Dealing with Out-Of-Vocabulary Problem in Sentence Alignment Using Word Similarity

Hai-Long Trieu
Japan Advanced Institute of Science and Technology
trieulh@jaist.ac.jp

Le-Minh Nguyen
Japan Advanced Institute of Science and Technology
nguyenml@jaist.ac.jp

Phuong-Thai Nguyen
Vietnam National University, Hanoi, Vietnam
thainp@vnu.edu.vn

Abstract

Sentence alignment plays an essential role in building bilingual corpora which are valuable resources for many applications like statistical machine translation. In various approaches of sentence alignment, length-and-word-based methods which are based on sentence length and word correspondences have been shown to be the most effective. Nevertheless a drawback of using bilingual dictionaries trained by IBM Models in length-and-word-based methods is the problem of out-of-vocabulary (OOV). We propose using word similarity learned from monolingual corpora to overcome the problem. Experimental results showed that our method can reduce the OOV ratio and achieve a better performance than some other length-and-word-based methods. This implies that using word similarity learned from monolingual data may help to deal with OOV problem in sentence alignment.

Keywords: sentence alignment, out-of-vocabulary, word similarity, monolingual data

1 Introduction

Sentence alignment plays an important role in building bilingual corpora for statistical machine translation and many other tasks. Given documents from two languages, the task is to align sentences which are translations of each other. There are three main methods in sentence alignment including length-based, word-based, and the combination of the first two methods. Length-based methods were proposed in (Brown et al., 1991; Gale and Church, 1993). (Wu, 1994) and (Melamed, 1996) introduced methods based on word correspondences. Length-based and word-based methods were also combined to make hybrid methods (Moore, 2002; Varga et al., 2007).

Length-based methods which are only based on the number of words or characters in sentence pairs can run very fast but show a low accuracy. Meanwhile, word-based methods which use bilingual lexicon gain high accuracy, but heavily depend on available lexical resources. The length-and-word-based methods which combine length-based and word-based methods (Moore, 2002; Varga et al., 2007) do not depend on lexical resources and overcome the problem of low accuracy in length-based methods. Nonetheless, a drawback of these length-and-word-based methods which trained a bilingual dictionary using IBM models is the OOV problem.

In this work, we propose an approach to deal with the OOV problem in sentence alignment based on word similarity learned from monolingual corpora. Words that were not contained in the bilingual dictionaries were replaced by their similar words from the monolingual corpora. Experiments conducted on English-Vietnamese sentence alignment showed that using word similarity learned from monolingual corpora can help to reduce the OOV ratio and lead to an improvement in comparison with some other length-and-word-based methods.

We describe phases used in our method in Section 2. Experimental results and discussions are analysed in Section 3. An overview of related researches is discussed in Section 4, and conclusions are drawn in Section 5.

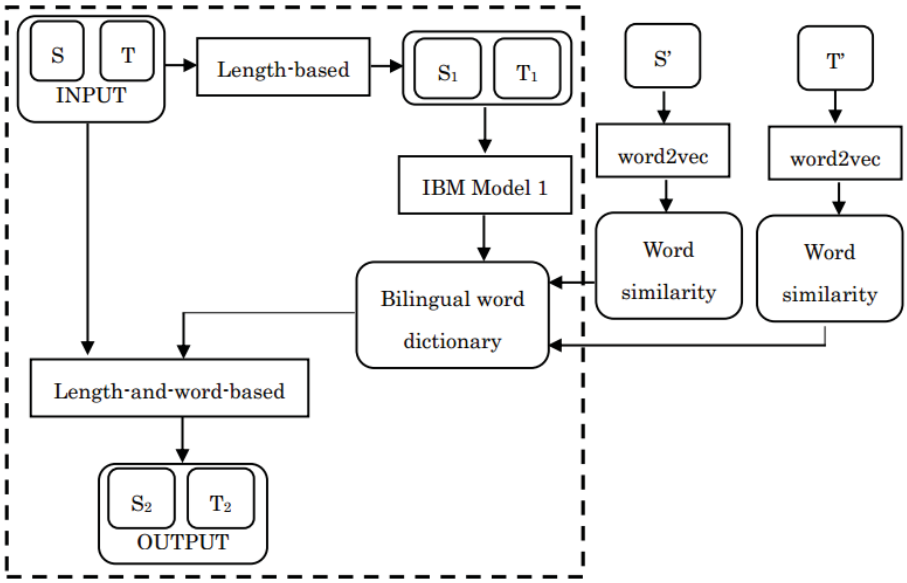

Figure 1: Phases in our model; *S*: the text of source language, *T*: the text of target language; S_1, T_1: sentences aligned by the length-based phase; S_2, T_2: sentences aligned by the length-and-word-based phase; *S'*, *T'*: monolingual corpora of the source and target languages, respectively. The components of the length-and-word-based method (Moore, 2002) are bounded by the dashed frame.

2 Method

In this section, we describe phases used in our method, which include four phases: the length-based phase, the training bilingual dictionaries, using word similarity to deal with the OOV problem, and the combination of length-based and word-based methods. The model is illustrated in Figure 1.

2.1 Length-based Phase

Let l_e and l_v be the lengths of English and Vietnamese sentences, respectively. Then, l_e and l_v varies according to Poisson distribution as follows:

$$P(l_v|l_e) = exp^{-l_v r}\frac{(l_e r)^{l_v}}{l_v!} \quad (1)$$

Where *r* is the ratio of the mean length of Vietnamese sentences to the mean length of English sentences.

As shown in (Moore, 2002), the length-based phase based on the Poisson distribution was slightly better than the Gaussian distribution proposed by (Brown et al., 1991).

$$P(l_v|l_e) = \alpha exp^{-\frac{log(\frac{l_v}{l_e}) - \mu)^2}{2\sigma^2}} \quad (2)$$

Where μ and σ^2 are the mean and variance of the Gaussian distribution, respectively. The length-based model based on the Poisson distribution was shown to be simpler to estimate than the model based on the Gaussian distribution which has to iteratively estimate the variance σ^2 using the expectation maximization (EM) algorithm.

Our model was based on the length-based model using the Poisson distribution.

2.2 Training IBM Model 1

Sentence pairs extracted from the length-based phase are then used to train IBM Model 1 (Brown et al., 1993) to build a bilingual dictionary.

Let e and v be English and Vietnamese sentences, respectively. The procedure of generating sentence v from a sentence e with the length of l_e is as follows:

1. Selecting a length l_v for the sentence v
2. For each word position j in { $1..l_v$} of v:

(a) Selecting a word e_i in e
(b) For each pair (j, e_i): choosing a word v_j to fill the position j

$$P(v|e) = \frac{\epsilon}{(l_e+1)^{l_v}} \prod_{j=1}^{l_v} \sum_{i=0}^{l_e} tr(v_j|e_i) \quad (3)$$

Where ϵ is the uniform probability for all possible lengths of v.

2.3 Using Word Similarity to Deal with OOV

In the sentence alignment task based on word correspondences, bilingual dictionaries trained on IBM models can help to produce highly accurate sentence pairs when they contain reliable word pairs with a high percentage of vocabulary coverage. The OOV problem appears when the bilingual dictionary does not contain word pairs which are necessary to produce a correct alignment of sentences. The higher the OOV ratio, the lower the performance. The bilingual dictionary can also be expanded by training IBM models on available bilingual data. However, such resources are very rare especially for low-resource language pairs like English-Vietnamese. Meanwhile, monolingual data is easy to acquire in an abundant amount. We propose using word similarity learned from monolingual corpora to overcome the OOV problem.

Monolingual corpora of English and Vietnamese were used to train two word similarity models separately using a continuous bag-of-word model. In continuous bag-of-words models, words are predicted based on their context, and words that appear in the same context tend to be clustered together as similar words. We used word2vec (Mikolov et al., 2013), a powerful continuous bag-of-words model to train word similarity. The word2vec model can run very fast and enables to train continuous vector representations of words on large data sets.

The word similarity models were then used to enrich the bilingual dictionary.

1. Let $(e_i - v_j)$ be a word pair in the dictionary in which e_i is the English word, and v_j is the Vietnamese word.

2. Let
   (a) $sim(e_i) = \{e'_{i_1}, ..., e'_{i_m}\}$
   (b) $sim(v_j) = \{v'_{j_1}, ..., v'_{j_n}\}$

   be sets of similar words of e_i and v_j, respectively.

3. The dictionary can be expanded as follows:
   (a) For e' in $sim(e_i)$: add pairs $(e' - v_j)$ to the dictionary
   (b) For v' in $sim(v_j)$: add pairs $(e_i - v')$ to the dictionary
   (c) $score(e' - v_j) = score(e_i - v_j) * cosine(e_i - e')$
   (d) $score(e_i - v') = score(e_i - v_j) * cosine(v_j - v')$

Where $score(a, b)$ is the word translation probability of the word pair (a, b) by training IBM Model 1. $cosine(a, b)$ is the cosine similarity between a and b from word similarity models.

The expanded dictionary can help to cover a higher ratio of vocabulary, which reduces the OOV ratio and improves overall performance.

2.4 Length-based and Word-based

The expanded dictionary was then combined with the length-based phase described in Section 2.1 to produce final alignments, which are described as follows:

$$P(e, v) = \frac{P_{1-1}(l_e, l_v)}{(l_e+1)^{l_v}} \left(\prod_{j=1}^{l_v} \sum_{i=0}^{l_e} tr(v_j|e_i)\right)\left(\sum_{i=1}^{l_e} f_u(e_i)\right) \quad (4)$$

Where f_u is the observed relative unigram frequency of the word in the text in the corresponding language.

3 Experiments

3.1 Setup

We conducted experiments on the sentence alignment task for English-Vietnamese, a low-resource language pair. We evaluated our method on the test set collected from the website.[1] After preprocessing the collected data, we conducted sentence alignment manually to achieve the reference data. We publish

[1] http://www.vietnamtourism.com/

these data sets on the website.[2] The statistics of these data sets are shown in Table 1.

Statistics	Test Data
Sentences (English)	1,705
Sentences (Vietnamese)	1,746
Average length (English)	22
Average length (Vietnamese)	22
Vocabulary Size (English)	6,144
Vocabulary Size (Vietnamese)	5,547
Reference Set	837

Table 1: Statistics of Test Corpus

In order to produce a more reliable bilingual dictionary, we added an available bilingual corpus to train IBM Model 1, which was collected from the IWSLT2015 workshop.[3] The dataset contains subtitles of TED talks (Cettolo et al., 2012). The IWSLT2015 training data is shown in Table 2.

Statistics	iwslt15
Sentences (English)	129,327
Sentences (Vietnamese)	129,327
Average length (English)	19
Average length (Vietnamese)	18
Vocabulary Size (English)	46,669
Vocabulary Size (Vietnamese)	50,667

Table 2: Statistics of the IWSLT15 Corpus

In the preprocessing steps, we tokenized these datasets using the tokenizer of Moses script[4] for English and JVnTextpro[5] for Vietnamese. The datasets were then lowercased. For Vietnamese, we conducted word segmentation using JVnTextpro.

For the sentence alignment algorithm, we reimplemented phases in the model (Moore, 2002) using Java.

To evaluate performance we used common metrics: Precision, Recall, and F-measure (Véronis and Langlais, 2000).

[2]https://github.com/nguyenlab/SentAlign-Similarity

[3]https://sites.google.com/site/iwsltevaluation2015/mt-track

[4]http://www.statmt.org/moses/?n=moses.baseline

[5]http://jvntextpro.sourceforge.net/

3.2 Training Word Similarity

In order to train word similarity models, we used English and Vietnamese monolingual corpora. For English we used the one-billion-words[6] dataset which contains almost 1B words. To build a huge monolingual corpus of Vietnamese, we extracted articles from the web (*www.baomoi.com*)[7]. The data set was then preprocessed to achieve 22 million Vietnamese sentences.

We used word2vec from gensim python[8] to train two word-similarity models on the monolingual corpora. We set the cbow model with configurations: window size=5, vector size=100, min count = 10. The word2vec trained model of Vietnamese is also available on the website.[2]

3.3 Result and Discussion

We compared our model with the two other length-and-word-based methods: M-align[9] (Moore, 2002) and Hun-align[10] (Varga et al., 2007). We showed how our method can deal with the OOV problem.

We setup the length-based phase's threshold to 0.99 to extract highest sentence pairs. Then in the length-and-word-based phase, we setup the threshold to 0.9 to ensure a high confidence. Experimental results are shown in Table 3.

Setup	M-align	Hun-align	OurMethod
Reference	837	837	837
Results	580	1373	609
Correct	412	616	433
Precision	71.03%	44.87%	71.10%
Recall	49.22%	73.60%	51.73%
F-measure	58.15%	55.75%	59.89%

Table 3: Experimental results. (*Reference*, *Results*, *Correct*: number of sentence pairs in reference set, results from systems, and correct sentences, respectively.)

Overall, the performance of our model slightly improved the M-align in all scores of precision, recall, and f-measure. Our model also gained higher

[6]http://www.statmt.org/lm-benchmark/

[7]http://www.baomoi.com/

[8]https://radimrehurek.com/gensim/models/word2vec.html

[9]http://research.microsoft.com/en-us/downloads/aafd5dcf-4dcc-49b2-8a22-f7055113e656/

[10]http://mokk.bme.hu/en/resources/hunalign/

performance than Hun-align. Although Hun-align can achieve the highest recall of 73.60% due to the approach that Hun-align constructs dictionaries, the method produced a number of error results, so this caused the lowest precision.

A problem of using the IBM Model 1 as in Moore's method was the OOV. When the dictionary cannot cover a high ratio of vocabulary, it decreases the contribution of the word-based phase. The average OOV ratio is shown in Table 4. In comparison with M-align, using word similarity in our model reduced the OOV ratio from 7.37% to 4.33% in English and from 7.74% to 6.80% in Vietnamese vocabulary. By using word similarity models we overcame the problem of OOV. The following discussion will show how the word similarity models helped to reduce the OOV ratio.

Setup	**Test**	**M-align**	**Our Model**
#vocab. en	1,705	27,872	28,371
#vocab. vi	1,746	25,326	25,481
OOV en	NA	7.37%	4.33%
OOV vi	NA	7.74%	6.80%

Table 4: Average OOV ratio.

We describe word similarity models using word2vec with examples. Tables 5 and 6 show examples of OOV words and their most similar words extracted from the word similarity models. The word similarity models can explore not only helpful similar words in terms of variants in morphology but also words that share the same meaning but different morphemes. There are useful similar words that can have the same meaning as the OOV words like word pairs ("*intends*" and "*aims*") or ("*honours*" and "*awards*"), ("*quát*", "*mắng*"), ("*ghe*", "*đò*"). However, because in the word2vec model words are predicted based on their context in terms of windows, some word pairs may contain different meanings like ("*bangkok*", "*jakarta*"), or ("*pagoda*", "*citadel*"), ("*phở*", "*cơm*"). Therefore extracting suitable similar words is also needed to be further investigated.

We show an example of how our method deals with the OOV problem in Table 7. The word pairs (*reunification-thống_nhất*) and (*impressively-mạnh_mẽ*) were not covered by the dictionary using IBM Model 1, and this became an example of

OOV Words	**Similar Words**	**Cosine Similarity**
intends	aims	0.74
intends	refuses	0.74
intends	plans	0.66
honours	honors	0.71
honours	prizes	0.65
honours	awards	0.62
bangkok	jerusalem	0.65
bangkok	jakarta	0.61
pagoda	temple	0.86
pagoda	tower	0.76
pagoda	citadel	0.73

Table 5: Examples of English Word Similarity Model

OOV Words	**Similar Words**	**Cosine Similarity**
quát (*to shout*)	mắng (*to scold*)	0.35
quát (*to shout*)	nạt (*to bully*)	0.32
hủy (*to destroy*)	hoại (*to ruin*)	0.50
hủy (*to destroy*)	dỡ (*to unload*)	0.42
hủy (*to destroy*)	phá (*to demolish*)	0.36
ghe (*junk*)	thuyền (*boat*)	0.64
ghe (*junk*)	xuồng (*whaleboat*)	0.61
ghe (*junk*)	đò (*ferry*)	0.56
phở (*noodle soup*)	cháo (*rice gruel*)	0.67
phở (*noodle soup*)	cơm (*rice*)	0.65

Table 6: Examples of Vietnamese Word Similarity Model. The italic words in brackets are corresponding English meaning which were translated by the authors.

Language	Sentence
English	since the reunification in 1975 , vietnam ' s architecture has been impressively developing .
Vietnamese	từ sau ngày đất_nước thống_nhất (1975) kiến_trúc việt_nam phát_triển khá mạnh_mẽ .
(*Translation*)	*After the country was unified (1975), vietnam's architecture has been developing rather impressively.*

Table 7: An example of English-Vietnamese OOV. The translations to English (italic) were conducted by the authors.

OOV Words	Similar Words	Cosine Similarity
reunification	independence	0.71
reunification	**unification**	**0.67**
reunification	peace	0.62
impressively	amazingly	0.74
impressively	**impressive**	**0.74**
impressively	exquisitely	0.72
impressively	brilliantly	0.71

Table 8: An example of similar word pairs trained on monolingual corpus

OOV. Examples of similar word pairs are shown in Table 8, and translation word pairs trained by IBM Model 1 are shown in Table 9. Because (*reunification-unification*) was a similar word pair, and the translation word pair (*unification-thống_nhất*) was contained in the dictionary, the new translation word pair (*reunification-thống_nhất*) was then created. Similarly, the new translation word pair (*impressively-mạnh_mẽ*) was created via the similar word pair (*impressively-impressive*) and the translation word pair (*impressive-mạnh_mẽ*). Table 10 shows induced translation word pairs. By using word similarity learned from monolingual corpora, a number of OOV words can be replaced by their similar words, which helped to reduce the OOV ratio and improve performance in overall.

4 Related Work

Sentence alignment is an essential task in natural language processing, which builds bilingual corpora, a valuable resource in many applications like statistical machine translation, word sense disambiguation, information retrieval, etc. The task can be solved based on the number of words or

Score	English	Vietnamese
0.597130	independence	độc_lập (*independent*)
0.051708	independence	sự_độc_lập (*independence*)
0.130447	**unification**	**thống_nhất** (*to unify*)
0.130447	unification	sự_thống_nhất (*unification*)
0.130446	unification	sự_hợp_nhất (*unify*)
0.551291	impressive	ấn_tượng (*impression*)
0.002927	**impressive**	**mạnh_mẽ** (*impressive*)
0.002440	impressive	kinh_ngạc (*amazed*)

Table 9: An example of bilingual dictionary trained by IBM Model 1 (*Score*: translation probability); the translations to English (italic) were conducted by the authors.

Score	English	Vietnamese
0.215471	reunification	thống_nhất (*to unify*)
0.369082	impressively	mạnh_mẽ (*impressive*)

Table 10: Induced translation word pairs; the translations to English (italic) were conducted by the authors.

characters(Brown et al., 1991; Gale and Church, 1993). These methods are fast and effective in some closed language pairs like English-French but achieve low performance in language pairs like English-Chinese. Word-based methods were proposed in (Kay and Röscheisen, 1993; Chen, 1993; Wu, 1994; Melamed, 1996; Ma, 2006), based on lexical resources. These methods showed better performance than length-based methods, but they depend on available linguistic resources, which are rare and expensive to achieve in almost all language pairs, especially in low-resource languages like English-Vietnamese. Hybrid methods which combine length-based and word-based methods as shown in (Moore, 2002; Varga et al., 2007) can overcome the low accuracy of length-based methods, and these methods also do not depend on lexical resources.

(Varga et al., 2007) proposed building bilingual corpora for medium-density languages. This can overcome the problem of the unavailability of bilingual resources of low-resource languages by building dictionaries and merge them to make a huge dictionary to cover a high ratio of vocabulary. However, because the method does not compute the score of word pairs in dictionaries, this leads to a low precision. Moore's method (Moore, 2002) can gain high accuracy, but the method has to deal with the OOV problem. Our model is similar to Moore's method, but we can overcome the OOV problem based on word similarity learned from monolingual corpora using a continuous bag-of-words model.

Continuous bag-of-words models were proposed in (Mikolov et al., 2013), which can learn word similarity on very monolingual data. The model also has been applied to learn phrase similarity on monolingual data to improve statistical machine translation (Zhao et al., 2015).

In using monolingual data for alignment tasks, (Trieu et al., 2014) proposed using word clustering trained on monolingual data to improve the Moore's method (Moore, 2002). In our model, we also based on word similarity learned from monolingual data, but we used a strong technique of word vector representation, word2vec, to learn word similarity. (Songyot and Chiang, 2014) proposed a method using word similarity from monolingual corpora to improve machine translation. In the work of (Songyot and Chiang, 2014), the word similarity is trained based on a word context model using a feedforward neural network and then applied to improve statistical machine translation.

The idea of using the word similarity model learned from monolingual data based on word2vec in our work is closed to the research of (Li et al., 2016). In (Li et al., 2016), the word similarity model is used to substitute rare words in neural machine translation. In our work, we adopted the word similarity model to overcome the out-of-vocabulary problem in sentence alignment.

5 Conclusion

In this work, we propose using word similarity to overcome the problem of OOV in sentence alignment. The word2vec model was trained on monolingual corpora to produce word-similarity models. These models were then combined with the bilingual word dictionary trained on IBM Model 1, which were integrated to length-and-word-based phase in a sentence alignment algorithm. Our method can reduce the OOV ratio with similar words learned from monolingual corpora, which leads to an improvement in comparison with some other length-and-word-based methods. Using word similarity trained on monolingual corpora based on a distributed word representation model like word2vec may help to reduce the OOV in sentence alignment. Some aspects of this work need to be more investigated in future work like: applying word similarity in sentence alignment in a large scale data; exploring the contribution of word2vec in this task like using both the cbow and skip-gram models. We also plan to further leverage monolingual corpora to sentence alignment and then apply to statistical machine translation, especially for low-resource languages.

References

Peter F Brown, Jennifer C Lai, and Robert L Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 169–176. Association for Computational Linguistics.

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.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268.

Stanley F Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 9–16. Association for Computational Linguistics.

William A Gale and Kenneth W Church. 1993. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Martin Kay and Martin Röscheisen. 1993. Text-translation alignment. *Computational linguistics*, 19(1):121–142.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. Towards zero unknown word in neural machine translation. In *Proceedings of the 25th International Conference on Artificial Intelligence*.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *LREC 2006: Fifth International Conference on Language Resources and Evaluation*, pages 489–492.

I Dan Melamed. 1996. A geometric approach to mapping bitext correspondence. *arXiv preprint cmp-lg/9609009*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Robert C Moore. 2002. *Fast and accurate sentence alignment of bilingual corpora*. Springer.

Theerawat Songyot and David Chiang. 2014. Improving word alignment using word similarity. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1840–1845.

Hai-Long Trieu, Phuong-Thai Nguyen, and Kim-Anh Nguyen. 2014. Improving moore's sentence alignment method using bilingual word clustering. In *Knowledge and Systems Engineering*, pages 149–160. Springer.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam studies in the theory and history of linguistic science series 4*, 292:247.

Jean Véronis and Philippe Langlais. 2000. Evaluation of parallel text alignment systems. In *Parallel text processing*, pages 369–388. Springer.

Dekai Wu. 1994. Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 80–87. Association for Computational Linguistics.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning translation models from monolingual continuous representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.