Word Embeddings for Code-Mixed Language Processing

Adithya Pratapa, Monojit Choudhury, Sunayana Sitaram
Microsoft Research, India
{t-pradi,monojitc,sunayana.sitaram}@microsoft.com

Abstract

We compare three existing bilingual word embedding approaches, and a novel approach of training skip-grams on synthetic code-mixed text generated through linguistic models of code-mixing, on two tasks - sentiment analysis and POS tagging for code-mixed text. Our results show that while CVM and CCA based embeddings perform as well as the proposed embedding technique on semantic and syntactic tasks respectively, the proposed approach provides the best performance for both tasks overall. Thus, this study demonstrates that existing bilingual embedding techniques are not ideal for code-mixed text processing and there is a need for learning multilingual word embedding from the code-mixed text.

1 Introduction

Word embeddings are useful for a variety of NLP tasks, as they allow to generalize the system on much larger corpora than the annotated dataset for the task. In recent times, there has been some interest in bilingual word embeddings, where words from two languages are embedded into the same space. The primary advantage of bilingual embeddings is in solving tasks involving reasoning across two languages, such as Machine Translation (Zou et al., 2013; Gouws et al., 2015; Vulić and Moens, 2016) and cross-lingual IR (Vulić and Moens, 2015), as well as allowing transfer of models learnt on a resource-rich language on to a resource poor language (Adams et al., 2017; Fang and Cohn, 2017). One of the potential, yet unexplored, applications of bilingual word embeddings is in the processing of code-mixed language.

Code-mixing (CM) refers to fluid alternation between two or more languages in a single conversation/sentence (Myers-Scotton, 1993). CM is a common phenomenon observed in almost all multilingual societies (Parshad et al., 2016; Rijhwani et al., 2017). Consequently, in recent times, processing of CM text and speech has been receiving a growing amount of interest and attention from the NLP community (Solorio and Liu, 2008; Li and Fung, 2014; Solorio et al., 2014; Sharma et al., 2016; Rudra et al., 2016). Since CM text draws words and linguistic structures from multiple languages, use of bilingual word embeddings for processing of such text could not only be useful, but also necessary. On the other hand, while there is some work that uses embeddings for CM text (Prabhu et al., 2016) (at sub-word level), we do not know of any study that systematically explores the usefulness of bilingual word embedding techniques in CM text processing.

Further, we argue that since all the standard bilingual word embedding techniques are designed to work on or across monolingual texts rather than on a mixture of the two languages, these techniques may not be ideal for learning embeddings for CM tasks. There are emergent syntactic structures and cross-lingual semantic associations in CM text, that do not exist in the individual monolingual corpora (Sec 3). Hence, ideally, word embeddings for CM tasks should be trained on real CM data.

In this paper, we compare three popular bilingual word embedding techniques (Sec 2): Bilingual correlation based embeddings (BiCCA) (Faruqui and Dyer, 2014), Bilingual compositional model (BiCVM) (Hermann and Blunsom, 2014) and Bilingual Skip-gram (BiSkip) (Luong et al., 2015) on two tasks for CM text - sentiment analysis, a semantic task, and POS tagging, a syntactic task. On the same tasks, we also compare word embeddings learnt from synthetic CM data (generated using linguistic models as proposed in a recent work (Pratapa et al., 2018)) (Sec 3). Note that Wick et al. (2016) use artificial code mixed data to learn

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3067–3072
Brussels, Belgium, October 31 - November 4, 2018.

multilingual embeddings for cross-lingual tasks, but their aim is to generate bilingual embeddings for monolingual or cross-lingual tasks.

Our study shows that even though in certain NLP tasks specific embeddings might perform well, in general bilingual embedding techniques like BiCCA, BiCVM and BiSkip are not ideal for processing CM language. Embeddings learnt from CM data, even if artificially generated, performs consistently better across tasks. Our initial results are promising and provide several interesting directions for further exploration.

2 Bilingual Embeddings

In the past few years, there has been a growing interest in learning bilingual embeddings (Upadhyay et al., 2016; Ruder et al., 2017) with a focus on cross-lingual transfer, which helps in building NLP models for low-resource languages. Upadhyay et al. (2016) provide an empirical comparison of four cross-lingual word embedding models varying in terms of the amount of supervision. Ruder et al. (2017) establishes the similarities among numerous cross-lingual word embedding models and shows that many models optimize for similar objectives. Along similar lines as Upadhyay et al. (2016), in this work, we chose the following three representative bilingual word embedding models for CM tasks. Training data is described later in Section 4.

2.1 Bilingual Correlation Based Embeddings (BiCCA)

Faruqui and Dyer (2014) proposed CCA based bilingual embeddings, where bilingual evidence is incorporated into word representations by performing canonical correlation analysis (CCA) on monolingual embeddings using a bilingual dictionary. The monolingual embedding matrices W_{L_1} and W_{L_2} can be of different dimensions and words with their translations from the dictionary are used to obtain matrices W'_{L_1} and W'_{L_2}. Using CCA, the individual projection matrices are computed, which are then utilized to project the embeddings into the same embedding space.

BiCCA embeddings were shown to perform well on syntactic tasks like cross lingual dependency parsing, but performed relatively poorly on cross lingual semantic tasks (Upadhyay et al., 2016). We learn the monolingual embeddings by training a skip-gram model (Mikolov et al., 2013) for 5 iterations with a window size of 5 and 10 negative samples. We built a bilingual dictionary of approximately 38k pairs by using word alignments. The dictionary contains word pairs (w_1, w_2), such that w_2 is aligned to w_1 the highest number of times and vice-versa. We use the `crosslingual-cca`[1] toolkit. To remain consistent with other embedding models, we choose the top (k=) 200 correlated dimensions.

2.2 Bilingual Compositional Model (BiCVM)

This approach, proposed by Hermann and Blunsom (2014), is based on the assumption that parallel sentences from different languages have equivalent meanings and thus should have similar sentence representations. Along with monolingual regularizers, the model optimizes for the aligned sentences to be closer to each other than randomly chosen negative samples, using a noise-contrastive update.

BiCVM is found to perform well on monolingual word similarity (SimLex-999, Upadhyay et al. (2016)) and is comparable to BiSkip on semantic tasks like cross-lingual document classification. We use the `bicvm`[2] toolkit to generate embeddings using the parallel English-Spanish data. We train an additive model for 100 iterations, with a hinge loss margin of 200, noise parameter of 10 and batch size of 50.

2.3 Bilingual Skip-gram Model (BiSkip)

The Skip-gram model proposed by Mikolov et al. (2013) has been adapted to the bilingual setting in Luong et al. (2015), where the model learns to predict word contexts cross-lingually. Along with the monolingual skip-gram with negative sampling objectives, BiSkip includes two more objectives L_{12} and L_{21} when predicting cross-lingually from L_1 to L_2 and vice-versa.

It has the best performance compared to the other embedding techniques on semantic tasks like cross-lingual dictionary induction and cross-lingual document classification. We use parallel sentences and word level alignments to train the `biskip`[3] model. We train the model using the toolkit for 5 iterations with 10 negative samples, high frequency threshold of 0.001, window size of 5 and cross-lingual weight as 1.

[1] https://github.com/mfaruqui/crosslingual-cca

[2] https://github.com/karlmoritz/bicvm/

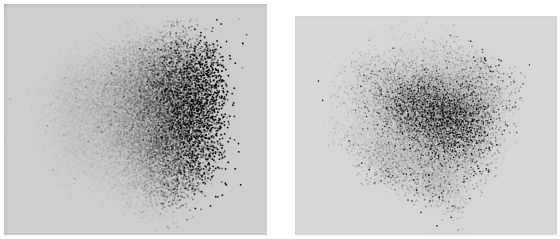
[3] https://github.com/lmthang/bivec

3 Synthetic CM data (gCM) based Embeddings

The aforementioned embedding techniques, owing to their ability to project the words of the two languages into a single space, are expected to be helpful in processing of CM text. However, we believe that CM text is distinct in its syntactic, semantic and statistical properties from the corresponding monolingual texts. For example, there has been a lot of work on understanding and establishing the grammatical constraints of CM (Joshi, 1985; Poplack, 1980; DiSciullo et al., 1986) and these syntactic constraints might introduce interesting collocations in the word space, such as that between English verbs and the Hindi verb "kar" (to do) in English-Hindi CM (Kachru, 1978).

In a recent work (Pratapa et al., 2018), we presented a methodology to generate linguistic theory based synthetic CM data (gCM) and showed its effectiveness in CM language modeling. Synthetic CM data was generated by employing Equivalence Constraint (EC) theory (Poplack, 1980; Sankoff, 1998). The generation model builds a lattice of grammatically valid CM sentences by imposing the EC theory constraints while utilizing monolingual parallel data, word-level alignments and the parse of English sentences. The Spanish parse is formed by projecting the English parse onto the Spanish sentence. For each monolingual pair of input sentences, a large number (generally exponential in terms of average input sentence length) of CM sentences were generated. We observed a non-uniform scaling of low and high frequency words and proposed two sampling techniques to overcome this frequency bias.

In the current work, we use this generation model (as proposed in (Pratapa et al., 2018)) to create training data for learning CM word embeddings. In fact, we also noticed the frequency bias in the word embedding space (Figure 1a). We adapt the two sampling strategies from the original paper, 1. Random sampling (χ-***gCM***), for every monolingual pair, sample random k CM sentences and, 2. SPF-based sampling (ρ-***gCM***), sample k CM sentences for each monolingual pair such that they have SPF (switch point fraction, i.e fraction of word boundaries where the language changes) distribution similar to real CM data (Pratapa et al., 2018). We were able to alleviate the frequency bias using the SPF-based sampling (Figure 1b).



(a) (b)

Figure 1: (a) PCA projections of skip-gram (with negative sampling) embeddings trained on entire gCM corpus, (b) ρ-sampled gCM. The color gradient (light $\rightarrow$ dark) is based on the frequency of word (low $\rightarrow$ high) in gCM corpus.

4 Evaluation

Data: Though CM is a common phenomenon, there is a scarcity of real code-mixed data in textual form. Hence, we learn CM embeddings from the English-Spanish parallel corpora (for BiCCA, BiCVM and BiSkip models) and the synthetic CM data obtained from Pratapa et al. (2018). This constitutes to approximately 4.5M parallel sentences and also bilingual supervision in the form of word and sentence alignments. The synthetic data creation procedure is described in the original paper. Both the sampling techniques, with k=2 result in a gCM corpus of approx. 8M sentences. We also combine monolingual data with these gCM corpus resulting in approx. 17M sentences. We train a skip-gram model for 10 iterations, with a window size of 5 and 5 negative samples, resulting in χ-***gCM-Skip*** and ρ-***gCM-Skip*** embeddings.

To quantitatively compare the embedding models, we chose two CM tasks, one semantic (Sentiment Analysis) and one syntactic task (POS tagging). Our choice of tasks is primarily motivated by the availability of annotated CM data. There has been prior work on CM sentiment identification (Vilares and Alonso, 2016; Joshi et al., 2016; Rudra et al., 2016; Prabhu et al., 2016) and POS tagging (Solorio and Liu, 2008; Vyas et al., 2014; AlGhamdi et al., 2016; Ghosh et al., 2016). But we are not aware of any work that utilizes pre-trained bilingual embeddings for these tasks.

4.1 Sentiment Analysis

Vilares and Alonso (2016) provide 2103 sentiment annotated CM tweets. The data contains 650 positive, 529 negative and 924 neutral tweets and we split the data in 8:1:1 ratio (train:validation:test)

Embedding	Sentiment			POS	
	CM Overall	SemEval 2014	TASS 2016	CM Overall	at SP
None	54.4 (1.3)	64.5 (0.6)	61.4 (1.0)	84.5 (0.3)	74.0 (0.7)
BiCCA	57.6 (3.0)	64.6 (1.0)	59.5 (1.8)	**84.7** (0.8)	**75.0** (1.8)
BiCVM	**64.3** (1.3)	66.8 (1.0))	61.9 (1.0)	82.0 (0.5)	70.6 (1.7)
BiSkip	61.5 (1.7)	66.6 (0.9)	**63.9** (1.2)	84.4 (0.7)	73.8 (0.9)
χ-gCM-Skip	62.0 (1.9)	**67.4** (1.3)	63.2 (1.5)	**84.8** (0.6)	74.0 (0.6)
ρ-gCM-Skip	**64.6** (2.0)	**67.7** (1.4)	**63.8** (2.2)	**84.9** (0.7)	**75.3** (1.7)

Table 1: The performance of different pre-trained embeddings on Sentiment (F1 score) and POS tasks (Accuracy). The reported values are mean and deviation (in parentheses) values computed over multiple runs.

while ensuring the sentiment distribution remains the same[4].

Model: We train a LSTM based sequence classifier with single hidden layer (dim=50) and dropout of 0.5. We use ADAM optimizer with learning rate of 0.001 and momentum parameter of 0.9. We train the model for a maximum of 10 epochs with a mini-batch size of 100. Our model is built using the Microsoft CNTK framework. We varied the CM sample size (k) (described in 3) over {1,2,5,10,20} and found the best performance with k=2. To check the robustness of CM embeddings, we also evaluate them on monolingual English[5] (SemEval 2014 Subtask B, (7177, 1199, 2865)) and Spanish[6] (TASS 2016, (6000, 1220, 1000)) tasks. The numbers in the parenthesis indicate the number of train, validation and test instances.

Results: Table 1 shows the results on Sentiment task. While all embedding models significantly improve over the baseline ('None'), ρ-gCM-Skip and BiCVM perform the best. Instead of linguistically motivated data, we tried with CM data created by random juxtaposition of monolingual fragments and it gave a F1 score of 56.0% and the minimal gain is possibly only because of the monolingual fragments in the embedding training data.

4.2 Part of Speech (POS) Tagging

Of the two corpora utilized in AlGhamdi et al. (2016), we chose Bangor Miami Corpus[7] over Spanglish Corpus (Solorio and Liu, 2008) owing to the larger size of the former corpus. Bangor Miami corpus consists of conversations of Spanish speakers in Florida but also fluent in English. In contrast to AlGhamdi et al. (2016), we only consider the code mixed utterances with significant ($\geq$30%) fraction of English and Spanish. This accounts to 982 sentences and 7705 tokens, split in 8:1:1 ratio.

Model: We use a bidirectional LSTM model with CRF as the output layer (Rei and Yannakoudakis, 2016)[8], with hidden layer dimension of 50 and dropout of 0.5. The model trains for a maximum of 20 epochs and terminates if there is no improvement in validation accuracy for 5 consecutive epochs. We use ADADELTA optimizer with a learning rate of 1.0.

Results: ρ-gCM-Skip, χ-gCM-Skip perform the best on the POS task (see Table. 1). Unlike the sentiment task, BiCCA performs close to the best while BiCVM does the worst. We also report accuracies at switch points (SP), which we believe is challenging for a CM POS tagger. As expected these accuracies are lower, but ρ-gCM-Skip and BiCCA still perform the best.

5 Conclusion

As we expected, pretrained bilingual word embeddings help improve syntactic and semantic CM processing tasks. Similar to (Upadhyay et al., 2016), we also note that BiCVM operating at sentence level performs better only on semantic tasks, while BiCCA does well only on syntactic tasks due to their usage of word alignments. Though ρ-gCM-Skip embeddings learnt from synthetic data

[4]These statistics are for the re-crawled tweets; some of the original tweets are no longer available.

[5]alt.qcri.org/semeval2014/task9/

[6]www.sepln.org/workshops/tass/2016/

[7]bangortalk.org.uk/speakers.php?c=miami

[8]https://github.com/marekrei/sequence-labeler

performed only marginally better than BiCVM and BiCCA on semantic and syntactic tasks respectively, it is the only embedding model that did consistently well across tasks. Thus, our study shows that standard bilingual embeddings are not well suited, in general, for CM tasks; embeddings learnt from CM data, either real or synthetic, seems much more useful. Further, along the lines of (Pratapa et al., 2018), our study also shows that synthetic CM data is a reasonably good proxy for real data.

While our experiments show a promising direction towards obtaining bilingual embeddings for CM tasks, there are several interesting ideas that are worth exploring. In particular, the linguistic model used for generating artificial CM data only addresses the syntactic constraints of CM, but not other kinds of constraints such as *lexical choice* which in a particular CM context might be overly skewed towards one language (like the English words 'school' and 'vote' are more common than their Hindi translations in English-Hindi CM (Bali et al., 2014)), and semantic/pragmatic constraints that make the choice of a particular language more common in some contexts (e.g., Hindi used more commonly for negative sentiment during English-Hindi CM (Rudra et al., 2016)). Similarly, the sense distribution of polysemous words can vary widely between a monolingual and CM corpus. For instance, the word 'school' in English has several meanings such as (a) an institute of education, (b) group of artists, writers etc., and (c) a large group of fish. However, in a Spanish dominant sentence or corpus, school is primarily, if not only, used in sense (a).

As future work, it will be interesting to explore techniques that can generate artificial CM data following the lexical, semantic and pragmatic constraints, or develop novel embedding techniques that can appropriately interpolate between real and artificial CM data to learn collocations that arise due to not only syntactic but also lexical, semantic and pragmatic aspects of code-mixing.

Acknowledgements

We would like to thank Kalika Bali, Gayatri Bhat and Sandipan Dandapat for their valuable suggestions and help in the creation of the synthetic CM corpus.

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 937–947.

Fahad AlGhamdi, Giovanni Molina, Mona Diab, Thamar Solorio, Abdelati Hawwari, Victor Soto, and Julia Hirschberg. 2016. Part of speech tagging for code switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 98–107.

Kalika Bali, Y. Vyas, J. Sharma, and M. Choudhury. 2014. "I am borrowing ya mixing?" An analysis of English-Hindi code mixing in Facebook. In *Proc. First Workshop on Computational Approaches to Code Switching, EMNLP*.

A.-M. DiSciullo, Pieter Muysken, and R. Singh. 1986. Government and code-mixing. *Journal of Linguistics*, 22:1–24.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 587–593.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*.

Souvick Ghosh, Satanu Ghosh, and Dipankar Das. 2016. Part-of-speech tagging of code-mixed social media text. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 90–97.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning*, pages 748–756.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributional Semantics. In *Proceedings of ACL*.

A. K. Joshi. 1985. Processing of Sentences with Intrasentential Code Switching. In D. R. Dowty, L. Karttunen, and A. M. Zwicky, editors, *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, pages 190–205. Cambridge University Press, Cambridge.

Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. 2016. Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491.

Braj B Kachru. 1978. Toward structuring code-mixing: An Indian perspective. *International Journal of the Sociology of Language*, 1978(16):27–46.

Ying Li and P Fung. 2014. Language modeling with functional head constraint for code switching speech recognition. In *EMNLP*.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *NAACL Workshop on Vector Space Modeling for NLP*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Carol Myers-Scotton. 1993. *Duelling Languages:Grammatical structure in Code-switching*. Clarendon Press, Oxford.

Rana D. Parshad, Suman Bhowmick, Vineeta Chand, Nitu Kumari, and Neha Sinha. 2016. What is India speaking? Exploring the "Hinglish" invasion. *Physica A*, 449:375–389.

Shana Poplack. 1980. Sometimes Ill start a sentence in Spanish y termino en espaol. *Linguistics*, 18:581–618.

Ameya Prabhu, Aditya Joshi, Manish Shrivastava, and Vasudeva Varma. 2016. Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553.

Marek Rei and Helen Yannakoudakis. 2016. Compositional sequence labeling models for error detection in learner writing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1181–1191.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982.

Sebastian Ruder, Ivan Vuli, and Anders Sgaard. 2017. A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902*.

Koustav Rudra, S Rijhwani, R Begum, K Bali, M Choudhury, and N Ganguly. 2016. Understanding language preference for expression of opinion and sentiment: What do Hindi-English speakers do on Twitter? In *EMNLP*, pages 1131–1141.

David Sankoff. 1998. A formal production-based explanation of the facts of code-switching. *Bilingualism: language and cognition*, 1(01):39–50.

A. Sharma, S. Gupta, R. Motlani, P. Bansal, M. Srivastava, R. Mamidi, and D.M Sharma. 2016. Shallow parsing pipeline for Hindi-English code-mixed social media text. In *Proceedings of NAACL-HLT*.

Thamar Solorio and Yang Liu. 2008. Part-of-speech tagging for English-Spanish code-switched text. In *Proc. of EMNLP*.

Thamar Solorio et al. 2014. Overview for the first shared task on language identification in code-switched data. *In 1st Workshop on Computational Approaches to Code Switching, EMNLP*, pages 62–72.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. *arXiv preprint arXiv:1604.00425*.

David Vilares and Miguel A Alonso. 2016. En-es-cs: An English-Spanish code-switching twitter corpus for multilingual sentiment analysis. In *LREC*.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 363–372. ACM.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. POS tagging of English-Hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979.

Michael Wick, Pallika Kanani, and Adam Craig Pocock. 2016. Minimally-constrained multilingual embeddings via artificial code-switching. In *AAAI*, pages 2849–2855.

Will Y Zou, Richard Socher, Daniel 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*, pages 1393–1398.