

Chapter 8

Current Challenges

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

Here, we examine a number of challenges to neural machine translation and give empirical results on how well the technology currently holds up, compared to traditional statistical machine translation. We show 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.

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.

Another challenge that we do not examine empirically: neural machine translation 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 neural machine translation.

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. Unless noted otherwise, we use default settings, such as beam search and single model

¹<https://www.nist.gov/itl/iad/mig/lorehlt16- evaluations>

decoding. The training data is processed with byte-pair encoding (Sennrich et al., 2016c) into subwords to fit a 50,000 word vocabulary limit.

Our statistical machine translation systems are trained using Moses² (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., 2016). While we consider here only 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.

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)³. For the domain experiments, we use the OPUS corpus⁴ (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.

8.1 Domain Mismatch

A known challenge in translation is that in different domains,⁵ 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 neural machine translation. 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 neural machine translation and statistical machine translation 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 8.1. 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 statistical machine translation and neural machine translation systems for all domains. All systems were trained for German-English, with tuning and test sets subsampled from the data (these were not used in training). A common byte-pair encoding is used for all training runs.

²<http://www.stat.org/moses/>

³<http://www.statmt.org/wmt17/>

⁴<http://opus.lingfil.uu.se/>

⁵We 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.

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



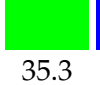
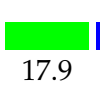




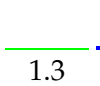




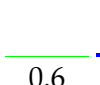




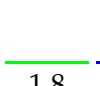



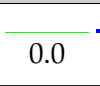
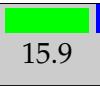



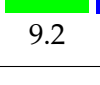
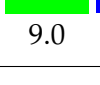

System ↓	Law	Medical	IT	Koran	Subtitles
All Data	 30.5 32.8	 45.1 42.2	 35.3 44.7	 17.9 17.9	 26.4 20.8
Law	 31.1 34.4	 12.1 18.2	 3.5 6.9	 1.3 2.2	 2.8 6.0
Medical	 3.9 10.2	 39.4 43.5	 2.0 8.5	 0.6 2.0	 1.4 5.8
IT	 1.9 3.7	 6.5 5.3	 42.1 39.8	 1.8 1.6	 3.9 4.7
Koran	 0.4 1.8	 0.0 2.1	 0.0 2.3	 15.9 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 8.1: Quality of systems (BLEU), when trained on one domain (rows) and tested on another domain (columns). Comparably, neural machine translation systems (left bars) show more degraded performance out of domain.

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

Figure 8.2: 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 neural machine translation.

See Figure 8.1 for results. While the in-domain neural and statistical machine translation systems are similar (neural machine translation is better for IT and Subtitles, statistical machine translation is better for Law, Medical, and Koran), the out-of-domain performance for the neural machine translation systems is worse in almost all cases, sometimes dramatically so. For instance the Medical system leads to a BLEU score of 3.9 (neural machine translation) vs. 10.2 (statistical machine translation) on the Law test set.

Figure 8.2 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 neural machine translation system. For instance, the translation from the IT system is *Switches to paused.*

Note that the output of the neural machine translation system is often quite fluent (e.g., *Take heed of your own souls.*) but completely unrelated to the input, while the statistical machine translation 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 misled by hallucinated content in the neural machine translation output.

8.2 Amount of Training Data

A well-known property of statistical systems is that increasing amounts of training data lead to better results. In statistical machine translation 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).

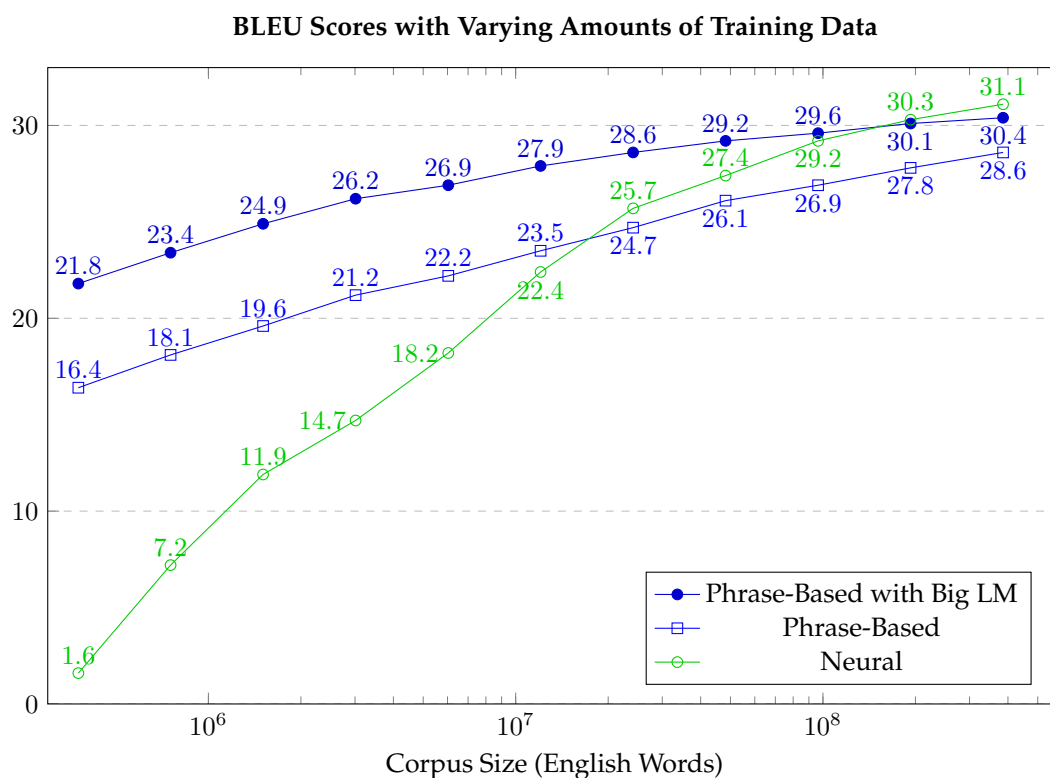


Figure 8.3: BLEU scores for English-Spanish systems trained on 0.4 million to 385.7 million words of parallel data. Quality for neural machine translation starts much lower, outperforms statistical machine translation at about 15 million words, and even beats a statistical machine translation system with a big 2 billion word in-domain language model under high-resource conditions.

Ratio	Words	Source: <i>A Republican strategy to counter the re-election of Obama</i>
$\frac{1}{1024}$	0.4 million	<i>Un órgano de coordinación para el anuncio de libre determinación</i>
$\frac{1}{512}$	0.8 million	<i>Lista de una estrategia para luchar contra la elección de hojas de Ohio</i>
$\frac{1}{256}$	1.5 million	<i>Explosión realiza una estrategia divisiva de luchar contra las elecciones de autor</i>
$\frac{1}{128}$	3.0 million	<i>Una estrategia republicana para la eliminación de la reelección de Obama</i>
$\frac{1}{64}$	6.0 million	<i>Estrategia siria para contrarrestar la reelección del Obama .</i>
$\frac{1}{32} +$	12.0 million	<i>Una estrategia republicana para contrarrestar la reelección de Obama</i>

Figure 8.4: Translations of the first sentence of the test set using neural machine translation system trained on varying amounts of training data. Under low resource conditions, neural machine translation produces fluent output unrelated to the input.

How do the data needs of statistical machine translation and neural machine translation compare? Neural machine translation promises both to generalize better (exploiting word similarity in embeddings) and condition on larger context (entire input and all prior output words).

We built English-Spanish systems on WMT data,⁶ 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 statistical machine translation, the language model was trained on the Spanish part of each subset, respectively. In addition to a neural and statistical machine translation system trained on each subset, we also used all additionally provided monolingual data for a big language model in contrastive statistical machine translation systems.

Results are shown in Figure 8.3. Neural machine translation 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 statistical machine translation 25.7 vs. 24.7 with $\frac{1}{16}$ of the data (24.1 million words), and even beating the statistical machine translation system with a big language model with the full data set (31.1 for neural machine translation, 28.4 for statistical machine translation, 30.4 for statistical with a big language model).

The contrast between the neural and statistical machine translation learning curves is quite striking. While neural machine translation 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 8.4. 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.

8.3 Noisy Data

Statistical machine translation is fairly robust to **noisy data**. The quality of systems holds up fairly well, even if large parts of the training data are corrupted in various ways, such as mis-

⁶Spanish was last represented in 2013, we used data from <http://statmt.org/wmt13/translation-task.html>

Ratio shuffled	0%	10%	20%	50%
SMT (BLEU)	32.7	32.7 (−0.0)	32.6 (−0.1)	32.0 (−0.7)
NMT (BLEU)	35.4	34.8 (−0.6)	32.1 (−3.3)	30.1 (−5.3)

Table 8.2: Impact of noise in the training data, with parts of the training corpus shuffled to contain mis-aligned sentence pairs. Neural machine translation degrades severely, while statistical machine translation holds up fairly well.

aligned sentences, content in wrong languages, badly translated sentences, etc. Statistical machine translation models are built on probability distributions estimated from many occurrences of words and phrases. Any unsystematic noise in the training only affects the tail end of the distribution.

Is this still the case for neural machine translation? Chen et al. (2016a) considered one kind of noise: misaligned sentence pairs in an experiments with a large English–French parallel corpus. They shuffle the target side of part of the training corpus, so that these sentence pairs are mis-aligned.

Table 8.2 shows the result. Statistical machine translation systems hold up fairly well. Even with 50% of the data perturbed, the quality only drops from 32.7 to 32.0 BLEU points, about what is to be expected with half the valid training data. However, the neural machine translation system degrades severely, from 35.4 to 30.1 BLEU points, a drop of 5.3 points, compared to the 0.7 point drop for statistical systems.

A possible explanation for this poor behavior of neural machine translation models is that its prediction has to find a good balance between language model and input context as the main driver. When training observes increasing ratios of training example, for which the input sentence is a meaningless distraction, it may generally learn to rely more on the output language model aspect, hence hallucinating fluent by inadequate output.

8.4 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.

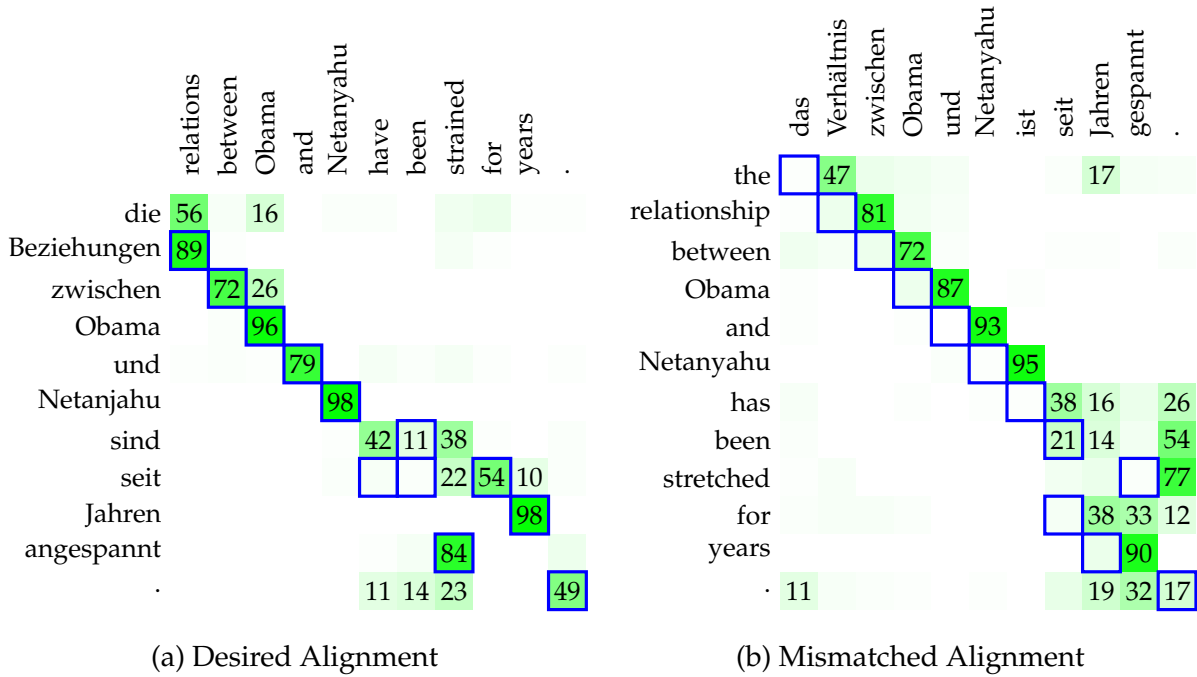


Figure 8.5: Word alignment for English–German: comparing the attention model states (green boxes with probability in percent if over 10) with alignments obtained from fast-align (blue outlines).

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., 2016b), 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 8.5a 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 8.5b 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 neural machine translation system match the alignments of fast-align with two metrics:

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

- 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 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 alignments⁷

In our experiment, we use the neural machine translation models provided by Edinburgh⁸ (Sennrich et al., 2016b). 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 neural machine translation system. Table 8.3 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., 2016b; Liu et al., 2016) where supervised word alignments (such as the ones produced by fast-align) are provided to model training.

8.5 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.

⁷(1) neural machine translation 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.

⁸<https://github.com/rsennrich/wmt16-scripts>

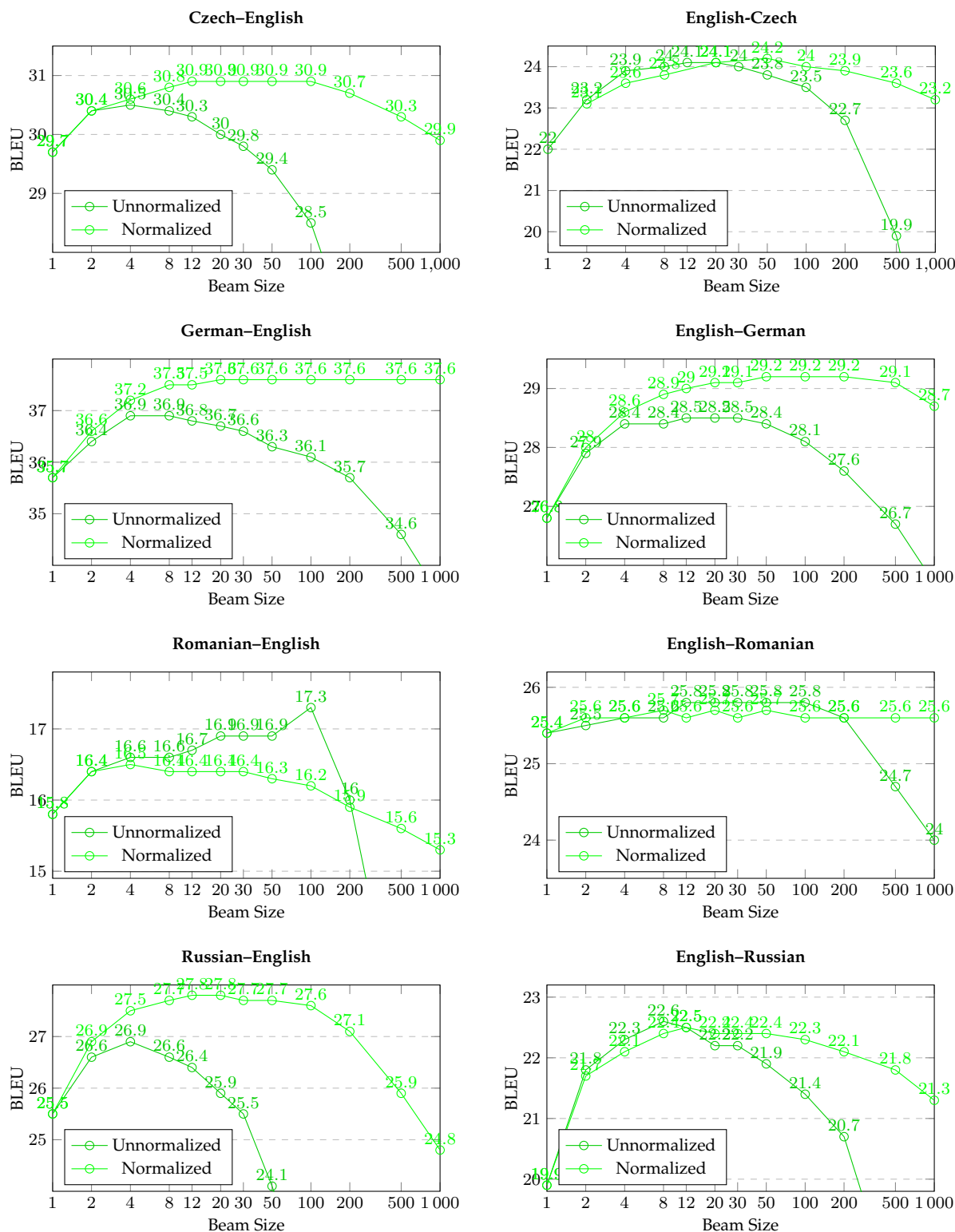


Figure 8.6: Translation quality with varying beam sizes. For large beams, quality decreases, especially when not normalizing scores by sentence length.

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 8.6 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 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.

8.6 Further Readings

Other studies have looked at the comparable performance of neural and statistical machine translation 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.

