English-Basque Statistical and Neural Machine Translation

Inigo Jauregi Unanue[1,2], Lierni Garmendia Arratibel[3], Ehsan Zare Borzeshi[2], Massimo Piccardi[1]

[1]University of Technology Sydney (UTS), Sydney, Australia
[2]Capital Markets Cooperative Research Center (CMCRC), Sydney, Australia
[3]Professional Translator, Barcelona, Spain

{ijauregi,ezborzeshi}@cmcrc.com, liernigarmendia@gmail.com, massimo.piccardi@uts.edu.au

Abstract

Neural Machine Translation (NMT) has attracted increasing attention in the recent years. However, it tends to require very large training corpora which could prove problematic for languages with low resources. For this reason, Statistical Machine Translation (SMT) continues to be a popular approach for low-resource language pairs. In this work, we address English-Basque translation and compare the performance of three contemporary statistical and neural machine translation systems: OpenNMT, Moses SMT and Google Translate. For evaluation, we employ an open-domain and an IT-domain corpora from the WMT16 resources for machine translation. In addition, we release a small dataset (*Berriak*) of 500 highly-accurate English-Basque translations of complex sentences useful for a thorough testing of the translation systems.

Keywords: Neural Machine Translation, Statistical Machine Translation, English-Basque, Basque

1. Introduction

The advent of deep neural networks in natural language processing (NLP) has led to significant progress in a variety of classification tasks, including named-entity recognition (Lample et al., 2016), answer sentence selection (Lowe et al., 2017) and natural language inference (Wang et al., 2017). Deep neural networks have also started to obtain promising results in NLP, mainly in Neural Machine Translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017; Wu et al., 2016). In NMT, the model receives a sentence in a *source* language as input and generates its translation word by word in a *target* language. NMT has outperformed previous translation systems in many language pairs (e.g., German-English, French-English).

However, in order to reach high accuracies, neural translation systems tend to require very large parallel training corpora (Koehn and Knowles, 2017). As a matter of fact, such corpora are not yet available for many language pairs. When the training data are relatively small, other, more traditional approaches such as Statistical Machine Translation (SMT) (Koehn et al., 2007) seem to be more accurate. Several ideas have been proposed in order to mollify this issue including multi-lingual systems with zero-shot translation (Johnson et al., 2016), transfer learning (Zoph et al., 2016) and back-translations (Sennrich et al., 2016). However, their general effectiveness still requires wider evaluation.

For this work, we have selected a low-resource language pair, English-Basque (abbreviation: en-eu), and used it as a case study for statistical and neural machine translation. Past machine translators for the Basque language have mainly used rule-based (Mayor et al., 2011) and statistical approaches (de Ilarraza et al., 2008; Del Gaudio et al., 2016; Stroppa et al., 2006). In our work, we compare three different systems: OpenNMT, an open-source NMT system; Moses SMT, an open-source SMT system; and Google Translate, a publicly-available commercial system which uses either SMT or NMT models depending on the language pair. The first two have been trained by us with dedicated datasets, while Google has just been used "as is" from its API. The three models have been tested over open-domain and Information Technology (IT) domain datasets from the WMT2016 IT helpdesk shared task. Moreover, we release a new, small en-eu corpus (named *Berriak*) useful for probing English-Basque machine translation. This corpus consists of 500 long and complex sentences translated from English to Basque by experienced translators and is much more realistic and challenging than the existing en-eu corpora. Due its small size, we have only used it as a test set.

The rest of this paper is organised as follows: Section 2. discusses the main characteristics of the Basque language. Section 3. describes the compared systems. Section 4. describes the datasets used in the experiments and presents the results. Section 5. concludes this paper.

2. The Basque Language

Basque (*Euskara*) is a language spoken in the Basque Country (northern Spain and southwestern France). It is not considered a member of the Indo-European language family and it remains isolated to date, meaning that researchers have not found any other language with similar characteristics. As stated by (Mayor et al., 2011), Basque is an *agglutinative* language with rich inflectional morphology. This means that a word may include several morphemes that change its inflectional category such as number, case, tense or person. Unlike in fusional languages such as Spanish and French, in agglutinative languages the boundaries between morphemes remain clear-cut (Aikhenvald, 2007). For agglutinative languages, rule-based systems have proved an effective translation approach in the past (Koehn and Monz, 2006).

During the military dictatorship of Spain (1939-1975), Basque became illegal and the number of speakers dropped drastically. However, in the 1970s a process for the formal standardization of the language (*Euskara Batua*) began and Basque teaching in schools was restored. Nowadays, the language is official in the Basque autonomous community in Spain and it has reached approximately 1M speakers (including Navarre and the French side). Yet, the persisting lack of available translation corpora makes the development of Basque machine translators a difficult task.

3. Methods

3.1. Moses SMT

SMT has been the state-of-the-art approach to machine translation for many years (Koehn et al., 2003). SMT systems are usually phrase-based system which first try to learn the alignments of phrases between different languages, and then predict the best composition of phrases for the translation with the help of a target language model (LM).

In this work, we have evaluated a popular open-source SMT toolkit called Moses (Koehn et al., 2007). First, a word alignment model between the source and target languages has been learned over the training data with the GIZA++ toolkit (Och and Ney, 2003). Then, an LM has been learned over the training data with KenLM (Heafield, 2011). Finally, based on these two models, the Moses decoder has been used to translate the sentences.

3.2. Google Translate

Google Translate is a publicly-available, well-known commercial machine translator. Recently, it has implemented the Google Neural Machine Translation (GNMT) (Johnson et al., 2016) over many language pairs and en-eu is one of them[1]. Google Translate does not train from a dedicated annotation of parallel text; rather, it crawls the Web to forage for "likely parallel" paragraphs (for instance, those marked with multiple HTML lang tags). In our work, we have used it from the convenient Google Cloud Translation API[2].

3.3. OpenNMT

We have trained an NMT model by using the OpenNMT toolkit (Klein et al., 2017) with the *seq2seq* architecture of (Sutskever et al., 2014). This architecture is formed by an encoder, which converts the source sentence into a sequence of numerical vectors, and a decoder, which predicts the target sentence based on the encoded source sentence. Both the encoder and the decoder are usually recurrent neural networks (RNNs). Additionally, an *attention mechanism* (Bahdanau et al., 2014; Luong et al., 2015) has been used to learn soft-alignments between the source and the target sentences.

In our model, we have used the Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) for both the encoder and the decoder. We have also used the *dot* attention mechanism (Luong et al., 2015) where weights over the encoded source sentence are provided by an auxiliary network. Like with any other neural models for NLP, prior to processing each unique word in the corpus needs to be mapped to a high-dimensional vector (*word embedding*). This mapping can be either random (the default) or based on user-provided pre-trained embeddings. In addition, the word embeddings can be kept constant during training, or updated alongside all other parameters to minimise the cost function. Since pre-trained embeddings have typically reported higher accuracies (Dernoncourt et al., 2017; Lample et al., 2016), we have trained Basque word embeddings using GloVe (Pennington et al., 2014) over the Basque Wikipedia. For English, we have used the available CommonCrawl pre-trained embeddings[3]. We have evaluated the use of these embeddings in two different ways: maintaining them fixed during both training and testing (f-emb) and updating them during the training stage (u-emb). The word embeddings have a dimension of 300. The remaining parameters of the network have been kept to their default values.

Data	***PaCo2_EnEu***	***WMT16_IT***	***Berriak***
training	125,356	89,983	—
test	5,000	1,000	500

Table 1: The number of samples in the *PaCo_EnEu*, *WMT16_IT* and *Berriak* datasets.

4. Experiments

4.1. Corpora

As mentioned previously, the en-eu language pair is considered to be low-resourced. To mitigate this issue, the WMT16 machine translation for IT domain shared task[4] provided a parallel en-eu corpus. This corpus includes both an IT-domain dataset and an open-domain dataset (*PaCo2_EnEu*). *PaCo2_EnEu* consists of approximately 130,000 en-eu translations crawled from the web (San Vicente et al., 2012). In the experiments, we have used 5,000 as a test set and the rest for training the models. We have also used the IT-domain data to evaluate the translators over a specialised domain. The IT-domain training set consists of 89,983 samples, but only 2,000 of them are proper sentences; the rest are translations of IT terms from Wikipedia and localization PO files. Consequently, the amount of "good quality" data in the training set is very limited. At its turn, the test set consists of 1,000 proper sentences.

By inspecting these resources, we had realised the lack of long and complex sentences, likely a major limitation for the realistic evaluation of this language pair. Such sentences appear in most professional translations and they are expected to prove far more challenging for automated translators. To provide a resource contribution, we have therefore collected and released a small, high-quality en-eu corpus called *Berriak* (*news* in Basque)[5,6]. To create a suitable corpus, we have randomly selected English sentences

[1] https://cloud.google.com/translate/docs/languages

[2] Google Translate API: https://cloud.google.com/translate/docs/

[3] GloVe: https://nlp.stanford.edu/projects/glove/

[4] WMT16:http://www.statmt.org/wmt16/it-translation-task.html

[5] ISLRN: 197-383-395-000-1

[6] https://github.com/ijauregiCMCRC/english_basque_MT

Model	***PaCo2_EnEu***	***Berriak***
en→eu		
Moses SMT	21.02	5.90
OpenNMT(r-emb)	20.07	1.49
OpenNMT(f-emb)	19.39	1.43
OpenNMT(u-emb)	21.18	1.84
Google Translate	9.12	9.91
eu→en		
Moses SMT	24.20	8.53
OpenNMT(r-emb)	19.44	3.35
OpenNMT(f-emb)	18.61	4.28
OpenNMT(u-emb)	22.42	5.53
Google Translate	16.66	20.80

Table 2: BLEU score of the models over the *PaCo_EnEu* and *Berriak* corpora.

Model	***WMT16_IT***
en→eu	
Moses SMT	11.74
Moses SMT+(PaCo train)	11.89
OpenNMT(r-emb)	11.87
OpenNMT(PaCo train)(r-emb)	12.42
OpenNMT(u-emb)	12.75
OpenNMT(PaCo train)(u-emb)	12.31
Google Translate	14.46
eu→en	
Moses SMT	19.06
Moses SMT+(PaCo train)	19.34
OpenNMT(r-emb)	15.46
OpenNMT(PaCo train)(r-emb)	16.93
OpenNMT(u-emb)	17.30
OpenNMT(PaCo train)(u-emb)	18.01
Google Translate	24.66

Table 3: BLEU score of the models over the *WMT16_IT* corpus.

from the English-German news corpus of WMT16 which meets the requirements. The sentences have been translated into Basque with the help of Librezale[7], an open group of highly-qualified volunteers who work to increase the use of the Basque language in the IT domain. To date, we have collected 500 en-eu translations for a total of 10,280 tokens. Due to the small size of this corpus, in this work we have only used it as a further test set for models trained with *PaCo2_EnEu*. The manual translations are still ongoing and we intend to release an extended version in the near future. Table 1 summarises the number of samples of the various datasets.

4.2. Experimental Settings and Results

We have conducted a number of experiments to evaluate the models in a variety of scenarios. In the first experiment, we have trained the SMT and NMT models with the *PaCo2_EnEu* training set and tested them with both the *PaCo2_EnEu* test set and *Berriak*. Google Translate has been used as is. Experiments have been conducted in both English-to-Basque (en→eu) and Basque-to-English (eu→en) to assess performance in both directions. In addition, for the NMT model we have experimented with updated random embeddings (r-emb), fixed pre-trained embeddings (f-emb), and updated pre-trained embeddings (u-emb). Table 2 reports the BLEU scores (Papineni et al., 2002) for the three models. The first remark is that all models generally perform worse with Basque as the target language. This suggests that its intrinsic difficulty is higher than English. As for the models' comparison, both Moses SMT and OpenNMT have remarkably outperformed Google Translate on the *PaCo2_EnEu* test set. The NMT model has achieved the highest BLEU score (21.18) in the en→eu direction, while the SMT model has achieved the highest BLEU score (24.20) in the opposite direction. For the NMT model, updating the pre-trained embeddings during training (u-emb) has invariably led to the highest accuracies, up to an improvement of 2.98 BLEU points over the random embeddings in the eu→en direction.

However, the performance ranking has changed drastically when testing on the more probing *Berriak* corpus. In this case, Google Translate has achieved the highest BLEU scores by a large extent. We believe that both Moses SMT and OpenNMT, which have been trained using only the *PaCo2_EnEu* training set, have obtained such low results because the training corpus does not contain the same kind of long sentences as *Berriak* and therefore the models could not learn to translate such challenging sentences. Between SMT and NMT, the former has clearly outperformed the latter, confirming that SMT generalises better when the training corpus is limited. On the other hand, the training corpus of Google Translate is certainly much bigger, and that has helped it achieve better results on *Berriak*. However, the BLEU score when Basque is the target is still very low (9.91) and significant improvements are an outstanding need. For what concerns NMT and word embedding, also in this case the updated pre-trained embeddings have led to an improvement (although slight) in score.

In a second experiment in the IT domain (Table 3), Google Translate has again obtained the best results. This can be explained with the fact that Moses SMT and OpenNMT have only been trained with 2,000 proper IT-domain sentences. Between these two models, OpenNMT has outperformed Moses SMT for Basque as the target language, and vice versa for English. To mollify the small training size issue, we have added the open-domain corpus to the training data (noted as *PaCo train* in Table 3). The results have slightly improved for both NMT and SMT, with a more noticeable improvement for NMT (12.42 in en→eu and 16.93 in eu→en). Larger relative improvements have been achieved with the use of the pre-trained embeddings (12.75 in en→eu and 17.30 in eu→en). Since the updated embeddings had proved clearly more accurate in the previous experiment, we have not used the fixed embeddings in this experiment. Finally, using both the open-domain data and the pre-trained embeddings has only improved the scores for English as the target language. Once again, all

[7]Librezale: https://librezale.eus/

English sentence	1. *How many people take part in The Sanfermin Bullrunnings - Sanfermin.com - Pamplona*
	2. *Search results as from 07/02/2011 in "Zarzuela"*
	3. *For this reason, after Madrid and Sydney, they plan to continue this international anti-bullfighting campaign in Croatia and Berlin.*
Ground Truth	1. *Sanferminetako entzierroa zenbat jendek egiten duen - Sanfermin.com - Pamplona*
	2. *Bilaketa emaitzak 2011/07/02 egunetik aurrera "Zarzuela"-(e)n*
	3. *Horregatik, zezenketen eta entzierroen aurkako nazioarteko protesta Madrilen eta Sidneyn egiteaz gain, Kroazia eta Berlinen ere izango da.*
Moses SMT	1. *Zenbat jende parte hartzeko Sanferminetako Bullrunnings - Sanfermin.com - Pamplona*
	2. *Bilaketa emaitzak 2011/07/02 egunetik aurrera "Zarzuela"-(e)*
	3. Hori dela eta, ondoren, Madrilera eta Sydney jarraitzen dute, plan horrek nazioarteko aurkako Zezenketetako @-@ kanpaina batean Kroazia eta Berlingo.
OpenNMT	1. *Nola bizi da Sanfermin Bullrunnings - Sanfermin.com - Pamplona*
	2. *Bilaketa emaitzak 2011/07/02 egunetik aurrera "Zarzuela"-(e)n*
	3. *Hori dela eta, Madrilen, Madrilen, Madrid, bullfighting eta Berlin, international eta Berlin.*
Google	1. *Zenbat pertsona parte hartu Sanferminetako entzierroetan - Sanfermin.com - Iruñean*
	2. *Bilaketaren emaitzak 2011/02/07 "Zarzuela" -en*
	3. *Horregatik, Madrilen eta Sydneyen ondoren, Kroazia eta Berlinen kontrako zezenketarako nazioarteko kanpaina aurrera eramateko asmoa dute.*

Table 4: Example of translations over the *PaCo2_EnEU* (en→eu) test set.

the models have performed significantly better with English as the target language, with an even bigger margin compared to the general-domain experiment. We speculate that this may be due to the fact that in the IT domain the Basque language does not have a vocabulary as comprehensive and developed as English does. In fact, many IT words and expressions are taken from English unchanged.

For a qualitative analysis, Table 4 shows three examples of translations provided by the different models alongside the ground truth from the *PaCo2_EnEu* test set, which is the dataset on which the trained NMT and SMT models have obtained the best accuracies. We can see that for sentence 2 the OpenNMT model has provided a translation identical to the ground truth, probably thanks to the fact that there are sentences with the same structure in the training corpus. However, the NMT model tends to directly bypass many words from the original English sentence into the prediction (see sentences 1 and 3; NB: OpenNMT allows the model to bypass words from the source sentence). More precisely, if the model is uncertain about which word from the target vocabulary should be predicted next, it will pass on the word with highest attention weight from the source sentence. This mechanism aims to help the prediction of words such as proper names, which are not likely to appear in the target vocabulary. As additional analisys, we have computed the percentage of words bypassed by the different NMT models. To this aim, we have counted the number of words in the test set's predictions which did not belong to the target vocabulary, and divided it by the total number of words predicted. Table 5 shows the computed percentages, averaged over all the different NMT models. We can observe that the numbers are clearly higher when Basque is the target language, which matches our intuition that Basque is more difficult to translate into. When comparing the different datasets, we observe a trend to bypass more words in *Berriak*, which is understandable as this dataset does not have a training corpus and has longer and more complex sentences. Conversely, *WMT16_IT* has the lowest percentages of source words in the predictions. This is likely due to the fact that this dataset is very domain-specific and has a smaller vocabulary size.

On a separate note, NMT tends to predict the same word repeatedly (see sentence 3), as often been reported for neural encoder-decoder architectures. On the other hand, the SMT model seems able to match more words correctly in each sentence, but it has difficulties to form grammatically-complete sentences (see all three examples). Finally, the sentences predicted by Google Translate contain synonyms of the words in the ground truth (e.g., *Iruñean* vs *Pamplona*) and errors in the inflectional morphemes (e.g., *entzierroa* vs *entzierroetan*, *Bilaketa* vs *Bilaketaren*, *zezenketen* vs *zezenketarako*).

Corpus	Bypassed (%)
en→eu	
PaCo2_EnEu	21.70
Berriak	36.60
WMT16_IT	3.78
eu→en	
PaCo2_EnEu	4.29
Berriak	7.59
WMT16_IT	1.07

Table 5: Average of the percentages of bypassed words by all the NMT models in each dataset and each direction.

5. Conclusion

This paper has presented a performance comparison of three contemporary MT approaches on a low-resourced language pair, English-Basque. The compared approaches include an NMT model (OpenNMT, with and LSTM encoder/decoder), an SMT model (Moses) and the popular Google Translate service.

The experimental results show that all the models have achieved worse results when Basque is the target language, confirming that languages with rich morphology are more difficult to translate into. The NMT and SMT models have outperformed Google Translate when using training and test data from the same corpus (*PaCo2_EnEu*). However, these models have not generalised well on long and complex sentences, in contrast to Google Translate. For the NMT model, initialising the word embeddings with pretrained embeddings (based on the Basque Wikipedia for Basque and CommonCrawl for English) and updating them during training has invariably led to the best BLEU scores. In absolute terms, the achieved BLEU scores suggest that machine translation for Basque still has large margins for improvement. As part of this research, we have released a new, small corpus (named *Berriak*) of highly accurate en-eu sentences translated by experienced human translators to be used as a probing test set for this language pair. In the future, we plan to enlarge the *Berriak* corpus for more extensive testing and explore ways to improve the accuracy of the translations without resorting to larger parallel datasets such as pivot languages and multi-lingual translators.

6. Acknowledgements

We would like to thank the Librezale association for the Basque translations for the new *Berriak* corpus. We would also like to acknowledge the financial support received from the Capital Markets Cooperative Research Centre (CMCRC), an applied research initiative of the Australian Government.

7. Bibliographical References

Aikhenvald, A. Y. (2007). Typological distinctions in word-formation. *Language typology and syntactic description*, 3:1–65.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

de Ilarraza, A. D., Labaka, G., and Sarasola, K. (2008). Statistical postediting: A valuable method in domain adaptation of rbmt systems for less-resourced languages. *Mixing Approaches to Machine Translation*, pages 35–40.

Del Gaudio, R., Labaka, G., Agirre, E., Osenova, P., Simov, K. I., Popel, M., Oele, D., van Noord, G., Gomes, L., Rodrigues, J. a. A., et al. (2016). Smt and hybrid systems of the qtleap project in the wmt16 it-task. In *WMT*, pages 435–441.

Dernoncourt, F., Lee, J. Y., Uzuner, O., and Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2016). Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the Workshop on Neural Machine Translation*. Association for Computational Linguistics.

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June. Association for Computational Linguistics.

Lowe, R. T., Pow, N., Serban, I. V., Charlin, L., Liu, C.-W., and Pineau, J. (2017). Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Mayor, A., Alegria, I., De Ilarraza, A. D., Labaka, G., Lersundi, M., and Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for basque. *Machine translation*, 25(1):53.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

San Vicente, I., Manterola, I., et al. (2012). Paco2: A fully automated tool for gathering parallel corpora from the web. In *Language Resources and Evaluation (LREC)*, pages 1–6.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Association for Computational Linguistics*.

Stroppa, N., Groves, D., Way, A., and Sarasola, K. (2006). Example-based machine translation of the basque language.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wang, Z., Hamza, W., and Florian, R. (2017). Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.