When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation?

Ye Qi, Devendra Singh Sachan, Matthieu Felix,
Sarguna Janani Padmanabhan, Graham Neubig
Language Technologies Institute, Carnegie Mellon University, USA
{yeq,dsachan,matthief,sjpadman,gneubig}@andrew.cmu.edu

Abstract

The performance of Neural Machine Translation (NMT) systems often suffers in low-resource scenarios where sufficiently large-scale parallel corpora cannot be obtained. Pre-trained word embeddings have proven to be invaluable for improving performance in natural language analysis tasks, which often suffer from paucity of data. However, their utility for NMT has not been extensively explored. In this work, we perform five sets of experiments that analyze when we can expect pre-trained word embeddings to help in NMT tasks. We show that such embeddings can be surprisingly effective in some cases – providing gains of up to 20 BLEU points in the most favorable setting.[1]

1 Introduction

Pre-trained word embeddings have proven to be highly useful in neural network models for NLP tasks such as sequence tagging (Lample et al., 2016; Ma and Hovy, 2016) and text classification (Kim, 2014). However, it is much less common to use such pre-training in NMT (Wu et al., 2016), largely because the large-scale training corpora used for tasks such as WMT[2] tend to be several orders of magnitude larger than the annotated data available for other tasks, such as the Penn Treebank (Marcus et al., 1993). However, for low-resource languages or domains, it is not necessarily the case that bilingual data is available in abundance, and therefore the effective use of monolingual data becomes a more desirable option.

Researchers have worked on a number of methods for using monolingual data in NMT systems (Cheng et al., 2016; He et al., 2016; Ramachandran et al., 2016). Among these, pre-trained word embeddings have been used either in standard translation systems (Neishi et al., 2017; Artetxe et al., 2017) or as a method for learning translation lexicons in an entirely unsupervised manner (Conneau et al., 2017; Gangi and Federico, 2017). Both methods show potential improvements in BLEU score when pre-training is properly integrated into the NMT system.

However, from these works, it is still not clear as to *when* we can expect pre-trained embeddings to be useful in NMT, or *why* they provide performance improvements. In this paper, we examine these questions more closely, conducting five sets of experiments to answer the following questions:

- Q1 Is the behavior of pre-training affected by language families and other linguistic features of source and target languages? (§3)
- Q2 Do pre-trained embeddings help more when the size of the training data is small? (§4)
- Q3 How much does the similarity of the source and target languages affect the efficacy of using pre-trained embeddings? (§5)
- Q4 Is it helpful to align the embedding spaces between the source and target languages? (§6)
- Q5 Do pre-trained embeddings help more in multilingual systems as compared to bilingual systems? (§7)

2 Experimental Setup

In order to perform experiments in a controlled, multilingual setting, we created a parallel corpus from TED talks transcripts.[3] Specifically, we prepare data between English (EN) and three pairs of languages, where the two languages in the pair are similar, with one being relatively low-resourced compared to the other: Galician (GL) and Portuguese (PT), Azerbaijani (AZ) and Turkish (TR), and Belarusian (BE) and Russian (RU).

[1] Scripts/data to replicate experiments are available at https://github.com/neulab/word-embeddings-for-nmt

[2] http://www.statmt.org/wmt17/

[3] https://www.ted.com/participate/translate

Proceedings of NAACL-HLT 2018, pages 529–535
New Orleans, Louisiana, June 1 - 6, 2018.

Dataset	train	dev	test
GL → EN	10,017	682	1,007
PT → EN	51,785	1,193	1,803
AZ → EN	5,946	671	903
TR → EN	182,450	4,045	5,029
BE → EN	4,509	248	664
RU → EN	208,106	4,805	5,476

Table 1: Number of sentences for each language pair.

The languages in each pair are similar in vocabulary, grammar and sentence structure (Matthews, 1997), which controls for language characteristics and also improves the possibility of transfer learning in multi-lingual models (in §7). They also represent different language families – GL/PT are Romance; AZ/TR are Turkic; BE/RU are Slavic – allowing for comparison across languages with different caracteristics. Tokenization was done using `Moses` tokenizer[4] and hard punctuation symbols were used to identify sentence boundaries. Table 1 shows data sizes.

For our experiments, we use a standard 1-layer encoder-decoder model with attention (Bahdanau et al., 2014) with a beam size of 5 implemented in `xnmt`[5] (Neubig et al., 2018). Training uses a batch size of 32 and the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.0002, decaying the learning rate by 0.5 when development loss decreases (Denkowski and Neubig, 2017). We evaluate the model's performance using BLEU metric (Papineni et al., 2002).

We use available pre-trained word embeddings (Bojanowski et al., 2016) trained using `fastText`[6] on Wikipedia[7] for each language. These word embeddings (Mikolov et al., 2017) incorporate character-level, phrase-level and positional information of words and are trained using CBOW algorithm (Mikolov et al., 2013). The dimension of word embeddings is set to 300. The embedding layer weights of our model are initialized using these pre-trained word vectors. In baseline models without pre-training, we use Glorot and Bengio (2010)'s uniform initialization.

3 Q1: Efficacy of Pre-training

In our first set of experiments, we examine the efficacy of pre-trained word embeddings across the various languages in our corpus. In addition to providing additional experimental evidence supporting the findings of other recent work on using pre-trained embeddings in NMT (Neishi et al., 2017; Artetxe et al., 2017; Gangi and Federico, 2017), we also examine whether pre-training is useful across a wider variety of language pairs and if it is more useful on the source or target side of a translation pair.

Src → / → Trg	std std	pre std	std pre	pre pre
GL → EN	2.2	**13.2**	2.8	12.8
PT → EN	26.2	**30.3**	26.1	**30.8**
AZ → EN	1.3	**2.0**	1.6	**2.0**
TR → EN	14.9	17.6	14.7	**17.9**
BE → EN	1.6	2.5	1.3	**3.0**
RU → EN	18.5	**21.2**	18.7	**21.1**

Table 2: Effect of pre-training on BLEU score over six languages. The systems use either random initialization (`std`) or pre-training (`pre`) on both the source and target sides.

The results in Table 2 clearly demonstrate that pre-training the word embeddings in the source and/or target languages helps to increase the BLEU scores to some degree. Comparing the second and third columns, we can see the increase is much more significant with pre-trained source language embeddings. This indicates that the majority of the gain from pre-trained word embeddings results from a better encoding of the source sentence.

The gains from pre-training in the higher-resource languages are consistent: ≈3 BLEU points for all three language pairs. In contrast, for the extremely low-resource languages, the gains are either quite small (AZ and BE) or very large, as in GL which achieves a gain of up to 11 BLEU points. This finding is interesting in that it indicates that word embeddings may be particularly useful to bootstrap models that are on the threshold of being able to produce reasonable translations, as is the case for GL in our experiments.

4 Q2: Effect of Training Data Size

The previous experiment had interesting implications regarding available data size and effect of pre-training. Our next series of experiments examines this effect in a more controlled environment by down-sampling the training data for the higher-resource languages to 1/2, 1/4 and 1/8 of their original sizes.

From the BLEU scores in Figure 1, we can see

[4] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl
[5] https://github.com/neulab/xnmt/
[6] https://github.com/facebookresearch/fastText/
[7] https://dumps.wikimedia.org/

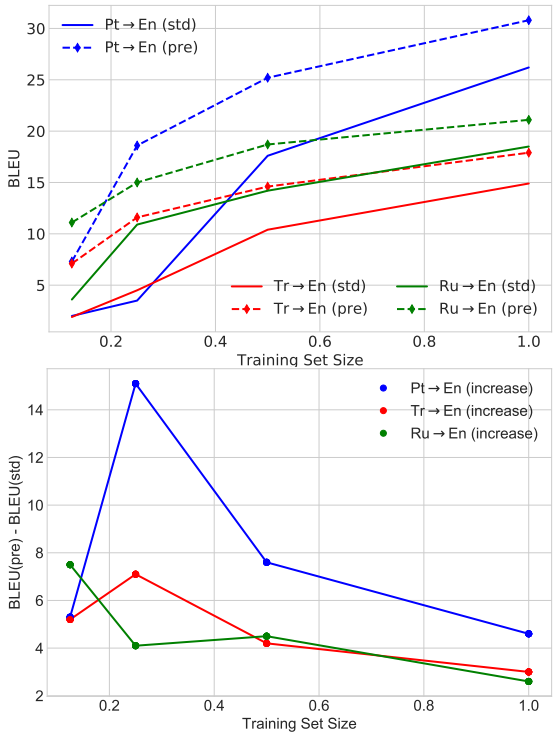

Figure 1: BLEU and BLEU gain by data size.

that for all three languages the gain in BLEU score demonstrates a similar trend to that found in GL in the previous section: the gain is highest when the baseline system is poor but not too poor, usually with a baseline BLEU score in the range of 3-4. This suggests that at least a moderately effective system is necessary before pre-training takes effect, but once there is enough data to capture the basic characteristics of the language, pre-training can be highly effective.

5 Q3: Effect of Language Similarity

The main intuitive hypothesis as to why pre-training works is that the embedding space becomes more consistent, with semantically similar words closer together. We can also make an additional hypothesis: if the two languages in the translation pair are more linguistically similar, the semantic neighborhoods will be more similar between the two languages (i.e. semantic distinctions or polysemy will likely manifest themselves in more similar ways across more similar languages). As a result, we may expect that the gain from pre-training of embeddings may be larger when the source and target languages are more similar. To examine this hypothesis, we selected Portuguese as the target language, which when following its language family tree from top to bottom, belongs to Indo-European, Romance,

Dataset	Lang. Family	std	pre
ES → PT	West-Iberian	17.8	24.8 (+**7.0**)
FR → PT	Western Romance	12.4	18.1 (+5.7)
IT → PT	Romance	14.5	19.2 (+4.7)
RU → PT	Indo-European	2.4	8.6 (+6.2)
HE → PT	*No Common*	3.0	11.9 (+**8.9**)

Table 3: Effect of linguistic similarity and pre-training on BLEU. The language family in the second column is the most recent common ancestor of source and target language.

Western Romance, and West-Iberian families. We then selected one source language from each family above.[8] To avoid the effects of training set size, all pairs were trained on 40,000 sentences.

From Table 3, we can see that the BLEU scores of ES, FR, and IT do generally follow this hypothesis. As we move to very different languages, RU and HE see larger accuracy gains than their more similar counterparts FR and IT. This can be largely attributed to the observation from the previous section that systems with larger headroom to improve tend to see larger increases; RU and HE have very low baseline BLEU scores, so it makes sense that their increases would be larger.

6 Q4: Effect of Word Embedding Alignment

Until now, we have been using embeddings that have been trained independently in the source and target languages, and as a result there will not necessarily be a direct correspondence between the embedding spaces in both languages. However, we can postulate that having consistent embedding spaces across the two languages may be beneficial, as it would allow the NMT system to more easily learn correspondences between the source and target. To test this hypothesis, we adopted the approach proposed by Smith et al. (2017) to learn orthogonal transformations that convert the word embeddings of multiple languages to a single space and used these aligned embeddings instead of independent ones.

From Table 4, we can see that somewhat surprisingly, the alignment of word embeddings was not beneficial for training, with gains or losses essentially being insignificant across all languages. This, in a way, is good news, as it indicates that *a priori* alignment of embeddings may not be neces-

[8] English was excluded because the TED talks were originally in English, which results in it having much higher BLEU scores than the other languages due to it being direct translation instead of pivoted through English like the others.

Dataset	unaligned	aligned
GL → EN	12.8	11.5 (−1.3)
PT → EN	30.8	30.6 (−0.2)
AZ → EN	2.0	2.1 (+0.1)
TR → EN	17.9	17.7 (−0.2)
BE → EN	3.0	3.0 (+0.0)
RU → EN	21.1	21.4 (+0.3)

Table 4: Correlation between word embedding alignment and BLEU score in bilingual translation task.

Train	Eval	bi	std	pre	align
GL + PT	GL	2.2	17.5	20.8	**22.4**
AZ + TR	AZ	1.3	5.4	5.9	**7.5**
BE + RU	BE	1.6	**10.0**	7.9	9.6

Table 5: Effect of pre-training on multilingual translation into English. bi is a bilingual system trained on only the eval source language and all others are multi-lingual systems trained on two similar source languages.

sary in the context of NMT, since the NMT system can already learn a reasonable projection of word embeddings during its normal training process.

7 Q5: Effect of Multilinguality

Finally, it is of interest to consider pre-training in multilingual translation systems that share an encoder or decoder between multiple languages (Johnson et al., 2016; Firat et al., 2016), which is another promising way to use additional data (this time from another language) as a way to improve NMT. Specifically, we train a model using our pairs of similar low-resource and higher-resource languages, and test on only the low-resource language. For those three pairs, the similarity of GL/PT is the highest while BE/RU is the lowest.

We report the results in Table 5. When applying pre-trained embeddings, the gains in each translation pair are roughly in order of their similarity, with GL/PT showing the largest gains, and BE/RU showing a small decrease. In addition, it is also interesting to note that as opposed to previous section, aligning the word embeddings helps to increase the BLEU scores for all three tasks. These increases are intuitive, as a single encoder is used for both of the source languages, and the encoder would have to learn a significantly more complicated transform of the input if the word embeddings for the languages were in a semantically separate space. Pre-training and alignment ensures that the word embeddings of the two source languages are put into similar vector spaces, allowing the model to learn in a similar fashion as it would if training on a single language.

Interestingly, BE → EN does not seem to benefit from pre-training in the multilingual scenario, which hypothesize is due to the fact that: 1) Belarusian and Russian are only partially mutually intelligible (Corbett and Comrie, 2003), i.e., they are not as similar; 2) the Slavic languages have comparatively rich morphology, making sparsity in the trained embeddings a larger problem.

8 Analysis

8.1 Qualitative Analysis

Finally, we perform a qualitative analysis of the translations from GL → EN, which showed one of the largest increases in quantitative numbers. As can be seen from Table 6, pre-training not only helps the model to capture rarer vocabulary but also generates sentences that are more grammatically well-formed. As highlighted in the table cells, the best system successfully translates a person's name ("*chris*") and two multi-word phrases ("*big lawyer*" and "*patent legislation*"), indicating the usefulness of pre-trained embeddings in providing a better representations of less frequent concepts when used with low-resource languages.

In contrast, the bilingual model without pre-trained embeddings substitutes these phrases for common ones ("*i*"), drops them entirely, or produces grammatically incorrect sentences. The incomprehension of core vocabulary causes deviation of the sentence semantics and thus increases the uncertainty in predicting next words, generating several phrasal loops which are typical in NMT systems.

8.2 Analysis of Frequently Generated n-grams.

We additionally performed pairwise comparisons between the top 10 n-grams that each system (selected from the task GL → EN) is better at generating, to further understand what kind of words pre-training is particularly helpful for.[9] The results displayed in Table 7 demonstrate that pre-training helps both with words of low frequency in the training corpus, and even with function words such as prepositions. On the other hand, the improvements in systems without pre-trained embed-

[9] Analysis was performed using `compare-mt.py` from https://github.com/neubig/util-scripts/.

source	*(risos)* e é que **chris** é un grande avogado , pero non sabía case nada sobre lexislación de patentes e absolutamente nada sobre xenética .
reference	*(laughter)* now **chris** is a really brilliant lawyer , but he knew almost nothing about patent law and certainly nothing about genetics .
bi:std	*(laughter)* and **i** 'm not a little bit of a little bit of a little bit of and (laughter) and i 'm going to be able to be a lot of years .
multi:pre-align	*(laughter)* and **chris** is a big lawyer , but i did n't know almost anything about patent legislation and absolutely nothing about genetic .

Table 6: Example translations of GL → EN.

bi:std		bi:pre	
) so	2/0	about	0/53
(laughter) i	2/0	people	0/49
) i	2/0	or	0/43
laughter) i	2/0	these	0/39
) and	2/0	with	0/38
they were	1/0	because	0/37
have to	5/2	like	0/36
a new	1/0	could	0/35
to do ,	1/0	all	0/34
`` and then	1/0	two	0/32

(a) Pairwise comparison between two bilingual models

multi:std		multi:pre+align	
here	6/0	on the	0/14
again ,	4/0	like	1/20
several	4/0	should	0/9
you 're going	4/0	court	0/9
've	4/0	judge	0/7
we 've	4/0	testosterone	0/6
you 're going to	4/0	patents	0/6
people ,	4/0	patent	0/6
what are	3/0	test	0/6
the room	3/0	with	1/12

(b) Pairwise comparison between two multilingual models

Table 7: Top 10 n-grams that one system did a better job of producing. The numbers in the figure, separated by a slash, indicate how many times each n-gram is generated by each of the two systems.

dings were not very consistent, and largely focused on high-frequency words.

8.3 F-measure of Target Words

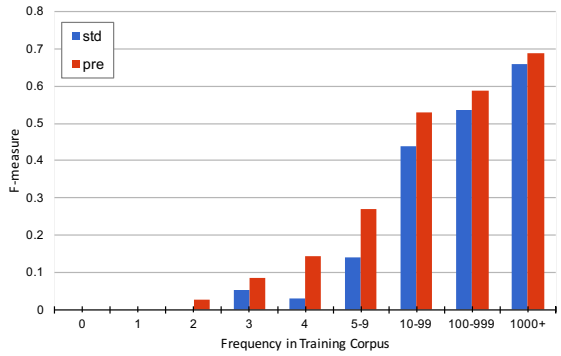

Figure 2: The f-measure of target words in bilingual translation task PT → EN

Finally, we performed a comparison of the f-measure of target words, bucketed by frequency in the training corpus. As displayed in Figure 2, this shows that pre-training manages to improve the accuracy of translation for the entire vocabulary, but particularly for words that are of low frequency in the training corpus.

9 Conclusion

This paper examined the utility of considering pre-trained word embeddings in NMT from a number of angles. Our conclusions have practical effects on the recommendations for when and why pre-trained embeddings may be effective in NMT, particularly in low-resource scenarios: (1) there is a sweet-spot where word embeddings are most effective, where there is very little training data but not so little that the system cannot be trained at all, (2) pre-trained embeddings seem to be more effective for more similar translation pairs, (3) *a priori* alignment of embeddings may not be necessary in bilingual scenarios, but is helpful in multi-lingual training scenarios.

Acknowledgements

Parts of this work were sponsored by Defense Advanced Research Projects Agency Information Innovation Office (I2O). Program: Low Resource Languages for Emergent Incidents (LORELEI). Issued by DARPA/I2O under Contract No. HR0011-15-C-0114. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041* .

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv e-prints* abs/1409.0473. https://arxiv.org/abs/1409.0473.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. *arXiv preprint arXiv:1606.04596* .

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087* .

Greville Corbett and Bernard Comrie. 2003. *The Slavonic Languages*. Routledge.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. *arXiv preprint arXiv:1706.09733* .

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073* .

Mattia Antonino Di Gangi and Marcello Federico. 2017. Monolingual embeddings for low resourced neural machine translation. In *International Workshop on Spoken Language Translation (IWSLT)*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. pages 249–256.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*. pages 820–828.

Melvin Johnson et al. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558* .

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *In EMNLP*. Citeseer.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *HLT-NAACL*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1064–1074. http://www.aclweb.org/anthology/P16-1101.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics* 19(2):313–330.

P.H. Matthews. 1997. *The Concise Oxford Dictionary of Linguistics.*. Oxford Paperback Reference / Oxford University Press, Oxford. Oxford University Press, Incorporated. https://books.google.com/books?id=aYoYAAAAIAAJ.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in pre-training distributed word representations .

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.

Masato Neishi, Jin Sakuma, Satoshi Tohda, Shonosuke Ishiwatari, Naoki Yoshinaga, and Masashi Toyoda. 2017. A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, pages 99–109.

Graham Neubig, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, Sarguna Padmanabhan, Ye Qi, Devendra Singh Sachan, Philip Arthur, Pierre Godard, John Hewitt, Rachid Riad, and Liming Wang. 2018. XNMT: The extensible neural machine translation toolkit. In *Conference of the Association for Machine Translation in the Americas (AMTA) Open Source Software Showcase*. Boston.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 311–318.

Prajit Ramachandran, Peter J Liu, and Quoc V Le. 2016. Unsupervised pretraining for sequence to sequence learning. *arXiv preprint arXiv:1611.02683* .

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859* .

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144.