The Impact of Preprocessing on Arabic-English Statistical and Neural Machine Translation

Mai Oudah, Amjad Almahairi[†] and Nizar Habash
Computational Approaches to Modeling Language Lab
New York University Abu Dhabi, UAE
[†]Element AI, Canada
{mai.oudah,nizar.habash}@nyu.edu
amjad.almahairi@elementai.com

Abstract

Neural networks have become the state-of-the-art approach for machine translation (MT) in many languages. While linguistically-motivated tokenization techniques were shown to have significant effects on the performance of statistical MT, it remains unclear if those techniques are well suited for neural MT. In this paper, we systematically compare neural and statistical MT models for Arabic-English translation on data preprecossed by various prominent tokenization schemes. Furthermore, we consider a range of data and vocabulary sizes and compare their effect on both approaches. Our empirical results show that the best choice of tokenization scheme is largely based on the type of model and the size of data. We also show that we can gain significant improvements using a system selection that combines the output from neural and statistical MT.

1 Introduction

Neural machine translation (NMT) has been rapidly attracting the attention of the research community for its promising results (Cho et al., 2014b; Bahdanau et al., 2014; Wu et al., 2016; Vaswani et al., 2017). NMT is composed of two neural networks, an encoder and a decoder, where the encoder is fed a sentence from the source language and the decoder generates its translation, word by word, in the target language. Recently, NMT

has been shown to outperform other MT systems in many language pairs, e.g. German-English, French-English and Basque-English (Escolano et al., 2017; Dahlmann et al., 2017; Unanue et al., 2018). While Arabic MT has been mostly developed under statistical MT (SMT), NMT has also been applied and studied recently (Habash and Sadat, 2006; Almahairi et al., 2016; Durrani et al., 2017).

Linguistically-motivated tokenization has shown to have a significant effect on SMT, particularly in the case of morphologically rich languages like Arabic (Habash and Sadat, 2006). However, it remains unclear if such techniques are well suited for NMT, where language-agnostic tokenizations, e.g. byte-pair encoding (BPE) (Sennrich et al., 2016), are widely used. Almahairi et al. (2016) has looked into Arabic SMT and NMT, achieving the highest accuracy using the Penn Arabic Treebank (ATB) tokenization, with 51.2 and 49.7 BLEU points for SMT and NMT, respectively.

In this paper, we study the impact of different preprocessing techniques in Arabic-English MT on both SMT and NMT, by examining various prominent tokenization schemes. We conduct learning curve experiments to study the interaction between data size and the choice of tokenization scheme. We study the performance under morphology-based and frequency-based tokenization schemes, provided by MADAMIRA (Pasha et al., 2014) and BPE, respectively, on in-domain data. In addition, we evaluate the best performing models on out-of-domain data. Our results show that the utilization of BPE for SMT can be effective and allows achieving a good performance even

with a small vocabulary size of 20K. Moreover, the results show that the performance of NMT is especially sensitive to the size of data. We notice that NMT suffers with long sentences, and thus, we utilize system selection, which yields significant improvements over both approaches. Our best system significantly outperforms previous results reported on the same in-domain test data by +4 BLEU points (Almahairi et al., 2016).

The rest of the paper is organized as follows. The related work is presented in Section 2. Section 3 describes our proposed approach. Section 4 illustrates the experimental settings. The results are reported in Section 5. In Section 6, we discuss our findings. Finally, we conclude the paper and mention the future work in Section 7.

2   Related Work

Many studies have compared the performance of different MT models on translation tasks (Unanue et al., 2018; Almahairi et al., 2016; Durrani et al., 2017). However, the data preprocessing was not unified across those models. For example, BPE is only applied to the training data utilized by the NMT system, but not SMT (Almahairi et al., 2016; Durrani et al., 2017).

Habash and Sadat (2006) investigated and compared across some preprocessing schemes for Arabic, describing and evaluating different methods for combining them. The main preprocessing schemes were Simple Tokenization, Decliticization (degrees 1 to 3), and Arabic Treebank Tokenization. Decliticization of degree 2 outperformed the rest when applied individually. They reported improvement in MT performance when combining different schemes together.

Almahairi et al. (2016) compared NMT and SMT on Arabic translation, and showed that NMT performs comparably to SMT. The best performance is achieved when Penn Arabic Treebank (ATB) tokenization is used with 51.19 and 49.70 BLEU points for SMT and NMT, respectively.

The idea of system selection for MT exists in the literature, but mostly for model selection under the same approach (SMT or NMT) (Devlin and Matsoukas, 2012; Salloum et al., 2014).

3   Approach

In our study, we systematically compare SMT and NMT on the following dimensions.

3.1   Source Language Tokenization

Much research has shown the importance of tokenization and orthographic normalization for SMT and NMT, as they deal with data sparsity (El Kholy and Habash, 2012; Habash and Sadat, 2006; Zalmout and Habash, 2017). Tokenization schemes can either be morphology-based or statistical/frequency-based (Pasha et al., 2014; Sennrich et al., 2016). We investigate both in the context of Arabic MT, both separately and in combination, to observe their interaction. We normalize Alif 'ا' and Ya 'ي' in all schemes, where Hamza is removed from the variants of Hamzated Alif (e.g.'أ','إ') to become 'ا', the Alif Maqsura 'ى' is replaced with Ya 'ي' and the diacritics are removed.

Morphology-based This tokenization scheme relies on the linguistic rules of the source language. We explore three schemes under this category (Habash and Sadat, 2006; Zalmout and Habash, 2017): 1) Simple Tokenization (Raw) that splits off punctuation and numbers; 2) Penn Arabic Treebank (ATB) Tokenization, which splits all clitics except definite articles; 3) Decliticization (D3), which splits all clitics.

Frequency-based We use byte-pair encoding (BPE) (Sennrich et al., 2016), which is an iterative compression approach that replaces the most frequent pair of characters in a sentence with a unique sequence of characters. It allows for a fixed-size vocabulary representation. Figure 1 shows an example across Raw and Tok schemes with/without BPE on top.

3.2   Training Data Size

We conduct a learning curve experiment to explore how much both Arabic-English SMT and NMT can benefit from adding more training data with each tokenization scheme. Habash and Sadat (2006) have conducted a similar learning curve study for SMT. Each tokenization scheme may result in a different number of tokens per sentence; hence, a sentence-length filter will discard more sentences from more verbose schemes. This would lead to some schemes having access to more words than others. Therefore, we adopt El Kholy and Habash (2012)'s approach of filtering training parallel data based on the D3 scheme as a reference scheme for selecting sentences of length up to 100

Setting	Sentence
Original	قال أحمد أبو حميدة أنّنا نرتاح اليوم ونوزع المهام على الأفراد.
Raw	قال احمد ابو حميدة اننا نرتاح اليوم و نوزع المهام علي الافراد .
ATB	قال احمد ابو حميدة ان +نا نرتاح اليوم و+ نوزع المهام علي الافراد .
D3	قال احمد ابو حميدة ان +نا نرتاح ال+ يوم و+ نوزع ال+ مهام علي ال+ افراد .
Raw+BPE	قال ا حمد ابو حمي@@ دة اننا نر@@ تاح اليوم ونو@@ زع المهام علي الافراد .
ATB+BPE	قال احمد ابو حمي@@ دة ان +نا ن@@ ر@@ تاح اليوم و+ نو@@ زع المهام علي الافراد .
D3+BPE	قال احمد ابو حميدة ان +نا ن@@ رت@@ اح ال+ يوم و+ نو@@ زع ال+ مهام علي ال+ افراد .
Translation	Today we are resting and distributing the new posts, said Ahmad Abou Hamida.

Figure 1: Tokenization schemes applied to an example.

tokens. Thus, the same sentences will be selected across different tokenization schemes.

3.3 Target Language Resources

We design the training so that both systems will have access to the same additional target language resources besides the target side of the training parallel corpus. In SMT, target language resources are used to build language models for fluency improvement. Whereas, many works have proven pretrained word embeddings to be useful in neural network models (Qi et al., 2018), and therefor, the same additional TLR are used to learn pretrained word embeddings that support the decoder in NMT. Here, the $tgt++$ designation next to the system name indicates the use of additional TLR.

3.4 Input Length and System Selection

Many have reported NMT performing worse with long sentences (Cho et al., 2014a; Koehn and Knowles, 2017), which was caught in our error analysis and thus we explored combining the two MT systems via a system selection approach, where the selection of either translation is based on which is closer to the input length as a criterion. Whereas the sentence BLEU score is the criterion in the Oracle system selection.

4 Experimental Settings

4.1 Datasets

The training dataset contains 1.2M sentence pairs in newswire (NW) domain from three Linguistic Data Consortium (LDC) resources: LDC2004T18, LDC2004T14, and LDC2007T08. For tuning, we use LDC2010T12 (MT04), which consists of

1,075 sentence pairs in NW and government documents. As for the in-domain testing, we use LDC2010T14 (MT05), which consists of 1,056 sentence pairs in NW, and has four English reference translations. We look into the performance of the systems in out-of-domain data using LDC2014T02 (MT12), which consists of 1,535 sentence pairs mostly web collection, and has four English reference translations.

4.2 Preprocessing

MADAMIRA (Pasha et al., 2014) is utilized for morphology-based tokenization of the source side. Sennrich et al. (2016)'s BPE implementation is used for learning and applying BPE models. We set vocabulary size to 20K in BPE learning after exploring multiple vocabulary sizes, including 10K, 20K and 30K, where the 20K setting achieved comparable results to the 30K and outperformed the 10K. Each BPE model is trained on source side of training data of the respective experiment. While Moses' (Koehn et al., 2007) tokenizer and lowercaser are used for preparing the target side.

4.3 SMT settings

We use Moses 3.0 (Koehn et al., 2007) to train SMT models with maximum phrase length of 8 tokens. Two versions of the language model are examined: 1) trained solely on the target side of the training dataset, and 2) trained on the target side and the English Gigaword 5th edition.

4.4 NMT settings

We use the encoder-decoder with the *general* global attention architecture as introduced by Lu-
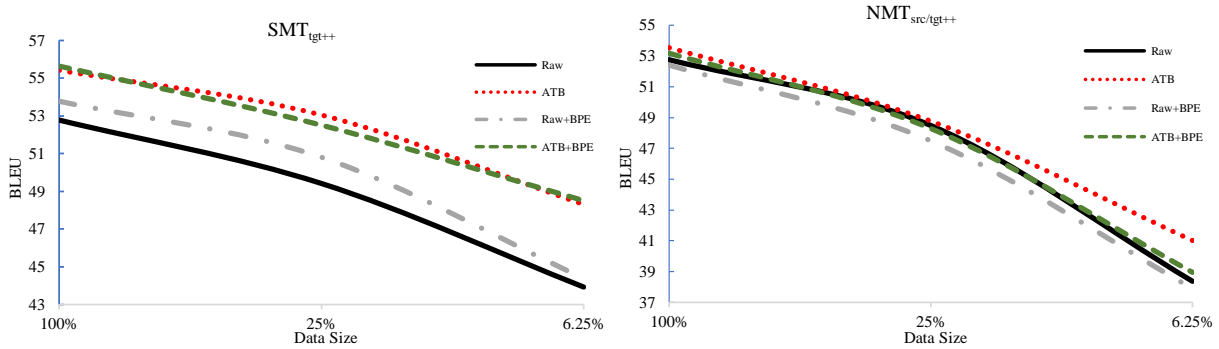
Figure 2: The performance on in-domain test (MT05) under different settings with different training data sizes.

	#Vocab	SMT_{tgt++}	CI	$\text{NMT}_{scr/tgt++}$	CI	P-value
Raw	331K	52.78	$\pm$ 0.98	52.76	$\pm$ 1.24	0.412
ATB	208K	55.42	$\pm$ 1.07	**53.54**	$\pm$ 1.20	0.002
D3	190K	54.66	$\pm$ 1.02	53.51	$\pm$ 1.20	0.027
Raw+BPE	20K	53.78	$\pm$ 1.10	52.41	$\pm$ 1.17	0.003
ATB+BPE	20K	**55.64**	$\pm$ 1.11	53.18	$\pm$ 1.15	0.001
D3+BPE	20K	54.59	$\pm$ 1.07	53.38	$\pm$ 1.16	0.018

Table 1: Comparing Raw, ATB and D3 Tokenized cases without/with BPE on in-domain test (MT05), in terms of BLEU scores, where the Confidence Interval (CI) and P-value are reported. Bold font highlights best results by SMT and NMT.

ong et al. (2015). All the NMT models have been trained using OpenNMT toolkit (Klein et al., 2017) with no restriction on input's vocabulary. We use long short-term memory units (LSTM) (Hochreiter and Schmidhuber, 1997), with hidden units of size 500 and two layers in both the encoder and decoder. The word embedding vector size for source/target is 300.

English pretrained word embeddings were trained as skip-gram model (Mikolov et al., 2013) via gensim tool (Rehurek and Sojka, 2010) with settings: (size=300, window=8, min count=5) on English Gigaword 5th edition (Graff and Cieri, 2003) dataset. Arabic embeddings were trained on the Arabic Gigaword 5th edition (Parker et al., 2011) via FastText (Bojanowski et al., 2017), which showed better performance with morphologically rich languages (Erdmann et al., 2018). We give the designation of $_{src/tgt++}$ to the system that uses both embeddings.

4.5 Evaluation Metrics

The evaluation results are reported in case insensitive BLEU scores (Papineni et al., 2002) with their confidence intervals (CI) and p-values. Bootstrap resampling is used to compute statistical significance intervals (Koehn, 2004).

5 Results

5.1 Preprocessing and Learning Curve

We examine Raw, ATB and D3 with and without BPE applied on top, across a learning curve where smaller sets of our training data (1.2M) are considered at 25% (300K) and 6.25% (75K) tokens. Figure 2 illustrates the learning curve results for Raw (baseline) and ATB (overall best), with and without BPE. Figure 2 shows the importance of training data availability, especially for NMT, and also that BPE impact can be seen in both systems, which we find interesting. Moreover, SMT is shown to be far more sensitive to preprocessing than NMT.

Table 1 shows the best systems' results when 100% of the training data is tokenized by Raw, ATB and D3, with and without BPE on top of it, for SMT and NMT. It shows ATB+BPE and ATB to achieve the best results for SMT and NMT, respectively, which we find interesting as BPE is usually associated with NMT. The p-value indicates whether the difference between SMT and NMT results under the same tokenization scheme is statistically significant or not. The statistical significance is illustrated with p-value < 0.05. SMT_{tgt++} and $\text{NMT}_{scr/tgt++}$ have comparable results at the baseline. As expected, using more data for LM
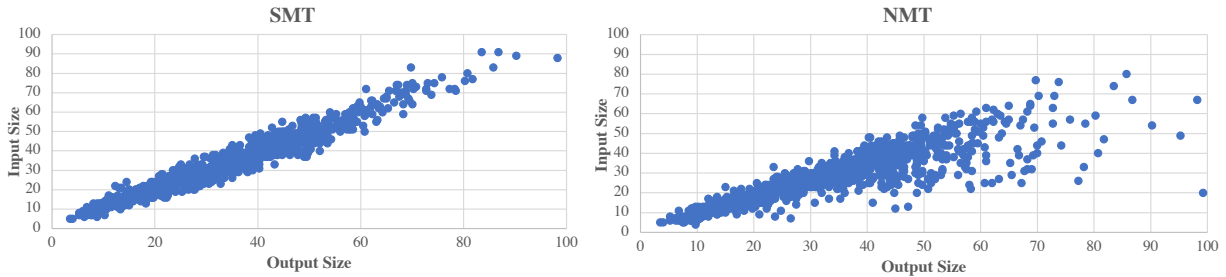
Figure 3: The input size vs. output size in SMT and NMT, respectively, on MT05 with ATB tokenization. We notice that in NMT parts of the input sentences are dropped and not translated at all, which motivates the length-based selection.

$\mathbf{SMT}_{tgt++}$		$\mathbf{NMT}_{scr/tgt++}$		System Selection	Oracle
Setting	BLEU	Scheme	BLEU	BLEU	BLEU
ATB+BPE	55.64	ATB	53.54	**56.18**	61.26

Table 2: BLEU score of the length-based system selection (using best models of SMT and NMT) when applied on in-domain test (MT05).

$\mathbf{SMT}_{tgt++}$		$\mathbf{NMT}_{scr/tgt++}$		System Selection	Oracle
Scheme	BLEU	Scheme	BLEU	BLEU	BLEU
ATB+BPE	35.11	ATB	36.56	**37.96**	39.11

Table 3: BLEU score of the length-based system selection (using best models of SMT and NMT) when applied on out-of-domain test (MT12).

produces better results as well as the increase in training data size. Using pretrained word embeddings for both languages improve the NMT results significantly compared to only target ones by two BLEU points. As shown in Table 3, the best $\mathrm{NMT}_{scr/tgt++}$ model (using ATB) outperforms SMT_{tgt++} model (using ATB+BPE) by 1.5 BLEU against MT12 in the out-of-domain testing.

5.2 Error analysis

Error analysis has shown that NMT output is more fluent than SMT's, especially with short sentences (< 50 tokens), in contrast to long sentences where coverage and accuracy drop, which support related work. Figure 3 shows dropping in NMT output size as the input size increases, especially < 40 tokens, while SMT keeps more consistent output to the input size. So we explore system selection based on the closeness to the input length as well as Oracle results, where the selection is based on the highest output BLEU score. Tables 2 and 3 show the results of length-based system selection on best models in SMT and NMT when applied to in-domain test (MT05) and out-of-domain testing (MT12), respectively, which illustrate im-

provements over the original BLEU scores.

Figure 4 illustrates five examples, where either SMT or NMT output is selected based on the output length compared to the source input size. In Example 1, the SMT output is selected over the NMT one as the NMT system drops the phrase after the comma and only translates the part before, however, fluently. Example 2, which represents much longer sentence, SMT output is selected over NMT, which translates the saying and drops the rest of the sentence. On the other hand, in Examples 3 and 4, NMT output is selected over SMT's for being the closest to the source input in terms of length. Furthermore, Example 5 represents a case where the system selection approach fails to select the better prediction (in terms of BLEU score) for the final output based on the source-output length comparison.

6 Discussion

We notice that morphology-based tokenization schemes improve the performance of MT systems regardless of the MT approach, but in different levels. The difference in scheme choice is less im-

	Example 1
Input	و+ كرر القول ان البرنامج النووي الايراني سلمي و+ ان ايران لا تسعي الي التزود ب+قنبلة ذرية , كما تؤكد الولايات المتحدة .
SMT*	and repeated that iran ' s nuclear program is peaceful and that iran is not seeking to acquire an atomic bomb , as the united states alleges .
NMT	he repeated that iran ' s nuclear program was peaceful and iran was not seeking to acquire atomic bomb .
Human	he repeated that iran ' s nuclear program is peaceful and that iran is not trying to acquire an atomic bomb as the united states claims .
	Example 2
Input	1 - 3 (اف ب) ـ قال الناطق الرسمي ب+اسم وزارة الخارجية الاردنية رجب السقيري ان جميع دول الجوار العراقي س+ تشارك في الاجتماع الذي س+يعقد في عمان الخميس المقبل علي مستوي وزراء الخارجية ب+استثناء ايران التي س+ تشارك ب+وفد يراس +ه مساعد وزير الخارجية للشوون الدولية و+ القانونية .
SMT*	oman 3 - 1 (afp) - the jordanian foreign ministry spokesman recep UNK that all iraqi neighboring countries will participate in the meeting , which will be held in amman next thursday at the level of foreign ministers , with the exception of iran , which would participate with a delegation led by assistant foreign minister UNK international and legal .
NMT	jordan will take part in a meeting of foreign ministers in amman on thursday , except for iran , a foreign ministry spokesman said .
Human	amman 1 - 3 (afp) - jordanian foreign ministry spokesperson rajab sukayri has said that all of iraq ' s neighboring countries will be taking part in thursday ' s meeting in amman at foreign minister level , with the exception of iran which will be represented by a delegation headed by the foreign minister ' s assistant for legal and international affairs .
	Example 3
Input	و+ لم يتسن التحقق من صحة البيان .
SMT	he could not verify the authenticity of the statement .
NMT*	the authenticity of the statement could not be verified .
Human	the authenticity of the statement could not be verified .
	Example 4
Input	و+ تهدف المحادثات النووية السداسية التي توقفت منذ سبتمبر الماضي الي حل القضية النووية سلميا علي شبه الجزيرة الكورية .
SMT	and designed six-party nuclear talks which have been stalled since last september to peacefully solve the nuclear issue on the korean peninsula .
NMT*	the six-party nuclear talks , which have been stalled since last september , are aimed at resolving the nuclear issue peacefully on the korean peninsula .
Human	the six-party nuclear talks , which stopped last september , are aimed at a peaceful settlement of the nuclear issue on the korean peninsula .
	Example 5
Input	و+ حضر الرئيسان السوداني عمر البشير و+ الجنوب افريقي ثابو مبيكي مراسم التوقيع
SMT*	the two presidents attended the omar al-beshir and south african president thabo mbeki the signing ceremony
NMT	sudanese president omar al-beshir and south african president thabo mbeki attended the signing ceremony
Human	the sudanese president umar bashir and the south african president thabo mbeki attended the signing ceremonies

Figure 4: Examples from MT05, with SMT and NMT outputs when ATB is used as a scheme. The * designation next to the system name indicates the decision of the system selection.

pactful on NMT; compared with SMT. The improvement range for NMT is 1.13 BLEU, while for SMT the range is 2.86 BLEU. While Raw results are almost the same for SMT and NMT; ATB improves both NMT and SMT; but the improvement is higher for SMT. Adding BPE helps SMT, while lowering vocabulary size. The effect of BPE on NMT is insignificant, which is a surprising result since BPE is often associated with NMT. Also, we significantly improve on Almahairi et al. (2016)'s results by more than three BLEU points.

Length-based system selection improves over both NMT and SMT results in in-domain and out-of-domain cases, significantly in the later, which indicates a hybrid MT system may be promising. Moreover, the huge jump in performance with Oracle selection shows that there is still room for potential improvement in system designs, for better accuracy and fluency. More TLR allow for better results in MT systems. When both Arabic and English pretrained word embeddings are used, the performance improves by more than two BLEU points compared to English only.

7 Conclusion and Future Work

In this paper, we study the impact of various preprocessing techniques to Arabic-English MT under SMT and NMT, where various prominent tokenization schemes are examined. We conduct a learning curve analysis of the different preprocessing settings with incremental training data size, where ATB scheme performs consistently well along the learning curve. Moreover, we implemented a length-based system selection to deal with NMT's struggle with short sentences, and significant improvements. The empirical results show that the choice of tokenization scheme can be optimized based on the type of model to train and the data available. We also gain significant improvements using length-based system selection that combines the output from neural and statistical MT. Our results significantly outperform the ones reported in the prior work when applied to in-domain test (MT05). As future work, we plan to examine training data of general domain with linguistically-motivated tokenization schemes to study further their impact on NMT under different neural models. Also, exploring sophisticated system selection schemes for potential improvement.

References

Almahairi, Amjad, Kyunghyun Cho, Nizar Habash, and Aaron Courville. 2016. First result on Arabic neural machine translation. *arXiv preprint arXiv:1606.02680*.

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

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cho, Kyunghyun, Bart Van, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8 Eighth Workshop on Syntax Semantics and Structure in Statistical Translation*, pages 103–111. Association for Computational Linguistics.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics.

Dahlmann, Leonard, Evgeny Matusov, Pavel Petrushkov, and Shahram Khadivi. 2017. Neural machine translation leveraging phrase-based models in a hybrid search. *CoRR*.

Devlin, Jacob and Spyros Matsoukas. 2012. Trait-based hypothesis selection for machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT 12, pages 528–532. Association for Computational Linguistics.

Durrani, Nadir, Fahim Dalvi, Hassan Sajjad, and Stephan Vogel. 2017. Qcri machine translation systems for iwslt 16. *CoRR*.

El Kholy, Ahmed and Nizar Habash. 2012. Orthographic and morphological processing for English–Arabic statistical machine translation. *Machine Translation*, 26(1-2):25–45.

Erdmann, Alexander, Nasser Zalmout, and Nizar Habash. 2018. Addressing noise in multidialectal word embeddings. In *Proceedings of Conference of the Association for Computational Linguistics*, Melbourne, Australia.

Escolano, Carlos, Marta Costa-jussa, and Jose Fonollosa. 2017. The talp-upc neural machine translation system for german/finnish-english using the inverse direction model in rescoring. In *Proceedings of the Second Conference on Machine Translation*, pages 283–287. Association for Computational Linguistics.

Graff, David and Christopher Cieri. 2003. English gigaword, ldc catalog no ldc2003t05. Linguistic Data Consortium, University of Pennsylvania.

Habash, Nizar and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *HLT-NAACL*.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017 System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39. Association for Computational Linguistics.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, Barcelona, Spain.

Luong, Thang, Hieu Pham, and Christopher Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics".

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Papineni, Kishore, 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Parker, Robert, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2011. Arabic Gigaword Fifth Edition. LDC catalog number No. LDC2011T11, ISBN 1-58563-595-2.

Pasha, Arfath, Mohamed Al-Badrashiny, Ahmed El Kholy, Ramy Eskander, Mona Diab, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *In Proceedings of LREC*.

Qi, Ye, Devendra Singh, Matthieu Felix, Sarguna Janani, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? *CoRR*.

Rehurek, Radim and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Salloum, Wael, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence level dialect identification for machine translation system selection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 772–778.

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

Unanue, Inigo, Lierni Arratibel, Ehsan Borzeshi, and Massimo Piccardi. 2018. English-basque statistical and neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, and Wolfgang Macherey. 2016. Googles neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Zalmout, Nasser and Nizar Habash. 2017. Optimizing Tokenization Choice for Machine Translation across Multiple Target Languages. *The Prague Bulletin of Mathematical Linguistics*, 108:257–270, June.