Six Challenges for Neural Machine Translation

Philipp Koehn
Computer Science Department
Johns Hopkins University
phi@jhu.edu

Rebecca Knowles
Computer Science Department
Johns Hopkins University
rknowles@jhu.edu

Abstract

We explore six challenges for neural machine translation: domain mismatch, amount of training data, rare words, long sentences, word alignment, and beam search. We show both deficiencies and improvements over the quality of phrase-based statistical machine translation.

1 Introduction

Neural machine translation has emerged as the most promising machine translation approach in recent years, showing superior performance on public benchmarks (Bojar et al., 2016) and rapid adoption in deployments by, e.g., Google (Wu et al., 2016), Systran (Crego et al., 2016), and WIPO (Junczys-Dowmunt et al., 2016). But there have also been reports of poor performance, such as the systems built under low-resource conditions in the DARPA LORELEI program.[1]

In this paper, we examine a number of challenges to neural machine translation (NMT) and give empirical results on how well the technology currently holds up, compared to traditional statistical machine translation (SMT).

We find that:

1. NMT systems have lower quality **out of domain**, to the point that they completely sacrifice adequacy for the sake of fluency.

2. NMT systems have a steeper learning curve with respect to the **amount of training data**, resulting in worse quality in low-resource settings, but better performance in high-resource settings.

3. NMT systems that operate at the sub-word level (e.g. with byte-pair encoding) perform better than SMT systems on extremely **low-frequency words**, but still show weakness in translating low-frequency words belonging to highly-inflected categories (e.g. verbs).

4. NMT systems have lower translation quality on very **long sentences**, but do comparably better up to a sentence length of about 60 words.

5. The attention model for NMT does not always fulfill the role of a **word alignment model**, but may in fact dramatically diverge.

6. **Beam search decoding** only improves translation quality for narrow beams and deteriorates when exposed to a larger search space.

We note a 7th challenge that we do not examine empirically: NMT systems are much less interpretable. The answer to the question of why the training data leads these systems to decide on specific word choices during decoding is buried in large matrices of real-numbered values. There is a clear need to develop better analytics for NMT.

Other studies have looked at the comparable performance of NMT and SMT systems. Bentivogli et al. (2016) considered different linguistic categories for English–German and Toral and Sánchez-Cartagena (2017) compared different broad aspects such as fluency and reordering for nine language directions.

2 Experimental Setup

We use common toolkits for neural machine translation (Nematus) and traditional phrase-based statistical machine translation (Moses) with common data sets, drawn from WMT and OPUS.

[1] https://www.nist.gov/itl/iad/mig/lorehlt16-evaluations

Proceedings of the First Workshop on Neural Machine Translation, pages 28–39,
Vancouver, Canada, August 4, 2017.

2.1 Neural Machine Translation

While a variety of neural machine translation approaches were initially proposed — such as the use of convolutional neural networks (Kalchbrenner and Blunsom, 2013) — practically all recent work has been focused on the attention-based encoder-decoder model (Bahdanau et al., 2015).

We use the toolkit Nematus[2] (Sennrich et al., 2017) which has been shown to give state-of-the-art results (Sennrich et al., 2016a) at the WMT 2016 evaluation campaign (Bojar et al., 2016).

Unless noted otherwise, we use default settings, such as beam search and single model decoding. The training data is processed with byte-pair encoding (Sennrich et al., 2016b) into subwords to fit a 50,000 word vocabulary limit.

2.2 Statistical Machine Translation

Our machine translation systems are trained using Moses[3] (Koehn et al., 2007). We build phrase-based systems using standard features that are commonly used in recent system submissions to WMT (Williams et al., 2016; Ding et al., 2016a).

While we use the shorthand SMT for these phrase-based systems, we note that there are other statistical machine translation approaches such as hierarchical phrase-based models (Chiang, 2007) and syntax-based models (Galley et al., 2004, 2006) that have been shown to give superior performance for language pairs such as Chinese–English and German–English.

2.3 Data Conditions

We carry out our experiments on English–Spanish and German–English. For these language pairs, large training data sets are available. We use datasets from the shared translation task organized alongside the Conference on Machine Translation (WMT)[4]. For the domain experiments, we use the OPUS corpus[5] (Tiedemann, 2012).

Except for the domain experiments, we use the WMT test sets composed of news stories, which are characterized by a broad range of topic, formal language, relatively long sentences (about 30 words on average), and high standards for grammar, orthography, and style.

[2] https://github.com/rsennrich/nematus/
[3] http://www.stat.org/moses/
[4] http://www.statmt.org/wmt17/
[5] http://opus.lingfil.uu.se/

Corpus	Words	Sentences	W/S
Law (Acquis)	18,128,173	715,372	25.3
Medical (EMEA)	14,301,472	1,104,752	12.9
IT	3,041,677	337,817	9.0
Koran (Tanzil)	9,848,539	480,421	20.5
Subtitles	114,371,754	13,873,398	8.2

Table 1: Corpora used to train domain-specific systems, taken from the OPUS repository. IT corpora are GNOME, KDE, PHP, Ubuntu, and OpenOffice.

3 Challenges

3.1 Domain Mismatch

A known challenge in translation is that in different domains,[6] words have different translations and meaning is expressed in different styles. Hence, a crucial step in developing machine translation systems targeted at a specific use case is domain adaptation. We expect that methods for domain adaptation will be developed for NMT. A currently popular approach is to train a general domain system, followed by training on in-domain data for a few epochs (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016).

Often, large amounts of training data are only available out of domain, but we still seek to have robust performance. To test how well NMT and SMT hold up, we trained five different systems using different corpora obtained from OPUS (Tiedemann, 2012). An additional system was trained on all the training data. Statistics about corpus sizes are shown in Table 1. Note that these domains are quite distant from each other, much more so than, say, Europarl, TED Talks, News Commentary, and Global Voices.

We trained both SMT and NMT systems for all domains. All systems were trained for German-English, with tuning and test sets sub-sampled from the data (these were not used in training). A common byte-pair encoding is used for all training runs.

See Figure 1 for results. While the in-domain NMT and SMT systems are similar (NMT is better for IT and Subtitles, SMT is better for Law, Medical, and Koran), the out-of-domain performance for the NMT systems is worse in almost all cases, sometimes dramatically so. For instance the Med-

[6] We use the customary definition of domain in machine translation: a *domain* is defined by a corpus from a specific source, and may differ from other *domains* in topic, genre, style, level of formality, etc.

System ↓	Law		Medical		IT		Koran		Subtitles	
All Data	30.5	32.8	45.1	42.2	35.3	44.7	17.9	17.9	26.4	20.8
Law	31.1	34.4	12.1	18.2	3.5	6.9	1.3	2.2	2.8	6.0
Medical	3.9	10.2	39.4	43.5	2.0	8.5	0.6	2.0	1.4	5.8
IT	1.9	3.7	6.5	5.3	42.1	39.8	1.8	1.6	3.9	4.7
Koran	0.4	1.8	0.0	2.1	0.0	2.3	15.9	18.8	1.0	5.5
Subtitles	7.0	9.9	9.3	17.8	9.2	13.6	9.0	8.4	25.9	22.1

Figure 1: Quality of systems (BLEU), when trained on one domain (rows) and tested on another domain (columns). Comparably, NMT systems (left bars) show more degraded performance out of domain.

ical system leads to a BLEU score of 3.9 (NMT) vs. 10.2 (SMT) on the Law test set.

Figure 2 displays an example. When translating the sentence *Schaue um dich herum.* (reference: *Look around you.*) from the Subtitles corpus, we see mostly non-sensical and completely unrelated output from the NMT system. For instance, the translation from the IT system is *Switches to paused.*

Note that the output of the NMT system is often quite fluent (e.g., *Take heed of your own souls.*) but completely unrelated to the input, while the SMT output betrays its difficulties with coping with the out-of-domain input by leaving some words untranslated (e.g., *Schaue by dich around.*). This is of particular concern when MT is used for information gisting — the user will be mislead by hallucinated content in the NMT output.

3.2 Amount of Training Data

A well-known property of statistical systems is that increasing amounts of training data lead to better results. In SMT systems, we have previously observed that doubling the amount of training data gives a fixed increase in BLEU scores. This holds true for both parallel and monolingual data (Turchi et al., 2008; Irvine and Callison-Burch, 2013).

Source	Schaue um dich herum.
Ref.	Look around you.
All	NMT: Look around you. SMT: Look around you.
Law	NMT: Sughum gravecorn. SMT: In order to implement dich Schaue .
Medical	NMT: EMEA / MB / 049 / 01-EN-Final Work progamme for 2002 SMT: Schaue by dich around .
IT	NMT: Switches to paused. SMT: To Schaue by itself . \t \t
Koran	NMT: Take heed of your own souls. SMT: And you see.
Subtitles	NMT: Look around you. SMT: Look around you .

Figure 2: Examples for the translation of a sentence from the Subtitles corpus, when translated with systems trained on different corpora. Performance out-of-domain is dramatically worse for NMT.

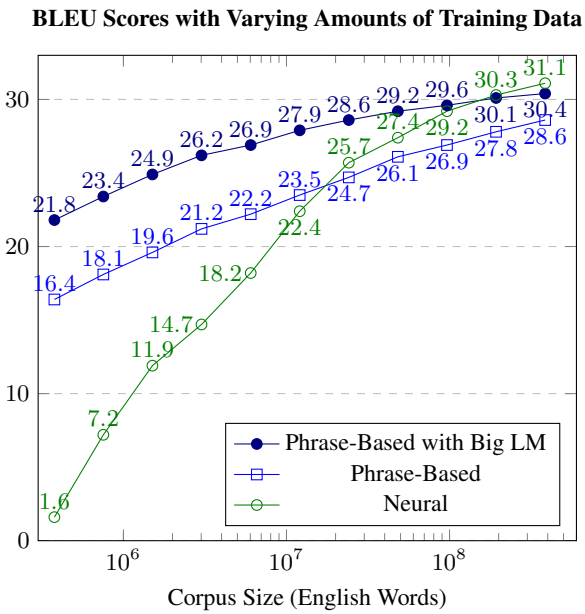

Figure 3: BLEU scores for English-Spanish systems trained on 0.4 million to 385.7 million words of parallel data. Quality for NMT starts much lower, outperforms SMT at about 15 million words, and even beats a SMT system with a big 2 billion word in-domain language model under high-resource conditions.

Src:	A Republican strategy to counter the re-election of Obama
$\frac{1}{1024}$	Un órgano de coordinación para el anuncio de libre determinación
$\frac{1}{512}$	Lista de una estrategia para luchar contra la elección de hojas de Ohio
$\frac{1}{256}$	Explosión realiza una estrategia divisiva de luchar contra las elecciones de autor
$\frac{1}{128}$	Una estrategia republicana para la eliminación de la reelección de Obama
$\frac{1}{64}$	Estrategia siria para contrarrestar la reelección del Obama .
$\frac{1}{32}+$	Una estrategia republicana para contrarrestar la reelección de Obama

Figure 4: Translations of the first sentence of the test set using NMT system trained on varying amounts of training data. Under low resource conditions, NMT produces fluent output unrelated to the input.

How do the data needs of SMT and NMT compare? NMT promises both to generalize better (exploiting word similary in embeddings) and condition on larger context (entire input and all prior output words).

We built English-Spanish systems on WMT data,[7] about 385.7 million English words paired with Spanish. To obtain a learning curve, we used $\frac{1}{1024}$, $\frac{1}{512}$, ..., $\frac{1}{2}$, and all of the data. For SMT, the language model was trained on the Spanish part of each subset, respectively. In addition to a NMT and SMT system trained on each subset, we also used all additionally provided monolingual data for a big language model in contrastive SMT systems.

Results are shown in Figure 3. NMT exhibits a much steeper learning curve, starting with abysmal results (BLEU score of 1.6 vs. 16.4 for $\frac{1}{1024}$ of the data), outperforming SMT 25.7 vs. 24.7 with $\frac{1}{16}$ of the data (24.1 million words), and even beating the SMT system with a big language model with the full data set (31.1 for NMT, 28.4 for SMT, 30.4 for SMT+BigLM).

The contrast between the NMT and SMT learning curves is quite striking. While NMT is able to exploit increasing amounts of training data more effectively, it is unable to get off the ground with training corpus sizes of a few million words or less.

To illustrate this, see Figure 4. With $\frac{1}{1024}$ of the training data, the output is completely unrelated to the input, some key words are properly translated with $\frac{1}{512}$ and $\frac{1}{256}$ of the data (*estrategia* for *strategy*, *elección* or *elecciones* for *election*), and starting with $\frac{1}{64}$ the translations become respectable.

3.3 Rare Words

Conventional wisdom states that neural machine translation models perform particularly poorly on rare words, (Luong et al., 2015; Sennrich et al., 2016b; Arthur et al., 2016) due in part to the smaller vocabularies used by NMT systems. We examine this claim by comparing performance on rare word translation between NMT and SMT systems of similar quality for German–English and find that NMT systems actually outperform SMT systems on translation of very infrequent words. However, both NMT and SMT systems do continue to have difficulty translating some infrequent words, particularly those belonging to highly-inflected categories.

For the neural machine translation model, we use a publicly available model[8] with the training settings of Edinburgh's WMT submission (Sennrich et al., 2016a). This was trained using Ne-

[7] Spanish was last represented in 2013, we used data from http://statmt.org/wmt13/translation-task.html

[8] https://github.com/rsennrich/wmt16-scripts/

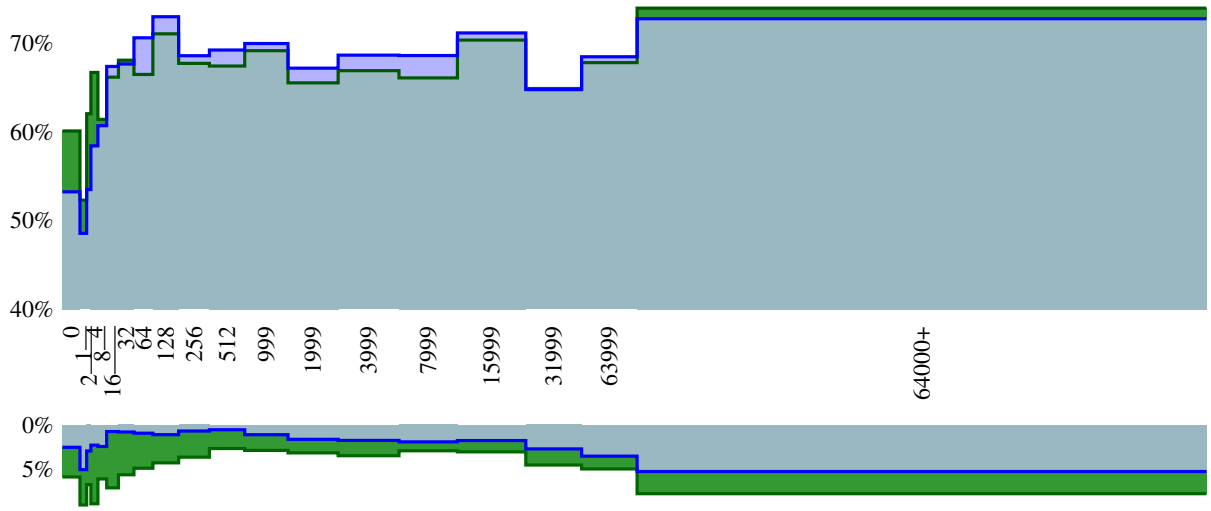

Figure 5: Precision of translation and deletion rates by source words type. SMT (light blue) and NMT (dark green). The horizontal axis represents the corpus frequency of the source types, with the axis labels showing the upper end of the bin. Bin width is proportional to the number of word types in that frequency range. The upper part of the graph shows the precision averaged across all word types in the bin. The lower part shows the proportion of source tokens in the bin that were deleted.

matus[9] (Sennrich et al., 2017), with byte-pair encodings (Sennrich et al., 2016b) to allow for open-vocabulary NMT.

The phrase-based model that we used was trained using Moses (Koehn et al., 2007), and the training data and parameters match those described in Johns Hopkins University's submission to the WMT shared task (Ding et al., 2016b).

Both models have case-sensitive BLEU scores of 34.5 on the WMT 2016 news test set (for the NMT model, this reflects the BLEU score resulting from translation with a beam size of 1). We use a single corpus for computing our lexical frequency counts (a concatenation of Common Crawl, Europarl, and News Commentary).

We follow the approach described by Koehn and Haddow (2012) for examining the effect of source word frequency on translation accuracy.[10]

The overall average precision is quite similar between the NMT and SMT systems, with the SMT system scoring 70.1% overall and the NMT system scoring 70.3%. This reflects the similar overall quality of the MT systems. Figure 5 gives a detailed breakdown. The values above the horizontal axis represent precisions, while the lower portion represents what proportion of the words were deleted. The first item of note is that the NMT system has an overall higher proportion of deleted words. Of the 64379 words examined, the NMT system is estimated to have deleted 3769 of them, while the SMT system deleted 2274. Both the NMT and SMT systems delete very frequent and very infrequent words at higher proportions than words that fall into the middle range. Across frequencies, the NMT systems delete a higher proportion of words than the SMT system does. (The related issue of translation length is discussed in more detail in Section 3.4.)

The next interesting observation is what happens with unknown words (words which were never observed in the training corpus). The SMT system translates these correctly 53.2% of the time, while the NMT system translates them correctly 60.1% of the time. This is reflected in Figure 5, where the SMT system shows a steep curve

[9] `https://github.com/rsennrich/nematus/`

[10] First, we automatically align the source sentence and the machine translation output. We use fast-align (Dyer et al., 2013) to align the full training corpus (source and reference) along with the test source and MT output. We use the suggested standard options for alignment and then symmetrize the alignment with grow-diag-final-and.

Each source word is either unaligned ("dropped") or aligned to one or more target language words. For each target word to which the source word is aligned, we check if that target word appears in the reference translation. If the target word appears the same number of times in the MT output as in the reference, we award that alignment a score of one. If the target word appears more times in the MT output than in the reference, we award fractional credit. If the target word does not appear in the reference, we award zero credit. We then average these scores over the full set of target words aligned to the given source word to compute the precision for that source word. Source words can then be binned by frequency and average translation precisions can be computed.

Label	Unobserved	Observed Once
Adjective	4	10
Named Entity	40	42
Noun	35	35
Number	12	4
Verb	3	6
Other	6	3

Table 2: Breakdown of the first 100 tokens that were unobserved in training or observed once in training, by hand-annotated category.

up from the unobserved words, while the NMT system does not see a great jump.

Both SMT and NMT systems actually have their worst performance on words that were observed a single time in the training corpus, dropping to 48.6% and 52.2%, respectively; even worse than for unobserved words. Table 2 shows a breakdown of the categories of words that were unobserved in the training corpus or observed only once. The most common categories across both are named entity (including entity and location names) and nouns. The named entities can often be passed through unchanged (for example, the surname "Elabdellaoui" is broken into "E@@ lab@@ d@@ ell@@ a@@ oui" by the byte-pair encoding and is correctly passed through unchanged by both the NMT and SMT systems). Many of the nouns are compound nouns; when these are correctly translated, it may be attributed to compound-splitting (SMT) or byte-pair encoding (NMT). The factored SMT system also has access to the stemmed form of words, which can also play a similar role to byte-pair encoding in enabling translation of unobserved inflected forms (e.g. adjectives, verbs). Unsurprisingly, there are many numbers that were unobserved in the training data; these tend to be translated correctly (with occasional errors due to formatting of commas and periods, resolvable by post-processing).

The categories which involve more extensive inflection (adjectives and verbs) are arguably the most interesting. Adjectives and verbs have worse accuracy rates and higher deletion rates than nouns across most word frequencies. We show examples in Figure 6 of situations where the NMT system succeeds and fails, and contrast it with the failures of the SMT system. In Example 1, the NMT system successfully translates the unobserved adjective *choreographiertes* (choreographed), while the SMT system does not. In Example 2, the SMT system simply passes the German verb

Src.	(1) ... **choreographiertes** Gesamtkunstwerk ... (2) ... die Polizei ihn **einkesselte**.
BPE	(1) **chore@@ ograph@@ iertes** (2) **ein@@ kes@@ sel@@ te**
NMT	(1) ... **choreographed** overall artwork ... (2) ... police **stabbed** him.
SMT	(1) ... **choreographiertes** total work of art ... (2) ... police **einkesselte** him.
Ref.	(1) ... **choreographed** complete work of art ... (2) ... police **closed in on** him.

Figure 6: Examples of words that were unobserved in the training corpus, their byte-pair encodings, and their translations.

einkesselte (closed in on) unchanged into the output, while the NMT system fails silently, selecting the fluent-sounding but semantically inappropriate "stabbed" instead.

While there remains room for improvement, NMT systems (at least those using byte-pair encoding) perform better on very low-frequency words then SMT systems do. Byte-pair encoding is sometimes sufficient (much like stemming or compound-splitting) to allow the successful translation of rare words even though it does not necessarily split words at morphological boundaries. As with the fluent-sounding but semantically inappropriate examples from domain-mismatch, NMT may sometimes fail similarly when it encounters unknown words even in-domain.

3.4 Long Sentences

A well-known flaw of early encoder-decoder NMT models was the inability to properly translate long sentences (Cho et al., 2014; Pouget-Abadie et al., 2014). The introduction of the attention model remedied this problem somewhat. But how well?

We used the large English-Spanish system from the learning curve experiments (Section 3.2), and used it to translate a collection of news test sets from the WMT shared tasks. We broke up these sets into buckets based on source sentence length (1-9 subword tokens, 10-19 subword tokens, etc.) and computed corpus-level BLEU scores for each.

Figure 7 shows the results. While overall NMT is better than SMT, the SMT system outperforms NMT on sentences of length 60 and higher. Quality for the two systems is relatively close, except for the very long sentences (80 and more tokens). The quality of the NMT system is dramatically lower for these since it produces too short translations (length ratio 0.859, opposed to 1.024).

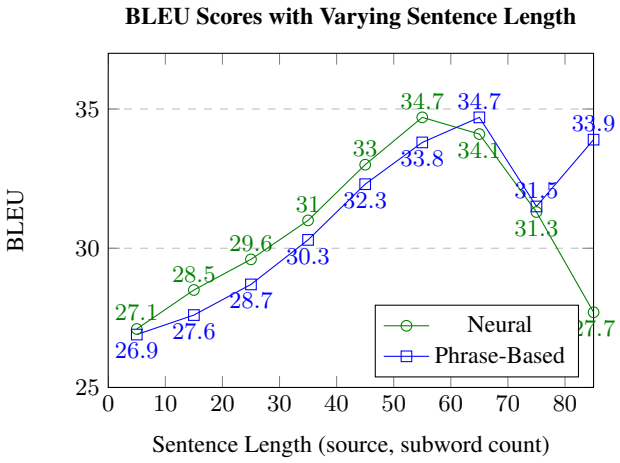

Figure 7: Quality of translations based on sentence length. SMT outperforms NMT for sentences longer than 60 subword tokens. For very long sentences (80+) quality is much worse due to too short output.

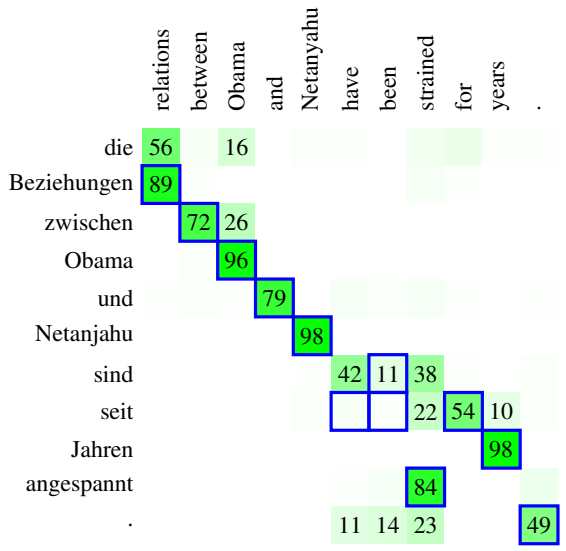

Figure 8: Word alignment for English–German: comparing the attention model states (green boxes with probability in percent if over 10) with alignments obtained from fast-align (blue outlines).

3.5 Word Alignment

The key contribution of the attention model in neural machine translation (Bahdanau et al., 2015) was the imposition of an alignment of the output words to the input words. This takes the shape of a probability distribution over the input words which is used to weigh them in a bag-of-words representation of the input sentence.

Arguably, this attention model does not functionally play the role of a word alignment between the source in the target, at least not in the same way as its analog in statistical machine translation. While in both cases, alignment is a latent variable that is used to obtain probability distributions over words or phrases, arguably the attention model has a broader role. For instance, when translating a verb, attention may also be paid to its subject and object since these may disambiguate it. To further complicate matters, the word representations are products of bidirectional gated recurrent neural networks that have the effect that each word representation is informed by the entire sentence context.

But there is a clear need for an alignment mechanism between source and target words. For instance, prior work used the alignments provided by the attention model to interpolate word translation decisions with traditional probabilistic dictionaries (Arthur et al., 2016), for the introduction of coverage and fertility models (Tu et al., 2016), etc.

But is the attention model in fact the proper means? To examine this, we compare the soft alignment matrix (the sequence of attention vectors) with word alignments obtained by traditional word alignment methods. We use incremental fast-align (Dyer et al., 2013) to align the input and output of the neural machine system.

See Figure 8 for an illustration. We compare the word attention states (green boxes) with the word alignments obtained with fast align (blue outlines). For most words, these match up pretty well. Both attention states and fast-align alignment points are a bit fuzzy around the function words *have-been/sind*.

However, the attention model may settle on alignments that do not correspond with our intuition or alignment points obtained with fast-align. See Figure 9 for the reverse language direction, German–English. All the alignment points appear to be off by one position. We are not aware of any intuitive explanation for this divergent behavior — the translation quality is high for both systems.

We measure how well the soft alignment (attention model) of the NMT system match the alignments of fast-align with two metrics:

- a **match score** that checks for each output if the aligned input word according to fast-align is indeed the input word that received the highest attention probability, and
- a **probability mass score** that sums up the

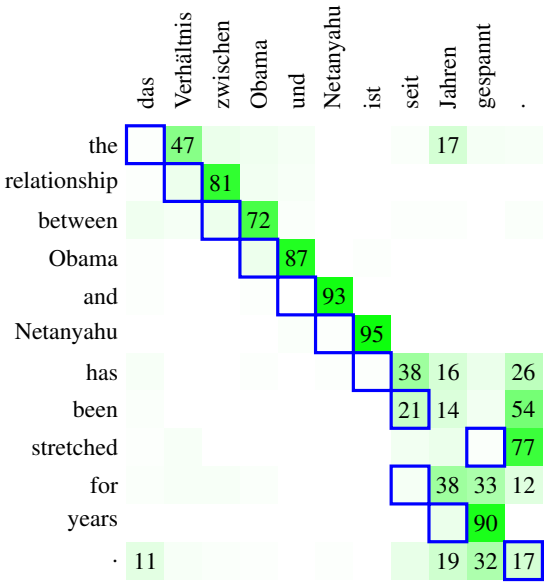

Figure 9: Mismatch between attention states and desired word alignments (German–English).

probability mass given to each alignment point obtained from fast-align.

In these scores, we have to handle byte pair encoding and many-to-many alignments[11]

In out experiment, we use the neural machine translation models provided by Edinburgh[12] (Sennrich et al., 2016a). We run fast-align on the same parallel data sets to obtain alignment models and used them to align the input and output of the NMT system. Table 3 shows alignment scores for the systems. The results suggest that, while drastic, the divergence for German–English is an outlier. We note, however, that we have seen such large a divergence also under different data conditions.

Note that the attention model may produce better word alignments by guided alignment training (Chen et al., 2016; Liu et al., 2016) where supervised word alignments (such as the ones produced by fast-align) are provided to model training.

[11] (1) NMT operates on subwords, but fast-align is run on full words. (2) If an input word is split into subwords by byte pair encoding, then we add their attention scores. (3) If an output word is split into subwords, then we take the average of their attention vectors. (4) The match scores and probability mass scores are computed as average over output word-level scores. (5) If an output word has no fast-align alignment point, it is ignored in this computation. (6) If an output word is fast-aligned to multiple input words, then (6a) for the match score: count it as correct if the n aligned words among the top n highest scoring words according to attention and (6b) for the probability mass score: add up their attention scores.

[12] `https://github.com/rsennrich/wmt16-scripts`

Language Pair	Match	Prob.
German–English	14.9%	16.0%
English–German	77.2%	63.2%
Czech–English	78.0%	63.3%
English–Czech	76.1%	59.7%
Russian–English	72.5%	65.0%
English–Russian	73.4%	64.1%

Table 3: Scores indicating overlap between attention probabilities and alignments obtained with fast-align.

3.6 Beam Search

The task of decoding is to find the full sentence translation with the highest probability. In statistical machine translation, this problem has been addressed with heuristic search techniques that explore a subset of the space of possible translation. A common feature of these search techniques is a beam size parameter that limits the number of partial translations maintained per input word.

There is typically a straightforward relationship between this beam size parameter and the model score of resulting translations and also their quality score (e.g., BLEU). While there are diminishing returns for increasing the beam parameter, typically improvements in these scores can be expected with larger beams.

Decoding in neural translation models can be set up in similar fashion. When predicting the next output word, we may not only commit to the highest scoring word prediction but also maintain the next best scoring words in a list of partial translations. We record with each partial translation the word translation probabilities (obtained from the softmax), extend each partial translation with subsequent word predictions and accumulate these scores. Since the number of partial translation explodes exponentially with each new output word, we prune them down to a beam of highest scoring partial translations.

As in traditional statistical machine translation decoding, increasing the beam size allows us to explore a larger set of the space of possible translation and hence find translations with better model scores.

However, as Figure 10 illustrates, increasing the beam size does not consistently improve translation quality. In fact, in almost all cases, worse translations are found beyond an optimal beam size setting (we are using again Edinburgh's WMT

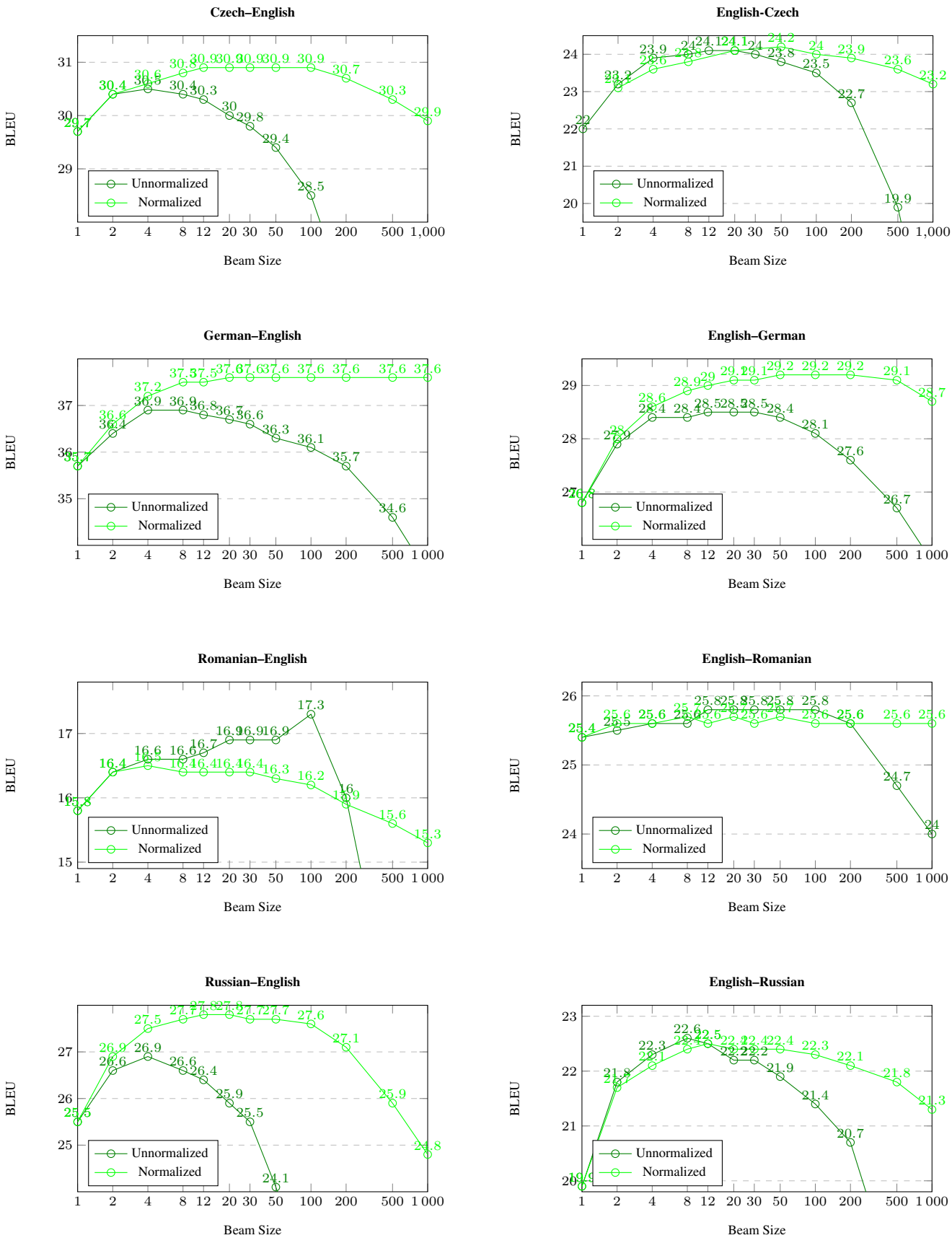

Figure 10: Translation quality with varying beam sizes. For large beams, quality decreases, especially when not normalizing scores by sentence length.

2016 systems). The optimal beam size varies from 4 (e.g., Czech–English) to around 30 (English–Romanian).

Normalizing sentence level model scores by length of the output alleviates the problem somewhat and also leads to better optimal quality in most cases (5 of the 8 language pairs investigated). Optimal beam sizes are in the range of 30–50 in almost all cases, but quality still drops with larger beams. The main cause of deteriorating quality are shorter translations under wider beams.

4 Conclusions

We showed that, despite its recent successes, neural machine translation still has to overcome various challenges, most notably performance out-of-domain and under low resource conditions. We hope that this paper motivates research to address these challenges.

What a lot of the problems have in common is that the neural translation models do not show robust behavior when confronted with conditions that differ significantly from training conditions — may it be due to limited exposure to training data, unusual input in case of out-of-domain test sentences, or unlikely initial word choices in beam search. The solution to these problems may hence lie in a more general approach of training that steps outside optimizing single word predictions given perfectly matching prior sequences.

Acknowledgment

This work was partially supported by a Amazon Research Award (to the first author) and a National Science Foundation Graduate Research Fellowship under Grant No. DGE-1232825 (to the second author).

References

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1557–1567. https://aclweb.org/anthology/D16-1162.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*. http://arxiv.org/pdf/1409.0473v6.pdf.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 257–267. https://aclweb.org/anthology/D16-1025.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198. http://www.aclweb.org/anthology/W/W16/W16-2301.

Wenhu Chen, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter. 2016. Guided alignment training for topic-aware neural machine translation. *CoRR* abs/1607.01628. http://arxiv.org/abs/1607.01628.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics* 33(2). http://www.aclweb.org/anthology-new/J/J07/J07-2003.pdf.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. 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*. Association for Computational Linguistics, Doha, Qatar, pages 103–111. http://www.aclweb.org/anthology/W14-4012.

Josep Maria Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. 2016. Systran's pure neural machine translation systems. *CoRR* abs/1610.05540. http://arxiv.org/abs/1610.05540.

Shuoyang Ding, Kevin Duh, Huda Khayrallah, Philipp Koehn, and Matt Post. 2016a. The jhu machine translation systems for wmt 2016. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 272–280. http://www.aclweb.org/anthology/W/W16/W16-2310.

Shuoyang Ding, Kevin Duh, Huda Khayrallah, Philipp Koehn, and Matt Post. 2016b. The JHU machine translation systems for WMT 2016. In *Proceedings of the First Conference on Machine Translation (WMT)*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, pages 644–648. http://www.aclweb.org/anthology/N13-1073.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897* .

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sydney, Australia, pages 961–968. http://www.aclweb.org/anthology/P/P06/P06-1121.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*. http://www.aclweb.org/anthology/N04-1035.pdf.

Ann Irvine and Chris Callison-Burch. 2013. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, pages 262–270. http://www.aclweb.org/anthology/W13-2233.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*. http://workshop2016.iwslt.org/downloads/IWSLT_2016_paper_4.pdf.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1700–1709. http://www.aclweb.org/anthology/D13-1176.

Philipp Koehn and Barry Haddow. 2012. Interpolated backoff for factored translation models. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej 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*. Association for Computational Linguistics, Prague, Czech Republic, pages 177–180. http://www.aclweb.org/anthology/P/P07/P07-2045.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Neural machine translation with supervised attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 3093–3102. http://aclweb.org/anthology/C16-1291.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 11–19. http://www.aclweb.org/anthology/P15-1002.

Jean Pouget-Abadie, Dzmitry Bahdanau, Bart van Merrienboer, KyungHyun Cho, and Yoshua Bengio. 2014. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. *CoRR* abs/1409.1257. http://arxiv.org/abs/1409.1257.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, pages 65–68. http://aclweb.org/anthology/E17-3017.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation (WMT)*. Association for Computational Linguistics, Berlin, Germany, pages 371–376.

http://www.aclweb.org/anthology/W/W16/W16-2323.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*. Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. http://www.aclweb.org/anthology/P16-1162.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 1063–1073. http://www.aclweb.org/anthology/E17-1100.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 76–85. http://www.aclweb.org/anthology/P16-1008.

Marco Turchi, Tijl De Bie, and Nello Cristianini. 2008. Learning performance of a machine translation system: a statistical and computational analysis. In *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Columbus, Ohio, pages 35–43. http://www.aclweb.org/anthology/W/W08/W08-0305.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, Barry Haddow, and Ondřej Bojar. 2016. Edinburgh's statistical machine translation systems for wmt16. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 399–410. http://www.aclweb.org/anthology/W/W16/W16-2327.

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, Greg 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. http://arxiv.org/abs/1609.08144.pdf.