On the Impact of Various Types of Noise on Neural Machine Translation

Huda Khayrallah
Center for Language & Speech Processing
Computer Science Department
Johns Hopkins University
huda@jhu.edu

Philipp Koehn
Center for Language & Speech Processing
Computer Science Department
Johns Hopkins University
phi@jhu.edu

Abstract

We examine how various types of noise in the parallel training data impact the quality of neural machine translation systems. We create five types of artificial noise and analyze how they degrade performance in neural and statistical machine translation. We find that neural models are generally more harmed by noise than statistical models. For one especially egregious type of noise they learn to just copy the input sentence.

1 Introduction

While neural machine translation (NMT) has shown large gains in quality over statistical machine translation (SMT) (Bojar et al., 2017), there are significant exceptions to this, such as low resource and domain mismatch data conditions (Koehn and Knowles, 2017).

In this work, we consider another challenge to neural machine translation: noisy parallel data. As a motivating example, consider the numbers in Table 1. Here, we add an equally sized noisy web crawled corpus to high quality training data provided by the shared task of the Conference on Machine Translation (WMT). This addition leads to a 1.2 BLEU point increase for the statistical machine translation system, but degrades the neural machine translation system by 9.9 BLEU.

The maxim *more data is better* that holds true for statistical machine translation does seem to come with some caveats for neural machine translation. The added data cannot be too noisy. But what kind of noise harms neural machine translation models?

	NMT	SMT
WMT17	27.2	24.0
+ noisy corpus	17.3 (–9.9)	25.2 (+1.2)

Table 1: Adding noisy web crawled data (raw data from paracrawl.eu) to a WMT 2017 German–English statistical system obtains small gains (+1.2 BLEU), a neural system falls apart (–9.9 BLEU).

In this paper, we explore several types of noise and assess their impact by adding synthetic noise to an existing parallel corpus. We find that for almost all types of noise, neural machine translation systems are harmed more than statistical machine translation systems. We discovered that one type of noise, copied source language segments, has a catastrophic impact on neural machine translation quality, leading it to learn a copying behavior that it then exceedingly applies.

2 Related Work

There is a robust body of work on filtering out noise in parallel data. For example: Taghipour et al. (2011) use an outlier detection algorithm to filter a parallel corpus; Xu and Koehn (2017) generate synthetic noisy data (inadequate and nonfluent translations) and use this data to train a classifier to identify good sentence pairs from a noisy corpus; and Cui et al. (2013) use a graph-based random walk algorithm and extract phrase pair scores to weight the phrase translation probabilities to bias towards more trustworthy ones.

Most of this work was done in the context of statistical machine translation, but more recent work (Carpuat et al., 2017) targets neural models. That work focuses on identifying semantic differences in translation pairs using cross-lingual textual entailment and additional length-based features, and demonstrates that removing such sentences improves neural machine translation performance.

Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 74–83
Melbourne, Australia, July 20, 2018. ©2018 Association for Computational Linguistics

As Rarrick et al. (2011) point out, one problem of parallel corpora extracted from the web is translations that have been created by machine translation. Venugopal et al. (2011) propose a method to watermark the output of machine translation systems to aid this distinction. Antonova and Misyurev (2011) report that rule-based machine translation output can be detected due to certain word choices, and statistical machine translation output due to lack of reordering.

In 2016, a shared task on sentence pair filtering was organized[1] (Barbu et al., 2016), albeit in the context of cleaning translation memories which tend to be cleaner than web crawled data. This year, a shared task is planned for the type of noise that we examine in this paper.[2]

Belinkov and Bisk (2017) investigate noise in neural machine translation, but they focus on creating systems that can *translate* the kinds of orthographic errors (typos, misspellings, etc.) that humans can comprehend. In contrast, we address noisy *training* data and focus on types of noise occurring in web-crawled corpora.

There is a rich literature on data selection which aims at sub-sampling parallel data relevant for a task-specific machine translation system (Axelrod et al., 2011). van der Wees et al. (2017) find that the existing data selection methods developed for statistical machine translation are less effective for neural machine translation. This is different from our goals of handling noise since those methods tend to discard perfectly fine sentence pairs (say, about cooking recipes) that are just not relevant for the targeted domain (say, software manuals). Our work is focused on noise that is harmful for all domains.

Since we begin with a clean parallel corpus and potentially noisy data to it, this work can be seen as a type of data augmentation. Sennrich et al. (2016a) incorporate monolingual corpora into NMT by first translating it using an NMT system trained in the opposite direction. While such a corpus has the potential to be noisy, the method is very effective. Currey et al. (2017) create additional parallel corpora by copying monolingual corpora in the target language into the source, and find it improves over back-translation for some language pairs. Fadaee et al. (2017) improve NMT performance in low-resource settings by altering existing sentences to create training data that includes rare words in different contexts.

Type of Noise	Count
Okay	23%
Misaligned sentences	41%
Third language	3%
Both English	10%
Both German	10%
Untranslated sentences	4%
Short segments ($\leq$2 tokens)	1%
Short segments (3–5 tokens)	5%
Non-linguistic characters	2%

Table 2: Noise in the raw Paracrawl corpus.

3 Real-World Noise

What types of noise are prevalent in crawled web data? We manually examined 200 sentence pairs of the above-mentioned Paracrawl corpus and classified them into several error categories. Obviously, the results of such a study depend very much on how crawling and extraction is executed, but the results (see Table 2) give some indication of what noise to expect.

We classified any pairs of German and English sentences that are not translations of each other as misaligned sentences. These may be caused by any problem in alignment processes (at the document level or the sentence level), or by forcing the alignment of content that is not indeed parallel. Such misaligned sentences are the biggest source of error (41%).

There are three types of wrong language content (totaling 23%): one or both sentences may be in a language different from German and English (3%), both sentences may be German (10%), or both languages may be English (10%).

4% of sentence pairs are untranslated, i.e., source and target are identical. 2% sentence pairs consist of random byte sequences, only HTML markup, or Javascript. A number of sentence pairs have very short German or English sentences, containing at most 2 tokens (1%) or 5 tokens (5%).

Since it is a very subjective value judgment what constitutes disfluent language, we do not classify these as errors. However, consider the following sentence pairs that we did count as okay, although they contain mostly untranslated names and numbers.

[1] NLP4TM 2016: rgcl.wlv.ac.uk/nlp4tm2016/shared-task

[2] statmt.org/wmt18/parallel-corpus-filtering.html

DE: Anonym 2 24.03.2010 um 20:55 314 Kommentare
EN: Anonymous 2 2010-03-24 at 20:55 314 Comments

DE: &lt; &lt; erste &lt; zurück Seite 3 mehr letzte &gt; &gt;
EN: &lt; &lt; first &lt; prev. page 3 next last &gt; &gt;

At first sight, some types of noise seem to be easier to automatically identify than others. However, consider, for instance, content in a wrong language. While there are established methods for language identification (typically based on character n-grams), these do not work well on a sentence-level basis, especially for short sentences. Or, take the apparently obvious problem of untranslated sentences. If they are completely identical, that is easy to spot — although even those may have value, such as the list of country names which are often spelled identical in different languages. However, there are many degrees of near-identical content of unclear utility.

4 Types of Noise

The goal of this paper is not to develop methods to detect noise but to ascertain the impact of different types of noise on translation quality when present in parallel data. We hope that our findings inform future work on parallel corpus cleaning.

We now formally define five types of naturally occurring noise and describe how we simulate them. By creating artificial noisy data, we avoid the hard problem of detecting specific types of noise but are still able to study their impact.

Misaligned Sentences As shown above, a common source of noise in parallel corpora is faulty document or sentence alignment. This results in sentences that are not matched to their translation. Such noise is rare in corpora such as Europarl where strong clues about debate topics and speaker turns reduce the scale of the task of alignment to paragraphs, but more common in the alignment of less structured web sites. We artificially create misaligned sentence data by randomly shuffling the order of sentences on one side of the original clean parallel training corpus.

Misordered Words Language may be disfluent in many ways. This may be the product of machine translation, poor human translation, or heavily specialized language use, such as bullet points in product descriptions (recall also the examples above). We consider one extreme case of disfluent language: sentences from the original corpus where the words are reordered randomly. We do this on the source or target side.

Wrong Language A parallel corpus may be polluted by text in a third language, say French in a German–English corpus. This may occur on the source or target side of the parallel corpus. To simulate this, we add French–English (bad source) or German–French (bad target) data to a German–English corpus.

Untranslated Sentences Especially in parallel corpora crawled from the web, there are often sentences that are untranslated from the source in the target. Examples are navigational elements or copyright notices in the footer. Purportedly multi-lingual web sites may be only partially translated, while some original text is copied. Again, this may show up on the source or the target side. We take sentences from either the source or target side of the original parallel corpus and simply copy them to the other side.

Short Segments Sometimes additional data comes in the form of bilingual dictionaries. Can we simply add them as additional sentence pairs, even if they consist of single words or short phrases? We simulate this kind of data by subsubsampling a parallel corpus to include only sentences of maximum length 2 or 5.

5 Experimental Setup

5.1 Neural Machine Translation

Our neural machine translation systems are trained using Marian (Junczys-Dowmunt et al., 2018).[3] We build shallow RNN-based encoder-decoder models with attention (Bahdanau et al., 2015). We train Byte-Pair Encoding segmentation models (BPE) (Sennrich et al., 2016b) with a vocab size of $50,000$ on both sides of the parallel corpus for each experiment. We apply drop-out with 20% probability on the RNNs, and with 10% probability on the source and target words. We stop training after convergence of cross-entropy on the development set, and we average the 4 highest performing models (as determined by development set BLEU performance) to use as an ensemble for decoding (checkpoint assembling). Training of

[3] marian-nmt.github.io

each system takes 2–4 days on a single GPU (GTX 1080ti).

While we focus on RNN-based models with attention as our NMT architecture, we note that different architectures have been proposed, including based on convolutional neural networks (Kalchbrenner and Blunsom, 2013; Gehring et al., 2017) and the self-attention based Transformer model (Vaswani et al., 2017).

5.2 Statistical Machine Translation

Our statistical machine translation systems are trained using Moses (Koehn et al., 2007).[4] We build phrase-based systems using standard features commonly used in recent system submissions to WMT (Haddow et al., 2015; Ding et al., 2016, 2017). We trained our systems with the following settings: a maximum sentence length of 80, grow-diag-final-and symmetrization of GIZA++ alignments, an interpolated Kneser-Ney smoothed 5-gram language model with KenLM (Heafield, 2011), hierarchical lexicalized reordering (Galley and Manning, 2008), a lexically-driven 5-gram operation sequence model (OSM) (Durrani et al., 2013), sparse domain indicator, phrase length, and count bin features (Blunsom and Osborne, 2008; Chiang et al., 2009), a maximum phrase-length of 5, compact phrase table (Junczys-Dowmunt, 2012) minimum Bayes risk decoding (Kumar and Byrne, 2004), cube pruning (Huang and Chiang, 2007), with a stack-size of 1000 during tuning. We optimize feature function weights with k-best MIRA (Cherry and Foster, 2012).

While we focus on phrase based systems as our SMT paradigm, 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 may have better performance in certain language pairs and in low resource conditions.

5.3 Clean Corpus

In our experiments, we translate from German to English. We use datasets from the shared translation task organized alongside the Conference on Machine Translation (WMT)[5] as clean training data. For our baseline we use: Europarl (Koehn, 2005),[6] News Commentary,[7] and the Rapid EU Press Release parallel corpus. The corpus size is about 83 million tokens per language. We use `newstest2015` for tuning SMT systems, `newstest2016` as a development set for NMT systems, and report results on `newstest2017`.

Note that we do not add monolingual data to our systems since this would make our study more complex. So, we always train our language model on the target side of the parallel corpus for that experiment. While using monolingual data for language modelling is standard practice in statistical machine translation, how to use such data for neural models is less obvious.

5.4 Noisy Corpora

For MISALIGNED SENTENCE and MISORDERED WORD noise, we use the clean corpus (above) and perturb the data. To create UNTRANSLATED SENTENCE noise, we also use the clean corpus and create pairs of identical sentences.

For WRONG LANGUAGE noise, we do not have French–English and German–French data of the same size. Hence, we use the EU Bookstore corpus (Skadiņš et al., 2014).[8]

The SHORT SEGMENTS are extracted from OPUS corpora (Tiedemann, 2009, 2012; Lison and Tiedemann, 2016):[9] EMEA (descriptions of medicines),[10] Tanzil (religious text),[11] Open Subtitles 2016,[12] Acquis (legislative text),[13] GNOME (software localization files),[14] KDE (localization files), PHP (technical manual),[15] Ubuntu (localization files),[16] and Open Office.[17] We use only pairs where both the English and German segments are at most 2 or 5 words long. Since this results in small data sets (2 million and 15 tokens per language, respectively), they are duplicated multiple times.

We also show the results for naturally occurring noisy web data from the raw 2016 ParaCrawl corpus (deduplicated raw set).[18]

[4] statmt.org/moses

[5] statmt.org/wmt17/

[6] statmt.org/europarl

[7] casmacat.eu/corpus/news-commentary.html

[8] opus.nlpl.eu/EUbookshop.php

[9] opus.nlpl.eu

[10] emea.europa.eu

[11] tanzil.net/trans

[12] opensubtitles.org

[13] ec.europa.eu/jrc/en/language-technologies/jrc-acquis

[14] l10n.gnome.org

[15] se.php.net/download-docs

[16] translations.launchpad.net

[17] openoffice.org

[18] paracrawl.eu

We sample the noisy corpus in an amount equal to 5%, 10%, 20%, 50%, and 100% of the clean corpus. This reflects the realistic situation where there is a clean corpus, and one would like to add additional data that has the potential to be noisy. For each experiment, we use the target side of the parallel corpus to train the SMT language model, including the noisy text.

6 Impact on Translation Quality

Table 3 shows the effect of adding each type of noise to the clean corpus.[19] For some types of noise NMT is harmed more than SMT: MISMATCHED SENTENCES (up to -1.9 for NMT, -0.6 for SMT), MISORDERED WORDS (source) (-1.7 vs. -0.3), WRONG LANGUAGE (target) (-2.2 vs. -0.6).

SHORT SEGMENTS, UNTRANSLATED SOURCE SENTENCES and WRONG SOURCE LANGUAGE have little impact on either (at most a degradation of -0.7). MISORDERED TARGET WORDS decreases BLEU scores for both SMT and NMT by just over 1 point (100% noise).

The most dramatic difference is UNTRANSLATED TARGET SENTENCE noise. When added at 5% of the original data, it degrades NMT performance by 9.6 BLEU, from 27.2 to 17.6. Adding this noise at 100% of the original data degrades performance by 24.0 BLEU, dropping the score from 27.2 to 3.2. In contrast, the SMT system only drops 2.9 BLEU, from 24.0 to 21.1.

6.1 Copied output

Since the noise type where the target side is a copy of the source has such a big impact, we examine the system output in more detail.

We report the percent of sentences in the evaluation set that are identical to the source for the UNTRANSLATED TARGET SENTENCE and RAW CRAWL data in Figures 1 and 2 (solid bars). The SMT systems output 0 or 1 sentences that are exact copies. However, with just 20% of the UNTRANSLATED TARGET SENTENCE noise, 60% of the NMT output sentences are identical to the source.

This suggests that the NMT systems learn to copy, which may be useful for named entities. However, with even a small amount of this data it is doing far more harm than good.

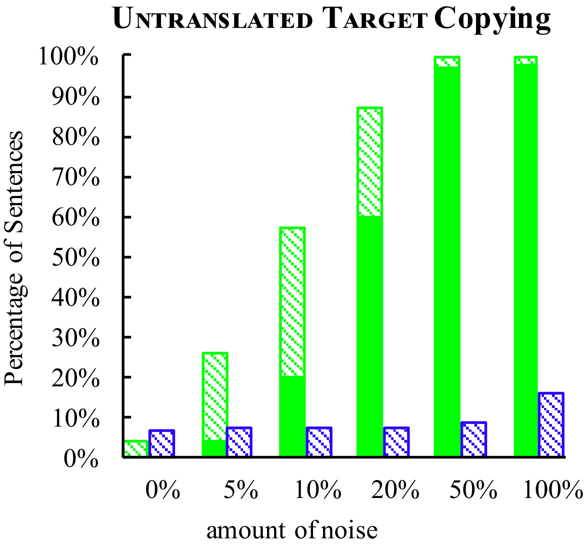

Figure 1: Copied sentences in the UNTRANSLATED (TARGET) experiments. NMT is the left green bars, SMT is the right blue bars. Sentences that are exact matches to the source are the solid bars, sentences that are more similar to the source than the target are the shaded bars.

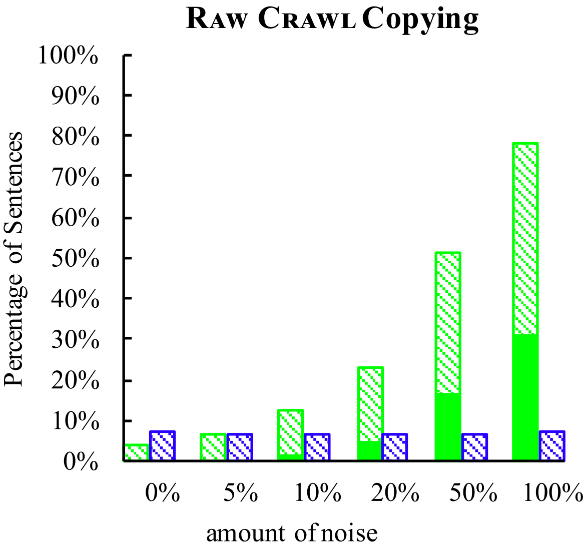

Figure 2: Copied sentences in the RAW CRAWL experiments. NMT is the left green bars, SMT is the right blue bars. Sentences that are exact matches to the source are the solid bars, sentences that are more similar to the source than the target are the shaded bars.

[19] We report case-sensitive detokenized BLEU (Papineni et al., 2002) calculated using mteval-v13a.pl.

	5%		10%		20%		50%		100%	
Misaligned sentences	26.5 -0.7	24.0 -0.0	26.5 -0.7	24.0 -0.0	26.3 -0.9	23.9 -0.1	26.1 -1.1	23.9 -0.1	25.3 -1.9	23.4 -0.6
Misordered words (source)	26.9 -0.3	24.0 -0.0	26.6 -0.6	23.6 -0.4	26.4 -0.8	23.9 -0.1	26.6 -0.6	23.6 -0.4	25.5 -1.7	23.7 -0.3
Misordered words (target)	27.0 -0.2	24.0 -0.0	26.8 -0.4	24.0 -0.0	26.4 -0.8	23.4 -0.6	26.7 -0.5	23.2 -0.8	26.1 -1.1	22.9 -1.1
Wrong language (French source)	26.9 -0.3	24.0 -0.0	26.8 -0.4	23.9 -0.1	26.8 -0.4	23.9 -0.1	26.8 -0.4	23.9 -0.1	26.8 -0.4	23.8 -0.2
Wrong language (French target)	26.7 -0.5	24.0 -0.0	26.6 -0.6	23.9 -0.1	26.7 -0.5	23.8 -0.2	26.2 -1.0	23.5 -0.5	25.0 -2.2	23.4 -0.6
Untranslated (English source)	27.2 -0.0	23.9 -0.1	27.0 -0.2	23.9 -0.1	26.7 -0.5	23.6 -0.4	26.8 -0.4	23.7 -0.3	26.9 -0.3	23.5 -0.5
Untranslated (German target)	17.6 -9.8	23.8 -0.2	11.2 -16.0	23.9 -0.1	5.6 -21.6	23.8 -0.2	3.2 -24.0	23.4 -0.6	3.2 -24.0	21.1 -2.9
Short segments (max 2)	27.1 -0.1	24.1 +0.1	26.5 -0.7	23.9 -0.1	26.7 -0.5	23.8 -0.2				
Short segments (max 5)	27.8 +0.6	24.2 +0.2	27.6 +0.4	24.5 +0.5	28.0 +0.8	24.5 +0.5	26.6 -0.6	24.2 +0.2		
Raw crawl data	27.4 +0.2	24.2 +0.2	26.6 -0.6	24.2 +0.2	24.7 -2.5	24.4 +0.4	20.9 -6.3	24.8 +0.8	17.3 -9.9	25.2 +1.2

Table 3: Results from adding different amounts of noise (ratio of original clean corpus) for various types of noise in German-English Translation. Generally neural machine translation (left green bars) is harmed more than statistical machine translation (right blue bars). The worst type of noise are segments in the source language copied untranslated into the target.

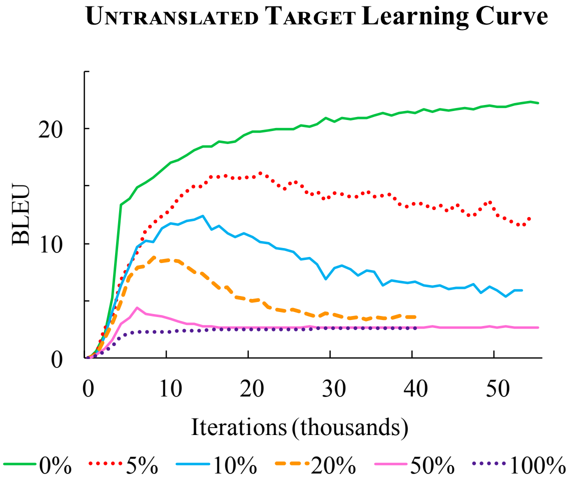

Figure 3: Learning curves for the NMT UNTRANSLATED TARGET SENTENCE experiments.

Figures 1 and 2 show the percent of sentences that have a worse TER score against the reference than against the source (shaded bars). This means that it would take fewer edits to transform the sentence into the source than it would to transform it into the target. When just 10% UNTRANSLATED TARGET SENTENCE data is added, 57% of the sentences are more similar to the source than to the reference, indicating partial copying.

This suggests that the NMT system is overfitting the copied portion of the training corpus. This is supported by Figure 3, which shows the learning curve on the development set for the UNTRANSLATED TARGET SENTENCE noise setup. The performance for the systems trained on noisy corpora begin to improve, before over-fitting to the copy portion of the training set. Note that we plot the BLEU performance on the development set with beam search, while the system is optimizing cross-entropy given a perfect prefix.

Other work has also considered copying in NMT. Currey et al. (2017) add copied data and back-translated data to a clean parallel corpus. They report improvements on EN ↔ RO when adding as much back-translated and copied data as they have parallel (1:1:1 ratio). For EN↔TR and EN↔DE, they add twice as much back translated and copied data as parallel data (1:2:2 ratio), and report improvements on EN↔TR but not on EN↔DE. However, their EN↔DE systems trained with the copied corpus did not perform worse than baseline systems. Ott et al. (2018) found that while copied training sentences represent less than 2.0% of their training data (WMT 14 EN↔DE and EN↔FR), copies are over-represented in the output of beam search. Using a subset of training data from WMT 17, they replace a subset of the true translations with a copy of the input. They analyze varying amounts of copied noise, and a variety of beam sizes. Larger beams are more effected by this kind of noise; however, for all beam sizes performance degrades completely with 50% copied sentences.[20]

6.2 Incorrect Language output

Another interesting case is when a German–French corpus is added to a German–English corpus (WRONG TARGET LANGUAGE). Both neural and statistical machine translation are surprisingly robust, even when these corpora are provided in equal amounts.

We performed a manual analysis of the neural machine translation experiments. For the each of the noise levels, we report the percentage of NMT output sentences in French (out of of 3004: 5%: 0.20%, 10%: 0.60%, 20%: 1.7%, 50%: 3.3%, 100%: 6.7%. Most NMT output sentences were either entirely French or English, with the exception of a few mis-translated cognates (e.g.: 'façade', 'accessibilité').

In the SMT experiment with 100% noisy data added, there are a couple of French words in mostly English sentences. These are much less frequent than unknown German words passed through. Only 1 sentence is mostly French.

It is surprising that such a small percentage of the output sentences were French, since up to half of the target data in training was in French. We attribute this to the domain of the added data differing from the test data. Source sentences in the test set are more similar to the domain-relevant clean parallel training corpus than the domain-divergent noise corpus.

7 Conclusion

We defined five types of noise in parallel data, motivated by a study of raw web crawl data. We found that neural machine translation is less robust to many types of noise than statistical machine translation. In the most extreme case, when the reference is an untranslated copy of the source data, neural machine translation may learn to excessively copy the input. These findings should inform future work on corpus cleaning.

[20] See Figure 3 in Ott et al. (2018).

Acknowledgements

This work has been partially supported by the DARPA LORELEI and the IARPA MATERIAL programs. It also benefited from the Paracrawl web crawling project which is funded by a Google faculty grant and the European Union (CEF).

References

Alexandra Antonova and Alexey Misyurev. 2011. Building a web-based parallel corpus and filtering out machine-translated text. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*. Association for Computational Linguistics, Portland, Oregon, pages 136–144. http://www.aclweb.org/anthology/W11-1218.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., pages 355–362. http://www.aclweb.org/anthology/D11-1033.

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.

Eduard Barbu, Carla Parra Escartín, Luisa Bentivogli, Matteo Negri, Marco Turchi, Constantin Orasan, and Marcello Federico. 2016. The first automatic translation memory cleaning shared task. *Machine Translation* 30(3):145–166. https://doi.org/10.1007/s10590-016-9183-x.

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *CoRR* abs/1711.02173. http://arxiv.org/abs/1711.02173.

Phil Blunsom and Miles Osborne. 2008. Probabilistic inference for machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, pages 215–223. http://www.aclweb.org/anthology/D08-1023.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark, pages 169–214.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, Vancouver, pages 69–79. http://www.aclweb.org/anthology/W17-3209.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, pages 427–436. http://www.aclweb.org/anthology/N12-1047.

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

David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Boulder, Colorado, pages 218–226. http://www.aclweb.org/anthology/N/N09/N09-1025.

Lei Cui, Dongdong Zhang, Shujie Liu, Mu Li, and Ming Zhou. 2013. Bilingual data cleaning for smt using graph-based random walk. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 340–345. http://www.aclweb.org/anthology/P13-2061.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Paper*. Association for Computational Linguistics, Copenhagen, Denmark, pages 148–156. http://www.aclweb.org/anthology/W17-4715.

Shuoyang Ding, Kevin Duh, Huda Khayrallah, Philipp Koehn, and Matt Post. 2016. 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, Huda Khayrallah, Philipp Koehn, Matt Post, Gaurav Kumar, and Kevin Duh. 2017. The jhu machine translation systems for wmt 2017. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared*

Task Papers. Association for Computational Linguistics, Copenhagen, Denmark, pages 276–282. http://www.aclweb.org/anthology/W17-4724.

Nadir Durrani, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013. Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT? In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sofia, Bulgaria.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 567–573. http://aclweb.org/anthology/P17-2090.

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.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, pages 848–856. http://www.aclweb.org/anthology/D08-1089.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 123–135. http://aclweb.org/anthology/P17-1012.

Barry Haddow, Matthias Huck, Alexandra Birch, Nikolay Bogoychev, and Philipp Koehn. 2015. The edinburgh/jhu phrase-based machine translation systems for wmt 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 126–133. http://aclweb.org/anthology/W15-3013.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, pages 187–197. http://www.aclweb.org/anthology/W11-2123.

Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 144–151. http://www.aclweb.org/anthology/P/P07/P07-1019.

M. Junczys-Dowmunt. 2012. A phrase table without phrases: Rank encoding for better phrase table compression. In Mauro Cettolo, Marcello Federico, Lucia Specia, and Andy Way, editors, *Proceedings of th 16th International Conference of the European Association for Machine Translation (EAMT)*. pages 245–252. http://www.mt-archive.info/EAMT-2012-Junczys-Dowmunt.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. *arXiv preprint arXiv:1804.00344* https://arxiv.org/abs/1804.00344.

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. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*. Phuket, Thailand. http://mt-archive.info/MTS-2005-Koehn.pdf.

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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, Vancouver, pages 28–39. http://www.aclweb.org/anthology/W17-3204.

Shankar Kumar and William J. Byrne. 2004. Minimum Bayes-Risk Decoding for Statistical Machine Translation. In *HLT-NAACL*. pages 169–176.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. *CoRR* abs/1803.00047. http://arxiv.org/abs/1803.00047.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Spencer Rarrick, Chris Quirk, and Will Lewis. 2011. MT detection in web-scraped parallel corpora. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*. International Association for Machine Translation, pages 422–430. http://www.mt-archive.info/MTS-2011-Rarrick.pdf.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 86–96. http://www.aclweb.org/anthology/P16-1009.

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*. Association for Computational Linguistics, Berlin, Germany. http://www.aclweb.org/anthology/P16-1162.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the eu bookshop corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

Kaveh Taghipour, Shahram Khadivi, and Jia Xu. 2011. Parallel corpus refinement as an outlier detection algorithm. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*. International Association for Machine Translation, pages 414–421. http://www.mt-archive.info/MTS-2011-Taghipour.pdf.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, volume V.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1411–1421. http://aclweb.org/anthology/D17-1148.

Ashish Vaswani, 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. http://arxiv.org/abs/1706.03762.

Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz Och, and Juri Ganitkevitch. 2011. Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., pages 1363–1372. http://www.aclweb.org/anthology/D11-1126.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2935–2940. http://aclweb.org/anthology/D17-1318.