Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 41-48 recibido 31-03-2019 revisado 07-05-2019 aceptado 15-05-2019

Morphological segmentation for extracting Spanish-Nahuatl bilingual lexicon

Segmentación morfológica para extraer un lexicón bilingüe español-náhuatl

Ximena Gutierrez-Vasques[1], **Alfonso Medina-Urrea**[2], **Gerardo Sierra**[3]
[1]Universidad Nacional Autónoma de México
xim@unam.mx
[2] El Colegio de México
amedinau@colmex.mx
[3]Universidad Nacional Autónoma de México
gsierram@iingen.unam.mx

Abstract: The aim of this work is to extract word translation pairs from a small parallel corpus and to measure the impact of dealing with morphology for improving this task. We focus on the language pair Spanish-Nahuatl, both languages are morphologically rich and distant from each other. We generate semi-supervised morphological segementation models and we compare two approaches (estimation, association) for extracting bilingual correspondences. We show that taking into account typological properties of the languages, such as the morphology, helps to counteract the negative effect of working with a low-resource language.
Keywords: Morphology, bilingual, translation, Nahuatl, Spanish

Resumen: El objetivo de este trabajo es extraer pares de traducción a partir de un corpus paralelo pequeño, así como medir el impacto de lidiar con la morfología para mejorar esta tarea. Nos enfocamos en el par de lenguas español-náhuatl, las dos lenguas son morfológicamente ricas y tipológicamente distantes. Generamos modelos semisupervisados de segmentación morfológica y comparamos dos enfoques (estimativo, asociativo) para extraer pares bilingües de palabras. Mostramos que tomar en cuenta las propiedades tipológicas de la lengua, como la morfología, ayuda a contrarrestar el efecto negativo de trabajar con una lengua de bajos recursos.
Palabras clave: Morfología, bilingüe, traducción, nahuatl, español

1 Introduction

In natural language processing (NLP), bilingual lexicon extraction is the task of obtaining a list of word pairs deemed to be word-level translations (Haghighi et al., 2008). This is an important task, since bilingual dictionaries are expensive resources that are not always available for all language pairs, specially for low-resource languages. Moreover, extracting lexical translations is an important step for building statistical machine translation (SMT) models. There is a wide variety of approaches to perform this task. However, most of them assume that there are large amounts of clean parallel corpora readily available.

The performance of downstream bilingual lexicon extraction methods tends to drop when they face small amount of data or a distant language pair. Working with low-resource languages in NLP implies several challenges due to the lack of language technologies and digitally available corpora. Traditional methods need to be adapted in order to deal with the scarcity.

Our case of study is the language pair Spanish-Nahuatl, which is spoken in the same country. These languages are typologically distant (Indo-European and Uto-Aztecan language families). In order to perform bilingual lexicon extraction for this language pair, there are several issues to address, i.e., it is difficult to obtain corpora, these languages have different morphological and syntactical phenomena, and it is not easy to find standardized annotated resources for Nahuatl.

Due to this, we explore the performance

ISSN 1135-5948. DOI 10.26342/2019-63-4

of widely used bilingual lexicon extraction methods when applied to a small parallel corpus of Spanish-Nahuatl. We conjecture that in order to find bilingual correspondences in a low-resource setting, it is not only important to pay attention to the methods but also to the morphology of the languages.

We apply different types of morphological processing to measure the impact of morphology in the quality of translation pairs. We trained our own semi-supervised morphological segmentation models.

Our work tries to be as unsupervised as possible, since it is difficult to rely on Nahuatl resources, due to the lack of orthographic norm, big dialectal variation and scarcity of resources.

The corpus used for our experiments is small in terms of the amount of text required by popular NLP models. However, we would like to highlight the challenges that arise when working with this language pair, and also to establish a first step in the development of automatic translation technology which is currently not available for this language pair.

The structure of the paper is as follows. Section 2 contains a brief overview regarding bilingual lexicon extraction methods and morphological phenomena of languages. Section 3 describes the dataset and methods used in our work. Finally, section 4 and 5 contain the discussion and conclusions based on the results.

2 Related work

2.1 Bilingual lexicon extraction

Bilingual lexicon extraction has been an active area of research for several years, especially with the availability of big amounts of parallel corpora that allow to model the relations between lexical units in one corpus and lexical units of the translated texts. This task became very important in terms of SMT systems, where word and phrase alignments are a fundamental step to translate a whole sentence.

Some of the most popular methods for word alignment are the IBM-models (Brown et al., 1993). They are based on an estimation approach that involves the use of probabilistic models that estimate parameters through a maximization process and produces probabilistic tables of lexical translations.

There are also methods that are based on an association approach, where word association or similarity measures are taken into account to find the word translations in a parallel corpus (Tufiş and Barbu, 2002; Ahrenberg, Andersson, and Merkel, 1998; Fung and Yee, 1998; Moore, 2005; Lardilleux, Lepage, and Yvon, 2011).

However, the quality of word alignment methods is heavily dependant on the amount of parallel data. There are alternative approaches, e.g., some works assume that if there is not enough parallel corpora for a language pair, 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 and temporal similarity (Schafer and Yarowsky, 2002), association measures, topical information (Mimno et al., 2009) and contextual features (Harris, 1954; Firth, 1957).

There are many works focused on the latter, the most recent ones use distributional and distributed vector representations. The idea is to build multilingual representations or to map monolingual vectors to the same space in order to find the closest translations (Lauly, Boulanger, and Larochelle, 2014; Hermann and Blunsom, 2014; Mikolov, Le, and Sutskever, 2013). These state of the art methods may not require parallel corpora but they are still based on huge amounts of monolingual or comparable corpora in order to work properly. It has been shown that when they face a low resource setting they can have even worst performance than less sophisticated methods (Levy et al., 2016).

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

2.2 Morphology

Morphology deals with the internal structure of words. Languages of the world have different word production processes. This morphological richness vary from language to language, depending on their linguistic typology.

In NLP, morphology is usually tackled by building morphological analysis/taggers tools. Commonly, lemmatization and stemming methods are used to reduce the morphological variation by converting words

forms to a standard form, i.e., a lemma or a stem. However, most of these technologies are focused in a reduced set of languages. For languages like English, with plenty of resources and relatively poor morphology, morphological processing may be considered solved.

However, this is not the case for all the languages. Specially for those with complex morphological phenomena where it is not enough to remove inflectional endings in order to obtain a stem.

Taking into account the morphological characteristics of languages is important for bilingual lexicon extraction tasks, since the alignment complexity between typologically different languages is far from the alignment complexity between similar languages (Cakmak, Acar, and Eryigit, 2012).

Moreover, when the morphological variation in a text is reduced, it can lead to an improvement of the performance of several NLP tasks, specially for low resource settings. For instance, Nießen and Ney (2004) incorporated the morphology of a low-resource language pair. By doing this, they significantly reduced the amount of parallel corpora needed to train a machine translation system.

Recently there has been a renewed interest in morphology from the NLP perspective, e.g., building vector representations of the morphs in order to improve the word representations (Lazaridou et al., 2013; Botha and Blunsom, 2014; Soricut and Och, 2015).

3 Methodology

3.1 The parallel corpus

Nahuatl is an indigenous language of Mexico with around 1.5M speakers. It is mostly spoken in central Mexico. Nahuatl does not have a web presence or text production comparable to Spanish. We decided to work with a parallel corpus. Since most of the documents that can be easily found in Nahuatl are translations, it seems easier to obtain parallel corpora than monolingual.

We used an existent digital parallel corpus that was created for this language pair (Gutierrez-Vasques, Sierra, and Pompa, 2016) and that is freely available through a search interface[1]. It is important to mention that this corpus was originally extracted from non digital books, therefore, it was digitized using an optical character recognition software (OCR) that could not properly identify Nahuatl words and made several types of mistakes (the authors performed a manual correction).

The documents gathered in this parallel corpus are not homogeneous, they come from different domains and there is dialectal, diachronic and orthographic variation. The Nahuatl language does not have a standardized orthography. The lack of an orthographic norm is an issue that has a negative impact in NLP tasks, since there can be many different word forms corresponding to the same word.

Although this corpus has around 1,186,662 tokens (taking into account the documents of both languages), we only used a subset (Table 1). We chose the documents that had more or less systematic writing, i.e., similar orthography regardless of the domain.

Language	**Tokens**	**Types**	**Sentences**
Spanish (ES)	118364	13233	5852
Nahuatl (NA)	81850	21207	5852

Table 1: Size of the parallel corpus

3.2 Morphological normalization and segmentation

We have mentioned before that taking into account typological properties of the morphology of the languages can help to improve the estimation of bilingual correspondences. In this sense, it is not only important to pay attention in the word alignment methods but also in the morphological text representation.

On one hand, reducing the morphological variation in a text can counteract the negative effect of working in a low-resource setting. The more productive the morphological inflection system of a language, the greater the number of different word forms in a text (Kelih, 2010), i.e., for highly productive languages it is less likely to find repeated words within a text (specially if there is corpora scarcity). Few repetitions of words can be problematic for extracting bilingual correspondences using statistical NLP methods, e.g., not enough contexts to model a word.

On the other hand, in order to find lexical

[1] Axolotl corpus http://www.corpus.unam.mx/axolotl

correspondences between Spanish and Nahuatl, it is important to notice their morphological phenomena. Nahuatl can agglutinate many different prefixes and suffixes to build complex words. Spanish uses mainly suffixes and it has a fusional behavior, where morphemes can be fused or overlaid into a single one. Table 2 shows an example[2].

ti-nech-caqui-z-nequi 2.SG.S-1.S.O-'hear'-FUT-'want' "Tú me quieres oír" (Spanish) *you want to hear me*
Lexical correspondence: oir-caqui

Table 2: Example of Nahuatl-Spanish

In order to normalize the texts and obtain morphological representations, we used different types of processing. For languages like Spanish, it may be enough to use lemmatization or stemming to reduce morphological variation. There are plenty of available tools, so we performed lemmatization and stemming of the Spanish texts using the FreeLing tool (Padró and Stanilovsky, 2012).

However, for Nahuatl, it is necessary to perform morphological segmentation. Since Nahuatl is a language that faces scarcity of data and of language technologies, we trained our own semi-supervised segmentation models using Morfessor 2.0 (Virpioja et al., 2013).

The development and test sets were built with the help of linguists that morphologically segmented a set of words. Morfessor works in a unsupervised way, i.e., it does not necessarily require a tagged training corpus to build a model. However, a tagged development set, with gold-standard segmentations, can be added in order to improve the generated model. For Nahuatl, we used small development and test sets, therefore, we used a semi-supervised setting.

Although it is not common to perform morphological segmentation of Spanish, we also trained a semisupervised segmentation model using Morfessor 2.0 for Spanish. In this case, there were already available development and test sets (Méndez-Cruz, Medina-Urrea, and Sierra, 2016). Table 3 shows the size of the datasets that were used for training morphological segmentation models.

[2]Leipzig Glossing Rules were used for interlinear morpheme-by-morpheme glosses

Spanish	**Tokens**	**Types**
Training	2175533	99564
Development	800	800
Test	792	792
Nahuatl	**Tokens**	**Types**
Training	83229	22174
Development	1379	1379
Test	288	288

Table 3: Datasets used for morphological segmentation

Since we aimed to achieve the best segmentation possible under these conditions, we optimized the main parameters involved in the segmentation models trained with Morfessor 2.0. We tried several values (from 0.1 to 10) of the unannotated corpus likelihood weight (α) that controls the oversegmentation or undersegmentation of the model (Smit et al., 2014). We also tried several ways of calculating the counts of words during the training phase (token based training, log of tokens, type based training).

Tables 4 and 5 show several of our segmentation models and their evaluation. BPR metric (Virpioja et al., 2011) was used, this is a popular metric that evaluates how correctly the morpheme boundaries are placed within the words, compared to a gold-standard reference.

Finally, we selected the best evaluated models. For Nahuatl we used the model obtained with α =0.8 and for Spanish the one with α =0.4.

3.3 Spanish-Nahuatl lexicon extraction

Since we have parallel corpus available, one natural step would be to train a SMT system. However, since we are facing a low-resource setting of distant languages, we decided to take a "first-things-first" approach (Monson et al., 2004), i.e., first explore how difficult would be to find word to word correspondences and the impact that different morphological representations can have for improving this task.

In order to automatically extract bilingual word pairs, we used two different approaches. We used an estimation approach, in particular, we used IBM model 1 (Brown et al., 1993). This model has proven to be still a

α	Pre	Rec	F
0.1	67.7%	89.3%	77%
0.2	69.6%	87.7%	77.5%
0.3	70.1%	84.9%	76.8%
0.4	70.9%	83.6%	76.7%
0.5	71.2%	82.3%	76.4%
0.6	72.5%	81.6%	76.8%
0.7	73.5%	79.6%	76.4%
0.8	**75.1**%	**80**%	**77.5**%
0.9	75.5%	77.4%	76.4%
1	76.6%	77.4%	77%
3	84.5%	49.6%	62.5%
10	97.6%	22.9%	37.1%

Table 4: Morphological segmentation of Nahuatl using Morfessor 2.0 (token based training)

α	Pre	Rec	F
0.1	77.1%	82.6%	79.7%
0.2	80.8%	82.1%	81.4%
0.3	82.2%	81.6%	81.9%
0.4	**84**%	**80.6**%	**82.3**%
0.5	83.9%	78.4%	81.1%
0.6	85%	77.8%	81.3%
0.7	86.3%	77%	81.4%
0.8	86%	75.8%	80.6%
0.9	86.4%	75.5%	80.6%
1	87.2%	75.5%	80.9%
3	89.8%	66.3%	76.2%
10	97.9%	24.4%	39.1%

Table 5: Morphological segmentation of Spanish using Morfessor 2.0 (logarithm of token frequency training)

strong baseline for bilingual lexicon extraction tasks (Levy et al., 2016).

On the other hand, we used an association based approach (Lardilleux and Lepage, 2009), where only those words that appear exactly in the same sentences are considered for alignment. The idea of this method is to produce more candidates, artificially, by creating many subcorpora of small sizes (subsampling). This method is usually known as sampling-based or Anymalign.

We chose these methods because they have similar performance buy they may not make the same mistakes, since they are based in different approaches.

It is not common to test these downstream methods in a low-resource setting. However, sample-based method (anymalign) has shown high accuracy for extracting translation candidates of low-frequency words (Kwon et al., 2014; Lardilleux, Lepage, and Yvon, 2011). On the other hand, IBM models have been used to extract bilingual lexicon from small quantities of parallel sentences of phonemic transcriptions (Adams et al., 2015).

We used different types of morphological text representations: Without any processing (ES-NA), Spanish lemmatized, Nahuatl segmented (ES_{lem}-NA_{morph}) and both languages morphologically segmented (ES_{morph}-NA_{morph}). An evaluation set was built by random sampling 150 Spanish words with frequency greater than 2 in the corpus, stopwords and grammatical words were excluded, the evaluation set mostly contains verbs, nouns and adjectives. For each of these words, several translation candidates were annotated, i.e., an expert annotated possible translations using the parallel corpus as a reference[3]. This evaluation, or test, lexicon required the help of a human annotator since it is difficult to rely in a single standardized Spanish-Nahuatl dictionary to extract evaluation word pairs (due to the orthographic and dialectal variation).

In order to make the results more comparable, we used the same set of words across the different morphological analyses, i.e., we took the random sample of words (without any morphological analysis) and we lemmatized and segmented them to build the different test sets.

Precision at 1 (p@1) and precision at 5 (p@5) were used for the evaluation, i.e., it was evaluated if a correct translation was within the top 1 or the top 5 candidates (Table 6).

It is important to mention that the translations were evaluated out of context. The evaluation, or test, lexicon was done by a human annotator since it is difficult to rely in a single standardized dictionary to extract evaluation word pairs.

Additionally, we were interested in performing a deeper quantitative and qualitative analysis. We focused on the analysis of

[3]The evaluation set and translation candidates can be found at https://github.com/ElotlMX/nahuatl

nouns and verbs. Table 7 shows, from the total amount of nouns and verbs in the evaluation dataset, what proportion was correctly translated.

In a more qualitative analysis, we noticed that there are some verbs in Spanish that, in spite of being very frequent, they were difficult to translate by the methods, e.g., copular or linking verbs (to be), auxiliar verbs (have). We conjecture that some copular and auxiliar verbs were difficult to align, since in Spanish these types of verbs are expressed in the syntax, while in Nahuatl they may be expressed in the morphology (agglutinated to another verb or a noun), or not appear at all.

	IBM %		ANYM %	
	p@1	p@5	p@1	p@5
ES-NA	48.9	73.1	43.8	61.3
ES_{lem}-NA_{morph}	54.6	78.6	**66.6**	**89.3**
ES_{morph}-NA_{morph}	49	73.9	57.4.8	79.9

Table 6: Bilingual lexicon extraction evaluation

	Verbs %		Nouns %	
ES-NA	p@1	p@5	p@1	p@5
IBM	41	66	50.9	75.4
ANYM	33.9	55.3	52.8	67.9
ES_{lem}-NA_{morph}	p@1	p@5	p@1	p@5
IBM	58.3	79.1	50	67.2
ANYM	70.8	**89.5**	53.4	74.1
ES_{morph}-NA_{morph}	p@1	p@5	p@1	p@5
IBM	38.2	61.7	54.3	**78.2**
ANYM	47	79.4	60.8	73.9

Table 7: Proportion of nouns and verbs in the dataset that were correctly translated

4 Discussion

Table 6 shows that all the types of morphological processing that we applied, helped to improve the bilingual lexicon extraction in our low-resource setting.

The most suitable text representation for obtaining translation pairs seems to be ES_{lem}-NA_{morph}. This type of morphological representation not only achieves better results compared to the methods applied to texts without any processing *(ES-NA)*, but the obtained pairs are closer to a Spanish-Nahuatl dictionary entry (word forms with few or none inflections).

Moreover, this setting is the one that is able to obtain the translation of more verbs (Table 7). This is important since verbs constitute the morphologically most complex word class in Nahuatl. Therefore, we would expect these translations were the more challenging to obtain.

The association based method (*ANYM*) seems to greatly benefit from the morphological analysis.

The evaluation of the ES_{morph}-NA_{morph} setting was not so straightforward. The precision in this setting was calculated taking into account less translation pairs, since we had to discard several problematic cases, e.g., Spanish single morphs that correspond to several grammatical functions. It is not always possible to find a morph to morph correspondence between the two languages, due to the differences between their morphological phenomena.

5 Conclusions

In this work, we present morphological segmentation models for Nahuatl and Spanish and we performed Spanish-Nahuatl bilingual lexicon extraction in a low-resource setting. Our conjecture was that morphology plays an important role to improve the performance of this task, specially when we deal with two morphologically-rich distant languages. We applied different types of morphological processing to the texts (stemming, lematization, segmentation) and we extracted bilingual correspondences using two popular approaches.

Using lemmas for Spanish and morphs for Nahuatl, combined with a sampling-based method, achieved the best performance. We showed that developing morphological segmentation tools is specially important for Nahuatl, this is an agglutinative and polysynthetic language that benefits from morphological segmentation, however, since Nahuatl is a low-resource language, there is scarcity of language technologies to process it.

As future work, we would like to extend and automatize the evaluation since relying on human annotators is expensive. We hope that this work could be useful for building translation technologies for this language pair.

Acknowledgements

This work was supported by the Mexican Council of Science and Technology (CONACYT), fund 2016- 01-2225.

References

Adams, O., G. Neubig, T. Cohn, and S. Bird. 2015. Inducing bilingual lexicons from small quantities of sentence-aligned phonemic transcriptions. In *12th International Workshop on Spoken Language Translation (IWSLT)*.

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

Botha, J. and P. Blunsom. 2014. Compositional morphology for word representations and language modelling. In *International Conference on Machine Learning*, pages 1899–1907.

Brown, P. F., V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Cakmak, M. T., S. Acar, and G. Eryigit. 2012. Word alignment for english-turkish language pair. In *LREC*, pages 2177–2180.

Firth, J. 1957. Papers in linguistics, 0xford: Oxford university press.

Fung, P. and L. Y. Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics*, pages 414–420. Association for Computational Linguistics.

Gutierrez-Vasques, X., G. Sierra, and I. H. Pompa. 2016. Axolotl: a web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Haghighi, A., P. Liang, T. Berg-Kirkpatrick, and D. Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*, pages 771–779.

Harris, Z. S. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Hermann, K. M. and P. Blunsom. 2014. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*.

Kelih, E. 2010. The type-token relationship in slavic parallel texts. *Glottometrics*, 20:1–11.

Kwon, H.-S., H.-W. Seo, M. Cheon, and J.-H. Kim. 2014. Iterative bilingual lexicon extraction from comparable corpora using a modified perceptron algorithm. *Journal of Contemporary Engineering Sciences*, 7(24):1335–1343.

Kwon, H.-s., H.-w. Seo, and J.-h. Kim. 2013. Bilingual lexicon extraction via pivot language and word alignment tool. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 11–15.

Lardilleux, A. and Y. Lepage. 2009. Sampling-based multilingual alignment. In *Recent Advances in Natural Language Processing*, pages 214–218.

Lardilleux, A., Y. Lepage, and F. Yvon. 2011. The contribution of low frequencies to multilingual sub-sentential alignment: a differential associative approach. *International Journal of Advanced Intelligence*, 3(2):189–217.

Lauly, S., A. Boulanger, and H. Larochelle. 2014. Learning multilingual word representations using a bag-of-words autoencoder. *arXiv preprint arXiv:1401.1803*.

Lazaridou, A., M. Marelli, R. Zamparelli, and M. Baroni. 2013. Compositionally derived representations of morphologically complex words in distributional semantics. In *ACL (1)*, pages 1517–1526.

Levy, O., A. Søgaard, Y. Goldberg, and I. Ramat-Gan. 2016. A strong baseline for learning cross-lingual word embeddings from sentence alignments. *arXiv preprint arXiv:1608.05426*.

Méndez-Cruz, C.-F., A. Medina-Urrea, and G. Sierra. 2016. Unsupervised morphological segmentation based on affixality measurements. *Pattern Recognition Letters*, 84:127–133.

Mikolov, T., Q. V. Le, and I. Sutskever. 2013. Exploiting similarities among languages

for machine translation. *arXiv preprint arXiv:1309.4168*.

Mimno, D., H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889. Association for Computational Linguistics.

Monson, C., A. Lavie, J. Carbonell, and L. Levin. 2004. Unsupervised induction of natural language morphology inflection classes. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology: Current Themes in Computational Phonology and Morphology*, pages 52–61. Association for Computational Linguistics.

Moore, R. C. 2005. Association-based bilingual word alignment. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 1–8. Association for Computational Linguistics.

Nießen, S. and H. Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational linguistics*, 30(2):181–204.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

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

Smit, P., S. Virpioja, S.-A. Grönroos, M. Kurimo, et al. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014*. Aalto University.

Soricut, R. and F. J. Och. 2015. Unsupervised morphology induction using word embeddings. In *HLT-NAACL*, pages 1627–1637.

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

Tufiş, D. and A.-M. Barbu. 2002. Lexical token alignment: Experiments, results and applications. In *Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas*, pages 458–465.

Virpioja, S., P. Smit, S.-A. Grönroos, M. Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

Virpioja, S., V. T. Turunen, S. Spiegler, O. Kohonen, and M. Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *TAL*, 52(2):45–90.

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