Lexicon Stratification for Translating Out-of-Vocabulary Words

Yulia Tsvetkov
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
ytsvetko@cs.cmu.edu

Chris Dyer
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
cdyer@cs.cmu.edu

Abstract

A language lexicon can be divided into four main strata, depending on origin of words: core vocabulary words, fully- and partially-assimilated foreign words, and unassimilated foreign words (or transliterations). This paper focuses on translation of fully- and partially-assimilated foreign words, called "borrowed words". Borrowed words (or loanwords) are content words found in nearly all languages, occupying up to 70% of the vocabulary. We use models of lexical borrowing in machine translation as a pivoting mechanism to obtain translations of out-of-vocabulary loanwords in a low-resource language. Our framework obtains substantial improvements (up to 1.6 BLEU) over standard baselines.

1 Introduction

Out-of-vocabulary (OOV) words are a ubiquitous and difficult problem in statistical machine translation (SMT). When a translation system encounters an OOV—a word that was not observed in the training data, and the trained system thus lacks its translation variants—it usually outputs the word just as it is in the source language, producing erroneous and disfluent translations.

All SMT systems, even when trained on billion-sentence-size parallel corpora, are prone to OOVs. These are often named entities and neologisms. However, OOV problem is much more serious in low-resource scenarios: there, OOVs are primarily not lexicon-peripheral items such as names and specialized/technical terms, but regular content words.

Procuring translations for OOVs has been a subject of active research for decades. Translation of named entities is usually generated using transliteration techniques (Al-Onaizan and Knight, 2002; Hermjakob et al., 2008; Habash, 2008). Extracting a translation lexicon for recovering OOV content words and phrases is done by mining bi-lingual and monolingual resources (Rapp, 1995; Callison-Burch et al., 2006; Haghighi et al., 2008; Marton et al., 2009; Razmara et al., 2013; Saluja et al., 2014; Zhao et al., 2015). In addition, OOV content words can be recovered by exploiting cognates, by transliterating and then pivoting via a closely-related resource-richer language, when such a language exists (Hajič et al., 2000; Mann and Yarowsky, 2001; Kondrak et al., 2003; De Gispert and Marino, 2006; Durrani et al., 2010; Wang et al., 2012; Nakov and Ng, 2012; Dholakia and Sarkar, 2014). Our work is similar in spirit to the latter line of research, but we show how to curate translations for OOV content words by pivoting via an unrelated, often typologically distant resource-rich languages. To achieve this goal, we replace transliteration by a new technique that captures more complex morpho-phonological transformations of historically-related words.

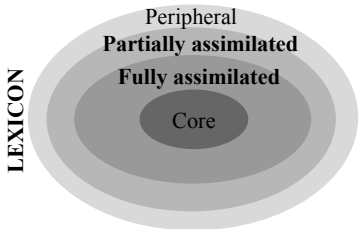

Figure 1: A language lexicon can be divided into four main strata, depending on origin of words. This work focuses on fully- and partially-assimilated foreign words, called borrowed words. Borrowed words (or loanwords) are content words found in all languages, occupying up to 70% of the vocabulary.

Our method is inspired by prior research in constraint-based phonology, advocating "lexicon stratification," i.e., splitting the language lexicon into separate strata, depending on origin of words and degree of their assimilation in the language (Itô and Mester, 1995). As shown in figure 1, there are four main strata: core vocabulary, foreign words that are fully assimilated, partially-assimilated for-

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 125–131,
Beijing, China, July 26-31, 2015.

eign words, and named entities which belong to the peripheral stratum. Our work focuses on the fully- and partially-assimilated foreign words, i.e., words that historically were *borrowed* from another language. Borrowing is the pervasive linguistic phenomenon of transferring and adapting linguistic constructions (lexical, phonological, morphological, and syntactic) from a "donor" language into a "recipient" language (Thomason and Kaufman, 2001). In this work, we advocate a pivoting mechanism exploiting lexical borrowing to bridge between resource-rich and resource-poor languages.

Our method (§2) employs a model of lexical borrowing to obtain cross-lingual links from loanwords in a low-resource language to their donors in a resource-rich language (§2.1). The donor language is used as pivot to obtain translations via triangulation of OOV loanwords (§2.2). We conduct experiments with two resource-poor setups: Swahili–English, pivoting via Arabic, and Romanian–English,pivoting via French (§3). We provide a systematic quantitative analysis of contribution of integrated OOV translations, relative to baselines and upper bounds, and on corpora of varying sizes (§4). The proposed approach yields substantial improvement (up to $+1.6$ BLEU) in Swahili–Arabic–English translation, and a small but statistically significant improvement ($+0.2$ BLEU) in Romanian–French–English.

2 Methodology

Our high-level solution is depicted in figure 2. Given an OOV word in resource-poor SMT, we plug it into a borrowing system (§2.1) that identifies the list of plausible donor words in the donor language. Then, using the resource-rich SMT, we translate the donor words to the same target language as in the resource-poor SMT (here, English). Finally, we integrate translation candidates in the resource-poor system (§2.2).

2.1 Models of Lexical Borrowing

Borrowed words (also called loanwords) are found in nearly all languages, and routinely account for 10–70% of the vocabulary (Haspelmath and Tadmor, 2009). Borrowing occurs across genetically and typologically unrelated languages, for example, about 40% of Swahili's vocabulary is borrowed from Arabic (Johnson, 1939). Importantly, since resource-rich languages are (historically) geopolitically important languages, borrowed words often

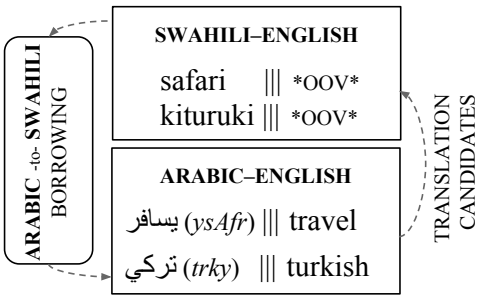

Figure 2: To improve a resource-poor Swahili–English SMT system, we extract translation candidates for OOV Swahili words borrowed from Arabic using the Swahili-to-Arabic borrowing system and Arabic–English resource-rich SMT.

bridge between resource-rich and resource-limited languages; we use this observation in our work.

Transliteration and cognate discovery models perform poorly in the task of loanword generation/identification (Tsvetkov et al., 2015). The main reason is that the recipient language, in which borrowed words are fully or partially assimilated, may have very different morpho-phonological properties from the donor language (e.g., 'orange' and 'sugar' are not perceived as foreign by native speakers, but these are English words borrowed from Arabic نارنج (*nArnj*)[1] and السكر (*Alskr*), respectively). Therefore, morpho-phonological loanword adaptation is more complex than is typically captured by transliteration or cognate models.

We employ a discriminative cross-lingual model of lexical borrowing to identify plausible donors given a loanword (Tsvetkov et al., 2015). The model is implemented in a cascade of finite-state transducers that first maps orthographic word forms in two languages into a common space of their phonetic representation (using IPA—the International Phonetic Alphabet), and then performs morphological and phonological updates to the input word in one language to identify its (donor/loan) counterpart in another language. Transduction operations include stripping donor language prefixes and suffixes, appending recipient affixes, insertion, deletion, and substitution of consonants and vowels. The output of the model, given an input loanword, is a n-best list of donor candidates, ranked by linguistic constraints of the donor and recipient languages.[2]

[1] We use Buckwalter notation to write Arabic glosses.

[2] In this work, we give as input into the borrowing system all OOV words, although, clearly, not all OOVs are loanwords, and not all loanword OOVs are borrowed from the donor language. However, an important property of the borrowing model is that its operations are not general, but specific to

2.2 Pivoting via Borrowing

We now discuss integrating translation candidates acquired via borrowing plus resource-rich translation. For each OOV, the borrowing system produces the n-best list of plausible donors; for each donor we then extract the k-best list of its translations.[3] Then, we pair the OOV with the resulting $n \times k$ translation candidates. The translation candidates are noisy: some of the generated donors may be erroneous, the errors are then propagated in translation. To allow the low-resource system to leverage good translations that are missing in the default phrase inventory, while being stable to noisy translation hypotheses, we integrate the acquired translation candidates as *synthetic phrases* (Tsvetkov et al., 2013; Chahuneau et al., 2013). Synthetic phrases is a strategy of integrating translated phrases directly in the MT translation model, rather than via pre- or post-processing MT inputs and outputs. Synthetic phrases are phrasal translations that are not directly extractable from the training data, generated by auxiliary translation and postediting processes (for example, extracted from a borrowing model). An important advantage of synthetic phrases is that they are recall-oriented, allowing the system to leverage good translations that are missing in the default phrase inventory, while being stable to noisy translation hypotheses.

To let the translation model learn whether to trust these phrases, the translation options obtained from the borrowing model are augmented with a boolean translation feature indicating that the phrase was generated externally. Additional features annotating the integrated OOV translations correspond to properties of the donor–loan words' relation; their goal is to provide an indication of plausibility of the pair (to mark possible errors in the outputs of the borrowing system).

We employ two types of features: phonetic and semantic. Since borrowing is primarily a phonological phenomenon, phonetic features will provide an indication of how typical (or atypical) pronunciation of the word in a language; loanwords are expected to be less typical than core vocabulary words. The goal of semantic features is to measure semantic similarity between donor and loan words: erroneous candidates and borrowed words that changed meaning over time are expected to have different meaning from the OOV.

Phonetic features. To compute phonetic features we first train a (5-gram) language model (LM) of IPA pronunciations of the donor/recipient language vocabulary (phoneLM). Then, we re-score pronunciations of the donor and loanword candidates using the LMs.[4] We hypothesize that in donor–loanword pairs the donor phoneLM score is higher but the loanword score is lower (i.e., the loanword phonology is atypical in the recipient language). We capture this intuition in three features: f_1=$P_{phoneLM}(donor)$, f_2=$P_{phoneLM}(loanword)$, and the harmonic mean between the two scores f_3=$\frac{2f_1f_2}{f_1+f_2}$.

Semantic features. We compute a semantic similarity feature between the candidate donor and the OOV loanword as follows. We first train, using large monolingual corpora, 100-dimensional word vector representations for donor and recipient language vocabularies.[5] Then, we employ canonical correlation analysis (CCA) with small donor–loanword dictionaries (training sets in the borrowing models) to project the word embeddings into 50-dimensional vectors with maximized correlation between their dimensions. The semantic feature annotating the synthetic translation candidates is cosine distance between the resulting donor and loanword vectors. We use the `word2vec` tool (Mikolov et al., 2013) to train monolingual vectors,[6] and the CCA-based tool (Faruqui and Dyer, 2014) for projecting word vectors.[7]

3 Experimental Setup

Datasets and software. The Swahili–English parallel corpus was crawled from the Global Voices project website[8]. To simulate resource-poor scenario for the Romanian–English language pair, we sample a parallel corpus of same size from the transcribed TED talks (Cettolo et al., 2012). To evalu-

the language-pair and reduced only to a small set of plausible changes that the donor word can undergo in the process of assimilation in the recipient language. Thus, the borrowing system only *minimally* overgenerates the set of output candidates given an input. If the borrowing system encounters an input word that was not borrowed from the target donor language, it usually (but not always) produces an empty output.

[3] We set n and k to 5, we did not experiment with other values.

[4] For Arabic and French we use the `GlobalPhone` pronunciation dictionaries (Schultz et al., 2013) (we manually convert them to IPA). For Swahili and Romanian we automatically construct pronunciation dictionaries using the Omniglot grapheme-to-IPA conversion rules at `www.omniglot.com`.

[5] We assume that while parallel data is limited in the recipient language, monolingual data is available.

[6] `code.google.com/p/word2vec`

[7] `github.com/mfaruqui/eacl14-cca`

[8] `sw.globalvoicesonline.org`

ate translation improvement on corpora of different sizes we conduct experiments with sub-sampled 4K, 8K, and 14K parallel sentences from the training corpora (the smaller the training corpus, the more OOVs it has). Corpora sizes along with statistics of source-side OOV tokens and types are given in tables 1 and 2. Statistics of the held-out dev and test sets used in all translation experiments are given in table 3.

	SW–EN		RO–EN	
	dev	test	dev	test
Sentences	1,552	1,732	2,687	2,265
Tokens	33,446	35,057	24,754	19,659
Types	7,008	7,180	5,141	4,328

Table 3: Dev and test corpora sizes.

In all the MT experiments, we use the `cdec`[9] toolkit (Dyer et al., 2010), and optimize parameters with MERT (Och, 2003). English 4-gram language models with Kneser-Ney smoothing (Kneser and Ney, 1995) are trained using KenLM (Heafield, 2011) on the target side of the parallel training corpora and on the Gigaword corpus (Parker et al., 2009). Results are reported using case-insensitive BLEU with a single reference (Papineni et al., 2002). We train three systems for each MT setup; reported BLEU scores are averaged over systems.

Upper bounds. The goal of our experiments is not only to evaluate the contribution of the OOV dictionaries that we extract when pivoting via borrowing, but also to understand the potential contribution of the lexicon stratification. What is the overall improvement that can be achieved if we correctly translate all OOVs that were borrowed from another language? What is the overall improvement that can be achieved if we correctly translate all OOVs? We answer this question by defining "upper bound" experiments. In the upper bound experiment we word-align all available parallel corpora, including dev and test sets, and extract from the alignments oracle translations of OOV words. Then, we append the extracted OOV dictionaries to the training corpora and re-train SMT setups without OOVs. Translation scores of the resulting system provide an upper bound of an improvement from correctly translating all OOVs. When we append oracle translations of the subset of OOV dictionaries, in particular translations of all OOVs for which the output of the borrowing system is not empty, we obtain an upper bound that can be achieved using our method (if the borrowing system provided perfect outputs). Understanding the upper bounds is relevant not only for our experiments, but for any experiments that involve augmenting translation dictionaries; however, we are not aware of prior work providing similar analysis of upper bounds, and we recommend this as a calibrating procedure for future work on OOV mitigation strategies.

Borrowing-augmented setups. As described in §2.2, we integrate translations of OOV loanwords in the translation model. Due to data sparsity, we conjecture that non-OOVs that occur only few times in the training corpus can also lack appropriate translation candidates, i.e., these are target-language OOVs. We therefore run the borrowing system on OOVs and non-OOV words that occur less than 3 times in the training corpus. We list in table 4 sizes of translated lexicons that we integrate in translation tables.

	4K	8K	14K
Loan OOVs in SW–EN	5,050	4,219	3,577
Loan OOVs in RO–EN	347	271	216

Table 4: Sizes of translated lexicons extracted using pivoting via borrowing and integrated in translation models.

Transliteration-augmented setups. In addition to the standard baselines, we evaluate transliteration-augmented setups, where we replace the borrowing model by a transliteration model (Ammar et al., 2012). The model is a linear-chain CRF where we label each source character with a sequence of target characters. The features are label unigrams and bigrams, separately or conjoined with a moving window of source characters. We employ the Swahili–Arabic and Romanian–French transliteration systems that were used as baselines in (Tsvetkov et al., 2015). As in the borrowing system, transliteration outputs are filtered to contain only target language lexicons. We list in table 5 sizes of obtained translated lexicons.

	4K	8K	14K
Translit. OOVs in SW–EN	49	32	22
Translit. OOVs in RO–EN	906	714	578

Table 5: Sizes of translated lexicons extracted using pivoting via transliteration and integrated in translation models.

[9] `www.cdec-decoder.org`

	4K	8K	14K
Tokens	84,764	170,493	300,648
Types	14,554	23,134	33,288
OOV tokens	4,465 (12.7%)	3,509 (10.0%)	2,965 (8.4%)
OOV types	3,610 (50.3%)	2,950 (41.1%)	2,523 (35.1%)

Table 1: Statistics of the Swahili–English corpora and source-side OOV for 4K, 8K, 14K parallel training sentences.

	4K	8K	14K
Tokens	35,978	71,584	121,718
Types	7,210	11,144	15,112
OOV tokens	3,268 (16.6%)	2,585 (13.1%)	2,177 (11.1%)
OOV types	2,382 (55.0%)	1,922 (44.4%)	1,649 (38.1%)

Table 2: Statistics of the Romanian–English corpora and source-side OOV for 4K, 8K, 14K parallel training sentences.

4 Results

Translation results are shown in tables 6 and 7. We evaluate separately the contribution of the integrated OOV translations, and the same translations annotated with phonetic and semantic features. We also provide upper bound scores for integrated loanword dictionaries as well as for recovering all OOVs.

	4K	8K	14K
Baseline	13.2	15.1	17.1
+ Translit. OOVs	13.4	15.3	17.2
+ Loan OOVs	14.3	15.7	18.2
+ Features	**14.8**	**16.4**	**18.4**
Upper bound loan	18.9	19.1	20.7
Upper bound all OOVs	19.2	20.4	21.1

Table 6: Swahili–English MT experiments.

	4K	8K	14K
Baseline	15.8	18.5	20.7
+ Translit. OOVs	15.8	**18.7**	**20.8**
+ Loan OOVs	**16.0**	**18.7**	20.7
+ Features	**16.0**	18.6	20.6
Upper bound loan	16.6	19.4	20.9
Upper bound all OOVs	28.0	28.8	30.4

Table 7: Romanian–English MT experiments.

Swahili–English MT performance is improved by up to $+1.6$ BLEU when we augment it with translated OOV loanwords leveraged from the Arabic–Swahili borrowing and then Arabic–English MT. The contribution of the borrowing dictionaries is $+0.6$–1.1 BLEU, and phonetic and semantic features contribute additional half BLEU. More importantly, upper bound results show that the system can be improved more substantially with better dictionaries of OOV loanwords. This result confirms that OOV borrowed words is an important type of OOVs, and with proper modeling it has the potential to improve translation by a large margin. Romanian–English systems obtain only small (but significant for 4K and 8K, $p < .01$) improvement. However, this is expected as the rate of borrowing from French into Romanian is smaller, and, as the result, the integrated loanword dictionaries are small. Transliteration baseline, conversely, is more effective in Romanian–French language pair, as two languages are related typologically, and have common cognates in addition to loanwords. Still, even with these dictionaries the translations with pivoting via borrowing/transliteration improve, and even almost approach the upper bounds results.

5 Conclusion

This paper focuses on fully- and partially-assimilated foreign words in the source lexicon—borrowed words—and a method for obtaining their translations. Our results substantially improve translation and confirm that OOV loanwords are important and merit further investigation. In addition, we propose a simple technique to calculate an upper bound of improvements that can be obtained from integrating OOV translations in SMT.

Acknowledgments

This work was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-10-1-0533. Computational resources were provided by Google Cloud Computing grant. We are grateful to Waleed Ammar for his help with transliteration, and to the anonymous reviewers.

References

Yaser Al-Onaizan and Kevin Knight. 2002. Machine transliteration of names in Arabic text. In *Proc. the ACL workshop on Computational Approaches to Semitic Languages*, pages 1–13.

Waleed Ammar, Chris Dyer, and Noah A. Smith. 2012. Transliteration by sequence labeling with lattice encodings and reranking. In *Proc. NEWS workshop at ACL*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proc. NAACL*, pages 17–24.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web inventory of transcribed and translated talks. In *Proc. EAMT*, pages 261–268.

Victor Chahuneau, Eva Schlinger, Noah A Smith, and Chris Dyer. 2013. Translating into morphologically rich languages with synthetic phrases. In *Proc. EMNLP*, pages 1677–1687.

Adrià De Gispert and Jose B Marino. 2006. Catalan-English statistical machine translation without parallel corpus: bridging through Spanish. In *Proc. LREC*, pages 65–68.

Rohit Dholakia and Anoop Sarkar. 2014. Pivot-based triangulation for low-resource languages. In *Proc. AMTA*, pages 315–328.

Nadir Durrani, Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2010. Hindi-to-Urdu machine translation through transliteration. In *Proc. ACL*, pages 465–474.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proc. ACL System Demonstrations*, pages 7–12.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proc. EACL*, pages 462–471.

Nizar Habash. 2008. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In *Proc. ACL*, pages 57–60.

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

Jan Hajič, Jan Hric, and Vladislav Kuboň. 2000. Machine translation of very close languages. In *Proc. ANLP*, pages 7–12.

Martin Haspelmath and Uri Tadmor, editors. 2009. *Loanwords in the World's Languages: A Comparative Handbook*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proc. WMT*, pages 187–197.

Ulf Hermjakob, Kevin Knight, and Hal Daumé III. 2008. Name translation in statistical machine translation-learning when to transliterate. In *Proc. ACL*, pages 389–397.

Junko Itô and Armin Mester. 1995. The core-periphery structure of the lexicon and constraints on reranking. *Papers in Optimality Theory*, 18:181–209.

Frederick Johnson. 1939. *Standard Swahili-English dictionary*. Oxford University Press.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proc. ICASSP*, volume 1, pages 181–184.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Proc. HLT-NAACL*, pages 46–48.

Gideon S Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proc. HLT-NAACL*, pages 1–8.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proc. EMNLP*, pages 381–390.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, pages 3111–3119.

Preslav Nakov and Hwee Tou Ng. 2012. Improving statistical machine translation for a resource-poor language using related resource-rich languages. *Journal of Artificial Intelligence Research*, pages 179–222.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English Gigaword fourth edition.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proc. ACL*, pages 320–322.

Majid Razmara, Maryam Siahbani, Reza Haffari, and Anoop Sarkar. 2013. Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proc. ACL*, pages 1105–1115.

Avneesh Saluja, Hany Hassan, Kristina Toutanova, and Chris Quirk. 2014. Graph-based semi-supervised learning of translation models from monolingual data. In *Proc. ACL*, pages 676–686.

Tanja Schultz, Ngoc Thang Vu, and Tim Schlippe. 2013. GlobalPhone: A multilingual text & speech database in 20 languages. In *Proc. ICASSP*, pages 8126–8130.

Sarah Grey Thomason and Terrence Kaufman. 2001. *Language contact*. Edinburgh University Press Edinburgh.

Yulia Tsvetkov, Chris Dyer, Lori Levin, and Archna Bhatia. 2013. Generating English determiners in phrase-based translation with synthetic translation options. In *Proc. WMT*, pages 271–280.

Yulia Tsvetkov, Waleed Ammar, and Chris Dyer. 2015. Constraint-based models of lexical borrowing. In *Proc. NAACL*, pages 598–608.

Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2012. Source language adaptation for resource-poor machine translation. In *Proc. EMNLP*, pages 286–296.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning translation models from monolingual continuous representations. In *Proc. NAACL*.