Kavitha Karimbi Mahesh

Master of Computer Applications

Augmenting Translation Lexica by Learning Generalised Translation Patterns

Dissertação para obtenção do Grau de Doutor em

Informática

Orientador: Doutor José Gabriel Pereira Lopes,
Investigador Principal Aposentado,
Universidade Nova de Lisboa

Júri

Presidente:	Doutor Nuno Manuel Robalo Correia
Arguentes:	Doutor Pavel Bernard Brazdil
	Doutor Pablo Gamallo Otero
Vogais:	Doutor Paulo Miguel Torres Duarte Quaresma
	Doutor Joaquim Francisco Ferreira da Silva
	Doutor José Gabriel Pereira Lopes

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Junho, 2017

Augmenting Translation Lexica by Learning Generalised Translation Patterns

To all the great personalities who adopt the scientific knowledge for the betterment of mankind.

Acknowledgements

It is a great pleasure for me to put on record a deep sense of gratitude to my guide Dr. José Gabriel Pereira Lopes, Principal Investigator, Department of Informatics (DI), Faculty of Science and Technology, New University of Lisbon (FCT-UNL), Portugal for his meritorious guidance, invaluable suggestions, meticulous reviews during the tenure of my research work as well as during the preparation of this thesis. His persistent encouragement, scholarly advice and constructive criticism has been an immense source of inspiration for me to carry out my research work. His scientific foresight, dynamism and leadership capacity have been instrumental in motivating me to take up research particularly in the field of translation classification and bilingual morphology learning, and I am fortunate enough to get the rare opportunity to work under him in this exciting field. The memory of my association with him will be cherished for ever. I am grateful to him for patiently teaching me the basic research techniques and providing me all the necessary facilities and every possible help for the completion of my research work and this thesis.

I take this opportunity to express my gratitude to the PhD panel committee members, Dr. Pavel Bernard Brazdil, Professor, FE-UP, Dr. Pablo Gamallo Otero, Investigador 'Ramón y Cajal', Faculdade de Filologia, Universidade de Santiago de Compostela, Dr. Joaquim Francisco Ferreira da Silva, Assistant Professor, DI, FCT-UNL and Dr. Paulo Miguel Torres Duarte Quaresma, Associate Professor, Department of Informatics, University of Évora for their valuable observations, suggestions and guidance which helped me to plan my work and contributed in improving the quality of my thesis.

I would like to acknowledge the help extended by Dr. Luís Manuel Marques da Costa Caires, Director, NOVA LINCS-Laboratory for Computer Science and Informatics, and HOD-DI, FCT-UNL during various phases of my research work. I am thankful to Dr. Nuno Manuel Robalo Correia, Professor, DI, FCT-UNL for his support in various capacities.

I acknowledge the PhD research fellowship grant by Fundação para a Ciência e Tecnologia (FCT/MCT) with Ref. no., SFRH/BD/64371/2009 and the funded research project ISTRION (Ref. PTDC/EIA-EIA/114521/2009) which facilitated me to initiate research in this novel area. Thanks are due to FCT/UNL for providing me the financial assistance in attending the conferences. This work was also supported by NOVA LINCS (ref. UID/CEC/04516/2013). I am sincerely thankful to these organisations.

It is my pleasure to acknowledge the help given by all the officers and the staff of the university administration, the Librarian and the staff of the University library and the staff of DI, for their kind co-operation and help in various capacities.

Thanks are due to my colleagues at NOVA-LINCS, and particularly to Luís Manuel dos Santos Gomes for his co-operation, collaboration and constructive discussions. Concern and timely help by Dr. Filipa Peleja will be remembered always.

I am ever thankful to Rev Fr Joseph Lobo, Director, St. Joseph Engineering College (SJEC), Mangaluru, Rev Fr Rohith D'costa, Assistant Director, SJEC and Dr Joseph Gonsalvis, Principal, SJEC, for their consistent support and encouragement whenever approached, ever since my tenure in SJEC. I express my deep sense of gratitude to Dr Rio D'Souza, Vice Principal and Dean MCA, SJEC for his understanding, timely help and the much needed support extended at various stages of my research after being part of SJEC. I thankfully acknowledge the help extended by Ms Sumangala N, in-charge HOD-MCA and I humbly thank all my colleagues in the Department of MCA, SJEC for their support and good wishes.

The emotional support and constant encouragement in every respect by my beloved mother, respectable father and sister will be cherished for ever. I am always thankful to my mother-in-law for all her support and blessings. I will always be indebted to my beloved husband Dr. Mahesh K. K. for standing by me through all walks of my life and helping me in maintaining a positive and balanced outlook all along the research tenure. He is instrumental in teaching research basics and inculcating me with perseverance throughout. It is apt to mention here that their affection and patience during these years have made this thesis a reality.

I take this opportunity to thank all those who have helped me in various capacities at different stages during the tenure of my research work but whom I have missed to mention by name.

Last, but not the least, I would like to acknowledge all the researchers who have made significant contributions to the field of machine learning and its applications, and the publishers who have made available all these important information, which helped me immensely in planning and carrying out the present thesis work.

Abstract

Bilingual Lexicons do improve quality: of parallel corpora alignment, of newly extracted translation pairs, of Machine Translation, of cross language information retrieval, among other applications. In this regard, the first problem addressed in this thesis pertains to the classification of automatically extracted translations from parallel corpora-collections of sentence pairs that are translations of each other. The second problem is concerned with machine learning of bilingual morphology with applications in the solution of first problem and in the generation of Out-Of-Vocabulary translations.

With respect to the problem of translation classification, two separate classifiers for handling multi-word and word-to-word translations are trained, using previously extracted and manually classified translation pairs as correct or incorrect. Several insights are useful for distinguishing the adequate multi-word candidates from those that are inadequate such as, lack or presence of parallelism, spurious terms at translation ends such as determiners, co-ordinated conjunctions, properties such as orthographic similarity between translations, the occurrence and co-occurrence frequency of the translation pairs. Morphological coverage reflecting stem and suffix agreements are explored as key features in classifying word-to-word translations. Given that the evaluation of extracted translation equivalents depends heavily on the human evaluator, incorporation of an automated filter for appropriate and inappropriate translation pairs prior to human evaluation contributes to tremendously reduce this work, thereby saving the time involved and progressively improving alignment and extraction quality. It can also be applied to filtering of translation tables used for training machine translation engines, and to detect bad translation choices made by translation engines, thus enabling significative productivity enhancements in the post-edition process of machine made translations.

An important attribute of the translation lexicon is the coverage it provides. Learning suffixes and suffixation operations from the lexicon or corpus of a language is an extensively researched task to tackle out-of-vocabulary terms. However, beyond mere words or word forms are the translations and their variants, a powerful source of information for automatic structural analysis, which is explored from the perspective of improving word-to-word translation coverage and constitutes the second part of this thesis. In this context, as a phase prior to the suggestion of out-of-vocabulary bilingual lexicon entries, an approach to automatically induce segmentation and learn bilingual morph-like units by identifying and pairing word stems and suffixes is proposed, using the bilingual corpus of translations automatically extracted from aligned parallel corpora, manually validated or automatically classified. Minimally supervised technique is proposed to enable bilingual morphology learning for language pairs whose bilingual lexicons are highly defective in what concerns word-to-word translations representing inflection diversity. Apart from the above mentioned applications in the classification of machine extracted translations and in the generation of Out-Of-Vocabulary translations, learned bilingual morph-units may also have a great impact on the establishment of correspondences of sub-word constituents in the cases of word-to-multi-word and multi-word-to-multi-word translations and in compression, full text indexing and retrieval applications.

Keywords: Translation lexicon filtering, SVM classification, bilingual morph-units, translation selection, classification, OOV lexicon entries, clustering, translation suggestion.

Resumo

Os léxicos bilingues são de extrema importância em diversas aplicações, permitindo aumentar a qualidade dos resultados obtidos com a sua utilização designadamente em: alinhamento de corpora paralelo; extracção automática de novas entradas para esses léxicos, tradução automática, etc. É neste contexto que se enquadra a solução de dois problemas que abordo nesta tese. O primeiro é um problema de classificação automática de traduções que foram extraídas automaticamente a partir de corpora paralelo (colecções de pares de frases, traduções umas das outras). O segundo é um problema de aprendizagem de morfologia bilingue com aplicações na resolução do primeiro problema e na geração de traduções de palavras inexistentes nos léxicos de que se partiu.

Relativamente ao problema de classificação automática de traduções são treinados dois tipos de classificadores: um para tratar de traduções de palavras simples que sejam palavras simples; e outro para os casos de traduções de multi-palavras. Em qualquer dos casos usam-se traduções previamente extraídas e manualmente etiquetadas como correctas ou incorrectas. Utilizam-se ainda propriedades dessas traduções como sejam a presença ou ausência de palavras de conteúdo não traduzidas e de artigos no final das traduções, semelhança ortográfica entre as traduções e frequência de ocorrência e de co-ocorrência. No caso das traduções de palavras simples por palavras simples, utiliza-se ainda a presença ou ausência de radicais e sufixos cujas correspondências foram aprendidas na resolução do segundo problema. A resolução deste problema tem um impacto enorme na produtividade de validadores humanos de traduções automaticamente extraídas, no aumento da velocidade e na qualidade dos processos de realinhamento subfrásico de frases paralelas e na qualidade de novas extracções de traduções. Tem ainda aplicações na filtragem de de tabelas de tradução utilizadas em tradução automática e na detecção de más escolhas de tradução feitas por motores de tradução, podendo assim contribuir para aumentos de produtividade no processo de pós-edição de traduções feitas pela máquina.

O segundo problema abordado nesta tese, de aprendizagem de morfologia bilingue, generaliza, ao nível bilingue, a aprendizagem que se faz a nível monolingue de sufixos e de operações de sufixação. Explora-se num léxico bilingue as diversas traduções de cada palavra e das suas possíveis formas e induz-se a segmentação em morfemas das palavras simples e das suas traduções (também palavras simples), identificando e estabelecendo um emparelhamento entre radicais e sufixos de palavras, utilizando técnicas de agrupamento (clustering) e classificação, utilizando também os corpora paralelos em que ocorrem. Isto é, descobre-se a tradução de radicais e de sufixos em cada par de línguas e, além disso, faz-se o agrupamento de sufixos visando uma geração de traduções fora-do-vocabulário com elevado grau de precisão no que respeita à flexão morfológica tanto das palavras de uma língua como das suas traduções na outra. No caso de línguas em que os léxicos bilingues disponíveis sejam altamente defectivos quanto à variedade e diversidade flexional das palavras e das suas traduções, propõe-se uma técnica de supervisão mínima para ultrapassar a falta de variedade. Além das aplicações já mencionadas de classificação automática de traduções e de geração de traduções inexistentes nos léxicos bilingues de que se parte, este trabalho terá ainda um grande impacto no estabelecimento de correspondências entre os constituintes sub-palavra de traduções de multi-palavras por palavras simples ou por multipalavras e em aplicações de compressão, indexação e pesquisa de texto.

Palavras-chave: Filtragem de léxicos de tradução, classificadores SVM, unidades morfológicas bilingues, selecção de traduções, classificação, geração de entradas lexicais bilingues inexistentes nesses léxicos, agrupamento, sugestão de traduções.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Context

Translation of texts from one language to another with the computer assistance has evolved substantially in the past few decades emerging as one of the major applications of Natural language processing and Artificial Intelligence. Direct, inter-lingual and transfer-based translation approaches were historically predominant. Owing to the richness of natural languages and the complexities involved in their analysis, machine translation led to new approaches, such as statistical and example-based, and this development was enabled by the growing availability and accessibility to large corpora of translated texts [Koe10]. Two texts are parallel if they are translations of each other. Such parallel texts being good sources of information to build bilingual translation dictionaries have proven to be crucial in training Statistical Machine Translation (SMT) systems.

Initial works on SMT considered word to word translation as a basic procedure [Bro+93]. But, soon the preference for phrase-based models [ON04] [Lop08] advanced the state of the art in SMT systems as it was evident that phrases (contiguous sequence of words) conveyed information that surpassed the meanings of the individual words in the expressions. Both word and phrase-based models incorporate as its essential component, an alignment function that maps words on either sides in a sentence pair. All phrase pairs that are consistent with the word alignment are extracted and compiled into a phrase table along with their associated probabilities, which in turn is utilised by a decoder for the translation of new text. Selection of appropriate translation pairs is done during the translation process by the decoder.

Unsupervised translation models always learn everything from scratch and has not evolved as much as one would expect it to be. We on the other hand believe that, no men or machine can ever learn and evolve without being corrected, without remembering where it /what failed earlier and where it acted correctly and same holds to Machine Translation systems as well. An approach in-line with this perspective, thus deviating from *completely automated training of translation models*, requires that extracted translation equivalents are validated [Air+09]. In such a scheme, a phrase-based aligner [GL09] is employed to align parallel texts at sub-sentence level using a bilingual translation lexicon. After the translations are extracted from aligned texts, they are manually validated and classified as 'accepted' or 'rejected' by the human evaluators and are added to the lexicon. Validation at this point is important as the extracted translations are used for re-alignment and translation extraction in subsequent iterations. Thus, to keep the alignment errors from being fed back to subsequent iterations, the extracted translations are validated. The approach hence involves a cycle of processes, viz., *alignment*, *extraction of yet unknown word or phrase translations*, and their *subsequent validation* and this series terminates when no further gains are achieved.

The approach being semi-supervised and iterative assumes two-fold advantages:

- an improved alignment precision while reducing uncertainty
- acquisition of more reliable translation lexicon.

As a matter of fact, the extracted translations are more reliable due to the human supervision involved at various levels of processing in the acquisition of the final translation phrase table.

Table 1.1: Sub-sentence Alignment with Known Correspondences

Source Segment (EN)	**Known**	**Target Segment (PT)**
use	*	utilizam
the	*	as
following	*	seguintes
bridging		
tables	*	tabelas
in their regular monitoring		
of	*	de
		correspondência
the	*	a
		o controlarem regularmente a
consistency	*	coerência
between	*	entre
the	*	a
eurosystem's		posição em
end	*	fim
...		...

The use of validated bilingual lexicon can be perceived from the Tables 1.1 through 1.3, illustrating how an ISTRION concordancer evolves with the acquired knowledge at various levels of processing. The entries marked with '*' in the column 'Known' represent known (correct) translations that already exist in the bilingual lexicon used for parallel corpora alignment. It might be observed that the accumulated knowledge also evolves over time (at various iterations of processing) which is used for re-alignment in subsequent iterations. As a matter of fact, the evolution from Table 1.1 to 1.2, was a consequence of manually entering *'in their regular monitoring of'* as a translation of *'a o controlarem regularmente'* that already appear in Table 1.1 aligned. From Table 1.1, we know that the aligner already knows that *'tables'* translate as *'tabelas'* and from Table 1.2, that it also knows that *'bridging'* can be translated as *'correspondência'*. However it did not know that *'bridging tables'* may translate as *'tabelas de correspondência'*. Some time later, the system had already extracted this translation equivalence using the 'fishnet method' described in the PhD Thesis of Luís Gomes [Gom16] and Table 1.3 is a result of incorporation of more knowledge, showing less alignment errors than the previous two tables.

Table 1.2: Sub-sentence Alignment with Known Correspondences at a later stage

Source Segment (EN)	**Known**	**Target Segment (PT)**
use	*	utilizam
the	*	as
following	*	segunites
bridging	*	tabelas de correspondência
tables		
in their regular monitoring of	*	a o controlarem regulamente
the	*	a
consistency	*	coerência
between	*	entre
the	*	a
eurosystem's		posição em
end	*	fim
...		...

While the supervision involved contributes towards a reliable lexicon, it would be important to speed up the process of categorising translation extractions as correct or incorrect by devising a linguistically-informed method. Automatic categorisation and hence selection of translation candidates is feasible when large amount of labelled positive and negative evidence is available. The availability of sufficient human-annotated translations marking their correctness/incorrectness leaves much scope for automating the validation process, which is one of the core areas of my work. My perspective is

Table 1.3: Sub-sentence Alignment with Known Correspondences at a much later stage

Source Segment (EN)	**Known**	**Target Segment (PT)**
use	*	utilizam
the	*	as
following	*	segunites
bridging tables	*	tabelas de correspondência
in their regular monitoring of	*	a o controlarem regulamente
the	*	a
consistency	*	coerência
between	*	entre
the	*	a
eurosystem's		posição em
end	*	fim
...		...

that, we can speed up the process of determining adequate translations for subsequent alignment iterations by mining the human-annotated translations to discover the nature of correct translations versus the extraction errors.

Classification of automatically extracted translations, by learning from previously extracted translations, that have been meanwhile classified as correct or incorrect by human users, and kept in a lexicon database, is the first problem I will address in this Thesis. This will be better explained in the Chapters 2, 3, 4 and 5 of Part I.

Henceforth, I introduce the context and challenges of my research pertaining to the another aspect of bilingual translation lexicons, i.e., its coverage. Translation of texts from one language to another is feasible and convenient when a term in first language appears in the translation lexicon of the system with its corresponding translation. Translating terms that are not registered in the system's lexicon or translation table could be challenging, as MT systems rely substantially on the translation table. An important characteristic of a translation lexicon, is therefore, the coverage it provides. However, several factors influence the translation lexicon coverage.

The coverage of a lexicon acquired automatically from parallel corpora might be *influenced by* and *restricted to* word or phrasal forms explicitly existing in the corpus being utilised for lexicon acquisition. No corpora will ever contain all possible translation patterns for any language pair and will not enable that all possible sentences will ever be produced. The difficulties in acquiring large-coverage lexicons is easily attributable to the problem of data sparseness and so requires careful attention to deal with it. For example, a word (*noun or verb*) in English (en) can have several different forms and it is unlikely that all the different forms of that word are seen during the translation extraction process. To instantiate, while looking for all the English forms acquired which include variations of word *demonstrate*, entries registered in our lexicon included those listed in Table 1.4.

Table 1.4: Translations containing some inflected forms of word *demonstrate*

Source Term (en)		Target Term (pt)
demonstrating	⇔	que demonstrem
demonstrated	⇔	demonstrou
as is demonstrated	⇔	como o provam
no toxicity was demonstrated	⇔	não se observe toxicidade
be demonstrated	⇔	demonstrar
pilot and demonstration projects	⇔	projectos - piloto e de demonstração
demonstration	⇔	demonstração
demonstration	⇔	a demonstração

Further, the term mostly appears in the context of larger text. However, no translations for independent word forms (single-word entries) such as, *demonstrate* or *demonstrates* exist in our lexicon.

Moreover, a term in English can have multiple translations in Portuguese and the scenario with the language pair English-Hindi is not different. Translation forms shown in Table 1.5 illustrates this for EN-HI, where the word *‘good’* (an adjective) in English translates as *‘acChA’*, *‘acChI’* and *‘acChe’* in Hindi. This is an instance of the source-target asymmetry, commonly seen in translations involving a morphologically rich language when the other language is morphologically poor. Thus, a lexicon should also be extensive with respect to the second language vocabulary by providing complete set of lexically and grammatically correct translations for every possible term in first language.

If a particular form is missing in the lexicon, the translation system is constrained to use only those forms that are recorded in the lexicon, unless the system is able to infer missing forms. In other words, if a MT system assumes different morphological forms of the word as independent entities, it is necessary to register all the possible translation patterns in the lexicon. As all the forms are hardly seen in the training data (or recorded in the translation table), this in turn triggers the need for identifying morphological similarities in the known example pairs[1]. In the referred example, the three translation forms, *‘good’* ⇔ *‘acChA’*, *‘good’* ⇔ *‘acChe’* and *‘good’* ⇔ *‘acChI’* share common bilingual segments, *‘good’* ⇔ *‘acCh’*, together with their bilingual extensions constituting dissimilar bilingual segments, *‘’* ⇔ *‘A’* | *‘e’* | *‘I’*. These bilingual extensions do appear as endings for other translation forms and hence serve in identifying bilingual suffix classes. In a broader sense, the separation of morphological suffixes from a bilingual pair conflates various forms of a translation, into a bilingual stem, which is a crucial source of information. On the other hand, the bilingual extensions (suffixes), that occur frequently with the translations belonging to similar class, could be utilised for generating unknown forms. Thus,

[1] As pointed out in [H.09], ‘words consist of high-frequency strings (‘affixes’) attached to low-frequency strings (‘stems’), as in the English play-ing’

Table 1.5: Possible multiple translations for words *boy* and *good*

ladDakA (boy)	acChA (good)
लड़का (ladDakA)	अच्छा (acChA)
लड़के (ladDake)	अच्छी (acChI)
	अच्छे (acChe)

using the bilingual morphological information, all the possible forms can be inferred by combining different component bilingual morphemes from different mappings learnt from the example bilingual pairs in the translation lexicon (training corpus). Thus, given that translation forms are composed of bilingual extensions inflecting a bilingual stem, and that any surface form can be obtained by productively combining such bilingual extensions with the bilingual stem, a system capable of analysing and generalising the frequently seen morph-like units, should enable inferring infrequent forms as well.

Yet another factor influencing the lexicon coverage concerns the limitations of the technique employed in translation extraction. Any particular extraction technique is not able to extract every possible translations or translation forms. For instance, the extraction method proposed by Aires et al. [Air+09], mainly focused on the translation extractions following pattern of the kind, expression aligned with nothing, followed by an expression aligned with another expression, followed by nothing aligned with another expression.

Table 1.6: Alignment and Term Translation Extraction

Source Segment (EN)	**Known**	**Target Segment (PT)**
Eurojust's		A Eurojust tem por missão apoiar
mission	*	missão
shall be to support		apoiar
and	*	e
strengthen		reforçar
coordination and cooperation	*	a coordenação e a cooperação
between	*	entre
national investigating and prosecuting		
authorities	*	as autoridades
		nacionais competentes para a investigação e o exercício de a acção penal
in relation to	*	em matéria de
		criminalidade
serious	*	grave
crime		
affecting	*	que afecte

Thus, for the aligned segments[2] depicted in Table 1.6, some of the possible term translations extractions include: *Eurojust's mission shall be to support ⇔ A Eurojust tem por missão apoiar*, *Eurojust's mission shall be to ⇔ A Eurojust tem por missão*, *support ⇔ apoiar*, *strengthen ⇔ reforçar*, *national investigating and prosecuting authorities ⇔ as autoridades nacionais competentes, para a investigação e o exercício de a acção penal*, *serious crime ⇔*

[2] * indicates previously known alignment

criminalidade grave [Air+09]. The approach taken alone is incapable of extracting other translation patterns that exist in the corpus. Thus, any particular technique for translation extraction has limitations with respect to the coverage provided.

Although the complete lexicon, the one that might register all word or phrase forms is impossible to attain, there have been various attempts to reduce the rate at which Out-of-Vocabulary (OOV) terms occurs in input. Each of these differ with respect to the approaches adopted, the type of translation units targeted, availability or accessibility of resources used for the purpose and their applicability across languages.

The usual way to deal with unseen segments during translation is either *to retain a copy* of such phrases in the target language or *to drop them out literally* [Koe+05]. He et al., [He+06] reconsidered translating named entities using a pre-translation support tool before discarding the remaining unknown terms. However, neither of the compromises for handling unknown words is acceptable, as they convey flawed interpretation when the translation of a larger portion of text containing unknown part is considered, thereby making translations less readable. If the lexicon (phrase table) of a MT system is not adequate, the quality of the whole system suffers. This determines how the presence of a phrase in, or its absence from, the lexicon affects the quality of MT system, which in turn will reflect in the overall output quality.

Transliteration rules have been suggested for translating special unknown words such as *proper nouns* and *technical terms* when the character set between source and target languages differ [Che+98; Hab08; Hua05; KG98; Zha+07].

Increased prevalence of monolingual and nonparallel, comparable electronic texts over parallel texts facilitated several approaches for learning translations for new words. The underlying idea behind these approaches is that, translations share similar contexts even in non-parallel corpora. As an example, to translate the word *stock*, the algorithm builds a context vector which might contain terms such as *capital*, *shares*, *business*, *money* etc. occurring in a four word window around the word *stock* [Rap95; Rap99] or in same sentence as the word *stock* [FY98] and the possible translations are the words in target language that have translations for these terms in surrounding context. The word pairs constituting the contexts are bridged using an incomplete bilingual lexicon and are referred as *seed words*. In extracting terminology translations, the study by Rapp [Rap95] revealed that association between a word and its close collocate are preserved in any language. Fung and Yee [FY98] further showed that association between a word and its context seed words are preserved in non-parallel comparable texts of different languages. They used *tf/idf* of contextual seed words (occurring in the same sentence as the word to be translated) to rank context words and similarity measures such as Dice, Jaccard coefficient to rank their translations. *Adjacent lexical words* were used by Rapp in building the context [Rap99]. Additional clues such as *frequency* and *orthographic similarity* were tested on top of the context in few of the other related works [KK02; SY02; TM99]. Haghighi et al. [Hag+08] suggested on learning a generative model from monolingual corpora using contextual and orthographic similarities for translating nouns. As an extension to the basic context model, Garera et al. [Gar+09] uses contexts derived from *tree positions* (parent and child relationships) consisting of head words in contrast to the contexts derived from adjacent lexical words. Use of *dependency parses* allowed dynamic context sizes, alleviated the reordering issues during vector projection for languages with different word orders. While this enabled significant improvement over the baseline approach, use of part-of-speech clustering allowed further improvements by expanding the scope of word types translated.

To handle unknown words encountered during decoding Chen et al. [Che+06], suggested dictionary lookup, by registering all found translations in the phrase table along with a certain probability. The translation model is then expected to choose the best translation. Remaining unknown words are handled using a rule-based Chinese-English translation system. As a final resort to handle the left unknown words, Eck et al. [Eck+08] suggests on finding translations for the *definitions of unknown words* in the source language dictionaries. The improved coverage and the easier accessibility to monolingual resources such as dictionaries and encyclopaedias makes the approach attractive, but remains limited by issues concerning the definitions that in turn end up containing unknown words. *First definitions* that conveyed most common meaning and those definitions that contained fewer numbers (less than 3) of unknown words is seen to be favorable, thereby demanding for a *definition selection task*. Moreover, *Specialized systems* are needed in situations where the style of definitions differ from the domain of actual translation. However, it is not given that adequate word forms are found.

“Whereas translation represents the preservation of meaning when an idea is rendered in the words of a different language, paraphrasing represents the preservation of meaning when an idea is expressed using different words in the same language" states Callison-Burch [CB07]. Recent studies [CB+06; GG08] have attempted to tackle the problem of unknown words or phrases by paraphrasing the source side with texts whose translations are known to the system. On the other hand, Cohn and Lapata [CL07] explored the use of a bridge language, that is different from the target language, for translating unknown words. As the paraphrase extraction procedures resemble the standard phrase extraction in SMT systems, large volumes of parallel data are required.

One of the other approaches attempted to handle the translations for unknown words relies on examples, referred to as the proportional analogy technique. Research reviews based on proportional analogies [LD05; Som+09], however point to some of the limitations that prevent the relevance of the approach in arriving at an adequate translation model. Increased processing times and over generations under increased example conditions are some of the major issues. However, the study suggests its appropriateness in handling special unknown words, such as named entities or technical terms. This is justified by the success of the experiments in proposing translations for ordinary unknown words, given that the validity of translation is around 80% as reported by Arora et al. [Aro+08]. Handling the input as strings of morphemes rather than as strings of words or characters is suggested as a possible alternative so as to deal with the underlying issues whenever proportional analogy is opted.

Transliteration based approaches discussed in the beginning address unknown terms that should be transliterated with their applicability restricted to specific types of unknown terms such as named entities. Some of the other approaches presented thereon either need comparable or large volumes of monolingual corpora and dictionaries.

Lexical inferences or morphological processing techniques [Gis+05; MT97; YK06] have been established to be interesting in handling unknown terms that are variations of known forms. These approaches have shown to provide good results in predicting translations and improving the translation quality, by specifically attending to unknown terms that are variants of known terms, requiring that reasonably close words are in the translation lexicon or a parallel corpus.

Learning suffixes and suffixation operations from a lexicon or corpus of a language improves word coverage by allowing new word forms to be generated [SK09]. However, beyond mere words/word forms are the translations or bilingual pairs, a powerful source of information for automatic structural analysis. In this context, taking advantage of the bilingual translation lexicon and extending the learning mechanism to those bilingual data, the challenge is to learn morphological similarities from translation examples, identify the mappings between morph units (mapping between bilingual stems and affixes) observed in the known translation forms across languages. Specific challenges include tackling one-to-one, one-to-many and many-to-many translation forms for identifying and generalising translation patterns. Focusing on one-to-one translations, the fundamental tasks and challenges to be tackled include the following:

- Morphological decomposition of bilingual pairs (translations) into the basic units of meaning or analysis of translations into bilingual morph-like units. For example, identifying the EN-PT bilingual pair, *ensuring* ⇔ *assegurando* and *ensured* ⇔ *assegurado* as sharing identical bilingual contexts *ensur* ⇔ *assegur* with varying bilingual suffixes *ing* ⇔ *ando* and *ed* ⇔ *ado*

- Learning substitution specifications or transformation rules from bilingual pairs of similar form, enabling one bilingual form to be inferred from another. For example, given the bilingual pair *declaring* ⇔ *declarando* and *ensured* ⇔ *assegurado*, we wish to identify bilingual substitution specifications *ing* ⇔ *ando* and *ed* ⇔ *ado*.

- Identifying clusters of bilingual pairs sharing similar substitution patterns, to allow translation generalisation.

- Infering new translation forms and hence be able to suggest OOV lexicon entries by classifying the new bilingual pairs into one of the pre-identified clusters and by applying the substitution patterns.

- Evaluating the generated translations for mis-generations and over-generations.

The similarities in translation forms illustrated in the above mentioned examples, should guide the learning model to automatically induce joint segmentation and alignment of morph-like units from a bilingual lexicon of parallel phrases/words. As the translation variants constitute the information source for the learning model, unlike the word-centric (monolingual) knowledge bases, these similarities in form should facilitate the model by constraining the space of joint segmentations.

The following facts pertaining to the existing lexicons further justifies the prospects for bilingual learning from translation pairs:

- A translation lexicon by itself can be considered as a parallel corpus.
- Currently existing lexicons are *validated* ones, with each bilingual pair *manually* tagged as *accepted* (positive examples), *rejected* (negative examples) or *left unverified*, hence making it reliable for the projected learning. Having a hugely high degree of certainty associated with each bilingual pair asserting its correctness, we may use it to learn or generalize translation patterns, infer new patterns and hence generate those unknown or unseen translation pairs that were not explicitly present in the training corpus used for the lexicon acquisition.
- The multilingual translation lexicon that we hold is sufficiently large and has an added advantage of including enough near word and phrase forms, translation patterns particularly for the language pairs, EN-PT, EN-FR, EN-ES and so forth, and hence it is suitable for morphological processing. Further details on the aspect of learning, anticipating its implications on the lexicon is elaborated on the respective chapters.
- In translating texts between EN-PT, the work by Aires [Air15] allowed improvements varying from 10 BLEU points higher in the PT-EN direction (for the DGT-TM) to 23 (for the DGT-TM) and 28.5 BLEU points higher (for the EUconst collection) for the EN-PT direction, as compared to the results obtained by Moses [Koe+07] when trained with the same corpus and in translating the same texts. This was achieved due to the use of a translation lexicon whose entries were manually validated, in addition to what the translating engine could learn from aligned parallel corpora [GL09].

Given the above, we aim towards an extensive lexicon that provides improved coverage over the existing one with respect to the vocabulary of both languages. This motivates us to come up with a tool that analyses the translation examples explicitly seen in the lexicon by investigating into the various aspects of relatedness between seen pairs and thereby learn to predict or generate new translations.

An insight into each of these is what the following chapters in Part II of this thesis elaborates on and what would be discussed in the light of the important observations discussed in the current section.

1.2 Research Objectives

The overall objective of the research presented in this document is to explore two major aspects pertaining to the automatically extracted bilingual translations lexicons - first, the quality of the lexicon entries, from the perspective of their use in subsequent iterations of parallel corpora alignment, new translation extraction and validation; second, the adaptability of the existing lexicon entries in suggesting OOV bilingual lexicon entries from the perspective of augmenting the automatically acquired lexicon in order to deal with OOV entries.

In the context of an iterative cycle involving alignment, extraction and validation, concerning the first aspect of correctness, I investigate the following questions: Which of the automatically extracted translations are relevant/re-usable in subsequent iterations? What are the common extraction/alignment errors? Does incorporating the knowledge of language/linguists help in selecting or segregating relevant candidates from inadequate extractions automatically? How can this linguistic knowledge, represented as human annotations, be used to learn automatic categorisation of the extracted candidates? Does the automated learning scheme benefit us in accelerating the human validation process? What are the features of the good translation equivalents that distinguish them from incorrect or inadequate translation candidates? Whether common features suffice for representation of unigram and multi-word translations? In what way does the enormity of linguistically annotated data influence the performance of the tool conceived for classification? In addressing each of these issues, my specific concern is to:

Objective 1: Apply machine learning technique such as classification so as to speed up and ease the manual process of validating automatically extracted translation candidates as correct or incorrect.

Concerning the problem of translating OOV terms, I consider the following questions prominent: How extensive is the existing lexicon? Can we improvise upon it by learning from what already exists? Can we employ the existing annotated lexicon entries in suggesting OOV translations automatically? How can we use the existing parallel resource of translations to learn and generalise translation patterns, infer new patterns and hence generate those OOV translations that were not explicitly present in the training corpus used for the lexicon acquisition? What are the prospects of simultaneous analysis of words and their translations? Does this joint analysis aid translation learning? How does the coverage of the existing lexicon influence the learning process? Can we come up with a tool that analyses the translation examples explicitly seen in the lexicon by investigating into the various aspects of relatedness between seen pairs and thereby learn to predict or generate new translations? What are the applications of such a bilingual learning scheme? In answering the above questions, the associated tasks involved are summarised as two objectives given below:

Objective 2: Automatically induce segmentation and bilingual morph-like units from a bilingual corpus of translations.

Objective 3: Apply the induced morph-like units to suggest OOV bilingual lexicon entries by productively combining the induced bilingual stems and suffixes.

While the first objective forms the subject of the next four chapters aggregated as Part I, the remaining two are discussed as separate chapters in the second half of the thesis, Part II.

1.3 Research Contributions

The major problems addressed in my thesis include :

- automatic validation of automatically extracted word and term translations, via classification [Kav+11; Kav+14b; Kav+15b], as correct or incorrect, with the aim of making it easier and quicker for human validation of those extracted translation pairs, enabling their usage in subsequent iterations of alignment and extraction, and providing an independent source of knowledge to classify the translations of text segments chosen by a machine translation engine as requiring greater attention from human post-editors (Objective 1).
- automatic learning of bilingual morph like units by identifying and pairing word stems and suffixes, using bilingual lexicons automatically extracted from parallel corpora (collections of texts and their translations), manually validated or automatically classified as above [Kav+14a; Kav+15a; Kav+15c] (Objective 2).
- improving bilingual lexicon coverage by suggesting unseen bilingual translation pairs (the so called Out-Of-Vocabulary, OOV, entries) [Kav+14a; Kav+15a; Kav+15c] (Objective 3).

The approaches adopted are characterised as follows :

- Manipulation of acquired and validated knowledge of the translations (bilingual pairs) by bilingual learning technique.
- Continual accommodation of the newly acquired knowledge in enhancing the learning process.

Main contributions that evolved during the study are published as articles in conference proceedings [Kav+11; Kav+14a; Kav+14b; Kav+15a; Kav+15b; Kav+15c] and are summarised in subsections below.

1.4 Thesis Outline and Organization

The first part of this thesis elaborates on augmenting the translation lexicon by automatic annotation of entries to indicate the quality (correctness or incorrectness) of automatically extracted translations as relevant for parallel corpora alignment, in the context of the iterative cycle involving alignment, extraction and validation. The Section 1.4.1 provides the summary of the related chapters.

One fundamental operation in Machine Translation (MT) is to find for a term[3] in first language its equivalent term in the second language. This is feasible when a term in first language appears in the translation lexicon of the system with its corresponding translation. But a major problem with MT systems is that they cannot translate terms that are not registered in the system's lexicon or translation table. In literature, these missing segments are referred to as *out-of-vocabulary* (OOV) terms. In the second half of my thesis, I present approaches for eventually improving the lexicon coverage by utilising the existing *correct* bilingual lexicon entries in suggesting translations for OOV words. Related chapters are summarized in Section 1.4.2.

1.4.1 Selection of Translation Equivalents for Parallel corpora Alignment

By incorporating human feedback in parallel corpora alignment and term translation extraction tasks, and by using all human validated term translation pairs that have been marked as *correct*, the alignment precision, term translation extraction quality and a bunch of closely correlated tasks improve. Moreover, such a labelled lexicon with entries tagged for correctness enables bilingual learning. Thus, beginning with an introduction to the nature of automatically extracted bilingual lexicon entries in Chapter 2, and the motivation for their classification, thereon in the Chapter 3 of this thesis, I elaborate the approach employed in classifying automatically extracted bilingual entries as '*correct*' or '*incorrect*', using a Support Vector Machine based classifier, SVMTEC trained on English-Portuguese (EN-PT) [Kav+11]. An evolution of this classifier augmented with additional features and experimental evaluations for three language pairs, English-Portuguese (EN-PT), English-French (EN-FR) and French-Portuguese (FR-PT) is discussed in Chapter 4 [Kav+15b].

Classification of automatically extracted *word-to-word* translations is the focus of the research presented in Chapter 5. Specifically, I examine and evaluate the use of bilingual stem, suffix correspondences and bilingual suffix classes in selecting correct word-to-word translations from among the automatically extracted ones. Experimental results for EN-PT and EN-FR show that by incorporating the morphological agreement information as features in training the classifier, the word-to-word classification accuracy improved, thereby leading to an overall improvement in the classifier accuracy when all translations (single- and multi-word translations) were considered [Kav+14b].

[3] term is being used either as a word, a phrase or a pattern in first language. Each language pair is ordered lexicographically (EN-FR, EN-HI, EN-PT, FR-PT and HI-PT), and both of the languages can be either a source or a target language.

1.4.2 Learning Bilingual Segments and Suffix Classes for Generating OOV Lexicon entries

In Chapter 6, the first of the chapters in part II of this thesis, I introduce and elaborate the concept of bilingual approach to learning morph-like units, instantiating how the known translation forms in a validated bilingual lexicon could be used to suggest similar translation forms, thereby addressing the issue of OOV terms in translation lexicons.

The bilingual learning approach for generating OOV lexicon entries itself forms the subject of the two chapters (Chapter 7 and 8) that follow. Based on the longest sequence common to a pair of orthographically similar translations, as for instance, *ensur* ⇔ *assegur* in the bilingual pair *ensuring* ⇔ *assegurando* and *ensured* ⇔ *assegurou*, the bilingual suffix transformations (replacement rules) of the form *ing* ⇔ *ando*, *ed* ⇔ *ou* are initially induced. Filtering of redundant analyses by examining the distribution of stem pairs and the associated transformations, grouping of bilingual suffixes conflating various translation forms and clustering of stem pairs sharing similar transformations serve as basis for the generative approach [Kav+14a]. Each of these phases are elaborated in Chapter 7, with experimental evaluations discussed for the language pair EN-PT.

Clustering the stem pairs sharing similar morphological extensions (transformations) is crucial for achieving safer translation generalisations. Two approaches for identifying bilingual suffix clusters are discussed in Section 7.2, one based on partition approach [Kav+14a] and the other based on bilingual suffix co-occurrence score [Kav+15c]. In the partition approach, the bilingual stems characterised by suffix pairs (features) are clustered using the clustering tool, CLUTO[4] which requires that the number of partitions are explicitly specified before clustering [Kav+14a]. Frequent associations between word suffixes have been observed to be crucial in inducing correct morphological paradigms [Des+14]. Extending this observation with respect to bilingual morphological extensions, as an alternative to the partition-based approach, the co-occurrence score between bilingual morphological extensions (suffixes) is used to determine if they should belong to the same cluster. Results are discussed for language pairs English-Portuguese (EN-PT) and English-Hindi (EN-HI) [Kav+15c].

Although, the aforementioned bilingual learning approach works well for training data having sufficient near translation forms, it proves insufficient under limited training data conditions. Thereby, in the Chapter 8, as an alternative to the approach mentioned in the Chapter 7, I present a minimally supervised method for situations where the lexicon is small without variety of translations covering different genders for adjectives, different numbers for nouns and different information for verbal inflection. In the context of bilingual learning, the proposed approach is discussed with experiments for a relatively small EN-HI translation data set and the main focus is on learning 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

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

classifier [Kav+15a].

To summarise, the approach presented in Chapter 7 was devised for a situation where the bilingual lexicon had enough variety of different translations, whereas, that discussed in Chapter 8 relies on minimal supervision and exploits the human usable dictionaries which lacks near translation forms, but are nevertheless used to deal with the limitations of the relatively small lexicons.

Finally, as a direct application of the use of bilingual suffix classes and bilingual segments, it is shown that the bilingual lexicon coverage can be improved by suggesting/generating OOV bilingual translation pairs that are similar to but different from the translations existing in the automatically acquired lexicon. Two aspects with respect to generation are dealt in - first, completing the bilingual lexicon for missing forms by simple concatenation of bilingual stems and suffixes that belong to the same bilingual suffix class, and second, given a new translation, suggesting new forms by first identifying the optimal split position, followed by classification of the split bilingual pair into one of the learnt bilingual suffix classes, and subsequent generation of all possible forms. The applicability of the bilingual segments and suffix classes in suggesting new translation is verified from the generation statistics for OOV unigram translations, which show that 90% of the generated translations were correct when both the bilingual segments (bilingual stem and bilingual suffix) in the bilingual pair being analysed are known to have occurred in the training data set. The generation phase is discussed in both the chapters 7 and 8 and is spread across various sections of the mentioned chapters.

1.4.3 Structure of this document

To summarise, the upcoming chapters are organised into two parts and are as enumerated below:

Chapter 2 through to Chapter 5 aggregates into part I, elaborating the translation selection problem.

Chapter 2 - Selection of Translation Equivalents for Parallel Corpora Alignment : Chapter 2 provides the background and motivation for the classification of bilingual pairs that were automatically extracted from aligned parallel corpora as '*correct*' or '*incorrect*'. The chapter also includes a general introduction to the nature of bilingual translation lexicon, the primary knowledge base for my study.

Chapter 3 - Selection of Word-word and Multi-word Translation Equivalents as Classification Problem: First Approach : In this chapter, I introduce classification as a means for validating automatically extracted EN-PT translations prior to human validation using a support vector machine based binary classifier, SVMTEC.

Chapter 4 - Selection of Word-word and Multi-word Translation Equivalents as Classification Problem: Second approach : The classifier discussed in this chapter is an enhanced version of that discussed in Chapter 3, with additional features and experimental evaluations extended to three different language pairs.

Chapter 5 - Selection of Word-to-Word Translation Equivalents as a Classification Problem: Third approach : In this chapter, classification is discussed from the perspective of word-to-word translations.

The remainder of the chapters, Chapter 6 through to Chapter 8, are dedicated to bilingual morphology learning and tackling of OOV bilingual terms and are consolidated in part II of this thesis.

Chapter 6 - Introduction to Bilingual Morphology Learning and Generation of OOV Lexicon entries : In this Chapter, the bilingual approach for learning morph-like units is introduced as a pre-phase to the generation of OOV lexicon entries.

Chapter 7 - Bilingual Morphology Learning using bilingual lexicons with diverse word inflections : The idea of bilingual learning from the translation lexicons having sufficient near translated forms is the subject of the Chapter 7.

Chapter 8 - Bilingual Morphology Learning using highly defective bilingual lexicons with limited inflection diversity : In this chapter, the learning approach is discussed considering highly defective lexicon with insufficient translation forms and with relatively smaller number of example translations.

Chapter 9 - Conclusions and Future Work : Towards the end of the thesis, in this Chapter, I present my concluding remarks and elaborate a bit on the scope for future work in the two main areas that form the subject of this thesis.

Part I

Selection of Translation Equivalents for Parallel Corpora Alignment

Chapter 2

Classification of Translations Automatically Extracted from Parallel Corpora: The Introduction

This chapter presents the background for classification and selection of bilingual translations that were automatically extracted from aligned parallel corpora. To begin with, in section 2.1, I introduce the characteristics of automatically extracted translations and the motivation for classification. In the following section (Section 2.2), the state-of-the-art work is presented.

2.1 Introduction and Background

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. 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.

A common approach for acquiring such a lexicon is based on aligning texts that are translations of each other (parallel texts) [Air+09], [GL09], [Koe+07]. The mainstream strategy for aligning parallel texts [Koe+07] is to apply a fully unsupervised machine learning[1] algorithm to learn the parameters (including alignment) of statistical translation models [Bro+93], [ON04]. 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 [GL09] and then extract new[2] term-pairs from those aligned texts [Air+09]. 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 translation lexicon with accepted and rejected entries, that was used in this study 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 [GL09] and enriches the lexicon, without the risk of decreasing its quality (because of manual validation).

This supervised strategy presents two-fold advantage of allowing an improved alignment precision while reducing uncertainty[3], 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 regard, in the Chapters 3, 4 and 5, I discuss automatic classifiers that segregate the extracted term-pairs as *correct* or *incorrect*, based on Support Vector Machine (SVM) trained upon a set of manually classified entries. The classification phase prior to validation improves validation productivity.

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

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

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

The decision on whether or not to incorporate the extracted pair of translation candidates into the bilingual lexicon as appropriate entries 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 [Air+09] [Koe+07] based on a machine learning approach is proposed. The experiments presented in forthcoming chapters 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 5,000[4] validated entries per day per validator, thereby contributing to significantly decrease the time consumed on manual validation. It should be stressed that I am not advocating that just this evaluated bilingual translation lexicon is used for translation. It would certainly decrease translation quality. Instead my focus is on the improvement of alignment precision and the subsequent extraction accuracy on each cycle of iteration.

2.2 Related Work

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 [Lop08] [ON04]. Such *completely automated training models* involve no human supervision and the selection of appropriate translation pairs is only done during the translation process by the decoder. Moreover, 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 [Joh+07].

The process of selecting translations, as discussed in literature, might be aimed from the perspective of improving the alignment precision and extraction quality or from the translation perspective itself [Kav+11; Tom+11]. Nevertheless, different researchers demonstrate varied views regarding the influence of alignment on translation quality, predominantly from the perspective of entries in a phrase table. It is observed that better alignment presents three-fold benefit that includes the advantage of producing a phrase table of manageable size with fewer phrase pairs, a reduced decoding time in searching the phrase table for the most probable translation, and a better quality of word or phrase level translation [Tia+14]. However, it is also observed that the decreased alignment error rate does not necessarily imply a significant increase in the translation quality [FM07; ON03; Vil+06]. I reiterate that I aim at an improved alignment precision and extraction accuracy in the context of the iterative cycle involving parallel corpora alignment, bilingual translation extraction and validation of automatically acquired bilingual pairs.

[4] validation of 3,000 entries per day is medically acceptable as the maximum limit for a human eye, in checking activities be it the shoes, bottles or any other articles.

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. The proposed approach falls down in the latter category, wherein, each pair of automatically extracted translation equivalent is classified into one of the pre-defined *accepted* or *rejected* categories.

2.2.1 Filtering Approaches

Melamed et al. [Mel95] introduced the filter-based approach for enhancing statistical translation models by inducing N-best translation lexicons with non-statistical sources of information. 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 be respectively dependent 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 [Mel95].

For discarding the most unlikely translation candidates extracted from parallel corpora, Tiedemann et al. [Tie98] suggested the use of automatic evaluation filters. 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 my experiments, the similarity and frequency based features have been used as baseline.

Aires et al. [Air+09] 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 for enhancing the quality of phrase tables is concerned, Chen et al. [Che+09] highlight the significance of association scores between phrase-pairs in parallel corpora and utilise 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 [Zha+04], word-based co-occurrence scores for re-ranking n-best list of translations [Che+05], significance testing of phrase pair co-occurrence with chosen threshold for removal of unlikely translation pairs [Joh+07] and the statistical independence measure namely Noise, for filtering phrase tables in Statistical Machine Translation System [Tom+09].

2.2.2 Approaches based on Support Vector Machines (SVM)

SVM, introduced by Vapnik is a learning machine based on the Structural Risk Minimisation principle and mapping of input vectors into high-dimensional feature space [Vap00]. Adequate feature identification that appropriately represent the knowledge implicit in data is fundamental to enable good learning.

SVMs had been successfully used for translation related tasks such as learning translation model for extracting word sequence correspondences (phrase translations) [SS03], automatic annotation of cognate pairs [BK07], extraction of bilingual terminologies [Ake+13] including others. The use of SVM based classifiers in selecting translation candidates, although rare, have been previously discussed in literature [Kut+05; Tom+11].

Kutsumi et al. [Kut+05], employed SVMs for selecting appropriate entries into a dictionary from aligned expressions and the work mainly targeted on complex proper noun phrases of the English-Japanese pair defined as proper noun phrases with prepositional phrases and/or co-ordinated phrases [Kut+05]. 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 translation. Morphemes, parts of speech, semantic markers obtained by consulting EDR concept dictionary, and upper-level semantic markers are used as means for representing the 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 classification experiments reported in this document, I introduce the concept of *translation mis-coverage* of bilingual translation entries that may compare with the common and difference feature proposed by [Kut+05]. Translation mis-coverage is learnt considering both the source and target sides of the bilingual pair [Cos+11]. I rely on naturally occurring positive and negative instances provided by the human specialists, based on appropriate consultations with the concordancer to locate all the possible translations where they occur before tagging an entry as *'Accepted'* or *'Rejected'*. While Kutsumi et al. target only complex proper noun phrase pairs in their experiments, I consider classifying all the automatically extracted bilingual pairs.

One-Class SVMs and the Mapping Convergence (MC) algorithm have been used to differentiate the usable and useless phrase pairs based on the confidence scores assigned by the classifier [Tom+11]. While the focus is on translation quality and avoiding alignment errors, the classifier is trained with a corpus that comprises of only useful instances.

All phrase pairs involved in best phrasal derivations[5] by the Oracle decoder are labeled as positive phrase pairs. Unlabelled examples of phrase pairs, however, are employed in addition to the positive examples in a semi-supervised framework[6] to improve the performance. I, on the other hand, view the task of selecting translation candidates as a supervised classification problem by utilising labeled training examples for both classes (positive and negative instances).

2.3 Summary

In the first half of the Chapter 1 (Section 1.1), I had introduced the significance of incorporating human feedback and the need for validating the automatically extracted translation equivalents in the context of iterative cycle involving alignment, extraction and validation. As a continuation, in this chapter, I have discussed the nature of the automatically acquired bilingual translation lexicon that forms the basis of my study, their implications on the subsequent cycles of parallel corpora alignment and hence extraction. As representatives of appropriate and inadequate candidates in the lexicon, I have exemplified a few positive and negative bilingual translations, thus characterising correct and incorrect translations. Thereby, I have provided the motivation for the translation classification and selection task.

Further, I have elaborated on the existing state-of-the-art literature prevalent in the area of translation selection. Certain approaches work to filter translation tables by simply discarding the wrong entries. Others employ SVMs for the task of translation selection. Instead of discarding the spurious translation candidates from the lexicon, we classify them into one of accepted or rejected classes. This provides scope for learning from those naturally occurring errors. While accepted translations could be used for bilingual learning and inducing OOV entries, as will be discussed in Part II of this thesis, the accepted and rejected candidates collectively serve in distinguishing between correct and incorrect translations, and also possibly, in helping the human post editors to tackle the mistranslated translation segments. Studies show that negative evidences are equally important as positive evidences in error detection and selection tasks. As a matter of fact, research on artificial generation of errors [FA09], [Dic10], [FY14] have been surfacing more recently. For instance, Foster et al. [FA09], explore the usefulness of error generation tool, GenERRate in creating synthetic training data to be used in grammatical error detection research.

In the chapters that follow, I will discuss the applicability of SVM based classifiers in classifying the word-to-word and multi-word translations with comparative results in relation to the selected techniques discussed in this chapter.

[5] one that maximises a combination of model score and translation quality metric

[6] MC algorithm

Chapter 3

Selection of Word-word and Multi-word Translation Equivalents as Classification Problem:First Approach

In the current chapter, I discuss the classification and selection of word-to-word and multi-word bilingual translations that were automatically extracted from aligned parallel corpora using the SVM based classifier, SVMTEC. The features characterizing the bilingual pairs, the corpora used in extracting the candidate translations, and the experimental results are elaborated in the Sections 3.2 and 3.3.

3.1 Validation as a Classification Problem

The task of validating the automatically extracted translation candidates is treated as a classification problem. In this regard, I discuss the use of SVM based classifier, SVMTEC used in segregating the extracted translation candidates as *accepted*, 'A' or *rejected*, 'R'. For classification, the training and test data representing bilingual data instances are formed from all of the automatically extracted and manually validated translations. The classifier for automatically extracted translations is trained on manually validated EN-PT translation pairs using a specific set of features of those translation pairs that characterise term frequency, co-occurrence frequency, orthographic similarity, translation coverage and translation endings [Kav+11]. Each bilingual pair is a data instance represented as a *feature vector*[1] and a target value known as the *class label*[2]. The learning function is trained with the scaled training data set, where each sample is represented as a feature vector with the label +1 ('A') or -1 ('R'). The estimated model is then used to predict the

[1] Vector of real numbers

[2] Positive and negative examples respectively labeled as +1 and -1. Data to be classified is labeled 0.

class for each of the unknown data instance kept aside for testing, represented in the same way as any sample in the training set, but with the class label 0.

I use the Radial Basis Function (RBF) kernel:

K(u, v) = $\exp(-\gamma * |u - v|^2)$;

parameterised by (C, γ), where u and v represent training and testing example respectively. $C > 0$ is the penalty parameter of the error term and $\gamma > 0$ is the kernel parameter.

RBF Kernel was chosen after experimenting with linear kernel beforehand. Although the results were not substantially different for EN-PT with the mentioned kernels, they were better for EN-FR and FR-PT with RBF, when an evolution of this classifier called SVMTEC-2 (discussed in Chapter 4) was used.

3.2 Features-SVMTEC

SVMTEC, the Support Vector Machine based classifier for automatically extracted translation equivalents takes into account various features of those translation pairs namely the individual term frequencies, co-occurrence frequency, writing similarity, translation mis-coverage, stemmed mis-coverage and translation endings. Below, I elaborate each of these features used in quantifying the bilingual candidates under consideration. Experiments are discussed for the language pair EN-PT.

3.2.1 Frequency and Orthographic Similarity based 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 [GL09]. Features derived using these frequencies, such as, the Dice coefficient of frequencies, the ratio of co-occurrence 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. Features reflecting orthographic similarity between the terms, computed based on Levenshtein edit distance [Lev66], longest common subsequence (LCS) [Mel95], longest common prefix (LCP) [Kon05] and length ratio are quantified as measurable characteristics and used as feature values to identify cognates. Each of these features are normalised by the length of the longest term in the bilingual pair under consideration and are respectively calculated using the equations 3.1 through 3.4 listed below, where *Len* denotes the length, *EditDist(X,Y)* is the edit distance between the term X in first language and the term Y in second language. In each of the equations 3.1 through 3.4, $|X|$ and $|Y|$ represents the length of X and Y measured in terms of the number of characters.

$$EditSim(X,Y) = 1.0 - EditDist(X,Y)/Max(|X|,|Y|) \tag{3.1}$$

$$LCSR(X,Y) = Len(LCS(X,Y))/Max(|X|,|Y|) \tag{3.2}$$

$$LCPR(X,Y) = Len(LCP(X,Y))/Max(|X|,|Y|) \tag{3.3}$$

$$LENR(X,Y) = Min(|X|,|Y|)/Max(|X|,|Y|) \tag{3.4}$$

EditSim in equation 3.1 represents the widely used edit-distance based similarity measure that returns the similarity between the term $|X|$ in first language and its translation $|Y|$ in second language. Equations 3.2, 3.3 and 3.4 respectively denotes normalised LCS, normalised LCP and the length ratio.

3.2.2 Determiners (DT) and Co-ordinated Conjunctions (CC)

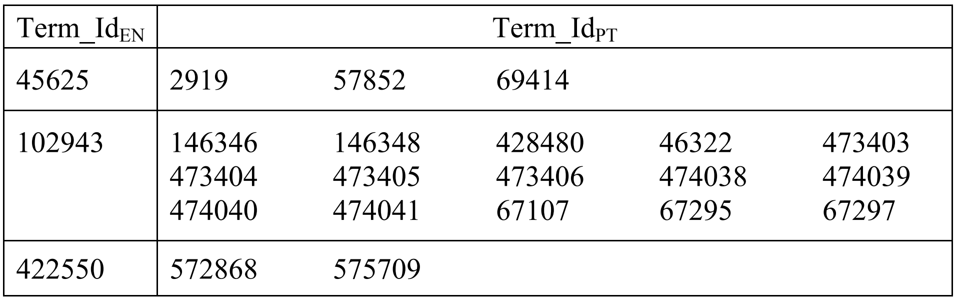
Terms ending with determiners such as, *'a'*, *'the'*, *'certain'* etc., in EN and *'os'*, *'uma'* in PT may not be considered as adequate candidates in the lexicon as was discussed in Section 2.1. In order to reflect this, binary-valued features discriminating translation pairs ending with the determiners are used. Further, 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 was intended to be captured. 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 ⇔ 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. The feature values are estimated using manually identified lists of determiners and co-ordinated conjunctions for each language under consideration.

3.2.3 Translation Mis-coverage (MC)

As was elaborated in section 2.1, the lexicon consisted of bilingual pairs wherein a term in one language differed from its counterpart (translation) 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-expressions 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 *incorrect* class. The features specify the *existence of translation gaps* or *missing translation segments* (or *mis-coverage* as it was named in [Cos+11]) 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*[3] 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 the feature value for corresponding term is set to 1. There is some mis-coverage.

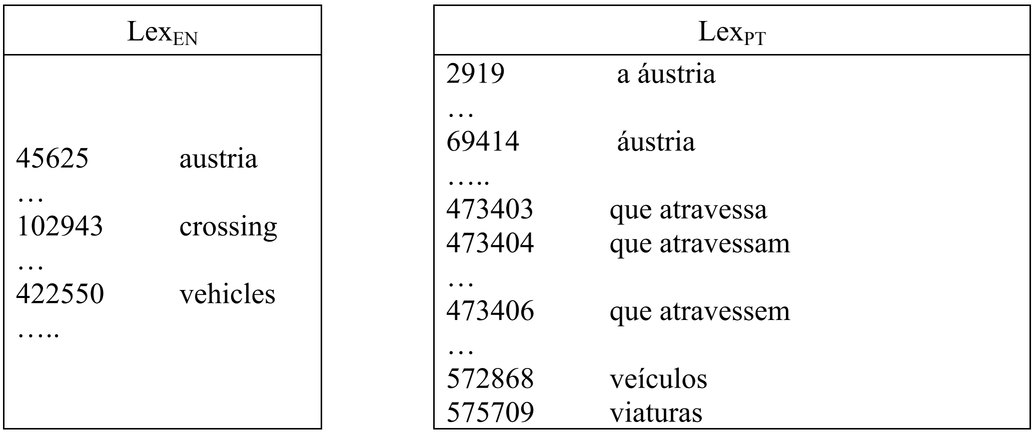
The procedure for identifying translation mis-coverage follows the Aho-corasick set-matching algorithm [Gus97] that checks if the terms in the key-word tree (constructed from the bilingual training data separately for EN and PT terms) occur as sub-expressions in the bilingual pair to be validated and if they occur are accepted translations.



Term_Id_{EN}	Term_Id_{PT}				
45625	2919	57852	69414		
102943	146346	146348	428480	46322	473403
	473404	473405	473406	474038	474039
	474040	474041	67107	67295	67297
422550	572868	575709			

Figure 3.1: Translation mappings for EN and PT

Figure 3.2: Snap shot of accepted terms in EN (L_{EN}) and PT (L_{PT})

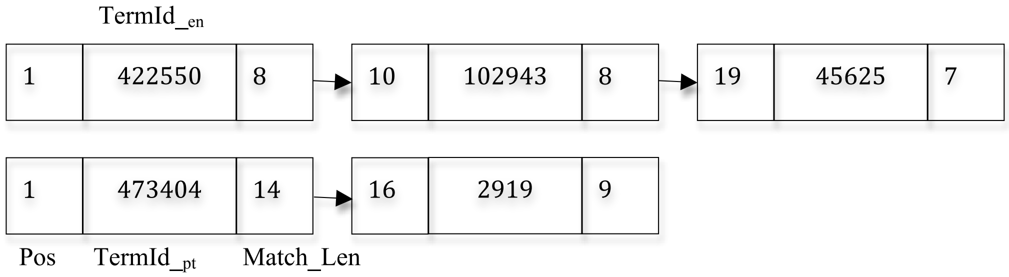


Lex_{EN}	
45625	austria
…	
102943	crossing
…	
422550	vehicles
…..	

Lex_{PT}	
2919	a áustria
…	
69414	áustria
…..	
473403	que atravessa
473404	que atravessam
…	
473406	que atravessem
…	
572868	veículos
575709	viaturas

[3] I look for words translating as multi-words, or multiwords translating as multiwords

Figure 3.3: Lists indicating occurrences of accepted terms as sub-expressions in the example bilingual pair *vehicles crossing austria* ⇔ *que atravessam a áustria* that is to be validated



The procedure that looks for translation coverage is summarised as below:

Let L_A be the lexicon of accepted bilingual translation equivalents with translation mappings as shown in Figure 3.1.
Let T_X=x_1, x_2.....x_m and T_Y=y_1, y_2.....y_n represent separate lists of terms *accepted* in language X and language Y respectively.
Let L_{UV} be the lexicon of translation candidates to be validated.

1. Construct separate keyword trees for terms in T_X and T_Y.

2. Apply the Aho-corasick set-matching algorithm [Gus97] to check if the terms in the key-word tree occur as sub-expressions in the bilingual pair from L_{UV}.

    - The result of this search are two lists (corresponding to the search performed on left-hand side term in language X and right-hand side term in language Y of the bilingual pair) representing occurrences of keyword tree terms as sub-expressions in the bilingual pair.

    - Each item in the list indicates the *occurrence position* of key-word tree terms in the term-pair, the *identifier* of matched key-word tree term and the *length* of the match (see figure 3.3; first and the second lists respectively represent occurrence of accepted EN and PT terms in the bilingual pair validated).

3. For each item representing x_i in language X in the initial list, retrieve all identifiers representing translations of x_i in language Y using the relation table. Construct an *extended list* representing those retrieved list of identifiers (see figure 3.4; the second field in each node represents translation identifier for term in language X, and the last field in each node are the retrieved list of identifiers representing possible translations for term x_i in language Y).

4. For each item in the *extended list* obtained for left-hand side term in language X, search if each of the retrieved identifiers is an item in the initial list obtained for right-hand side term in language Y. Construct the output list (sorted by *occurrence position*) by appending the matched items from *extended list* for left-hand side term in language X and initial list for right-hand side term in language Y (see figure 3.5).

5. Repeat steps 4 and 5 for each item in the list obtained for the right-hand side term in language Y as well.

Figure 3.4: Extended List corresponding to the term *vehicles crossing austria* in EN indicating occurrences with corresponding mappings from translation table

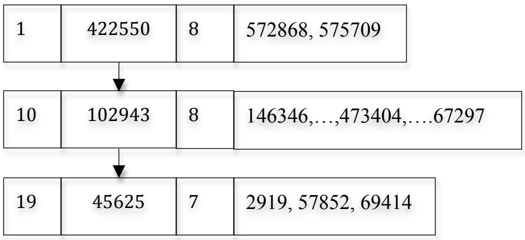
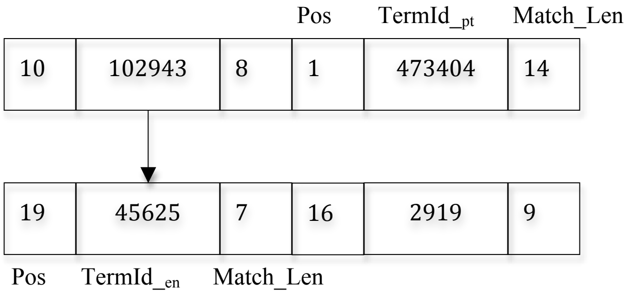


Figure 3.5: List indicating translations in PT for sub-expressions of left-hand side term *vehicles crossing austria* in EN



In Table 3.1, the feature values[4] representing coverage for EN-PT bilingual pairs are illustrated. The first two examples with gap_{EN}, gap_{PT} set to 0 illustrate correctly translated bilingual pairs having complete coverage. However, the bilingual candidate '*vehicles crossing austria ⇔ que atravessam a áustria*' is an example of incorrect bilingual entry. It could be seen that, no translation exists for the English sub-expression '*vehicles*' in Portuguese. Hence, I indicate this missing translation for '*vehicles*' by gap_{EN} set to 1. However, looking for coverage from the right-hand side term, gap_{PT} should be set to 0 as '*que atravessam ⇔ crossing*' and '*a áustria ⇔ austria*'. This asymmetry could be further 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, words of shorter length such as *de*, *a* (functional words) in PT are ignored. In other words, I neglect missing translations for sub-expression that are not content words, provided they are encapsulated between content words that

[4] 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

Table 3.1: Example of features indicating translation coverage for *EN-PT*

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)
bronchitically	*bronquiticamente*	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

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*', the translation candidate is considered to have a full coverage and hence gap_{EN} and gap_{PT} are both set to 0.

3.2.4 Reprocessing Translation Gaps

When both gap_{EN} and gap_{PT} are set to 1 indicating translation mis-coverage on either ends, it is important to examine if the sub-expressions indicating missing translations in English and Portuguese parts are *possible translations*. This is achieved in two ways: firstly, using the stemmed lexicon of accepted English and Portuguese terms and secondly, by considering the orthographic similarities of expressions showing mis-coverage on either ends.

3.2.4.1 Use of Stemmed Training Data

To illustrate the use of stemmed training data in determining parallelism for those sub-expressions indicating missing translations, consider the bilingual pair '*watertight compartment* ⇔ *compartimento estanque*'. For the example considered, 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 stems 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 fully covered [5] by the lexicon, the feature values for gap_{EN} and gap_{PT} are set to 0.5, which is a neutral value reflecting the lack of support for deciding whether to accept or to reject that pair. All such pairs are as well subjected to further processing to select from among them, those entries that might represent correct translations. For example, the pair *'accrued ⇔ vencidas'*, although does not appear in training data causing the features values to be initially set to 0.5 representing complete mis-coverage, nevertheless, the feature values are reset to 0 after reprocessing, as other similar forms *'accrued ⇔ vencida'* and *'accrued ⇔ vencidos'* do exist in the training data set. As explained in the previous paragraph, the use of stemmed training data enables this setting. Further, it will be learned in Chapter 7 that, *'accru ⇔ 'venc'*, and *'ed ⇔ 'ido'* | *'ida'* | *'idos'* | *'idas'* enabling us to infer that *'accrued'* might be a translation of *'vendido'*, *'vencida'*, *'vencidos'* and *'vencidas'*.

3.2.4.2 The Spelling Similarity Measure - SpSim [GL11]

Apart from the use of stemmed training data in determining the parallelism for sub-expressions representing mis-coverage, additionally, the orthographic similarities of such expressions are taken into account using a similarity measure based on the edit distance. The measure is based on estimating the similarity between words (SpSim) [GL11] computed as in equation 3.1, with the exception that the characteristic spelling differences that were learnt previously are discounted in computing the edit distance. Examples of such spelling differences include (ph | f) and (on | ão) found in English-Portuguese cognates, such as, (phase | fase) and (photon | fotão) [GL11]. SpSim for a pair of terms (X,Y) is calculated as shown in the Equation 3.5 below:

$$SpSim(X,Y) = 1.0 - \frac{D(X,Y)}{Max(|X|,|Y|)} \tag{3.5}$$

where the distance function *D(X,Y)* is the *EditDist* between words, discounting characteristic spelling differences that were learnt previously. $|X|$ and $|Y|$ represents the length of X and Y measured in terms of the number of characters.

Accepted word pairs (cognates) that have EdSim (Edit-Distance based similarity) greater than 0.5 as observed in the training data set was used for training SpSim. Using a lower threshold to select examples causes selection of non-cognates for training, which would make SpSim assign high scores for some non-cognates. A dictionary containing the substitution patterns is thus learnt. For instance, the substitution pattern

[5] This is specifically the case with unigram translations

extracted from EN-PT cognate word pair 'phase' and 'fase' is ('^ph', '^f'), after eliminating all matched (aligned) characters, 'a' ⇔ 'a', 's' ⇔ 's' and 'e' ⇔ 'e'. The caret (^), at the beginning of the aligned strings distinguishes that the patterns appears as a prefix.

3.3 Experimental Setup-SVMTEC

SVM based tool namely LIBSVM[6] [5] was used to learn the binary 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.

3.3.1 Data sets

The translation candidates that were extracted from aligned parallel corpora [7] [Ste+06] for language pairs EN-PT was used for intended experiments. Data consisted of a set of bilingual pairs representing terms in first language and its equivalent in second language, collected from an existing bilingual lexicon whose entries are manually tagged as being accepted (positive examples), rejected (negative examples) or left unverified. Translations constituting of (first/second language) terms ranging in length from one word to a maximum of 7 words were used. The alignment and extractions follow the procedures as those proposed in [GL09] and [Air+09] respectively.

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.

Table 3.2: Overview of the Training and Test data set

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

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 3.3) were constructed from the training data presented in Table 3.2. The classifier was trained 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 3.2.

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

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

3.3.2 Baselines

The classifier trained with frequency-based features such as Dice coefficient of frequencies, the ratio of co-occurrence frequency to source term frequency, ratio of the co-occurrence frequency to target term frequency, minimum to maximum frequency ratio and the features reflecting orthographic similarity between the terms, computed based on Levenshtein edit distance [Lev66], longest common subsequence (LCS) [Mel95], longest common prefix (LCP) [Kon05] and length ratio is used as baseline.

3.4 Results and Evaluation-SVMTEC

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:

$$Precision_{Rej} = t_n/(t_n + f_n) \tag{3.6}$$

$$Precision_{Acc} = t_p/(t_p + f_p) \tag{3.7}$$

$$Recall_{Acc} = t_p/(t_p + f_n) \tag{3.8}$$

$$Recall_{Rej} = t_n/(t_n + f_p) \tag{3.9}$$

$$Accuracy = (t_p + t_n)/(t_p + f_p + t_n + f_n) \tag{3.10}$$

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 attained in classifying the bilingual pairs for each of the classes over various data sets are as shown in Figure 3.6, where 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.

Also, in order 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 3.11 through 3.13 below.

$$\mu_P = (Precision_{Acc} + Precision_{Rej})/2 \tag{3.11}$$

$$\mu_R = (Recall_{Acc} + Recall_{Rej})/2 \tag{3.12}$$

$$\mu_F = 2 * \mu_P * \mu_R/(\mu_P + \mu_R) \tag{3.13}$$

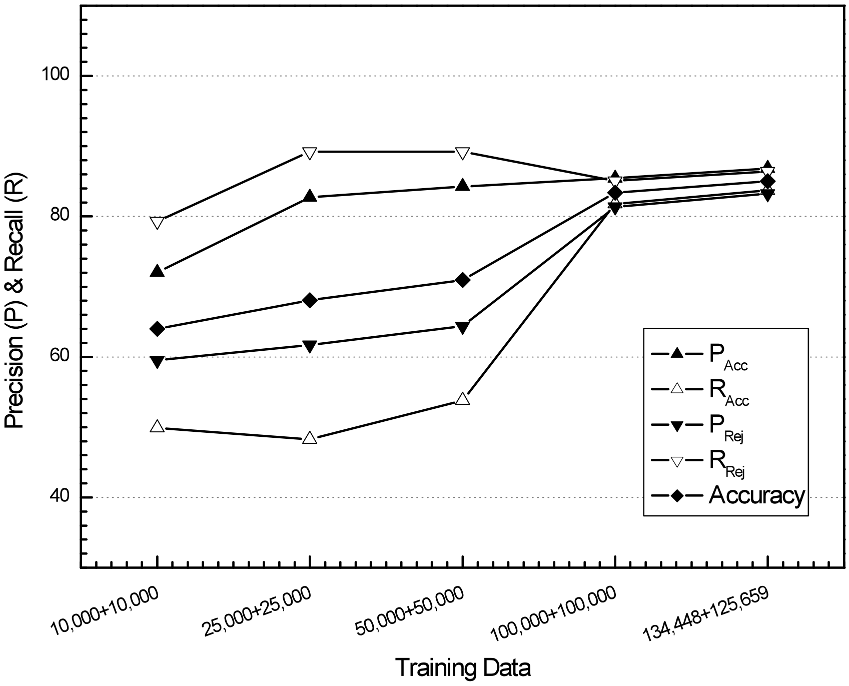


Figure 3.6: Precision and Recall for different classes

Table 3.3 shows the μ_P, μ_R and μ_F obtained in classifying the bilingual pairs together with the chosen kernel function and corresponding parameter values for different training sets. The classification approach discussed above, enabled a micro-average f-measure of 85.06% compared to that attained using the scoring functions proposed by Aires et al. [Air+09].

Table 3.3: Performance results for different training data sets

Training Data	**Kernel = RBF**	**Type = C-SVC**	μ_P	μ_R	μ_F	**Accuracy**
(Positive + Negative)	**Gamma (g)**	**Cost (C)**				
10,000 + 10,000	0.125	32	65.80	64.52	65.15	64.02
25,000 + 25,000	0.125	32	72.22	68.74	70.44	68.05
50,000 + 50,000	0.5	32	74.32	71.54	72.90	70.94
100,000 + 100,000	0.5	32	83.45	82.39	82.92	83.39
134,448 + 125,659	0.5	32	85.04	85.08	85.06	85.03

The accuracy of the estimated classifier in predicting classes for the EN-PT bilingual pairs using various features are presented in Table 3.4. The information indicating term pairs ending in determiners (DT) and co-ordinated conjunctions (CC) proved beneficial in discarding unproductive bilingual pairs that would otherwise contribute to a huge lexicon. A rough alignment based method looking for translation mis-coverage (MC) from either sides of the bilingual pair provided significant improvement in discriminating the classes. The underlying notion was to utilise the available knowledge about highly reliable translation pairs in deciding if newly extracted bilingual pairs are correct. Using the stemmed lexicon obtained from the training data with accepted bilingual pairs (stemmed positive examples) in further processing of the segments representing translation gaps, a remarkable overall improvement (almost 10%) was observed in the classification results.

Table 3.4: Performance of classifier on EN-PT bilingual pairs for different features over the entire training data set

Features	μ_P	μ_R	μ_F	**Accuracy**
Orthogonal Similarity + Frequency (Baseline-BL)	54.13	53.95	54.04	54.30
BL + DT + CC	67.10	66.73	66.91	66.96
BL + DT + CC + MC	75.47	74.93	75.19	75.16
BL + DT + CC + MC + Reprocessing Gaps	85.04	85.08	85.06	85.03

3.5 Summary

In this chapter, I have discussed the classification approach as a means for selecting appropriate and adequate candidates for parallel corpora alignment using the classifier model, SVMTEC. Experiments on EN-PT show that the best micro-average f-measure of 85.03% was attained when the classifier was trained with all of the features.

Several insights are useful in distinguishing the adequate candidates from inadequate ones such as, lack (presence) of parallelism, spurious terms at translation ends and the base properties (similarity and occurrence frequency) of the translation pairs.

The work reported in this chapter is motivated by the need for a system that evaluates the automatically extracted translation candidates, prior to their submission for human validation, where the extraction method proposed by Aires et al. [Aires+09] was employed. An evolution of this classifier forms the subject of the next chapter that follows.

Chapter 4

Selection of Word-word and Multi-word Translation Equivalents as Classification Problem: Second approach

In this chapter, I discuss the experiments on classifying automatically extracted word-word and multi-word translations carried out using the classifier SVMTEC-2, which might be considered as an evolution of the classifier SVMTEC discussed in the Chapter 3. In Section 4.2, features characterising the bilingual pairs are elaborated. The experimental settings with the datasets are discussed in Section 4.3.

4.1 SVMTEC-2

SVMTEC-2, the SVM-based classifier for automatically extracted translations, differs from the classifier SVMTEC mainly with respect to the number of features used. The performance of the classifier is tested on three language pairs EN-PT, EN-FR and PT-FR and experimental results are discussed for each of these language pairs. Further, the usability of the classifier trained on one language pair in classifying translations belonging to different language pairs are tested. Specifically, the classifier trained with EN-PT, is used in classifying the translations extracted for the language pairs EN-FR and FR-PT. Analogously, the applicability of classifiers trained with EN-FR and FR-PT are tested in classifying other mentioned language pairs and the results are presented.

4.2 Features-SVMTEC-2

An overview of the features used in the classification model, SVMTEC-2 is discussed in this section. I limited the various features reflecting orthographic similarity (LCPR, LCSR, LENR and EditSim) and the term frequencies (F_X, F_Y and F_{XY}) to two each, as these features (orthographic similarity - EditSim, SpSim and frequency based - Dice Coefficient, MinMaxRatio) effectively substituted the remaining features discussed in Chapter 3, that were dropped out. Instead of relying on manually identified patterns representing determiners and co-ordinated conjunctions, two approaches are used in automatically identifying the bad translation endings, one based on stop words and the other based on patterns observed as translation endings in rejected training data, but absent in accepted training data. Given the better performance of the classifier SVMTEC, when trained with the feature that examines for the translation mis-coverage, thus enabling the identification of incomplete translations, the translation mis-coverage feature is retained in SVMTEC-2 as well. Stemmed mis-coverage is determined as was discussed for the classifier SVMTEC and is used to overcome the lack of sufficient surface word-to-word translation forms.

Features derived using the orthographic similarity measures (strsim) and the frequency measures (freq) ($BL_{strsim+freq}$) discussed in the sections 4.2.1 and 4.2.2 are used in training the baseline classifier.

4.2.1 Orthographic Similarity

Two orthographic similarity measures based on edit distance are used to quantify the similarity between terms on either sides of a bilingual pair: the Levenshtein Edit Distance [Lev66] (Equation 3.1), discussed earlier in Section 3.2, and the Phrase_SpSim [Gom16] discussed in the Section 4.2.1.1 below.

4.2.1.1 The Phrase Similarity Measure - Phrase_SpSim [GL11]

Phrase_SpSim measure is an evolution of SpSim [GL11] (Equation 3.5 discussed in Subsection 3.2.4), the spelling similarity measure for computing the similarity score between a pair of words taking into account spelling differences that are characteristic of each language pair and learnt from example cognates previously known. While SpSim is suitable for computing the spelling similarity between pair of words, Phrase_SpSim measures the orthographic similarity by applying SpSim to a pair of phrases. Phrase_SpSim works by considering all possible word-word pairings in the given expressions and choosing the word-level alignment that has lower overall distance (which means higher similarity).

To instantiate, consider the bilingual pair, *'general indifference'* and *'indiferença geral'*. The algorithm considers two possible word alignments (say A1 and A2).

$A1 = (\text{'general'}, \text{'indiferença'}), (\text{'indifference'}, \text{'geral'})$

$A2 = (\text{'general'}, \text{'geral'}), (\text{'indifference'}, \text{'indiferença'})$

Thus, the total SpSim scores for those two alignments are as discussed below:

$Phrase\_SpSim(A1) = 1-(spchardist(\text{'general'}, \text{'indiferença'})+spchardist(\text{'indifference'}, \text{'geral'}))/L)$

$Phrase\_SpSim(A2) = 1-(spchardist(\text{'general'}, \text{'geral'})+spchardist(\text{'indifference'}, \text{'indiferença'}))/L)$

In the above equations, L represents the length (in characters) of the longest phrase (which in this case is *'general indifference'*) excluding the spaces between words. The distance function spchardist(X, Y) is related to the original measure SpSim(X, Y) but, differs in that it returns the absolute distance (in terms of characters) rather than the similarity. The relationship between the two (one can be computed from the other) is as shown below:

$spchardist(X,Y) = spdist(X,Y) * max(len(X), len(Y)),$

where, len(X) and len(Y) respectively represent the length of the word X and Y respectively, and spdist(X,Y) is the (relative) spelling distance of the two words. It varies between 0 and 1 and is given by:

$spdist(X,Y) = 1 - spsim(X,Y)$

From the two alternative alignments for the example considered, the algorithm would select A2 because $Phrase\_SpSim(A2) > Phrase\_SpSim(A1)$.

The character-level alignments and the substitution patterns found by SpSim when comparing *'general'* with *'geral'* and *'indifference'* with *'indiferença'* is as shown below. Note the extra begin and end markers (^ and $, respectively):

Table 4.1: The character-level alignments for *'general'* with *'geral'* and *'indifference'* with *'indiferença'* using SpSim

^	g	e	n	e	r	a	l	$	i	n	d	i	f	f	e	r	e	n	c	e	$
^	g	e			r	a	l	$	i	n	d	i	f		e	r	e	n	ç	a	$

The substitution patterns from these alignments are (*'ener'*, *'er'*) for (*'general'*, *'geral'*) and (*'ffe'*, *'fe'*) and (*nce$'*, *'nça$'*) for (*'indifference'*, *'indiferença'*). Unlike the first pattern (which is not very common), the following two patterns are relatively common, as they are found in bilingual pairs such as, (*'effect'* - *'efeito'*), (*'Florence'* - *'Florença'*) and so forth.

Table 4.2: The Orthographic Similarity Scores

Term$_{EN}$	**Term**$_{PT}$	**EdSim**	**Phrase_SpSim**
general indifference	indiferença geral	0.15	1.00
official	comercial	0.56	0.66
commitments	compromissos de crédito	0.29	0.24
limits of the	limites de a	0.54	0.82
impact on the	impacto em a indústria	0.39	0.47

4.2.2 Frequency of occurrence

To represent the translational equivalence, based on the frequencies of terms on both sides of the bilingual pair, two measures are used: the Dice association measure (Equation 4.1) and the MinMaxRatio (Equation 4.2).

The Dice association measure for a pair of terms *(X,Y)* takes into account the co-occurrence frequency of the terms on either sides of the bilingual pair in addition to the frequency of term in first and second language and is given by the equation,

$$Dice(X,Y) = \frac{2 * F_{XY}}{F_X + F_Y} \tag{4.1}$$

where F_X is the frequency of the term X in the first language text and F_Y is the frequency of the term Y in the second language text; F_{XY} is the joint frequency of the terms in aligned parallel texts.

Another measure that efficiently substitutes the individual frequencies F_X and F_Y is the minimum to maximum frequency ratio given by the equation,

$$MinMaxRatio(X,Y) = \frac{Min(F_X,F_Y)}{Max(F_X,F_Y)} \tag{4.2}$$

4.2.3 Bad Ends

The bilingual pair *'limits of the ⇔ limites de a'* instantiates a particular type of inadequate translation wherein, the term (on either sides) ends with a determiner following which a noun or a noun phrase is anticipated. It is the absence of the noun or a noun phrase after the determiner that makes the translation incomplete. By allowing this entry into the lexicon as a correct translation, we cannot refrain other entries ending with *'o'*, *'os'*, and so forth from accommodating the determiner's position. Such translations with inadequate endings are referred to as having bad ends (BE). To keep check over such entries, two binary valued features are used, signifying whether the terms in the bilingual pair ends with a determiner (1) or not (0). Each of the two features represents the goodness of the translation endings on each side of the bilingual pair.

Table 4.3: The Bad Ending Scores

Term$_{EN}$	**Term**$_{PT}$	**BE**$_{SW}$	**BE**$_{Pat_{R-A}}$
general indifference	indiferença geral	(0.00, 0.00)	(0.00, 0.00)
official	comercial	(0.00, 0.00)	(0.00, 0.00)
commitments	compromissos de crédito	(0.00, 0.00)	(0.00, 0.00)
limits of the	limites de a	(1.00, 1.00)	(1.00, 1.00)
impact on the	impacto em a indústria	(1.00, 0.00)	(1.00, 0.00)

In the Chapter 3, it was pointed that the manually identified lists of determiners and co-ordinated conjunctions were used to determine if bilingual pairs had inadequate endings. On the other hand, in training SVMTEC2, two different approaches are employed to identify bad ends: One set of two features are extracted based on endings that are stop words (BE_{SW}) and the other set of two features based on endings seen in the rejected training data, but absent in the accepted training data ($BE_{Pat_{R-A}}$). I considered only those endings that occurred more than 5 times in the rejected training dataset, but not in accepted training set. Also, to make sure that content words are not considered as bad endings, I restricted the term length to less than 5 characters. Table 4.3 shows the scores for few of the randomly selected bilingual pairs.

4.2.4 Translation Mis-coverage (MC)

A typical error observed in extracted candidates is the lack of parallelism with respect to content words, and was introduced in Chapter 3. For instance, for the bilingual pair *'commitments ⇔ compromissos de crédito'* to be considered as correct, *'crédito'* needs to be translated as either *'lending'* or *'loan'* in EN. So the correct term translation would be *'lending commitments ⇔ compromissos de crédito'* or *'loan commitments ⇔ compromissos de crédito'*. Likewise, the bilingual pair *'union level ⇔ união'* is an incorrect translation because no translation exists on the right hand side for the English word *'level'*. Also, it should be noted that sometimes a word in English may be translated by two or three words in Portuguese, as for instance with the bilingual pair *'superhighway ⇔ super auto estrada'*.

In general, the lexicon consisted of bilingual pairs wherein a term in one language differed from their counterparts (translations) by the *number of content words*. 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.

To assess the bilingual candidates for parallelism, I introduce two features. A translation candidate is considered to have a translation mis-coverage with respect to first language (gap_{L1}=1) when the term in the first language does not have a translation in second language in whole or in parts and vice versa. Lack of parallelism implies a mis-coverage in translation. The features specify *mis-coverage* or *missing translation segments* in each of the first and second language terms, where mis-coverage characterises a sub-expression of the term in one language for which there is no known translation equivalent in the term of the other language.

4.2.5 Stemmed Mis-coverage ($\mathbf{MC}_{Stm}$)

In the Chapter 3, it was discussed that while looking for coverage, if the expressions on either sides are not covered by the lexicon, the features gap_{L1} and gap_{L2} would be set to 0.5, a neutral value reflecting the lack of support in deciding whether to accept or to reject that pair, based on known word translations. Further, only those segments exhibiting mis-coverage were reprocessed using SpSim and stemmed training data.

The difference here is that, to deal with such situations reflecting the lack of support, additionally, I extract two features reflecting mis-coverage by repeating the experiment for the entire dataset (all bilingual pairs to be validated) using the stemmed training data. These features work in the same way as discussed above for MC, except for the fact that, mis-coverage here is determined by considering stemmed training and test datasets rather than the original datasets. To instantiate, while looking for coverage, for the bilingual pair *'bronchitically ⇔ bronquiticamente'*, its stemmed version *'bronchit ⇔ bronquit'* is used, as the coverage is examined using the stemmed training and test sets. If the training data contains the term *'bronchit'* in EN and *'bronquit'* in PT, then (gap_{L1}, gap_{L2}) would be (0.0, 0.0). This feature would find less gaps in translations that are indeed parallel, and thus decrease the number of false negatives (i.e., good translations that are classified as bad).

For identifying the translation mis-coverage, I use the Aho-corasick set-matching algorithm and check if the terms in the key-word tree[1] occur as sub-expressions in the bilingual pair to be validated and if they occur, are accepted translations [Gus97]. Similarly, to find the stemmed coverage, I use the stemmed training and test sets, obtained using the Snowball stemmer. Here, each keyword tree is constructed using the stemmed part of the term. Translation gaps are identified using the Aho-corasick set-matching algorithm as elaborated in Section 3.2.3 of the Chapter 3.

4.3 Experimental Setup-SVMTEC-2

I used LIBSVM[2], an SVM based tool to learn the binary classifier. Data is scaled in the range [0 1]. I perform a grid-search on RBF kernel parameters, (C, γ) using cross-validation, so that the classifier can accurately predict unknown data (testing data).

4.3.1 Data sets

The translation candidates used in my experiments were acquired using various extraction techniques applied on a (sub-)sentence aligned parallel corpora[3] [Air+09; Bro+93; GL11; LL09]. I experimented with 3 language pairs, EN-PT, EN-FR and FR-PT. The suffix array based phrase translation extraction technique [Air+09] was employed only for the language pair EN-PT and was excluded in extracting EN-FR and FR-PT bilingual pairs. The statistics of the training and test datasets (validated bilingual lexicon) are as shown in Table 4.4. Randomly selected 5% of the validated lexicon is set aside as the test set. I repeat experiments for comparing the experimental results related to the size of the training corpus by taking into account randomly extracted 50%, 75%, 80%, 90% and the entire 95% of the training set.

[1] constructed separately using the first and second language terms in the accepted bilingual training data

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

[3] 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/

Table 4.4: Training and Testing Data Statistics

Data Sets		EN-PT			EN-FR			FR-PT		
		Accepted	**Rejected**	**Total**	**Accepted**	**Rejected**	**Total**	**Accepted**	**Rejected**	**Total**
Training	95%	853,452	575,951	1,429,403	362,017	51,054	413,071	372,306	78,754	451,060
	90%	768,105	518,356	1,286,461	342,963	48,370	391,333	352,711	74,609	427,320
	80%	682,761	460,761	1,143,522	304,856	42,996	347,852	313,521	66,319	379,840
	75%	640,088	431,963	1,072,051	285,803	40,308	326,111	293,926	62,174	356,100
	50%	426,725	287,976	714,701	181,009	26,871	207,880	195,952	41,449	237,401
Test	5%	44,920	30,312	75,232	19,053	2,687	21,740	19,595	4,145	23,740

4.3.2 Baseline

The baseline classifier is trained using the bilingual pairs characterised by the orthographic similarity measures such as the Edit distance based similarity measure and Phrase Similarity Measure, and using the frequency based features such as Dice coefficient and minimum to maximum frequency ratio.

4.4 Results and Evaluation-SVMTEC-2

In the current section, I discuss the classification results and the performance of the classifier with respect to various features using the complete data set (95%) introduced in the Section 4.3.1 for each of the language pairs EN-PT, EN-FR and FR-PT. In the Table 4.5 and Figures 4.1 through 4.3, $\mathrm{BL}_{strsim+freq}$ represents the baseline, constituting of string similarity measures (discussed in Section 4.2.1) and frequency based measures (discussed in Section 4.2.2). BE_{SW} and $\mathrm{BE}_{\mathbf{Pat_{R\text{-}A}}}$ respectively represent the bad translation endings that are stop words and those endings frequently observed in rejected training data that are absent as translation endings in the accepted training data set. MC and MC_{stem} represent translation mis-coverage estimated using the training data set and the stemmed training data set respectively.

The Table 4.5 shows the precision (P_{Acc}, P_{Rej}), recall (R_{Acc}, R_{Rej}) and the accuracy of the estimated classifier in predicting each of the classes (*Acc* and *Rej*) while using different features. Micro-average Recall (μ_R), Micro-average Precision (μ_P), and Micro-average f-measure (μ_F) are used to assess the global performance over both classes. Each of these evaluation metrics are computed as discussed in Chapter 3.

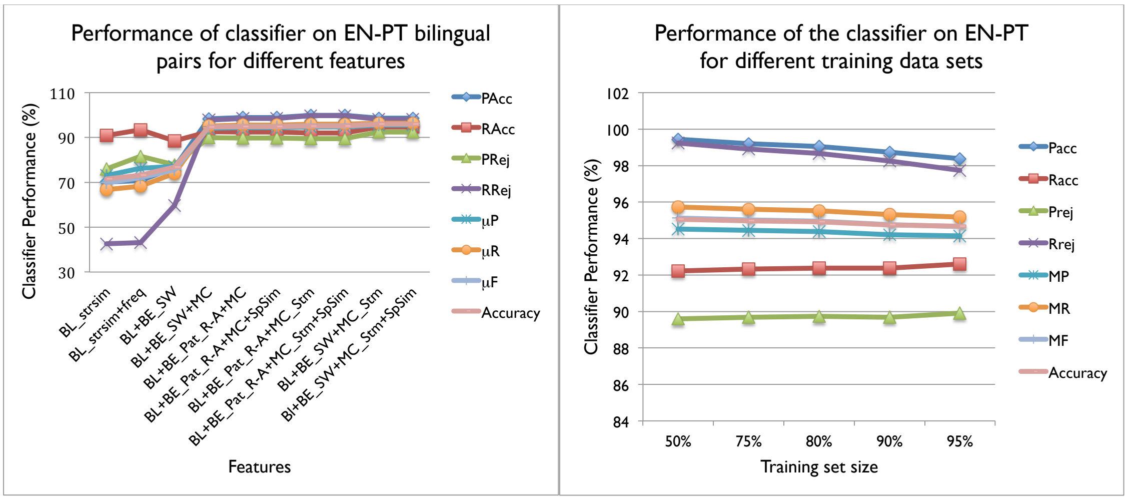
As might be seen from the Table 4.5 and Figure 4.1 (left), for EN-PT, substantial improvement is achieved by using the feature that looks for translation coverage on both sides of the bilingual pair. An increase in μ_F of 22.85% is observed over the base line and 19.32% over a combination of the features representing baseline and bad ends. Best μ_F is obtained when the stemmed[4] lexicon is used to look for stem coverage rather than the original lexicon. However, for EN-FR, training with stemmed lexicon did not show a meaningful improvement.

[4] stemmed using the snowball stemmer

Table 4.5: Classifier Results using different features for EN-PT, EN-FR and FR-PT

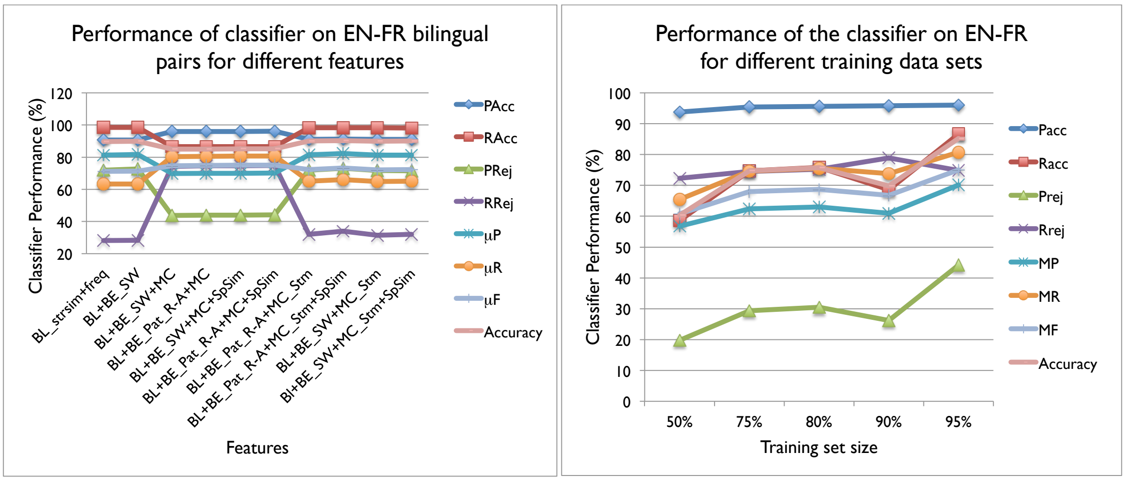
Language Pairs	Features	P_{Acc}	R_{Acc}	P_{Rej}	R_{Rej}	μ_P	μ_R	μ_F	Accuracy
EN-PT	$\mathbf{BL_{strsim+freq}}$	70.87	93.47	81.66	43.08	76.27	68.28	72.05	73.17
	$\mathbf{BL + BE_{SW}}$	76.50	88.47	77.76	59.73	77.13	74.10	75.58	76.89
	$\mathbf{BL + BE_{Pat_{R\text{-}A}} + MC}$	98.93	92.41	89.75	98.52	94.34	95.47	94.90	94.87
	$\mathbf{BL + BE_{Pat_{R\text{-}A}} + MC_{Stm}}$	99.85	92.03	89.42	99.80	94.64	95.92	95.27	95.16
	$\mathbf{BL + BE}_{SW} \mathbf{+ MC}_{Stm}$	98.64	94.63	92.50	98.06	95.57	96.35	**95.96**	96.02
EN-FR	$\mathbf{BL_{strsim+freq}}$	90.67	98.45	71.89	28.17	81.28	63.31	71.18	89.76
	$\mathbf{BL + BE_{SW}}$	90.69	98.50	72.73	28.28	81.71	63.39	71.39	89.83
	$\mathbf{BL + BE_{Pat_{R\text{-}A}} + MC}$	96.03	86.56	43.92	74.62	69.98	80.59	74.91	85.09
	$\mathbf{BL + BE_{Pat_{R\text{-}A}} + MC + SpSim}$	96.07	86.63	44.11	74.84	70.09	80.74	**75.04**	85.17
	$\mathbf{BL + BE_{Pat_{R\text{-}A}} + MC_{Stm}}$	91.10	98.25	71.98	31.93	81.54	65.09	72.39	90.05
	$\mathbf{BL + BE_{Pat_{R\text{-}A}} + MC_{Stm} + SpSim}$	91.34	98.23	73.04	33.98	82.19	66.11	73.26	90.29
FR-PT	$\mathbf{BL_{strsim+freq}}$	85.12	97.85	65.30	19.16	75.21	58.51	65.81	84.11
	$\mathbf{BL + BE_{SW}}$	85.12	97.83	65.05	19.13	75.09	58.48	65.75	84.09
	$\mathbf{BL + BE_{SW} + MC}$	88.80	74.55	31.58	55.54	60.19	65.05	62.52	71.23
	$\mathbf{BL + BE_{Pat_{R\text{-}A}} + MC + SpSim}$	88.87	75.54	32.35	55.30	60.61	65.42	62.92	72.01
	$\mathbf{BL + BE_{Pat_{R\text{-}A}} + MC_{Stm}}$	85.12	97.83	65.05	19.13	75.09	58.48	65.75	84.09
	$\mathbf{BL + BE_{Pat_{R\text{-}A}} + MC_{Stm} + SpSim}$	85.13	97.87	65.54	19.18	75.34	58.53	**65.87**	84.13

Figure 4.1: Performance of the Classifier for EN-PT using different features (left) and for different training sets (right).



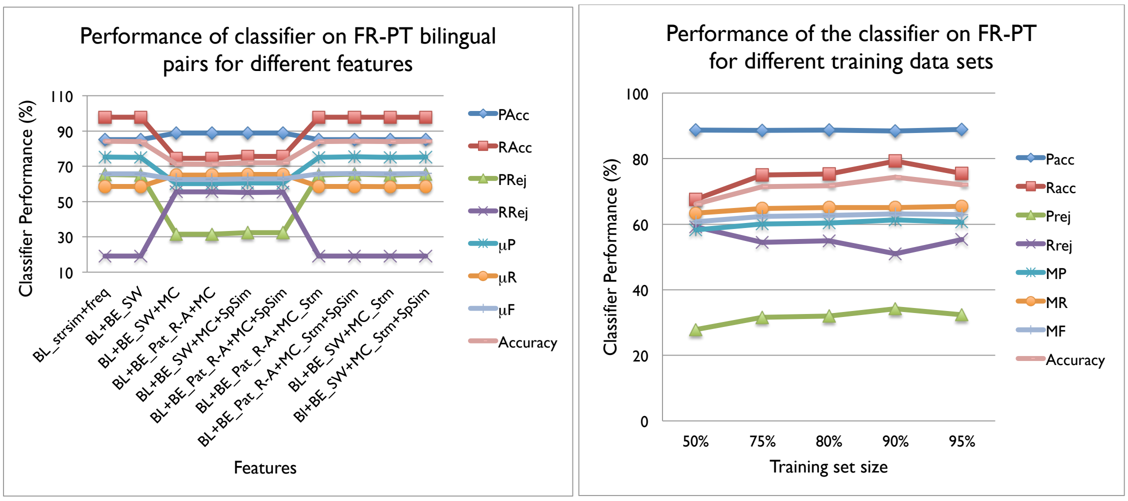
FR-PT results are worse than the results obtained for other language pairs: the best μ_F and accuracy of 65.87% and 84.13% respectively are obtained when a combination of features BL+$BE_{Pat_{R-A}} + MC_{Stm} + SpSim$ is used. However, the improvement is negligible (approximately ranging from 0.01% - 0.14%) against the baseline ($BL_{strsim+freq}$) in every terms (precision, recall and micro f-measure) over both classes. This may be explained because the number of 'single word - single word' pairs is comparatively larger than for the other language pairs and the number of 'multi-word - multi-word' pairs is small (50,552 for the accepted). Approximately 250K French multi-words are paired with single Portuguese words and approximately 9K Portuguese multi-words are paired with

Figure 4.2: Performance of the Classifier for EN-FR using different features (left) and for different training sets (right).



single French words. Moreover, approximately 130K are single word pairs for this pair of languages which is quite different from the EN-PT scenario.

Figure 4.3: Performance of the Classifier for FR-PT using different features (left) and for different training sets (right).



Also, patterns indicating bad ends that are stop words (BE_{SW}) are substantially few in number with respect to FR-PT and EN-FR lexicon corpus as opposed to EN-PT. This is because extractions for these language pairs use all of the techniques mentioned in section 4.3 except for the suffix array based extraction technique [Air+09]. Hence EN-FR and FR-PT were much cleaner. Most frequent patterns representing bad ends are shown in Table 4.6.

Table 4.6: Patterns representing bad ends for EN-PT, EN-FR and FR-PT

Language Pairs	**#Patterns**	**#Patterns**	**Frequent Pattern**	**#Occurrences**	**Frequent Pattern**	**#Occurrences**
(L_1-L_2)	**(L_1)**	**(L_2)**	**(L_1)**	**(L_1)**	**(L_2)**	**(L_2)**
EN-PT	112	86	the	27,455	a	22,242
EN-FR	43	15	to	210	pas	237
FR-PT	5	8	de	27	de	43

4.4.1 Classifier Performance by Training Set Size

I analysed the impact of varying the size of training datasets on the improvement given by various features. Table 4.7 and the plots shown in figures 4.1, 4.2 and 4.3 (right), respectively, show the results obtained using the features $BL_{strsim+freq}$+BE_{SW}+MC (EN-PT) and $BL_{strsim+freq}$+$BE_{Pat_{R-A}} + MC + SpSim$ (EN-FR and FR-PT).

Table 4.7: Classifier Results for EN-PT, EN-FR and FR-PT by training set sizes

Language Pairs	**Training**	**P_{Acc}**	**R_{Acc}**	**P_{Rej}**	**R_{Rej}**	**μ_P**	**μ_R**	**μ_F**	**Accuracy**
(Test Set)	**Data set**								
EN-PT	50%	**99.45**	92.22	89.59	**99.24**	**94.52**	**95.73**	**95.12**	**95.05**
	75%	99.21	92.32	89.68	98.90	94.45	95.61	95.02	94.97
	80%	99.04	92.38	89.73	98.67	94.39	95.53	94.95	94.91
	90%	98.74	92.38	89.69	98.25	94.22	95.32	94.76	94.74
	95%	98.38	**92.60**	**89.91**	97.74	94.15	95.17	94.65	94.67
EN-FR	50%	93.75	58.62	19.77	72.31	56.76	65.47	60.80	60.31
	75%	95.41	74.77	29.39	74.47	62.40	74.62	67.97	74.73
	80%	95.59	75.82	30.49	75.21	63.04	75.52	68.72	75.74
	90%	95.85	68.59	26.17	**78.94**	61.01	73.77	66.78	69.87
	95%	**96.07**	**86.63**	44.11	74.84	**70.09**	**80.74**	**75.04**	**85.17**
FR-PT	50%	88.68	67.57	27.86	**59.20**	58.27	63.39	60.72	66.11
	75%	88.62	75.05	31.59	54.45	60.11	64.75	62.34	71.46
	80%	88.76	75.29	31.99	54.93	60.38	65.11	62.65	71.74
	90%	88.43	79.29	**34.22**	50.95	**61.33**	65.12	**63.17**	**74.34**
	95%	**88.87**	**75.54**	32.25	55.30	60.61	**65.42**	62.92	72.01

Looking at the classification results for EN-PT using SVM and the training set, it is observed that the larger the training set larger the recall (R_{Acc} is 92.6% against 92.22%) for the *'Accepted'* class. Meanwhile, when the training set is augmented, precision falls from 99.45% to 98.38%. However, by augmenting the training set, the precision improved (R_{Acc} from 89.59% to 89.91%) for the *'Rejected'* class, whereas the recall dropped (R_{Rej} from 99.24% to 97.74%). As the training set is much larger than for other language pairs (95% of the corpus), not necessarily much is gained. Thus, precision and recall for EN-PT does evolve in a way, such that, while one augments the other tends to decrease, partially deviating from the trend observed in my earlier experiments [Kav+11].

Unlike EN-PT, for the language pairs EN-FR and FR-PT, with larger training sets the performance of the trained classifier improved. For the features listed in Table 4.5, best results were obtained with 95% and 90% of the training set.

4.4.2 Classifier trained on one language pair in classifying others

Motivated by the classifier performance for language pairs EN-PT, I conducted few more experiments: I trained the classifier using the full set of features on one language pair, and tested on the other. Training on EN-PT data and testing on EN-FR and FR-PT resulted in μ_F of 55.64% and 54.99%, far below the baseline for EN-FR (a drop by approximately 15% from 71.18%) and FR-PT (a drop by approximately 11% against 65.81%) respectively. Training the system with EN-FR and testing on FR-PT did even worse, leading to a micro-average f-measure of 52.96%. Training on FR-PT data and testing on EN-FR, led to a μ_F of 47.8%. This lets me to conclude that it does not make any sense to use a classifier trained on one language pair in classifying the data from other language pairs. The related results are shown in Table 4.8.

Table 4.8: Performance of Classifier trained on one language pair when tested on others.

Language Pairs (Test Set)	Classifier Trained	P_{Acc}	R_{Acc}	P_{Rej}	R_{Rej}	μ_P	μ_R	μ_F	Accuracy
EN-FR	Train with EN-FR model	96.03	86.56	43.92	74.62	69.98	80.59	74.91	85.09
	Train with EN-PT model	89.07	88.55	22.02	22.93	55.55	55.74	55.64	80.44
	Train with FR-PT model	86.85	70.55	10.39	24.23	48.62	47.39	47.80	64.82
FR-PT	Train with FR-PT model	88.80	74.55	31.58	55.54	60.19	65.05	62.52	71.23
	Train with EN-PT model	85.71	59.66	21.74	52.98	53.73	56.32	54.99	58.49
	Train with EN-FR model	84.96	46.00	19.42	61.52	52.19	53.76	52.96	48.71

4.5 Summary

In this chapter, I have discussed the classifier model SVMTEC-2, an evolution of the classifier SVMTEC. Experimental results demonstrate the use of the classifier SVMTEC-2 on EN-PT, EN-FR and FR-PT language pairs under small, medium and large data conditions.

Automatically extracted bilingual translations after human validation, are subsequently used for realigning the parallel corpora and extracting new translations forming an indefinite cycle of iterations. Automatic classification prior to validation contributes to speed up the process of distinguishing the correct translations from naturally occurring alignment and extraction errors. The positive side effect is an enriched annotated lexicon suitable for machine learning systems such as bilingual morphology learning and translation suggestion tool, apart from its primary use as an aid in alignment, extraction and translation.

Chapter 5

Selection of Word-to-word Translation Equivalents as a Classification Problem: a third approach

In the discussions presented in Chapter 2, it was shown that not all of the extracted translations are good enough to be used for subsequent learning. A few inadequate extractions (multi-word translations) were illustrated in the Section 2.1 of the Chapter 2. In the current chapter, I shift my focus on identifying inadequate word-to-word translations and their subsequent segregation as *'incorrect'* class. In the section 5.1, I introduce the characteristics of automatically extracted word-to-word translations and the motivation for classification. Features specific to word-to-word translations are elaborated in Section 5.2. The datasets, experimental evaluations and the results for language pairs EN-PT and FR-PT are discussed in Sections 5.3 and 5.4.

5.1 Word-to-Word Translations

In the context of word-to-word translations, inadequacy is attributed to syntactic (inflectional) or/and semantic (stem context) disagreements, a few of which are as discussed below.

The extracted translation candidate '*transported* ⇔ *transportar*', labelled manually as '*rejected*', is an example indicating disagreement in the morphological inflection. On the other hand, '*transported* ⇔ *transformação*' is an example indicating inappropriate stem contexts. The correct translations being '*transported*' ⇔ '*transportou*' | '*transportaram*' | '*transportados*' | '*transportado*' | '*transportada*' | '*transportadas*', or '*transport*' ⇔ '*transportar*' (from the translation perspective of the word in PT), the first example points out to the underlying disagreement between the suffix, '*ed*' in first language (EN) and the suffix, *'ar'* in second language (PT) and can be used as a feature characterising *'rejected'* entries. The latter example indicates a disagreement in the stem translation (or semantics), as, *'transported'* should translate into one of the above mentioned forms (or the translation *'transformação'* is acceptable provided its EN counterpart is *'transformation'*). Further, the candidate translation extracted, *'observation'* ⇔ *'observações'*, manually labelled as *'rejected'*, instantiates translation candidates, where, inflections do not match in number (singular noun vs plural noun). These observations stress on the need to bilingual morphological learning and the associated feature extraction (using both the accepted and rejected translations) that will enable the classification of the extracted translations with the accuracy nearing that achieved with human validation.

In the Chapters 3 and 4, I had elaborated on the classification of extracted translations (both word-to-word and multi-word translations) using the SVM-based classifiers SVMTEC and SVMTEC-2, trained with the features, such as, occurrence frequencies of terms in the aligned segments, orthographic similarity measured using Levenshtein Distance, SpSim and Phrase_SpSim. Further, the features indicating translation mis-coverage and stemmed mis-coverage were used to uncover the existence of translation gaps (missing translations in one language with respect to the other language and vice-versa) [Kav+11; Kav+15b]. However, certain translation candidates (such as those discussed above), manually labelled as 'rejected', were misclassified by the classifier as 'accepted'. This is because, the morphological based features, reflecting the underlying (dis)agreements between bilingual stems and suffixes, were not considered in training the classifier, leading to false positives. To avoid such classifier errors, I adapt the translation coverage based features to represent the morphological coverage in candidate translations considering bilingual stem correspondence and the suffix correspondence induced from the bilingual morphology learning, a subject matter that is explicitly treated in Chapter 7[1]. I include four additional features (hereafter, morphological coverage feature), that looks for the stem and its translation (likewise, considering suffix and its translation) to reflect whether the morphological gap exists in the bilingual pair to be validated. Further, whether the bilingual stems and suffixes belong to the same cluster is indicated using a binary valued feature. The performance of the classifier trained with these additional features when tested on word-to-word translations is evaluated [Kav+14b].

5.2 Features

The classifier for word-to-word translations was trained using the orthographic similarity based features discussed in Chapter 3, apart from the features regarding stem pairs, suffix pairs and suffix classes each of which are elaborated below.

[1] If you feel a bit lost, please read first the chapter 7 as there the learning procedure is explained, as well as the results obtained in the generation of Out-Of-Vocabulary word-to-word bilingual entries. Then you may better understand the results that are used in this chapter.

5.2.1 Morphological Coverage

The morphological coverage of a bilingual pair refers to the contextual (stem) and inflectional (suffix) coverage exhibited by the bilingual pair under consideration. More precisely, the coverage is determined as the agreement between morphological units comprising of stem in one language and its translation in another language and between suffix in one language and its translation in other language, respectively. The features are binary valued, each representing stem coverage and suffix coverage, thereby adding two features characterising the bilingual pair to be validated.

The bilingual morph-units and suffix classes identified using the method discussed in Chapter 7 [Kav+14a] are used to extract features for training the SVM based classifier in addition to the orthographic similarity based features (here used as abaseline) adopted in training the classifier SVMTEC [Kav+11]. To represent the morphological coverage, first, two separate key-word trees are constructed for the words in EN and in PT using the bilingual pairs labelled as accepted in the training data. Each keyword tree is constructed using the *stem part* of the word learnt using the bilingual learning approach (see Chapter 7) [Kav+14a]. Similarly, the training and the test sets are represented using their stems. The procedure for identifying coverage (with respect to stems) follows the Aho-corasick set-matching algorithm [Gus97]. The set of all stems in first language is matched against the left hand side term in first language of the bilingual pair to check if a stem in the keyword tree (constructed from the bilingual training data separately for first language) occurs in the left hand side of the bilingual pair. Similarly, by matching all stems in second language, it is checked if a stem in the keyword tree (constructed from the bilingual training data separately for second language) occurs in the right hand side of the bilingual pair. 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 bilingual pair is said to be covered with respect to stem or is said to share same contexts.

To represent the coverage based on suffixes, two more key-word trees are constructed separately for suffixes in first and second language terms, that are learnt from the accepted training data [Kav+14a] (refer Chapter 7). Analogous to the stem set matching discussed in the previous paragraph, using the Aho-corasick set-matching algorithm, it is checked if the the bilingual pair to be validated ends with the suffix in the keyword tree (separately constructed from the bilingual suffixes learnt for EN and PT terms). If the bilingual pair ends with the suffixes in the keyword tree and the matched suffixes form valid bilingual suffix pairs, then the bilingual pair satisfies the suffix agreement requirement and hence is covered with respect to suffix.

5.2.2 Stem-Suffix Agreement

Apart from verifying the morphological coverage, it is checked if the bilingual morph-like units ((stem$_{L1}$, stem$_{L2}$) and (suffix$_{L1}$, suffix$_{L2}$)) that constitute the bilingual pair belong to the same bilingual suffix class.

5.3 Experimental Setup

SVM based tool namely LIBSVM[2] was used to learn the classifier. In the experiments discussed, the radial basis function kernel, with parameters g=32, C=0.5 was used. The values presented for g and C reflect the best cross-validation rate.

5.3.1 Data sets

Data consists of a set of bilingual pairs in EN-PT and FR-PT representing word-to-word translations taken from existing bilingual lexicon with the entries manually tagged as being accepted or rejected. The extraction techniques and the parallel corpora used in acquiring the bilingual lexicon are as elaborated in Section 4.3.1. The details of the training and test sets are shown in Table 5.1. For the language pair EN-PT, the training and test data sets were formed from a total of 209,739 accepted and 72,138 rejected single word translations. Similarly, a total of 122,759 accepted and 48,599 rejected word-to-word pairs were used in framing the test sets for FR-PT. In each of the cases, the training data constituted of 95% accepted and rejected pairs, while the remaining 5% each of the accepted and rejected pairs formed the test sets.

Table 5.1: Training and Testing Data Statistics for Word-to-Word Translations

Data Sets		EN-PT			FR-PT		
		Accepted	Rejected	Total	Accepted	Rejected	Total
Training	95%	199,253	68,529	267,782	116,621	46,169	162,790
Test	5%	10,486	3,609	14,093	6,138	2,430	8,568

5.3.2 Baseline

The baseline classifier for the word-to-word translations was trained using the orthographic similarity based features discussed in Chapter 3.

5.4 Results and Evaluation

Figures 5.1 and 5.2 show the precision (P_{acc}, P_{rej}), recall (R_{acc}, R_{rej}) and the accuracy of the estimated classifier in predicting each of the classes (accepted, *acc* and rejected, *rej*) when trained with different features for the language pairs EN-PT and FR-PT.

Global performance over both classes is estimated by computing the Micro-average Recall (μ_R), Micro-average Precision (μ_P), and Micro-average f-measure (μ_F). For the two language pairs considered, the Table 5.2 shows the μ_P, μ_R and μ_F obtained in classifying

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

Figure 5.1: Precision-Recall graph showing performance of the classifier for EN-PT word-to-word translations

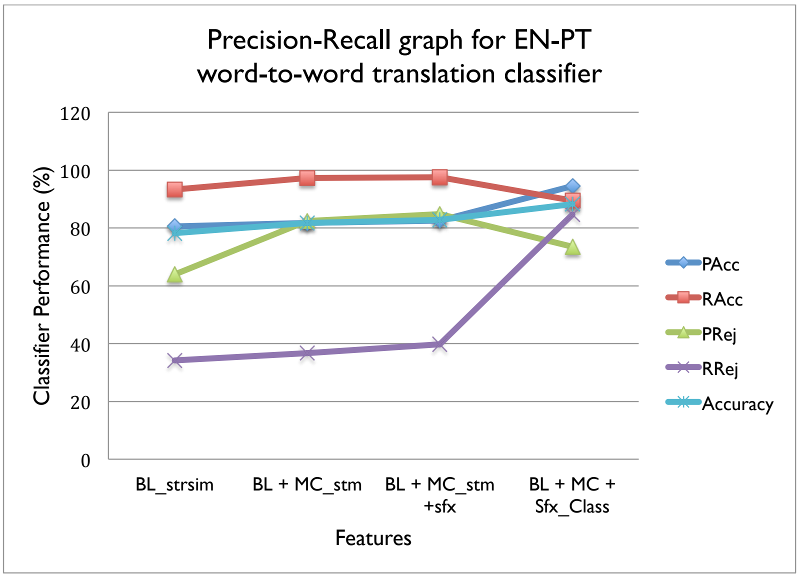
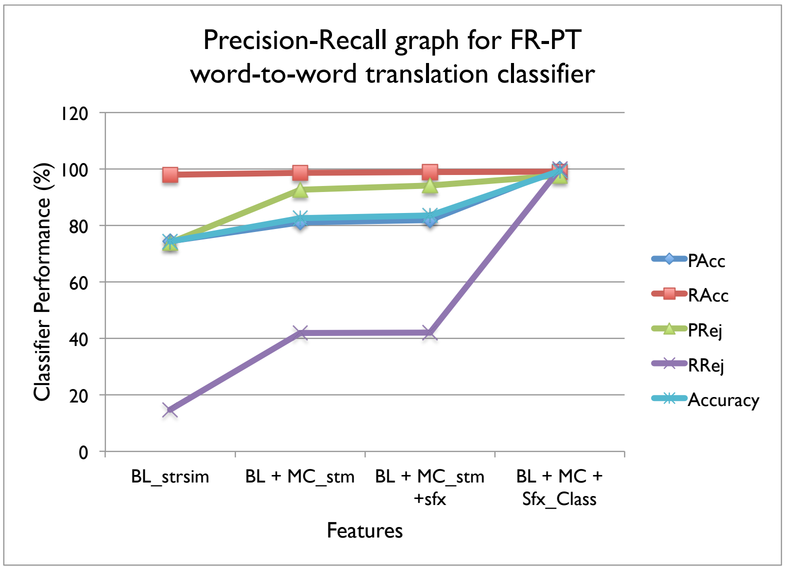


Figure 5.2: Precision-Recall graph showing performance of the classifier for FR-PT word-to-word translations



the word-to-word translations using various features. These evaluation metrics were defined in Section 3.4 of Chapter 3.

By adapting the features indicating morphological coverage with respect to bilingual stems and suffixes, the micro-average f-measure attained is 75.39%, which shows an improvement of 6.01% over the baseline (BL) for the language pair EN-PT. The morphological coverage added with the suffix class feature enabled a micro-average f-measure of 85.51%, almost 17.79% above BL. Similarly, in the case of FR-PT, the morphogical coverage feature enabled substantial improvement in the global performance of the classifier,

Table 5.2: Performance of the classifier trained on EN-PT and FR-PT word-word-translations for different features

Language Pairs	Features	$\mathrm{P}_{\mathrm{Acc}}$	$\mathrm{R}_{\mathrm{Acc}}$	$\mathrm{P}_{\mathrm{Rej}}$	$\mathrm{R}_{\mathrm{Rej}}$	μ_{P}	μ_{R}	μ_{F}	Accuracy
EN-PT	**BL$_{\mathbf{strsim}}$**	80.48	93.35	63.89	34.18	72.19	63.77	67.72	78.21
	BL + MC$_{stm}$	81.73	97.29	82.36	36.76	82.05	67.03	73.78	81.80
	BL + MC$_{stm}$ + MC$_{sfx}$	82.47	97.54	84.74	39.73	83.61	68.64	75.39	82.74
	BL + MC + SuffixClass	94.45	89.48	73.48	84.72	83.97	87.1	85.51	88.26
FR-PT	**BL$_{\mathbf{strsim}}$**	74.38	97.93	73.87	14.77	74.13	56.35	64.03	74.35
	BL + MC$_{\mathbf{stm}}$	81.10	98.70	92.71	41.89	86.91	70.3	77.73	82.59
	BL + MC$_{\mathbf{stm}}$+MC$_{\mathbf{sfx}}$	81.96	98.89	94.15	42.02	88.06	71.96	79.20	83.61
	BL + MC + SuffixClass	100	99.07	97.71	100	98.86	99.54	99.20	99.33

contributing to a micro-average f-measure of 79.20%, an increase by 15.17% over the BL.

5.4.1 Summary

In the current section, I have discussed the use of bilingual stem and suffix correspondences in classifying EN-PT and FR-PT word-to-word translations. The features discussed in Chapters 3 and 4 do not identify morphological disagreements in bilingual pairs, thereby resulting in false positives while classifying word-to-word translations. The morphological coverage feature discussed in this chapter might be viewed as a variation of the translation (mis-)coverage feature, where sub-expressions correspond to stems and suffixes. The feature values are extracted by looking for stem and suffix agreements in the bilingual pair under consideration using the bilingual morph-units and the suffix classes identified by employing the methods discussed in Part II (Chapters 7 and 8), where a better explanation on how such bilingual morph-units are learnt with the motivation (Chapter 6) could be found.

By adapting the morphological coverage feature to classify the word-to-word translations, a substantial improvement was observed in the classifier performance over the BL for both language pairs. This led to an overall improvement in the classifier performance over that obtained when all the translations (both word-to-word and multi-word) were considered [Kav+11].

Part II

Learning Bilingual Segments and Suffix Classes for Generating OOV Lexicon Entries

Chapter 6

Introduction to Bilingual Morphology Learning and Generation of OOV Lexicon entries

This chapter serves as a general introduction to the two approaches for bilingual learning discussed in forth-coming chapters, one of which is unsupervised (Chapter 7) and the other is minimally supervised (Section 8). An overview of the existing lexicon with respect to the availability of near translation forms and the need for dealing with OOV lexicon entries by bilingual learning is presented in Section 6.1. Related research methods are reviewed in Section 6.2.

6.1 Introduction and Background

In Part I of this thesis, one aspect of the translation lexicons concerning the quality of bilingual entries from the perspective of their subsequent usage in parallel corpora alignment was discussed. In this chapter, as a scope for continual improvement, the coverage aspect of the translation lexicon is attended to and elaborated with the objective of tackling OOV lexicon entries utilising the available word-to-word translation forms.

Fundamental to the discussions in this part of the thesis are the bilingual translation lexicons acquired from an aligned parallel corpora[1] [GL09] using various extraction methods [Air+09; Bro+93; GL11; LL09]. However, the lexicon thus acquired is not complete as it does not contain all possible translation pairs. Table 6.1 shows the accepted translations extracted for each of the word forms corresponding to *ensure*. The translations seem exhaustive with respect to the first 2 columns. However, certain missing translation forms are, *garantam* for *ensure*, *garantiu*, *garantiram*, *permitidas*, *permitido*, *permitidos*,

[1] A collection of pair of texts that are translations of each other.

permitiu, *permitiram* for *ensured*, which can also be considered as possible translations. All the translation forms in the 3rd column are missing. This is because, the extraction techniques cannot handle what is not in the parallel corpora used for extracting translations, unless we care about automatically learning and generalising word and multi-word structures. Moreover, they are not able to extract everything.

Table 6.1: Translation Patterns in the extracted lexicon for the language pair EN-PT

Term (EN)	**Term (PT)**			
ensure	assegurar asseguram assegurem	zelar	garantir garantem	permitir permitam permitem
ensures	assegura assegure		garanta garante	permite permita
ensured	asseguradas assegurados assegurado assegurou asseguraram		garantidas garantido garantidos	
ensuring	assegurando		garantindo	permitindo

On a whole, although it is evident that the existing lexicon is reasonably extensive, acknowledging its incompleteness with respect to the vocabularies in either or both languages involved, accommodating most of the possible patterns demand learning the translation structure. I focus on generating word-to-word translations, by treating a translation lexicon itself as a parallel corpus. Having a hugely high degree of certainty associated with each bilingual pair asserting its correctness, I use it to learn and generalise translation patterns, infer new patterns and hence generate those OOV translation pairs that were not explicitly present in the training corpus used for the lexicon acquisition.

From Table 6.1, it might be observed that each of the terms in EN share the same set of suffixes *-e*, *-es*, *- ed* and *-ing* and stem *ensur*. Including their corresponding translations, one can observe that a term in EN ending with *-ed* is translated to a term ending with *-adas*, *- ados*, *-ado*, *-ou*, *-aram*. Likewise, the translations for *declared*[2] follow similar pattern and the translations end with *-ada*, *-adas*, *-ado*, *-ados*, *-aram*, *-ava*, *-ou*. Equivalently they share the stems *assegur* and *declar*[3]. Knowing that *ensured* and *declared* share suffixes *-ed* and by considering the intersection of suffixes corresponding to their translation endings in PT, it could be seen that a term with suffix *-ed* in EN might be translated to terms with suffixes *-adas*, *-ados*, *-ado*, *-ados*, *-aram*, *-ou* in PT. This simple knowledge allows new translation pairs to be generated based on the similarities observed from the known examples. But, in the 4th column of the Table 6.1, the translation endings such as, *-ido*, *-idas*, *-idos* corresponding to *-ed* are related and may be used to generate *permitido*,

[2] declared ⇔ *declarada, declaradas, declarado, declarados, declararam, declarava, declarou*

[3] Stem can be determined as longest common sequence of characters

permitidas, permitidos thereby partially completing the translations for *ensured* in column 5[4].

Moreover, clusters of suffixes in English may translate as different clusters of suffixes in Portuguese and hence it is necessary to identify, for a specific case the best selection. Referring Table 6.1, it might be observed that suffix *-e* in English maps to *-am, -em, -ar* (inflections for root assegur), *-ir, -em* (inflections for root garant) and *-am, -em, -ir* (inflections for root permit). Suffixes *-am, -em* are shared by all the three verb forms, while the suffix *-ar* discriminates the verbs in *-ar* group from the verbs in *-ir* group. Equivalent phenomena occurs for the other suffixes. Generally, it is seen that verbs belonging to *-e, -es, - ed, -ing* in English could be mapped to verbs belonging to one of the three Portuguese conjugation classes *-ar, -ir* or *-er* with the classes being discriminated by the ending of their infinitive forms. It is to be noted that, by chance, the suffix *-ou* corresponding to *-ed* in column 1 of the Table 6.1, is also a discriminator for Portuguese verbs belonging to *-ar* group.

A snapshot of the bilingual lexicon for EN-PT, depicted in Table 6.1, very well demonstrates a huge amount of inflectional variations for the PT language, despite the scarcity of inflections for EN. The bilingual suffixation approach discussed in the Section 7.1 of the Chapter 7 is particularly suitable for capturing bilingual morph-like units from such lexicons. In contrast to the lexicons demonstrated in Table 6.1, is the EN-HI bilingual lexicon, having very few inflectional variants for words or their inflectionally modified translations. The major challenges that has to be addressed in this setting pertain to the handling of previously unseen bilingual pairs and dealing with the segmentation ambiguities due to rather limited training data available for EN-HI, apart from the problem of limited translation forms, as opposed to the large lexicon available for EN-PT with the added advantage of sufficient near translation forms[5].

The approach discussed in the Chapter 8 of this thesis represents a minimally supervised strategy for the discovery of bilingual morph-like units that is intended to adapt reasonably well under limited training data conditions. The motive behind categorising it as a minimally-supervised approach is the very fact that the segmentations and clusters identified using the approach similar to that discussed in Sections 7.1 and 7.2 of the Chapter 7 serve as training data in learning the segmentations for previously unseen bilingual pairs. However, manual validation is incorporated over those automatically learnt segmentations and clusters. This strategy may not be considered as a practical approach for word segmentation task on its own, as, the segmentation standards do vary depending on the final objective[6], and therefore under the said scenario, labeled examples are employed for achieving an acceptable satisfactory performance. Furthermore, I rely on the supervised model owing to its proved significantly lower error rates when

[4] This follows as verbs *garantir* and *permitir* belong to the 3^{rd} verb inflection class, unlike *assegurar* which belongs to the 1^{st} verb inflection class.

[5] EN-HI training data set constitutes of just one-fourth of the entries in EN-PT lexicon

[6] My intention is suggesting OOV Translations

sufficient labelled data is provided. Features represent global distribution of morph-like units with respect to parameters such as the length of bilingual pairs, apart from the instance-specific features derived from lexicon entries used as training data.

6.2 Related Work

The fact that 'words consist of high-frequency strings (affixes) attached to low-frequency strings (stems)' has motivated several researches ranging from text analysis for acquisition of morphology, to learning suffixes and suffixation operations for improving word coverage and for allowing word generation. Certain approaches are unsupervised [Déj98]. 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 [Lin+09]. Unsupervised, Minimum Description Length based models focus on finding a better compressed representation for lexicon of words [Gol01], [CL02]. Other unsupervised approaches mainly address language specific issues such as data and resource sparseness [HB11], agglutination [Mon+09] and so forth. In each of the mentioned studies, morphological segmentation is determined considering monolingual data.

6.2.1 Monolingual Approaches

Lexical inference or morphological processing techniques have been established to be interesting in suggesting translations for OOV words that are variations of known forms. Below, I discuss a few of them from a monolingual perspective.

The hierarchical back-off model for translating unseen forms stand out to act on highly inflectional languages such as German and Finnish, particularly under the scenario of limited training data [YK06]. Morphological decompositions mainly include alternative layers of *stemming* and *compound splitting*, requiring that "a more specific form (a form closer to the full word form) is chosen before a more general form (a form that has undergone morphological processing)". Yang et al. [YK06] investigate all possible ways of segmentation with the only constraint that each part has a minimum length of 3 characters. Acceptance of the segmentation is subjected to the appearance of subparts as individual items in the training data vocabulary. The approach relies on translation probabilities derived from stemmed or split versions of the word in its phrasal context. Experiments with varied amount of training data reveal its appropriateness under limited training data conditions and adaptability to highly inflected languages.

Gispert et al. [Gis+05] show that translations for unseen verb forms can be generated by generalising them using the verb forms seen in training data. Verbs are identified using rules incorporating word forms, POS-tags and word lemmas and are classified to the lemma of their head verb, such that they belong to only one class with such a classification done for each language separately. The instance model is estimated based on the relative frequency of each instance across all tuples (pair of source and target classes) that share the same source phrase. To translate an unseen verb form, the verb is classified into the lemma of its head word and all the tuples representing translation of that class of verbs (in the training data) are identified. If the verb form to be translated is not found among all seen instances of the identified tuples (after excluding personal pronoun or verb suffix), identical instances in terms of words, POS tags and lemmas are looked for in each tuple. For each identical instance found, new target verb form is generated by replacing the personal pronoun in the seen form with the personal pronoun in the expression to be translated. The suggested translation is weighed based on the frequency of its occurrence in the training data. In case of any ambiguity in generalisation of verb forms (a personal pronoun such as *'you'* translates in more than one way), the approach over-generates all possible forms, leaving the target language model to decide on the best translation alternative.

The need for dealing with the language-specific problems while translating from English to morphologically rich languages by identifying those morphological relationships that are left un-captured by current SMT models, the possibilities of handling these by morphology derivation, independent of the translation model, are discussed by Gispert and Marino [GM08]. Proper derivations are introduced into the texts by simplifying morphological information (or parts of it), followed by morphology generation by means of a classification model which makes use of a set of relevant features for each simplified morphology word and its context. The study reveals that the main source of potential improvement lies in the verb form morphology as this morphological category is seen to exhibit more derivation in Romance languages.

The discriminative log-linear model proposed by Poon et al. [Poo+09] relies on overlapping features, such as, morphemes and their contexts to boost the segmentation decisions. The model incorporates two MDL inspired priors, the lexicon prior: an exponential prior with negative weight on the length of the morpheme lexicon, and the corpus prior: an exponential prior on the number of morphemes used to segment each word in the corpus, for penalising over-segmentation. Viewing the segmentation to be a set of morpheme strings and their contexts, for a corpus W (a set of words) and segmentation S splitting each word in W into prefixes, a stem and suffixes, the model defines a joint probability distribution over a restricted set of W and S:

$P_\theta(W,S) = 1/Z \cdot u_\theta(W,S)$, where

$u_\theta(W,S) = exp(\sum_\sigma \lambda_\sigma f_\sigma(S) + \sum_c \lambda_c f_c(S) + \alpha \cdot \sum_{\sigma \epsilon Pref(W,S)} L(\sigma) + \alpha \cdot \sum_{\sigma \epsilon Stem(W,S)} L(\sigma) + \alpha \cdot \sum_{\sigma \epsilon suff(W,S)} L(\sigma) + \beta \cdot \sum_{w \epsilon W} M_S(w)/L(w))$

Pref(W,S), Stem(W,S) and Suff(W,S) respectively represent the lexicon of prefixes, stems, and suffixes induced by S for W. $f_\sigma(S)$ and $f_c(S)$ respectively represent the occurrence frequency for morphemes and contexts under S, and $\theta = (\lambda_\sigma, \lambda_c : \sigma, c)$ are their feature weights. α and β are the weights for the priors. Z is the normalization constant.

The availability of labeled data makes the model applicable to supervised or semi-supervised learning. However, though the system does not rely on bilingual information, it depends on the use of mono-lingual morpheme contexts.

6.2.2 Approaches that simultaneously exploit bilingual or multi-lingual data

Unlike the approaches listed in the previous subsection that deal with monolingual data, I take advantage of bilingual data to deal with the ambiguities and complexities in decomposition by exploiting the *'frequent forms occurring in translations rather than words in one language'*. The focus is to learn bilingual suffixation, treating the bilingual lexicon as a parallel resource. Thus a major difference is that the result is a bilingual learning model, not skewed on one language at a time. No annotations on data are assumed in either languages, although the semi-supervised learning algorithm (discussed in Chapter 8) requires that the automatically segmented translations used as training data for the classifier is validated. Further, I do not rely on additional language-dependent resources, such as taggers.

Below I discuss few of the related works which exploit bilingual and multi-lingual correspondences in morphology learning to handle the problem of translating unknown terms.

6.2.2.1 Predicting Translation for Unknown Words - An Inductive Learning Mechanism [MT97]

Predicting translation for unknown words based on inductive learning mechanism is one of the earliest discussed works [MT97]. 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. For instance[7], with the known translation pairs *favorite* ⇔ *favoritos* and *favorable* ⇔ *favorável*, extracting the common parts of English terms yields PW-e1: *favor* γ[8] and different parts yields PW-e2: *ite* and PW-e3: *able*. Similarly, considering the common and different parts of target strings yield PW-p1: *favor* γ, PW-p2: *itos*, PW-p3: *ável*. Bilingual pair of common parts give PPW-1: *favor* γ ⇔ *favor* γ, and different parts give PPW-2: *ite* ⇔ *itos* and PPW-3: *able* ⇔ *ável*. Now given an unknown word *insecure* in English to be translated, the words for translation are predicted from *insecure* using PPWs in the PPW prediction process, which might be PPW-1: *in* γ ⇔ *in* γ and PPW-2: *secure* ⇔ *seguro* thereby predicting *inseguro* as a possible translation. The predicted translation(s) are ranked based on the correct or erroneous prediction frequency of PPW [9]. However, the approach suffers from an inadequate PPW dictionary.

[7] although the study targets English-Japanese, I consider English-Portuguese for illustration.

[8] γ represents a variable

[9] number of times PPW has been used in correct or erroneous prediction.

Similar to the approach discussed by Momouchi et al. [MT97], the bilingual approach proposed in this thesis is based on identifying common and different bilingual segments occurring in existing translation examples and employing them in generating new translations. I restrict the bilingual segments only to two parts, interpreting the first part as the bilingual stem and the second part as bilingual suffix. However, selection of candidate bilingual morph-units is driven by the frequency distribution of bilingual stems and suffixes. A pair of bilingual suffixes attached to the same bilingual stem indicate the suffix replacement option and hence motivates translation generation. Unlike the prediction approach adopted by Momouchi et al. [MT97], to enable generalisation, clusters of bilingual stems sharing same transformations are identified. Moreover, suggestion of new translations relies on the identified clusters and the suggested translations are manually validated. Momouchi et al. [MT97] on the other hand does not use clusters and relies on ranking of predicted translations in a feedback process.

6.2.2.2 Translating German Compounds by Compound Splitting - [KK03]

Koehn and Knight [KK03] report morphological processing as a means to learn translations for unknown German compounds from the translation of their parts. The splitting options are guided by parallel texts in such a way that all the parts should have been observed as whole word translations in the training corpus. Such a guidance is opted based on the observation that a frequency based splitting metric[10] does not allow a compound to be broken, particularly if it occurs more frequently than its parts, irrespective of it being translated in parts into English[11]. The guidance from parallel corpus relies on two translation lexicons with correspondences learnt using the toolkit Giza. While the first lexicon is learnt from original versions of parallel texts, the second is learnt from the parallel corpus with split German and unchanged English text versions in order to learn specific translations for compound parts[12]. The two lexicons are jointly used to guide the splitting process. While the approach records 99.1% accuracy with an improvement of 0.039 BLEU in German-English noun phrase translation task, any reported failure directs towards unseen parts due to the lack of training data.

The task is focussed exclusively on splitting German compound words to enable translation of compounds by the translation of their parts. However, to avoid prefixes and suffixes from splitting off, the parts are restricted only to content words such as nouns, adverbs, adjectives, and verbs by using POS tags, thus excluding the prepositions or determiners. Contrarily, segmentations in our approach are guided by common and

[10] solely defined in terms of German word frequencies. Given a compound, the split with highest geometric mean of word frequencies of its parts is chosen.

[11] Aktionsplan occurs more frequently than *aktion* and *plan* but is not split, while the term that should not be split, such as *Freitag (friday)* is broken down into *frei (free)* and *tag (day)*

[12] The translation learnt for *Grund* from original corpus containing the term *Grundrechte* was *reason* and *foundation*, but the parallel corpus with the split German text enabled translations *basic* and *fundamental* for *Grund*.

different parts in translation pairs, further refined based on the frequency distributions of candidate bilingual stems and suffixes.

6.2.2.3 Simultaneous Morphology Learning - A Multi-lingual Approach - [SB08]

Snyder and Barzilay [SB08] proposed simultaneous morphology learning from multiple languages for the discovery of cross-lingual morpheme patterns, also termed as abstract morphemes. A multi-lingual corpus of short parallel phrases serves as knowledge base for automatic segmentation and morpheme alignment. For instance, given the parallel phrase *'in my land'* in multiple languages such as English, Arabic, Hebrew and Aramaic, the non-parametric Bayesian model used jointly induces morpheme segmentation and alignment of the form: *in* ⇔ *fy* ⇔ *b* ⇔ *b*, *land* ⇔ *ard* ⇔ *ars* ⇔ *ar* and *my* ⇔ *y* ⇔ *y* ⇔ *y*. Similarities in form representing cognates (the word '*land* in Arabic, Hebrew and Aramaic is derived from a common ancestor), identical suffixes (*my* ⇔ *y*) are key elements that guide simultaneous morpheme alignment, further facilitating the model by constraining the space of joint segmentation. Probabilistic dependencies across languages as well as morpheme distributions within each language are modelled using a hierarchical Bayesian model. While the segmentation model relies on stable recurring string patterns within words as representatives of morphemes, it induces single *'abstract morpheme'* by joining the frequently occurring bilingual morpheme pairs, in addition to the induction of independent morpheme patterns for each language. The model works well in inducing morphological segmentation by exploiting cross-lingual patterns. The underlying benefit is that morphological structure ambiguous in one language is explicitly marked in another language.

The *bilingual morph-like units* referred to in my study roughly corresponds to the abstract morphemes discussed in Snyder and Barzilay's [SB08] work. Induction of bilingual morphemes in my work is based on *pairing bilingual pairs or translations*, while the segmentation model proposed by Snyder and Barzilay is guided by pairing words in two or multiple languages, where a pair represents a translation. Furthermore, the induction of bilingual morph-like units is driven by highly frequent bilingual stem and suffix patterns, and is analogous to the preference for high frequency cross-lingual morpheme patterns discussed in their model. However, as the ultimate objective of my work is translation suggestion, clustering is further preferred as a means to enable reliable translation generation.

6.3 Summary

Majority of the existing approaches for morphology based lexical inferences are monolingual. In what concerns translation coverage, having sufficient near translation forms can certainly influence translation generation. In the upcoming chapters, I will elaborate my experiences on the influence of bilingual word forms in word-to-word translation generation task under high and low density dataset scenarios. Specifically, in the Chapter 7, I propose a method suitable for learning from translation lexicon that is adequate in near translation forms (inflected forms) and is language independent. Furthermore, in reflecting its applicability under large and limitations under small data conditions, and in order to attend to the scenarios with limited bilingual data, a minimally-supervised approach is discussed in the Chapter 8.

Chapter 7

Bilingual Morphology Learning using bilingual lexicons with diverse word inflections

In this chapter, I discuss how the known translations in a validated bilingual lexicon could be used to suggest similar but different translation forms, thereby addressing the issue of OOV terms in translation lexicons. As a pre-phase for generating OOV lexicon entries, first and foremost, in Section 7.1, I discuss the bilingual learning approach that identifies bilingual morph-like units consisting of bilingual stems and suffixes. Clustering is crucial to allow safer translation generalisations. Thus, set of bilingual suffixes representing bilingual extensions for a set of bilingual stems, referred to as bilingual suffix classes/clusters needs to be identified. In this context, I experimented with an open source clustering tool kit, CLUTO for identifying bilingual suffix clusters and the approach follows the partition based clustering. Alternatively, I present experiments using the bilingual suffix co-occurrence score for achieving bilingual suffix clustering. The background, approaches and related experimental results are presented in Section 7.2.

As a direct application of the bilingual morph-like units, in the Section 7.3 of this chapter, I present a simple concatenation technique used to generate OOV word-to-word lexicon entries, along with related experiments. In doing so, I elaborate on the use of bilingual suffix classes and the bilingual segments (bilingual stems and suffixes) in generating OOV lexicon entries.

7.1 Learning Bilingual Segments by Bilingual Suffixation

Given as input, a bilingual lexicon of word-to-word translations extracted from the aligned parallel corpora[1], the approach discussed in this section returns all probable bilingual stems and suffixes along with their frequencies as observed in the input. Also, the bilingual suffix replacement rules that allow one translation form to be obtained from the other are identified, by grouping all the bilingual suffixes that associate with the 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 automatic translation suggestion task discussed in this chapter.

7.1.1 Decomposition of Bilingual Pairs

Given a lexicon of bilingual entries (word-to-word translations), I first look for orthographically similar translations. Translations are considered similar if they begin with the same substring[2].

Based on the longest sequence common to pair of similar translations, first, the bilingual stems and the pair of bilingual suffixes attached to it are identified. For example, considering the translation forms *ensuring* ⇔ *assegurando* and *ensured* ⇔ *assegurou*, the bilingual stem obtained is *ensur* ⇔ *assegur* with a pair of bilingual suffixes, (*ing* ⇔ *ando*, *ed* ⇔ *ou*).

To determine their validity, the induced bilingual segments are analysed with respect to their occurrences as bilingual stems and bilingual suffixes. The following conditions are to be satisfied with respect to the bilingual segments:

- Each candidate bilingual stem should attach to at least two unique morphological extensions (pair of bilingual suffixes).

  For example, the bilingual stem (*ensur* ⇔ *assegur*) is considered as a candidate bilingual stem as it attaches to the pair of bilingual suffixes, (*ing* ⇔ *ando*, *e* ⇔ *em*) induced from the translations *ensuring* ⇔ *assegurando* and *ensure* ⇔ *assegurem* with the decomposition as below:

  [*ensur* ⇔ *assegur*] + [*ing* ⇔ *ando*] and [*ensur* ⇔ *assegur*] + [*e* ⇔ *em*],

- Similarly, each pair of bilingual suffixes should have been attached to at least two unique bilingual stems.

[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/

[2] same with respect to the first and the second language, where the minimum substring length is 3 characters

For example, (*ing* ⇔ *ando*, *ed* ⇔ *ou*) is a valid pair of bilingual suffixes, as it attaches to another bilingual stem such as (*declar* ⇔ *declar*) induced from translations *declaring* ⇔ *declarando* and *declared* ⇔ *declarou* with the following decompositions:

[*declar* ⇔ *declar*] + [*ing* ⇔ *ando*] and [*declar* ⇔ *declar*] + [*ed* ⇔ *ou*].

Note that (*ing* ⇔ *ando*, *ed* ⇔ *ou*) attaches to *ensur* ⇔ *assegur* as well.

7.1.2 Filtering

For each of the bilingual stems obtained, all the bilingual suffixes attaching to it are gathered. For example, the candidate bilingual suffixes that associate with the candidate bilingual stems ('ensur', 'assegur') is as follows:
('ensur', 'assegur') : ('e', 'em'), ('ing', 'ando'), ('ed', 'ou').

Each such grouping indicates the bilingual suffix replacement rules that enable one translation form to be obtained using the other. For instance, from the above grouping, it follows that replacing the suffix *'e'* with *'ed'* and the suffix *'em'* with *'ou'* in the bilingual pair *ensure* ⇔ *assegurem*, yields *ensured* ⇔ *assegurou*.

A few among the identified groups are redundant. The bilingual stems and the associated bilingual suffixes listed below exemplify such redundancies.
('ensur', 'assegur') : ('e', 'ar'), ('ed', 'ado'), ('ed', 'ados'), ('ed', 'ada'), ('ed', 'adas'), ('ing', 'ando'), ('es', 'e'), ('es', 'a'), ('e', 'am'), ('e', 'em'), ('ed', 'aram'), ('ed', 'ou').
('ensure', 'assegur') : ('', 'ar'), ('d', 'ado'), ('d', 'ados'), ('d', 'ada'), ('d', 'adas'), ('s', 'e'), ('s', 'a'), ('', 'am'), ('', 'em'), ('d', 'aram')
('ensur', 'assegura') : ('e', 'r'), ('ed', 'do'), ('ed', 'dos'), ('ed', 'da'), ('ed', 'das'), ('ing', 'ndo'), ('es', ''), ('e', 'm'), ('ed', 'ram')
('ensure', 'assegura') : ('', 'r'), ('d', 'do'), ('d', 'dos'), ('d', 'da'), ('d', 'das'), ('s', ''), ('', 'm'), ('d', 'ram')

The redundant groups where the bilingual stems vary by single character in the boundary are discarded, while retaining the bilingual stems that allow higher number of transformations. This is done by counting the number of unique translations in the lexicon that begin with each of the bilingual stems (see Table 7.1). To handle multiple such instances shorter bilingual stems are preferred over longer pairs. In the examples listed above, the first group is retained.

Table 7.1: Occurrence frequencies of induced bilingual stems with respect to the translations in the bilingual lexicon

Stem pair	Frequency	Stem pair	Frequency
'accord', 'acord'	3	'abandon', 'abandon'	17
'accord', 'acorde'	2	'abandon', 'abandona'	2

7.2 Identification of Bilingual Suffix Clusters

A set of bilingual stems that share same bilingual suffix transformations form a cluster. The main intention behind clustering is to generalise the bilingual suffix replacement rules, by looking for other stem pairs that go through the same transformation. Two approaches were experimented: Partition approach and Bilingual Suffix Co-occurrence Score based approach.

7.2.1 Clustering by Partition Approach

The bilingual stems and the suffix pairs identified after the filtering phase are clustered using the clustering tool, CLUTO[3]. The toolkit provides three different classes of clustering algorithms such as, partition, agglomerative and graph-partitioning, to enable the clustering of low and high dimensional data sets. The partition approach for clustering was adapted in my experiments. The partition clustering is driven by total of seven different criterion functions [ZK02]. Each bilingual stem is characterised by its associated bilingual suffixes.

7.2.2 Experimental Setup - Partition Approach

7.2.2.1 Datasets and Pre-processing

The resources generated as output from the filtering phase include the list of bilingual stems, bilingual suffixes and bilingual suffixes grouped by bilingual stems, which constitute the input data for the clustering. Table 7.2 provides an overview of the unique bilingual segments identified from the different training data sets.

Table 7.2: Statistics of unique bilingual segments identified with different training sets

Training Data	Unique Bilingual Stems	Unique Bilingual Suffixes
35,891	6,644	224
209,739	24,223	232

A selected list of the frequent bilingual suffixes identified from the existing lexicon entries for EN-PT is shown in Table 7.3. Table 7.4 illustrates the suffix pairs associated with the automatically identified bilingual stems *'answer'* ⇔ *'respond'* and *'encourag'* ⇔ *'estimul'*, where the first group represents ('', 'er') group suffixes, while the second shows ('e', 'ar') group suffixes.

To prepare the data for clustering, the *doc2mat*[4] tool is used, which provides the necessary conversion of data into matrix form that is compatible with CLUTO's clustering algorithms.

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

[4] http://glaros.dtc.umn.edu/gkhome/files/fs/sw/cluto/doc2mat.html

Table 7.3: Highly frequent Bilingual Suffixes identified for EN-PT bilingual bases with different training sets

Training Set$_1$		**Training Set$_2$**	
Suffix Pair	**freq$_{lexicon}$**	**Suffix Pair**	**freq$_{lexicon}$**
(‘’, ‘o’)	4,644	(‘’, ‘o’)	15,006
(‘’, ‘a’)	2,866	(‘’, ‘a’)	9,887
(‘e’, ‘o’)	1,685	(‘’, ‘as’)	5,840
(‘’, ‘os’)	1,362	(‘’, ‘os’)	5,697
(‘’, ‘as’)	1,339	(‘ed’, ‘ado’)	4,760
(‘e’, ‘a’)	1,297	(‘ed’, ‘ados’)	4,221
(‘ed’, ‘ado’)	1,001	(‘ed’, ‘ada’)	4,193
(‘ed’, ‘ada’)	868	(‘e’, ‘o’)	4,159
(‘ed’, ‘ados’)	814	(‘ed’, ‘adas’)	4,051
(‘ation’, ‘ação’)	658	(‘e’, ‘a’)	3,158

Table 7.4: Bilingual suffixes grouped by bilingual stems for EN-PT

Suffix pairs	**Stem pairs**
(”, er), (”, erem), (”, am), (”, em), (s, e), (s, a), (ed, ida), (ed, idas), (ed, ido), (ed, idos), (ed, eram), (ed, eu), (ing, endo), (ing, er)	answer ⇔ respond
(e, ar), (e, arem), (e, am), (e, em), (es, e), (es, a), (ed, ada), (ed, adas), (ed, ado), (ed, ados), (ed, aram), (ed, ou), (ing, ando), (ing, ar)	encourag ⇔ estimul

Experiments were carried out with 10, 15, 20, 50 and 100 way clustering and the best results were obtained with 50 clusters. The clustering results were further analysed manually to remove outliers (bilingual suffixes) from each cluster and to identify the sub-clusters from among the clustered results.

7.2.3 Clustering Results - Partition Approach

Presented below are a few randomly chosen clusters along with the discriminating features (bilingual suffixes) and a few example bilingual stems under each of the verb, noun and adjective classes.

Verb-ar Cluster: (‘e’, ‘ar’), (‘e’, ‘arem’), (‘e’, ‘am’), (‘e’, ‘em’), (‘es’, ‘e’), (‘es’, ‘a’), (‘ed’, ‘ada’), (‘ed’, ‘adas’), (‘ed’, ‘ado’), (‘ed’, ‘ados’), (‘ed’, ‘aram’), (‘ed’, ‘ou’), (‘ing’, ‘ando’), (‘ing’, ‘ar’)

Example Bilingual Stems: *toggl* ⇔ *comut*, *argu* ⇔ *afirm*, *shuffl* ⇔ *baralh*

Verb-er Cluster: (‘’, ‘er’), (‘’, ‘erem’), (‘’, ‘am’), (‘’, ‘em’), (‘s’, ‘e’), (‘s’, ‘a’), (‘ed’, ‘ida’), (‘ed’, ‘idas’), (‘ed’, ‘ido’), (‘ed’, ‘idos’), (‘ed’, ‘eram’), (‘ed’, ‘eu’), (‘ing’, ‘endo’), (‘ing’, ‘er’)

Example Bilingual Stems: *spend* ⇔ *dispend*, *reply* ⇔ *respond*, *answer* ⇔ *respond*

Verb-ir Cluster: (‘’, ‘ir’), (‘’, ‘irem’), (‘’, ‘am’), (‘’, ‘em’), (‘s’, ‘e’), (‘s’, ‘a’),(‘ed’, ‘ida’), (‘ed’, ‘idas’), (‘ed’, ‘ido’), (‘ed’, ‘idos’), (‘ed’, ‘iram’), (‘ed’, ‘iu’), (‘ing’, ‘indo’), (‘ing’, ‘ir’)
Example Bilingual Stems: *expand* ⇔ *expand*, *acclaim* ⇔ *aplaud*, *reopen* ⇔ *reabr*
Adjective-al Cluster: (’al’, ’ais’), (’al’, ’al’)
Example Bilingual Stems: *accident* ⇔ *acident*, *cervic* ⇔ *cervic*, *environment* ⇔ *ambient*
Noun-ence Cluster: (‘ence’, ‘ência’), (‘ences’, ‘ências’)
Example Bilingual Stems: compet ⇔ *compet*, *recurr* ⇔ *reocorr*, *transfer* ⇔ *transfer*
Noun-ist Cluster: (‘ist’, ‘ista’), (‘ists’, ‘istas’)
Example Bilingual Stems: *journal* ⇔ *colun*, *baloon* ⇔ *ascension*

Table 7.5 shows the classification statistics for the lexicon entries under each of the randomly chosen clusters representing verbs, nouns and adjectives.

Table 7.5: Clustering statistics for bilingual entries with randomly selected classes

Cluster	**# of bilingual stems**
Verb-*(‘’,‘ar’) + (‘e’,‘ar’)*	689 + 570
Verb-*(‘’,‘er’) + (‘e’,‘er’)*	77+ 82
Verb-*(‘’,‘ir’) + (‘e’, ‘ir’)*	175 + 109
Adjective-*‘ent’*	174
Adjective-*‘al’*	568
Noun-*‘ence’*	65
Noun-*‘ment’*	181

7.2.4 Clustering by Bilingual Suffix Co-occurrence Score based Approach

The approach discussed in afore-mentioned sections (Section 7.2.1 through Section 7.2.3), identifies clusters of bilingual stems characterised by suffix pairs (features) using the partition based clustering technique provided in the clustering tool, CLUTO[5]. The clustering is based on partition approach and requires that the number of partitions are explicitly specified before clustering [Kav+14a].

In the current section, I focus on clustering word-to-word translations, using the bilingual suffix co-occurrence score, a measure representing the number of times a bilingual suffix co-occurs with another bilingual suffix in the input lexicon. The measure bilingual suffix co-occurrence is discussed with examples in Section 7.2.4.2. The degree of co-occurrence between two bilingual morphological extensions (bilingual suffixes) with reference to common bilingual stems determine if each of them should fall in the same cluster. As the bilingual suffix-pair based co-occurrence statistics is used in clustering the bilingual translations, the number of partitions need not be anticipated in advance. Experiments are presented for two language pairs EN-HI and EN-PT.

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

7.2.4.1 Bilingual Segments as Input Resources

In the Section 7.1, the bilingual approach for learning morph-like units was introduced as a fundamental step for suggesting new translations. In this section, I re-illustrate the approach and the output resources learnt for the language pair EN-HI. The approach involves identification and extraction of orthographically and semantically similar bilingual segments, as for instance, *'good'* ⇔ *'acCh'*, occurring in known translation examples (here, EN-HI), such as, *'good'* ⇔ *'acChA'*, *'good'* ⇔ *'acChe'* and *'good'* ⇔ *'acChI'* together with their bilingual extensions constituting dissimilar bilingual segments (bilingual suffixes), *''* ⇔ *'A'* | *'e'* | *'I'*[6]. The common part of translations that conflates all its bilingual variants[7] represents a bilingual stem (*'good'* ⇔ *'acCh'*). The different parts of the translations contributing to various surface forms represent bilingual suffixes (*''* ⇔ *'A'* | *'e'* | *'I'*). A pair of such extensions represent bilingual suffix replacement rules. Further, set of bilingual suffixes representing bilingual extensions for a set of bilingual stems together form bilingual suffix clusters[8], hence allowing safer translation generalisations.

As evident from the previous section on bilingual learning, by applying the bilingual approach to learning bilingual morph-units on a bilingual lexicon [Kav+14a], the resources listed below are produced as output.

List of Bilingual stem and suffix pairs: This represents the list of bilingual stems and suffixes with their observed frequencies in the training dataset. For instance, in the Table 7.7, *'plant'* ⇔ *'paudh'* is a bilingual stem conflating two different surface translation forms *'plant'* ⇔ *'paudhA'* and *'plants'* ⇔ *'paudhoM'*, while ('nation', 'rAShTr') is a bilingual stem that attaches to four different bilingual pairs ('national', 'rAShTrIya'), ('nationalism', 'rAShTrIyatA'), ('nationality', 'rAShTrIyatA') and ('nationalist', 'rAShTrIyatAvAdI') (2^{nd} line in each row shows transliterations for HI terms). *('', 'I')*, *('', 'A')*, *('s', 'oM')* and so forth in the Table 7.6 are automatically identified bilingual suffixes which are attached to 10,743, 29,529 and 226 different bilingual pairs respectively. These lists aid in identifying the bilingual stems and bilingual suffixes when a new translation is given, for which all possible inflected forms should be suggested.

Bilingual suffixes grouped by bilingual stems: This represents which set of bilingual suffixes attach to which bilingual stem. In the Table 7.7 for instance, *('', 'A')*, *('s', 'oM')* are bilingual suffixes which attach to the same bilingual stem *'plant'* ⇔ *'paudh'* contributing to the surface forms *'plant'* ⇔ *'paudhA'* and *'plants'* ⇔ *'paudhoM'*.

Each such grouping indicates the bilingual suffix replacement rules that enable one translation form to be obtained using the other. For instance, from the above grouping, it follows that replacing the null suffix *''* with *'s'* and the suffix *'A'* with *'oM'* in the bilingual pair *'plant'* ⇔ *'paudhA'*, the bilingual pair *'plants'* ⇔ *'paudhoM'* can be obtained. In other words, in the bilingual pair *'plant'* ⇔ *'paudhA'*, appending *'s'* at the end of the EN word

[6] Note the null suffix in EN corresponding to gender and number suffixes in HI.

[7] Translations that are lexically similar.

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

Table 7.6: Bilingual Suffixes undergoing frequent replacements in EN-HI

Bilingual Suffixes	Bilingual Suffixes (Hindi Suffixes transliterated)	Frequency
('', '◌ी')	(‘’, ‘I’)	10743
('', '◌ा')	(‘’, ‘A’)	29529
('ion', '◌ा')	(‘ion’, ‘A’)	457
('er', '◌ा')	('er', ‘A')	428
('ity', '◌ा')	('ity', ‘A')	286
('s', '◌ो◌ं')	('s', ‘oM')	226
('ity', 'ता')	('ity', 'tA')	223

Table 7.7: Bilingual suffixes grouped by bilingual stems for EN-HI

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

‘plant’ and replacing the suffix *‘A’* with *‘oM’* in its translated form *‘paudhA’* (in HI), yields the bilingual pair *‘plants’* ⇔ *‘paudhoM’*.

The resources just mentioned are fundamental for the clustering experiments based on bilingual suffix co-occurrence score, that I discuss in the subsections below. A cluster is typically made of bilingual suffixes that attach to a set of bilingual stems. After all the bilingual suffixes that attach to a bilingual stem have been grouped as mentioned earlier (see Table 7.7 for EN-HI examples), all bilingual stem pairs sharing same set of bilingual suffixes (and hence undergoing similar transformations) are further grouped forming a cluster. Clusters can thus be obtained by grouping the bilingual stems sharing identical bilingual suffix replacement rules (as discussed in the Subsection 7.2.1).

An alternative is to group the bilingual suffixes based on the frequency of their association with common bilingual stems and identify all stem pairs sharing those bilingual suffixes. Below, I discuss the use of Bilingual Suffix Co-occurrence score in learning bilingual suffix clusters.

7.2.4.2 Bilingual Suffix Co-occurrence Score

Bilingual suffix co-occurrence score represents the number of times a bilingual suffix $(s_{i_{L1}}, s_{i_{L2}})$ has co-occurred with another bilingual suffix $(s_{j_{L1}}, s_{j_{L2}})$ in the bilingual lexicon. Two bilingual suffixes are said to co-occur if they attach to a common bilingual stem.

The co-occurrence scores between different bilingual suffixes in EN-HI are shown in the Table 7.8; the co-occurrence score between the bilingual suffixes *('', 'A')* and *('s', 'oM')* is 27 implying that they co-occur with 27 distinct bilingual stems. Similarly, in the Table 7.9, for the language pair EN-PT, the co-occurrence score between *('ence', 'ência')* and *('ences', 'ências')* is 65.

Based on the notion that the bilingual suffixes that co-occur more frequently are likely to be good candidates for a cluster, the candidate bilingual suffixes are determined for each cluster. The bilingual suffix co-occurrence score between two bilingual suffixes is required to be above the set threshold, if they should belong to the same cluster. For EN-HI, the threshold set at 3 yielded better results while for EN-PT, slightly higher threshold set at 5 improved the generation performance. The Algorithm 1 shows the steps involved in clustering.

Definitions Let L be a Bilingual Lexicon.
Let L1, L2 be languages with alphabet set Σ_1, Σ_2.
Let $S_{StemPair}$ represent the set of bilingual stems and $S_{SuffixPair}$ be the set of bilingual suffixes.

In the algorithm, *'Suffix-Class-String'* represents a string consisting of all bilingual suffixes ($(s_{i_{L1}}, s_{i_{L2}})$, 1 ≤ i ≤ n, separated by commas preceding the dot in Step 18 of Algorithm 1) along with the bilingual stem ($(p_{i_{L1}} p_{i_{L2}})$ following the dot in Step 18 Algorithm 1) to which those suffixes attach. Each row in Table 7.7 may thus be interpreted in the above specified form as follows: ('', 'A'), ('er', 'k'), ('ers', 'koM') . ('test', 'parIksh'). Then, $S_{Suffix-Class}$ simply represents set of such strings.

'Merged-Suffix-Class-string' represents a *'Suffix-Class-String'* consisting of a set of grouped bilingual suffixes along with all the bilingual stems sharing those suffixes. An example for the Merged-Suffix-Class-string is, ('', 'A'), ('er', 'k'), ('ers', 'koM') . ('test', 'parIksh').('print', 'mudr'). Here, ('print', 'mudr') is another bilingual stem that shares the same transformations ('', 'A'), ('er', 'k'), ('ers', 'koM') as the bilingual stem ('test', 'parIksh').

7.2.5 Experimental Setup - Co-occurrence based Clustering Approach

7.2.5.1 Data set

Bilingual pairs taken from EN-HI bilingual lexicon representing single-word translations form 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 (SCP) based statistical measure [DSL99] from the aligned parallel corpora[10]. The manually extracted translation pairs

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

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

Algorithm 1 Learning Bilingual Suffix Clusters

```
procedure LEARN–BILINGUALSUFFIXCLUSTER
    for each input bilingual pair $(a_{L1}, a_{L2}) \in L$, where,
        $(a_{L1}= p_{1_{L1}}+s_{1_{L1}})$, $(a_{L2}= p_{1_{L2}}+s_{1_{L2}})$, and
        $(p_{1_{L1}} p_{1_{L2}}) \epsilon S_{StemPair}$ and $(s_{1_{L1}}, s_{1_{L2}}) \epsilon S_{SuffixPair}$ do
            Set Suffix-Class-String= $(s_{1_{L1}}, s_{1_{L2}})$

        for every bilingual pair, $(b_{L1}, b_{L2}) \in L$, such that,
            $b_{L1}= p_{1_{L1}}+s_{2_{L1}}$ and $b_{L2}= p_{1_{L2}}+s_{2_{L2}}$, where,
            $(p_{1_{L1}} p_{1_{L2}}) \epsilon S_{StemPair}$ and $(s_{2_{L1}}, s_{2_{L2}}) \epsilon S_{SuffixPair}$ do

            if Co-occurence-Score$((s_{1_{L1}}, s_{1_{L2}}),(s_{2_{L1}}, s_{2_{L2}})) \geq$ threshold then
                Set Suffix-Class-String=Suffix-Class-String.$(s_{2_{L1}}, s_{2_{L2}})$
                Add Suffix-Class-String to $S_{Suffix-Class}$, bilingual suffix set
            end if
        end for
    end for
    for each Suffix-Class-String $S_1 \in S_{Suffix-Class}$, where,
        $S_1=((s_{1_{L1}}, s_{1_{L2}}), (s_{2_{L1}}, s_{2_{L2}}), ..., (s_{n_{L1}}, s_{n_{L2}})).(p_{1_{L1}}, p_{1_{L2}})$, and
        $((s_{1_{L1}}, s_{1_{L2}}), (s_{2_{L1}}, s_{2_{L2}}), ..., (s_{n_{L1}}, s_{n_{L2}})) \epsilon S_{SuffixPair}$,
        $(p_{1_{L1}} p_{1_{L2}}) \epsilon S_{StemPair}$ do
            Set Merged-Suffix-Class-string = $((s_{1_{L1}}, s_{1_{L2}}), (s_{2_{L1}}, s_{2_{L2}}), ................,$
                $(s_{n_{L1}}, s_{n_{L2}})).(p_{1_{L1}}.p_{1_{L2}})$

        if $\exists$ suffix class string $S_2 \in S_{Suffix-Class}$, such that,
            $S_2=((s_{1_{L1}}, s_{1_{L2}}), (s_{2_{L1}}, s_{2_{L2}}), ..., (s_{n_{L1}}, s_{n_{L2}})).(p_{2_{L1}} p_{2_{L2}})$,
            and $(p_{2_{L1}} p_{2_{L2}}) \epsilon S_{StemPair}$ then
                Merged-Suffix-Class-string=Merged-Suffix-Class-string.$(p_{2_{L1}} p_{2_{L2}})$
                Add Merged-Suffix-Class-string to $S_{Cluster}$,
                 the set of Bilingual Suffix clusters.
        end if
    end for
end procedure
```

were extracted from the Bible parallel corpora using a tool that worked on the aligner by Tiago et al. [IL05].

For EN-PT, the lexicon of bilingual entries (word-to-word translations) was extracted from the aligned parallel corpora[11] using various extraction techniques [Air+09; Bro+93; GL11; LL09]. In the experiments discussed, I used 52K samples as training data for EN-HI and 210K samples for EN-PT. Input to the clustering constitutes of the three resources: the bilingual stem list, the bilingual suffix list and the bilingual suffix groups. For EN-HI, these resources are learnt using the approach that will be discussed in the forthcoming chapter.

[11] 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/

7.2.6 Clustering Results - Co-occurrence based Clustering

A total of 143 clusters for EN-HI and 63 clusters for EN-PT were identified. For both EN-PT and EN-HI, the smallest cluster consisted of only one suffix replacement rule, i.e, a pair of bilingual suffixes. For EN-PT, the largest cluster representing the *('', 'er')* group consisted of 15 different bilingual suffixes that are attached to 717 different bilingual stems. For EN-HI, the largest cluster comprised of 5 different bilingual suffixes. Tables 7.8 and 7.9 respectively show randomly chosen clusters (with partial entries of bilingual suffixes for EN-PT[12]) for each of the language pairs EN-HI and EN-PT.

Table 7.8: Suffix co-occurrence scores for bilingual suffix pairs representing highly frequent bilingual suffix replacement rules

Bilingual Suffixes	Bilingual Suffix 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')
('', '◌ा'), ('er', 'क') ('', 'A'), ('er', 'k')	32	('test', 'परीक्ष') ('test', 'parIksh')	('print', 'मुद्र') ('print', 'mudr')
('', '◌ा'), ('s', '◌े') ('', 'A'), ('s', 'e')	10	('month', 'महीन') ('month', 'mahIn')	('curtain', 'पर्द') ('curtain', 'pard')

Table 7.9: Translation patterns representing bilingual suffix classes and the bilingual suffix co-occurrence scores for EN-PT

Bilingual Suffixes	Bilingual Suffix Co-occurrence Score	Bilingual Stem Instances	
(ence, ência), (ences, ências)	65	(prefer, prefer)	(recurr, ocorr)
(al, ais), (al, al)	568	(compartment, compartiment)	(department, departament)
(e, er), (’ ’, ir)	0	-	-
(ed, ida), (ed, idas)	75	(acclaim, aplaud)	(dismiss, demit)
(ed, ada), (ed, adas)	318	(affirm, confirm)	(adjust, ajust)

7.3 Using Bilingual Morph-units and Bilingual Suffix Clusters in Generating OOV Lexicon Entries - Concatenative Approach

In sections 7.1 through 7.2, approaches for learning bilingual segments and suffix classes as a pre-phase to translation generation were discussed. As a direct application of these

[12] In the Table 7.9, only two bilingual suffixes are shown per cluster although the original clusters contains varying number of bilingual suffixes ranging from 2 to 15 for EN-PT and from 2 to 5 for EN-HI

bilingual morph-like units, in the current section, I present a simple concatenation technique used to generate OOV word-to-word lexicon entries and the related experimental results.

7.3.1 Completing the Lexicon for Missing Forms

The resources generated out of the learning phase includes known list of bilingual stems, bilingual suffixes along with their observed frequencies in the training data set. Further, information about which set of bilingual suffixes attach to which set of bilingual stems is also known. The underlying approach for suggesting new translations relies on the clusters formed using the bilingual stems, suffixes and their groupings identified during the learning phase. New translations are generated by direct concatenation of stem pair and suffix pair belonging to the same cluster. Thus, the approach is said to complete the lexicon for missing forms by generating missing word-to-word translation forms that are similar to those translations existing in the lexicon. These newly generated pairs, upon validation, further serve as additional training data for the subsequent iterations.

7.3.2 Generation results and Evaluation

Table 7.10 provides an overview of the data sets used in training and the associated generation statistics. The bilingual translation lexicon used in this study is acquired from an aligned parallel corpora[13] [GL09] using various extraction methods [Air+09], [Bro+93], [LL09], [GL11].

Table 7.10: Overview of the generation results for EN-PT with different training sets

Training Data	Unique Bilingual Stems	Unique Bilingual Suffixes	Generated Pairs	Correct Generations	Incorrect Generations
35,891	6,644	224	4,279	3,862	306
209,739	24,223	232	14,530	2,283/2,334	20/2,334

With a training data of approximately 209K bilingual pairs, about 15K new translations were generated. Among the 2,334 validated entries, 2283 were accepted, 27 were inadequate (*accept-*) indicating incomplete/inadequate translations and 20 were rejected (*reject*). Table 7.11 shows the statistics for the generated translations (correct, accepted) in the parallel corpora, where the co-occurrence frequency is less than 10. Among the generated entries, 9034 bilingual pairs did not occur in the parallel corpora. When both the bilingual stem and the bilingual suffix in the bilingual pair to be analysed are known, 90% of the generated translations were correct, with the first data set.

[13] 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/

Table 7.11: Co-occurrence frequency for the generated translations in the parallel corpora

Co-occurrence Frequency	# of generated bilingual pairs	Co-occurrence Frequency	# of generated bilingual pairs
9	45	4	148
8	62	3	207
7	64	2	324
6	80	1	489
5	102	-	-

The evaluation results for clustering based on the applicability of induced segments and clusters in generating new translations is shown in the column 4 of Table 7.12.

The precision for generated translations is calculated as the fraction of correctly generated bilingual pairs to total number of bilingual pairs generated. In completing the translation lexicon for missing forms, where both bilingual stems and bilingual suffixes are known, the precision achieved for translation generation reaches 84.02% when compared to the precision of 81.31% obtained using the Kavitha's et al. approach [Kav+14a] for EN-HI and 88% for EN-PT, showing a drop by 2% in precision as opposed to the partition approach [Kav+14a].

Table 7.12: Clustering Statistics for EN-HI and EN-PT language pairs using Partition and Co-occurrence Score Approach

Language Pairs	Clustering Approach	Number of Clusters	Generation Precision
EN-HI	Partition	224	0.81
	Bilingual Suffix Co-occurrence Score	143	0.84
EN-PT	Partition	50	0.90
	Bilingual Suffix Co-occurrence Score	63	0.88

7.3.3 Error Analysis

Analysing the newly generated translations, it is observed that certain translations are incomplete (examples labelled as *accept-*), and some are incorrect (examples labelled as *reject*). Translation candidates such as '*intend* ⇔ *pretendem*' are inadequate as the correct translations require '*intend*' to be followed by '*to*'. Similarly, '*include*' should be translated either by '*contam-se*' or '*se contam*' and so '*include* ⇔ *contam*' is classified as *accept-*, meaning that the translations in PT are shorter than necessary. It is incorrect, but this classification may enable other kind of learning.

Other generated entries, such as, '*collector* ⇔ *coleccionadores*', '*advisor* ⇔ *consultores*', '*rector* ⇔ *reitores*', '*elector* ⇔ *eleitores*' are instances wherein the noun acts as an adjective that is translated either by adding '*de*' before the plural noun translation in PT, eventually with an article after *'de'* as in *'de os'*. Again, the bilingual pair generated, *'wholesales ⇔ grossistas'* misses the noun *'vendas'* as in *'vendas grossistas'*. The English noun is compounded in this case.

Generation errors labelled as *'reject'* in Table 7.13, are a consequence of incorrect generalisations. Verbs in PT ending in *'uir'* form past participle forms adding *'iu'*, as for instance with the word forms *'construir'* and *'construiu. 'wants'* is an irregular verb that is translated either by *'quer'* or *'queira'* or *'quizer'*.

Table 7.13: Generated Translations for EN-PT

Accept-	**Reject**
languages ⇔ linguísticas	rights ⇔ adequados
instructor ⇔ instrutores	replaced ⇔ substituida / -idas / -idos / -ido
ambassador ⇔ embaixadores	several ⇔ vário
include ⇔ contam	wants ⇔ quere
emerged ⇔ resultados	electrical ⇔ electrica

7.4 Summary

In this Chapter, I have presented an approach for identifying bilingual segments consisting of stem pairs and suffix pairs for translation generation. The stem pairs represent bilingual morph-units that conflate various inflected forms, while the suffix pairs represent bilingual morphological extensions of the identified stem pairs. A pair of such bilingual morphological extensions represent transformation rules, enabling one inflected form of a translation to be obtained from another.

In order to generalise the transformation rules, two clustering approaches have been discussed. In the first of these, the partition approach provided in the CLUTO toolkit is used to identify clusters of bilingual stems and suffixes. However, the downside of the partition approach is the need to pre-suppose the number of partitions into which the input data has to be clustered. As an alternative, I have discussed the use of co-occurrence score between two bilingual suffixes as a means to determine if two bilingual suffixes should fall in the same cluster [Kav+15c]. In clustering the bilingual translations, this enables bilingual suffixes to be grouped without having to suppose the number of clusters prior to clustering. Further, evaluation based on generating new missing translations suggest precision closer to that achieved using the partition based approach [Kav+14a]. Later in the chapter, I have discussed the simple concatenation scheme for translation suggestion. Experimental results for EN-PT show that when both the bilingual stem and the bilingual suffix in the bilingual pair to be analysed are known, approximately 90% of the generated translations are correct. However, this attainment is seen to drop for language pairs with fewer inflected forms and with smaller lexicons.

The approach discussed being purely bilingual suffixation based, does not handle irregular forms and does not capture stem changes prior to suffixation. The bilingual learning approach works well for training data having sufficient near translation forms, but performs poorly under conditions of limited near translation forms (inflected forms).

The motivation for this study is the fact that extraction techniques cannot handle what is not in a parallel corpora and they cannot extract everything. Above all, the way in which translations are extracted and evaluated does not guarantee that most of the possible translation pairs not found in parallel corpora might be automatically suggested for a translation engine or as bilingual entries.

Chapter 8

Bilingual Morphology Learning using highly defective bilingual lexicons with limited inflection diversity

The bilingual learning approach presented in Chapter 7 particularly works well for training data having sufficient near translation forms. In a different scenario, we have the EN-HI bilingual lexicon that has very few inflectional variants for the words and their inflectionally modified translations. Majority of the entries are from a normal dictionary designed for human translation that has no information about how words inflect in both languages. Further, the number of bilingual pairs that serve as knowledge base (for the learning) is limited. In this setting, dealing with segmentation ambiguity is difficult. Also, handling of previously unseen bilingual pairs is a challenge. In this chapter, a minimally supervised approach is discussed to enable learning from lexicons with limited inflection forms covering different genders for adjectives, different numbers for nouns and different information for verbal inflection. Nevertheless, except for the limitations in their applicability with respect to varying data set sizes and the level of supervision involved, the approaches are similar in that both are purely based on bilingual suffixation.

Further, the generation technique discussed in the previous chapter relies on the concatenation of bilingual morph-units (bilingual stems and suffixes) belonging to one of the bilingual suffix classes identified using the learning scheme. Alternatively, in this chapter it is shown that, as new translations are added to the lexicon all possible inflection forms can be suggested, provided we are first able to predict for the newly added translation, the bilingual stem and bilingual suffix by segmentation and subsequently the bilingual suffix class to which that translation belongs. Upon segmentation, generation of new translations can be done by following one of the steps below:

- If both the bilingual stems and the bilingual suffixes are known[1], check whether they belong to the same cluster. If so, each of the new translations are suggested by concatenating the stem pair with the remaining associated suffix pairs in the identified cluster.

- If only the bilingual suffix is known,

    1. determine the bilingual suffix class to which it belongs.
    2. generate new translations by replacing the identified bilingual suffix with other bilingual suffixes that have been observed to co-occur with the identified bilingual suffix.

However, without sufficient (in the worst case, without any) near forms in the training data set, predicting the segmentation boundary and hence the bilingual morph-units can be difficult. In this regard, from the perspective of learning near translation forms under the said context, I propose the minimally supervised approach for learning bilingual segments and the multi-label classification scheme for determination of bilingual suffix classes.

8.1 Introduction and Background

In the previous chapter, morphological splits for bilingual pairs or translations were derived by pairing orthographically similar translations and extracting parts sharing longest common stems and different suffixes. When the lexicon adopted for bilingual learning is relatively small, further limited by very few near translation forms (inflection forms), the segmentation precision drops substantially. Owing to the fact that the automatically acquired translation lexicon is relatively small for EN-HI with very few near translation forms, to boost the learning process I additionally rely upon a translation dictionary.

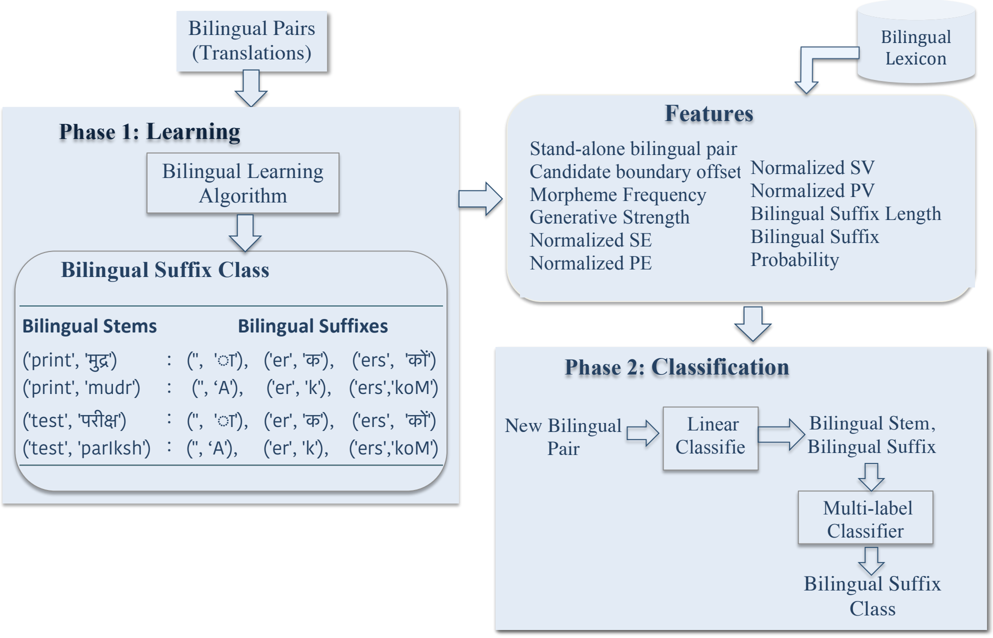
The translation dictionary, however, was neither specifically meant for machine translation nor bilingual morphology learning. As a matter of fact, this *specific parallel corpora of entries of translation dictionary* (EN-HI), is simply a human usable dictionary, where for instance, it is known that '*good*' translates as '*acChA*' but, no information exists at the dictionary level that there are other similar forms that are also translations of '*good*' namely '*acChI*' (singular feminine form), '*acChe*' (plural masculine/feminine form) apart from other meanings. In other words, I use a translation dictionary where there is no information on how word forms translate each other. By using parallel corpora, I assume that there is no need to have texts semantically tagged, as, the translation functions as semantic tags in other languages. Having the parallel texts (translations of each other), we have a quite natural way of semantical tagging, the translation proper, that was not done with the objective of semantically tagging. The tags do not exist in any ontology because they were quite natural translations of the content.

[1] known to have occurred in the training dataset

8.2 Minimally Supervised Bilingual Learning Approach - An Overview

The proposed segmentation strategy operates in 2 stages: the *learning phase* for identifying bilingual stems, suffixes and suffix classes that partially serves as the training data (an adaptation of the approach discussed in Chapter 7) and the *classification phase* for deciding segmentation (Section 8.2.2). Figure 8.1 depicts the proposed framework.

Figure 8.1: Architecture of Supervised Segmentation and Suffix Class Determination


Bilingual Pairs (Translations)
Bilingual Lexicon
Phase 1: Learning
Bilingual Learning Algorithm
Bilingual Suffix Class
Bilingual Stems
Bilingual Suffixes
('print', 'मुद्र') : ('', 'ा'), ('er', 'क'), ('ers', 'कों')
('print', 'mudr') : ('', 'A'), ('er', 'k'), ('ers','koM')
('test', 'परीक्ष') : ('', 'ा'), ('er', 'क'), ('ers', 'कों')
('test', 'parIksh') : ('', 'A'), ('er', 'k'), ('ers','koM')
Features
Stand-alone bilingual pair
Candidate boundary offset
Morpheme Frequency
Generative Strength
Normalized SE
Normalized PE
Normalized SV
Normalized PV
Bilingual Suffix Length
Bilingual Suffix Probability
Phase 2: Classification
New Bilingual Pair
Linear Classifie
Bilingual Stem, Bilingual Suffix
Multi-label Classifier
Bilingual Suffix Class

Fundamental to the segmentation and generation strategy is the learning phase, intended to prepare the partial training data needed in deciding upon the subsequent segmentation for any unseen bilingual pair. In this regard, in Section 8.2.1, I discuss the modified bilingual approach for learning morph-like units [Kav+14a], used in preparing the partial training data.

As was elaborated in the Chapter 7, the approach involves identification and extraction of orthographically and semantically similar bilingual segments (as for instance, *'good'* ⇔ *'acCh'*) occurring in known translation examples (*'good'* ⇔ *'acChA'*, *'good'* ⇔ *'acChI'* and *'good'* ⇔ *''acChe'*), together with their bilingual extensions constituting dissimilar bilingual segments (bilingual suffixes) (*''* ⇔ *'A'* | *'e'* | *'I'*)[2]. However, the learning approach discussed in the current section slightly differs from the previous approach (Chapter 7) [Kav+14a] in that, here, true compound bilingual suffixes (a combination of multiple candidate bilingual suffixes) are retained based on the observation that the strength [DN07] of a compound bilingual suffix is less than the strengths of the bilingual suffixes composing it. Suffix containment check is performed to avoid over-segmentation

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

and involves looking for one candidate bilingual suffix enclosed within another. For example, in translations such as (‘nationalist’, ‘rAShTrIyatAvAdI’), the suffix pair (‘alist’, ‘IyatAvAdI’) is composite made of suffix pairs (‘al’, ‘IyatA’) and (‘ist’, ‘vAdI’). If (‘al’, ‘IyatA’), (‘ist’, ‘vAdI’) and (‘alist’, ‘IyatAvAdI’) ϵ Suffix list and (‘nation’, ‘rAShTr’) ϵ Stem list then, (‘alist’, ‘IyatAvAdI’) is retained provided the strength((‘alist’, ‘IyatAvAdI’)) < strength((‘ist’, ‘vAdI’)) and strength((‘al’, ‘IyatA’)), .

Similar to the earlier discussed approach, 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 morphological extensions for a set of bilingual stems form bilingual suffix classes[4]. The bilingual suffix classes thus learnt along with the bilingual lexicon constitutes the training data set for the classification phase employed in arriving at a segmentation decision for any new bilingual pair. Validating the bilingual resources generated in the learning phase allows in an improved segmentation precision as these serve as clues in deciding the segmentations for new unseen bilingual pairs.

To infer all possible translation forms (inflected forms) for any given (new) bilingual pair, first a valid segmentation boundary needs to be determined. After the determination of segmentation boundary for the given bilingual pair, depending on the bilingual suffix and the stem surfaced, the bilingual pair is subsequently classified into one of the bilingual suffix classes identified in the training phase. While word segmentation measures are used as clues for deciding segmentation of an unseen bilingual pair (discussed in Section 8.2.2), the SVM based multi-label classifier is used in determining the bilingual suffix class, post-segmentation (discussed in Section 8.2.3.1).

8.2.1 The Learning Phase

In the Section 7.1, the approach for learning bilingual suffixation operations by utilising the translation lexicon as a parallel resource was discussed in detail [Kav+14a]. As a prephase 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 Momouchi *et al.* [MT97], the bilingual approach [Kav+14a] allows identification of common (bilingual stems) and different (bilingual suffixes) bilingual segments occurring in translation examples, which are then used in generating new translations, which I adapt here for preparing the partial training data.

[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.

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 [Kav+14a] (refer Algorithm 2). In the approach presented in Section 7.1 [Kav+14a], redundant groups were eliminated by retaining shorter stem pairs and those sharing higher number of transformations. In the current approach, suffix containment check ensures that longer suffixes enclosing other suffixes are retained provided the strength [DN07] of the enclosing suffix is lower than that of the sub suffixes (see steps 11 thru 17 of Algorithm 2.).

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 similarly for language 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 8.1, examples illustrate bilingual variants: *noun_singular* forms (columns 1, 2 in 1st 7 rows) – *noun_plural* forms (column 3, 4 in 1st 7 rows) and *adjective* forms (columns 1, 2 in last 4 rows) – *adverb* forms (columns 3, 4 in last 4 rows).

Output :

1. ***List of Bilingual stems and suffixes***: These include the list of bilingual stems and suffixes with their observed frequencies in the training dataset. These lists aid in identifying bilingual stems and bilingual suffixes, given a new translation.

2. ***Bilingual suffixes grouped by bilingual stems***: This represents which set of bilingual suffixes attach to which bilingual stem.

3. ***Bilingual Suffix Clusters***: A set of bilingual stems that share same suffix transformations form a cluster or a bilingual suffix class.

Table 8.1: Bilingual variants in EN-HI Lexicon

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)

The first two of these resources were illustrated in the Tables 7.7 and 7.6 of Chapter 7. For instance, in the Table 7.7, *‘plant’ ⇔ ‘paudh’* is a bilingual stem conflating two different surface translation forms *‘plant’ ⇔ ‘paudhA’* and *‘plants’ ⇔ ‘paudhoM’* (bilingual variants shown in the third row of the Table 8.1). *(‘’, ‘A’), (‘s’, ‘oM’)* in the Table 7.6 are automatically identified bilingual suffixes which are attached to 29,529 and 226 different bilingual pairs respectively, of which two bilingual pairs are *‘plant’ ⇔ ‘paudhA’* and *‘plants’ ⇔ ‘paudhoM’*.

As an example for bilingual suffix clusters, consider the Table 8.5. In the 1st row, *(‘’, ‘A’)* and *(‘s’, ‘oM’)* represent bilingual suffixes that combine with bilingual stems, *‘plant’ ⇔ ‘paudh’*, *‘boy’ ⇔ ‘laDak’* and many more, thus forming a cluster. These allow new translation forms to be subsequently suggested upon identification of bilingual stems and suffixes in an unseen translation given as input.

8.2.2 Segmentation of Unseen Bilingual Pair as Classification

Given a translation form, one can conveniently infer similar translation forms by deducing first the productive bilingual segments constituting of bilingual stems and suffixes and by further determining to which bilingual suffix classes the segmented bilingual pair belongs. In this regard, I discuss the use of SVM based linear classifier[5] [R.+08] in predicting if a given segmentation option corresponds to a valid boundary or not, from among all possible segmentations constituting of bilingual stems and suffixes that could possibly be considered for a bilingual pair.

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

Algorithm 2 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 $S_{SuffixPair}$.
        end if
    end for
    for each suffix pair $(s_{i_{L1}},s_{i_{L2}}) \epsilon S_{SuffixPair}$ do
        if $\exists$ $m,n$ such that $(ms_{i_{L1}},ns_{i_{L2}}) \epsilon S_{SuffixPair}$, $u_2m = u_1, v_2n = v_1$
            and bilingual stem $(u_1,v_1)$ and $(u_2,v_2)$ $\epsilon S_{StemPair}$, then
                retain $(ms_{i_{L1}},ns_{i_{L2}})$ iff
                $Strength(s_{i_{L1}},s_{i_{L2}})$ or $Strength(m,n) > Strength(ms_{i_{L1}},ns_{i_{L2}})$.
        end if
    end for
    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 $(p_{i_{L1}},p_{i_{L2}})$.
        end if
    end for
    Cluster the stem pairs sharing similar suffix transformations into bilingual suffix
        classes.
end procedure
```

$$(p_{1_{L1}},p_{1_{L2}})(s_{1_{L1}},s_{1_{L2}}),(p_{2_{L1}},p_{2_{L2}})(s_{2_{L1}},s_{2_{L2}}),.................,(p_{n_{L1}},p_{n_{L2}})(s_{n_{L1}},s_{n_{L2}}) \qquad (8.1)$$

In Equation 8.1, 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, representing instances in the test set). A binary classifier is trained using the features identified from the training dataset made up 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).

8.2.2.1 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, with all real values undergoing normalisation.

Stand-alone Bilingual Pair Stand-alone Bilingual Pair represents 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.* [HW74]. Instances of bilingual stems appearing as stand-alone bilingual pairs in the lexicon for language pair EN-HI 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 [Gol01]. 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 [Pop+10]. Further, the index values have been used as features in correcting the problem with predecessor variety values resulting from normalisation [Çöl10]. 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, *'plants'* ⇔ *'paudhoM'*, with a candidate bilingual stem *'plant'* ⇔ *'paudh'*, the offsets[6] are 5 and 3 EN and HI characters, respectively. For the bilingual pair, *'boys'* ⇔ *'laDakoM'*, with a candidate bilingual stem *'boy'* ⇔ *'laDak'*, the offsets are 3 and 4.
- A pair of integer-valued features corresponding to the offsets from the end of the bilingual pair (with respect to candidate boundary). For each of the above mentioned examples, the offsets are 1 and 2 EN and HI character, respectively.

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

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

$$H(p_{L1},p_{L2}) \quad = \quad - \sum_{(s_{L1},s_{L2}) \in succ(p_{L1},p_{L2})} \frac{f(p_{L1}s_{L1},p_{L2}s_{L2})}{f(p_{L1},p_{L2})}.log_2 \frac{f(p_{L1}s_{L1},p_{L2}s_{L2})}{f(p_{L1},p_{L2})} \tag{8.2}$$

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.

Normalised 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 8.2 by replacing successor with predecessor and switching the concatenation order.

Normalised Successor Variety (NSV) and Normalised Predecessor Variety (NPV) The successor variety value is defined 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 [Har70] is employed in numerous word-segmentation tasks [AS+05; Bor08; Déj98; SP07]. Further, researches show how this measure could be utilised in improving the segmentation results [Çöl10; HW74].

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 provide an elaborate analysis of the problems concerning SV values and the suggested improvements using normalised SV scores [Çöl10].

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.

$$bmf(m_{L1},m_{L2}) = Number\ of\ unique\ bilingual\ pairs\ (m_{L1},m_{L2})\ attaches\ to. \tag{8.3}$$

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 [DN07]. 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.

$$gs(m_{L1}, m_{L2}) = \sum_{(w_{i_{L1}}, w_{i_{L2}})} Strength(w_{i_{L1}}, w_{i_{L2}}). \tag{8.4}$$

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 [PS08; Zem08].

Table 8.4 (columns 3 and 4) shows the scores for frequent bilingual suffixes using each of the above mentioned frequency based scoring functions.

8.2.2.2 Clusters as Training Data

The clusters (bilingual suffix classes) generated in the learning phase are additionally used as training data to model the bilingual suffixes for classification. Each of the possible candidate segmentation boundaries is characterised by the below listed cluster-based features.

Cluster-based Bilingual Suffix Length (CBSL) This is calculated as the number of times a bilingual pair which is (l_1, l_2) characters long contains an (sl_1, sl_2) character long bilingual suffix, normalised by the total number of bilingual pairs with length (l_1, l_2) [BK15].

Cluster-based Bilingual Suffix Probability (CBSP) This represents the probability that a candidate bilingual morphological extension is a correct bilingual suffix 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 $(s_{i_{L1}}, s_{i_{L2}})$ [BK15].

8.2.3 Generation of OOV Entries - Split All and Generate

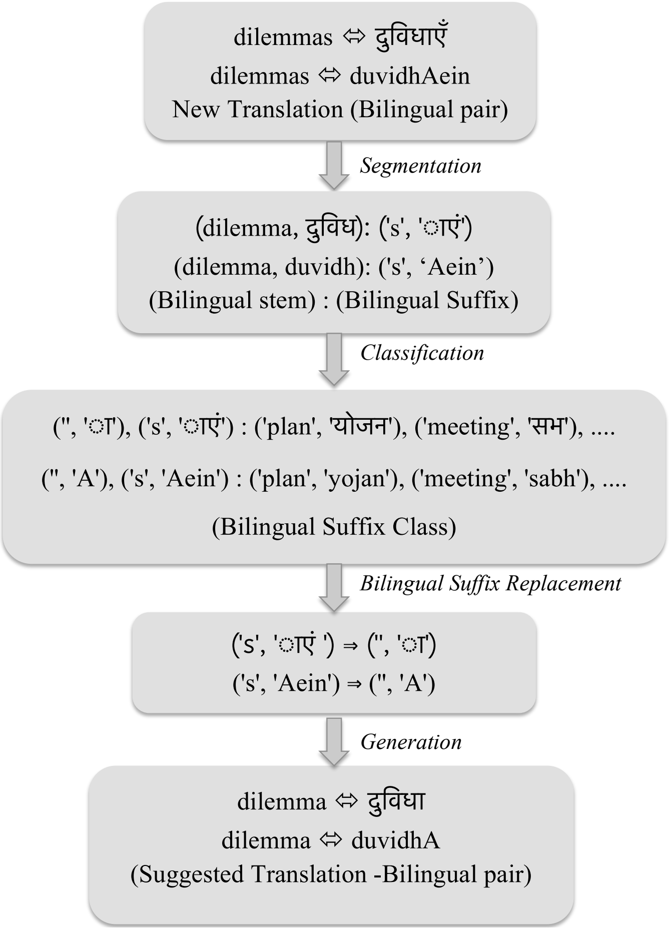
Whenever a new bilingual pair that was automatically extracted and added to the lexicon needs to be analysed for segmentation boundary consisting of stems and suffixes, I consider all possible splits restricting the first part (bilingual stem) to a minimum of 3 characters. In the previous section (Section 8.2.2), I have discussed the supervised approach for selecting bilingual segments by classification, from among all possible segmentation options for any given bilingual pair. The approach allows identification of the segmentation boundary, and hence bilingual segments made of stems and suffixes. Further, upon identification of bilingual suffix classes, all translation forms different from those existing in the lexicon can be inferred for the given bilingual pair, by applying the bilingual suffix replacement rules learnt and as discussed in the Section 7.3.

The steps involved in suffix class determination and translation generation, post-segmentation of a bilingual pair, is discussed with an illustration below.

8.2.3.1 Suffix Class Determination and Translation Generation

Segmentation boundary (after classification) for any bilingual pair (translation equivalent) corresponds to the split position yielding bilingual stem and bilingual suffix. Generation of new forms is possible if it is possible to deduce first, to which bilingual suffix classes the segmented bilingual pair belongs. In other words, the problem is that of classifying the bilingual pair to the most appropriate suffix class based on the bilingual morph-units surfaced as a result of segmentation. The problem hence, might be tackled as a multi-label classification, where the bilingual suffix classes correspond to target labels each of which is characterised by bilingual stems and suffixes (features).

Figure 8.2: Sample generation


dilemmas ⇔ दुविधाएँ
dilemmas ⇔ duvidhAein
New Translation (Bilingual pair)
Segmentation
(dilemma, दुविध): ('s', 'ाएं')
(dilemma, duvidh): ('s', 'Aein')
(Bilingual stem) : (Bilingual Suffix)
Classification
('', 'ा'), ('s', 'ाएं') : ('plan', 'योजन'), ('meeting', 'सभ'), ...
('', 'A'), ('s', 'Aein') : ('plan', 'yojan'), ('meeting', 'sabh'), ...
(Bilingual Suffix Class)
Bilingual Suffix Replacement
('s', 'ाएं ') ⇒ ('', 'ा')
('s', 'Aein') ⇒ ('', 'A')
Generation
dilemma ⇔ दुविधा
dilemma ⇔ duvidhA
(Suggested Translation -Bilingual pair)

An Illustration Consider that the new bilingual pair whose class is to be determined is, *'dilemmas'* ⇔ *'duvidhAein'*. Further it is learnt that, for the bilingual pair *'dilemmas'* ⇔ *'duvidhAein'* (Figure 8.2), the bilingual suffix resulting from segmentation is ('s', 'Aein'). Since ('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'*.

8.2.4 Experimental Setup-Minimally Supervised Generation

8.2.4.1 Data set

Table 8.2: Statistics of the Data set

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

Bilingual pairs taken from EN-HI bilingual lexicon representing single-word translations form the training data set. Approximately 90% of the entries in the lexicon were acquired from the dictionary[7]. The remaining (10%) entries were partly compiled manually and partially using the Symmetric Conditional Probability based statistical measure from the aligned parallel corpora[8] [DSL99]. The manually extracted translation pairs were extracted from Bible parallel corpora using a tool that worked on the aligner by Tiago et al. [IL05]. The details of the data set and the sources for EN-HI are as shown in the Table 8.2 and in Table 8.3.

Table 8.3: Dataset Sources

Sources	Status	#Entries
hindilearner.com	unverified	154
automatic (hipairs)	accepted	275
automatic (hipairs)	unverified	34
sanskritdocuments.org	accepted	5
sanskritdocuments.org	unverified	50,193
www.dicts.info	accepted	28
www.dicts.info	unverified	1,443
manual	accepted	5,197

[7] http://sanskritdocuments.org/hindi/dict/eng-hin_unic.html, www.dicts.info, hindilearner.com

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

8.2.4.2 Multi-label Classifier - LIBSVM

SVM based tool namely LIBSVM[9] was used to learn the multi-label classifier. A class is represented as a set of features represented by feature-value pairs 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).

After the bilingual suffix class for a translation is determined based on the split, new translations are suggested by applying the suffix replacement rules as was introduced earlier in this chapter.

8.2.4.3 BaseLine - Longest Bilingual Suffix Match (LBSM)

The LBSM technique is used as a baseline for identifying bilingual suffixes. After the learning phase, different sets of bilingual stems that have been grouped according to their bilingual inflectional classes are available. Such sets are referred to as Bilingual Suffix Classes. For each translation in the test set, its bilingual inflections (suffixes) and the associated bilingual suffix class needs to be determined. As baseline, I classify each new (unseen) translation in the test set into the class of longest matching bilingual suffix from the bilingual suffix list [Lin+09]. For instance, the longest bilingual suffix matching the bilingual pair *'conservative'* ⇔ *'rakshAtmak'* is *'ative'* ⇔ *'Atmak'* yielding the bilingual stem *'conserv'* ⇔ *'raksh'*.

8.2.5 Results and Evaluation - Learning Phase

The bilingual suffixes (frequently undergoing transformations) recognised using the approach discussed in Section 8.2.1 are shown in Table 8.4. For each of the bilingual suffixes undergoing frequent replacements, the generative strength statistics (gs) is as shown in column 4. Table 8.5[10] 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'*. The association scores (discussed in Section 7.2.4.2 of Chapter 7) between bilingual suffixes in the chosen clusters are as well shown in column 2 of Table 8.5.

A few of the induced bilingual suffix class based morphological patterns were incomplete as not all the translation forms were seen in the lexicon used as training data. In spite of using the human-usable bilingual dictionary that generally has no inflectional variants of words and their translations with inflectional modifications, few of the translations still did not have near forms. In certain cases, distinct surface translation forms

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

[10] *Number of times a bilingual suffix co-occurs with another bilingual suffix in the input lexicon [Kav+15c]

Table 8.4: Bilingual Suffixes with frequent replacements

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

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')

due to inflection classes resulted in distinct bilingual suffix classes some of which need to be collapsed.

The bilingual segments and clustering results are evaluated by examining the applicability of induced segments in generating new translations. The translation lexicon is first completed 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 (specifically, using the bilingual segments derived out of them) in the training data for the chosen bilingual suffix classes [Kav+14a]. The generated translations are then evaluated. Table 8.6 shows the results of the learning phase. The precision for generated translations is calculated as the fraction of correctly generated bilingual pairs to the 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 [Kav+14a].

Table 8.6: Clustering Statistics using different Learning Schemes

Learning Approach	Unique Bilingual Stem Count	Unique Bilingual Suffix Count	Number of Clusters	Generation Precision
[Kav+14a]	12,603	781	224	81.31
[Kav+15a]	10,224	426	143	86.52

8.2.5.1 Generation Results

Table 8.7[11] shows suggested translation examples. The generated translations fall into one of the 3 classes (separated by thick border) based on the degree of correctness. First 3 rows represent acceptable translations (Accept). The following row shows 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 8.7) 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'*.

Table 8.7: Generated Translations for EN-HI

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')

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'*. Similarly, the correct translation for the word *'weeks'* is *'saptahoM'*. The adverbial variant *'weekly'* ⇔ *'sAptahik'* is obtained from *'weeks'* ⇔ *'saptahoM'* by replacing *'s'* with *'ly'* in EN and *'oM'* with *'ik'* in HI, with the stem undergoing a change from *'a'* to *'A'*.

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

8.2.6 Results and Evaluation - Minimally supervised learning

The results of segmentation by classification were 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:

$$P = t_p/(t_p + f_p) \tag{8.5}$$

$$R = t_p/(t_p + f_n) \tag{8.6}$$

$$F_m = 2 * P * R/(P + R) \tag{8.7}$$

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 8.8. When new translations are given as inputs, the best f-measure of 70.88% is achieved. This fall in the precision by 10% may be attributed to misclassification of the given new pair into a bilingual suffix classes and due to relatively fewer number of similar surface forms in the dictionary used as training data, which is not the case when both the bilingual segments are known to have occurred in the training data set.

Table 8.8: Results of minimally supervised learning

Features	Precision	Recall	F-measure
Longest Bilingual Suffix Match	74.71	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

8.3 Summary

In this chapter the minimally supervised approach for generation of OOV translations was discussed. The training data prepared using the bilingual learning approach partially serves as the basis for supervised segmentation along with the bilingual lexicon [Kav+14a]. 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%.

Chapter 9

Conclusion and Future Work

9.1 Overview

Improving the quality of machine translation systems calls for much work to be invested on the translation lexicon, firstly, in ensuring careful selection of automatically extracted translations without hindering the re-alignment precision and extraction accuracy and secondly, in maintaining a high-coverage lexicon providing maximum possible translation coverage. The lexicon quality issue i.e., the correctness of lexicon entries, in the context of their iterative usage in realignment and extraction is handled through prior classification followed by human validation. The coverage issue with respect to word-to-word translations is currently handled through translation suggestion via bilingual learning of suffixation operations and generalisation of previously known word-to-word translation forms.

9.2 Classification

The need for automatically extracted translation equivalents to be manually validated, prior to its use for iteratively aligning, extracting and validating new translation pairs, when the extraction method proposed by Aires et al. [Aires+09] was our main extraction process, has been the main motivation behind the classification work discussed in the earlier part of this thesis. 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 reduces this work, thereby saving the time involved and progressively improving alignment and extraction quality, and hence contributing to improve translation quality. When the first classifier, SVMTEC was trained we had approximately 120,000 entries to validate. By classifying them, we obtained 80,000 entries classified as incorrect and about 40,000 classified as correct. From results obtained in a controlled environment we knew that about 15% of them were incorrectly classified. Validation of these entries were made by one person in about 30 days, not exceeding 5 work hours per day.

Highly precise classifiers discussed in Part I of this thesis in combination with the *Lexicon Validation Interface* allows the validators to select several bilingual entries with minimum mouse clicks, as the validator needs to select the unverified entries tagged as 'accepted' by the classifier, using a specific validation interface that enables the manual marking of a chosen number of those entries as 'correct', followed by an individual checking of those entries and eventually the consultation of a bilingual concordancer. Categorisation of translations were mostly centred on linguistic and/or translation properties for multi-word translations and for unigram translations based on segregating grammatical errors from grammatically correct unigram translations.

Classification with human validation resulted in an enriched annotated lexicon suitable for machine learning systems such as bilingual morphology learning, translation suggestion tool, besides its primary use in alignment, extraction and machine translation.

9.2.1 Future Applications of Classification

Currently, the classifier is evaluated for language pairs with similar character set. The adaptability of the classifier features in classifying translations with different character sets can be explored in future.

Another area for exploration that directly follows from classification of extracted translation equivalents, discussed in Part I of the thesis is its possible use in helping the human post editors to tackle the mistranslated translation segments. It is evident that the translation produced by a translation engine is the result of splitting each sentence into segments and using the translations of the segments that maximises a function of the translation models used. So, knowing the segmentations made by an engine, which is the prevalent scenario, we know the translation used for each of the segments. So, by classifying each of those segments and the translations used, we get translated segments that according to the classifier are wrongly translated and others that are correctly translated. By highlighting those classifications by some means (colour, for instance), human post-editors may look first at the wrongly translated sub-segments and try to first rectify those translations, thus supporting post-edition operations. Sub-sentence alignment proposed by Luís Gomes [Gom16], having a different knowledge base may also bring out a different perspective of mistranslation made by the translation engine under use.

9.3 Bilingual Learning and Generation

Moving on to the latter aspect of lexicon usage, I explored morphological similarities between the seen word-to-word translation forms as a means to generalise the existing examples and based on these learn to induce new translation forms that are infrequent or have never been encountered in the parallel corpus used for translation (lexicon) extraction. In this context, using known bilingual translation forms from the existing lexicon, bilingual morph-like units comprising of bilingual stems and suffixes are identified and are productively combined in suggesting unseen translation forms. New translation forms that are identical to but different from existing translations are suggested with precision as high as 90%. 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. Handling irregular forms remain an open issue, as the approach is purely bilingual suffixation based and does not capture stem changes prior to suffixation.

9.3.1 Future Applications of Bilingual Learning

9.3.1.1 Bigram Translation Generation

One of the applications of the bilingual stem mappings and bilingual suffix classes is in generating bilingual translation forms. Personal[1] pronouns (with number[2] variations) such as *'I'* or *'we'* appearing prior to the verb in English, as in *'I declared'*, are reflected as conjugation suffixes in Portuguese as in translation pairs *'I declare'* ⇔ *'declaro'*, *'we declare'* ⇔ *'declaramos'* and these personal pronouns may be omitted in Portuguese. Further, the inflections in the target vocabulary is also an indicative of *tenses*[3] and *moods*[4] represented in the source side. Similar patterns include *'have declared'* ⇔ *'declararam'*, *'has declared'* ⇔ *'declarou'*, *'will (shall) declare'* ⇔ *'declarará'*. These observations further allow the generation of bigram translations of the form, for example, *'I demonstrated'* ⇔ *'(Eu) demonstrei'*, *'we demonstrate'* ⇔ *'(nós) demonstramos'*, *'shall demonstrate'* ⇔ *'demonstrará, demonstrarão, demonstra'* or *'demonstram'*, *'have demonstrated'* ⇔ *'demonstraram'*, and so forth. Further, generation of multi-word translation forms remains unexplored in the current work.

9.3.1.2 Compression

Full text indexing using suffixes, stems, white spaces and un-analised word forms will bring greater repetition, so also, compression will be much higher than currently obtained [Cos+13; Cos+15]. Instead of using word-based compression as proposed by Costa et al. [Cos+13; Cos+15], a representation taking into account learned stem-pairs and suffix pairs aligned might be more compact as the number of stems is lower than the number of words of a language and the number of suffixes, a few hundreds, due to much higher repetition and compression may be higher. Additionally, one may obtain and search for patterns as $1/ing ⇔ que $1/am/em/a/e where '/' denotes concatenation of a stem with

[1] Three persons namely, first, second, third

[2] The two numbers represent singular and plural

[3] Six (five) tenses, namely, present, preterite, imperfect, pluperfect, future, and conditional in Portuguese.

[4] The four (three) moods include indicative, subjunctive, imperative (and conditional) in Portuguese

suffixes ‘am’, ‘em’, ‘a’ and ‘e’ for the Portuguese side and suffix ‘ing’ for the English side. This is further explored in the Thesis of Luís Gomes [Gom16].

9.3.1.3 Pivoting

Machine Translation Systems relying on Translation Lexicons, to be multilingual demand that translation lexicons are maintained for multiple language pairs. For this purpose, pivoting may be used to induce missing entries of a bilingual lexicon for the language pair ($L_2 - L_3$) using automatically acquired and manually validated bilingual lexicons for language pairs ($L_1 - L_2$) and ($L_1 - L_3$). While the lexicon is derived by transitivity [Gom16], the correspondences are identified based on previously learnt bilingual stems and suffixes rather than surface translation forms. Induced pairs are automatically validated using a binary classifier trained on an existing, automatically acquired, manually validated bilingual translation lexicon for language pair ($L_2 - L_3$).

In the preliminary experiments, EN-FR and EN-PT lexicon of word-to-word translations with EN as pivot language was employed, and used to generate word-to-word translations for the language pair FR-PT. The classifier trained on morphological and similarity-based features (discussed in the Chapter 5, and the classification results for FR-PT are shown in the figure 5.2 and Table 5.2) enables the automatic validation of induced pairs. However, experiments were not conclusive and on-going work in this area will be published in the near future.

9.4 Summary

The content of this thesis is centred on two major aspects pertaining to the automatically extracted bilingual translations lexicons- First, the selection and usability of the lexicon entries for subsequent iterations of parallel corpora alignment, new translation extraction and validation; second, the adaptability of the existing lexicon entries in suggesting OOV bilingual lexicon entries from the perspective of augmenting the automatically acquired lexicon in order to deal with OOV entries.

Concerning the first aspect, the main objective was to:

Objective 1: Apply machine learning technique such as classification so as to speed up and ease the manual process of validating automatically extracted translation candidates as correct or incorrect.

In retrospecting the questions posed summarised as Objective 1, I conclude that SVM-based binary classifiers trained to automatically segregate correct and incorrect lexicon entries do benefit us in accelerating the human validation process, enabling 5,000 entries to be validated per day, thereby saving the time involved in manual validation. Few of the translations (less than 5% for EN-PT) are incorrectly classified, requiring only those to be manually labeled as correct or incorrect. It was evident that for two terms to be considered an adequate translation pair, parallelism is a key criteria, which was successfully judged using the translation mis-coverage feature. Further, it was also learnt that adequate word-to-word translations should agree not only with respect to the stems that conveys the meaning but also suffixes which affirms grammatically correct translations. The size of manual annotated data does influence the performance of the classification tool, so also the distribution of positive and negative examples. This is very well understood from the experiments on EN-PT, FR-PT vs EN-FR word-to-word translation classification (EN-FR word-to-word translation classification failed due to insufficient negative examples).

Concerning the adaptability of the existing lexicon entries in continual lexicon augmentation the below mentioned objectives were formulated:

Objective 2: Automatically induce segmentation and bilingual morph-like units from a bilingual corpus of translations.

Objective 3: Apply the induced morph-like units to suggest OOV bilingual lexicon entries by productively combining the induced bilingual stems and suffixes.

The plausibility of improvising upon the existing lexicon entries by learning from what is already known was examined from the perspective of word-to-word translations and of learning bilingual suffixation operations from those translations. It does turn out that separation of morphological suffixes conflates various forms of a translation into a bilingual stem which is a crucial source of information and that bilingual suffix transformation rules can be learnt by identifying frequent bilingual suffixes and their association with the bilingual stems. Partition approach and the Suffix Co-occurrence score based clustering were experimented in identification of bilingual suffix paradigms, with the intention of generalizing the rules for suggesting OOV translations. Experiments under sufficient training data conditions reveal that by productively combining the bilingual stems and bilingual suffixes and by duly considering the bilingual suffix paradigms, translation suggestion can be achieved with accuracy nearing 90%.

Following the experimentations on EN-HI translation lexicon, it was learnt that the success of the bilingual learning and generation approach relies on having sufficient near translation forms (Chapter 7). Hence, to cope up with the limited training data conditions, as with language pairs such as EN-HI, supervised learning strategy was proposed to infer segmentations for new bilingual pairs. Additionally, human-usable dictionaries were employed to boost the learning process. Instead of relying merely on instance-specific features, features representing global distribution of morph-like units with respect to parameters such as the length of bilingual pairs were used in training the classifier.

Thus, as a direct application of bilingual morph-units, experiments on generating OOV translations were discussed under poor and adequate data conditions.

To summarize, the contributions of the study are :

1. a fully automated classifier for translations automatically acquired from aligned parallel corpora

2. a bilingual learning technique to infer bilingual stems (conflating various translation forms), their bilingual morphological extensions and the bilingual suffix

paradigms from a given automatically extracted and validated bilingual translation lexicon.

3. a OOV translation suggestion scheme employing the knowledge bases learnt in (2).

The first of these contributions consists of learning a binary classifier model based on SVMs for classifying word-to-word and multi-word translations. The second involves a three step learning scheme comprising of: first, bilingual pair decomposition for identification of candidate bilingual stems and suffixes; second, filtering, where bilingual suffixes are grouped with respect to their association with the bilingual stems, subsequently followed by the elimination of redundant groups; and third, clustering approach for identification of bilingual suffix clusters. Finally, the bilingual resources learnt in (2) are employed in two generation schemes: first, generation by simple concatenation of bilingual stems and suffixes belonging to same suffix class; second, generation scheme involving three phases comprising of segmentation via classification, suffix class determination via multi-label classification and concatenation of bilingual stems and suffixes.

Bibliography

[Air+09] J. Aires, G. P. Lopes, and L. Gomes. "Phrase translation extraction from aligned parallel corpora using suffix arrays and related structures". In: *Progress in Artificial Intelligence* (2009), pp. 587–597.

[Air15] J. Aires. "Genuine phrase-based statistical machine translation with supervision". PhD thesis. 2015. URL: http://hdl.handle.net/10362/18153.

[Ake+13] A. Aker, M. L. Paramita, and R. J Gaizauskas. "Extracting bilingual terminologies from comparable corpora." In: *Proceedings of the 51st Annual Meeting for Computational linguistics*. Vol. 2. 2013, pp. 402–411.

[AS+05] R. Al-Shalabi, G. Kannan, I. Hilat, A. Ababneh, and A. Al-Zubi. "Experiments with the successor variety algorithm using the cutoff and entropy methods". In: *Information Technology Journal* 4.1 (2005), pp. 55–62.

[Aro+08] K. Arora, M. Paul, and E. Sumita. "Translation of unknown words in phrase-based statistical machine translation for languages of rich morphology". In: *Proceedings of SLTU* (2008).

[BK07] S. Bergsma and G. Kondrak. "Alignment-based discriminative string similarity". In: *Annual meeting-ACL*. Vol. 45. 1. 2007, p. 656.

[Bor08] S. Bordag. "Unsupervised and knowledge-free morpheme segmentation and analysis". In: *Advances in Multilingual and Multimodal Information Retrieval*. Springer, 2008, pp. 881–891.

[Bro+93] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. "The mathematics of statistical machine translation: Parameter estimation". In: *Computational linguistics* 19.2 (1993), pp. 263–311. ISSN: 0891-2017.

[BK15] T. Brychcín and M. Konopík. "HPS: High precision stemmer". In: *Information Processing & Management* 51.1 (2015), pp. 68–91.

[CB07] C. Callison-Burch. "Paraphrasing and translation". PhD thesis. University of Edinburgh, 2007.

[CB+06] C. Callison-Burch, P. Koehn, and M. Osborne. "Improved statistical machine translation using paraphrases". In: *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics. 2006, pp. 17–24.

[Che+05] B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico. "The ITC-irst SMT system for IWSLT-2005". In: *Proceeding of IWSLT* (2005), pp. 98–104.

[Che+09] B. Chen, G. Foster, and R. Kuhn. "Phrase translation model enhanced with association based features". In: *Proceedings of MT-Summit XII* (2009).

[Che+98] H. Chen, S. Hueng, Y. Ding, and S. Tsai. "Proper name translation in cross-language information retrieval". In: *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics. 1998, pp. 232–236.

[Che+06] Y. Chen, X. Shi, and C. Zhou. "The XMU phrase-based statistical machine translation system for IWSLT 2006". In: *Proc. of the International Workshop on Spoken Language Translation*. Kyoto, Japan, 2006, pp. 153–157.

[CL07] T. Cohn and M. Lapata. "Machine translation by triangulation: Making effective use of multi-parallel corpora". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguists*. Vol. 45. Prague, Czech Republic, 2007, pp. 728–735.

[Çöl10] Ç. Çöltekin. "Improving Successor Variety for Morphological Segmentation". In: *LOT Occasional Series* 16 (2010), pp. 13–28.

[Cos+11] J. Costa, L. Gomes, G. Lopes, and L. Russo. "Managing and Querying a Bilingual Lexicon with Suffix Trees". In: *EPIA 2011*. APPIA, Portuguese Association for Artificial Intelligence, 2011, 675–689.

[Cos+13] J. Costa, L. Gomes, G. P. Lopes, L. M. S. Russo, and N. R. Brisaboa. "Compact and Fast Indexes for Translation Related Tasks". In: *Progress in Artificial Intelligence*. Ed. by L. P. R. Luis Correia and J. Cascalho. Lecture Notes in Computer Science 8154. Springer-Verlag, Sept. 2013, pp. 504–515.

[Cos+15] J. Costa, L. Gomes, G. P. Lopes, and L. M. S. Russo. "Improving Bilingual Search Performance Using Compact Full-Text Indices". In: *Computational Linguistics and Intelligent Text Processing*. Springer, 2015, pp. 582–595.

[CL02] M. Creutz and K. Lagus. "Unsupervised discovery of morphemes". In: *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*. ACL. 2002, pp. 21–30.

[DSL99] J. F. Da Silva and G. P. Lopes. "Extracting multiword terms from document collections". In: *Proceedings of the VExTAL: Venezia per il Trattamento Automatico delle Lingue*. 1999, pp. 22–24.

[DN07] S. Dasgupta and V. Ng. "Unsupervised word segmentation for Bangla". In: *Proceedings of ICON*. 2007, pp. 15–24.

[Déj98] H. Déjean. "Morphemes as necessary concept for structures discovery from untagged corpora". In: *Proceedings of the NeMLaP3/CoNLL98*. ACL. 1998, pp. 295–298.

[Des+14] S. Desai, J. Pawar, and P. Bhattacharyya. "A Framework for Learning Morphology using Suffix Association Matrix". In: *WSSANLP-2014*. 2014, pp. 28–36.

[Dic10] M. Dickinson. "Generating learner-like morphological errors in Russian". In: *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics. 2010, pp. 259–267.

[Eck+08] M. Eck, S. Vogel, and A. Waibel. "Communicating unknown words in machine translation". In: *Proceedings of LREC*. 2008.

[FY14] M. Felice and Z. Yuan. "Generating artificial errors for grammatical error correction". In: *EACL*. 2014, pp. 116–126.

[FA09] J. Foster and Ø. E. Andersen. "GenERRate: generating errors for use in grammatical error detection". In: *Proceedings of the fourth workshop on innovative use of nlp for building educational applications*. Association for Computational Linguistics. 2009, pp. 82–90.

[FM07] A. Fraser and D. Marcu. "Measuring word alignment quality for statistical machine translation". In: *Computational Linguistics* 33.3 (2007), pp. 293–303.

[FY98] P. Fung and L. Yee. "An IR approach for translating new words from nonparallel, comparable texts". In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics. 1998, pp. 414–420.

[Gar+09] N. Garera, C. Callison-Burch, and D. Yarowsky. "Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences". In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics. 2009, pp. 129–137.

[GM08] A. de Gispert and J. B. Marino. "On the impact of morphology in English to Spanish statistical MT". In: *Speech Communication* 50.11-12 (2008), pp. 1034–1046. ISSN: 0167-6393.

[Gis+05] A. de Gispert, J. Mariño, and J. Crego. “Improving statistical machine translation by classifying and generalizing inflected verb forms”. In: *Proceedings of 9th European Conference on Speech Communication and Technology*. Lisboa, Portugal, 2005, pp. 3193–3196.

[Gol01] J. Goldsmith. “Unsupervised learning of the morphology of a natural language”. In: *Computational linguistics* 27.2 (2001), pp. 153–198.

[GL09] L. Gomes and G. P. Lopes. “Parallel texts alignment”. In: *New Trends in Artificial Intelligence, 14th Portuguese Conference in Artificial Intelligence, EPIA 2009*. Aveiro, 2009, pp. 513–524.

[Gom16] L. Gomes. “Translation Alignment and Extraction within a Human-Machine Interactive Workflow (to appear)”. PhD thesis. Monte da Caparica: Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (FCT-UNL), 2016.

[GL11] L. Gomes and G. P. Lopes. “Measuring Spelling Similarity for Cognate Identification”. In: *Progress in Artificial Intelligence — 15th Portuguese Conference on Artificial Intelligence, EPIA 2011*. Lisbon, Portugal: Springer, 2011, pp. 624–633.

[Gus97] D. Gusfield. *Algorithms on strings, trees, and sequences: computer science and computational biology*. pages 52–61. Cambridge Univ Pr, 1997.

[GG08] F. Guzmán and L. Garrido. “Translation paraphrases in phrase-based machine translation”. In: *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*. Springer-Verlag. 2008, pp. 388–398.

[H.09] H. H. “Unsupervised learning of morphology and the languages of the world”. PhD thesis. Gothenburg: Chalmers University of Technology and Göteborg, 2009.

[Hab08] N. Habash. “Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation”. In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics. 2008, pp. 57–60.

[HW74] M. A Hafer and S. F. Weiss. “Word segmentation by letter successor varieties”. In: *Information storage and retrieval* 10.11 (1974), pp. 371–385.

[Hag+08] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein. “Learning bilingual lexicons from monolingual corpora”. In: *Proceedings of ACL-08: HLT* (2008), pp. 771–779.

[HB11] H. Hammarström and L. Borin. “Unsupervised learning of morphology”. In: *Computational Linguistics* 37.2 (2011), pp. 309–350.

[Har70] Z. S. Harris. "From phoneme to morpheme". In: *Papers in Structural and Transformational Linguistics - Formal Linguistics Series* (1970), pp. 32–67.

[He+06] Z. He, Y. Liu, D. Xiong, H. Hou, and Q. Liu. "Ict system description for the 2006 tc-star run# 2 slt evaluation". In: *Proceedings of TCSTAR Workshop on Speech-to-Speech Translation*. Citeseer. 2006, pp. 63–68.

[Hua05] F. Huang. "Cluster-specific name transliteration". In: *Proceedings of the HLT-EMNLP*. 2005.

[IL05] T. Ildefonso and G. P. Lopes. "Longest sorted sequence algorithm for parallel text alignment". In: *Tenth International Conference on Computer Aided Systems Theory— Revised Selected Papers*. Lecture Notes in Computer Science 3643. Springer. 2005, pp. 81–90.

[Joh+07] J. H. Johnson, J. Martin, G. Foster, and R. Kuhn. "Improving translation quality by discarding most of the phrasetable". In: *Proceedings of EMNLP-CoNLL*. 2007, pp. 967–975.

[Kav+11] K. M. Kavitha, L. Gomes, and G. P. Lopes. "Using SVMs for Filtering Translation Tables for Parallel corpora Alignment". In: *15th Portuguese Conference in Arificial Intelligence, EPIA 2011*. 2011, pp. 690–702.

[Kav+14a] K. M. Kavitha, L. Gomes, and J. G. P. Lopes. "Identification of Bilingual Segments for Translation Generation". In: *Advances in Intelligent Data Analysis XIII*. Vol. 8819. LNCS. Springer International Publishing, 2014, pp. 167–178.

[Kav+14b] K. M. Kavitha, L. Gomes, and J. G. P. Lopes. "Identification of Bilingual Suffix Classes for Classification and Translation Generation". In: *Advances in Artificial Intelligence, IBERAMIA 2014*. LNCS. Springer, 2014, pp. 154–166.

[Kav+15a] K. M. Kavitha, L. Gomes, and J. G. P. Lopes. "Bilingually motivated segmentation and generation of word translations using relatively small translation data sets". In: *Proceedings of 29th Pacific Asia Conference on Language, Information and Computation: Posters*. PACLIC29, 2015, 187–196.

[Kav+15b] K. M. Kavitha, L. Gomes, J. Aires, and J. G. P. Lopes. "Classification and Selection of Translation Candidates for Parallel Corpora Alignment". In: *EPIA 2015, Progress in Artificial Intelligence*. Ed. by F. Pereira, P. Machado, E. Costa, and A. Cardoso. Vol. 9273. Lecture Notes in Computer Science. Springer International Publishing, 2015, pp. 723–734.

[Kav+15c] K. M. Kavitha, L. Gomes, and J. G. P. Lopes. "Learning Clusters of Bilingual Suffixes using Bilingual Translation Lexicon". In: *Mining Intelligence and Knowledge Exploration*. Springer, 2015, 607–615.

[KG98] K. Knight and J. Graehl. “Machine transliteration”. In: *Computational Linguistics* 24.4 (1998), p. 612. ISSN: 0891-2017.

[Koe10] P. Koehn. *Statistical machine translation*. Cambridge University Press, 2010.

[KK02] P. Koehn and K. Knight. “Learning a translation lexicon from monolingual corpora”. In: *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*. Association for Computational Linguistics. 2002, pp. 9–16.

[KK03] P. Koehn and K. Knight. “Empirical methods for compound splitting”. In: *Proceedings of the tenth conference on EACL-Volume 1*. ACL. 2003, pp. 187–193.

[Koe+05] P. Koehn, A. Axelrod, A. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. “Edinburgh system description for the 2005 IWSLT speech translation evaluation”. In: *International Workshop on Spoken Language Translation*. Citeseer. 2005.

[Koe+07] 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*. Association for Computational Linguistics. 2007, pp. 177–180.

[Kon05] G. Kondrak. “Cognates and word alignment in bitexts”. In: *Proceedings of the 10th Machine Translation Summit*. 2005, pp. 305–312.

[Kut+05] 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*. 2005.

[LL09] A. Lardilleux and Y. Lepage. “Sampling-based multilingual alignment”. In: *Proceedings of Recent Advances in Natural Language Processing*. 2009, pp. 214–218.

[LD05] Y. Lepage and E. Denoual. “Purest ever example-based machine translation: Detailed presentation and assessment”. In: *Machine Translation* 19.3 (2005), pp. 251–282. ISSN: 0922-6567.

[Lev66] V. Levenshtein. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet Physics Doklady*. Vol. 10. 8. 1966, pp. 707–710.

[Lin+09] K. Lindén, M. Silfverberg, and T. Pirinen. “HFST Tools for Morphology – An Efficient Open-Source Package for Construction of Morphological Analyzers”. In: *State of the Art in Computational Morphology*. Vol. 41. CCIS. Springer, 2009, pp. 28–47.

[Lop08] A. Lopez. “Statistical machine translation”. In: *ACM Computing Surveys (CSUR)* 40.3 (2008), pp. 1–49. ISSN: 0360-0300.

[Mel95] I. Melamed. “Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons”. In: *Proceedings of the Third Workshop on Very Large Corpora*. Boston, MA. 1995, pp. 184–198.

[MT97] H. Momouchi and K. Tochinai. “Prediction method of word for translation of unknown word”. In: *Proceedings of the IASTED International Conference, Artificial Intelligence and Soft Computing, July 27 to August 1 1997, Banff, Canada*. Acta Pr. 1997, p. 228.

[Mon+09] C. Monson, J. Carbonell, A. Lavie, and L. Levin. “Paramor and morpho challenge 2008”. In: *Evaluating Systems for Multilingual and Multimodal Information Access*. Springer, 2009, pp. 967–974.

[ON03] F. J. Och and H. Ney. “A systematic comparison of various statistical alignment models”. In: *Computational linguistics* 29.1 (2003), pp. 19–51.

[ON04] F. J. Och and H. Ney. “The alignment template approach to statistical machine translation”. In: *Computational Linguistics* 30.4 (2004), pp. 417–449. ISSN: 0891-2017.

[PS08] A. K. Pandey and T. J. Siddiqui. “An unsupervised Hindi stemmer with heuristic improvements”. In: *Proceedings of the second workshop on Analytics for noisy unstructured text data*. ACM. 2008, pp. 99–105.

[Poo+09] H. Poon, C. Cherry, and K. Toutanova. “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*. ACL. 2009, pp. 209–217.

[Pop+10] K. Popat, P. P., and P. Bhattacharyya. “Hybrid Stemmer for Gujarati”. In: *23rd International Conference on Computational Linguistics*. 2010, p. 51.

[R.+08] F. R., K. Chang, C. Hsieh, X. Wang, and C. Lin. “LIBLINEAR: A library for large linear classification”. In: *The Journal of Machine Learning Research* 9 (2008), pp. 1871–1874.

[Rap95] R. Rapp. “Identifying word translations in non-parallel texts”. In: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 1995, pp. 320–322.

[Rap99] R. Rapp. “Automatic identification of word translations from unrelated English and German corpora”. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics. 1999, pp. 519–526.

[SS03] K. Sato and H. Saito. "Extracting Word Sequence Correspondences Based on Support Vector Machines". In: *Journal of Natural Language Processing* 10.4 (2003), pp. 109–124. ISSN: 1340-7619.

[SY02] C. Schafer and D. Yarowsky. "Inducing translation lexicons via diverse similarity measures and bridge languages". In: *proceedings of the 6th conference on Natural language learning-Volume 20*. Association for Computational Linguistics. 2002, pp. 1–7.

[SB08] B. Snyder and R. Barzilay. "Unsupervised Multilingual Learning for Morphological Segmentation." In: ACL. 2008, pp. 737–745.

[Som+09] H. Somers, S. Dandapat, and S. Naskar. "A review of EBMT using proportional analogies". In: *3rd International Workshop on Example-Based Machine Translation*. Dublin, 2009, pp. 53–60.

[SP07] B. Stein and M. Potthast. "Putting successor variety stemming to work". In: *Advances in Data Analysis*. Springer, 2007, pp. 367–374.

[Ste+06] 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)*. 2006, pp. 2142–2147.

[SK09] K. Sunitha and N. Kalyani. "Improving word coverage using unsupervised morphological analyser". In: *Sadhana* 34.5 (2009), pp. 703–715. ISSN: 0256-2499.

[TM99] T. Tanaka and Y. Matsuo. "Extraction of translation equivalents from non-parallel corpora". In: *Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*. Citeseer. 1999, pp. 109–19.

[Tia+14] L. Tian, D. F. Wong, L. S. Chao, and F Oliveira. "A Relationship: Word Alignment, Phrase Table, and Translation Quality". In: *The Scientific World Journal* 2014 (2014).

[Tie98] J. Tiedemann. "Extraction of translation equivalents from parallel corpora". In: *Proceedings of the 11th NoDaLiDa*. 1998, pp. 120–128.

[Tom+09] N. Tomeh, N. Cancedda, and M. Dymetman. "Complexity-based phrase-table filtering for statistical machine translation". In: (2009), 144–151.

[Tom+11] N. Tomeh, M. Turchi, A. Allauzen, and F. Yvon. "How good are your phrases? Assessing phrase quality with single class classification." In: *IWSLT*. 2011, pp. 261–268.

[Vap00] V. Vapnik. "The Nature of Statistical Learning Theory". In: *Data Mining and Knowledge Discovery* (2000), pp. 1–47.

[Vil+06] D. Vilar, M. Popovic, and H. Ney. “AER: Do we need to “improve" our alignments?” In: *IWSLT*. 2006, pp. 205–212.

[YK06] M. Yang and K. Kirchhoff. “Phrase-based backoff models for machine translation of highly inflected languages”. In: *Proceedings of EACL*. 2006, pp. 41–48.

[Zem08] D. Zeman. “Unsupervised acquiring of morphological paradigms from tokenized text”. In: *Advances in Multilingual and Multimodal Information Retrieval* (2008), pp. 892–899.

[Zha+04] B. Zhao, V. S., and A. Waibel. “Phrase pair rescoring with term weightings for statistical machine translation”. In: (2004).

[Zha+07] B. Zhao, N. Bach, I. Lane, and S. Vogel. “A log-linear block transliteration model based on bi-stream HMMs”. In: *Proceedings of NAACL HLT*. 2007, pp. 364–371.

[ZK02] Y. Zhao and G. Karypis. “Evaluation of hierarchical clustering algorithms for document datasets”. In: *Proceedings of the eleventh international conference on Information and knowledge management*. ACM. 2002, pp. 515–524.