Phrase Table Expansion for Statistical Machine Translation with reduced parallel corpora: the Chinese-Spanish case

Jingyi Han

TESI DOCTORAL UPF / ANY 2017

DIRECTORA DE LA TESI

Dra. Núria Bel

Departament de Traducció i Ciències del Llenguatge
Institut Universitari de Lingüística Aplicada

Universitat
Pompeu Fabra
Barcelona

Acknowledgments

This dissertation would not have been possible without the kind help and support that I have received from so many great people over the years. I would like to express my most heartfelt gratitude to the following people.

First and foremost, I would like to express my sincere appreciation to my supervisor, Nuria Bel, for giving me the opportunity to pursue this PhD program, and also for her excellent guidance, patience and encouragement. The insightfulness and enthusiasm she has for her researches really motivate me for my future career. Thanks for guiding me to the right path.

A special thank you to Professor Mercè Lorente Casafont for accepting my PhD proposal at the beginning of this journey and helping me with the scholarship application.

My appreciation also extends to all my companions of TRL group: Marina Fomicheva, Silvia Necsulescu, Lauren Romeo, Marco del Tredici, Cao Shuyuan, Jorge Diz Pico and Joel Merce, who have been accompanying me on the road and sharing all the nice moments together.

For the wonderful research stay in Dublin, I am truly thankful to Professor Andy Way for welcoming me into ADAPT centre, and also a big thank you to all the guys from Dublin City University: Jinhua Du, Xiaojun Zhang, Jian Zhang, Antonio Toral, Longyue

Wang, Xiaofeng Wu, Liangyou Li and Wei Li, for their kind help and for making me feel at home.

I would also like to thank all the members of IULA for the support that they gave me throughout the whole adventure. Special thanks are reserved for Vanessa Alonso and Sylvie Hochart, who generously helped me with the complicated procedure and documentation.

More than anybody, I want to thank my parents, Liu Yang and Wei Han, for always being there through my ups and downs, and also a big thank you to my whole lovely family for always being supportive from the very beginning.

A special mention goes to Lingxian Bao for being such a great and 100% supportive boyfriend. He can always see the best in me and encourage me to pursue my dreams no matter how crazy they seem. Thanks for starting this adventure with me and he is always my *super-chu*!

Finally, I gratefully acknowledge the funding source that made my PhD work possible: the FI-DGR scholarship of the Secretary of Universities and Investigation (Secretaría de Universidades e Investigación) of the Department of Economy and Knowledge (Departamento de Economía y Conocimiento) of the Government of Catalonia (Generalitat de Catalunya), of the European Social Fund (Fons Social Europeu).

Abstract

Parallel data scarcity problem is a major challenge faced by Statistical Machine Translation (SMT). The aim of this thesis is to enrich a SMT system by adding more morphological variants and new translation lexicon automatically generated out of monolingual data. To induce bilingual lexicon, instead of taking advantages of comparable corpora or parallel data, we proposed a supervised classifier trained using monolingual features (e.g. word embedding vectors, plus Brown clustering or word frequency information) of only a small amount of translation equivalent word pairs. The classifier is able to predict whether a new word pair is under a translation relation or not.

Our experiments of SMT phrase table expansion were conducted on Chinese and Spanish, since we realized that although they are two of the most widely spoken languages of the world, this language pair is suffering from a data scarcity situation. In addition to the problems caused by the words that are not included in the training corpus, the inflection differences between this language pair make the translation even more challenging when only limited parallel data are available.

The obtained results demonstrate that, on the one hand, with the method of morphology expansion, the SMT system achieves an improvement of up to + 0.61 BLEU compared to the results of a low resource Chinese-Spanish phrase-based SMT baseline. On the

other hand, our supervised classifier reaches a 0.94 F1-score and the SMT experiment results show an improvement of up to +0.70 BLEU with the resulting bilingual lexicon, demonstrating that the errors of the classifier are ultimately controlled by the SMT system.

Resumen

La escasez de datos paralelos es un problema importante para la Traducción Automática Estadística (TAE). El objetivo de esta tesis es enriquecer un sistema de TAE añadiendo más variantes morfológicas y un nuevo léxico de traducción generado automáticamente desde datos monolingües. Para inducir el léxico bilingüe, en lugar de depender de corpus comparables o de datos paralelos, proponemos un clasificador supervisado entrenado con representaciones monolingües (por ejemplo, vectores distribuidos, agrupaciones de Brown e información de la frecuencia de palabras) de sólo una pequeña cantidad de traducciones. El clasificador es capaz de predecir si un nuevo par de palabras es una traducción la una de la otra, o no.

Realizamos los experimentos para enriquecer el sistema de TAE con chino y español, porque a pesar de que estas lenguas son dos de las más habladas del mundo, este par de idiomas sufre de escasez de datos paralelos. Además de los problemas causados por las palabras que no están incluidas en el corpus de entrenamiento, las diferencias de flexión morfológica entre este par de idiomas hace que la traducción sea de peor calidad cuando se dispone de pocos recursos paralelos.

Los resultados obtenidos demuestran que, por un lado, con el método de expansión morfológica, el sistema de TAE logra una mejora de hasta + 0,61 BLEU en comparación con los resultados obtenidos con un sistema básico chino-español con poco corpus. Por otro lado, nuestro clasificador supervisado, que alcanza una F1 de 0,94, proporciona nuevos pares de traducción que resultan en una mejora de hasta +0,70 BLEU con respecto al sistema básico, demostrando que los errores del clasificador son, en último término, controlados por el sistema de TAE.

Resum

L'escassetat de dades paral·leles és un problem important per a la Traducció Automàtica Estadística (TAE). L'objectiu d'aquesta tesi és enriquir el sistema de TAE afegint més variants morfològiques i un nou lèxic de traducció generat automàticament des de dades monolingües. Per induir el lèxic bilingüe, en lloc de dependre de corpus comparables o de dades paral·leles addicionals, proposem un classificador supervisat entrenat amb representacions monolingües (per exemple, vectors distribuïts, agrupacions de Brown i informació de la freqüència de paraules) de només una petita quantitat de traduccions. El classificador és capaç de predir si un nou parell de paraules és un la traducció de l'altre, o no.

Hem realitzat experiments per enriquir un sistema de TAE entre xinès i espanyol, perquè tot i que aquestes llengües són dues de les més parlades del món, aquest parell d'idiomes està patint l'escassetat de dades paral·leles. A més dels problemes causats per les paraules que no estan al corpus d'entrenament, les diferències de flexió morfològica entre aquest parell d'idiomes fa que la traducció sigui de poca qualitat quan es disposa de pocs recursos paral·lels.

Els resultats obtinguts demostren que, per una banda amb el mètode d'expansió morfològica, el sistema de TAE aconsegueix una millora de fins a + 0,61 BLEU en comparació amb els resultats obtinguts per un sistema bàsic xinès-espanyol entrenat amb pocs recursos. Per altra banda, el nostre classificador supervisat, que assoleix una F1 de 0,94, proporciona nous parells de traducció que resulten en una millora de fins a + 0,70 BLEU respecte al sistema bàsic, demostrant que els errors del classificador són, en últim terme, controlats pel sistema de TAE.

Contents

List of Figures

List of Tables

Chapter 1

1. INTRODUCTION

Machine translation (MT) has been one of the most paradigmatic and valuable applications in natural language processing (NLP). Looking back to the seventeenth century, Leibniz and Descartes proposed a series of codes which would relate words between languages. However, all of these proposals remained theoretical, and none resulted in the development of actual machine translation systems in our time. From the late twentieth century, the availability of language resources and the improvement of the NLP technologies brought MT to a new scenario, making this research area inspiring and challenging. Besides the scientific research interest, the profitability of MT as a business has attracted more and more companies and institutions to exploit MT technologies, because nowadays factors like budgets, staffing and time consumption, always make organizations shy away from translating even a small fraction of the information they have at hand. To help with such situation, MT can offer significant improvements in translation efficiency and maximize the amount of information available to customers with different language backgrounds.

Although MT has achieved significant advancements in recent times, it is still challenging for computers to translate our human languages because of the following general reasons:

- Much of the difficulty of natural language processing in general, and MT in particular, arises from the ambiguity problem. The same word may have different meanings regarding different contexts. For instance, the Chinese word *水分* means 'water' in the sentence *健康的肌肤需要充足的* ***水分*** ('healthy skin requires plenty of water'), but in the sentence *这个报道有些****水分*** ('this report seems somewhat exaggerated'), it means 'exaggeration'. Since in both cases, the word *水分* is used as a noun, it is difficult to disambiguate them.

- Differences of word order, morphological richness level and grammatical structure between languages also pose problems for MT. For example, translation errors due to inflection frequently occur when translating from Chinese to Spanish, since Chinese nouns and adjectives have no gender and number, and adjectives typically come before the nouns they modify, but in Spanish it is the other way around.

- Language is highly variable with respect to several dimensions: style, genre, domain and topics, etc. Even apparently small differences in domain might result in significant deviations in the underlying statistical models. For example, current technologies require large language samples for statistical machine translation (SMT), linguistic variability would indeed suggest to consider many alternative data sources as well (Bertoldi and Federico, 2009).

- Machine can not perform common-sense reasoning which involves literally millions of facts about the world. The task of coding up the vast amount of knowledge required is daunting (Arnold, 2003). In practice, most of what we understand by “common-sense reasoning” is far beyond the reach of modern algorithms. For instance, to translate the sentence *I will bring my bike tomorrow if it looks nice in the morning*, one must be aware that *it* refers to the weather because people usually ride bikes when the weather is nice and bikes do not usually change their aspect in the mornings (Sánchez-Cartagena, 2015).

At the present stage, the main MT technologies are: rule-based machine translation (RBMT), which is based on explicit linguistic knowledge; statistical machine translation (SMT), which learns translation correspondences from parallel text corpora; and neural machine translation (NMT), a new approach to MT in which a large neural network is trained to maximize the translation performance.

In this work, we mainly focussed on the statistical approach because it is still the most widely used technique in MT. This conclusion is based on two main issues: (1) in comparison with rule-based systems, development effort for SMT is dramatically reduced and, in general terms, the achieved translation quality is considered to be of acceptable quality; (2) NMT is still in its infancy and is computationally expensive both in training and in translation inference, thus, a lot of effort remains to be done for transforming it into a truly robust platform.

1.1 Key issues in SMT from Chinese to Spanish

Chinese has often been described as ideographic or pictographic language, that is, consisting of graphic symbols rather than letters which represent an idea or concept, and its writing system does not reflect pronunciation (Casas-Tost and Rovira-Esteva, 2014). It has its own non-alphabetic symbols, known as orthographic characters. Its writing system contains tens of thousands of characters. Unlike letters in alphabetic writing systems of western languages, Chinese characters are not arranged in a linear fashion. Each character corresponds to a single syllable that is also the smallest meaning-bearing unit — morpheme. DeFrancis (1984) claimed that Chinese writing is morphosyllabic. However, it should be made clear that characters can not be equated with the concept of ‘word’ in Chinese. In Chinese writing system, there are two relevant linguistic units — morpheme (词素) and word (词):

- **Morpheme**

Similar to western languages, the smallest meaningful elements are named as morphemes as well in Chinese, and they can be combined to form words. A morpheme can not be further analysed into smaller parts and still remain meaningful. In fact, the vast majority of Chinese morphemes have a lexical nature. They have certain semantic properties, but can also be bound in some cases. For instance, the character *火* means ‘fire’ in Chinese, and it can be treated as a word of only one morpheme, but in other cases, it can also be combined with other characters to form new words, such as *火柴*（‘matches’）, *火焰*（‘flame’）and *火龙果* (‘Pitaya’). Arcodia (2007) compared this type of Chinese morphemes to the “neoclassical constituents” of Standard Average European languages (henceforth, SAE), such as *philo-*, -*logy* or *–phobia*, having lexical (rather than grammatical) meaning and always bound to some other constituent. The difference is that most of these "Chinese bound morphemes" can be used independently as well. Besides the morphemes with semantic properties, there is another type of morphemes which play an important role in syntactic function, but barely has semantic significance and can not be used independently in a sentence, such as 了, 化 and 的. For instance, in the two sentences shown below, the extra 了 in (2) changes the tense of the sentence (1):

(1) 他 去 超市。 → He goes to the supermarket.
(2) 他 去 超市 了。 → He went to the supermarket.

- **Word**

Packard (2000) defined word as the smallest syntactically free form that can stand as an independent occupant of a syntactic form class slot. For instance, in English, *eat* and *for* both are words since they normally occupy a verb slot and a preposition slot in sentence respectively, and they both have their own semantic and syntactic significance. Although some morphemes (free morphemes) can occur alone in a sentence, the morpheme units are not identical to word. The principal difference is that a morpheme may or may not stand alone, whereas a word is freestanding.

When dealing with Chinese in natural language processing, the notion of word is hard to determine. In western languages, words are mostly well defined since there are spaces between them in a sentence, whereas in Chinese there are no delimiters between characters to indicate word boundary (Chang, 2009). This characteristic brings an extra challenge for SMT systems that involve Chinese. The most obvious obstacle is the segmentation error. For instance, the Chinese phrase 最胖的和尚未吃饭的 is segmented in both ways by different segmenters:

(3) 最胖/的/和尚/未/吃饭/的。

(The fattest monk has not eaten.)

(4) 最胖/的/和/尚未/吃饭/的。

(The fattest one and the ones that have not eaten)

However, in case (3), the Spanish translation is “El monje más gordo todavía no ha comido”, while in case (4), the phrase is translated as “los más gordos y los que no han comido”.

Besides the segmentation problem, another important challenge of Chinese is that Chinese has little or no morphological complexity within a word or in grammatical relations (Li and Thompson, 1981). Specifically, it has no case, gender or number markers for nouns and no subject-verb agreement or tense markers for verbs. Chinese grammatical relationships are expressed either by word order or by the use of independent grammatical particles (Norman, 1988). Due to the non-inflectional property of Chinese, the SMT from Chinese to Spanish becomes more difficult. We show several examples below to demonstrate the morphological differences between Chinese and Spanish.

- **No tense**

In Chinese, verb behaves in a unique word form for different time references. Normally, tense of a sentence can be determined by time markers such as time adverbials. Observing the Chinese example in (5) and (6), the verb 住 in both sentences is the same, while in Spanish, the verb *vive* changes from the present tense to the past tense (*vivió*).

(5) CH: 她 (She) 现在 (now) 住 (lives) 在 (in) 巴塞罗那 (Barcelona)。

ES: Ella ahora **vive** en Barcelona.

(6) CH: 她 (She) 去年 (last year) **住** (lived) 在 (in) 巴塞罗那 (Barcelona)。

ES: Ella **vivió** en Barcelona el año pasado.

- **No subject-verb agreement**

Besides tense, in Spanish, verb word forms also vary depending on the person of its related subject. Chinese, by contrast, has no such agreement. In (7) and (8), the same verb *querer* behaves conjugated in different forms (*quiero* and *quieren*) according to the subject, while in Chinese the verb 想 does not change.

(7) CH: **我**(I) **想** (want to) 跳舞(dance)。

ES: **Yo quiero** bailar.

(8) CH: 他们 (They) 想 (want to) 跳舞 (dance)。

ES: **Ellos quieren** bailar.

- **No number and gender marking**

Chinese nouns have no distinction between singular and plural, while in Spanish plural nouns normally end with *-s* /*-es*. Thus, sometimes in a Chinese sentence, number may not be clearly defined as shown in (9) and (10). Besides number, Chinese has no gender as well, whereas in Spanish, the gender of an adjective should be consistent with the noun that it modifies as shown in (11) and (12).

(9) CH: 他 (He) 想要 (wants) **我的** (my) **书** (book/books)。

ES: El quiere **mi libro**.

(10) CH: 他 (He) 想要 (wants) **我的** (my) **书** (book/books)。

ES: El quiere **mis libros**.

(11) CH: **美味的**(delicious)意面(pasta)。

ES: pasta **deliciosa**.

(12) CH: **美味的**(delicious) 佳肴 (dish)。

ES: plato **delicioso**.

1.2 Problems addressed in SMT with reduced parallel corpora

The performance of SMT systems depends on the size of available training data. The more parallel texts are used to train the models, the better the system can approximate the final translation probability among a large enough number of expressions. However, large enough parallel corpora are not that easy to gather and for some language pairs, it is not even possible. As a result of this challenge, research on statistical machine translation with limited-size parallel corpus is receiving more and more attention.

The most common problem of SMT is Out-of-Vocabulary (OOV) words, since with a reduced parallel corpus, the expressions that do not occur in training data will be missing causing errors when translating new documents. For instance, in our experiment, some of the source OOVs were directly output to the final translation as shown in (13); in other cases, our low resource SMT baseline just ignored the expressions that could not be translated as shown in (14):

(13)

Source: *文化多样性*

Reference: *diversidad cultural*

Low resource SMT baseline: *多样性. A la cultura*

(14)

Source: *负面影响*

Reference: consecuencias negativas

Low resource SMT baseline: *la negativo*

In addition to OOV, language pairs with inflection differences are more challenging when very limited parallel data are available, since isolating languages like Chinese and Vietnamese, never use inflections while synthetic languages like Greek and Spanish can be highly inflected. For instance, in the Chinese-Spanish case, *我今天买了一本很棒的书* is translated as *hoy compré un libro fantástico*. In the Chinese sentence, *买了* (*have bought*), *一本* (*a*), *很棒的* (*fantastic*) and *书* (*book*), all the words do not have inflectional morphology. Unlike in Spanish, the verb *compré* (third-person/singular/pretérito), the quantifier *un* (masculine/ singular), the noun *libro* (singular/masculine) and the adjective *fantástico* (masculine/singular), all have an appropriate grammatical form corresponding to the context. So when translating from a non-inflected language to a highly inflected language, if the correct word form is not included in the training data, the system can not produce the correct translation result as shown in the following example (15) from our experiments:

(15)

Source: 当他看到我时。

Reference: *cuando me vio.*

Low resource SMT baseline: *cuando lo vi*.

1.3 Objectives

This thesis addresses the parallel data shortage problem in SMT. Our main goal was to devise methods to supplement available parallel data with new translation pairs automatically generated out of monolingual resources. Although these automatically generated resources might contain wrong translations, the SMT is expected to handle them successfully, that is, eventually using the good translation pairs and discarding the wrong ones, and therefore improving the quality of the output.

Along this line, the objectives pillaring our work were the following:

- To assess the quality improvements of a SMT trained with limited resources when introducing morphological variants of target language words occurring in the translation table.

- To enlarge the coverage of the SMT system by adding new translation pairs which are automatically induced from monolingual corpora.

- To investigate whether a supervised classifier trained with features extracted from monolingual corpora can provide such new translation pairs.

- And to investigate whether a SMT can handle the output of the classifier by eventually pruning bad translation pairs.

Our experiments were conducted on Chinese and Spanish, since we realize that although they are two of the most widely spoken languages of the world, as far as we know, this language pair is currently suffering a data scarcity situation (Costa-jussa et al., 2012a), and there has not been much investigation done on machine translation for this pair.

1.4 Thesis Outline

The rest of this thesis is structured as follows:

Chapter 2: State of the art addresses the research frameworks that underpin our research and reviews the major techniques developed for phrase table expansion of SMT with limited parallel resources. Section ***Machine Translation approaches*** describes a classification of machine translation methods, and Section ***Statistical Machine translation (SMT) phrase table expansion methods*** focuses on the related works that aim at enriching the SMT phrase table by adding more morphological variants and monolingually-derived bilingual lexicons.

Chapter 3: Morphology Expansion for SMT describes an experiment to enrich the SMT phrase table by adding more inflected forms that are derived from a lexical resource. The goal is to provide more word forms as translation options to the system. Different from the bilingual lexicon induction method described in Chapter 4, the morphological expansion is based on the entries already occurring in a baseline system phrase table. At the end of the section, we carry out the analysis of the experiment results and discuss its advantages and limitations.

Chapter 4: Supervised bilingual lexicon induction for SMT presents and discusses several approaches to word representation for a supervised classifier to induce new translation word pairs from monolingual corpora. Section ***New translation lexicon generation with word embeddings (WE)*** describes a classifier trained with WE vectors of a small amount of translation equivalents. Based on the experimental results, we discuss the limitations and possible drawbacks of our WE-based approach. In Section ***Improving WE-based classifier with additional word frequency information***, we added word frequency information, which was also learned from monolingual corpora, to the word embedding vectors for improving the performance of the classifier. In Section ***Improving WE-based classifier with additional Brown clustering***, instead of word frequency, we incorporated a Brown clustering representation to the WE vectors for alleviating the identified limitations of the basic classifier. The obtained results demonstrate that both word frequency and Brown cluster features positively affect the performance of WE-based classifier. Section ***SMT phrase table expansion***

using induced bilingual lexicons describes the application of the classifier to induce translation lexicons from monolingual corpora for enriching the SMT phrase table. We carried out an error analysis based on the translation results produced by the classifier and also discuss the advantages and limitations of this method.

Chapter 5: Conclusions and future works section draws the main conclusions and contributions of this dissertation and discusses different directions for future works.

Chapter 2

2. STATE OF THE ART

In this chapter, we explain in general the most commonly used MT approaches and the previous methods that related to our work.

2.1 Machine Translation Approaches

On a basic level, it can be said that MT performs simple substitution of words in one natural language for words in another (Albat, 2012). There are different paradigms in MT, we visualize them in Figure 2.1. In the rest part of this section, we present in details the most commonly used ones: Rule-based MT, Statistical MT and Neural MT.

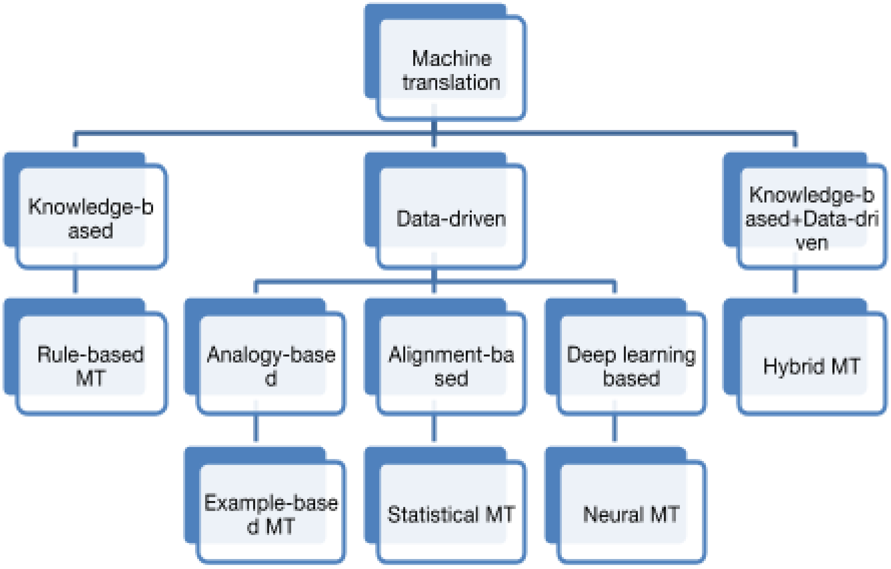


Figure 2.1: Machine translation approaches

2.1.1 Rule-based Machine Translation

Rule-Based Machine Translation (RBMT) refers to the MT systems based on linguistic rules retrieved from bilingual dictionaries and grammars covering the main semantic, morphological, and syntactic regularities of source and target language. Given source input sentences, an RBMT system generates them to target output sentences on the basis of morphological, syntactic, and semantic analysis of both the source and the target languages involved in a concrete translation task (Okpor, 2014). The architecture of a rule-based machine translation system can be observed through the

Vauquois Triangle, which illustrates different levels of analysis (shown in Figure 2.2).

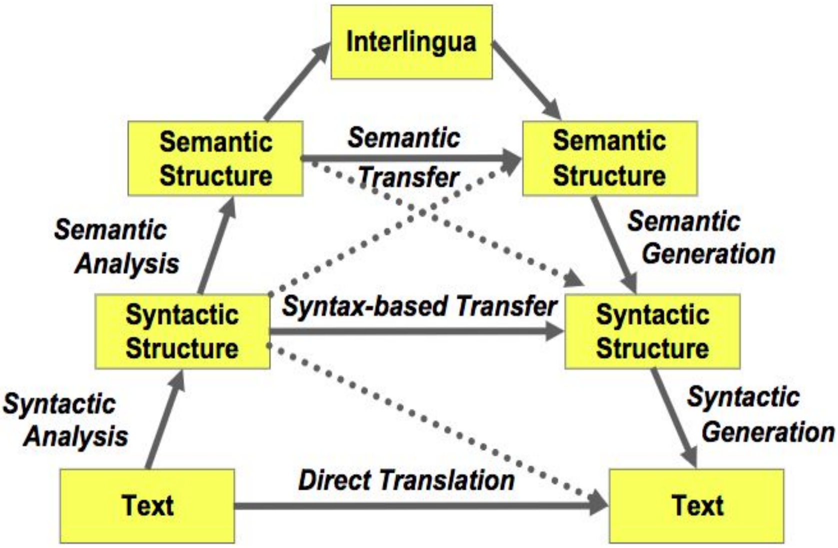


Figure 2.2: Vauquois Triangle(Vauquois, 1976)

So what are the advantages of RBMT? One of its main merits is that RBMT is totally based on linguistic theories so that systems can achieve high translation accuracy with reduced resources, but without requiring expensive computation processing. Besides, RBMT is better suited to post-editing and durable changes, hence the translation performance is predictable and can be well controlled.

However, RBMT entails a huge human effort as well as long time for preparing a large amount of linguistics rules. The following are the shortcomings that are associated with RBMT approach (Okpor, 2014):

- It is difficult and expensive to gather sufficient amount of high quality dictionaries and linguistic transfer rules.
- It is hard to deal with rule interactions in big systems, ambiguity, and idiomatic expressions.
- It is hard to adapt to new domains. Although RBMT systems provide mechanisms to extend and adapt new lexicon, changes are usually very costly and the results, frequently, do not pay off.

Comparing RBMT with SMT from linguistic perspective, Costa-Jussa et al. (2012b) demonstrated that, orthographic and morphological errors tend to be lower in the rule-based machine translation systems, while the performance at the semantic level is better in the statistical systems.

2.1.2 Neural Machine Translation

Neural machine translation is an end-to-end approach which aims at building a single neural network that can be jointly tuned to maximize the translation performance (Bahdanau et al., 2014). Most of the neural network machine translation models (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014a) consist of an encoder and a decoder. The encoder extracts a fixed-length vector representation from a variable-length input sentence, and from this representation the decoder generates a correct, variable-length target translation (Cho et al., 2014b). The architecture of a simplified diagram of NMT system (See et al., 2016) is described below in Figure 2.3.

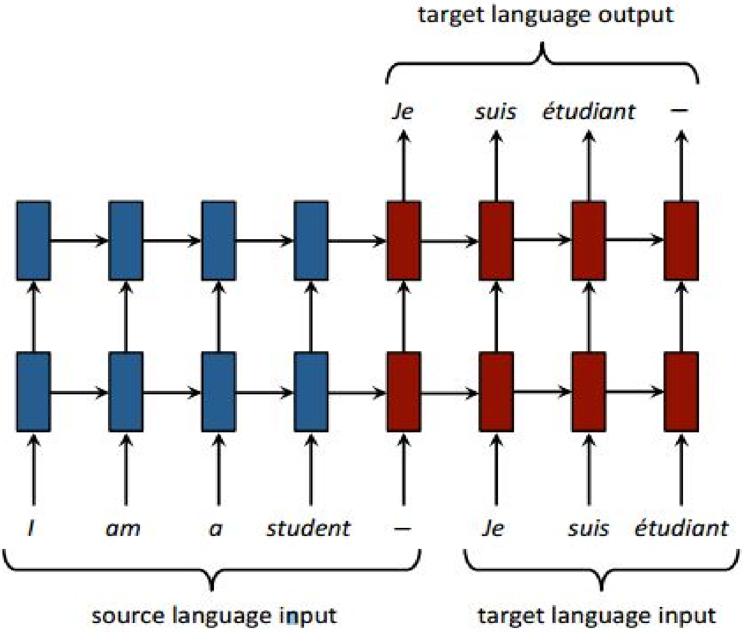


Figure 2.3: A simplified diagram of neural machine translation (See et al., 2016).

NMT has been popular nowadays because it sidesteps many brittle design limitations of traditional machine translation methods. Its main merits over the most widely used phrase-based SMT are (Jean et al., 2015):

- NMT requires a minimal set of domain knowledge and does not assume any linguistic property in both source and target sentences except that they are sequences of words.
- The whole NMT system is jointly trained to maximize the translation performance, unlike the existing phrase-based systems which consist of many separately trained features whose weights are then tuned jointly.
- Memory footprint of the NMT model is often much smaller than SMT systems which rely on maintaining large tables of phrase pairs.

The comparative case study with SMT and NMT carried out by Bentivogli et al. (2016) showed that the NMT system significantly reduce translation problems in some linguistic phenomena (morphological error, lexical errors and word order errors) compared to SMT systems.

However, in practice, NMT is still in its infancy compared to SMT, especially when training on very large-scale datasets as used for the very best publicly available translation systems. The main weaknesses of Neural Machine Translation are (Wu et al., 2016):

- A considerable amount of time and computational resources are needed for the training on a large-scale translation dataset. For inference they are generally much slower than phrase-based systems due to the large number of parameters used.

- NMT lacks robustness in translating rare words.

- NMT systems sometimes produce output sentences that do not translate all parts of the input sentence – in other words, they fail to completely "cover" the input, which can result in surprising translations.

2.1.3 Statistical Machine Translation

Unlike other MT approaches, instead of providing answers to questions of what representations to use and what steps to perform to translate, the goal of statistical machine translation (SMT) is to produce translation by constructing probabilistic models of adequacy and fluency, then combining these models to select the "highest scored" translation result. According to the first definition of SMT due to Brown at al. (1993), the translation process from a source language sentence $f_1^J = f_1, ..., f_J$ to a target language sentence $e_1^I = e_1, ..., e_I$ can be modeled by applying the Bayes rule as:

$$P(e_1^I | f_1^J) = \frac{P(e_1^I) \cdot P(f_1^J | e_1^I)}{P(f_1^J)} \qquad (1)$$

Note that the denominator $P(f_1^J)$ can be ignored since the intuition is to choose the best target sentence for a fixed source sentence f_1^J, hence $P(f_1^J)$ is a constant. So the final translation can be described as:

$$\widehat{e}_1^{\widehat{I}} = argmax_{I,\, e_1^I} \left\{ P(e_1^I) \cdot P(f_1^J | e_1^I) \right\} \qquad (2)$$

In regard to (2), there are three principal components in SMT:

- A language model to compute $P(e_1^I)$.
- A translation model to compute $P(f_1^J | e_1^I)$.
- A decoder to produce the most probable translation $\widehat{e}_1^I$.

As an alternative to the classical source-channel approach (Brown et al. 1993) is to directly compute the posterior probability $P(e_1^I | f_1^J)$ by using a log-linear combination of different models

(Papineni et al., 1998; Och and Ney, 2002). It has the advantage that additional features can be easily added into the overall system. In this framework, there are a set of feature functions $h_m(e_1^I | f_1^J)$, m=1,..., M . For each feature function, there exists a model parameter λ_m , m=1,..., M.

$$P(e_1^I | f_1^J) = p_{\lambda_1^M}(e_1^I | f_1^J)$$

$$= \frac{exp\left[\Sigma_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right]}{\Sigma_{e'^I_1} exp\left[\Sigma_{m=1}^M \lambda_m h_m(e'^I_1, f_1^J)\right]} \qquad (3)$$

Since the denominator represents a normalization factor that depends only on the source sentence, we can ignore it during the search process. Thus the result of a linear combination is computed as:

$$\widehat{e}_1^I = argmax_{e_1^I} \left\{P(e_1^I | f_1^J)\right\}$$

$$= argmax_{e_1^I} \left\{\Sigma_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right\} \qquad (4)$$

The architecture of a statistical machine translation system is described below in Figure 2.4. Beside the language model and translation model, two additional components can also be included in the process: the preprocessing step performs a pre-editing of the input text, adapting it to the translation system; and the post-processing modifies the translation result, making it more acceptable for humans. For the statistical approach, the more data are used to train the models, the better the result will be. However, bilingual corpora large enough to build competitive SMT systems are not always easy to gather.

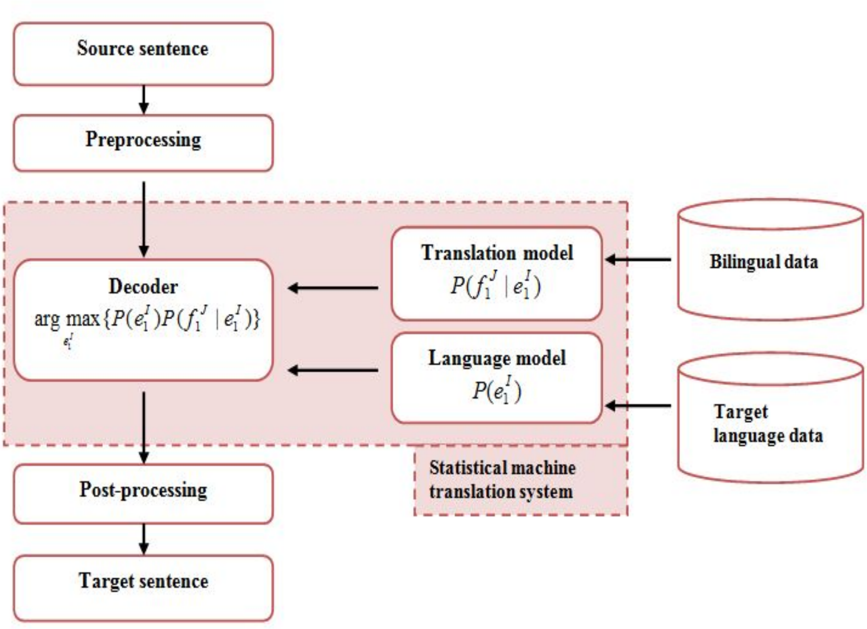


Figure 2.4: Illustration of the generation process of a statistical machine translation.

- **Main statistical machine translation approaches**

Statistical machine translation approaches are quite diverse, and can be roughly classified along several axes, including: word-based translation, phrase-based translation, syntax-based approach and hierarchical phrase-based approach. Word-based machine translation only focuses on the translation of lexical level (Brown et al. 1993). Simple word-based system is limited to translate consecutive phrases, because it could map a single word to multiple words, but not in reverse. Moreover, compared to the other approaches of SMT, word-based model does not take contextual information into account for the translation decisions, so word-based approach is not widely used today.

While word-based SMT estimates translation probabilities by only considering how each individual word is translated, Phrase-based SMT is based on the intuition that a better way to calculate translation probability is by taking into account phrases. Therefore, the phrase-based models considers phrases as atomic units. Note that the phrases in phrase-based systems are not linguistic phrases as in constituents of parse trees, but are chunks or strings of words that are identified statistically from symmetrized word alignments in bi-texts (Och and Ney, 2003). There are mainly three steps in its generation procedure: 1) segmenting foreign input in phrases; 2) translating each source phrase into a target phrase; 3) reordering each of the target phrases. Compared to the word-based models, memorizing larger units enables the phrase-based model to handle many-to-one translation pairs, and also to take into account local context information in translation. So it is more common nowadays. However, this model also has some limitations. For instance, it has issues handling long-distance reordering; it also has the spurious phrasal segmentation problem which allows multiple derivations of a bilingual sentence pair having different model scores for each segmentation (Durrani et al., 2013).

In addition to the previously described approaches, syntax-based models are based on the idea of translating syntactic units, rather than single words or strings of words. The goal of this type of model is to incorporate an explicit representation of syntax into the statistical systems in order to improve the translation between language pairs with very different structure order. However, the problem with syntax-based model is the cost of decoding, which is mostly modeled as a parsing problem (Ahmed and Hanneman, 2005). It relies on a good parser, but a good parser is not available in all languages, especially not in resource-poor languages.

Hierarchical phrase-based machine translation combines fundamental ideas from both syntax-based and phrase-based approaches. This model uses phrases as basic units for translation and applies synchronous context-free grammars (SCFG) as rules.

But the weakness of this model is that although the model captures global reordering by SCFG, it does not explicitly introduce reordering model to constrain word order (Hayashi et al., 2010).

- **Phrase-based statistical machine translation**

In the present work, we focus on the phrase-based statistical machine translation (PBSMT) (Koehn et al. 2003), because it is one of the most widely used approaches in statistical machine translation. The performance of a PBSMT mostly depends on the induction of a good phrase translation table. There are different ways to acquire such a table. The most common methods is to create a word alignment between each sentence pair of the parallel corpus and then extract phrase pairs that are consistent with this word alignment as we describe below.

- **Word Alignment**

Word alignment is a mapping between the source words and the target words estimated from parallel corpus. A set of alignment algorithms are applied in the literature such as IBM Models and HMM Model. For training the alignment model, expectation maximization (EM) algorithm is applied to search for the maximum likelihood in the models. The intuition of EM training is that in the E-step, expected counts are computed for the *t* parameter based on summing over the hidden variable (the alignment), while in the

M-step, the maximum likelihood estimate of the *t* probability from these counts is calculated. Therefore the procedure is (Jurafsky, 2009):

E-step 1: Compute the expected counts $E\,[count\,(t\,(f \mid e))]$ for all word pairs $(f_j,\ e_{a_j})$.

E-step 1a: First calculate $P\ (a,\ f\ \mid\ e)$ by multiplying all the t probabilities as following:

$$P(A,F \mid E) = \prod_{j=1}^{J} t(f_j \mid e_{a_j})$$

E-step 1b: Normalize $P(a,\ f \mid e\ )$ to get $P(\ a \mid e,\ f\ )$, using:

$$P(a,\ f \mid e\ ) = \frac{P(a \mid e, f\ )}{\Sigma_a P(a, f \mid e\ )}$$

E-step 1c: Compute expected (fractional) counts, by weighting each count by $P(a \mid e,\ f)$.

M-step 1: Compute the MLE probability parameters by normalizing the t count to sum to 1.

E-step 2a: recalculate $P(a,\ f \mid e)$ again by multiplying the *t* probabilities, by following:

$$P(A,F|E)=\prod_{j=1}^{J}t(f_j|e_{a_j})$$

This procedure shows how EM is used to learn the parameters for a simplified version of model1. It is also applied in the form of the Baum-Welch algorithm, for learning the parameters of the HMM model. However, these models only allow that each source word to be aligned with at most one target language word. To overcome this limitation, symmetrization is performed by applying a heuristic post-processing step that combines the alignments in both translation directions. It starts with the intersection of the two alignments and then adds neighboring alignment points from the union and unaligned points to the intersection. In Figure 2.5 we show an alignment examples of symmetrization heuristic (Koehn, 2010)

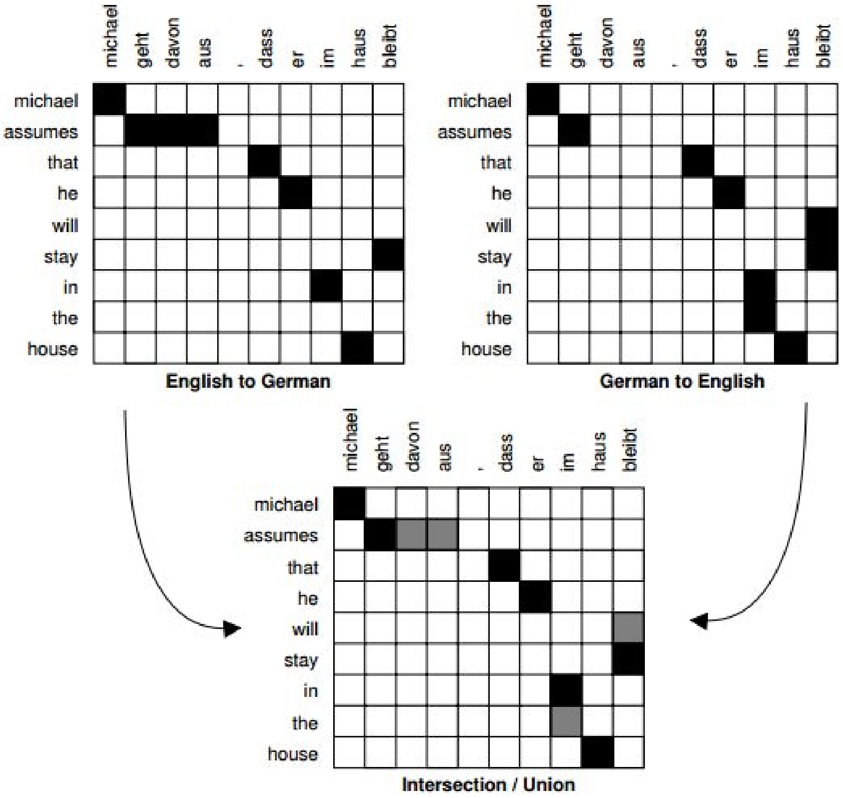


Figure 2.5: Alignment example of symmetrization heuristic(Koehn, 2010).

- **Learning phrase translation pairs**

Once all the alignments are collected, the next step is to extract phrase translation pairs based on the symmetrized alignment. Phrase pairs extracted from a parallel corpus need to be consistent with obtained word alignment matrix. Two phrases are considered to be translations of each other if the words are aligned only to each other, and not to words outside. Different alignment templates were proposed for learning phrase translations. For instance, Marcu and Wong (2002) introduced a phrase-based joint probability model that simultaneously generates both the source and target sentences in a parallel corpus. Och and Ney (2003) proposed a heuristic approach to refine the alignments obtained from GIZA++. At a minimum, all alignment points of the intersection of the two alignments are maintained. At a maximum, the points of the union of the two alignments are considered. Similarly, the method of Koehn et al. (2003) starts with intersection of the two word alignments, but only new alignment points that exist in the union of two word alignments are added. The new alignment point is required to be connected with at least one previously unaligned word. Following Koehn et al. (2003), in Figure 2.6, we show all possible phrase pairs extracted based on the symmetrized alignment example in Figure 2.5. The phrase translation probability is estimated by the relative frequency of phrases pairs:

$$\phi(\bar{f} \mid \bar{e}) = \frac{count(\bar{e}, \bar{f})}{\Sigma_{f_i} count(\bar{e}, \bar{f}_i)}$$

The inverse phrase translation probability is computed in the same way. Besides the direct and inverse phrase translation probabilities, the state-of-the-art PBSMT systems (Koehn et al., 2007b) include the following features as well: direct and inverse lexical weighting(estimated by using the word-based IBM Model of each phrase pair); word penalty(refers to target translation hypothesis length. With this feature, we are able to adjust the sentence length); phrase penalty(depends on phrase length. This feature is set by the user to the same value ρ for each phrase, If ρ > e, longer phrases will be preferred over shorter ones. Conversely, if ρ < e, shorter phrases will be preferred); reordering probability (computed based on the distance from the end position of a phrase to the start position of the next phrase); probability obtained from language model (In addition to the distortion penalty, a standard n-gram language model is used for ensuring the fluency of target sentences).

```
                                     michael   |||   michael
                             michael assumes   |||   michael geht davon aus
                             michael assumes   |||   michael geht davon aus ,
                        michael assumes that   |||   michael geht davon aus , dass
                     michael assumes that he   |||   michael geht davon aus , dass er
michael assumes that he will stay in the house   |||   michael geht davon aus , dass er im haus bleibt
                                     assumes   |||   geht davon aus
                                     assumes   |||   geht davon aus ,
                                assumes that   |||   geht davon aus , dass
                             assumes that he   |||   geht davon aus, dass er
assumes that he will stay in the house ||| geht davon aus , dass er im haus bleibt
                                  that ||| dass
                                  that ||| , dass
                               that he ||| dass er
                               that he ||| , dass er
         that he will stay in the house ||| dass er im haus bleibt
         that he will stay in the house ||| , dass er im haus bleibt
                                    he ||| er
              he will stay in the house ||| er im haus bleibt
                             will stay ||| bleibt
                 will stay in the house ||| im haus bleibt
                                in the ||| im
                          in the house ||| im haus
                                 house ||| haus
```

Figure 2.6: Example of phrase pairs extraction based on the symmetrized alignment in Figure 2.5 (Koehn, 2010).

- Decoding

Decoder is the final component of a SMT system that produces the best target translation for a given foreign source sentence. According to the State-of-the-art PBSMT systems (Koehn et al., 2007), during the decoding, given an input string of words, a number of translation options learned from the phrase alignment matrix are applied for a beam search algorithm to find the best translation output. The search process starts from an initial state where no foreign input words are translated and no translation output words have been generated. Then new hypotheses are generated by extending the target translation output with a phrasal translation that covers some of the foreign input words not yet translated. The cost of the new hypothesis is computed by multiplying the cost of the original state with the translation, distortion and language model costs of the added phrasal translation. The final states in the search are hypotheses that cover all source words. The hypothesis with the lowest cost is selected as the best translation.

Using beam algorithm to search the best translation among all possible translation candidates could be problematic in practical situation since it is computationally expensive. To speed up the translation process, **pruning** is applied to filter out bad hypotheses based on their partial score. There are generally two approaches to pruning (Koehn, 2010):

1) **Histogram pruning** keeps a maximum number N of hypotheses in the stack. it is a simple way of limiting the beam size compared to other pruning strategies. The stack size N has direct relation to decoding

speed. According to this method, all stacks are filled and all translation options are applicable all the time. The downside is that, in some cases there is a big difference in score between the best and worst hypothesis in the stack, while in other cases they are close. So this method is inconsistent with regard to pruning out bad hypothese.

2) **Threshold pruning** proposes a fixed threshold *a,* by which if the score of a hypothesis is *a* times worse than the best one, it is pruned out. An advantage of threshold pruning is that it adjusts itself to the amount of ambiguity. However, its drawback is that there is no upper limit on the number of hypotheses. If many hypotheses have similar score, all these hypotheses are kept in the beam. Thus, the beam could get arbitrarily large. In practice, today's machine translation decoders use both histogram pruning and threshold pruning.

So in conclusion, SMT systems work by building statistical models from parallel data. The translation quality heavily depends on the availability of parallel training corpus. the more data is used to estimate the parameters of the system, the better translation performance can be obtained. However, for many language pairs, parallel corpora are difficult to obtain, hence most of the current researches focus on SMT in the absence of large parallel resources.

2.2 SMT Phrase table expansion methods

The use of monolingual resources to enrich translation model for SMT in low resource condition has been proposed by many researches. Basically two different methods are applied in the literature: (1) deriving new translation options (inflected forms and lexical variants) based on the existing knowledge of low resource SMT phrase table; (2) inducing new translation entries generated from monolingual data. In the present work, we enriched our low resource SMT system in both directions: morphological variant integration and bilingual lexicon induction. In the rest of this section, we review the main works related to SMT phrase table expansion approaches in the literature.

2.2.1. Morphology and paraphrase expansion

Recent researches have largely focused on translating from non-inflected or weakly inflected languages (e.g. Chinese, Vietnamese and English etc.) to rich morphology languages (e.g. Spanish, Greek and Arabic). In a study of translation quality for languages in the Europarl corpus, Avramidis and Koehn (2005) demonstrated that translating into morphologically richer languages is more difficult than translating from them.

In Figure 2.7, Avramidis and Koehn (2008) reported the error analysis of their English to Greek baseline system, which is similar (from morphologically poor language to high-inflected language) to our Chinese to Spanish language pair. From their error analysis and classification, we can tell that 43.7% of the errors are translations in incorrect word forms. Regarding this challenge, in this work, one of our objectives is to alleviate morphological variant translation problem.

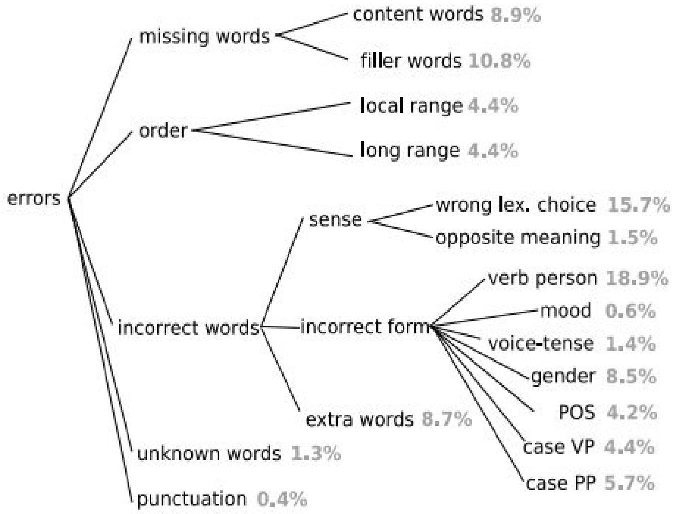


Figure 2.7: Error analysis on an English to Greek baseline system (Avramidis and Koehn, 2008).

There have been a number of related works done in this area. For instance, Habash (2008) analyzed the OOVs into lexeme and features. Though the morphological analysis, they matched the OOV word with those in-domain (phrases in the baseline phrase table) words which could be its possible morphological variants and added them into the baseline translation model. To do so, first they cluster all the single-word source entries in the baseline phrase table that (a) translate into the same target phrase and (b) have the same lexeme analysis. From these clusters they learned which morphological inflectional features in the source language word are irrelevant to the target language word. Then a set of morphological inflection rules were created to map OOV words with in-domain words. Phrases associated with the in-domain token in the phrase table are "recycled" to create new phrases in which the in-domain word is replaced with the OOV word. The translation weights of the in-domain phrase are inherited by the new phrase. In this method, Habash (2008) only considered the morphological expansion for OOV, which is limited for improving the translation performance on inflected variants when testing with new datasets.

Some other approaches (e.g. Durgar El-Kahlout and Oflazer, 2006; Habash, 2007; Koehn and Hoang 2007a; Birch et al., 2007 and Almaghout et al., 2010; Toutanova et al., 2008) applied morphological and syntactic features (e.g. POS tag, morph stems and Combinatorial Categorial Grammar supertags) on the source or target language for a better performance on translation from morphologically poor languages to rich morphology languages. For instance, Toutanova et al. (2008) developed a Maximum Entropy Markov model that predicts word forms from their stems using lexical and syntactic information from both the source and target languages. More specifically, morphological resources were used to analyse (into stem and morphological variants) the target words of phrase pairs from base MT system and generate an inflection set for each target word based on the training data (only inflections covered in training corpus were included). The task was designed as follows: given a source sentence, its translation is formed by (1) a sequence of stems in the target language, and (2) their corresponding inflection form (selected from the generated inflection set) which defined by the morpho-syntactic annotations (POS tags and word dependency structure) derived from the source input sentence. To combine the inflection prediction component with the base MT system, three different methods were tried:

- **Method 1**: The MT baseline was trained with normal parallel corpora (fully inflected word form), the inflection model is applied to re-inflect the 1-best or m-best translations and to select an output translation. The hypotheses in the m-best output from the base MT system are stemmed and the scores of the stemmed hypotheses are assumed to be equal to the scores of the original ones.

- **Method 2**: Word alignment is performed using fully inflected target language sentences. After the alignment, the target language is stemmed and the base MT systems' sub-models are trained using this stemmed input and alignment. In addition to this word-aligned corpus the MT systems use another product of word alignment: the IBM model 1 translation tables. Because the trained translation tables of IBM model 1 use fully inflected target words, the stemmed versions of the translation tables was generated by applying the rules of probability.

- **Method 3**: In this method the base MT system produces sequences of target stems. It is trained in the same way as the baseline MT system, except its input parallel training data are preprocessed to be stemmed. Then the corresponding inflection form was defined by the morpho-syntactic annotations derived from the source input sentence.

Another similar approach by Chahuneau et al. (2013) proposed to deal with the problem of lexical inflection translation was based on two phases: First, a discriminative model was learned to predict inflections of target words from rich source-side annotations; then, this model was used to create additional sentence-specific word/phrase-level translations that were further added to a standard translation model as “synthetic” phrases. To generate these synthetic phrases with new inflections, they created an additional phrase-based translation model using the training parallel corpus that was preprocessed to replace inflected surface words with their stems. Then, a set of non-gappy phrases for each sentence were extracted and the target side of each such phrase was re-inflected, conditioned on the source sentence, using the inflection discriminative model. The original features extracted for the stemmed phrase are conserved, and the following features are added to help the decoder select good synthetic phrases: (a) a binary feature indicating that the phrase is synthetic; (b) the log-probability of the inflected forms according to the model; and (c) the count of words that have been inflected, with a separate feature for each morphological category in the supervised case. A class-based n-gram language model was used to capture some basic agreement patterns. As results, with the class-based language model, they obtained an improvement of around 0.8 BLEU compared with the baseline. After adding their inflected synthetic phrases, an improvement of up to 0.7 BLEU was obtained compared with the model enhanced by the language model.

Reviewing the previous methods proposed by Toutanova et al. (2008) and Chahuneau et al. (2013), they both aimed at re-inflecting the existing phrase pairs of baseline phrase table based on the available training parallel data with the help of source and target linguistic resources. However, their methods do not consider those inflected variants for words that are not present in the training data, and this is the main problem for low resource SMT.

Turchi and Ehrmann (2011) proposed to expand the existing entries of a phrase-based SMT system using external morphological resources. According to this method, a set of new pairs made out of possible morphological variations of source and target phrases are created and added to the phrase table and reordering model learnt during the training process. Their method is based on the assessment of a similarity measure between the morphosyntactic tags of the bilingual pair candidates to be associated. The similarity score is afterwards used, together with the probabilities of the original association, to weight the probability of new associated phrases. This method significantly improved the translation quality of their baseline system. However, this is not a solution for pairs of languages that involve a non-inflected language. For instance, in our case, Chinese has no grammatical inflection, the similarity based on the morphosyntactic features can not be obtained.

Sánchez-Cartagena et al. (2011)[1] and Sánchez-Cartagena et al. (2015) proposed to enrich the phrase table of a PBSMT system with bilingual phrase pairs matching dictionary entries and transfer rules from the Apertium shallow-transfer MT system. To generate translation pairs from the bilingual dictionary, all the source language (SL) surface forms recognised by the shallow-transfer MT system and their corresponding SL intermediate representations (IR) are listed; then, these SL IRs are translated with the bilingual dictionary to obtain their corresponding target language (TL) IR; finally, the corresponding TL word forms are obtained by means of the RBMT generation module. If the TL IR contains missing values for morphological inflection attributes, a different TL phrase for each possible value of the attribute is generated. To expand the phrase table, they joined synthetic phrase pairs and corpus-extracted phrase pairs and calculate the phrase translation probabilities by means of relative frequency as usual. The phrase translation probabilities of the resulting phrase table are therefore computed as follows (for both directions):

[1] svn://svn.code.sf.net/p//apertium/svn/trunk/apertium-transfer-tools

$$\phi(t \mid s) = \frac{count_{corpus(s,\, t\,)} + count_{synth(s,\, t\,)}}{\Sigma_{t_i}(count_{corpus\,(s,\, t_i)} + count_{synth(s,\, t_i)})}$$

Lexical translation model was obtained from the concatenation of the training parallel corpus and the synthetic phrase pairs generated from the RBMT bilingual dictionary. The lexical weighting scores are then computed using the word alignments obtained by statistical methods. However, this technique heavily relies on the available morphological analyzers and bilingual dictionaries from the Apertium shallow transfer rule-based MT platform. The data of some language pairs may be too limited to support the expansion of phrase table. For instance, the bilingual dictionary of Spanish –German only contains 615 words; and the pair English-Albanian contains even less, 581 words.

The main differences between approaches described above and our methods are:

- **Morphological inflection generation**. Unlike previous approaches, instead of only depending on available morphological resource or RBMT, in our case, inflected variant generation involves the following methods: (a) A morphological resource (includes stem and inflected form information) on the target side was used to return all possible inflected forms of target words based on the translation rules of existing phrase table. Thus a set of derived translation pairs that contain new inflections were generated. (2) Our supervised approach for bilingual lexicon induction described in Chapter 4 is also effective for introducing translations with inflectional variants. Combining both strategies, a lot of inflectional variant pairs that do not appear in training data can be covered during that training process.

- **Integration of new inflection pairs**. Unlike previous approaches which integrated new translation rules by manipulating the baseline phrase table or creating new phrase table with externally invented features, we directly appended our new translation rules into the parallel training data. Schwenk et al. (2009) claimed that quality of the alignments obtained can be improved when the words in the bilingual dictionary appear in other sentences of the parallel corpus. Besides, the lexical weighting of existing aligned word pairs can be re-distributed and the phrase extraction algorithm may split the resulting bilingual phrase pairs into smaller units (Sánchez-Cartagena et al., 2015). The experimental results show that, within the newly acquired translation options, language model can help to select the appropriate inflection form in most of the cases.

2.2.2 Bilingual lexicon induction

Using bilingual lexicon induced from monolingual data to enrich translation model of SMT has been proposed by many researches. One approach was to explore semantically related translation candidates from monolingual data by applying paraphrasing techniques. For instance, Marton (2009) proposed augmenting the training data with word paraphrases generated by using distributional techniques on a large monolingual corpus. His system constructed monolingual distributional profiles of OOV words in the source language for the translation model. It then generated paraphrase candidates from phrases that co-occur in similar contexts, and estimated their semantic similarity to the paraphrased term by applying distributional semantic distance measures. An improvement of 0.7 BLEU score was obtained compared to their Spanish-English baseline. Marton et al. (2010) extended the word paraphrasing method by combining a distributional semantic distance measure with a shallow linguistic resource to create a hybrid semantic distance measure for handling not only words but phrases too. However, one potential disadvantage of relying on distributional profile is that items that are distributionally similar may not necessarily end up being paraphrastic. For instance, elements of the pairs (boys-girls), (cats-dogs), (high-low) can occur in similar contexts but are not semantically equivalent (Madnani and Dorr, 2010).

Another line of work exploited graph propagation-based methods to generate new translations for unknown words. For instance, Razmara et al. (2013) proposed to induce lexicons by constructing a graph on source language monolingual text. According to this method, nodes that have related meanings were connected together and nodes for which they have translations in the phrase table were annotated with target side translations and their feature values. A graph propagation algorithm was then used to propagate translations from labeled nodes to unlabeled nodes (OOV). Their approach differs from previous approaches by adopting a graph propagation approach that takes into account not only one-step (from OOV directly to a source language phrase that has a translation) but multi-step paraphrases from OOV source language words to other source language phrases and eventually to target language translations. They obtained an increase of up to 0.46 BLEU compared to the French-English baseline.

Saluja et al. (2014) presented a semi-supervised graph-based approach for generating new translation rules that leverages bilingual and monolingual data. The proposed technique first constructed phrase graphs using both source and target language monolingual corpora together with the baseline phrase table. Next, graph propagation identified translations of phrases that were not observed in the bilingual corpus, assuming that similar phrases have similar translations. This approach significantly improved the performance over an Arabic-English and an Urdu-English phrase based systems, leading to consistent improvements between 1 and 4 BLEU points on standard evaluation sets. Unlike Razmara et al. (2013), they used higher order n-grams instead of restricting to unigrams, since this approach was expected to go beyond OOV mitigation and could enrich the entire translation model by using evidence from monolingual text. Nevertheless, this method relies on pairwise mutual information between any pair of phrases in the monolingual corpus, which is very expensive to compute, even for moderately sized corpora.

Over the years, there has been a surge of interest in learning bilingual lexicon from monolingual data. The first work in this area by Rapp (1995) was based on the hypothesis that translation equivalents in two languages have similar distributions or co-occurrence patterns. Following this idea, (Koehn and Knight, 2002; Haghighi et al., 2008; Schafer and Yarowsky, 2002; Irvine and Callison-Burch, 2013) combined context information and other monolingual features (e.g., relative frequency and orthographic substrings,etc.) of source and target language words to learn translation pairs from monolingual corpora. For instance, Irvine and Callison-Burch (2013) used a log-linear classifier trained on various signals of translation equivalence (e.g., contextual similarity, temporal similarity, orthographic similarity and topic similarity) to induce word translation pairs from monolingual corpora.

Besides previously described approaches, distributional semantic models (DSMs) have also been used as representations to induce translation lexicons.

In general, DSMs can be described as follows:

- **Traditional distributional semantics models**

  Within the framework of DSMs, Turney and Pantel (2010) suggested to classify traditional DSMs into three different subclasses based on the structure of the matrix in a vector space model:

  - **Term-document matrix**. When dealing with a large collection of documents, term-document matrix is used to describe the frequency of terms that occur in a collection of documents. The row vectors of the matrix correspond to terms and the column vectors correspond to documents. One of the most common usage of this matrix is term frequency—inverse document frequency, which is applied in information retrieval.

- **Word-context matrix**. Traditional DSMs of this class are based on the assumption that the meaning of a word can be inferred from its distribution in text and that words appearing in similar contexts tend to have similar meanings (Harris, 1954). This has given rise to many word representation methods in the NLP literature. Most of these methods can be seen as a matrix *M* in which each row *i* corresponds to a word, each column *j* to a context in which the word appeared, and each matrix entry M_{ij} corresponds to some association measure between the word and the context. Words are then represented as rows in *M* or in a dimensionality-reduced matrix based on *M* (Levy and Goldberg, 2014b). These strategies are frequently applied to measure the relatedness between words in many natural language processing tasks, such as word sense disambiguation, paraphrasing techniques and thesaurus compilation etc.

- **Pair-pattern matrix**. In a pair–pattern matrix, row vectors correspond to pairs of words, and column vectors correspond to the patterns in which the pairs co-occur, such as “X cuts Y ” and “X works with Y”. These strategies are normally used for measuring the semantic similarity of patterns and the relations between word pairs (Turney and Pantel, 2010).

- **Word embedding models**

More recently, there has been a surge of work proposing a new generation of DSMs that frame the vector estimation problem directly as a supervised task, where the weights in a word vector are set to maximize the probability of the contexts in which the word occur (Bengio et al., 2003; Collobert and Weston, 2008; Mikolov et al., 2013a; Pennington et al., 2014). Compared to these new approaches, the traditional construction of context vectors is turned on its head: Instead of first collecting context vectors and then reweighting these vectors based on various criteria, the vector weights are directly set to optimally predict the contexts in which the corresponding words tend to appear. Since similar words occur in similar contexts, the system naturally learns to assign similar vectors to similar words (Baroni et al., 2014). Thus, this new generation is described as “predictive models”, and the representations produced by these new models are normally referred as “word embeddings”. Baroni et al. (2014) demonstrated that, in most of the tasks, predictive models consistently outperform count-based models.

In our experiments, we applied the word embedding strategy proposed by Mikolov et al. (2013a) which contains two model architectures for learning distributed representations:

- **Continuous Bag-of-Words Model**. The theory behind Continuous Bag-of-Words **(**CBOW) model is similar to the feedforward neural network language model, where the non-linear hidden layer is removed and the projection layer is shared for all words; thus, all words get projected into the same position. Since the order of words in the history does not influence the projection, it is called bag-of-words model. The training objective of this model is to predict the the word in the middle from a window of surrounding context words.

- ***Continuous Skip-gram Model***. The skip-gram model is similar to CBOW, but instead of using the context to predict the middle word, the distributed representation of the input word is used to predict its context in the same sentence. More precisely, each input word is applied to a log-linear classifier with continuous projection layer to predict words within a certain window (words occur before and after the input word). Compared to Skip-gram, CBOW is faster and more suitable for larger datasets (Mikolov et al., 2013a). Therefore, in our experiments, we trained our word embedding model using CBOW.

Since both CBOW and Skip-gram are trained using a simple neural network architecture, with these models, very accurate high dimensional word vectors can be obtained from a larger amount of data in much less time compared to the popular neural network models (both feedforward and recurrent). Their model architectures are shown in Figure 2.8.

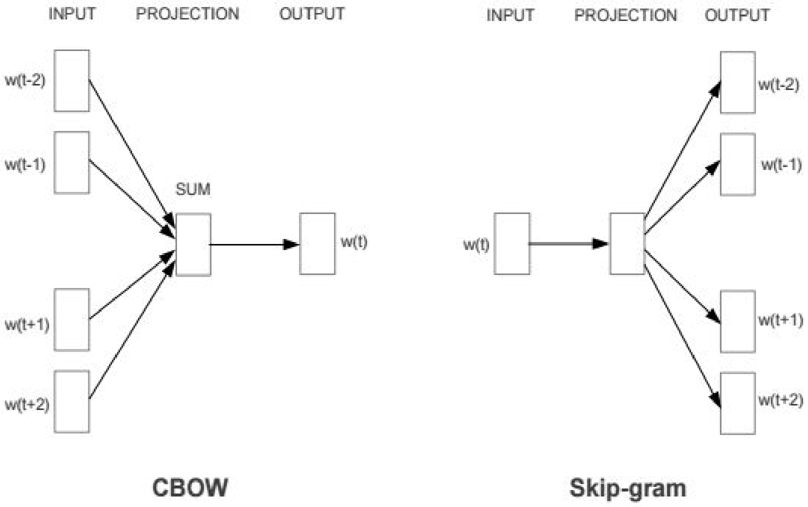


Figure 2.8: CBOW and Skip-gram model architectures (Mikolov et al., 2013a).

Mikolov et al. (2013a) showed that word embeddings can project word semantics into a vector space from their distributional characteristics. More interestingly, it is claimed that the relationship between vector spaces that represent different language word semantics can be captured by a linear transformation, since same concepts in different languages share similar geometric arrangement in vector spaces as the examples shown in Figure 2.9. According to the method proposed by Mikolov et al. (2013b), the process of generating dictionaries was automated by learning a linear transformation between vector spaces of two particular languages on a 5K seed dictionary. At test time, a new word can be translated by projecting its vector representation from the source language space to the target language space. Once the vector in the target language space is obtained, similar target language word vectors (found by cosine similarity assessment) are ranked as possible translations. The translation matrix is found via optimization with a stochastic gradient descent algorithm. Their results in the form of ranked lists are further refined with a confidence threshold that tries to balance precision and recall, i.e. coverage. Thus, the highest coverage achieved for the pair English-Spanish is 92.5%, but precision at top position is 53%. Best precision reported is 78% (better results are obtained when refining with edit distance) but with a coverage of 17%. However the limitation of the approach of Mikolov et al. (2013b), as well as other similar models (Faruqui and Dyer, 2014; Dinu et al., 2015; Lazaridou et al., 2015; Vulic and Korhonen, 2016) is that they all rely on readily available seed lexicons of highly frequent words to learn the mapping.

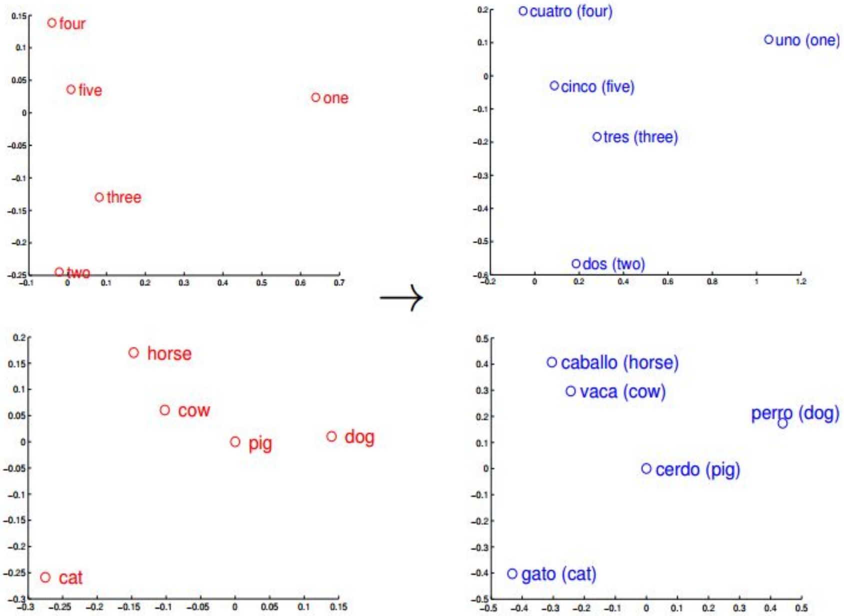


Figure 2.9: Distributed word vector representations of numbers and animals in English (left) and Spanish (right) (Mikolov et al., 2013b) .

Besides the approaches presented above, there are also many other interesting bilingual word embedding (BWE) strategies. According to Vulic and Korhonen (2016) and Upadhyay et al. (2016), we cluster the rest of the BWE models into four different types based on bilingual signals used for training, and two properties: P1 regarding leveraging large monolingual training sets tied together through a bilingual signal and P2 regarding the use of inexpensive bilingual signal to learn shared bilingual word embedding space

(SBWES) in a scalable and widely applicable manner across languages and domains:

- ***Parallel-Only***: This group of BWE models relies on sentence-aligned and/or word-aligned parallel data as the only data source (Zou et al., 2013; Hermann and Blunsom, 2014a; Kociský et al., 2014; Hermann and Blunsom, 2014b; Chandar et al., 2014). In addition to an expensive bilingual signal (colliding with P2), these models do not leverage larger monolingual datasets for training (not satisfying P1).

- ***Joint Bilingual Training***: These models jointly optimize two monolingual objectives, together with a cross-lingual objective acting as a cross-lingual regularizer during training (Klementiev et al., 2012; Gouws et al., 2015; Soyer et al., 2015; Shi et al., 2015; Coulmance et al., 2015). The main disadvantage of this approach is the costly parallel data needed for the bilingual signal (thus colliding with P2).

- ***Pseudo-Bilingual Training***: This set of models requires document alignments as bilingual signal to induce a SBWES. For instance, Vulic and Moens

(2016) created a collection of pseudo-bilingual documents by merging every pair of aligned documents in training data. With these pseudo-bilingual documents, the “context” of a word is redefined as a mixture of neighbouring words (in the original language) and words that appeared in the same region of the document (in the ”foreign” language). The bilingual contexts for each word in each document steer the final model towards constructing a SBWES. Compared to the selected baselines models[2], The relative increase over the best scoring baseline BLI models from comparable data is 19.4% for the ES-EN pair, 6.1% for IT-EN and 65.4% for NL-EN. The advantage of these models lies in exploiting weaker document-level bilingual signals (satisfying P2), but these models are unable to exploit monolingual corpora during training (thus colliding with P1).

- ***Matching Term inspired by IBM Model 1***: Following the spirit of IBM Model 1, this set of models learns bilingual lexicons/phrases from

[2] Baseline models: BiLDA-BLI model proposed by Vulic et al. (2011); Assoc-BLI model proposed by Vulic and Moens (2013a); and PPMI+cos model proposed by Bullinaria and Levy (2007).

non-parallel corpora by optimizing a matching term using Viterbi EM algorithm (satisfying P1) . For instance, Dong et al. (2015) proposed a joint model for iteratively learning parallel lexicons and phrases from non-parallel corpora. The model was trained using a Viterbi EM algorithm that alternates between constructing parallel phrases using lexicons and updating lexicons based on the constructed parallel phrases. Zhang et al. (2017) designed a similar matching mechanism into bilingual word representation learning. One crucial difference is the parametrization of the matching probability. Dong et al. (2015) used the standard IBM model 1 to define the phrase translation probability and their model did not involve continuous representation of words, which in turn leads to different optimization procedure. Since both models can iteratively improve bilingual lexicons learning through benefiting the newly acquired translation pairs derived from monolingual data, only small parallel corpora or seed dictionary are required for the training (satisfying P2).

In this dissertation, we decided to induce bilingual lexicons from monolingual data using the word embedding vector proposed by Mikolov et al. (2013a). However, instead of optimizing a transformation matrix across language spaces, we treat bilingual lexicon generation as a binary classification problem: given a source word, the classifier predicts whether a target language word is its translation or not. This method was inspired by the recent evidence of Necsulescu et al. (2015) that indicated that simple concatenation of word embeddings is effective for finding lexical semantic relations (i.e. hyponymy, hyperonymy, meronymy, attribution and properties) holding in word pairs with supervised methods. Our task is accordingly defined as whether the concatenated vector of source and target word could be useful for a classifier to learn the translation relation between them.

Recently, several researches applied similar strategies of bilingual lexicons induction to enrich SMT phrase table. For instance, Zhao et al. (2015) proposed a method that uses monolingual phrase representations (via simple element-wise addition of word vectors) to generate translation rules for infrequent, or phrases that do not appear in the bilingual data, which are called by these authors ‘unlabelled phrases. The general idea behind this method is to identify phrases for which translation rules are known to be similar to an unlabeled phrase, and to use them to induce translation rules for the unlabeled phrase. Their approach was similar to Mikolov et al. (2013a), but instead of learning a single global linear projection matrix to capture the mapping relationship between the source and target spaces, they proposed to learn many local linear projections which are individually trained for each unlabeled source phrase. More specifically, for each unlabeled source phrase f, they learned a mapping $W_f \in R^{d \times d}$ based on the translations of m of f's labeled neighbors:

$$(f_1, e_1),\ (f_2, e_2), \ldots,\ (f_m, e_m),\ f_i \in N(f),\ 1 \leq i \leq m,\ m \geq d.$$

An additional k-NN query was required to find the labeled neighbors for each unlabeled source phrase. Then based on Saluja et al. (2014) approach, structured label propagation (SLP) was applied to propagate translation candidates from frequent source phrases that are labeled to unlabeled neighbors that are infrequent as: for a known translation rule $(f`', e')$, SLP propagates the target side phrases $e \in N(e')$, that are similar to e', to the unlabeled source phrases $f \in N(f')$, that are similar to f', as new translation rules. To calculate the propagation probability of source and target phrases, costly PMI statistics (according to Saluja et al., 2014) was replaced by continuous phrase representations. It was computed as: $sim(\overline{e} \mid e) = \dfrac{1}{1 + \| \overline{e} - e \|}$, where $\overline{e}$ is the projected point of foreign phrase f, and $\| \overline{e} - e \|$ is the Euclidean distance between vectors $\overline{e}$ and e. Then the phrase translation probability for each candidate $e \in N(\overline{e})$ was calculated as:

$$P(e \mid f) = \frac{exp\{sim(e, \overline{e})\}}{\Sigma_{e' \in N(\overline{e})} exp\{sim(e', \overline{e})\}}$$

The backward translation probability was computed using Bayes' Theorem. Similar to Saluja et al. (2014), forward and backward lexicalized weightings were obtained by using a baseline lexical model. Their approach improved a phrase-based baseline by up to 1.6 BLEU on Arabic-English translation. However, the limitation is that the expansion is restricted by the existing translation pairs of the baseline phrase table.

Irvine and Callison (2014 and 2016) enriched SMT low resource systems with bilingual lexicons extracted from monolingual corpora. Their bilingual lexicons were learned by a log-linear classifier (Irvine and Callison, 2013) that was trained on many different monolingual signals (contextual similarity, temporal similarity, orthographic similarity and topic similarity, frequency similarity and burstiness similarity). To produce new translation candidates, they paired and scored all source language unigrams in the tuning and test sets with target language unigrams that appear in a comparable corpora. Then, for each source language unigram, the log-linear model scores were used to rerank target language translation candidates, only top-k (k=1, 2, 5, 25, 200) translation candidates were chosen for the phrase table expansion. Since much noise was introduced, besides the standard phrase-based MT feature set (phrase and lexical translation probabilities and a lexicalized reordering weighting), 30 monolingually-derived signals were also needed to be applied as further translation table features to prune the new phrase pairs. The experiments were conducted on seven different language pairs. They achieved an average of 0.8 BLEU improvement compared to the low resource baselines. However, in this approach, comparable corpus is still required for learning new translation pairs and the 30 new phrase table features make the method too complicated. Besides, extracting a large number of monolingual signals is computationally expensive.

Chapter 3

3. MORPHOLOGY EXPANSION FOR SMT

Parallel corpora are the key resource that supports SMT to learn translation correspondences at the level of words (Brown et al., 1993), phrases (Koehn, 2003) and treelets (Galley et al., 2006). The more data is used to estimate the parameters of the translation model, the better it can approximate translation probabilities that in turn will deliver better translations.

Although nowadays large parallel corpora are easily available for some language pairs such as English-Spanish and English-French, it is still difficult to get, or even doesn't exist, for most others. This is the case of the Chinese-Spanish language pair, for which to find more parallel corpora to train a SMT system is clearly insufficient. One way of approaching the lack of corpora is to provide the system with more inflectional translation options for alleviating the translation problem of morphological variants.

According to the error analysis and error classification of a English-Greek SMT system (shown in Section 2.2) given by

Avramidis and Koehn (2008), more than 40% of the errors are due to the incorrect translation of word forms. This problem is particularly severe when translating from a morphologically poor language to a high-inflected language, as in our case, from Chinese to Spanish. As explained in Section 1.1, Chinese has no morphological complexity within a word, but Spanish, in contrast, is highly inflected. So when a system suffers from the parallel data scarcity problem, it is not capable of producing a translation of the correct word form. So in Chapter 3, based on the knowledge of our Chinese-Spanish SMT baseline system, we generated the missing morphological variants using a Spanish lexical resource and applied the newly acquired inflectional translation pairs to enrich the baseline phrase table.

According to the Spanish morphological resource[3] that we used for the generation of morphological variants, we plot in Figure 3.1 the number of inflectional variants regarding each lemma of noun, verb and adjective [4]. The x-axis represents each of the lemmas (noun, verb and adjective) in the morphological resources and the y-axis indicates how many variants each lemma has. The three sparkline charts (of noun, verb and adjective) share the same x-axis, but with different y-axis.

[3] IULA spanish lexical resource: https://www.upf.edu/web/iula/recursos-corpus-i-eines

[4] Our variants expansion was only based on noun, verb and adjective.

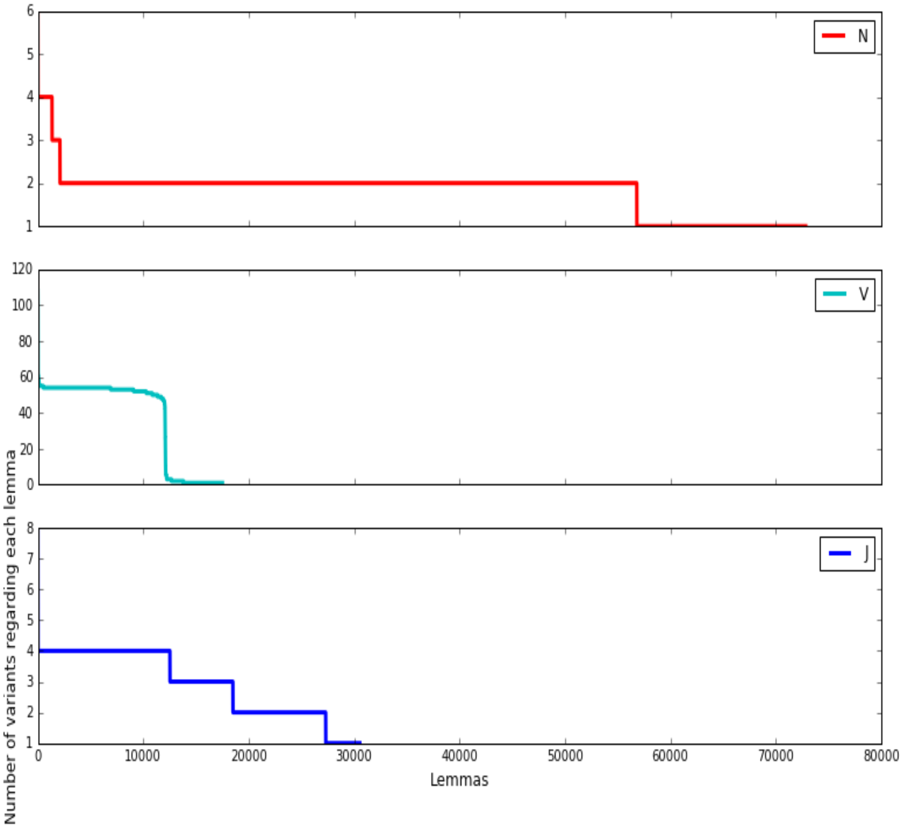


Figure 3.1: Number of morphological variants regarding the corresponding lemma in the Spanish morphological resource.

In the feature above, it can be seen that the lemmas of noun and adjective can be inflected in up to 8 different different forms. In case of verb, there can be more than 100 different morphological variants regarding a single lemma. Therefore, the inflection difference could be a big problem for SMT between Chinese and

Spanish when there is not enough parallel corpora available for training the translation model.

As referenced in Section 2.2.1, to reduce the morphological translation errors in SMT, many works focus on using extra syntactic or lexical resources on source side or target side to improve the mapping between the source word and its corresponding translation variant. However, without having all possible inflection variants in the translation model, these methods would be of no effect. So the goal of this experiment is to provide more inflectional translation options to the system.

3.1. Morphology generation with lexical resource

To generate new inflectional variants, first, all the unigram translation pairs of nouns, adjectives and verbs were collected from the baseline phrase table including the bad translations. We focus on nouns, adjectives and verbs because these are the word classes in which might occur morphological translation errors. Given a bilingual entry, the Spanish morphological resource was used for lexical lookup to return all the possible inflectional variants that share the same lemma with the target Spanish word. Since the Chinese source word has no inflection, all the target morphological variants share the same source word. Note that we only take those variants which could be found in the language model, since some very low frequency (rarely used) variants were included in the morphological resource such as *viésemos* (the first-person plural imperfect subjunctive form of *ver*) and *circunstanciadamente* ('according to the situation'). The generation process of the new association can be described as shown in Figure 3.2.

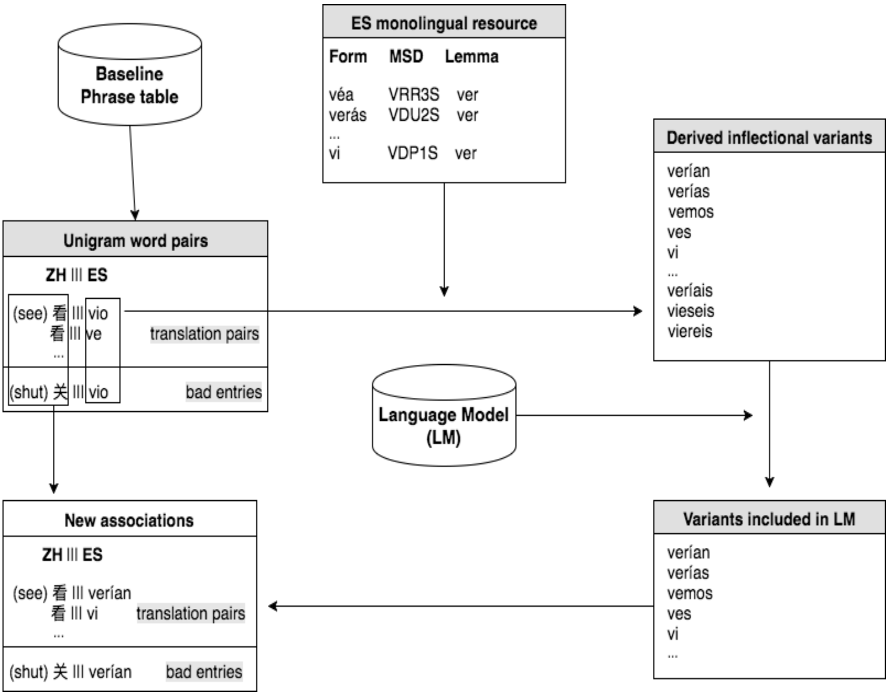


Figure 3.2: Generation process of new morphological associations.

Before moving to the step of SMT phrase table expansion, we would like to clarify the following issues:

- In the baseline phrase table, there are some translation pairs that contain same source word but with different translation inflections such as *看_vio* and *看_ve,* so some inflectional translation pairs may be repeatedly generated. To avoid such situation, once obtained all the new associations, the repeated entries were removed. Namely, only unique entries were kept for the phrase table expansion.

- Our new inflection associations were generated based on all the unigram translation pairs (only nouns, adjectives and verb) of the baseline phrase table including bad translations without any selection threshold. Since when the parallel training corpora were limited, the correct translation candidate could rank in any position making it hard to properly define a specific threshold. As a result of this fact, for each Chinese source word, we decided to consider all the target candidates delivered by the baseline for the generation of the new translation variants. Although in this way a lot of noise entries were generated and delivered to the SMT system, the experimental results demonstrate that the language model can handle them to an acceptable extent.

3.2. Experimental setup

This experiment was conducted on the language pair Chinese and Spanish. The parallel corpora used to train the SMT baseline system are: Chinese-Spanish OpenSubtitles 2013[5] (1M sentences). For tuning, we randomly sampled 1K sentence pairs from the News Commentaries[6] parallel corpora released by Tiedemann (2012). For testing, we used TAUS translation memory[7] (2K sentences) and a subset of UN parallel corpora[8] (2K sentences). To build the language model, since we need all the newly acquired inflectional variants to be included, the Spanish Wikipedia corpus[9] (150M words, 2006 dump) was combined with OpenSubtitles 2013 target corpus for the training of the language model. We used Stanford PoS tagger[10] to collect all the noun, adjective and verb unigram translations from the baseline phrase table. In total, 0.89 M word pairs were used to generate morphological variants.

The morphological resource used to generate new morphological variants includes three information: word form, morphosyntactic description and lemma. In total, there are 72K lemmas of noun, 17K lemmas of verb and 30K lemmas of adjective. For the phrase table

[5] http://opus.lingfil.uu.se/OpenSubtitles2013.php
[6] http://opus.lingfil.uu.se/News-Commentary.php
[7] http://www.tauslabs.com/
[8] http://opus.lingfil.uu.se/UN.php
[9] http://hdl.handle.net/10230/20047
[10] https://nlp.stanford.edu/software/tagger.shtml

expansion, we only generate inflectional variants for those target words that are present in the baseline phrase table. Figure 3.3 demonstrates the number of inflectional variants regarding each lemma in our unigram bilingual lexicon. After the filter of the language model, 4.4K noun lemma, 2.1K verb lemma and 1.8K adjective lemma were applied to enrich the SMT phrase table.

- **Phrase-based SMT setup**

To build our SMT system, we used Moses phrase-based MT framework (Koehn et al., 2007b) and the standard phrase-based MT feature set, including phrase and lexical translation probabilities (direct and inverse) and reordering score produced by a lexicalized reordering model (Koehn et al., 2005). We applied *mgiza* (Gao and Vogel, 2008) to align parallel corpora and *KenLM* (Heafield, 2011) to train a 3-gram language model. For the evaluation, we used BLEU (Papineni et al., 2002) and METEOR metric (Banerjee and Lavie, 2005).

Note that the parameter *Good Turing*[11](Gale, 1995) was applied in order to reduce overestimated translation probabilities. Since the parallel corpus contains many new unigram translation pairs, the translation probabilities of these new word pairs might be higher than other translation pairs of the baseline phrase table. *Good*

[11] http://www.statmt.org/moses/?n=FactoredTraining.ScorePhrases

Turing is a popular count smoothing technique which provides a principled way to adjust count. To obtain better phrase translation probabilities, the observed counts may be reduced by expected counts which takes unobserved events into account. Borrowing a method from language model estimation, *Good Turing* discounting can be used to reduce the actual counts to a more realistic number. The value of the adjusted count is determined by an analysis of the number of singleton, twice-occurring, thrice-occurring, etc. phrase pairs that were extracted (Koehn, 2010). In order to fairly measure the impact of the newly derived translation variants on the SMT system, the *Good Turing* parameter was applied to the training processes of both baseline and expanded phrase table.

3.3. Experimental results

In this section, we present the experimental results of the expanded SMT system (from Chinese to Spanish). Table 3.1 depicts the evaluations by BLEU and METEOR metric on UN and TAUS test sets.

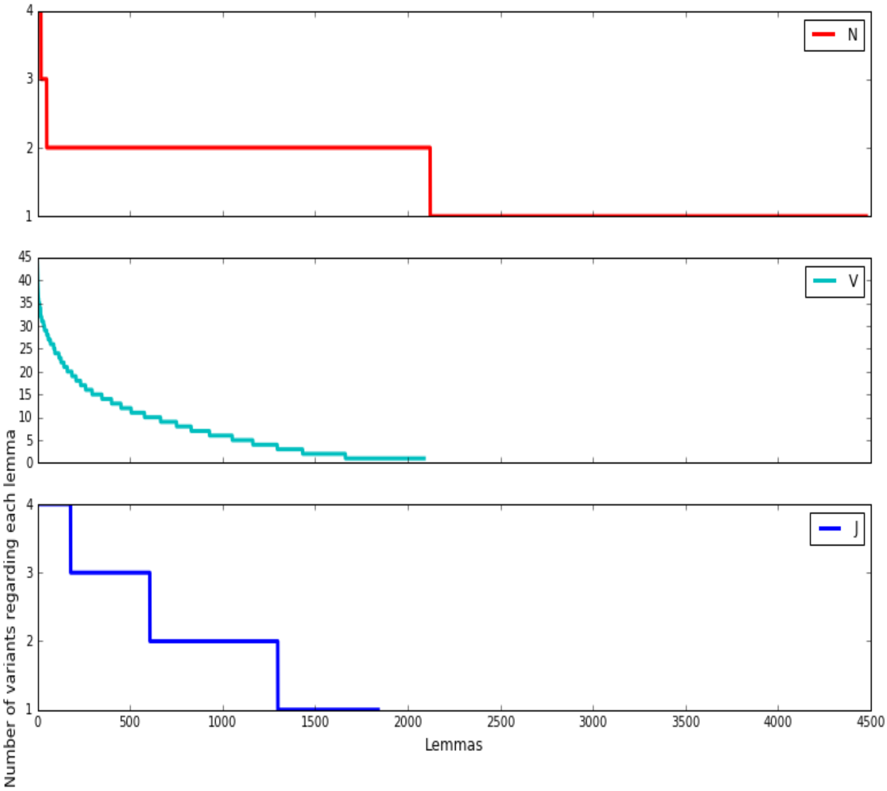


Figure 3.3: Number of morphological variants regarding the corresponding lemma for inflectional variants expansion.

	TAUS		UN	
	BLEU	METEOR	BLEU	METEOR
Baseline	8.80	0.2929	10.81	0.3075
PhT with morphological expansion	9.39	0.3075	11.42	0.3198

Table 3.1: Test results of the baseline and expanded phrase table.

According to Table 3.1, the system with the expanded phrase table outperforms the baseline by an improvement of 6% BLEU score and 5% METEOR score on TAUS test set; and an improvement of 5% BLEU score and 4% METEOR score on UN test set. The improvements are statistically significant according to the paired student t-test at the level of $p < 0.05$. Figure 3.4 and Figure 3.5 show the METEOR score regarding each sentence pair of UN test set and TAUS test set, respectively. After incorporating new translation lexicons with inflections, 67% sentence pairs outperform the baseline in UN test set and 63% sentence pairs outperform the baseline in TAUS test set.

3.4. Discussion

Herein we present the observations and analysis based on the translation results produced by the baseline and the expanded system. As shown in Table 3.1, the expanded system outperforms the baseline, demonstrating that our new morphological variants indeed help the system to produce better translations. One of the most direct impact was that the system can produce more correct word form after the expansion. For instance, in the following example, with the baseline, the verb 決定 was translated as a noun, but the expanded system was capable of delivering the correct verb word form.

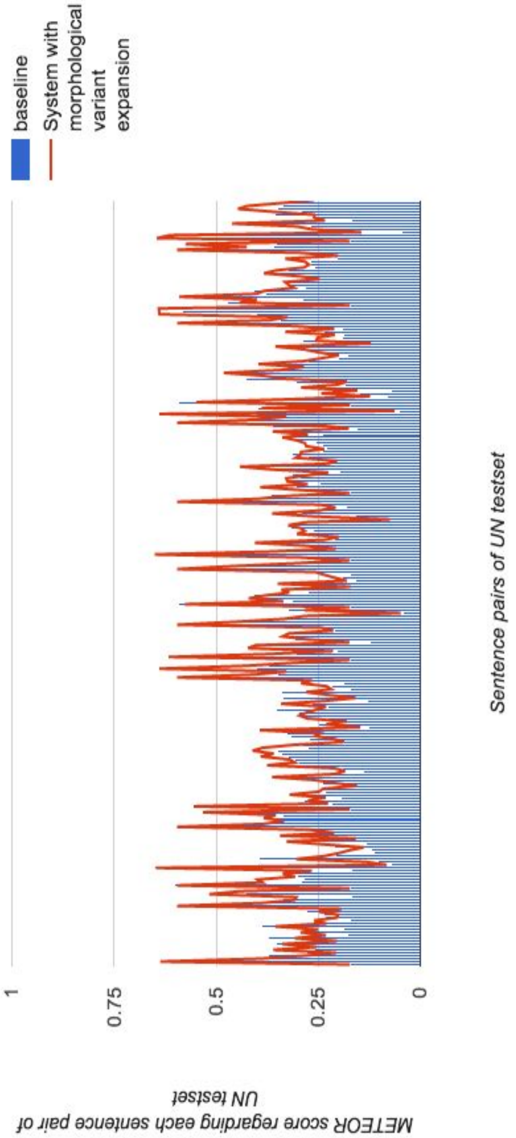


Figure 3.4: METEOR score regarding each sentence pair of UN test set.

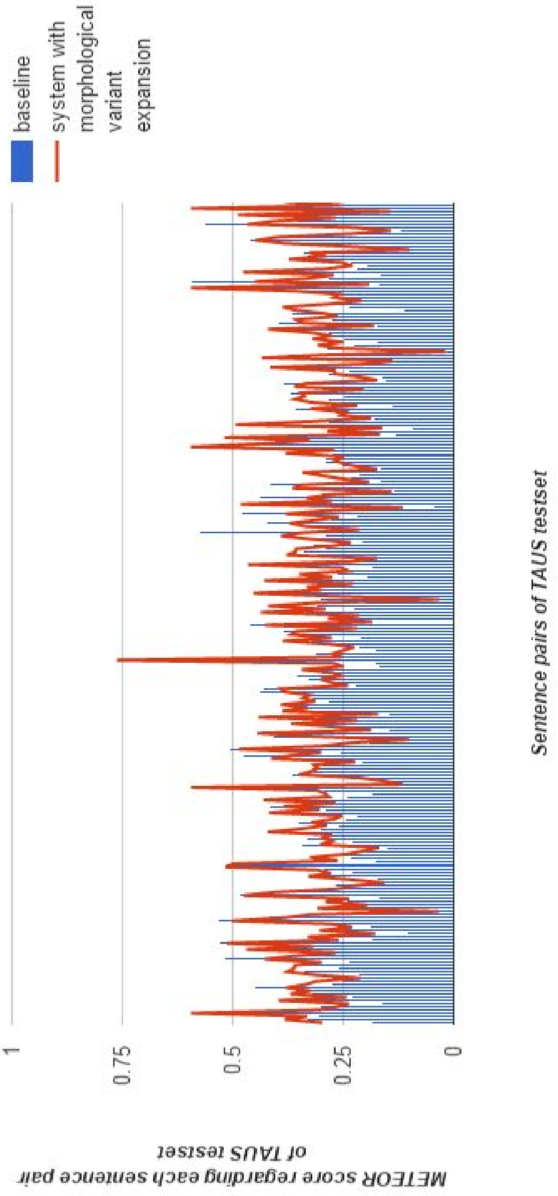


Figure 3.5: METEOR score regarding each sentence pair of TAUS test set.

(1)

Source:

决定任命一名负有一定任务的特别报告员。

(Decided to appoint a special rapporteur with a certain mandate.)

Reference:

decidió nombrar un relator especial con un mandato expreso.

Baseline:

la decisión del Comité para una misión especial.

Expanded system:

decidió nombrar un especial de la misión.

We notice that both the baseline and the enriched system have the *omission* problem (Vilar et al., 2006; Costa et al., 2015), according to which the translations of some words of the original source sentence are missing. Observing the translation result produced by the baseline, in some cases, although the source word was present with several translation options in the baseline phrase table, the system prefered not to translate it if it was not the right case. Compared to the baseline, the expanded system performs slightly better due to the addition of new infections.

(2)

Source:

以 确保 其 ***<u>有效</u>*** 运作。

(to ensure it effectively works.)

Reference:

con objeto de asegurar su funcionamiento efectivo.

Baseline:

para asegurar . Funciona

Expanded system:

para asegurar ***<u>efectivamente</u>*** funciona.

(3)

Source:

该 国 人民 ***<u>清楚 地</u>*** 决定 选择 民主。

(The people of the country have clearly decided to prefer democracy.)

Reference:

la población ha decidido ***<u>claramente</u>*** preferir la democracia.

Baseline:

el pueblo ... decidí elegir entre la democracia.

Expanded system:

el país ***<u>claramente</u>*** decide elegir a la democracia.

In both examples above, instead of giving an incorrect translation word form, the baseline omitted translating the adverbs 有效 and 清楚地. The expanded system, on the contrary, delivered the correct translations after the phrase table expansion, hence alleviated the *omission* problem to some extent. However, observing the example (3), the verb 决定 was translated differently as *decidí* and *decide* by the baseline and expanded system, respectively. We realize that simply adding all morphological variants in the baseline system can not solve all the translation problem of word inflections. It is particularly serious in our case, since Chinese has no case, gender or number markers for nouns and no subject-verb agreement or tense for verbs. So when a subject or a tense is not explicitly expressed, it may not be translated appropriately regarding the Chinese source sentence in some cases. For instance, in both Chinese examples below, there is no subject or time marker to determine the person and tense of the verbs 表示. If simply judging from these sentences without any context, they can be interpreted either in present or in past tense. Therefore, in the examples above, we consider the “ incorrect translations” of the verb tense as an acceptable translation option rather than translation errors. Regarding the lexical choice, the expanded system performs better than the baseline as shown in the following examples.

(4)

Source:

表示关切 在宣言通过四十年之后竟然仍存在一些非自治领土。

(Expressing its concern that, after forty years of the adoption of the Declaration, there are still some Non-Self-Governing Territories.)

Reference:

expresando su preocupación por que cuarenta años después de la aprobación de la Declaración, aún siga habiendo territorios no autónomos.

Baseline:

significa que su en el manifiesto de 40 años después de que todavía existe algo que 自治 territorio.

Expanded system:

expresó ***preocupación*** en la Declaración de cuarenta años después de que todavía existen algunos territorios 自治.

(5)

Source:

回顾 其关于援助巴勒斯坦难民的 1948年 11月 19日 第 212 (III) 号 决议。

(Recalling its resolution 212 (III) of 19 November 1948 on assistance to Palestine refugees.)

Reference:

recordando su resolución 212 (III), de 19 de noviembre de

1948, relativa a la ayuda a los refugiados de Palestina.

Baseline:
en su ayuda a los refugiados de Palestina, el 19 de noviembre de 1948年 第212 lll, el Estado.

Expanded system:
recuerda la ayuda de los refugiados de Palestina 19 de noviembre de 1948年 第212 lll ; . el asunto.

3.5. Summary

In this section we described an approach to enrich our Chinese-Spanish SMT phrase table by adding new translation pairs with morphological variants that derived from a Spanish lexical resource. The experimental results showed improvements of the translation quality in different aspects as demonstrated in Section 3.4. However, we notice that simply adding all the inflectional variants into the parallel training corpora is not sufficient to handle some translation problem of verb conjugation due to the unclearness property of Chinese. To enhance this situation, as future work, combine our method with the approach that incorporates syntax information of morphologically poor language (Avramidis and Koehn, 2008) may enable a further improvement, as they claimed that their method could be limited by the data sparsity problem.

Chapter 4

4. BILINGUAL LEXICON INDUCTION FOR SMT

In the experiments described in the previous chapter, the morphology expansion was based on the existing knowledge of the baseline phrase table. In this chapter, we induce new translation lexicons from monolingual corpora with the intention of alleviating OOV problem. Different researches have been done on the topic of automatic construction of bilingual lexicons. As reported in Section 2.2, (Tanaka and Umemura, 1994; Bond et al., 2001; Nerima and Wehrli, 2008) learned new translation pairs by combining existing bilingual dictionaries that share a common language. Some other approaches (Fung, 1995; Chiao and Zweigenbaum, 2002; Yu and Tsujii, 2009) consist in extracting translation equivalents from comparable corpora rather than parallel corpora. More recently, (Mikolov et al.,2013a; Vulic and Moens, 2015; Chandar et al., 2014;Wang et al., 2016) proposed cross-lingual word embedding strategies to map words from a source language vector space to a target language vector space, and also demonstrated its application to bilingual lexicon induction.

The methods presented in this section are similar to the end-to-end experiments of Irvine and Callison-Burch (2014) and Irvine and Callison-Burch (2016), which generated bilingual lexica by training a classifier using a large variety of monolingual signals. In the present experiments, we trained classifiers using only word embedding vectors and two other monolingual features, but the classifier is also capable of learning translation lexicons from monolingual data. The first classifier described in Section 4.1 was trained only with word embedding vector. Then to improve the performance of word embedding-based model, in Section 4.2 and Section 4.3, we proposed two enhanced models trained with additional word frequency information and Brown clustering, respectively.

4.1 New translation lexicon generation with word embeddings

In the experiments described in this chapter, we built a Support Vector Machine (SVM) to induce bilingual lexicon from monolingual corpora for SMT phrase table expansion. The goal of SVM model is to find the best hyperplane that maximizes the margin between data points of one class from those of the other class as shown in Figure 4.1 (Cortes and Vapnik, 1995). The basic idea behind SVM is: the input vectors of training examples are mapped into a higher dimensional feature space through some nonlinear mapping chosen a priori. In this space a linear decision hyperplane is constructed with the maximal margin. To determine the margin, only a small amount of the training data, the so called support vectors, are taken into account.

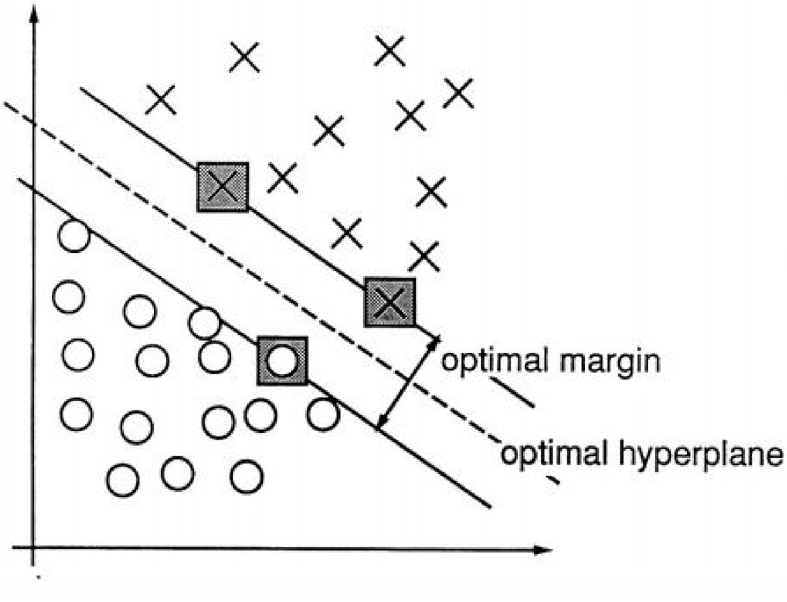


Figure 4.1: An example of a separable problem in a 2 dimensional space. The support vectors, marked with grey squares, define the margin of largest separation between the two classes (Cortes and Vapnik, 1995).

However, training a SVM requires the solution of a very large quadratic programming (QP) optimization problem, which means training algorithms for SVM can be very slow, especially for large data. To alleviate such problem, Platt (1998) proposed the Sequential Minimal Optimization (SMO) algorithm that can break this large QP problem into a series of smallest possible QP problems. These small QP problems are solved analytically, which avoids using a time-consuming numerical QP optimization as an inner loop. In general, SMO is faster, and has better scaling properties for difficult SVM problems than the standard SVM training algorithm (Platt, 1998). Therefore, in this work, we trained our SVM classifier using SMO.

4.1.1. Approach

The task of this experiment was to train a SVM-SMO binary classifier with vectors made of concatenated word embeddings of translation equivalents. In testing mode, new word pairs were classified as being one the translation of the other or not. The classifier is trained to recognize a translation relation.

Our intuition was that although the word embedding models were trained on unrelated monolingual corpora, the concatenated word embedding vector of source and target word does encode useful information for the classifier to decide whether a word pair is under a translation relation or not. In Figure 4.2 we visualize the geometric arrangement of our 6K training word pairs (*right* *translation* and *no translation*) represented by word embedding vector in a 3-dimensional space.

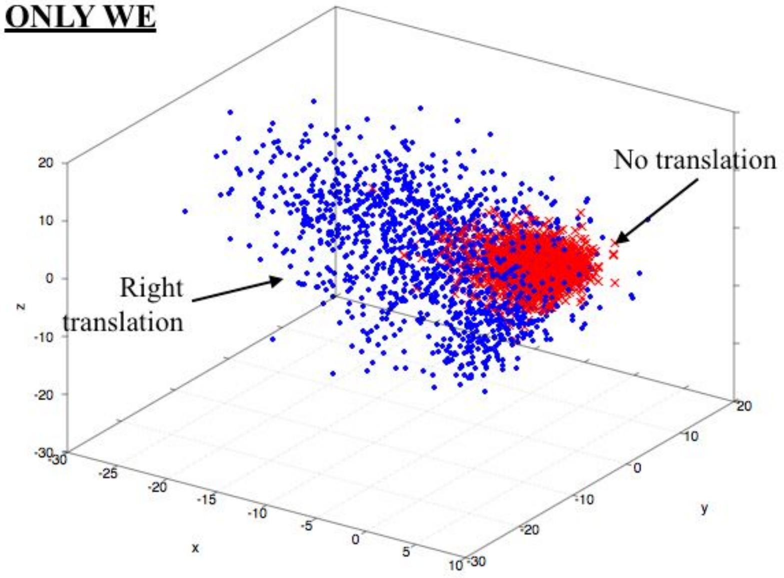


Figure 4.2: Distributed representations of 6K word pairs (1K *right translation* and 5K *no translation*) with WE of 400 dimensions. We used PCA to project high dimensional vector representations down into a 3-dimensional space.

For the training and testing set, each translation pair was represented by concatenating the word embedding vector of the source word and of its corresponding translation for positive examples. For negative examples, we randomly combined a number of source word with target words. Formally, given a translation word pair (x, y), x being a source and y a target word, whose vector features are $v(x) = (x_1, x_2, ..., x_n)$ and $v(y) = (y_1, y_2, ..., y_m)$ respectively, then $v(x, y)$ is defined as the concatenation of $v(x)$ and $v(y)$: $v(x, y) = (x_1, x_2, ..., x_n, y_1, y_2, ..., y_m)$. This experiment was only conducted with unigrams of three word classes: noun, verb and adjective.

The classifier was evaluated in two different scenarios:

- ***Proof-of-Concept***. For the *proof-of-concept* evaluation, the test set was prepared in the same way as the training set. Each source word embedding was paired with one target word embedding. Given a set of testing word pairs, the trained model was used to classify whether these testing word pairs are translation equivalents or not.
- ***SMT simulation***. This evaluation scenario was similar to how SMT system produce a phrase table where pairs of words are extracted from all possible combinations of words occurring in a given set of aligned sentences and the probability of a particular word being the translation of other is estimated. In this evaluation scenario, each of the source word representation from the test set was concatenated with all target word (of same POS) representations from the target monolingual corpus. Then all translation pair

candidates were ranked by the confidence score produced by our classifier.

4.1.2 Experimental setup

In this section we describe the experimental settings of our SVM-SMO classifier trained using gaussian kernel that was intended to learn the translation relation. The classifier is binary: it learns whether a word pair is *right translation* or *no translation*. The general outline of the experiments is: (i) Generation of the right and wrong translation lists. (ii) Obtaining the corresponding word vector representation from monolingual word embedding models. (iii) Concatenation of the vector representations of the source word and its translation equivalent (or random word for negative instances). (iv)Training a SVM-SMO classifier[1] using the previous concatenated representation. (v) Evaluation of the classifier.

- **Data sets**

The word embedding-based classifier was evaluated on two quite distinct language pairs: Chinese and Spanish (ZH-ES) and English and Spanish (EN-ES). The monolingual corpora used for the *Proof-of-Concept* evaluation scenario were: Chinese Wikipedia Dump corpus[2] (149M words); Spanish Wikipedia corpus[3] (150M,

[1] SMO algorithm （Platt, 1998) as implemented in WEKA (Hall et al., 2009).

[2] https://archive.org/details/zhwiki_20100610

2006 dump); and for English, the BNC[4] (100M). For the *SMT simulation* evaluation task, the corpora that we used were: WMT 11[5] text data of English (59M) and Spanish (59M).

To obtain the positive training set (*right translation*), a translation list was produced by, first randomly extracting a list of about 1K nouns, verbs and adjectives[6] (word frequency rank plot in Figure 4.3 and Figure 4.4) from the ZH monolingual corpus and EN monolingual corpus. Then these randomly selected words were translated to ES using on-line Google Translator and all the translation outputs delivered by Google Translator were manually evaluated. Since not all the produced translations could be found in the target monolingual corpus, we removed from our datasets those words whose corresponding translation was not in the target corpus. To build the negative training set (*no translation*), as mentioned before, we randomly selected non-related words from the monolingual corpus of each language and randomly combined them.

Since Zhao et al. (2015) pointed out that the frequency of the positive training examples has great impact on the performance of

[3] http://hdl.handle.net/10230/20047

[4] http://www.natcorp.ox.ac.uk/

[5] http://www.statmt.org/wmt11/training-monolingual.tgz

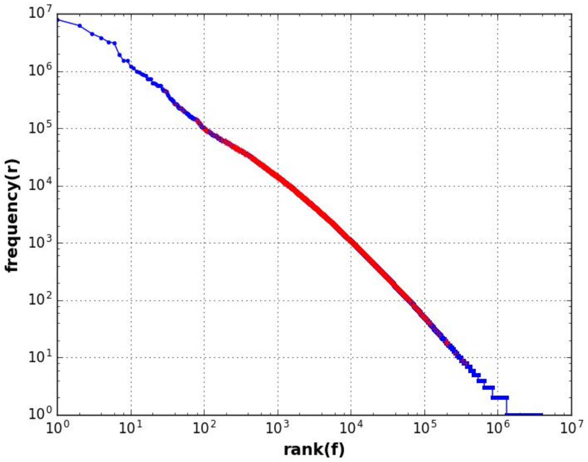
[6] For PoS tagging of all corpora, we used Stanford PoS Tagger (Toutanova et al., 2003)

bilingual lexicon induction. Figure 4.3 and Figure 4.4 show the zipf plot for word counts in our monolingual corpora, where we highlight the word frequency distribution of our positive training examples in red color.

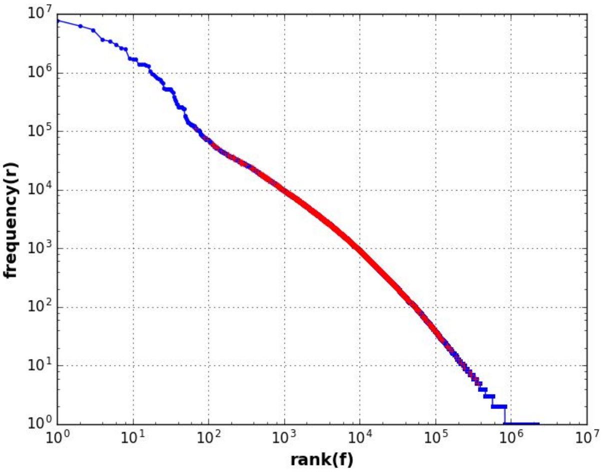
This dataset was divided into training and testing sets. Final figures of the datasets are provided in Table 4.1.

	ZH-ES				EN-ES			
	Training		Testing		Training		Testing	
	YES	NO	YES	NO	YES	NO	YES	NO
Noun	451	2390	99	449	449	2379	94	469
Adj.	302	1492	71	398	300	1500	99	500
Verb	400	1999	113	599	300	1500	99	500
Total	1153	5881	283	1446	1049	5379	292	1469

Table 4.1: Translation pair datasets for ZH-ES and EN-ES.



(1)



(2)

Figure 4.3: Zipf plot for word counts in monolingual corpora (blue dot) in ZH-ES experiment: ZH Wikipedia corpus (1) and ES Wikipedia corpus (2), and their corresponding word frequency distribution of positive training examples (red cross) .

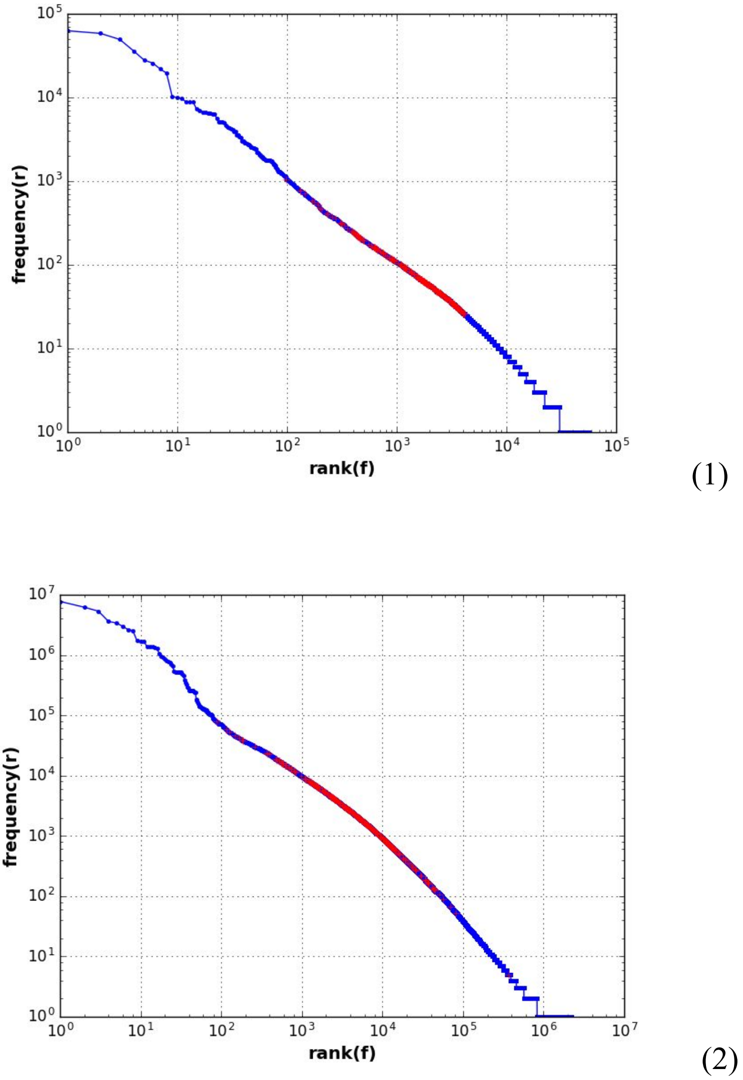


Figure 4.4: Zipf plot for word counts in monolingual corpora (blue dot) in EN-ES experiment: EN BNC corpus (1) and ES Wikipedia corpus (2), and their corresponding word frequency distribution of positive training examples (red cross) .

- **Word Embeddings**

We obtained word embeddings from the monolingual corpora described above for our Spanish, English and Chinese words in the *right translation* and *no translation* lists using the Continuous Bag-of-words (CBOW) method as implemented in word2vec[7] tool, because it is faster and more suitable for larger datasets (Mikolov et al., 2013a). To train the CBOW models we used the parameters with window size 8, minimum word frequency 5 and 200 dimensions for both source and target vectors. For the ranking experiment, we used 300 dimensions for all vectors.

4.1.3 *Proof-of-Concept* evaluation scenario

For the *proof-of-concept* evaluation, we experimented with different options: using WE obtained from raw corpora and from lemmatized corpora, and training the classifier with different ratios of positive and negative training sets.

We first experimented with the ratio of 5 negative training samples to each positive sample. We chose this imbalanced ratio to approach the actual distribution of the data when inducing bilingual lexicons from monolingual corpora, since there will be many more *no translation* than *right translation* pairs. However, it should be made

[7] https://code.google.com/archive/p/word2vec/

clear that our classifier also performs well on the balanced data as demonstrated in Table 4.6 and Table 4.7.

We trained and tested SVM-SMO classifiers on EN-ES and ZH-ES for three word categories: noun (N), adjective (Adj) and verb (V), and another for the three categories together. Note that this thesis focuses on the phrase table expansion techniques for ZH-ES SMT system. The word embedding-based classifiers described in this section were also evaluated on EN-ES because the *SMT simulation* evaluation scenario is similar to the EN-ES ranking experiment proposed by Mikolov et al.(2013b). We conducted the experiments on the same EN-ES dataset as used by Mikolov et al. (2013b) with the intention to compare their ranking task with our method.

The evaluation was double, as we performed a 10 fold cross-validation with the training set and we tested again the model with the held-out test set. The results are presented in terms of precision (P), recall (R) and F1-measure (F1). The tables below show the P, R and F1-score of both classes (*right translation* and *no translation*) separately of the experiments.

For the first experiment, we learned our ZH, EN and ES word embedding models using lemmatized monolingual corpora, thus all the word pairs were represented by lemmas. The ratio was 5 negatives to each positive. Table 4.2 and Table 4.3 show the test results on the language pairs ZH-ES and EN-ES respectively.

	10 cross-validation			Held-out test set		
	P	R	F1	P	R	F1
YES	0.845	0.796	0.82	0.83	0.809	0.819
NO	0.948	0.962	0.955	0.963	0.967	0.965

Table 4.2: *Proof-of-Concept* test results on lemmatized corpora for ZH-ES with the ratio 5:1.

	10 cross-validation			Held-out test set		
	P	R	F1	P	R	F1
YES	0.804	0.708	0.753	0.782	0.736	0.758
NO	0.944	0.966	0.955	0.948	0.959	0.954

Table 4.3: *Proof-of-Concept* test result on lemmatized corpora for EN-ES with the ratio 5:1.

As shown in Table 4.2 and Table 4.3, for ZH-ES we achieved a F1-score of around 0.82 for *right translation*, and 0.96 for *no translation*; for EN-ES, results are slightly worse with a F1-score of around 0.75 for *right translation* and 0.95 for *no translation*. In terms of accuracy, we obtained 92.8% for ZH-ES and 92.4% for EN-ES.

To evaluate whether the classifier performs better on lemmatized corpora or on corpora in word form, we also conducted an experiment on the same datasets and experimental settings that used for the previous experiment, but with word embedding models trained on the monolingual corpora in word form. The results are demonstrated in Table 4.4 and Table 4.5.

	10 cross-validation			Held-out test set		
	P	R	F1	P	R	F1
YES	0.937	0.919	0.928	0.926	0.871	0.898
NO	0.984	0.988	0.986	0.976	0.987	0.981

Table 4.4: *Proof-of-Concept* test result on corpora in word form for ZH-ES with the ratio 5:1.

	10 cross-validation			Held-out test set		
	P	R	F1	P	R	F1
YES	0.926	0.888	0.907	0.97	0.883	0.924
NO	0.974	0.983	0.979	0.932	0.983	0.957

Table 4.5: *Proof-of-Concept* test result on corpora in word form for EN-ES with the ratio 5:1.

Compared to the results of lemma-based experiment given in Table 4.2 and Table 4.3, the classifier trained on corpora in word form delivers better P, R and F1-score as shown in Table 4.4 and Table 4.5 for both language pairs. The accuracy increased to 96.8% for ZH-ES and 96.5% for EN-ES, demonstrating that our word embedding based classifier performs better on corpora in word form than in lemmatized corpora.

To explore the relation between the performance of the classifier and the number of training instances, Figure 4.5 plots the learning curves (F1-score and kappa value) over different percentage of positive training instances from 100 (10%) to 900 (90%), with negative instances from 500 to 4500, for the language pair Chinese and Spanish. It shows that the classifier achieved stable and good results with around 50% of training instances, and it could benefit from using more training samples.

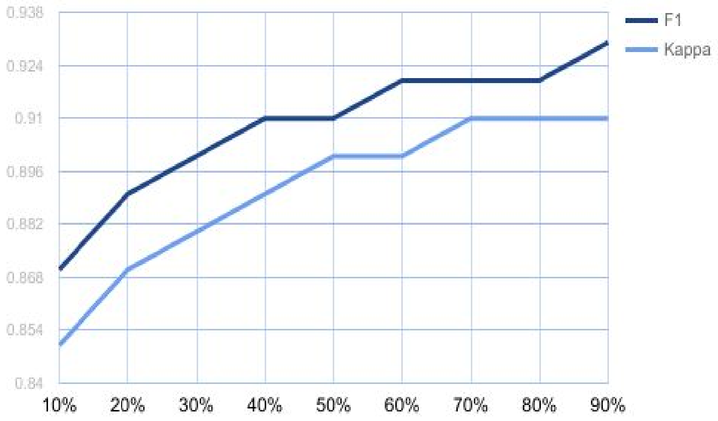


Figure 4.5: Learning curve over different percentage of training data for Chinese and Spanish.

To evaluate whether our classifier can maintain similar perform with balanced training data, we also conducted an experiment on the corpora in word form using the same datasets but with balanced training dataset. The experimental results are demonstrated in Table 4.6 and Table 4.7.

	10 cross-validation			Held-out test set		
	P	R	F1	P	R	F1
YES	0.973	0.96	0.966	0.844	0.934	0.887
NO	0.961	0.973	0.967	0.987	0.967	0.977

Table 4.6: Test results on corpora in word form with balanced dataset for Chinese and Spanish.

	10 cross-validation			Held-out test set		
	P	R	F1	P	R	F1
YES	0.959	0.924	0.941	0.927	0.91	0.918
NO	0.927	0.961	0.943	0.946	0.956	0.951

Table 4.7: Test results on corpora in word form with balanced dataset for English and Spanish.

Compared to results of the experiment of ratio 1:5 given in Table 4.4 and Table 4.5, the classifier trained with balanced positive and negative instances performs similarly on the *right translation* class, but worse on the *no translation* class in all evaluation scenarios for both language pairs as shown in Table 4.6 and Table 4.7. We show several examples of translation equivalents that are correctly classified by our classifier in Table 4.8

CH-ES	EN-ES
友好 (friendly) - amistoso	economic - económico
古老 (old) - antiguo	efficient - eficiente
头 (head) - cabeza	attractive - atractivo
特征 (characteristic) - característica	activity - actividad
烧伤 (burn) - quemar	bed - cama
导致 (provoke) - provocar	language - idioma

Table 4.8: Examples of translation pairs correctly classified by the classifiers.

- **Discussion**

The evaluation results of the *proof-of-concept* experiments showed that a classifier as the proposed one was able to generalize to a large extent. We carry out an error analysis to assess whether the results could be considered an upper bound limit for the task or there was room for improvements. Note that error analysis, however, is hindered by the nature of the used vectors: being word embeddings a projection, no special feature selection study can be easily performed (Levy et al. 2014a). Thus, we mainly looked at the false negative (FN) cases produced by the ZH-ES noun classifier.

After manual inspection, we show the following issues that might have negative impacts on our classifier:

- Word embedding representation obtained with low frequency words is insufficient for the classifier to learn translation relationship between words, since Schnabel et al. (2015) demonstrated word frequency plays a significant role on the quality of word embedding. In line with this reasoning, the following low frequency candidates seem engaged with this problem:

    - Foreign words were normally misclassified, for example: "number" (an English word in the Spanish corpus) was present in the test set with the corresponding translation " 号 " ('number') in Chinese;

- Words that have a wrong, or very unusual PoS tag could only learn its distribution from just few occurrences. This is the case for pairs such as "católica" ('catholic', normally an adjective but was tagged as a noun in our corpus) and "天主教 "('catholicism', a noun in Chinese).

- Word pairs that contained low frequency words (e.g. frequency lower than 100 occurrences), such as *autonómico- 区域性* ('regional' in a geopolitical sense in Spanish) and *carnívoro- 肉食性* ('carnivorous') were also misclassified. We checked whether among the correctly classified pairs there were similar low frequent words, and indeed it was not the case.

❖ For some other errors the explanation is less obvious. We found that, although the translation provided by Google translate could be correct in very particular contexts, there is a semantic difference between the members of the pair: the Chinese word is more general than the Spanish one, or the other way around: "pueblo" (inhabited place or group of people) was paired with "村" (only inhabited place, i.e. village), "reflexivo" ('thoughtful' or 'reflective') was paired with " 反光" ('light reflective'), or "enlace" ('link' but also 'wedding') with "链接" (only 'link' in Chinese). In these cases, the classifier did not find the pair to hold the translation relation.

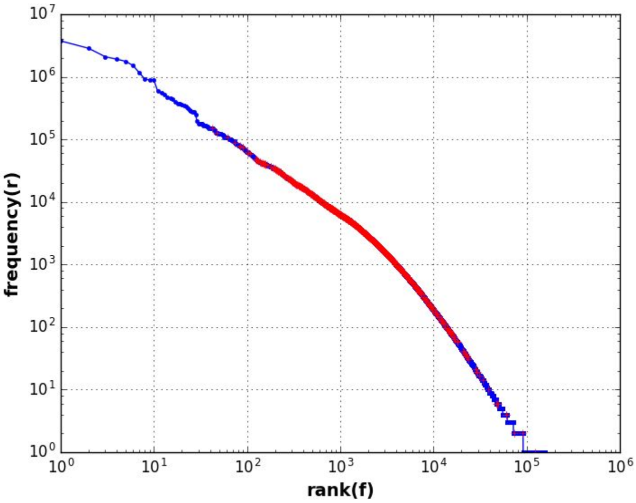
4.1.4. *SMT simulation* evaluation scenario

In the *SMT simulation* evaluation scenario, instead of giving a relatively controlled set of word pairs for testing, each of the source word representation from the test set was concatenated with all target word representations from the corpus following the way that SMT systems produce phrase tables. Each member of the source language test set was paired with all the target language words as possible translation pairs. Then all the concatenated candidates were tested and all the *right translation* pairs delivered by the classifier for each source word were ranked based on the confidence score provided by the classifier, since it is the reliability on the classification decision that ranges from 0 to 1, for a particular instance to belong to a particular class. This scenario is similar to Mikolov et al. (2013b) (described in Section 2.2.2) which evaluated their transformation matrix by finding the top-k nearest candidates for the projected vector from the target corpus.

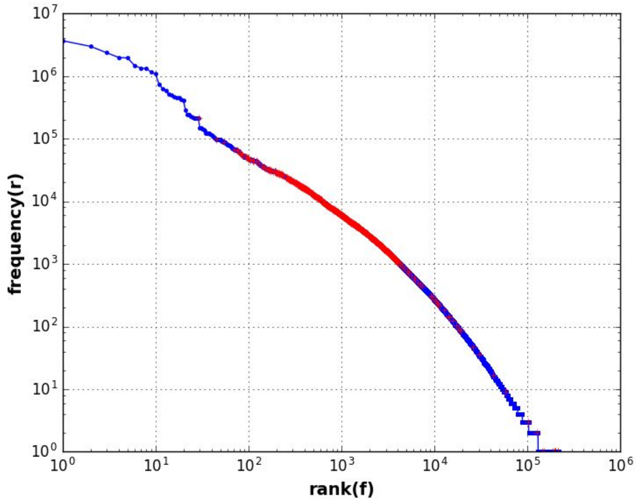
To validate whether our method can locate the correct translation at the top ranking position, we trained a new classifier with EN-ES WMT 11 datasets as used by Mikolov et al. (2013b) following the outline of our previous experiments. The classifier was evaluated in two different ways: the precision, recall and F1-score results of the binary classification task and top-10 ranking task according to its corresponding confidence score. For the binary classification, we again experimented with two different ratios of positive and negative training samples: balanced dataset of positive and negative examples, and five negative instances for each positive example. The experiment was conducted only with nouns. The datasets for training and testing are shown in Table 4.9. To show the frequency range of our positive training examples, Figure 4.6 shows the Zipf plot of word count distribution of the positive training examples (represented in red colour) in the monolingual corpora.

	Training		Testing	
	YES	NO	YES	NO
1:1	990	990	434	832
1:5	990	4950	434	832

Table 4.9: Translation pair datasets for EN-ES on *SMT simulation* evaluation scenario.



(1)



(2)

Figure 4.6: Zipf plot for word counts in monolingual corpora (blue dot) in WMT EN-ES experiment: EN WMT11 corpus (1) and ES WMT11 corpus (2), and their corresponding word frequency distribution of positive training examples (red cross) .

Our experimental results of the binary classification testing using two different ratios are provided in Tables 4.10 and 4.11. Compared to the results of previous classification experiments on the language pair ES-EN (described in Section 4.1.3), we obtained similar performance on 10-cross validation with the same ratio 1:5. However, for the held-out test set, results of both classes decreased unexpectedly due to some low frequent test data, such as *soy*, *signaling*, *underuse* and *skepticism*, whose frequency are only 6, 10, 5 and 78, respectively.

	10 cross-validation			Held-out test set		
	P	R	F1	P	R	F1
YES	0.878	0.895	0.886	0.794	0.71	0.75
NO	0.893	0.876	0.884	0.857	0.904	0.88

Table 4.10: Binary classification results on *SMT simulation* evaluation scenario with balanced dataset for English-Spanish.

	10 cross-validation			Held-out test set		
	P	R	F1	P	R	F1
YES	0.825	0.835	0.83	0.963	0.533	0.686
NO	0.967	0.964	0.966	0.802	0.989	0.886

Table 4.11: Binary classification results on *SMT simulation* evaluation scenario with the ratio 1:5 for English-Spanish.

For the ranking task, our intention was to compare our method with the results of the transformation matrix proposed by Mikolov et al. (2013b). To do so, we created a new test set, in which each source word was paired with all the target language words. We used only nouns, both the source and target side, in order to reduce the computational load, resulting in 24,706 pairs for each source noun. The trained model was used to classify the test set and the confidence score was used to rank all translation pairs classified as *right translation*, expecting to find the correct translation pair ranked in top positions. However, we realized that many word pairs obtained the same confidence score making it impossible to properly set up the ranking list in our case. To better understand the results of our ranking experiment, we give some examples of our test result in Table 4.12. Note that the reported ranking position is with respect to the candidate translations for each source word.

- **Discussion**

Observing the results of *right translation* predicted by our classifier, we noticed that the classifier trained using WE was not efficient in the following cases:

Translation pairs	Ranking position	Confidence score
sugar_azúcar	6	0.978
shipyard_astillero	163	1
square_plaza	197	0.922
tribune_tribuna	362	1
sphere_esfera	512	1
sir_señor	694	0.99

Table 4.12: Examples of the ranking experiment.

(i) **Semantically related candidates.**

Words that always occur in similar contexts or nearby tended to be confusing for the classifier to make the right decision. For instance, the classifier assigned both *turista* ('tourist') and *turismo* ('tourism') as possible translations for the source word *旅游业* ('tourism').

(ii) **Candidates affected by *hubness* problem.**

After the error analysis, we realized that a small group of target words were repeatedly assigned as possible translations to many different source words, such as *parte* ('part'), *nombre* ('name') and *tiempo* ('time'). This could be a consequence of the '*hubness problem*'. (Radovanovic et al., 2010a; Radovanovic et al., 2010b) demonstrated that high-dimensional spaces contain certain elements – *hubs* – that are near many other points in space without being similar to the latter in any meaningful way. The mechanisms behind the *hubness* is that points that are located closer to the mean of the data distribution are, on average, closer to all other points. As dimensionality increases, stronger correlation emerges, implying that points closer to the mean tend to become *hubs*. As recently noted by Dinu et al. (2015), the *hubness problem* is greatly exacerbated when one looks at the nearest neighbours of vectors that have been mapped across spaces with ridge. Lazaridou et al. (2015) also addressed this problem in cross-space (cross-modal and cross linguistic) mapping function learning, and they proposed to improve the performance by replacing the ridge estimation by max-margin. In our method, instead of looking for the nearest neighbours of projected vectors in the target language spaces, we treated the translation process as a simple binary classification problem. Radovanovic et al. (2010a) examined the influence of bad *hubs*[8] on

[8] Radovanovic et al. (2010a) defined bad *hubs* as points with high *BNk (x)*, which is the number of "bad" k-occurrences of x. Namely, the number of points from the data set for which *x* is among the first *k* NNs, but the labels of *x* and the points in question do not match.

classification algorithm as well, such as SVMs (the same algorithm as we used). They claimed that for high-dimensional data, points with high *BN$_k$* can comprise good support vectors since their experimental results show that the accuracy significantly drops with removal by *BN$_k$*, indicating that bad *hubs* are important for SVMs.

4.1.5. Summary

In this section, we have proposed a novel method to learn bilingual lexicon from monolingual corpora by training a supervised classifier. On average, we obtained quite good results on *Proof-of-Concept* evaluation scenario. However, we could not compare the results of the ranking task as proposed by Mikolov et al. (2013b), since many word pairs obtained the same confidence score making it impossible to properly set up the ranking list. Besides, due to the *hubness* problem, a number of particular target words were repeatedly classified as possible translation of many different source words with high confidence score.

Despite the fact that the confidence score supplied by the classifier is insufficient to tackle the ranking task, judging from its outstanding performance of the classification experiment, it was expected to be useful for being applied to expand phrase tables of

SMT systems when no parallel or comparable corpora is available. In the next section, we report on two experiments adding further additional monolingual information to the word embedding-based classifier for improving the classification performance before applying this method to enrich our baseline phrase table.

4.2 Improving Word Embedding-based classifier with additional word frequency information

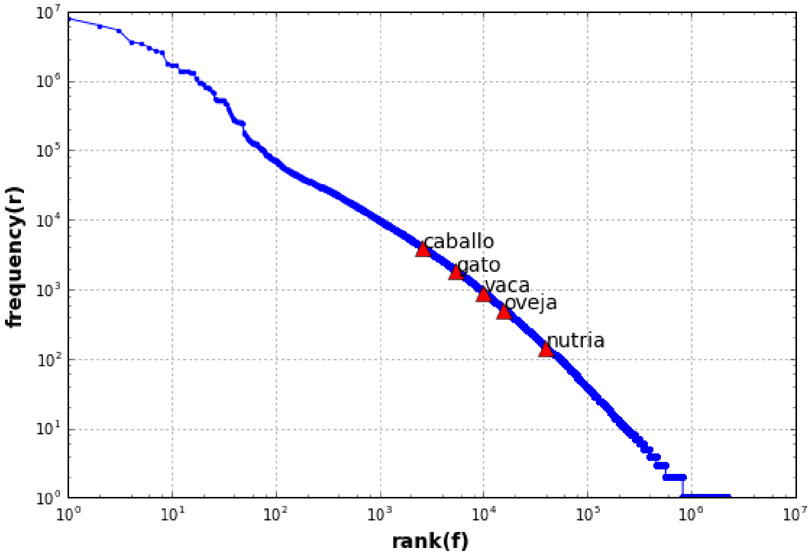
In Section 4.1 we proposed a word embedding-based supervised classifier to induce a bilingual lexicon from monolingual corpora. According to the errors delivered by the model trained with word embedding, *hubs* were found to be one of the problems that affect the classifier performance. According to (Newman et al., 1983; Newman and Rinott, 1985 and Radovanovic et al., 2010b), the *hubness* phenomenon is an inherent property of data distributions in high-dimensional space under widely used assumptions, and not an artefact of a finite sample or specific properties of a particular data set. Radovanovic et al. (2010a) also claimed that for SVM classifiers, those bad *hubs* with high *BNk* (described in section 4.1.4) could be good support vectors. As decreasing *BNk* can not promise a better performance for our classifier, we decided to alleviate the situation by adding additional monolingual features to the word embedding vectors. In the experiment described in this section, we report on how we incorporated word frequency information to our word embedding vectors during the training process and explored the impact of word frequency feature on bilingual lexicon induction.

Observing the results delivered by the classifier described in Section 4.1, we realized that the source and target word of many bad translation pairs are quite different in their frequencies[9]. These results give rise to a question: within non-parallel corpora, is the source word frequency related to the frequency of its relative translation? One of the interesting facts about human language is that given a natural language text corpus, the frequency of any word is inversely proportional to its rank position and the distribution of the word frequency (WC) roughly follows the mechanisms of Zipf's law. From the point of view of our human languages, words are used to convey an intended meaning. Therefore, the word frequency distribution can be seen as "need distribution" for how often we need to communicate each meaning (Piantadosi, 2014), hence it is a general property of word distribution across different languages. Calude and Pagel (2011) examined Swadesh lists[10] of 17 languages

[9] Note that we are not trying to prove that word frequency is one of the crucial factors that cause the *hubness problem*. Instead, we just give the reason why we decided to incorporate word frequency information.

[10] Swadesh lists provides translations of simple, frequent words like "mother", "happy" across many languages; they are often used to do historical reconstruction.

from six language families and compared frequencies of words on the list. They reported an average inter-language correlation in log frequency of $R^2 = 0.53$ ($p < 0.0001$)[11] for these common words, indicating that word frequencies are surprisingly robust across languages and predictable from their meanings. Smith (2008) also showed that word distributions are observed universally in languages, even in extinct and yet-untranslated languages like Meroitic. To verify the frequency-rank relationship across languages, Figure 4.7 gives the frequency-rank plot of several translation pair examples in our source and target monolingual corpora.



[11] Note that Swadesh words will tend to be high-frequency, so the estimated R2 is almost certain to be lower for less frequent words(Calude and Pagel, 2011).

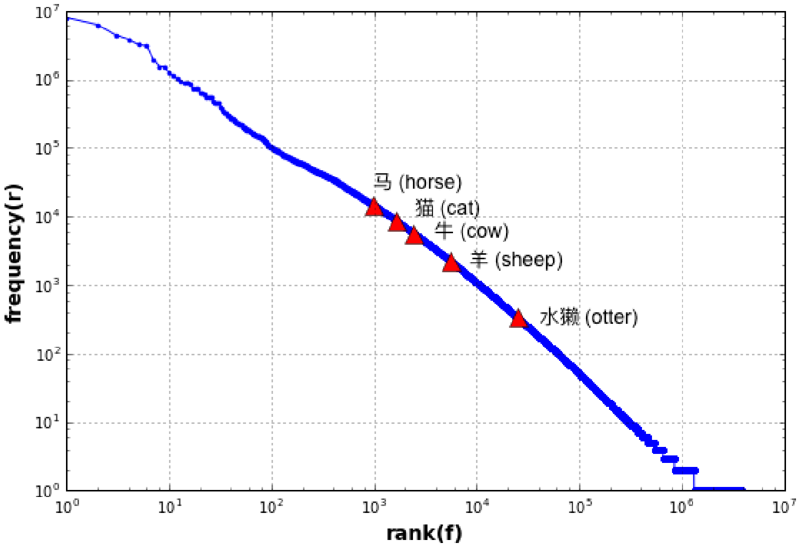


Figure 4.7: word frequency-ranking plot of translation examples in our source and target monolingual corpora.

As these examples reveal, the frequency distribution is similar across languages. We are inclined to believe that even with monolingual data, the frequency of words in the source monolingual corpus is somehow correlated to the frequency of their relative translations in the target corpus. In this section, we explored how monolingual word frequency can be used to improve the performance of word embedding-based classifier.

There have been several works that suggested to use word frequency information for learning translations from non-parallel data. For instance, Koehn and Knight (2002) combined word frequency together with various other clues, such as cognate, similar context and preservation of word similarity, to induce a word-level translation lexicon from monolingual corpora. They assumed that for most words, especially occurring in comparable corpora, there is a considerable correlation between the frequencies of a word and its translation. The frequency measurement was defined as a ratio of both word frequencies, normalized by the corpus size. Koehn and Knight (2002) used a greedy search to look for the best translation for a given source word. First they searched for the highest score for any word pair. This word pair was added to the lexicon, and not included in future searches. Then they repeated the first step, searched for the highest score and added the corresponding word pair, dropped these words from further search, and so on. This was done iteratively, until all words were used up. The experimental results reveal that only using word frequency clue was too imprecise to pinpoint the search to the correct translations, but when combining the word frequency to some other feature, such as spelling, it indeed provided valuable information for learning bilingual lexicons out of monolingual data.

Schafer and Yarowsky (2002) also explored the usage of word frequency together with other features to induce a translation lexicon from non-parallel corpora. They worked with the hypothesis that a word and its translation are likely to have a similar relative frequency in the corpora of their respective languages. For their experiment, instead of using directly the word count or the relative frequency, they used a word frequency similarity score. It was found that a simple ratio of logs of frequencies correlate well with translational compatibility and was proved to be an improvement for the ranking task when searching for the correct translation among candidates for a given source word.

As mentioned in section 2.2.2, Irvine and Callison-Burch (2013) applied a supervised classifier to induce translation pairs by using monolingual word frequency together with other signals such as orthography, topic, temporal signature (from time-stamped web crawl data) and context information. They also assumed that words that are translations of one another are likely to have similar relative frequencies in their respective monolingual corpora. To include word frequency information, they calculated the frequency similarity of two words as the absolute value of the difference between the logs of their relative monolingual corpus frequencies. The experimental result demonstrates that word frequency feature indeed has positive impact on bilingual lexicon induction.

The experiment presented in this section was inspired by these previous researches. In our case, instead of using word relative frequency and frequency similarity between word pairs to enhance our classifier, we explored using word frequency standard deviation of the source and target monolingual corpus to define our frequency feature.

4.2.1. Approach

Standard Deviation is a measure that is used to quantify the dispersion of a set of data values (Bland and Altman, 1996). Since it has been proved that word frequency distribution is similar across languages, we assumed that the frequency dispersion degree of the same concept in different languages should be similar as well. Figure 4.8 shows the word frequency distribution of several translation examples based on the standard deviation of word frequencies of our source and target monolingual corpora.

According to the word frequencies shown in the Figure 4.8, translations of the words that classified into the same sigma in the source corpus, are also classified into a corresponding sigma in the target corpus. Therefore, our hypothesis was that word frequency can help the classifier to improve the classification performance on the basis that if a target candidate is not the translation equivalent it might have a very different frequency distribution compared to the paired source word.

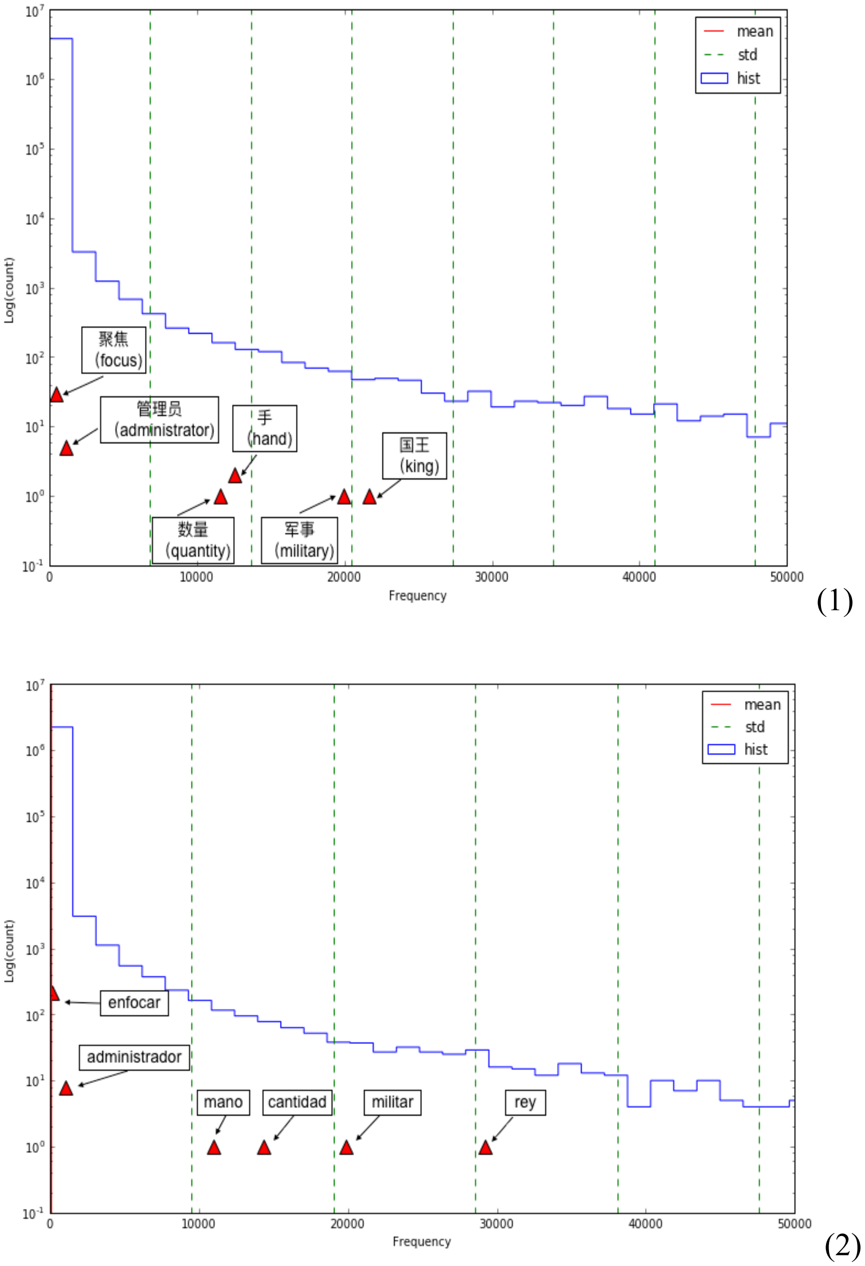


Figure 4.8: Word frequency distribution of several translation examples in the Chinese (1) and Spanish (2) monolingual corpus.

In this experiment, we used standard deviation (sigma) to define the grouping of each source and target word based on their word frequency distribution. The calculation of the word frequency standard deviation of monolingual corpus can be described by Equation (1):

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^{2}}$$

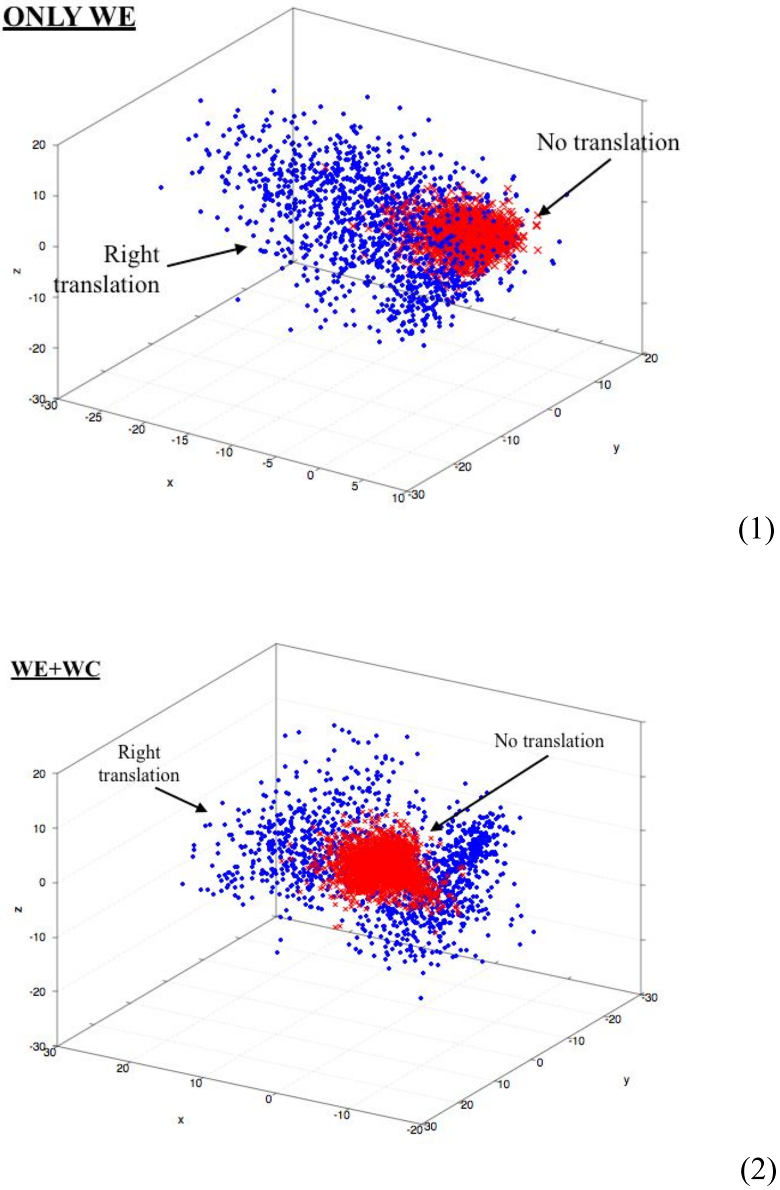
Where, σ is the standard deviation and N is the number of words in the monolingual corpus, x_i and $\bar{x}$ refer to the raw frequency of each word and the mean of all frequencies, respectively.

We assigned each word to its corresponding frequency group according to the sigma from left to right as $g = g_1, g_2, g_3 ... g_j$. Then the definition of g_j can be described as the Equation (2):

$$g_j = \begin{cases} \left\lceil \frac{x_i}{\sigma} \right\rceil, \text{if } g_1 = \sigma & \textcircled{1} \\ \\ \left\lceil \frac{(x_i - g_1)}{\sigma} \right\rceil + 1, \text{if } g_1 < \sigma \text{ and } x_i > g_1 & \textcircled{2} \end{cases}$$

As shown in the equation, the grouping process was divided into two different situations: (a) when the frequency range of the first group g_1 is equal to the sigma σ, to calculate to which group a given word belongs, we simply divided its raw frequency x_i by the sigma σ (as shown in ①). If the obtained result was not divisible by the sigma, we rounded it up to get an integer; (b) when the range of the first group g_1 is smaller than the sigma σ and $x_i > g_1$, we defined the word grouping as calculated in ②. For instance, given a word frequency $x_i = 8$, if the range of the first group $g_1 = 1$, sigma $\sigma = 3$, then the corresponding grouping of the given word should be: $g_j = \lceil (8 - 1) / 3 \rceil + 1 = 4$. Note that if a word frequency $x_i < g_1$, it directly belongs to the the first group g_1.

To visualize the impact of word frequency on translation classification, in Figure 4.9, we compare the geometric arrangement of 6K word pairs (*right translation* and *no translation*) represented by only WE vectors (demonstrated in Section 4.1) and with additional WC information in a 3-dimensional space. It demonstrates that the joint representation indeed encodes relevant information for the classification.



(1)

(2)

Figure 4.9 Distributed representations of 6K word pairs (1K *right translation* and 5K *no translation*) with WE of 400 dimensions (1) and with combination of WE and WC of 2048 dimensions (2). We used PCA to project high dimensional vector representations down into a 3-dimensional space.

4.2.2 Experimental setup

This experiment was conducted on the language pair Chinese and Spanish. The outline can be described as following: (i) Calculating the word frequency Standard Deviation of source and target monolingual corpora; (ii) Obtaining the corresponding word frequency feature (as described in Section 4.2.1) and word embedding vector for the positive and negative training word pairs (the examples used in Section 4.1) from the monolingual corpora; (iii) Training a SVM classifier using the concatenated features (WE+WC and WC) of source word and its translation equivalent (or random word for negative instances); (iv) Evaluating the classifier.

- **Classifier datasets**

We used the same ZH-ES training and testing sets as used for the word embedding based experiment described in Section 4.1 so that the impact of word frequency feature can be fairly compared and evaluated. We chose the imbalanced ratio (5 negative instances for each positive one) to train the classifier, since in Section 4.1, it has been proved that the average performance (on both *right translation* and *no translation)* of imbalanced datasets was better than the results produced with balanced datasets.

Following the previous translation induction experiment conducted in Section 4.1, we built the new classifier applying SVM-SMO with gaussian kernel using the joint representation of word embedding and word frequency information as features. The model was evaluated by both 10 fold cross-validation and a held-out test set.

- **Word frequency feature and word embedding**

To train word embedding models, we used the monolingual corpora that used in the experiments in Section 4.1: Chinese Wikipedia Dump corpus (149M words) and Spanish Wikipedia corpus (150M words, 2006 dump). Same parameter settings were applied: window size 8, minimum word frequency 5 and 200 dimensions for both source and target vectors.

The word frequency features were induced from the same monolingual corpora that used for learning word embedding vector. The mean and standard deviation of the Chinese monolingual corpus are 38.89 and 6832.95, respectively. Regarding the Spanish monolingual corpus, the mean and standard deviation are 58.43 and 9519.84, respectively.

In order to include word frequency distribution feature in word pair representations, we used one-hot encoding to represent each standard deviation. In concrete, each word embedding concatenated vectors were added 2048 binary features: 1289 for the source word and 759 for the target word. Each component represents one of the 1289 sigmas for each source word and one of the 759 sigmas for each target word.

4.2.3. Evaluation

Table 4.13 shows the results delivered by the classifier trained with the joint representation of word frequency (WC) and word embedding (WE) in terms of precision (P), recall (R) and F1-score (F). Besides, for a better observation and comparison, we also give the results delivered by the classifier trained only with WC and WE (results given in Table 4.4) separately.

		10 cross-cross validation			Held-out testset		
		P	R	F1	P	R	F1
WE	YES	0.937	0.919	0.928	0.926	0.871	0.898
	NO	0.984	0.988	0.986	0.976	0.987	0.981
WC	YES	0.707	0.932	0.804	0.745	0.92	0.823
	NO	0.989	0.939	0.963	0.984	0.939	0.961
WE+WC	YES	0.96	0.94	0.95	0.958	0.92	0.939
	NO	0.988	0.992	0.99	0.983	0.991	0.987

Table 4.13: Test results of the Chinese-Spanish classifier trained with WE, WC and the combination of WE and WC.

As shown in Table 4.13, the model trained with the joint representation of WE and WC outperforms the WE-based classifier and WC-based classifier. After adding WC to WE, both precision and recall were improved. We achieved F1-score of 0.939 for *right translation*, and 0.987 for *no translation*. In terms of accuracy, the performance was increased from 96.8 to 98.3 compared to the model trained with WE, demonstrating that adding word frequency information indeed encode useful information for discarding those word pairs with large difference in their frequencies.

4.2.4 Discussion

Before moving to the result analysis, we would like to clarify the following questions that people might have regarding this experiment:

1. ***Challenge of high dimensional feature space.***

   In this experiment, we used in total 2449 dimensional features to train our WE+WC classifier. According to Koppen (2000) and Duda et al. (2001), the accuracy of classification algorithms tends to deteriorate in high dimensions due a phenomenon called the *curse of dimensionality*. Besides, a reduced number of training samples in high dimensional data settings cause the classification model to overfit to the training data, thereby having poor generalization ability for the model, especially for those conventional classifiers such as logistic regression, maximum likelihood classification etc. (Pappu and Pardalos, 2014).

However, in our method, the classifier was trained using SVM. It has been proved by many researches (Pal and Mather, 2005; Joachims, 1998; Vapnik and Chapelle, 2000) that SVM is more effective in high dimensional data space compared to other traditional classifiers. To explain why SVM can overcome the overfitting problem and the curse of dimensionality, Vapnik et al.(2000) proposed the concept of error bonds, according to which the generation ability of a SVM classifier does not depend on the dimensionality of the input space, instead, it is inversely dependent on the sample size and the margin between the parallel planes. So we believe that our high dimensional features do not have a negative impact on our model.

2. ***Negative impacts of high dimensional one-hot encoding?***

In the previous point we have clarified the capacity of SVM to handle high dimensional feature space. Another possible concern might be that a large number of zeros might render ineffective the class membership decision of the classifier. However, in our case, kernel-based SVM does not operate on the source feature space directly, instead, the data are mapped into a higher dimensional feature space through a particular nonlinear transformation, then the hyperplane is achieved by using the inner products in the transformed space. So we are inclined to believe that high dimensional one-hot encoding feature does not have a negative impact and it indeed encode relevant information into our word embedding-based classifier since our obtained results shown in Section 4.2.2 confirm this assumption as well.

Looking back to the obtained results, the classifier enhanced with the additional word frequency information achieved better recall, especially for the *right translation* class. For instance, with word embedding-based classifier, the translation pairs *番木瓜_papaya, 降水量_precipitación* and *发音_pronunciación* were classified as *no translation*, but after adding word frequency feature, they were correctly classified since according to their frequency feature, the source word is distributed very similarly to its corresponding translation. Regarding the performance of the enhanced model on *no translation* class, although the recall was increased as well, and most of the false positives delivered by the word embedding-based classifier were correctly classified, a few new false positives were produced by the WE+WC classifier, and most of them were low frequency words. So the word frequency feature might not be efficient to handle very low frequency candidates.

4.2.5. Summary

In this section, we proposed to improve the performance of the word embedding-based classifier by adding frequency features of word pair. This assumption is based on large source and target monolingual corpus of general domain. The experimental results reveal that after adding word frequency information to word embedding vectors, both the precision and recall were improved, especially in terms of recall of *right translation* which was increased from 0.871 to 0.92. The obtained results confirmed the hypothesis that the distribution of word frequencies extracted from monolingual corpora is useful for the classifier to learn the translation relationship between words, as well as to filter out those word pairs with large difference in their frequencies.

4.3 Improving Word embedding-based classifier with additional Brown cluster features

In Section 4.2 we proposed to add word frequency feature to the word embedding vector to improve the classification performance. In this experiment, we incorporated Brown clustering information to our word embedding vectors and evaluated its impacts compared to the word embedding-based classifier.

Brown clustering is a representation of word semantics learned from monolingual corpus (Turian et al., 2010) in an unsupervised way. It uses mutual information to determine distributional similarity, placing similar words in the same cluster and similar clusters are located in a hierarchical binary tree. To obtain a Brown cluster model, the input is a text and the output is a binary tree, in which the leaves of the tree correspond to the words in the vocabulary, and the roots correspond to the clusters. Intermediate nodes of the tree can be interpreted as groups containing the words in the subtrees. Figure 4.10 shows some of the substructures in the binary tree given by Brown et al. (1992). Initially, the algorithm starts with each word in its own cluster. As long as there are at least two clusters left, the algorithm merges the two clusters that maximizes the quality of the resulting clustering. Note that each word belongs to only one cluster.

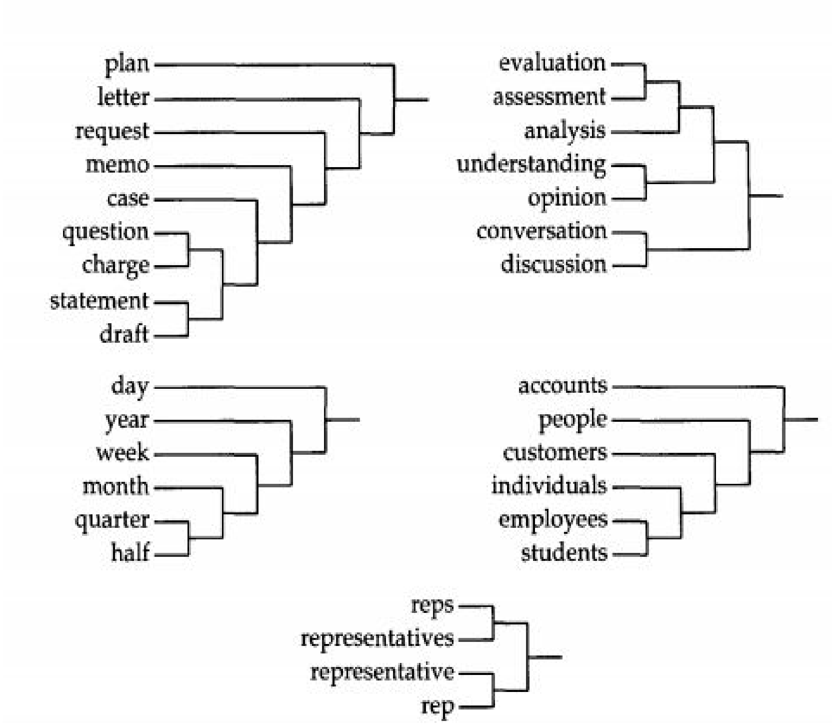


Figure 4.10: Sample subtrees given by Brown et al.(1992).

There have been several works (Zhao et al. 2005; Täckström and McDonald, 2012) to demonstrate that word clustering provides relevant information for cross-lingual tasks. For instance, Och (1999) developed an optimization criterion based on a maximum likelihood approach and described a clustering algorithm to determine bilingual word clustering suitable for statistical machine translation. Birch et al. (2013) explored using source and target

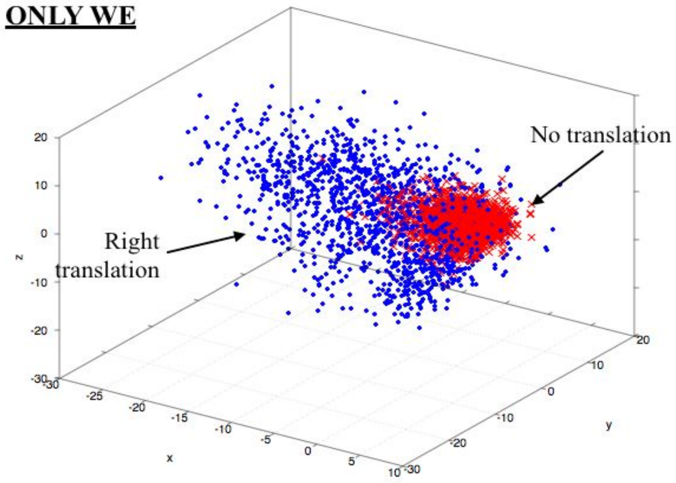
Brown clusters in factored-based phrase translation model and the operation sequence model. Their experimental results showed that the integration of Brown clustering information consistently enhanced the baseline system giving significant improvements in most cases. Therefore, in this section, we evaluated the impact of the additional Brown clustering information on our word embedding-based classifier.

Observing our data, semantically related words in the source monolingual corpus are grouped into the same class, while their translations belong to a corresponding class in the target monolingual corpus as well. Table 4.14 lists several example translation pairs from our monolingual datasets, with their respective word clustering (c=200) information. According to the examples, similar Chinese source words are grouped into same classes, as well as their Spanish translations.

01011001 英格兰 (England)	10111100 inglaterra
01011001 巴塞罗那 (Barcelona)	10111100 barcelona
01011001 日本 (Japan)	10111100 japón
011111110110 演员 (actor)	11010100 actor
011111110110 记者 (journalist)	11010100 periodista
01010100 开心 (happy)	1010010101 agradable
01010100 悲伤 (sad)	1010010101 triste

Table 4.14: Brown clusters of several translation pair examples.

To visualize the impact of using BC, in Figure 4.11, we compare the geometric arrangement of 6K word pairs (*right translation* and *no translation*) represented by only WE vectors (demonstrated in Section 4.1) and with additional Brown cluster (BC) information in a 3-dimensional space. It demonstrates that the joint representation indeed encodes relevant information for the classification.



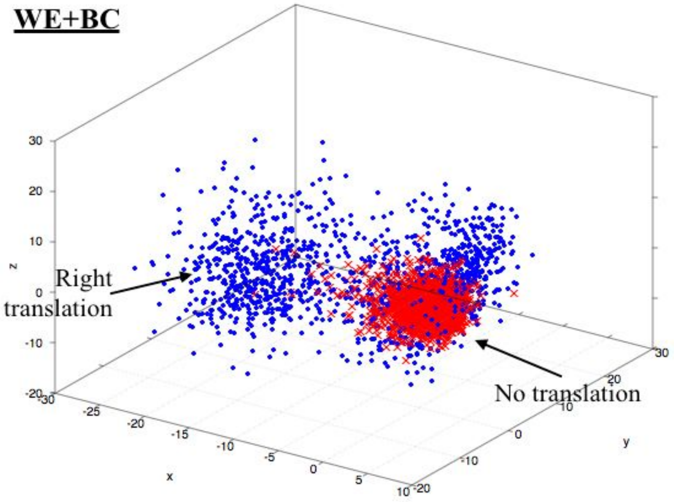


Figure 4.11: Distributed representations of 6K word pairs (1K *right translation* and 5K *no translation*) with WE of 400 dimensions (1) and with combination of WE and BC of 800 dimensions (2). We used PCA to project high dimensional vector representations down into a 3-dimensional space.

4.3.1 Experimental Setup

The outline of this experiment can be described as: (i) Obtaining the corresponding word embedding vector and (ii) Brown clusters for the positive and negative training examples from monolingual corpora. (iii) Concatenating the representation features of the source word and its translation equivalent (or random word for negative instances). (iv) Training a SVM-SMO classifier using the previously concatenated representation. (v) Evaluating the classifier.

- **Classifier datasets**

This experiment was conducted on the language pair Chinese-Spanish. We used the same ZH-ES training and testing sets as used for the word embedding-based experiments described in Section 4.1 so that the impact of Brown cluster feature can be fairly evaluated. We chose the imbalanced ratio (5 negative instances for each positive one) to train the classifier, since based on the results shown in Section 4.1, the average performance (on both *right translation* and *no translation)* of imbalanced dataset were better than the results produced with balanced dataset.

- **Word embedding vectors and Brown clusters**

To train word embedding model, we adopted the monolingual corpora that were used in word embedding based experiment: Chinese Wikipedia Dump corpus (149M words) and Spanish Wikipedia corpus (150M words, 2006 dump). Same parameter settings were applied: window size 8, minimum word frequency 5 and 200 dimensions for both source and target vectors. Our Brown clustering representation were induced from the same monolingual corpora that used for learning word embedding vector. We set c=200 for computational cost savings, although with larger number of clusters it might perform better. In order to include Brown clustering in word pair representations, instead of using directly the bit path, we used one-hot encoding. More specifically, 400 binary features were added to the word embedding concatenated vectors: 200 for each word. Each component represents one of the 200 word clusters for each source and target word.

- **SVM Classifier**

We built and tested a SVM classifier on the language pair Chinese-Spanish using the joint representation of word embedding and Brown clustering. Our SVM classifier was trained with the gaussian kernel using SMO following the same experimental settings as applied in word embedding-based experiment described in Section 4.1. The model was evaluated by both 10 fold cross-validation and a held-out test set.

4.3.2. Evaluation

In this section, we present the evaluation results of the classifier trained with word embedding vector plus Brown clustering feature. For a better observation, Table 4.15 shows the comparison of evaluation results delivered by our classifier trained with WE (given in Section 4.1.3, Table 4.4), BC and with the combination of WE and BC in terms of precision (P), recall (R) and F1-score (F).

		10 cross-cross validation			Held-out test set		
		P	R	F1	P	R	F1
WE	YES	0.937	0.919	0.928	0.926	0.871	0.898
	NO	0.984	0.988	0.986	0.976	0.987	0.981
BC	YES	0.781	0.792	0.787	0.742	0.821	0.779
	NO	0.957	0.955	0.956	0.965	0.945	0.955
WE+BC	YES	0.955	0.935	0.945	0.955	0.92	0.937
	NO	0.987	0.991	0.989	0.985	0.992	0.988

Table 4.15: Test results of the Chinese-Spanish classifier trained with WE, BC and the combination of WE and BC.

As shown in the table, compared to the classifier trained only with WE and only with BC, the classifier trained with the joint representation of WE and BC improved both the precision and recall. In terms of accuracy, the performance improved from 96.8 to 97.6.

In order to verify whether the classifier was learning that particular BCs were associated to right or wrong translation categories, we checked clusters' distribution in our datasets. Table 4.16

demonstrates the distribution of Brown clusters in *right translation* (Yes), *no translation* (No) and in both classes)Yes∩No).

	Training			Testing		
	Yes	No	Yes∩No	Yes	No	Yes∩No
ZH	60	176	57	26	140	23
ES	53	152	52	39	113	31

Table 4.16: Brown clusters distribution in datasets.

4.3.3. Discussion

Analysing the results delivered by the WE+BC classifier, many false negatives delivered by the word embedding-based classifier were correctly classified by the enhanced model resulting in a considerable improvement of the recall of *right translation* class. For instance, with only WE, the translation pairs *诗人_poeta, 王子_príncipe, 心理学家_psicólogo and 扩张_expandir* were classified as *no translation*, while the WE+BC model was capable of delivering the correct classification results. Regarding the performance of *no translation* class, the situation is similar to the WE+WC classifier: although the recall was increased and most of the false positives delivered by the WE classifier were correctly classified, the WE+BC classifier also generated several new false positives which are mostly different from the false positives delivered by the WE classifier and WE+WC classifier. Such situation might be caused by the irregular result of Brown clustering model, since according to Derczynski et al. (2015), normally Brown clustering with classes of more than 1K obtain relatively stable and good performance on different NLP tasks, but in our case, we only used BC with 200 classes.

Comparing the classification performance of WE+WC classifier and WE+BC classifier, for the *right translation* class, the recall increased from 0.871 to 0.92 in both cases. For the *no translation* class, WE+BC classifier increased the recall from 0.987 to 0.992, while the WE+WC classifier delivered an improvement from 0.987 to 0.991. Considering that two enhanced models performed similarly, and there is no obvious evidence to explain how those new false positives arose in each case, we chose the WE+BC model to enrich our SMT baseline phrase table[12].

4.3.4. Summary

In this section, we proposed to improve the performance of the word embedding-based classifier by adding Brown clustering

[12] Our goal here is to evaluate whether the wrong translation candidates delivered by the classifier can be well handled by the SMT system. Since two improved models (one trained with WE and BC and one trained with WE and WC) performed similarly, we chose one of them to expand the baseline phrase table simply for time saving reason.

information of word pairs. In general, both the precision and recall were improved, especially in term of recall which was increased from 0.871 to 0.92. The obtained results addressed the positive impact of using monolingual Brown clustering of source and target word on the task of bilingual lexicon induction. The better recall ensures that more reliable right translations can be delivered to expand SMT phrase table.

4.4. SMT phrase table expansion using induced bilingual lexicon

In this section, we report on how we applied the translation lexicon delivered by the WE+BC classifier to expand the SMT baseline phrase table. As described in Section 2.2, there have been many researches that aim at enriching SMT phrase table using monolingual resources, most of which add new translation pairs directly into the phrase table by creating pairwise probabilities depending on related translation pairs of the baseline system. However, in the present work, our bilingual lexicon was delivered by the classifier from extra monolingual corpora, there is no bilingually estimated translation probability available for our new translation pairs. Therefore, we decided to append the new translation lexicon directly to the training corpus so that the translation probabilities of new translation pairs can be generated by the SMT system itself.

4.4.1. Experimental setup

This experiment was conducted on the language pair Chinese-Spanish. The model trained with WE and BC (described in Section 4.3) was used to produce the new translation lexicon from monolingual corpora following the *SMT simulation* evaluation scenario described in Section 4.1.4: the representation of each given source word was concatenated with all target word representations (with same PoS) from the corpus. Then all the concatenated candidates were tested and the *right translation* delivered by the classifier were finally added to the parallel training data for the phrase tale expansion.

- **Phrase-based SMT setup**

Our SMT system was built using Moses phrase-based MT framework (Koehn et al., 2007b). We used *mgiza* (Gao and Vogel, 2008) to align parallel corpora and *KenLM* (Heafield, 2011) to train a 3-gram language model. We applied standard phrase-based MT feature sets, including direct and inverse phrase and lexical translation probabilities. Reordering score was produced by a lexicalized reordering model (Koehn et al., 2005). Similar to the morphology expansion experiment described in Section 3, the parameter *Good Turing* was applied in order to reduce overestimated translation probabilities. For the evaluation, we used BLEU metric (Papineni et al., 2002).

The parallel corpora used were: Chinese-Spanish OpenSubtitles2013[13] (1M sentences) for training; TAUS translation memory[14] (2K sentences) and UN corpus[15] (2K sentences) for testing. For tuning, we randomly sampled 1K sentence pairs from the News Commentaries parallel corpora released by Tiedemann (2012)[16]. To train the language model, we combined Spanish Wikipedia corpus with OpenSubtitles2013 target corpus.

The classifier was used to deliver, for each of about 3K selected source words (the most frequent words that were not present in the baseline phrase table), all the possible translation candidates as found in the combination with the 30K target language vocabulary of the same PoS (for computational savings). All word pairs classified as *right translation* were then appended to the existing parallel corpora for training a new SMT system. Algorithm 4.1 shows the generation and integration of the new induced bilingual lexicon.

[13] http://opus.lingfil.uu.se/OpenSubtitles2013.php
[14] http://www.tauslabs.com/
[15] http://opus.lingfil.uu.se/UN.php
[16] http://opus.lingfil.uu.se/News-Commentary.php

```
Input: Vector representations of 3K source  words S1;
          Vector representation of all target words T1;
          Supervised classifier model M;
          Parallel corpora for SMT baseline L1
Output: Expanded parallel corpora L2
for each source word vector V(x) in S1 do
   for each target word vector V(y) in T1 do
      if PoS of source word x and target word y are the same then
         concatenate V(x) with V(y);
         append concatenation V(x,y) to C;
      end
   end
end
for each concatenation V(x, y) in C do
   test V(x, y) using M;
   if V(x, y) is classified as right translation then
      append the word pair (x, y) to L1;
   end
 end
```

Algorithm 4.1: Algorithm for generation and integration of supervised bilingual lexicon.

4.4.2 Experimental results and discussion

Table 4.17 shows experimental results of the SMT system trained using the enriched parallel corpora. The system was tested on two different test sets (described in Section 4.4.1) and measured by BLEU metric (Papineni et al., 2002) and Out of Vocabulary rate (OOV).

Setup	TAUS		UN	
	BLEU	OOV	BLEU	OOV
Baseline PhT	8.8	9.6%	10.81	6.8%
PhT + 3K SBL	9.58	8.7%	11.42	5.9%

Table 4.17: Test results of the baseline system and the expanded SMT system.

According to the results shown in the table, with the new translation candidates given by our classifier, the performance of the SMT system improved with respect to the baseline by up to +0.70 and +0.61 BLEU scores, and the OOV[17] rate of the baseline system was reduced around 0.9% for both test sets. The improvements are statistically significant according to the paired student t-test at the level of $p < 0.05$.

As demonstrated in the error analysis of the *SMT simulation* evaluation scenario (described in Section 4.1.4), when this method was applied to search the correct translation for a given source word among all target word candidates, a small group of target words were repeatedly assigned as possible translation to many different source words. Compared to the bilingual lexicon delivered by the word embedding-based classifier, the incorporation of BC feature resulted in a considerable reduction of word pairs classified as *right*

[17] The OOV words were generated as shown in http://www.statmt.org/moses/?n=Advanced.OOVs

translation. Note that 88.65 M word pairs were presented to the classifier from 2955 source words combined with 30K target words. The WE classifier delivered a 7% word pairs classified as *right translation*, while the WE+BC classifier delivered only a 2.7%. Observing the results delivered by the WE+BC classifier for the phrase table expansion, some of the highly repeated bad *hubs* produced by the WE classifier such as *parte* ('part'), *nombre* ('name') and *tiempo* ('time') were successfully discarded after adding the BC feature. The obtained results demonstrated that *hubness* problem can be alleviated to some extent by adding BC information.

- **Discussion**

Different from the morphology expansion experiment described in Chapter 3, the phrase table expansion with induced bilingual lexicon was not dependent on the knowledge of the baseline phrase table. To evaluate the impact of these new translations generated from extra monolingual corpora on the baseline system, we performed a systematic analysis of the translation results produced by the enriched SMT system and compared it with the translation output delivered by the baseline system. Inspired by the work of (Vilar et al., 2006; Costa et al., 2016), we applied the following taxonomy for the analysis:

- **Lexical level**

This category includes all errors related to the way each word, as a whole, is translated. The following types of lexical errors are taken into account: *omission* (described in Section 3.4), *untranslated item* [18] and *wrong lexical choice*[19]. Among all the errors of this level, *omission* and *wrong lexical choice* were found to be the most dominant for both systems (baseline and the expanded SMT system). For instance, in the example (1), the source word 建议 was ignored by both systems. But compared to the baseline, the expanded system performed relatively better in some cases. For instance, in the example (2), the source phrase 继续审查 was well translated as *sigue la revisión de* by the expanded system, and the baseline, by contrast, omitted to translate the word 继续. This result reveals that although a certain amount of bad entries were induced by the classifier, the alignment of the original training corpus was improved by the new bilingual lexicon to some extent since the phrase pairs were splitted into smaller units.

[18] *Untranslated item* refers to Out-of-vocabulary word.

[19] *Wrong lexical choice* errors are found when the system is unable to find the correct translation of a given word. Note that in this case, we only take into account the incorrect word that has no semantic relation with the source word.

(1)

Source:
2000年12月4日第81次全体会议根据委员会**的建议**（
A/55/602/ Add.3，第49段） 经记录表决，以67票赞成，
54票反对，46票弃权通过.
(Approved in the 81st plenary meeting, on 4 December 2000, by a recorded vote of 67 to 54, with 46 abstentions, on the recommendation of the Committee (A / 55/602 / Add.3, para. 49.)

Reference:
Aprobada en la 81a. sesión plenaria , celebrada el 4 de diciembre de 2000 , **por** **recomendación** **de** la Comisión (A / 55 / 602 / Add.3 , párr . 49) , en votación registrada de 67 votos contra 54 y 46 abstenciones.

Baseline:
el 4 de diciembre de 2000 , 81 veces a la reunión de comité de (A / 55 / 602 / Add.3 , 第49) , con un récord de votación , 67 , 54 votos contra 46 abstenciones.

Expanded system:
el 4 de diciembre de 2000 a la reunión del Comité de 81 veces (A / 55 / , / Add.3 第49 602) con un récord de votación , con 67 entradas, 54 votos contra , 46 abstenciones.

(2)

Source:
继续 审查
(continue reviewing)

Reference:
continúe examinando

Baseline:
revisión

Expanded system:
sigue la revisión de

Compared to the *omission* error, the problem of *untranslated items* and *wrong lexical choice*, in many cases, were improved to some extent after the introduction of the bilingual lexicon. As shown in the examples (3), 强烈谴责 was translated to a semantically unrelated phrase in target language by the baseline, while the expanded system delivered a reasonable translation. In the example (4) and (5), 多样性 and 支助 are OOV words for the baseline, but after inducing the new translation lexicon, the expanded system provided quite acceptable translations.

(3)

Source:
强烈谴责.
(Strongly condemned.)

Reference:
Condena enérgicamente.

Baseline:
recomiendo aprobar.

Expanded system:
una fuerte condena.

(4)

Source:
文化多样性
(cultural diversity)

Reference:
diversidad cultural

Baseline:
多样性 . a la cultura

Expanded system:
diversidad cultural

(5)

Source:
继续支助

(continue supporting)

Reference:
continúen apoyando

Baseline:
seguir 支助

Expanded system:
estado manteniendo

- **Grammatical level**

Vilar et al. (2006) and Costa et al. (2016) described grammar errors as deviations in the morphological and syntactical aspects of language. They are distinguished between *misselection* and *misordering*. *Misselection* problem occurs when the system is not able to produce the correct form of a word, although the translation of the base form was correct. *Misordering* problem refers to the error that the system fails to generate the correct order of output sentence, and this problem can be distinguished between local and long range reorderings. In this experiment, *misselection* errors were found to be frequent with both systems, especially when translating verbs. For instance, in our test set, the word 认为 was constantly translated as *creo que* (first-person singular) by the systems, while the correct translation according to the contexts should be *cree que* or *considera que* (third-person singular).

Regarding *misordering* problem, after the induction of our new unigram translation pairs, translation results were more literal and the system was found to prioritize local reordering than long range reordering. For instance, in the following example, *经济，社会和文化发展* was translated as *la economía de desarrollo cultural y social* by the expanded system, while the baseline delivered a correct translation.

(6)

Source:

有助于实现讨论***经济 、 社会和文化发展*** 各方面问题的联合国全球会议的目标。

(Contribute to the achievement of the objectives of the global conferences of United Nations on economic, social and cultural development in all aspects)

Reference:

contribuye al logro de los objetivos de las conferencias mundiales de las Naciones Unidas relativos a diversos aspectos del ***desarrollo económico , social y cultural***.

Baseline:

y de ***desarrollo económico, social y cultural*** de la ONU en cada aspecto de la reunión.

Expanded system:

ayudar a realizar la ***economía de desarrollo cultural y social***, los aspectos de las Naciones Unidas en la reunión del objetivo.

- **Semantic level**

Semantic errors address problems that regard the wrong selection of word meaning. Different from the *wrong lexical choice*, semantic errors focus on confusion of senses. As in many cases, a source can be translated into different target words. In the example (7), our expanded system performs better than the baseline, since ***表示*** can be translated as *expresar* and *significar* in Spanish, but according to this context, it should be translated as *expresar* (as delivered by the expanded system) instead of *significar* (translation by the baseline). Besides, the expanded system also provided acceptable translation *pobreza* and *analfabeto* for the baseline OOVs 赤贫 and 文盲, respectively.

(7)

Source:

表示关切 在《 世界人权宣言 》 通过五十多年后， ***赤贫*** 、饥饿 、疾病 、没有 适当住房 、 ***文盲*** 和无望等不可接受的境况仍是十几亿人的命运。

(Expressing concern that the unacceptable situation of extreme poverty, hunger, disease, lack of adequate shelter, illiteracy and despair is still the fate of more than a billion people after more than 50 years of the adoption of the Universal Declaration of Human Rights.)

Reference:

Observando ***con*** ***preocupación*** ***que***, más de cincuenta años después de la aprobación de la Declaración Universal de Derechos Humanos, una situación inaceptable de pobreza absoluta, hambre, enfermedades, falta de viviendas adecuadas, analfabetismo y desesperanza siga siendo la suerte de más de 1.000 millones de personas.

Baseline:

significa que su en el mundo de la Declaración &amp; 五十多 años después , ***赤贫*** , hambre , las enfermedades , - no viajara ni se relacionara ***文盲*** viviendas , y no de 十几亿 境况 es el destino .

Expanded system:

ha ***expresado*** ***su*** ***interés*** ***en*** el mundo a través de la Declaración de derechos humanos &amp; 五十多, años después, la ***pobreza***, el hambre y enfermedades, y que no, espera. prácticamente ***analfabeto***. no había esperanza de 境况 十几亿 es el destino.

4.4.3. Summary

This section described how we applied a machine learning supervised method to induce new translation lexicon from monolingual corpora for enriching the phrase table of our Chinese-Spanish SMT baseline. The experiment shows an improvement of +0.7 BLEU score was achieved even though an average of 800 translation pairs per source word were added to the existing parallel corpus. The high recall of our classifier ensures that more reliable translation candidates can be introduced to the SMT system and the language model component is able to handle the selection of the correct one, hence delivering a better translation output.

Chapter 5

5. CONCLUSIONS AND FUTURE WORKS

5.1. Conclusions

The main research directions of this dissertation were addressed on alleviating the OOV problem and inappropriate translation of lexical inflections to reduce the inflectional translation errors that frequently occur with low resource SMT. Along this line, we conducted a series of experiments to enhance a ZH-ES PBSMT system trained with the available corpora, which was of a reduced size compared with others. The phrase table expansion was based on two different methods: (1) deriving more morphological variant translations based on the knowledge of the baseline phrase table as described in Chapter 3, and (2) inducing new bilingual lexicon from extra monolingual corpora as described in Chapter 4.

The experimental results showed improvements of translation quality in different aspects. In the following we give a brief summary of our approaches, as well as the obtained conclusions.

- **Morphology expansion for SMT**

For our first objective which was to enrich a baseline system with inflected variants obtained in a lexical resource, we conducted an experiment to expand the translation model by adding inflected word forms based on the translation rules of a baseline phrase table as described in Chapter 3. To do so, a Spanish lexical resource was used to return all possible morphological variants for the Spanish target word of given translation rules from the baseline phrase table. Once obtained all the new inflected translation pairs, we directly appended our new translation rules, including noise, into the parallel training data. The expanded system obtained an improvement of 0.59 BLEU score compared to the baseline system, demonstrating that these new translation pairs can effectively enrich the phrase table since the system can produce more inflected translations after the phrase table expansion. However, after analyzing the translation outputs, we realized that the enriched system is still not sufficient to handle the translation problem of Spanish verb conjugation when a subject or a time marker is not explicitly expressed in Chinese source sentence.

- **Supervised bilingual lexicon induction for SMT**

To achieve our second objective that focuses on learning new translation lexicon from monolingual data for alleviating OOV problem of the baseline system, in Chapter 4, a supervision method which only needs monolingual features was used to induce new translation lexicon for enriching the SMT phrase table. The model was first trained with word embedding vectors of a small amount of translation equivalents as described in Section 4.1 and then was improved by adding other monolingual features such as word frequency information and Brown clustering as described in Section 4.2 and 4.3, respectively.

- **Word embedding-based classification experiment**

  To generate new translations, in the first experiment, a bilingual lexicon was induced with a SVM classifier trained using word embedding vectors of 1K translation equivalents and 5K randomly paired word pairs. We evaluated this method on two quite distinct language pairs Chinese-Spanish and English-Spanish in two different scenarios: *Proof-of-Concept evaluation* and *SMT simulation scenario*.

  a) In *Proof-of-Concept evaluation,* the classifiers achieved around 0.90 F1-score for the *right translation* class for both language pairs, and obtained an accuracy of up to 96%. According to our experimental results, a small seed lexicon of about

500 translation pairs has proved to be sufficient for our classifier to obtain a relatively stable performance (F1-score 0.80 +) on predicting whether a new word pair is in a translation relation or not. Besides, we also evaluated the model with different types of corpora (lemmatized corpora and corpora in word form) and different ratio of positive and negative examples. The empirical results showed that the model trained using corpora in word form with the imbalanced ratio of 5 negatives to each positive obtained the best performance.

b) In what we called the *SMT simulation scenario*, we applied the trained model to rank for every particular source test word the obtained right translations among all target vocabulary words using the obtained confidence score expecting to find the correct translation pair ranked in top positions. However, the results showed that many word pairs obtained same confidence score making it impossible to properly set up the ranking list.

The experimental results demonstrated that word embedding vectors obtained in the source and target monolingual corpora are sufficient to train the binary classifier for determining the translation relationship between words without depending on extra parallel or comparable data. However, the performance of the word embedding-based classifier suffered from the *hubness* problem which leads to a certain number of bad words that were repeatedly classified as the *right translation* for many different source words.

- **Improving word embedding-based classifier using word frequency information**

  To enhance the word embedding-based classifier, in Section 4.2, we added the information of word frequency distribution to word embedding vectors. The hypothesis behind this experiment was that the additional word frequency features can be useful for the classifier to learn the translation relationship between words, as well as to filter out those word pairs with large difference in their frequencies. Unlike most of other approaches, instead of using raw frequency or relative frequency, we computed the standard deviation of word frequencies of the monolingual corpus, and used it to divide the word frequency range into *n* different subranges. One-hot encoding was used to encode the corresponding subrange of each source word and target word. Our hypothesis was confirmed by the results shown in Section 4.2.3: after integrating word frequency feature, our classifier obtained a F1-score improved, from 0.91 to 0.94, for the *right translation* class, demonstrating that the encoded word frequency representations indeed provided useful information for improving the performance of the word embedding-based classifier.

- **Improving word embedding-based classifier using Brown clustering**

  Besides the word frequency feature, we also tried to combined word embedding vector with Brown clustering information to improve the classification performance. Our assumption was that adding Brown clustering feature to the word embedding vector can be useful for filtering out those candidates that are not semantically related to the corresponding source word. So in the experiment presented in Section 4.3, we trained the classifier using the joint representation of WE and BC of translation equivalents. The results show that, after adding BC, the classifiers achieved the F1-score of around 0.94 for the *right translation* class, demonstrating that although the word clustering information was learned from the source and target monolingual corpora, it had positive impact on the word embedding-based classifier. Besides, the *hubness* problem of WE classifier was alleviated by the additional BC feature to some extent.

- **SMT phrase table expansion with an induced bilingual lexicon**

  Since the model trained with WE and BC and the model trained with WE and WC obtained similar performances, in Section 4.4 we applied the model trained with WE and BC to generate new translations from monolingual corpora for phrase table expansion. The classifier was used to deliver, for each given source word, all the possible translation candidates as found in the combination with all the target words of the same PoS. As results, the WE-based classifier delivered a 7% word pairs classified as *right translation*, while the classifier trained using WE and BC delivered only a 2.7%, confirming again the assumption proposed in Section 4.3 that the monolingual word clustering information is useful for discarding the candidates that are not semantically related to the given source word.

  After the expansion with new translation lexicons, an improvement of up to 0.7 BLEU score was achieved even though an average of 800 translation pairs per source word were added. The high recall of our classifier ensures that more reliable translation candidates can be introduced to the SMT system and the language model component is able to handle the selection of the correct one, hence delivering a better translation output. At the end of the section, we also carried out a series of result analysis regarding the possible impacts of the new translation lexicon on the baseline system.

5.2. Contributions

The main contribution of this dissertation is that we have designed a number of approaches to enrich a low resource SMT phrase table by only exploiting monolingual resources. The following are some detailed achievements resulting from the work:

- A word embedding-based classifier to automatically induce bilingual lexicon from unrelated monolingual corpora. Instead of learning complicated transformation matrix to project source language word to target language space, Our method treats the translation relation as a simple binary classification problem. To train the classifier we only need a small dictionary of 500 translation pairs instead of parallel corpus which is not available for many language pairs. With only word embedding features, the classifier is able to achieve an accuracy of up to 96.8%.

- An enhanced classification approach in which the classifier was trained using the joint representation of word embedding and word frequency feature. After adding word frequency information, the accuracy of the classifier increased from 96.8% to 98.3% compared to word embedding-based classifier.

- An improved classification solution in which the classifier was trained on the joint representation of word embedding and Brown clusters of translation equivalents. With additional Brown clustering information, the obtained accuracy improved from 96.8% to 97.6%, resulting in a considerable reduction of word pairs classified as *right translation*.

- A systematic analysis regarding the translation results delivered by the expanded systems compared to the performance of the baseline. We classified the errors into lexical level, grammatical level and semantic level and explained the advantages and limitations of our methods.

The experiments presented in Chapter 4 have been accepted and published in the following conferences:

- Han, Jingyi; Bel, Núria (2017). "Enriching low resource SMT using induced bilingual lexicons". In Proceedings of the 33th Conference of Spanish Society for Natural Language Processing ‘17. Murcia: SEPLN (accepted).

- Han, Jingyi; Bel, Núria. (2016). "Towards producing bilingual lexica from monolingual corpora". In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).

- Han, Jingyi; Bel, Núria (2015). "Towards phrase table expansion using automatically induced bilingual lexica". In Proceedings of the International Workshop on Embeddings and Semantics SEPLN ‘15. Alacant: SEPLN.

5.3 Future works

In the light of the results and conclusions obtained in the previous experiments, some future research lines are identified:

- Although in Section 4.2 and Section 4.3, the incorporation of Brown Cluster and Word frequency information improved the performance of the word embedding-based classifier to some extents, there are still a certain amount of wrongly classified *right translations* which remain to be filtered out. To reduce those false positives delivered by the classifier, another avenue for future work is using ensemble learning to combine our model with models separately trained with other monolingual features. Set up ranking list with ensemble modeling may lead to a higher accuracy, since our classifiers has been capable of reducing more than 95% of no translation.

- It has been proved that the binary classifier trained using the joint representation of WE and additional monolingual features (BC and WC) can alleviate the *hubness* problem to some extent. It would be interesting to evaluate the impact of the Brown cluster and word frequency distribution feature on the linear mapping function between language vector spaces, since according to Dinu et al. (2015), the *hubness problem* is one of the main obstacles when one looks at the nearest neighbours of vectors that have been mapped across spaces with ridge.

- Neural machine translation has recently achieved impressive results while learning from raw, sentence-aligned parallel text. Sennrich and Haddow (2016) proposed to improve NMT by exploiting monolingual linguistic features such as subword tags, morphological information, POS tags and dependency labels. They obtained an improvement of up to

1.5 BLEU scores. In our experiments, it has been proved adding Brown cluster and word frequency feature can improve the performance of our binary classifier. It would be interesting to incorporate these features into neural machine translation and evaluate their impact on the translation performance.

REFERENCES

Ahmed, Amr, and Greg Hanneman. 2005. "Syntax-Based Statistical Machine Translation: A Review." *Computational Linguistics*.

Albat, Thomas Fritz. 2015. "Systems and Methods for Automatically Estimating a Translation Time." *US Patent 0185235*.

Almaghout, Hala, Jie Jiang, and Andy Way. 2010. "CCG Augmented Hierarchical Phrase Based Machine-Translation." In *The 7th International Workshop on Spoken Language Translation*. Paris, France.

Arcodia, Giorgio Francesco. 2007. "Chinese: A Language of Compound Words?" In *Selected Proceedings of the 5th Décembrettes: Morphology in Toulouse, Ed. Fabio Montermini, Gilles Boyé, and Nabil Hathout*, 79–90. MA: Cascadilla Proceedings Project.

Arnold, Doug. 2003. "Why Translation Is Difficult for Computers." In *Computers and Translation: A Translator's Guide*, 119–142. Amsterdam.

Avramidis, Eleftherios, and Philipp Koehn. 2008. "Enriching Morphologically Poor Languages for Statistical Machine Translation." In *Association for Computational Linguistics*, 763–770.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. “Neural Machine Translation by Jointly Learning to Align and Translate.” In *ICLR 2015*.

Banerjee, Satanjeev, and Alon Lavie. 2005. “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.” In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization*, 65–72. Kluwer Academic Publishers. doi:10.1007/s10590-009-9059-4.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. “Don’t Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors.” In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 238–247. Baltimore, Maryland, USA.

Bel, Núria. 2006. “Handling of Missing Values in Lexical Acquisition.” In *LREC*.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, Christian Jauvin, Jaz Kandola, Thomas Hofmann, Tomaso Poggio, and John Shawe-Taylor. 2003. “A Neural Probabilistic Language Model.” *Journal of Machine Learning Research* 3: 1137–1155.

Bentivogli Fbk, Luisa, Trento Italy, Arianna Bisazza, and Marcello Federico. 2016. “Neural versus Phrase-Based Machine Translation Quality: A Case Study.” In *Proceedings of the 2016 Conference on Empirical*

Methods in Natural Language Processing, 257–267. Texas.

Bertoldi, Nicola, and Marcello Federico. 2009. "Domain Adaptation for Statistical Machine Translation with Monolingual Resources." In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 182–189. Athens, Greece: Association for Computational Linguistics.

Birch, Alexandra, Nadir Durrani, and Philipp Koehn. 2013. "Edinburgh SLT and MT System Description for the IWSLT 2013 Evaluation." In *N Proceedings of the 10th International Workshop on Spoken Language Translation*, 40–48.

Birch, Alexandra, Miles Osborne, and Philipp Koehn. 2007. "CCG Supertags in Factored Statistical Machine Translation." In *Proceedings of the Second Workshop on Statistical Machine Translation*, 9–16. Association for Computational Linguistics.

Bland, J Martin, and Douglas G Altman. 1996. "Statistics Notes: Measurement Error." *BMJ* 313 (7059).

Bond, Francis, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. 2001. "Design and Construction of a Machine-Tractable Japanese-Malay Dictionary." In *In Proc. of MT Summit VIII*, 53–58.

Brown, Peter E, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. "The Mathematics of

Statistical Machine Translation: Parameter Estimation.” *Association for Computational Linguistic*.

Brown, Peter F., Peter V. DeSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. “Class-Based N-Gram Models of Natural Language.” *Computational Linguistics* 18: 467--479.

Bullinaria, John A, and Joseph P Levy. 2012. “Extracting Semantic Representations from Word Co-Occurrence Statistics: Stop-Lists, Stemming and SVD.” *Behavior Research Methods*.

Callison-Burch, Chris, Philipp Koehn, and Miles Osborne. 2006. *Improved Statistical Machine Translation Using Paraphrases*. *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. Vol. 1. ACL. doi:10.3115/1220835.1220838.

Calude, Andreea S, and Mark Pagel. 2011. “How Do We Use Language? Shared Patterns in the Frequency of Word Use across 17 World Languages.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 366 (1567). The Royal Society: 1101–7. doi:10.1098/rstb.2010.0315.

Casas-Tost, Helena, and Sara Rovira-Esteva. 2014. “New Models, Old Patterns? The Implementation of the Common European Framework of Reference for Languages for Chinese.” *Linguistics and Education* 27 (September): 30–38. doi:10.1016/j.linged.2014.07.001.

Chahuneau, Victor, Eva Schlinger, Noah A Smith, and Chris Dyer. 2013. “Translating into Morphologically Rich

Languages with Synthetic Phrases." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1677–1687. Seattle, Washington, USA: Association for Computational Linguistics.

Chandar, A P Sarath, Stanislas Lauly, Hugo Larochelle, Mitesh M Khapra, Balaraman Ravindran, Vikas Raykar, and Amrita Saha. 2014. "An Autoencoder Approach to Learning Bilingual Word Representations." In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 1853–1861. MIT Press.

Chang, and Pi-Chuan. 2009. "Improving Chinese-English Machine Translation through Better Source-Side Linguistic Processing." Stanford University.

Cho, Kyunghyun, Bart Van Merriënboer, and Dzmitry Bahdanau. 2014. "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches." In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 103–111. Doha.

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014a. "Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1724–1734. Doha.

Collobert, Ronan, and Jason Weston. 2008. "A Unified Architecture for Natural Language Processing." In

Proceedings of the 25th International Conference on Machine Learning - ICML '08, 160–167. New York, New York, USA: ACM Press. doi:10.1145/1390156.1390177.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20: 273–297.

Costa, Angela, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. "A Linguistically Motivated Taxonomy for Machine Translation Error Analysis." *Machine Translation* 29 (2). Springer Netherlands: 127–161. doi:10.1007/s10590-015-9169-0.

Costa-jussà, Marta R, Mireia Farrús, José B Mariño, and José A R Fonollosa. 2012b. "STUDY AND COMPARISON OF RULE-BASED AND STATISTICAL CATALAN-SPANISH MACHINE TRANSLATION SYSTEMS." *Computing and Informatics* 31: 245–270.

Costa-jussà, Marta R, Carlos A Henríquez, and Rafael E Banchs. 2012a. "Evaluating Indirect Strategies for Chinese–Spanish Statistical Machine Translation." *Journal of Artificial Intelligence Research* 45: 761–780.

Coulmance, Jocelyn, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. "Trans-Gram, Fast Cross-Lingual Word-Embeddings." In *In EMNLP*, 1109–1113.

DeFrancis, John. 1984. *The Chinese Language : Fact and Fantasy*. University of Hawaii Press.

Derczynski, Leon, Sean Chester, and Kenneth S Bøgh. 2015. “Tune Your Brown Clustering, Please.” In *Conference: Recent Advances in Natural Language Processing*.

Dinu, Georgiana, Angeliki Lazaridou, and Marco Baroni. 2014. “Improving Zero-Shot Learning by Mitigating the Hubness Problem.” *arXiv:1412.6568*, December.

Dong, Meiping, Yang Liu, Huanbo Luan, Maosong Sun, Tatsuya Izuha, and Dakun Zhang. 2015. “Iterative Learning of Parallel Lexicons and Phrases from Non-Parallel Corpora.” In *Proceedings of the 24th International Conference on Artificial Intelligence*, 1250–1256. The Association for the Advancement of Artificial Intelligence Press.

Duda, Richard O., Peter E. Hart, and David G. Stork. 2001. *Pattern Classification*. Wiley.

Durgar El-Kahlout, Ilknur, and Kemal Oflazer. 2006. “Initial Explorations in English to Turkish Statistical Machine Translation.” In *Proceedings of the Workshop on Statistical Machine Translation*, 7–14.

Durrani, Nadir, Alexander Fraser, and Helmut Schmid. 2013. “Model With Minimal Translation Units, But Decode With Phrases.” In *Proceedings of NAACL-HLT*, 1–11.

Faruqui, Manaal, and Chris Dyer. 2014. “Improving Vector Space Word Representations Using Multilingual Correlation.” In *Proceeding of EACL*.

Gale, William A. 1995. “Good-Turing Smoothing without Tears.” *Journal of Quantitative Linguistics* 2.

Galley, Michel, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. “Scalable Inference and Training of Context-Rich Syntactic Translation Models.” In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, 961–68. Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1220175.1220296.

Gao, Qin, and Stephan Vogel. 2008. “Parallel Implementations of Word Alignment Tool.” In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, 49–57.

Gouws, Stephan, Yoshua Bengio, and Greg Corrado. 2015. “BilBOWA: Fast Bilingual Distributed Representations without Word Alignments.” In *Proceedings of the 32nd International Conference on Machine Learning*, 37:748–756.

Habash, Nizar. 2007. “Arabic Morphological Representations for Machine Translation.” *Arabic Computational Morphology*, 263–285.

Habash, Nizar. 2008. “Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation.” In *Proceedings of ACL-08: HLT, Short Papers (Companion Volume)*, 57–60. Association for Computational Linguistics.

Haghighi, Aria, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. “Learning Bilingual Lexicons from

Monolingual Corpora." In *Proceedings of ACL-08: HLT*, 771–779. Columbus, Ohio, USA.

Harris, Zellig S. 1954. "Distributional Structure." *WORD* 10 (2–3). Routledge: 146–62. doi:10.1080/00437956.1954.11659520.

Hayashi, Katsuhiko, Hajime Tsukada, Katsuhito Sudoh, Kevin Duh, and Seiichi Yamamoto. 2010. "Hierarchical Phrase-Based Machine Translation with Word-Based Reordering Model." In *Proceedings of the 23rd International Conference on Computational Linguistics*, 439–446. Beijing.

Heafield, Kenneth. 2011. "KenLM: Faster and Smaller Language Model Queries." In *Proceedings of the 6th Workshop on Statistical Machine Translation*, 187–197.

Hermann, Karl Moritz, and Phil Blunsom. 2014a. "Multilingual Distributed Representations without Word Alignment." *arXiv:1312.6173*, December.

Hermann, Moritz Karl, and Phil Blunsom. 2014b. "Multilingual Models for Compositional Distributed Semantics." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 58–68.

Huang, Eric H., Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. "Improving Word Representations via Global Context and Multiple Word Prototypes." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. Association for Computational Linguistics.

Irvine, Ann, and Chris Callison-Burch. 2014. "Hallucinating Phrase Translations for Low Resource MT." In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

Irvine, Ann, and Chris Callison-Burch. 2013. "Supervised Bilingual Lexicon Induction with Multiple Monolingual Signals." In *Proceedings of HLT-NAACL*, 518–523.

Irvine, Ann, and Chris Callison-Burch. 2016. "End-to-End Statistical Machine Translation with Zero or Small Parallel Texts." *Natural Language Engineering* 22 (4): 517–548.

Jean, Sébastien, Kyunghyun Cho, and Roland Memisevic. 2015. "On Using Very Large Target Vocabulary for Neural Machine Translation." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1–10. Beijing.

Joachims, Thorsten. 1998. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." In *Proceedings of the European Conference on Machine Learning*, 137–142. Springer, Berlin, Heidelberg. doi:10.1007/BFb0026683.

Jurafsky, Daniel, and James H Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Kalchbrenner, Nal, and Phil Blunsom. 2013. "Recurrent Convolutional Neural Networks for Discourse

Compositionality." In *Proceedings of the Workshop on Continuous Vector Space Models and Their Compositionality*, 119–126.

Klementiev, Alexandre, Ivan Titov, and Binod Bhattarai. 2012. "Inducing Crosslingual Distributed Representations of Words." In *Proceedings of COLING*, 1459–1474.

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge University Press.

Koehn, Philipp, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. "Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation." In *In IWSLT*.

Koehn, Philipp, and Hieu Hoang. 2007a. "Factored Translation Models." In *EMNLP-CoNLL*, 868–876.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, et al. 2007b. "Moses: Open Source Toolkit for Statistical Machine Translation." In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 177–180. Association for Computational Linguistics.

Koehn, Philipp, and Kevin Knight. 2002. "Learning a Translation Lexicon from Monolingual Corpora." In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, 9–16.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. "Statistical Phrase-Based Translation." In *Proceedings of* *the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, 1:48–54. Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1073445.1073462.

Köppen, Mario. 2000. "The Curse of Dimensionality." In *Proceedings of the 5th Online World Conference on Soft Computing in Industrial Applications*, 4–8.

Lazaridou, Angeliki, Georgiana Dinu, and Marco Baroni. 2015. "Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 270–280.

Levy, Omer, and Yoav Goldberg. 2014b. "Linguistic Regularities in Sparse and Explicit Word Representations." In *Proceedings of the Eighteenth Conference on Computational Language Learning*, 171–80. Baltimore, Maryland USA: Association for Computational Linguistics.

Levy, Omer, and Yoav Goldberg. 2014a. "Dependency-Based Word Embeddings." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 302–8. Baltimore, Maryland, USA.

Li, Charles N., and Sandra A. Thompson. 1981. *Mandarin Chinese : A Functional Reference Grammar*. University of California Press.

Madnani, Nitin, and Bonnie J. Dorr. 2010. "Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods." *Computational Linguistics* 36 (3). MIT Press: 341–387. doi:10.1162/coli_a_00002.

Marcu, Daniel, and William Wong. 2002. "A Phrase-Based, Joint Probability Model for Statistical Machine Translation." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 133–139.

Marton, Yuval. 2010. "Improved Statistical Machine Translation with Hybrid Phrasal Paraphrases Derived from Monolingual Text and a Shallow Lexical Resource." *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Marton, Yuval, Chris Callison-Burch, and Philip Resnik. 2009. "Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 381–390. ACL.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. "Efficient Estimation of Word Representations in Vector Space." *arXiv:1301.3781*, January.

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. 2013b. "Exploiting Similarities among Languages for Machine Translation." *arXiv:1309.4168*, September.

Mikolov, Tomas, Wen-Tau Yih, and Geoffrey Zweig. 2013c. "Linguistic Regularities in Continuous Space Word

Representations.” In *Proceedings of NAACL-HLT*, 746–51. Association for Computational Linguistics.

Neculescu, Silvia, Nuria Bel, Sara Mendes, David Jurgens, and Roberto Navigli. 2015. “Reading Between the Lines: Overcoming Data Sparsity for Accurate Classification of Lexical Relationships.” In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, 182–192.

Nerima, Luka, and Eric Wehrli. 2008. “Generating Bilingual Dictionaries by Transitivity.” In *In Proceeding of LREC*.

Newman, C. M., Y. Rinott, and A. Tversky. 1983. “Nearest Neighbors and Voronoi Regions in Certain Point Processes.” *Advances in Applied Probability* 15 (4). Applied Probability Trust: 726–751. doi:10.2307/1427321.

Newman, Charles M., and Yosef Rinott. 1985. “Nearest Neighbors and Voronoi Volumes in High-Dimensional Point Processes with Various Distance Functions.” *Advances in Applied Probability* 17 (4). Applied Probability Trust: 794. doi:10.2307/1427088.

Norman, Jerry. 1988. *Chinese*. Cambridge: Cambridge University Press.

Och, Franz Josef. 1999. “An Efficient Method for Determining Bilingual Word Classes.” In *Proceeding of EACL*.

Och, Franz Josef, and Hermann Ney. 2002. “Discriminative Training and Maximum Entropy Models for Statistical Machine Translation.” In *Proceedings of the 40th Annual*

Meeting of the Association for Computational Linguistics, 295–302. Philadelphia.

Och, Franz Josef, and Hermann Ney. 2003. "A Systematic Comparison of Various Statistical Alignment Models." *Computational Linguistics* 29 (1). MIT Press: 19–51. doi:10.1162/089120103321337421.

Okpor, M D. 2014. "Machine Translation Approaches: Issues and Challenges." *International Journal of Computer Science Issues* 11 (5).

Packard, Jerome L. 2000. *The Morphology of Chinese: A Linguistic and Cognitive Approach*. Cambridge University Press.

Pal, M., and P. M. Mather. 2005. "Support Vector Machines for Classification in Remote Sensing." *International Journal of Remote Sensing* 26 (5). Taylor & Francis Group: 1007–11. doi:10.1080/01431160512331314083.

Papineni, K.A., S. Roukos, and R.T. Ward. 1998. "Maximum Likelihood and Discriminative Training of Direct Translation Models." In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, 1:189–92. Seattle: IEEE. doi:10.1109/ICASSP.1998.674399.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "BLEU: A Method for Automatic Evaluation of Machine Translation." In *Proceedings of the 40th*

Annual Meeting of the Association for Computational Linguistics, 311–318.

Pappu, Vijay, and Panos M. Pardalos. 2014. "High-Dimensional Data Classification." *Clusters, Orders and Trees: Methods and Applications*. doi:10.1007/978-1-4939-0742-7_8.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Piantadosi, Steven T. 2014. "Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions." *Psychonomic Bulletin & Review*, 1112–1130. doi:10.3758/s13423-014-0585-6.

Platt, John. 1998. "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines."

Radovanovi, Miloš, Alexandros Nanopoulos, and Mirjana Ivanovi. 2010b. "On the Existence of Obstinate Results in Vector Space Models." In *33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 186–193.

Radovanovi, Miloš, Alexandros Nanopoulos, and Mirjana Ivanovic. 2010a. "Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data." *Journal of Machine Learning Research* 11: 2487–2531.

Rapp, Reinhard. 1995. “Identifying Word Translations in Non-Parallel Texts.” In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, 320–322. Association for Computational Linguistics.

Saluja, Avneesh, Hany Hassan, Kristina Toutanova, and Chris Quirk. 2014. “Graph-Based Semi-Supervised Learning of Translation Models from Monolingual Data.” *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 676–686.

Sánchez Cartagena, Víctor Manuel. 2015. “Building Machine Translation Systems for Language Pairs with Scarce Resources.” Universidad de Alicante.

Sánchez-Cartagena, Víctor M, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2011. “Integrating Shallow-Transfer Rules into Phrase-Based Statistical Machine Translation.” In *Machine Translation Summit*.

Schafer, Charles, and David Yarowsky. 2002. “Inducing Translation Lexicons via Diverse Similarity Measures and Bridge Languages.” In *Proceeding of the 6th Conference on Natural Language Learning*, 20:1–7. Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1118853.1118879.

Schepens, Job, Ton Dijkstra, Franc Grootjen, and Walter J B van Heuven. 2013. “Cross-Language Distributions of High Frequency and Phonetically Similar Cognates.” *PloS One* 8 (5). Public Library of Science: e63006. doi:10.1371/journal.pone.0063006.

Schnabel, Tobias, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. "Evaluation Methods for Unsupervised Word Embeddings." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 298–307. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D15-1036.

Schwenk, Holger, Sadaf Abdul-Rauf, Loc Barrault, and Jean Senellart. 2009. "SMT and SPE Machine Translation Systems for WMT'09." In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 130–134. Athens, Greece.

See, Abigail, Minh-Thang Luong, and Christopher D. Manning. 2016. "Compression of Neural Machine Translation Models via Pruning." In *CoNLL 2016*.

Sennrich, Rico, and Barry Haddow. 2016. "Linguistic Input Features Improve Neural Machine Translation." *arXiv:1606.02892*.

Shi, Tianze, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. "Learning Cross-Lingual Word Embeddings via Matrix Co-Factorization." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 567–572.

Smith, Reginald, Bouchet-Franklin Institute, and Ga Decatur. 2007. "Investigation of the Zipf-Plot of the Extinct Meroitic Language." *Glottometrics* 15: 53–61.

Somers, Harold. 2003. *Computers and Translation : A Translator's Guide*. John Benjamins Publishing.

Soyer, Hubert, Pontus Stenetorp, and Akiko Aizawa. 2015. "Leveraging Monolingual Data for Crosslingual Compositional Word Representations." In *In ICLR*.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. "Sequence to Sequence Learning with Neural Networks." In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 3104–3112. MIT Press.

Täckström, Oscar, Ryan McDonald, and Jakob Uszkoreit. 2012. "Proceedings of the 2012 Conference of the North American Chapter Of." In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 477–487. Association for Computational Linguistics.

Tanaka, Kumiko, and Kyoji Umemura. 1994. "Construction of a Bilingual Dictionary Intermediated by a Third Language." *arXiv:cmp-lg/9410020*.

Tiedemann, Jörg. 2012. "Parallel Data, Tools and Interfaces in OPUS." In *Workshop Abstracts : Eighth International Conference on Language Resources and Evaluation*, s. 2214-2218. ELRA.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." In *Proceedings of the Conference of the North American*

Chapter of the Association for Computational Linguistics on Human Language Technology, 173–180. Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1073445.1073478.

Toutanova, Kristina, Hisami Suzuki, and Achim Ruopp. 2008. “Applying Morphology Generation Models to Machine Translation.” In *Proceedings of ACL-08: HLT*, 514–522. Columbus, Ohio, USA: Association for Computational Linguistics.

Turchi, Marco, and Maud Ehrmann. 2011. “Knowledge Expansion of a Statistical Machine Translation System Using Morphological Resources.” *Polibits*, no. 43: 37–43.

Turian, Joseph, Lev Ratinov, and Yoshua Bengio. 2010. “Word Representations: A Simple and General Method for Semi-Supervised Learning.” In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384–394.

Turney, Peter D., and Patrick Pantel. 2010. “From Frequency to Meaning: Vector Space Models of Semantics.” *Journal of Artificial Intelligence Research*, March, 141–188. doi:10.1613/jair.2934.

Tversky, Amos, Yosef Rinott, and Charles M Newman. 1983. “Nearest Neighbor Analysis of Point Processes: Applications to Multidimensional Scaling.” *Journal of Mathematical Psychology* 27 (3): 235–250. doi:10.1016/0022-2496(83)90008-1.

Upadhyay, Shyam, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. “Cross-Lingual Models of Word

Embeddings: An Empirical Comparison." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1661–1670. Berlin, Germany: Association for Computational Linguistics.

Vapnik, Vladimir, and Olivier Chapellé. 2000. "Bounds on Error Expectation for Support Vector Machines." *Neural Computation*, 2013–2036.

Vauquois, B. 1976. "Automatic Translation — a Survey of Different Approaches." In *Proceedings of the 6th International Conference on Computational Linguistics*, 127–135. Ottawa.

Vilar, D., J. Xu, L. D'haro, and H. Ney. 2006. "Error Analysis of Statistical Machine Translation Output." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Vuli, Ivan, Wim De Smet, and Marie-Francine Moens. 2011. "Identifying Word Translations from Comparable Corpora Using Latent Topic Models." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpaper*, 479–484.

Vuli, Ivan, and Anna Korhonen. 2016. "On the Role of Seed Lexicons in Learning Bilingual Word Embeddings." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 247–257. Berlin, Germany: Association for Computational Linguistics.

Vulic, Ivan, and Marie-Francine Moens. 2015. “Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction.” In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, 719–725. Beijing, China,.

Vulic, Ivan, and Marie-Francine Moens. 2013a. “A Study on Bootstrapping Bilingual Vector Spaces from Non-Parallel Data (and Nothing Else).” In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1613–1624.

Vulic, Ivan, and Mar1ie-Francine Moens. 2013. “Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses.” In *Proceedings of NAACL-HLT*, 106–116.

Vulic, Ivan, and Marie-Francine Moens. 2016. “Bilingual Distributed Word Representations from Document-Aligned Comparable Data.” *Journal of Artificial Intelligence Research* 55 (1). AI Access Foundation: 953–994.

Wong, Kam-Fai, Wenji Li, Ruifeng Xu, and Zheng-sheng Zhang. 2009. *Introduction to Chinese Natural Language Processing*. Morgan & Claypool Publishers.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, et al. 2016. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." *arXiv Preprint arXiv:1609.08144*, September.

Yang, Qiang., Michael J. Wooldridge, Huanbo Luan, Maosong Sun, Tatsuya Izuha, and Dakun Zhang. 2015. "Iterative Learning of Parallel Lexicons and Phrases from Non-Parallel Corpora." In *Proceedings of the 24th International Conference on Artificial Intelligence*, 1250–56. Buenos Aires, Argentina: The Association for the Advancement of Artificial Intelligence Press.

Zhang, Meng, Haoruo Peng, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. "Bilingual Lexicon Induction From Non-Parallel Data With Minimal Supervision." *Thirty-First AAAI Conference on Artificial Intelligence*.

Zhao, Bing, Eric P Xing, and Alex Waibel. 2005. "Bilingual Word Spectral Clustering for Statistical Machine Translation." In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, 25–32.

Zhao, Kai, Hany Hassan, and Michael Auli. 2015. "Learning Translation Models from Monolingual Continuous Representations." In *Proceedings of HLT-NAACL*.

Zou, Will Y, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. "Bilingual Word Embeddings for Phrase-Based Machine Translation." In *Proceedings of EMNLP*.