Bilingually motivated segmentation and generation of word translations using relatively small translation data sets

K. M. Kavitha[1,3] **Luís Gomes**[1,2] **José Gabriel P. 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

Out-of-vocabulary (OOV) bilingual lexicon entries is still a problem for many applications, including translation. We propose a method for machine learning of bilingual stem and suffix translations that are then used in deciding segmentations for new translations. Various state-of-the-art measures used to segment words into their sub-constituents are adopted in this work as features to be used by an SVM based linear classifier for deciding appropriate segmentations of bilingual pairs, specifically, in learning bilingual suffixation.

1 Introduction

OOV bilingual lexicon entries still remain an open problem and the approach proposed in this paper will contribute to solve this by machine learning of bilingual stem and suffix pairs using a very small English (EN)-Hindi (HI) bilingual lexicon. These bilingual segments are then used in deciding segmentations for unseen translations and also in generating new ones. We examine a combination of commonly used segmentation measures as clues for bilingual suffixation of unseen translations in a minimally supervised framework.

No translation extraction technique guarantees the extraction of all possible translation pairs specially when not found or are infrequent in parallel corpora. Source-target asymmetry further adds to the problem for morphologically poor and rich language pairs, as is the case of English and Hindi. Noun, verb or adjective forms in English tend to have multiple translations in Hindi. Consider the English term *'good'* with 3 possible translations: *'acChA'*, *'acChI'* and *''acChe'* in Hindi. Each of these represent variants of the basic word form *'acChA'*, where *'-A'* and *'-I'* represent singular masculine and feminine, while *'-e'* denotes a plural adjective suffix. As all the forms might hardly be seen in the training data, there is a need for identifying morphological similarities in the known example pairs[1]. In the referred example, the three word forms share the stem *'acCh'* and differ in the endings *'-A'*, *'-I'*, *'-e'*. As these inflections appear as endings for other words they serve in identifying word classes. Thus, the separation of morphological suffixes conflates various forms of a word, into a stem which is a crucial source of information. On the other hand, suffixes that occur frequently with words belonging to similar class, could be utilised for generating unknown forms. Hence, by using the morphological information, all possible forms can be inferred by combining different component morphemes from different mappings learnt using the example pairs in the translation lexicon.

We discuss a generative approach for suggesting new translations based on the morphological similarities learnt from translation examples seen in the existing bilingual lexicon. The approach is distinguishing as we rely on the frequent forms (suffixes) occurring in translations rather than on words in just one language. Fundamental to this generation strategy, we have 2 phases involving learning and classification. Firstly, the bilingual approach to learning morph-like units is used in preparing the training data (Mahesh et al., 2014). This involves identification and extraction of orthographically and semantically similar bilingual segments (as for instance, *'good'* ⇔ *'acCh'*) occurring in known translation

[1]Words consist of high-frequency strings (affixes) attached to low-frequency strings (stems) (Hammarström, 2009)

29th Pacific Asia Conference on Language, Information and Computation: Posters, pages 187 - 196
Shanghai, China, October 30 - November 1, 2015

examples (*'acChA', 'acChI'* and *''acChe'*), together with their bilingual extensions constituting dissimilar bilingual segments (bilingual suffixes) (*'' ⇔ 'A' | 'e' | 'I'*)[2]. The common part of translations that conflates all its bilingual variants[3] represents a bilingual stem (*'good' ⇔ 'acCh'*). The different parts of the translations contributing to various surface forms represent bilingual suffixes or bilingual morphological extensions (*'' ⇔ 'A' | 'e' | 'I'*). Further, bilingual suffixes representing bilingual extensions for a set of bilingual stems form bilingual suffix classes[4], hence allowing safer translation generalisation. The bilingual suffix classes thus learnt along with the bilingual lexicon constitutes the training data set for the classification phase. Upon identification of the segmentation boundary (by classification), depending on the bilingual suffix and the stem surfaced for the given unseen translation, the bilingual pair is then classified into one of the bilingual suffix classes identified in the training phase. New translations are then suggested by simple concatenation of bilingual stems and suffixes belonging to the identified class.

2 Related Work

2.1 Monolingual Approaches

Lexical inference or morphological processing techniques have been established for handling unknown terms that are variations of known forms. Moreover, learning suffixes and suffixation operations from the corpus or lexicon of a language allows new words to be generated. Such approaches are categorised as supervised (Déjean, 1998), semi-supervised (Lindén et al., 2009) and unsupervised (Goldsmith, 2001; Creutz and Lagus, 2005; Monson et al., 2009).

The state-of-the-art approaches to unsupervised morphology learning are overviewed and discussed with sufficient level of detail by Wicentowski and Yarowsky (2002) and Hammarström and Borin (2011), respectively. A most recent work integrates orthographic and semantic view of words and models word formation in terms of morphological chains (Narasimhan et al., 2015). To address the morphological segmentation problem, Kirschenbaum (2015) suggests the use of segmentations derived from words sharing similar distribution and form in analysing less frequent words.

Partially supervised strategies for morphology learning may be viewed as classification tasks. The classifier trained on known paradigms classifies the unseen words into paradigms or induces new paradigms (Lindén et al., 2009).

2.2 Bilingual Learning Approaches

Sasaoka *et al.*, proposed bilingual inductive learning mechanism for predicting translations for unknown words (Sasaoka et al., 1997). Common and different parts of strings between known words and their translations represent the example strings, referred as Piece of Word (PW) and Pair of Piece of Word (PPW). The bilingual pairs of these extracted example strings maintained as a Pair of Piece of Word (PPW) dictionary form the basis of the prediction process.

Snyder and Barzilay (2008) proposed simultaneous morphology learning for discovery of abstract morphemes using multiple languages. To boost the segmentation decisions, Poon et al. (2009), proposed discriminative log-linear model employing overlapping contextual features.

In our previous work, we proposed an approach for learning bilingual suffixation operations by utilising the translation lexicon as a parallel resource (Mahesh et al., 2014). As a pre-phase to translation generation, bilingual morph-like units conflating various translation forms are learnt and consequently clustered into bilingual suffix classes. Frequent forms occurring in translations rather than in word forms (in a language) are used in arriving at the segmentation decision. The ambiguities and complexities in decompositions are reduced as the translation forms impose a restricted subset over the entire universe of word forms from which segmentation decisions are made. Similar to the approach proposed by Sasaoka *et al.*, (Sasaoka et al., 1997), our approach (Mahesh et al., 2014) that we adapt here for preparing the partial training data, allows identification of common (bilingual stems) and different (bilingual suffixes) bilingual segments occurring in translation examples, which are then used in generating new translations.

3 Proposed approach

Much of the research ranging from text analysis for acquisition of morphology, to learning suffixes and

[2]Note the null suffix in the English side corresponding to gender and number suffixes in the Hindi side.

[3]Translations that are lexically similar.

[4]A *suffix class* may or may not correspond to Part-of-Speech such as noun or adjective but there are cases where the same suffix class aggregates nouns, adjectives and adverbs.

suffixation operations for partially overcoming OOV bilingual entries and generating necessary trustable bilingual entries, is driven by the fact that word is made up of high-frequency affixes attached to low-frequency stems (Hammarström, 2009). Extending this observation, we interpret a bilingual pair to be constituted by frequent bilingual suffixes attached to less frequent bilingual stems. The proposed approach operates in 2 stages: the *learning phase* for identifying bilingual suffix classes that partially serves as the training data and the *classification phase* for deciding segmentation.

3.1 Learning Phase

Learning bilingual segments using translation variants and their mapping into morphologically related classes closely follows the bilingual learning approach and involves learning bilingual suffixes and suffixation operations (Mahesh et al., 2014) (refer Algorithm 1).

Definitions Let L be a Bilingual Lexicon.
Let L1, L2 be languages with alphabet set Σ_1, Σ_2.
T=$\{(w_{L1}, w_{L2})|(w_{L1}, w_{L2}) \subset L\}$ be set of valid bilingual pairs (translations) in L.
S=$\{p_{i_{L1}}, s_{i_{L1}}, p_{i_{L2}}, s_{i_{L2}}|p_{i_{L1}}s_{i_{L1}} = w_{i_{L1}}; p_{i_{L2}}s_{i_{L2}} = w_{i_{L2}}; p_{i_{L1}}, s_{i_{L1}}\epsilon\Sigma_1, p_{i_{L2}}, s_{i_{L2}}\epsilon\Sigma_2\}$ be the set of substrings of $w_{i_{L1}}, w_{i_{L2}}$, where $p_{i_{L1}}s_{i_{L1}}$ denotes the concatenation of stem $p_{i_{L1}}$ and suffix $s_{i_{L1}}$ in languages L1 and L2.
Let $S_{SuffixPair}$ be the set of bilingual suffix pairs and $S_{StemPair}$ be the set of bilingual stem pairs.
Two translations $(w_{1_{L1}}, w_{1_{L2}})$ and $(w_{2_{L1}}, w_{2_{L2}}) \in L$ are said to be *similar* if $|lcp(w_{1_{L1}}, w_{2_{L1}})| \geq 3$ and $|lcp(w_{1_{L2}}, w_{2_{L2}})| \geq 3$, where lcp is the longest common prefix of the strings under consideration.

Input - Bilingual/Translation Lexicon (L):

Translation lexicon refers to a dictionary which contains a term (taken as a single word - any contiguous sequence of characters) in the first language cross-listed with the corresponding term in the second language such that they share the same meaning or are usable in equivalent contexts. In Table 1, sample entries illustrate bilingual variants: $noun\_singular$ forms (columns 1, 2 in 1^{st} 7 rows) – $noun\_plural$ forms (column 3, 4 in 1^{st} 7 rows) and *adjective* forms (columns 1, 2 in last 4 rows) – *adverb* forms (columns 3, 4 in last 4 rows).

Output :
List of Bilingual stem and suffix pairs: These include the list of bilingual stems (columns 3, 4 in Table 5) and suffixes (Table 4) with their observed frequencies in the training dataset. Sample bilingual stems include *'plant'* ⇔ *'paudh'*, *'boy'* ⇔ *'laDak'*. Sample bilingual suffixes are *('', 'I'), ('', 'A'), ('ion', 'A')* and are attached to 10,743, 29,529 and 457 different bilingual pairs respectively. These lists aid in identifying bilingual stems and bilingual suffixes, given a new translation.

Term (EN)	Term (HI)	Term (EN)	Term (HI)
process	प्रक्रिया (prakriyA)	processes	प्रक्रियाओं (prakriyAoM)
proof	प्रमाण (pramAN)	proofs	प्रमाणों (pramANoM)
plant	पौधा (paudhA)	plants	पौधों (paudhoM)
proceeding	कार्यवाही (kAryavAhI)	proceedings	कार्यवाहियों (kAryavAhiyoM)
plan	योजना (yojanA)	plans	योजनाएं (yojanAeM)
prayer	प्राथना (prArthanA)	prayers	प्राथनाएँ (prArthanAeN)
promise	वाद (vAd)	promises	वादे (vAde)
usual	सामान्य /साधारण (sAmAny/sAdharaN)	usually	सामान्यतः /साधारणतः (sAmAnyatH/sAdharaNatH)
chief	प्रधान (pradhAn)	chiefly	प्रधानतः (pradhanataH)
rapid	शीघ्र (shIghr)	rapidly	शीघ्रता (shIghratA)
weak	दुर्बल (durbal)	weakly	दुर्बलता (durbalatA)

Table 1: Bilingual variants in EN-HI Lexicon

Bilingual suffixes grouped by bilingual stems: This represents which set of bilingual suffixes attach to which bilingual stem. In Table 2[5], the bilingual suffixes, *('s', 'oM')* and *('ous', 'I')* attach to the same bilingual stem *('mountain', 'pahAD')* yielding the surface forms *'mountains'* ⇔ *'pahADoM'* and *'mountainous'* ⇔ *'pahADI'*.

Bilingual Stems	Bilingual Suffixes
('nation', 'राष्ट्र')	: ('al', 'ीय'), ('alism', 'ीयता'), ('ality', 'ीयता'), ('alist', 'ीयतावादी')
('nation', 'rAShTr')	: ('al', 'Iya'), ('alism', 'IyatA'), ('ality', 'IyatA'), ('alist', 'IyatAvAdI')
('mountain', 'पहाड')	: ('s', 'ों'), ('ous', 'ी')
('mountain', 'pahAD')	: ('s', 'oM'), ('ous', 'I')

Table 2: Bilingual suffixes grouped by bilingual stems

Bilingual Suffix Classes: A set of bilingual stems that share same suffix transformations form a cluster or a bilingual suffix class. In the 1^{st} row of Table 5, *('', 'A')* and *('s', 'oM')* represent bilingual suffixes that combine with bilingual stems, *'plant'* ⇔ *'paudh'*, *'boy'* ⇔ *'laDak'* and many more. These allow new translation forms to be subsequently suggested upon identification of bilingual stems and suffixes in an unseen translation given as input.

[5] 2^{nd} line in each row shows transliterations for HI terms

Algorithm 1 Learning Bilingual Suffix Classes

procedure LEARNBILINGUALSUFFIXCLASS
  for each translation $(a_{L1}, a_{L2}) \in L$ **do**
    if $\exists$ (b_{L1}, b_{L2}) *similar* to (a_{L1}, a_{L2}), and (c_{L1}, c_{L2}) *similar* to $(d_{L1}, d_{L2}) \in L$,
      where $p_{1_{L1}}, p_{1_{L2}}, p_{2_{L1}}, p_{2_{L2}}, s_{1_{L1}}, s_{1_{L2}}, s_{2_{L1}}, s_{2_{L2}} \epsilon S$, and
      (a_{L1}, a_{L2})=$((p_{1_{L1}} s_{1_{L1}}), (p_{1_{L2}} s_{1_{L2}}))$; (b_{L1}, b_{L2})=$((p_{1_{L1}} s_{2_{L1}}), (p_{1_{L2}} s_{2_{L2}}))$,
      $(c_{L1} c_{L2})$=$((p_{2_{L1}} s_{1_{L1}}), (p_{2_{L2}} s_{1_{L2}}))$; (d_{L1}, d_{L2})=$((p_{2_{L1}} s_{2_{L1}}), (p_{2_{L2}} s_{2_{L2}}))$ **then**
        add $(p_{1_{L1}}, p_{1_{L2}})$ to the list of bilingual stems $S_{StemPair}$.
        add $((s_{1_{L1}}, s_{1_{L2}}), (s_{2_{L1}}, s_{2_{L2}}))$ to the list of bilingual suffixes $S_{SuffixPair}$.
  for each suffix pair $(s_{i_{L1}}, s_{i_{L2}}) \epsilon S_{SuffixPair}$ **do**
    if $\exists$ m, n such that $(ms_{i_{L1}}, ns_{i_{L2}}) \epsilon S_{SuffixPair}$, $u_2 m = u_1, v_2 n = v_1$
      and bilingual stem (u_1, v_1) and $(u_2, v_2) \epsilon S_{StemPair}$, **then**
        replace (u_1, v_1) by (u_2, v_2) and $(s_{i_{L1}}, s_{i_{L2}})$ by $(ms_{i_{L1}}, ns_{i_{L2}})$ **iff**
        $Strength(s_{i_{L1}}, s_{i_{L2}})$ or $Strength(m, n) > Strength(ms_{i_{L1}}, ns_{i_{L2}})$.
  for each stem pair $(p_{i_{L1}}, p_{i_{L2}}) \epsilon S_{StemPair}$, where $((p_{i_{L1}} s_{i_{L1}}), (p_{i_{L2}} s_{i_{L2}}))$= $(w_{i_{L1}}, w_{i_{L2}}) \epsilon L$ **do**
    if $(s_{i_{L1}}, s_{i_{L2}})$ is not in the list of bilingual suffixes
      associated with the bilingual stem $(p_{i_{L1}}, p_{i_{L2}})$ **then**
      append $(s_{i_{L1}}, s_{i_{L2}})$ to the suffix list associated with $(p_{i_{L1}}, p_{i_{L2}})$.
  Cluster the stem pairs sharing similar suffix transformations into bilingual suffix classes.
end procedure

3.2 Classification

In this section, we discuss the use of SVM based linear classifier[6] (Fan et al., 2008) in predicting if a given segmentation option corresponds to a valid boundary or not.

$$(p_{1_{L1}}, p_{1_{L2}})(s_{1_{L1}}, s_{1_{L2}}), (p_{2_{L1}}, p_{2_{L2}}) \\ (s_{2_{L1}}, s_{2_{L2}}), ..., (p_{n_{L1}}, p_{n_{L2}})(s_{n_{L1}}, s_{n_{L2}}) \quad (1)$$

In Equation 1, all possible bilingual stems and suffixes associated with a given bilingual word pair $(w_{i_{L1}}, w_{i_{L2}})$ are represented, where $(p_{i_{L1}}, p_{i_{L2}})(s_{i_{L1}}, s_{i_{L2}})$ represents a candidate for the bilingual stem and suffix (a possible segmentation boundary). The principle of classification involves learning a function, to infer a binary decision for each split, given all possible segmentations comprising of bilingual stems and suffixes for any given unseen translation.

Each of the possible segmentations (constituting bilingual stem and bilingual suffixes) is a data instance, represented as a feature vector and a target value indicating if the corresponding segmentation is valid (+1), invalid (-1) or unknown (0). We train a binary classifier using the features identified from the training dataset made of the bilingual lexicon and the clusters (bilingual suffix classes) identified during the learning phase. Segmentation boundaries identified for each of the bilingual pairs during the learning phase represent positive samples and all other possible segmentation options for the bilingual pair represent negative samples. Given all possible splits for a new bilingual pair, the estimated model should predict if each of the candidate segmentations represents a valid boundary (+1) or not (-1).

[6] http://www.csie.ntu.edu.tw/ cjlin/papers/liblinear.pdf

3.2.1 Lexicon as Training Data

The measures discussed below, used in segmenting words into substituent morphemes, are adopted in bilingual framework and are used to derive features to minimally supervise the segmentation.

Stand-alone Bilingual Pair We use a binary valued feature indicating if each candidate bilingual stem appears as a stand-alone translation in the lexicon with respect to the candidate segmentation boundary. This knowledge is frequently used in several word-based models and in one of the best performing approaches selected by Hafer *et al.* (Hafer and Weiss, 1974). Instances of bilingual stems appearing as stand-alone bilingual pairs in the lexicon are *'mountain'* ⇔ *'pahAD'* and *'region'* ⇔ *'kShetr'*.

Candidate Boundary Offset (BO) A pair of index numbers indicating the position of the candidate boundary relative to the beginning and end of the bilingual pair characterises the boundary points. Single-character suffixes, or generally short suffixes are often observed to be spurious than the long ones (Goldsmith, 2001). Index values have been used as multipliers in the function reflecting optimal split

position to deal with the disparity with respect to the frequency of shorter stems and suffixes vs longer ones (Patel et al., 2010). Further, the index values have been used as features in correcting the problem with predecessor variety values resulting from normalisation (Çöltekin, 2010). This knowledge is represented by 4 additional features:

- A pair of integer-valued features corresponding to the offsets from the beginning of the bilingual pair (with respect to candidate boundary). For the bilingual pair, *'boys'* ⇔ *'laDakoM'*, with a candidate bilingual stem *'boy'* ⇔ *'laDak'*, the offsets[7] are 3 and 3 EN and HI characters, respectively.
- A pair of integer-valued features corresponding to the offsets from the end of the bilingual pair (with respect to candidate boundary). For the example above, the offsets are 1 and 1 EN and HI character, respectively.

Normalised Successor Entropy (NSE) The successor entropy is calculated for each stem pair as :

$$H(p_{L1}, p_{L2}) = -\sum_{(s_{L1}, s_{L2}) \in succ(p_{L1}, p_{L2})} \frac{f(p_{L1}s_{L1}, p_{L2}s_{L2})}{f(p_{L1}, p_{L2})} \cdot log_2 \frac{f(p_{L1}s_{L1}, p_{L2}s_{L2})}{f(p_{L1}, p_{L2})} \quad (2)$$

where, $(p_{L1}s_{L1}, p_{L2}s_{L2})$ is the bilingual string that is formed by concatenation of s_{L1} to p_{L1} and s_{L2} to p_{L2}, f() represents the frequency of the bilingual pairs starting with the given bilingual stem (prefix pair), and succ() returns all bilingual suffixes (suffix pairs) for the given bilingual stem (p_{L1}, p_{L2}).

NSE for a candidate stem pair is obtained by dividing the calculated entropy value by the expected value (considering bilingual stems having same length as the candidate stem pair) corresponding to the split position.

Normalized Predecessor Entropy (NPE) NPE for a candidate suffix pair is obtained by dividing the calculated predecessor entropy (PE) value by the expected value (considering the bilingual suffixes having same length as the candidate suffix pair) with respect to the split position. PE can be obtained using the Equation 2 by replacing successor with predecessor and switching the concatenation order.

[7]Transcription of HI characters to Latin ones is not character number conservative. But as we work with both character types, offsets must obey the character set in question.

Normalized Successor Variety (NSV) and Normalized Predecessor Variety (NPV) We define successor variety as the number of distinct bilingual suffixes that follow a candidate bilingual stem. This count is calculated for each candidate bilingual stem in the training data set. The SV segmentation measure initially proposed by Harris (1970) is employed in numerous word-segmentation tasks (Déjean, 1998; R. et al., 2005; Stein and Potthast, 2007; Bordag, 2008). Further, researches show how this measure could be utilised in improving the segmentation results (Hafer and Weiss, 1974; Çöltekin, 2010).

The variety values are normalised by dividing the calculated value by the expected value (based on the equi-lengthed bilingual stems) with respect to the split position. The NPV value for a candidate bilingual suffix may be calculated similarly. Çöltekin (2010) provide an elaborate analysis of the problems concerning SV values and the suggested improvements using normalized SV scores.

Bilingual Morpheme Frequency (BMF) This measure quantifies a candidate bilingual morpheme by the number of distinct translations to which it attaches in the bilingual lexicon.

$$bmf(m_{L1}, m_{L2}) = Number\ of\ unique\ bilingual pairs\ (m_{L1}, m_{L2})\ attaches\ to. \quad (3)$$

where (m_{L1}, m_{L2}) is the candidate bilingual morpheme (a bilingual stem or a bilingual suffix). This adds 2 features, corresponding to each candidate bilingual stem and the candidate bilingual suffix.

Generative Strength (GS) Instead of placing same weight on each bilingual pair when scoring a morpheme, each bilingual pair might be assigned weight based on its generative strength (Dasgupta and Ng, 2007). The generative strength of a bilingual pair is estimated by calculating how many distinct induced bilingual morphemes attach to that bilingual pair. The score of a bilingual morpheme is defined to be the sum of the strengths of the bilingual pairs to which it attaches.

$$gs(m_{L1}, m_{L2}) = \sum_{(w_{i_{L1}}, w_{i_{L2}})} Strength(w_{i_{L1}}, w_{i_{L2}}). \quad (4)$$

where $(w_{i_{L1}}, w_{i_{L2}})$ represents the bilingual pair to which the candidate bilingual morpheme (m_{L1}, m_{L2}) attaches. The heuristic has been used in various word-based segmentation tasks to select from among multiple suffixes while stemming a

word form (Pandey and Siddiqui, 2008; Zeman, 2008).

Table 4 (columns 3 and 4) shows the scores for frequent bilingual suffixes using each of the above mentioned scoring functions.

3.2.2 Clusters as Training Data

The clusters (bilingual suffix classes) generated in the learning phase is additionally used as training data to model the bilingual suffixes for classification.

Cluster-based Bilingual Suffix Length (CBSL) This is calculated as the number of times a bilingual pair which is (l_1, l_2) characters contains an (sl_1, sl_2) character long bilingual suffix, normalized by the total number of bilingual pairs with length (l_1, l_2) (Brychcín and Konopík, 2015).

Cluster-based Bilingual Suffix Probability (CBSP) This represents the probability that a candidate bilingual morphological extension is a correct bilingual suffix. The clusters generated in learning phase are used to estimate this and is calculated as the number of times the bilingual suffix $(s_{i_{L1}}, s_{i_{L2}})$ follows the bilingual stem of a translation $(w_{i_{L1}}, w_{i_{L2}})$ (for each bilingual pair in each cluster), divided by the number of all times $(w_{i_{L1}}, w_{i_{L2}})$ ends with $(s_{i_{L1}}, s_{i_{L2}})$ (Brychcín and Konopík, 2015).

3.3 Suffix Class Determination and Translation Generation

Given a new translation, upon identification of the segmentation boundary (after classification), we need to identify to which bilingual suffix class the surfaced bilingual suffix and hence the translation belongs. Depending on the bilingual suffix and the stem identified for the given translation, the bilingual pair is classified into one of the bilingual suffix classes identified in the training phase. This is approached as a multi-label classification problem.

SVM based tool namely LIBSVM[8] was used to learn the multi-label classifier. A class is represented as a set of features represented by a feature-value pair and a label. The features are bilingual suffixes that are representatives of a class. For any class, the value in a feature-value pair simply indicates whether the bilingual suffix is a representative of that class (if so, 1) or not (if not, 0).

[8] A library for SVMs - Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

After the bilingual suffix class for a translation is determined based on the split, new translations are suggested by applying the suffix replacement rules to the identified bilingual stem. For example, given a new bilingual pair *'dilemmas'* ⇔ *'duvidhAein'* (Figure 1), the bilingual suffix resulting from segmentation is ('s', 'Aein'). As ('s', 'Aein') is classified as belonging to the bilingual suffix class ('', 'A'), ('s', 'Aein'), the new translation is generated by replacing *'s'* with *''* and *'Aein'* with *'A'*, giving rise to the new bilingual variant 'dilemma' ⇔ *'duvidhA'*.

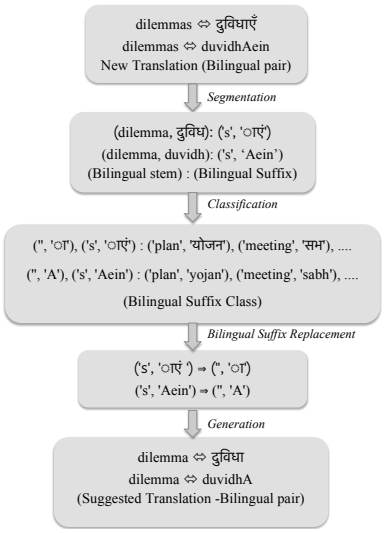

Figure 1: Sample generation

3.4 Longest Bilingual Suffix Match (LBSM)

The LBSM technique is used as baseline for identifying bilingual suffixes. After the learning phase, we have different sets of bilingual stems that have been grouped according to their bilingual inflectional classes. We call such sets as Bilingual Suffix Classes. For each translation in the test set, we wish to determine their bilingual inflections (suffixes) and the associated bilingual suffix class. As baseline, we classify each new (unseen) translation in the test set into the class of longest matching bilingual suffix from the bilingual suffix list. For instance, the longest bilingual suffix matching the bilingual pair *'conservative'* ⇔ *'rakshAtmak'* is *'ative'* ⇔ *'Atmak'* yielding the bilingual stem *'conserv'* ⇔ *'raksh'*.

4 Experimental Results and Discussion

4.1 Data set

We used bilingual pairs taken from EN-HI bilingual lexicon representing single-word translations as the

training data set. Approximately 90% of the entries in the lexicon were acquired from the dictionary[9]. The remaining (10%) entries were partly compiled manually and partially using the Symmetric Conditional Probability based statistical measure from the aligned parallel corpora[10] (Da Silva and Lopes, 1999). The details are as shown in the Table 3.

Description	Total	Training	Test
Bilingual Pairs	58,048	52K	6K
Minimum Length (EN-HI)	3, 3		
Maximum Length (EN-HI)	18, 10		

Table 3: Statistics of the Data set

4.2 Bilingual Learning and Generation

The bilingual suffixes (frequently undergoing transformations) recognised using the approach discussed in Section 3.1 are shown in Table 4. Table 5[11] presents the bilingual suffix transformation rules which enable one translation form to be obtained using the other. The grouping in row 1 implies that replacing the suffix *'s'* with *''* and the suffix *'oM'* with *'A'* in the bilingual pair *'boys'* ⇔ *'laDakoM'*, yields its bilingual variant *'boy'* ⇔ *'laDakA'*.

Bilingual Suffixes	Bilingual Suffixes (Hindi Suffixes transliterated)	Frequency (bmf)	Generative Strength (gs)
('', 'ी')	(‘’, ‘I’)	10,743	11,240
('', 'ा')	(‘’, ‘A’)	29,529	30,635
('ion', 'ा')	(‘ion’, ‘A’)	457	567
('er', 'ा')	('er', ‘A')	428	515
('ity', 'ा')	('ity', ‘A')	286	340

Table 4: Bilingual Suffixes with frequent replacements

To avoid over-segmentation, we perform the suffix containment check, looking for one candidate bilingual suffix enclosed within another. A true compound bilingual suffix (a combination of multiple candidate bilingual suffixes) is retained based on the observation that the strength of a compound bilingual suffix is less than the strengths of the bilingual suffixes composing it (Dasgupta and Ng, 2007).

Evaluation A few of the induced bilingual suffix class based morphological patterns are incomplete as not all the translation forms are seen in the lexicon. Further, distinct surface translation forms due to inflection classes result in distinct bilingual suffix classes some of which should be collapsed.

[9]http://sanskritdocuments.org/hindi/dict/eng-hin˙unic.html, www.dicts.info, hindilearner.com

[10]EMILLE Corpus - http://www.emille.lancs.ac.uk/

[11]*Number of times a bilingual suffix co-occurs with another bilingual suffix in the input lexicon (Mahesh et al., 2015)

Bilingual Suffixes	Suffix pair Co-occurrence Score*	Bilingual Stems	
('', 'ा'), ('s', 'ोे') ('', 'A'), ('s', 'oM')	27	('plant', 'पौध') ('plant', 'paudh')	('boy', 'लड़क') ('boy', 'laDak')
('', 'ी'), ('s', 'ोे') ('', 'I'), ('s', 'oM')	27	('job', 'नौकर') ('job', 'naukar')	('archer', 'धनुषधार') ('archer', 'dhanuShadhaar')
('s', 'ोे'), ('ous', 'ी') ('s', 'oM'), ('ous', 'I')	8	('mountain', 'पर्वत') ('mountain', 'parvat')	('mountain', 'पहाड') ('mountain', 'pahAD')
('', 'ा'), ('s', 'ाएं') ('', 'A'), ('s', 'AeM')	3	('plan', 'योजन') ('plan', 'yojan')	('meeting', 'सभ') ('meeting', 'saB')

Table 5: Highly (top 2), less (bottom 2) frequent bilingual suffix replacement rules

We evaluate the bilingual segments and clustering results indirectly by examining the applicability of induced segments in generating new translations. We first complete the translation lexicon with missing bilingual pairs using bilingual stems and bilingual suffixes learnt using the known bilingual pairs. Generation of missing translation is purely concatenative and is done using the translations in the training data for the chosen bilingual suffix classes (Mahesh et al., 2014). The generated translations are then evaluated. Table 6 shows the results of the learning phase. We calculate the precision for generated translations as the fraction of correctly generated bilingual pairs to total number of bilingual pairs generated. In completing the translation lexicon for missing forms, when both bilingual stems and bilingual suffixes are known, the precision achieved for translation generation reaches 86.52% when compared to the precision of 81.31% obtained using the bilingual learning approach (Mahesh et al., 2014).

Learning Approach	Unique Bilingual Stem Count	Unique Bilingual Suffix Count	Number of Clusters	Generation Precision
IDA2014 (Kavitha *et al.*, 2014)	12,603	781	224	81.31
Proposed-Phase 1	10,224	426	143	86.52

Table 6: Clustering statistics

Table 7[12] shows suggested translation examples. We categorise the generated translations into 3 classes (separated by thick border) based on the degree of correctness. First 3 rows represent acceptable translations (Accept). The following row shows

[12]Two bilingual suffixes are shown per class, though they range from 2 to 5

translation errors (Reject) and the last row represents an inadequate translation (Inadequate). Mentioned errors are briefly explained below:

Inadequate: The bilingual pair *'Russians'* ⇔ *'rUsiyoM'* (last row of the Table 7) is inadequate, as in actual usage, both the singular and plural variants *'Russian'* and *'Russians'* are translated as *'rUsI'*. An alternate correct translation would be *'rUs vAsI'*.

Generated Translations	Existing Lexicon Entry	Rule used
cleverly ⇔ निपुणता	cleverness ⇔ निपुणता	('ly', 'ता'), ('ness', 'ता')
capitalist ⇔ पूँजीवादी materialist ⇔ भौतिकवादी	capitalism ⇔ पूंजीवाद materialism ⇔ भौतिकवाद	('ism', 'वाद'), ('ist', 'वादी')
framework ⇔ ढाँचा	frameworks ⇔ ढाँचे	('', 'ा'), ('s', 'े')
world ⇔ लौक (lauk) weeks ⇔ साप्ताहों (sAptahoM)	worldly ⇔ लौकिक (laukik) weekly ⇔ साप्ताहिक (sAptahik)	('ly', 'िक), ('s', 'ोों')
Russians ⇔ रूसियों (rUsiyoM)	Russian ⇔ रूसी (rUsI)	('ian', 'ी'), ('ians', 'ियों')

Table 7: Generated Translations

Reject: Incorrect generations are a result of incorrect generalisations. Typical errors correspond to irregular translation forms, specifically, the stem changes before suffixation and misclassifications due to insufficient translation forms. An example for the former class of errors is the generated translation *'world'* ⇔ *'lauk'* (row 5), as the correct translated form should be *'world'* ⇔ *'lok'*. The surface variant *'worldly'* ⇔ *'laukik'* is obtained from the stem pair *'world'* ⇔ *'lok'* by appending *'ly'* ⇔ *'ik'* at the end of the word pair 'world' ⇔ *'lok'*. Further, the stem undergoes a change from *'o'* to *'au'*.

Our approach being purely bilingual suffixation based, does not handle irregular forms and does not capture stem changes prior suffixation.

4.3 Minimally supervised learning

The results of segmentation by classification were indirectly evaluated by examining what the induced bilingual segments is expected to facilitate, specifically, in suggesting or generating new translations. In evaluating the generated translations, the Precision (P), Recall (R) and F-measure (F_m) are computed as given below:

$$P = t_p/(t_p + f_p),\ R = t_p/(t_p + f_n),$$
$$F_m = 2 * P * R/(P + R) \quad (5)$$

where, t_p denotes the number of times the generated translations were correct, f_p denotes the number of times the generated translations were incorrect and f_n denotes the number of times a possible correct translation suggestion was missed. The results for various features are shown in Table 8. When new translations are given as inputs, the best f-measure of 70.88% is achieved.

Features	Precision	Recall	F-measure
Longest Bilingual Suffix Match	74.41	47.32	57.85
NSV + NPV + BO + Stand-alone pair	75.23	52.54	61.87
NPE + NSE + BO + Stand-alone pair	70.14	57.22	63.02
BMF + GS + CBSP + CBSL + BO + Stand-alone pair	76.21	66.24	70.88

Table 8: Results of minimally supervised learning

5 Conclusion and Future Work

We have discussed a minimally supervised approach for learning bilingual segments. The training data prepared using the bilingual learning approach partially serves as the basis for segmentation along with the bilingual lexicon (Mahesh et al., 2014). Various measures used in word segmentation tasks are used as features to represent a boundary/non-boundary condition in a bilingual framework. The segmentation boundary identified for a bilingual pair during the learning phase represent a positive sample and all other possible segmentation options for the bilingual pair represent negative samples. Experiments with distant language pairs and limited training data show that knowing both bilingual stems and bilingual suffixes, missing forms could be generated with the precision of 86.52%. For new translations, the precision falls by 10%.

As future work, direct evaluations should be done by comparing the learned bilingual segments and suffix classes to those in the grammar descriptions for the language pairs under consideration. Learning from bigram equivalents to predict translations for verb forms shall be addressed in the future work.

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, the funded research project ISTRION (Ref. PTDC/EIA-EIA/114521/2009) that provided other means for the research carried out. The authors thank NOVA LINCS, FCT/UNL for providing partial financial assistance to participate in PACLIC 2015, and ISTRION BOX - Translation & Revision, Lda., for providing the valuable consultation.

References

Stefan Bordag. 2008. Unsupervised and knowledge-free morpheme segmentation and analysis. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 881–891. Springer.

Tomáš Brychcín and Miloslav Konopík. 2015. HPS: High precision stemmer. *Information Processing & Management*, 51(1):68–91.

Çağrı Çöltekin. 2010. Improving successor variety for morphological segmentation. *LOT Occasional Series*, 16:13–28.

Mathias Creutz and Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology.

Joaquim Ferreira Da Silva and Gabriel Pereira Lopes. 1999. Extracting multiword terms from document collections. In *Proceedings of the VExTAL: Venezia per il Trattamento Automatico delle Lingue*, pages 22–24.

Sajib Dasgupta and Vincent Ng. 2007. Unsupervised word segmentation for bangla. *Proceedings of ICON*, pages 15–24.

Hervé Déjean. 1998. Morphemes as necessary concept for structures discovery from untagged corpora. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, pages 295–298. Association for Computational Linguistics.

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

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational linguistics*, 27(2):153–198.

Margaret A Hafer and Stephen F Weiss. 1974. Word segmentation by letter successor varieties. *Information storage and retrieval*, 10(11):371–385.

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350.

Harald Hammarström. 2009. *Unsupervised learning of morphology and the languages of the world*. Ph.D. thesis, Chalmers University of Technology and Göteborg, Gothenburg, December.

Zellig S Harris. 1970. *From phoneme to morpheme*. Springer.

Amit Kirschenbaum. 2015. To split or not, and if so, where? Theoretical and empirical aspects of unsupervised morphological segmentation. In *Computational Linguistics and Intelligent Text Processing*, volume 9041 of *LNCS*, pages 139–150. Springer.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. HFST tools for morphology – An efficient open-source package for construction of morphological analyzers. In *State of the Art in Computational Morphology*, volume 41 of *CCIS*, pages 28–47. Springer.

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

Kavitha Karimbi Mahesh, Luís Gomes, and José Gabriel P Lopes. 2015. Learning clusters of bilingual suffixes using bilingual translation lexicon. In *Mining Intelligence and Knowledge Exploration (Accepted)*. Springer.

Christian Monson, Jaime Carbonell, Alon Lavie, and Lori Levin. 2009. Paramor and morpho challenge 2008. In *Evaluating Systems for Multilingual and Multimodal Information Access*, volume 5706 of *Lecture Notes in Computer Science*, pages 967–974. Springer.

Karthik Narasimhan, Regina Barzilay, and Tommi Jaakkola. 2015. An unsupervised method for uncovering morphological chains. *ArXiv preprint arXiv:1503.02335*.

Amaresh Kumar Pandey and Tanveer J Siddiqui. 2008. An unsupervised hindi stemmer with heuristic improvements. In *Proceedings of the second workshop on Analytics for noisy unstructured text data*, pages 99–105. ACM.

Pratikkumar Patel, Kashyap Popat, and Pushpak Bhattacharyya. 2010. Hybrid stemmer for gujarati. In *23rd International Conference on Computational Linguistics*, page 51.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 209–217. Association for Computational Linguistics.

Al-Shalabi R., Ghassan Kannan, Iyad Hilat, Ahmad Ababneh, and Ahmad Al-Zubi. 2005. Experiments with the successor variety algorithm using the cutoff and entropy methods. *Information Technology Journal*, 4(1):55–62.

Hisayuki Sasaoka, Kenji Aaraki, Yoshio Momouchi, and Koji Tochinai. 1997. Prediction method of word for translation of unknown word. In *Proceedings of the IASTED International Conference, Artificial Intelligence and Soft Computing, Banff, Canada*, page 228. Acta Pr.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. *ACL-08: HLT*, page 737.

Benno Stein and Martin Potthast. 2007. Putting successor variety stemming to work. In *Advances in Data Analysis*, pages 367–374. Springer.

Richard Wicentowski and David Yarowsky. 2002. *Modeling and learning multilingual inflectional morphology in a minimally supervised framework*. Ph.D. thesis, Ph. D. Thesis. Johns Hopkins University, Baltimore, Maryland.

Daniel Zeman. 2008. Unsupervised acquiring of morphological paradigms from tokenized text. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 892–899. Springer.