Learning Bilingual Projections of Embeddings for Vocabulary Expansion in Machine Translation

Pranava Swaroop Madhyastha*
Department of Computer Science
University of Sheffield
Sheffield, S1 4DP, UK
p.madhyastha@sheffield.ac.uk

Cristina España-Bonet*
University of Saarland
DFKI, German Research Center
for Artificial Intelligence
Saarbrücken, Germany
cristinae@dfki.de

Abstract

We propose a simple log-bilinear softmax-based model to deal with vocabulary expansion in machine translation. Our model uses word embeddings trained on significantly large unlabelled monolingual corpora and learns over a fairly small, word-to-word bilingual dictionary. Given an out-of-vocabulary source word, the model generates a probabilistic list of possible translations in the target language using the trained bilingual embeddings. We integrate these translation options into a standard phrase-based statistical machine translation system and obtain consistent improvements in translation quality on the English–Spanish language pair. When tested over an out-of-domain test-set, we get a significant improvement of 3.9 BLEU points.

1 Introduction

Data-driven machine translation systems are able to translate words that have been seen in the training parallel corpora, however translating unseen words is still a major challenge for even the best performing systems. The amount of parallel data is finite (and sometimes scarce) and, therefore, word types like named entities, domain specific content words, or infrequent terms are rare. This lack of information can potentially result in incomplete or erroneous translations.

This problem has been actively studied in the field of machine translation (MT) (Habash, 2008; Daumé III and Jagarlamudi, 2011; Marton et al., 2009; Rapp, 1999; Dou and Knight, 2012; Irvine and Callison-Burch, 2013). Lexicon-based resources have been used for resolving unseen content words by exploiting a combination of monolingual and bilingual resources (Rapp, 1999; Callison-Burch et al., 2006; Zhao et al., 2015). In this context, distributed word representations, or word embeddings (WE), have been recently applied to resolve unseen word related problems (Mikolov et al., 2013b; Zou et al., 2013). In general, word representations capture rich linguistic relationships and several works (Gouws et al., 2015; Wu et al., 2014) try to use them to improve MT systems. However, very few approaches use them directly to resolve the out-of-vocabulary (OOV) problem in MT systems.

Previous research in MT systems suggests that a significant number of named entities (NE) can be handled by using simple pre or post-processing methods, e.g., transliteration techniques (Hermjakob et al., 2008; Al-Onaizan and Knight, 2002). However, a change in domain results in a significant increase in the number of unseen content words for which simple pre or post-processing methods are sub-optimal (Zhang et al., 2012).

Our work is inspired by the recent advances (Zou et al., 2013; Zhang et al., 2014) in applications of word embeddings to the task of vocabulary expansion in the context of statistical machine translation (SMT). Our focus in this paper is to resolve unseen content words by using continuous word embeddings on both the languages and learn a model over a small seed lexicon to map the embedding spaces. To this extent, our work is similar to Ishiwatari et al. (2016) where the authors map distributional representations using a linear regression method similar to Mikolov et al. (2013b) and insert a new feature based on cosine similarity metric into the MT system. On the other hand, there is a rich body of recent literature that focuses on obtaining bilingual word

* This work was done while the authors were in TALP Research Center, Universitat Politècnica de Catalunya, Barcelona.

Proceedings of the 2nd Workshop on Representation Learning for NLP, pages 139–145,
Vancouver, Canada, August 3, 2017.

embeddings using either sentence aligned or document aligned corpora (Bhattarai, 2012; Gouws et al., 2015; Kočiský et al., 2014). Our approach is significantly different as we obtain embeddings separately on monolingual corpora and then use supervision in the form of a small sparse bilingual dictionary, in some terms similar to Faruqui and Dyer (2014). We use a simple yet principled method to obtain a probabilistic conditional distribution of words directly and these probabilities allow us to expand the translation model for new words.

The rest of the paper is organised as follows. Section 2 presents the log-bilinear softmax model, and its integration into an SMT system. The experimental work is described in Section 3. Finally, we conclude and sketch some avenues for future work.

2 Mapping Continuous Word Representations using a Bilinear Model

Definitions. Let $\mathcal{E}$ and $\mathcal{F}$ be the vocabularies of the two languages, source and target, and let $e \in \mathcal{E}$ and $f \in \mathcal{F}$ be words in these languages respectively. Let us assume, we have a source word to target word $e \rightarrow f$ dictionary. We also assume that we have access to some kind of distributed word embeddings in both languages, ϕ_s for the source and ϕ_t for the target, where $\phi(.) \rightarrow \mathbb{R}^n$ denotes the n-dimensional distributed representation of the words. The task we are interested in is to learn a model for the conditional probability distribution $\Pr(f|e)$. That is, given a word in a source language, say English (e), we want to get a conditional probability distribution of all the words in a foreign language (f).

Log-Bilinear Softmax Model. We formulate the problem as a bilinear prediction task as proposed by Madhyastha et al. (2014a) and extend it for the bilingual setting. The proposed model makes use of word embeddings on both languages with no additional features. The basic function is formulated as log-bilinear softmax model and takes the following form:

$$\Pr(f|e;W) = \frac{\exp\{\phi_s(e)^\top W \phi_t(f)\}}{\sum_{f' \in \mathcal{F}} \exp\{\phi_s(e)^\top W \phi_t(f')\}} \quad (1)$$

Essentially, our problem reduces to: a) first obtaining the corresponding word embeddings of the vocabularies from both the languages using a significantly large monolingual corpus and b) estimating W given a relatively small dictionary. That is, to learn W we use the source word to target word dictionary as training supervision. The dictionary can be a true bilingual dictionary or the word alignments generated by the SMT system, therefore, no additional resources to the training parallel corpus are needed.

We learn W by minimizing the negative log-likelihood of the dictionary using a regularized (relaxed low-rank regularization based) objective as: $L(W) = -\sum_{e,f} \log(\Pr(f|e;W)) + \lambda \|W\|_p$. λ is the constant that controls the capacity of W. To find the optimum, we follow previous (Madhyastha et al., 2014b) work and use an optimization scheme based on Forward-Backward Splitting (FOBOS) (Singer and Duchi, 2009).

We experiment with two regularization schemes, $p = 2$ or the ℓ_2 regularizer and $p = *$ or the ℓ_* (nuclear norm) regularizer. We find that both norms have approximately similar performance, however the trace norm regularized W has lower capacity and hence, smaller number of parameters. This is also observed by (Bach, 2008; Madhyastha et al., 2014a,b). In general, we can apply the ideas used by Mikolov et al. (2013b) to speed up the training as this model is equivalent to a softmax model. We can obtain models with similar properties if we change the loss from bilinear log softmax to a bilinear margin based loss. We leave this exploration for future work.

A by-product of regularizing with ℓ_* norm is a lower-dimensional, language aligned, and compressed embeddings for both languages. This is possible because of the induced low-dimensional properties of W. That is, assume W has rank k, where $k < n$, such that $W \approx U_k V_k^\top$, then the product:

$$\phi_s(e)^\top U_k V_k^\top \phi_t(f) \quad (2)$$

gives us $\phi_s(e)^\top U_k$ and $V_k^\top \phi_t(f)$ compressed embeddings with shared properties. These are similar to the CCA based projections obtained in Faruqui and Dyer (2014).

Integrating the Probabilistic List into the SMT System. We integrate the probabilistic list of translation options into the phrase-based decoder using the standard log-linear approach (Och and Ney, 2002). Consider a word pair (e, f), where the decoder searches for a foreign word $\hat{f}$ that maxi-

Table 1: Top-10 accuracy (in percentage) for bilingual dictionary induction for English–German and English–French.

		Strong supervision		Soft supervision			
l_1	l_2	BiSkip	BiCVM	BiCCA	BiVCD	Ours-300	Ours-100
en	de	79.7	74.5	72.4	62.5	73.8	71.1
en	fr	78.9	72.9	70.1	68.8	72.1	69.7

mizes a linear combination of feature functions:

$$\widehat{f} = \text{argmax}_f \{\sum \lambda_i \log(h_i(f,e)) + \lambda_{oov} \log(\Pr(f,e))\}$$

here, λ_i is the weight associated with feature $h_i(f,e)$ and λ_{oov} is the weight associated with the unseen word.

3 Empirical Analysis

Quality of the Learned Embeddings. To understand the performance of the embedding projections in our model, we perform experiments to compute the top-10 accuracy of our models in the same setting provided in Upadhyay et al. (2016) for cross-lingual dictionary induction[1]. The evaluation task judges how good cross-lingual embeddings are at detecting word pairs that are semantically similar across languages. Similarly to Upadhyay et al. (2016), we compare against BiSkip embeddings (Luong et al., 2015a), BiCVM (Hermann and Blunsom, 2014), BiCCA (Faruqui and Dyer, 2014) and BiVCD (Vulic and Moens, 2015). We experiment with English–German and English–French language pairs, so that we can induce the dictionaries for the five systems. As seen in Table 1, our full 300-dimensional embeddings perform better than the BiCCA-based model, whereas 100-dimensional compressed embedding perform slightly worse, but still are competitive. Since our model and BiCCA use similar supervision, we obtain similar results and differ in a similar way to those that use stronger supervision like BiCVM and BiSkip based embeddings.

MT Data and System Settings. For estimating the monolingual WE, we use the CBOW algorithm as implemented in the `Word2Vec` package (Mikolov et al., 2013a) using a 5-token window. We obtain 300 dimension vectors for English and Spanish from a Wikipedia dump of 2015 and the Quest data[2]. The final corpus contains 2.27 billion tokens for English and 0.84 for Spanish. We remove any occurrence of sentences from the test set that are contained in our corpus. The coverage in our test sets is of 97% of the words.

To train the log-bilinear softmax based model, we use the dictionary from the `Apertium project`[3] (Forcada et al., 2011). The dictionary contains 37651 words, 70% of them are used for training and 30% as a development set for model selection. The average precision @1 is 86% for the best model over the development set.

A state-of-the-art phrase-based SMT system is trained on the Europarl corpus (Koehn, 2005) for the English-to-Spanish language pair. We use a 5-gram language model that is estimated on the target side of the corpus using interpolated Kneser-Ney discounting with `SRILM` (Stolcke, 2002). Additional monolingual data available within Quest corpora is used to build a larger language model with the same characteristics. Word alignment is done with `GIZA++` (Och and Ney, 2003) and both phrase extraction and decoding are done with the `Moses` package (Koehn et al., 2007). At decoding time, `Moses` allows to include additional translation pairs with their associated probabilities to selected words via xml markup. We take advantage of this feature to add our probabilistic estimations to each OOV. Since, by definition, OOV words do no appear in the parallel training corpus, they are not present in the translation model either and the new translation options only interact with the language model. The optimization of the weights of the model with the additional translation options is trained with MERT (Och, 2003) against the BLEU (Papineni et al., 2002) evaluation metric on the NewsCommentaries 2012[4] (NewsDev) set. We test our systems on the NewsCommentaries 2013 set (NewsTest) for an in-domain evaluation and on a test set

[1] We also used the script provided here: https://github.com/shyamupa/biling-survey

[2] http://statmt.org/~buck/wmt13qe/wmt13qe_t13_t2_MT_corpus.tgz

[3] The bilingual dictionary can be downloaded here: http://goo.gl/TjH31q.

[4] http://www.statmt.org/wmt13/translation-task.html

Table 2: OOVs on the dev and test sets.

	Sent.	Tokens	OOV_{all}	OOV_{CW}
NewsDev	3003	72988	1920 (2.6%)	378 (0.5%)
NewsTest	3000	64810	1590 (2.5%)	296 (0.5%)
WikiTest	500	11069	798 (7.2%)	201 (1.8%)

extracted from Wikipedia by Smith *et. al.* (2010) for an out-of-domain evaluation (WikiTest).

The *domainess* of the test set is established with respect to the number of OOVs. Table 2 shows the figures of these sets paying special attention to the OOVs in the basic SMT system. Less than a 3% of the tokens are OOVs for News data (OOV_{all}), whereas it is more than a 7% for Wikipedia's. In our experiments, we distinguish between OOVs that are named entities and the rest of content words (OOV_{CW}). Only about 0.5% (NewsTest) and 1.8% (WikiTest) of the tokens fall into this category, but we show that they are relevant for the final performance.

MT Experiments. We consider two baseline systems, the first one does not output any translation for OOVs (*noOOV*), it just ignores the token; the second one outputs a verbatim copy of the OOV as a translation (*verbatimOOV*). Table 3 shows the performance of these systems under three widely used evaluation metrics TER (Snover et al., 2006), BLEU and METEOR (MTR) (Banerjee and Lavie, 2005). Including the verbatim copy improves all the lexical evaluation metrics. Specially for NEs and acronyms (the 80% of OOVs in our sets), this is a hard baseline to be compared to as in most cases the same word is the correct translation.

We then enrich the systems with information gathered from the large monolingual corpora in two ways, using a bigger language model (*BLM*) and using our newly proposed log-bilinear model that uses word embeddings (*BWE*). BLMs are important to improve the fluency of the translations, however they may not be helpful for resolving OOVs as they can only promote translations available in the translation model. On the other hand, BWEs are important to make available to the decoder new vocabulary on the topic of the otherwise OOVs. Given the large percentage of NEs in the test sets (Table 2), our models add the source word as an additional option to the list of target words to mimic the *verbatimOOV* system.

Table 3 includes seven systems with the additional monolingual information. Three of them add, at decoding time, the top-n translation options given by the BWE for a OOV. *BWE* system uses the top-50 for all the OOVs, BWE_{CW50} also uses the top-50 but only for content words other than named entities[5], and BWE_{CW10} limits the list to 10 elements. *BLM* is the same as the baseline system *verbatimOOV* but with the large language model. *BLM+BWE*, $BLM+BWE_{50}$ and $BLM+BWE_{10}$ combine the three BWE systems with the large language model.

Table 3: Automatic evaluation of the translation systems defined in Section 3. The best system is bold-faced (see text for statistical significance).

	NewsTest			WikiTest		
	TER	BLEU	MTR	TER	BLEU	MTR
noOOV	58.21	21.94	45.79	61.26	16.24	38.76
verbatimOOV	57.90	22.89	47.06	58.55	21.90	45.77
BWE	58.33	22.23	45.76	58.38	21.96	44.84
BWE_{CW50}	57.66	23.09	47.14	56.19	24.16	48.49
BWE_{CW10}	57.85	23.06	47.11	55.64	24.71	49.05
BLM	55.37	25.83	**49.19**	52.60	30.63	51.04
BLM+BWE	55.89	24.92	47.84	51.02	32.20	52.09
$BLM+BWE_{50}$	55.55	25.61	49.01	49.50	33.94	54.93
$BLM+BWE_{10}$	**55.31**	**25.86**	49.04	**49.12**	**34.58**	**55.52**

In the NewsTest, most of unseen words are named entities and using BWEs to translate them barely improves the translation. The reason is that embeddings of related NEs are usually equivalent. This affects the overall integration of the scores into the decoder and induces ambiguity in the system. However, we observe that the decoder benefits from the information on content words, specially for the out-of-domain WikiTest set. In this case, given the constrained list of alternative translations (BWE_{CW10}) one achieves 2.75 BLEU points of improvement.

The addition of the large language model improves the results significantly. When combined with the BWEs we observe that the BWEs clearly help in the translation of WikiTest but do not seem as relevant in the in-domain set. We achieve a statistically significant improvement of 3.9 points of BLEU with the BLM and BWE combo system – $BLM+BWE_{10}$ with respect to *BLM*– in WikiTest ($p<0.001$); the improvement in the NewsTest is not statistically significant (p-value=0.29). The number of translation options in the list is also

[5]We consider a named entity any word that begins with a capital letter and is not after a punctuation mark, and any fully capitalized word.

Table 4: Top-n list of translations obtained with the bilingual embeddings.

GALAXY	NYMPHS	STUART	FOLKSONG
galaxia	**ninfas**	William	música
planeta	ninfa	Henry	**folclore**
universo	crías	John	literatura
planetas	diosa	Charles	himno
galaxias	dioses	Thomas	folklore
...	...	Estuardo (#48)	**canción (#7)**

relevant, and for *BLM+BWE*$_{\text{CW50}}$ we have a significant but smaller improvement of 3.3 points on BLEU in WikiTest. All these results are consistent among different evaluation metrics.

In order to estimate the relevance of the bilingual embeddings into the final translation, we have manually evaluated the translation of WikiTest using the *BWE*$_{\text{CW50}}$ model. For the translation of the OOVs, we obtain an accuracy of a 68%, that is, the BWE gives the correct translation option at least 68% of the times. We note that, even if the correct translation option is in the translation list obtained by the BWE, the decoder may choose not to consider it.

In general, we observe that when our model fails, in most of the cases, the words in the translated language happened to be either a multiword expression or a named entity. In Table 4 we present some of the these examples. The first two examples *galaxy* and *nymphs* are nouns where we obtain the first option as the correct translation. The problem is harder for named entities as we observe in the table, the name *Stuart* in English has *William* as most probable translation in Spanish, the correct translation *Estuardo* however appears as the 48th choice. Our model is also unable to generate multiword expressions, as shown in the table for the english word *folksong*, the correct translation being *canción folk*. This would need two words in Spanish in order to be translated correctly, however, our model does obtain words: *canción* and *folclore* as the most probable translation options.

4 Conclusions

We have presented a method for resolving OOVs in SMT that performs vocabulary expansion by using a simple log-bilinear softmax based model. The model estimates bilingual word embeddings and, as a by-product, generates low-dimensional compressed embeddings for both languages.The addition of the new translation options to a mere 1.8% of the words has allowed the system to obtain a relative improvement of a 13% in BLEU (3.9 points) for out-of-domain data. For in-domain data, where the number of content words is small, improvements are more moderate.

The analysis of the results shows how the performance is damaged by not considering multiword expressions. The automatic detection of these elements in the monolingual corpus together with the addition of one-to-many dictionary entries for learning the W matrix can alleviate this problem and will be considered in future work.

We also note that this approach can be extended directly within neural machine translation systems, where its effects could be even larger due to the limited vocabulary. While one of the popular approaches to deal with OOVs is to use subword units (Sennrich et al., 2016) in order to resolve of unknown words, dictionary-based approaches, where an unknown word is translated by its corresponding translation in a dictionary or a (SMT) translation table, have also been used (Luong et al., 2015b). Our method can go further in the latter direction by learning correspondences of source and target vocabularies using large monolingual corpora and either a small dictionary or the word alignments.

References

Yaser Al-Onaizan and Kevin Knight. 2002. Machine transliteration of names in arabic text. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*. Association for Computational Linguistics, pages 1–13.

Francis R Bach. 2008. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research* 9:1179–1225.

Satanjeev Banerjee 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*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 65–72.

Alexandre Klementiev Ivan Titov Binod Bhattarai. 2012. Inducing Crosslingual Distributed Representations of Words. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, pages 1459–1474.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. HLT-NAACL '06, pages 17–24.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Association for Computational Linguistics*. Portland, OR, pages 407–412.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 266–275.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*. pages 462–471.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim ORegan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation* 25(2):127–144.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. pages 748–756.

Nizar Habash. 2008. Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. pages 57–60.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Baltimore, Maryland, pages 58–68.

Ulf Hermjakob, Kevin Knight, and Hal Daumé III. 2008. Name translation in statistical machine translation: Learning when to transliterate. In *Proceedings of ACL-08: HLT*. pages 389–397.

Ann Irvine and Chris Callison-Burch. 2013. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. pages 262–270.

Shonosuke Ishiwatari, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2016. Instant translation model adaptation by translating unseen words in continuous vector space. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016). Konya, Turkey*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the Tenth Machine Translation Summit*. AAMT, AAMT, Phuket, Thailand, pages 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computation Linguistics (ACL), Demonstration Session*. pages 177–180.

Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Baltimore, Maryland, pages 224–229.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015a. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. pages 151–159.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015b. Addressing the rare word problem in 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 (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 11–19.

Pranava Swaroop Madhyastha, Xavier Carreras, and Ariadna Quattoni. 2014a. Learning task-specific bilexical embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, pages 161–171.

Pranava Swaroop Madhyastha, Xavier Carreras, and Ariadna Quattoni. 2014b. Tailoring word embeddings for bilexical predictions: An experimental comparison. *International Conference on Learning Representations 2015, Workshop Track* .

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 381–390.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR* abs/1309.4168. http://arxiv.org/abs/1309.4168.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics*. Sapporo, Japan, pages 160–167.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL 2002, pages 295–302.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Association of Computational Linguistics*. pages 311–318.

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*. Association for Computational Linguistics, pages 519–526.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. pages 1715–1725.

Yoram Singer and John C Duchi. 2009. Efficient learning using forward-backward splitting. In *Advances in Neural Information Processing Systems*. pages 495–503.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora Using Document Level Alignment. In *In Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Cha pter of the Association for Computational Linguistics (NAACL-HLT 2010)*. pages 403–411.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas (AMTA 2006)*. Cambridge, Massachusetts, USA, pages 223–231.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the Seventh International Conference of Spoken Language Processing (ICSLP2002)*. Denver, Colorado, USA, pages 901–904.

Shyam Upadhyay, 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, ACL 2016, August 7-12, 2016, Berlin, Germany*. pages 1661–1670.

Ivan Vulic 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 of the Asian Federation of Natural Language Processing, ACL 2015*. pages 719–725.

Haiyang Wu, Daxiang Dong, Xiaoguang Hu, Dianhai Yu, Wei He, Hua Wu, Haifeng Wang, and Ting Liu. 2014. Improve Statistical Machine Translation with Context-Sensitive Bilingual Semantic Embedding Model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 142–146.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually-constrained phrase embeddings for machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Baltimore, Maryland, pages 111–121.

Jiajun Zhang, Feifei Zhai, and Chengqing Zong. 2012. Handling unknown words in statistical machine translation from a new perspective. In *Natural Language Processing and Chinese Computing*, Springer, pages 176–187.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning Translation Models from Monolingual Continuous Representations. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2015*. pages 1527–1536.

Will Y. Zou, Richard Socher, Daniel M. Cer, and Christopher D. Manning. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*. pages 1393–1398.