct and incorrect bilingual pairs are briefed in the Subsection 4.3.1.

4.3.1 Stem and Suffix Coverage

We view a bilingual translation to be composed of two bilingual morphological segments, the bilingual stem and bilingual suffix. The stem and suffix coverage refers to the content (stem) and inflectional (suffix) coverage exhibited by the bilingual pair under evaluation. The coverage is determined as the agreement between morphological units comprising of stem in one language and its translation in another language and between their morphological extensions, respectively. The features are binary valued, each representing the *stem coverage* (MC_{stm}) and *suffix coverage* (MC_{sfx}), thus characterising the bilingual pair to be validated. For any bilingual pair, a feature value ‘0’ indicates coverage, while ‘1’ indicates mis-coverage, with respect to the morph-unit under evaluation (stem or suffix). The two features may be collectively referred to as the *morphological coverage* feature ($MC_{stm+sfx}$).

To check for parallelism with respect to stems, the left hand side term of the stem pair (FR-PT) to be validated is matched against the set of all stems in first language (FR), learnt from the validated lexicon (training dataset of correct translations) for FR-PT. Similarly, we check if a match is found for the right hand side of the candidate stem pair in the set of known stems for PT. If matched stems are found with respect to first and second languages and further happen to be translations of one another (i.e., bilingual stem pairs), then the candidate translation under test is said to be covered with respect to stem. The set of stems for FR and PT are represented as separate keyword trees (Gusfield, 1997) and are learnt by applying the bilingual learning approach (Karimbi Mahesh et al., 2014a) on the FR-PT lexicon of word-to-word translations. Alternatively, existing stemmer may be employed for the purpose. Aho-corasick set-matching algorithm (Gusfield, 1997) is applied to allow faster search over the known stems represented as keyword tree.

As elaborated in Section 4.1, the bilingual suffixes that attach to a bilingual stem are chosen based on their occurrence frequencies in the bilingual suffix list learnt from FR-PT validated lexicon. Hence, naturally, the bilingual pair satisfies the suffix agreement requirement and hence is covered with respect to suffix.

For instance, consider the newly induced correspondences (‘délivr’, ‘emit’) and their associated bilingual suffixes, (‘er’, ‘ir’), (‘é’, ‘ido’) with surface forms (‘délivrer’, ‘emitir’), (‘délivré’, ‘emitido’). We check if ‘délivr’ matches the set of stems in FR represented as a keyword tree. If so, we check if ‘emit’ matches the set of stems in PT. If a match is found in both the languages, we check if (‘délivr’, ‘emit’) appears as valid stem pair in the set of bilingual stems learnt from FR-PT validated lexicon. MC_{stm} is set to 0 if the candidate stem pair is found, else is set to 1. For any induced translation, the feature value representing the suffix coverage is set to 0. This is because, for each bilingual stem induced via pivoting, its bilingual extensions are those bilingual suffixes (transitive correspondences) that are observed at least three times in the bilingual suffix list learnt from training dataset for FR-PT.

5 Experimental Setup

5.1 Datasets for Pivoted Induction

The bilingual segments (stems and suffixes) used for pivoted induction were learnt from validated EN-FR and EN-PT bilingual lexicons. The statistics of the bilingual resources for EN-PT and EN-FR used in pivot based lexicon induction is as shown in Table 2. For each of the language pairs listed in first column, the second column shows the count of manually accepted word-to-word translations used in acquiring the bilingual resources comprising of stem pairs and suffix pairs. Similarly, the third, fourth and fifth columns respectively show the statistics of bilingual stems, suffixes and bilingual suffix classes learnt. A suffix class corresponds to set of bilingual suffixes representing bilingual extensions for a set of bilingual stems. It may or may not correspond to Part-of-Speech such as noun, verb, adverb or adjective. However, there are cases where the same suffix class aggregates nouns, adjectives and adverbs.

Table 2: Word-to-word translations for EN-FR and EN-PT with the bilingual stem and suffix statistics

Language Pair	Word-word Tanslations	Bilingual Stems	Bilingual Suffixes	Bilingual Suffix Classes
EN-FR	148,441	18,095	261	77
EN-PT	209,739	24,223	232	136

5.2 Datasets for Classification

In order to train a binary classifier capable of evaluating the newly induced FR-PT bilingual pairs, a total of 162,790 word-to-word bilingual pairs were used as the training dataset. 116,621 accepted word-to-word translations were used as positive examples while 46,169 rejected entries formed the negative examples. The FR-PT training and test datasets used in training and testing the classifier were extracted using the approaches mentioned in the Section 3.

Table 3: Training and test data statistics for FR-PT classifier

Dataset	Accepted	Rejected	Total
Training	116,621	46,169	162,790
Test	6,138	2,430	8,568

5.3 SVM-based Binary Classifier

SVM based tool, LIBSVM[4] was used to learn the binary classifier. Grid-search was performed on RBF kernel parameters (C, γ) using cross-validation to enable accurate predictions for the test data.

Table 4: Performance of the FR-PT word-word classifier on FR-PT test set

Features	$\mathbf{P_{Acc}}$	$\mathbf{R_{Acc}}$	$\mathbf{P_{Rej}}$	$\mathbf{R_{Rej}}$	$\mu_{\mathbf{P}}$	$\mu_{\mathbf{R}}$	$\mu_{\mathbf{F}}$	**Accuracy**
StrSim	74.38	97.93	73.87	14.77	74.13	56.35	64.03	74.35
StrSim + MC$_{\text{stm}}$	81.10	98.70	92.71	41.89	86.91	70.3	77.73	82.59
StrSim + MC$_{\text{stm+sfx}}$	81.96	98.89	94.15	42.02	88.06	71.96	79.20	83.61

The SVM-based FR-PT classifier for word-word translations with a micro average f-measure (Equation 8) approximating 80% (last row of the Table 4) when tested on the test dataset shown in Table 3, trained with the features elaborated in Section 4.3, was used in classifying the newly induced bilingual morph-units. The classifier was trained using the string similarity based features (StrSim), apart from the features regarding stem pairs and suffix pairs. Orthographic similarity measure based on edit distance (Levenshtein, 1966) was used to quantify the similarity between terms on either sides of the bilingual pair (surface form).

6 Evaluation

The evaluation metrics for the classifier and the translation suggestion task are elaborated in the Subsections 6.1 and 6.2 respectively.

6.1 Classification

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

$$P_{Rej} = t_n/(t_n + f_n) \tag{1}$$

$$P_{Acc} = t_p/(t_p + f_p) \tag{2}$$

$$R_{Acc} = t_p/(t_p + f_n) \tag{3}$$

$$R_{Rej} = t_n/(t_n + f_p) \tag{4}$$

$$Accuracy = (t_p + t_n)/(t_p + f_p + t_n + f_n) \tag{5}$$

In the equations 1 through 5, 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*. P_{Acc} and R_{Acc} denotes precision and recall for the accepted class, and P_{Rej} and R_{Rej} represents precision and recall for the rejected class.

To assess the global performance over both classes, the Micro-average Precision (μ_P), Micro-average Recall (μ_R) and Micro-average f-measure (μ_F) were used, and calculated as shown in equations 6 through 8 below:

$$\mu_P = (P_{Acc} + P_{Rej})/2 \tag{6}$$

[4] A library for support vector machines - Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

$$\mu_R = (R_{Acc} + R_{Rej})/2 \tag{7}$$

$$\mu_F = 2 * \mu_P * \mu_R/(\mu_P + \mu_R) \tag{8}$$

6.2 Generation

The precision for generated bilingual pairs[5] is calculated as the fraction of correctly generated bilingual pairs to the total number of bilingual pairs generated.

6.2.1 Manual Evaluation

Manual classifications are based on the observations that certain translations are wrong (incomplete or inadequate) (examples labelled as 'Reject' in Table 7). For instance, some of the newly suggested translations are inadequate as they miss an auxiliary verb form in French or a relative pronoun in Portuguese or a negation expression "n'...pas" in French that is also missing in Portuguese ('não'). Despite adjective gender differences in French and Portuguese, no one will be able to know a priori the chosen translation for any noun and the adjective number and gender in French (or in Portuguese) will depend on the chosen French (Portuguese) noun translation.

7 Results and Discussion

As summarised in Table 5, experiments using EN-FR and EN-PT lexicons enabled induction of 28,755 unique transitive bilingual stem correspondences with 1,047 unique bilingual suffix correspondences for FR-PT, contributing to a total of 272,193 unique word-to-word surface translation candidates.

Table 5: Statistics for pivoted bilingual stems, suffixes and translations induced for FR-PT

Description	**Bilingual Stems**	**Bilingual Suffixes**	**Word-Word Translations**
Unique Correspondences	28,755	1,047	272,193

Exclusive automatic validation of newly induced stem pairs using FR-PT binary classifier show that 1,022 candidate bilingual stems matched with the validated and accepted bilingual stems learnt (Karimbi Mahesh et al., 2014a) from the training dataset (word to word translations) for FR-PT. 2,016 stem pairs were orthographically similar (cognates). 19,946 bilingual stem correspondences did not exist in the training data used for classification.

Table 6: Results of classification on FR-PT bilingual morph units

Features	**Bilingual Stems**
Matching Correspondences (Pivot && Bilingually learnt)	1,022
Orthographically Similar	2,016
Total Correspondences automatically validated	3,038

Evaluation of 126 pivoted FR-PT bilingual suffixes-only show that 108 of them were correct and 21 were incorrect, yielding the precision 84%. Further, manual validation of induced pairs (surface translations) indicates precision approximating 60%.

Among the 272,193 FR-PT inflected word-to-word translations generated via pivoted induction, it was observed that 234,000 were new entries that had not been extracted by any other methods. 38,000 entries had already been extracted by other methods. 39,000 of the new translations generated by pivoted induction (that have not been extracted by any other method) did occur in the parallel corpora, with 13,000 entries co-occurring only once, 7,000 co-occurring twice and 2,000 co-occurring three times.

[5] bilingual stems, bilingual suffixes and bilingual surface forms

It is to be noted that the above stated results were achieved using all of the bilingual stems and suffixes learnt from the EN-PT and EN-FR lexicon. The automatically learnt bilingual resources comprising of bilingual stems and suffixes that served as knowledge bases for pivoting FR-PT translations were evaluated indirectly in terms of the generation precision considering new translations (surface forms) suggested. Generation precision (computed as specified in Subsection 6.2) was respectively 90%[6] for EN-PT and 81.55%[7] for EN-FR. As all of the automatically learnt bilingual stems and suffixes were used in our experiments, restricting the knowledge bases used in pivoting to only correct bilingual segments would further improve the results.

7.1 Error Analysis

Some of the French suffixes 'é', 'ée', 'és', 'ées', 'u', 'ue' (and others) were wrongly paired with 'ou', 'aram', 'eu', 'eram', 'iu', 'iram' in PT. Generally, these past participle French forms need an auxiliary verb in French, 'a' or 'ont', to give rise to a form in Portuguese ending in 'ou', 'aram', 'eu', 'eram', 'iu', 'iram' and these correspond to verb forms in English ending in 'ed' that sometimes occur with auxiliary verb forms 'has' or 'have'. Examples include 'a soutenu' (FR) ⇔ 'supported' (EN) ⇔ 'has supported' (EN) ⇔ 'apoiou' (PT) and 'ont suscité' (FR) ⇔ 'provoked' (EN) ⇔ 'have provoked' (EN) ⇔ 'provocaram' (PT). It is the generation of single word form in English that gives rise to those errors. An infinitive in French (as 'soutenir') never translates as a present indicative form, either in subjunctive mood (as 'apoiem') or in indicative mood (as 'apoiam') in Portuguese and requires some extensions both in French (a preposition as 'à', 'de', etc.) and in Portuguese (a relative pronoun as 'que').

Table 7: Manual classifications for newly generated translations using the pivoted induction approach (FR-PT). The columns 'Accept' and 'Reject' show correct and wrong translations respectively. The column 'Corresponding Correct Forms' just illustrates some of the correct translations into Portuguese corresponding to wrong translation inducted from FR-PT

FR-PT		Corresponding Correct Forms
Accept	Reject	
soutenir ⇔ apoiarem	soutenir ⇔ apoiam	à soutenir ⇔ que apoiam
soutenir ⇔ apoiar	soutenir ⇔ apoiem	de soutenir ⇔ que apoiem
soutenu ⇔ apoiado	soutenu ⇔ apoiou	a soutenu ⇔ apoiou
soutenue ⇔ apoiado	soutenue ⇔ apoiou	este soutenue ⇔ apoiou-se
suscité ⇔ provocado	suscité ⇔ provocaram	ont suscité ⇔ provocaram
suscitées ⇔ provocados	suscité ⇔ provocou	a suscité ⇔ provocou
suscitées ⇔ provocadas	suscitées ⇔ provocaram	ont été suscitées ⇔ provocaram-se
adaptant ⇔ adaptarem	adaptant ⇔ adaptem	adaptant ⇔ que adaptem
adaptant ⇔ adaptando	adaptant ⇔ adaptem	n'adaptant pas ⇔ que não adaptem

In what regards suffixes 'é', 'ée', 'és', 'ées', when auxiliary verb in French is 'est' or 'sont' we have a passive form that generally translates as a passive form in English, while in Portuguese it requires either an auxiliary verb form, as 'é' or 'são', or requires a passive clitic 'se'. Even French may use the clitic 'se' or 's' and auxiliary verb 'être' ('ést' and 'sont'), or the clitic 'on' and auxiliary 'a' or 'ont' (for singular and plural). Suffixes 'é', 'ée', 'és', 'ées' in French belong to a verbal group ending in 'er', as is the case of 'susciter'.

[6] Precision shown corresponds to 2,334 evaluated EN-PT surface forms out of a total of 14,530 pairs generated, where 2,283 were correct and 20 were incorrect

[7] Among the evaluated 4254 entries, out of a total of 18,095 EN-FR bilingual pairs generated, 3469 were correct and 785 were incorrect.

8 Conclusion

In this paper, we have explored the possibility of inducing a bilingual lexicon for the language pair FR-PT by learning transitive correspondences between bilingual stems and suffixes for the language pair EN-FR and EN-PT. Unlike the traditional induction scheme using surface translation forms, we used resources comprising of bilingual stems and suffixes as basis for the pivoted induction. Our approach relies on initially learning suffixes and suffixation operations from validated bilingual lexicons of word-to-word translations using a bilingual learning framework. The bilingual segments thus learnt are then utilised in suggesting new translations using pivoted induction strategy.

Newly induced pairs were validated using an SVM-based binary classifier trained on morphological and similarity based features learnt from validated FR-PT bilingual translation lexicon. Manual validation of the induced surface forms shows precision approximating 60%. The results may be improved by using only those bilingual segments that have been classified as 'accepted'. As future work, we intend to experiment with other language pairs such as EN-PT and EN-HI, EN-LT and EN-PT. Experiments on pivoted induction with morphologically rich language as pivot needs to examined. The bilingual morph-units may enable compact representation of bilingual lexicon, apart from their applicability in inducing surface inflected forms.

Acknowledgements

K. M. Kavitha and Luís Gomes acknowledge the Research Fellowship by FCT/MCTES with Ref. nos., SFRH/BD/64371/2009 and SFRH/BD/65059/2009, respectively. This work was further supported by the funded research project ISTRION (Ref. PTDC/EIA-EIA/114521/2009), NOVA LINCS (ref. UID/CEC/04516/2013), FCT-UNL and ISTRION BOX - Translation & Revision, Lda..

References

Judit Ács. 2014. Pivot-based multilingual dictionary building using wiktionary. In *LREC*, pages 1938–1942. ELRA.

José Aires, José Gabriel Pereira Lopes, and Luís Gomes. 2009. Phrase translation extraction from aligned parallel corpora using suffix arrays and related structures. In *Progress in Artificial Intelligence*, pages 587–597. Springer.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Jorge Costa, Luís Gomes, Gabriel Pereira Lopes, and Luís MS Russo. 2015. Improving bilingual search performance using compact full-text indices. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 582–595. Springer.

Luís Gomes and José Gabriel Pereira Lopes. 2011. Measuring Spelling Similarity for Cognate Identification. In *Progress in Artificial Intelligence — 15th Portuguese Conference on Artificial Intelligence, EPIA 2011*, pages 624–633, Lisbon, Portugal, October. Springer.

Luís Gomes. 2009. Parallel texts alignment. Master's thesis, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (FCT-UNL), Monte da Caparica.

Dan Gusfield. 1997. *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge Univ Pr. pages 52–61.

Hiroyuki Kaji and Dashtseren Erdenebat. 2008. Automatic construction of a japanese-chinese dictionary via english. In *LREC*, pages 699–706.

Kavitha Karimbi Mahesh, Luís Gomes, and José Gabriel Pereira Lopes. 2014a. Identification of bilingual segments for translation generation. In *Advances in Intelligent Data Analysis XIII*, volume 8819 of *LNCS*, pages 167–178. Springer International Publishing.

Kavitha Karimbi Mahesh, Luís Gomes, and José Gabriel Pereira Lopes. 2014b. Identification of bilingual suffix classes for classification and translation generation. In *Advances in Artificial Intelligence, IBERAMIA 2014*, LNCS, pages 154–166. Springer.

Adrien Lardilleux and Yves Lepage. 2009. Sampling-based multilingual alignment. In *Proceedings of Recent Advances in Natural Language Processing*, pages 214–218.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710.

Luka Nerima and Eric Wehrli. 2008. Generating bilingual dictionaries by transitivity. In *LREC*.

Kyonghee Paik, Satoshi Shirai, and Hiromi Nakaiwa. 2004. Automatic construction of a transfer dictionary considering directionality. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, pages 31–38. Association for Computational Linguistics.

Xabier Saralegi, Iker Manterola, and Iñaki San Vicente. 2011. Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 846–856, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xabier Saralegi, Iker Manterola, and Iñaki San Vicente. 2012. Building a basque-chinese dictionary by using english as pivot. In *LREC*, pages 1443–1447.

Daphna Shezaf and Ari Rappoport. 2010. Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 98–107. Association for Computational Linguistics.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, COLING '94, pages 297–303, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mairidan Wushouer, Tomoyuki Ishida, and Donghui Lin. 2013. A heuristic framework for pivot-based bilingual dictionary induction. In *Culture and Computing (Culture Computing), 2013 International Conference on*, pages 111–116. IEEE.

Mairidan Wushouer, Toru Ishida, Donghui Lin, and Katsutoshi Hirayama. 2014a. Bilingual dictionary induction as an optimization problem. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 2122–2129.

Mairidan Wushouer, Donghui Lin, Toru Ishida, and Katsutoshi Hirayama. 2014b. Pivot-based bilingual dictionary extraction from multiple dictionary resources. In *PRICAI 2014: Trends in Artificial Intelligence*, pages 221–234. Springer.