Bilingual lexicon extraction for a distant language pair using a small parallel corpus

Ximena Gutierrez-Vasques
GIL IINGEN
UNAM
Mexico City, Mexico
xim@unam.mx

Abstract

The aim of this thesis proposal is to perform bilingual lexicon extraction for cases in which small parallel corpora are available and it is not easy to obtain monolingual corpus for at least one of the languages. Moreover, the languages are typologically distant and there is no bilingual seed lexicon available. We focus on the language pair Spanish-Nahuatl, we propose to work with morpheme based representations in order to reduce the sparseness and to facilitate the task of finding lexical correspondences between a highly agglutinative language and a fusional one. We take into account contextual information but instead of using a precompiled seed dictionary, we use the distribution and dispersion of the positions of the morphological units as cues to compare the contextual vectors and obtaining the translation candidates.

1 Introduction

Parallel corpora are a rich source of bilingual lexical information, they are a valuable resource that allows the development of several language technologies such as automatic construction of bilingual lexicons and statistical machine translation systems (SMT). Automatic construction of bilingual lexicons is useful since bilingual dictionaries are expensive resources and not many are available when one of the languages is resource-poor.

One way to perform bilingual lexical extraction from a parallel corpus is through word alignment. However, most of the methods to perform word-alignment, and in general the approaches to SMT, require huge amounts of parallel data. The task of extracting bilingual lexicon becomes even harder when we are dealing with very different languages, i.e., languages from different linguistic families that do not share orthographic, morphological or syntactic similarity.

The goal of this thesis is to propose a method for bilingual lexicon extraction that could be suitable for low-resource settings like the mentioned above. We work with the language pair Spanish-Nahuatl which are languages distant from each other (Indo-European and Uto-Aztecan language families) with different morphological phenomena. Nahuatl is an agglutinative language with polysynthetic tendency, this means that it can agglutinate many different morphemes to build highly complex words. On the other hand, Spanish can be classified as a fusional language where the words don't contain many different morphemes since several morphemes can be fused or overlaid into one encoding several meanings.

Although both languages are spoken in the same country, there is scarcity of parallel and monolingual corpora for Nahuatl. It is not easy to find general standard dictionaries due to the big dialectal variation and the lack of orthographical normalization of Nahuatl. Automatic extraction of a bilingual lexicon could be useful for contributing with machine-readable resources for the language pair that we are studying. Spanish is one of the most widely spoken languages in the world but, in the case of Nahuatl, few digital resources are available even though there exist around two million speakers of this language.

Our proposal aims to explore which information

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 154–160,
Denver, Colorado, June 1, 2015. ©2015 Association for Computational Linguistics

can be combined in order to estimate the bilingual correspondences and therefore building a bilingual lexicon. We plan to take into account correlation measures, positional cues and contextual information. Many of the methods that exploit contextual information require a precompiled digital seed dictionary or lexicon. We would like to propose a way to leave aside this language dependent requirement since many language pairs can face the same situation in which it is not easy to obtain a precompiled digital dictionary.

Unlike other approaches, we plan to take into account morphological information for building the word representations. The motivation behind is that morpheme-based representations can be useful to overcome the sparseness problem when building semantic vectors for morphologically rich languages with small corpus available.

The structure of the paper is as follows: Section 2 contains a general overview of the existing methods that tackle the bilingual extraction task and a description of our particular problem. In section 3, we describe the dataset and our proposal to address the bilingual lexical extraction for our low-resource setting. Finally, section 4 contains the conclusions.

2 Research Problem

2.1 Bilingual Lexicon Extraction

Bilingual lexicon extraction is the task of obtaining a list of word pairs deemed to be word-level translations (Haghighi et al., 2008). This has been an active area of research for several years, especially with the availability of big amounts of parallel corpuora that allow to model the relations between lexical units of the translated texts. One direct way to perform bilingual lexicon extraction is through word alignment from a parallel corpus. Word alignment is a fundamental part of SMT systems which build probabilistic translation models, based on several millions of parallel sentences, in order to estimate word and phrase level alignments (Brown et al., 1993).

However, the quality of word alignment methods used in SMT are heavily dependant on the amount of data and they require even more parallel data if we are dealing with very different languages. Since most of the language pairs do not have large amounts of clean parallel corpora readily available, there are alternative approaches for extracting multilingual information. Some methods rely on association and similarity measures to estimate the lexical correspondences, e.g., log-likelihood measures (Tufiş and Barbu, 2002), t-scores (Ahrenberg et al., 1998), positional difference between two successive occurrences of a word (Fung, 2000), just to mention some.

2.2 The low-resource setting

If there is not enough parallel corpora for a language pair, another alternative is to assume that there is enough comparable corpora or monolingual corpora for each of the languages. In these approaches bilingual lexicons are induced by taking into account several features, e.g, orthographic similarity, temporal similarity (Schafer and Yarowsky, 2002), association measures, topical information (Mimno et al., 2009) and contextual features. There are many works focused on the latter, they are based on the distributional notion (Harris, 1954) that a word that occurs in a given context in a language should have a translation that occurs in a similar context in the other language.

The general approach for using contextual information includes: 1. building a context vector for each lexical unit in both languages 2. Translating or projecting these context vectors to a common space using a seed dictionary or lexicon 3. Computing the similarity between the source and target words to find the translation candidates. There are several works that use contextual information, they vary in the way they represent the contexts and how they measure the similarity of the contextual vectors to extract translation candidates. (Rapp, 1995; Fung and Yee, 1998; Rapp, 1999; Diab and Finch, 2000; Déjean et al., 2002; Gaussier et al., 2004; Haghighi et al., 2008; Shezaf and Rappoport, 2010; Laroche and Langlais, 2010)

Another alternative is to use pivot languages as an intermediary language to extract bilingual lexicon (Tanaka and Umemura, 1994; Wu and Wang, 2007; Tsunakawa et al., 2008; Seo and Kim, 2013).

Lately there has been interest in multilingual distributed representation learning (Klementiev et al., 2012; Zou et al., 2013). These approaches are related with the ones that transfer information between languages using distributed representations and deep learning techniques (Lauly et al., 2014; Hermann and Blunsom, 2014). These approaches have the potential of semantic transfer into low-resource languages.

2.3 Our case of study

We focus on the language pair Spanish-Nahuatl, this represents a setting in which there is a small parallel corpus available, the two languages are very distant from each other and it is not easy to obtain comparable corpora or monolingual corpora for one of the languages.

These two languages are spoken in same country but Nahuatl does not have a web presence or text production comparable to Spanish. Most of the documents that can be easily found in Nahuatl are translations, that is why it is easier to obtain parallel corpora than monolingual. Although there are existing dictionaries for this language pair, not all of them are machine readable, the most extensive ones were made several centuries ago causing that some Spanish entries do not correspond anymore to the language spoken nowadays. Moreover, there is a big dialectal variation that complicates having one standard dictionary.

Under these conditions traditional statistical methods for word alignment are not the most suitable, in fact, to our knowledge it does not exist a SMT system yet for this language pair. We cannot rely either on orthographic similarity and there is no a pivot language that could be useful. On the other hand, practically all the methods based on contextual information require at some point a seed bilingual dictionary. This represents a chicken-egg problem (Koehn and Knight, 2002): If we have a bilingual lexicon we can translate the context vectors but we can only generate a bilingual lexicon with these methods if we are able to translate the context vectors.

The transfer based approaches have the potential of transferring semantic knowledge to low resource languages, e.g., alignment between sentences or phrases. However, they need to be trained with resource fortunate languages, usually requiring some supervised signal like word alignments to learn the bilingual embeddings.

We aim to address our low resource setting by combining several sources of information, mainly contextual features and association measures. In order to counteract the sparseness derived from working with a small parallel corpus of morphologically rich languages, we aim to use to morpheme representations instead of words. For the contextual approach, we prefer not to use the available noisy dictionaries as seed lexicon. Instead, we would like to explore features like the distribution and the dispersion of the positions of a morpheme in a text in order to be able to compare two contextual vectors representing lexical units in different languages.

Our conjecture is that the combination of several features, some of them usually applied for extracting lexicon from comparable corpora, could be suitable for a small, noisy parallel corpus of a distant language pair. Unlike other methods, our proposal aims to prescind from prior knowledge, e.g., a precompiled seed lexicon.

3 Methodology

3.1 The parallel corpus

To our knowledge, it did not exist a digital Spanish-Nahuatl parallel corpus publicly available. We had to build one, most of the sources were non digital books. As we have mentioned before, for some languages is not easy to extract parallel content from the typical web sources. Working with a low resource language sometimes implies difficulties that are not common when working with other languages, e.g., we had to perform a manual correction of the texts after being digitized since the OCR software confused several character patterns (it was probably assuming that it was processing a different language).

The documents of the parallel corpus are not quite homogeneous in the sense that there is dialectal, diachronic and orthographical variation. This variation can represent noise for many of the statistical methods, as an attempt to reduce it we performed an orthographic normalization. It does not exist a general agreement regarding to the appropriate way to write nahuatl language. We chose a set of normalization rules (around 270) proposed by linguists to normalize classical nahuatl (Thouvenot and Maynez, 2008) in order to obtain a more systematic writing. We implement them in FOMA (Hulden, 2009) a finite state toolkit used mainly for computational morphology. The set of rules that we used reduces the variation of many of the texts but unfortunately not from all them.

The total size of the corpus is around 1 million tokens (included both languages) which is still very small for the SMT approaches. To this scarcity, we have to add the fact that we will only work with a subset of documents, those that do not have a big dialectal or orthographical variation

3.2 Morphology

In order to perform the bilingual lexicon extraction, we would like to take into account the morphology of the language pair since the alignment complexity between typologically different languages is far away from the alignment complexity between similar languages (Cakmak et al., 2012).

Nahuatl is a polysynthetic language that allows compact nominal and verbal constructions where the complements, adjectives and adverbs can agglutinate with the verbal or nominal roots. This language also has incorporation and some other morphological phenomena. In contrast, Spanish is a fusional language in which a single morpheme can simultaneously encode several meanings. Regarding to the word order, Nahuatl and Spanish are relativetely flexible, specially Nahuatl.

Dealing with the morphology could be important to reduce the negative impact of sparseness and therefore having better representations of the lexical units. Specially in cases like ours where the corpus is small and the languages are morphologically rich, this may cause many different word types but few repetitions of them in the documents. If we have few contexts characterizing a word, then the contextuals vectors will not have a good quality, affecting the performance of the methods that exploit contextual features. Building morpheme based representations could be also useful for pairing the bilingual lexical units, since in agglutinative languages a single word can correspond to many in another language. The next example shows a morphologically segmented word in nahuatl and its correspondence to Spanish:

ti- nech - maca- z - nequi
2SG.S-1S.O-'give'-FUT-'want'
"Tu me quieres dar" (Spanish)
"You want to give me"

Recent approaches take into account morphology and investigate how compositional morphological distributed representations can improve word representations (Lazaridou et al., 2013) and language models (Botha and Blunsom, 2014; El-Desoky Mousa et al., 2013; Luong et al., 2013).

We aim to use, already implemented, unsupervised methods to perform morphological segmentation. Software like Morfessor (Creutz and Lagus, 2005) that seems to work well for agglutinative languages could be useful as well for languages like Nahuatl. Additionally, there is a morphological analysis tool based on rules for classical nahuatl (Thouvenot, 2011) that could be used to improve the unsupervised morphological segmentation. As for the Spanish case, there are unsupervised approaches that have proven to be successful in discovering Spanish affixes (Urrea, 2000; Medina-Urrea, 2008).

Once we have the segmented morphemes, we can build morpheme-based representations to extract the bilingual correspondences. Initially we plan to focus in extracting bilingual lexicon only for words with lexical meaning and not the grammatical ones.

At this moment, we have not still decided if we will work only with vector representations of each morpheme or with a composed representation of the words based on the morphemes.

3.3 Bilingual lexicon extraction without using a seed dictionary

For the bilingual lexical extraction we aim to combine several cues including correlation measures and contextual information. As we have mentioned before, most of the contextual methods have in common a need for a seed lexicon of translations to efficiently bridge the gap between languages. We would like to prescind from this requirement.

Seed lexicons are necessary to compare the contexts between the word representations in different languages. Few works have tried to circumvent this requirement, e.g., building a seed lexicon based on spelling and cognate cues (Koehn and Knight, 2002), using punctuation marks as a small seed lexicon and find alignments by measuring intralingual association between words (Diab and Finch, 2000). Lately some works have explored training a cross-language topic model on comparable corpora in order to obtain a seed lexicon without prior knowledge (Vulić and Moens, 2012).

We would like to explore the positions in which a word occurs in a text and the dispersion of these positions as cues for finding similar words in both languages and being able to compare the context vectors that characterize the words in both languages. The hypothesis is that words that are translations of each other tend to occur in similar positions of a parallel text and the distributions have similar dispersions. It is noteworthy that in our case we attempt to work at the morpheme level instead of the word level.

For each type in the text, in our case morphemes, we can store a vector of offsets, i.e. the positions in which the type occurs relative to the size of corpus. After recollecting all the positions for a lexical unit we can also measure the dispersion by calculating the variance or the standard deviation.

We conjecture that those lexical units between languages that obtain high similarity in their position distributions and their dispersion, are useful to compare the context vectors. They can be seen as a sort of initial seed lexicon constructed in a language independent way. The similarity can be calculated in terms of measurements like cosine similarity or measurements that take into account correlations or divergence between distributions.

Regarding to the construction of vectors encoding contextual information of the lexical units, we plan to try different experimental setups, examining different representations of word contexts, i.e., different association measures and weighting schemes for building the semantic vectors, different sizes of context windows and other important parameters that must be taken into account when working with distributional semantic representations.

Once we have the contextual vectors that represent the lexical units (in our case representations based on morphology) translation candidates can be obtained. Based on the contexts that are similar between the two vectors we can compare a source and a target contextual vector using different techniques or projecting them into a joint space and calculate the distance between them.

Taking into account the contexts and positions of the words in the whole document could be useful for noisy parallel corpora where there is not always a one to one correspondence between sentences. This is the case of some of the texts of our parallel corpus.

3.4 Combination of features and evaluation

It is very common for bilingual extraction methods to use a diverse set of cues and then combine them in order to obtain better translation candidates (Koehn and Knight, 2002; Tiedemann, 2003; Irvine, 2013). We will not use some of the typical cues like orthographic similarity or temporal, but we would like to combine the contextual information explained in the above section with some association measures between words or morphemes. Our intention is to propose a weighting scheme that allows to combine the several criteria and to obtain a rank of the translation candidates.

Once the translation candidates are extracted, we can establish a baseline by using some of the methods suitable for parallel corpora, e.g., the typical word alignment methods used in SMT. Additionally, it would be interesting to try different language pairs with more resources, in order to evaluate if our method can be competitive to more downstream approaches that rely on more data. For instance, we can evaluate in resource fortunate distant pairs like Spanish-German, since German is also morphologically rich with extensive use of compounds.

4 Conclusions

In this work we have presented a thesis proposal where the goal is to extract a bilingual lexicon under a particular low-resource setting in which is difficult to obtain big amounts of parallel or monolingual corpora and also is not easy to have an extensive standard electronical dictionary. The particularities of the methods are not completely defined since the work is in progress, we propose to combine morpheme based representations with contextual and association features in order to obtain translation candidates for the lexical units.

In our proposal we try to circumvent the need of a bilingual electronic dictionary which can be hard to obtain when working with low-resource languages. Although we focus in a particular language pair, the proposed methods are language independent and they could be used for languages with similar settings or even for comparable corpora.

Some of the aspects that are missing to tackle are the problems that may arise when dealing with synonyms and polysemic words.

5 Acknowledgements

This work is supported by the Mexican Council of Science and Technology (CONACYT), funds 370713 and CB-2012/178248. Thanks to the reviewers for their valuable comments.

References

Lars Ahrenberg, Mikael Andersson, and Magnus Merkel. 1998. A simple hybrid aligner for generating lexical correspondences in parallel texts. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 29–35. Association for Computational Linguistics.

Jan A Botha and Phil Blunsom. 2014. Compositional morphology for word representations and language modelling. *arXiv preprint arXiv:1405.4273*.

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

Mehmet Talha Cakmak, Süleyman Acar, and Gülsen Eryigit. 2012. Word alignment for english-turkish language pair. In *LREC*, pages 2177–2180.

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

Hervé Déjean, Éric Gaussier, and Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Mona Diab and Steve Finch. 2000. A statistical word-level translation model for comparable corpora. Technical report, DTIC Document.

A El-Desoky Mousa, H-KJ Kuo, Lidia Mangu, and Hagen Soltau. 2013. Morpheme-based feature-rich language models using deep neural networks for lvcsr of egyptian arabic. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8435–8439. IEEE.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 414–420. Association for Computational Linguistics.

Pascale Fung. 2000. A statistical view on bilingual lexicon extraction. In *Parallel Text Processing*, pages 219–236. Springer.

Eric Gaussier, J-M Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 526. Association for Computational Linguistics.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*, volume 2008, pages 771–779.

Zellig S Harris. 1954. Distributional structure. *Word*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, pages 29–32. Association for Computational Linguistics.

Ann Irvine. 2013. Statistical machine translation in low resource settings. In *HLT-NAACL*, pages 54–61.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*, pages 9–16. Association for Computational Linguistics.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd international conference on computational linguistics*, pages 617–625. Association for Computational Linguistics.

Stanislas Lauly, Alex Boulanger, and Hugo Larochelle. 2014. Learning multilingual word representations using a bag-of-words autoencoder. *arXiv preprint arXiv:1401.1803*.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *ACL (1)*, pages 1517–1526. Citeseer.

Minh-Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. *CoNLL-2013*, 104.

Alfonso Medina-Urrea. 2008. Affix discovery based on entropy and economy measurements. *Computational Linguistics for Less-Studied Languages*, 10:99–112.

David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 880–889. Association for Computational Linguistics.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322. Association for Computational Linguistics.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics.

Hong-Seok Kwon Hyeong-Won Seo and Jae-Hoon Kim. 2013. Bilingual lexicon extraction via pivot language and word alignment tool. *ACL 2013*, page 11.

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

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 297–303. Association for Computational Linguistics.

Marc Thouvenot and Romero-Galvan Ruben Maynez, Pilar. 2008. La normalizacion grafica del codice florentino. In *El universo de Sahagun pasado y presente*. UNAM.

Marc Thouvenot. 2011. Chachalaca en cen, juntamente. In *Compendio Enciclopedico del Nahuatl, DVD*. INAH.

Jörg Tiedemann. 2003. Combining clues for word alignment. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 339–346. Association for Computational Linguistics.

Takashi Tsunakawa, Naoaki Okazaki, and Jun'ichi Tsujii. 2008. Building a bilingual lexicon using phrase-based statistical machine translation via a pivot language. In *COLING (Posters)*, pages 127–130.

Dan Tufiş and Ana-Maria Barbu. 2002. Lexical token alignment: Experiments, results and applications. In *Proceedings from The Third International Conference on Language Resources anrd Evaluation (LREC-2002), Las Palmas, Spain*, pages 458–465.

Alfonso Medina Urrea. 2000. Automatic discovery of affixes by means of a corpus: A catalog of spanish affixes. *Journal of quantitative linguistics*, 7(2):97–114.

Ivan Vulić and Marie-Francine Moens. 2012. Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 449–459. Association for Computational Linguistics.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pages 1393–1398.