Extracting In-domain Training Corpora for Neural Machine Translation Using Data Selection Methods

Catarina Cruz Silva
Unbabel
R. Visc. de Santarém 67B,
1000-286 Lisboa, Portugal
catarina@unbabel.com

Chao-Hong Liu, Alberto Poncelas, Andy Way
ADAPT Centre, School of Computing,
Dublin City University
Dublin 9, Ireland
{chaohong.liu, alberto.poncelas,
andy.way}@adaptcentre.ie

Abstract

Data selection is a process used in selecting a subset of parallel data for the training of machine translation (MT) systems, so that 1) resources for training might be reduced, 2) trained models could perform better than those trained with the whole corpus, and/or 3) trained models are more tailored to specific domains. It has been shown that for statistical MT (SMT), the use of data selection helps improve the MT performance significantly. In this study, we reviewed three data selection approaches for MT, namely Term Frequency–Inverse Document Frequency, Cross-Entropy Difference and Feature Decay Algorithm, and conducted experiments on Neural Machine Translation (NMT) with the selected data using the three approaches. The results showed that for NMT systems, using data selection also improved the performance, though the gain is not as much as for SMT systems.

1 Introduction

Data selection is a technology used to improve Machine Translation (MT) performance by choosing a subset of the corpus for the training of MT systems (Chen et al., 2016). There are additional benefits using subsets instead of the whole corpus for MT training. Firstly, the training time could be reduced significantly. In some application scenarios, a much shorter training time would be very useful. Secondly, we could select data with the aim to make trained systems perform well for specific domains. In MT, models built with in-domain data perform better, as the vocabulary and sentence structures used in one domain (e.g. legal) differs from another unrelated domain (e.g. biotechnology).

There are several studies on data selection methods for SMT, showing good improvements over the baselines in which the whole corpora were used for training (Chen et al., 2016). A popular data selection method is cross-entropy difference (CED) (Moore and Lewis, 2010). In particular its bilingual variant (Axelrod et al., 2011) showed a positive impact of data selection for MT.

Term Frequency-Inverse Document Frequency (TF-IDF) (Salton and Yang, 1973) has also been used as a baseline data selection method in the literature. Data selection with cleaning was proposed to improve the robustness of training with divergent sentences (Carpuat et al., 2017).

Feature Decay Algorithms (FDA) are data selection methods that try to extract the subset of sentences by which the coverage of target language features is maximized (Biçici and Yuret, 2011). It has been used to select sentences from parallel data for SMT and NMT (Poncelas et al., 2018) in order to obtain a subset of data that is more tailored to a given test set.

Most of these results focused on comparing training of models from scratch for use in specific domains. The aforementioned papers do not include a focus on the impact of such techniques in fine-tuning the resulting trained model, which could be useful in the case where a baseline model works as an initialization and can be reused for any domain and thus reduce the time required to train the models for specific domains (van der Wees et al., 2017).

In this paper we evaluate the impact of data selection methods on Neural Machine Translation (NMT) systems. We would like to answer the following questions: Do data selection approaches improve domain NMT performance? Which of the three commonly used methods delivers the best results on data selection for NMT? How does the size of the seed and the selected training sentences affect the performance?

The paper is organised as follows. In Section 2, we give an overview of data selection approaches.

Proceedings of the Third Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 224–231

https://doi.org/10.18653/v1/W18-64023

Experimental setup and results are presented in Section 3 and Section 4. Conclusions and future work are given in Section 5.

2 Data Selection Methods

In order to train an MT model for a specific domain, it is best to use those sentences in a data set that are the most related to that domain. We use different data selection techniques to retrieve the sentences. These techniques aim to extract a subset of data from large datasets. The application of these techniques can be used to limit the amount of resource consumption, removing noise and/or adapting the data to a particular domain.

Among different data selection techniques (Eetemadi et al., 2015), in this work, we focus on three particular methods: Cross Entropy Difference (Section 2.1), TF-IDF Data Selection (Section 2.2), and Feature Decay Algorithms (Section 2.3).

2.1 Cross Entropy Difference

The Cross Entropy Difference method was first introduced by (Moore and Lewis, 2010) as a way to build more accurate in-domain Language Models for use in several tasks. The method is a variant of scoring by perplexity, since cross-entropy and perplexity are tightly coupled as shown in 1, where b is the used base.

$$b^{-\sum_x \cdot p(x) \log q(x)} = b^{H(p,q)} \tag{1}$$

Given a general language model LM_G, built with out-of-domain data, and an in-domain language-model LM_D, the method ranks sentences s using the cross-entropy difference in both language models, as in (2):

$$CED(s) = H_D(s) - H_G(s) \tag{2}$$

Although different ranking methods have been introduced, this method still remains popular among data selection approaches, having been used in recent work such as for the selection of monolingual data (Junczys-Dowmunt and Grundkiewicz, 2016), and for the selection of conversational data (Lewis and Federmann, 2015). Some work was also published on the use of neural language models for this purpose, such as Duh et al. (2013), but this applied to Statistical Machine Translation.

In our experiments, we built n-gram language models of order 5 using the KenLM tool[1] (Heafield, 2011). We then use the language model probability scores normalized by sentence length to compute the cross-entropy difference and rank the entire generic corpus.

2.2 TF-IDF data selection

The TF-IDF (Salton and Yang, 1973) method is widely known for its use in several information retrieval applications. It is defined in (3), where $\text{tf}_{t,d}$ is the term frequency in the document, i.e. the ratio between the number of times the term appears in the sentence and the total number of terms, and $\text{idf}_{t,d}$ is the inverse document frequency, the ratio between the total number of documents and the number of documents containing the term.

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \cdot \frac{N}{\text{df}_t} \tag{3}$$

To compute the TF-IDF measure in our experiments, we apply tokenization, remove punctuation and common stopwords in the texts, and finally truecase the sentences. We then consider every sentence in the domain corpus as a query sentence, and every sentence in the generic corpus as a document. Then, we obtain for each query a ranking of the documents, computed with cosine-similarity.

This ranking is stored for every query sentence and used to retrieve the K-nearest neighbours (KNN) necessary to obtain different data selection sizes.

2.3 Feature Decay Algorithms

Feature Decay Algorithms (FDA) (Biçici and Yuret, 2011; Biçici, 2013) are methods of data selection that try to extract, from a set of sentences, those that better represent a seed. It has been used in SMT to extract sentences from parallel corpora in order to obtain a subset of data more adapted to a given test set. These methods select sentences based on two criteria: a) the similarity with the seed (the more sequence of words it shares with the seed the better); and b) the variability of the words (the occurrences of the words shared with the seed should be well distributed, and avoid having too many occurrences of a few words).

These algorithms extract the n-grams from the seed as features. Each feature is assigned an initial value, indicating the relevance of being selected, and the sentences are scored as the normalized sum of values of contained features. Then, the sentences are iteratively selected. Each time a sentence is selected, the values of contained features

[1] https://github.com/kpu/kenlm

are decayed. Accordingly, it promotes selecting features that have not been previously selected in the process.

The decay function is defined in Equation (4):

$$decay(f) = init(f)\frac{d^{C_L(f)}}{(1 + C_L(f))^c} \quad (4)$$

where L is the set of selected sentences and $C_L(f)$ is the count of the feature f in L. $init(f)$ is an initialization function. The variables $d \in (0, 1]$ and $c \in [0, \infty)$ are parameters that regulate how much the value of the feature f should decay. These values are by default (Biçici and Yuret, 2011) 0.5 and 0.0 for d and c, respectively (so, by using default values the decay function in Equation (4) is $decay(f) = init(f)0.5^{C_L(f)}$). There are alternative ways of setting the values (Poncelas et al., 2016, 2017) that can obtain better results. However, in this work we used the default configuration of $d = 0.5$, $c = 0.0$ and used trigrams as features.

3 Experimental Setup

3.1 Data description

For the experiments we use English–French parallel data from two different domains/corpora: EMEA[2] and DGT[3] from the Open Parallel Corpus (OPUS) (Tiedemann, 2009). The first consists of medical data and the second a translation memory in the legal domain. We chose these domains in particular because they are categories more distant from the generic data, which is comprised of news data. The MultiUN corpus (Ziemski et al., 2016) is used for the training of generic models. Moreover, we use only its 6-way subset corpora, to be able to run the experiments in a more comparable setting.

3.2 Seed preparation

Although each data selection method has provided its own approach to select subsets from large corpora, in practice they would better perform if given a good initial subset (i.e. seed) to start with.

To prepare such an initial seed (the same seed is used in the three data selection algorithms), we remove noisy sentences considering punctuation and numerical character. In particular, we remove sentences where:

1. a source (or target) sentence contains fewer than t_{chars} non-punctuation characters,
2. a source (or target) sentence contains fewer than t_{words} words,
3. the source (or target) sentence ratio between punctuation characters and non-punctuation characters is above t_{ratio}.

where t_{chars}, t_{words} and t_{ratio} are thresholds. For both domains and language pairs, t_{chars}=5, t_{words}=2 and t_{ratio}=0.5 are used. We then removed duplicates using the source as reference and compile the remaining sentences into three parts: a validation set (2000 lines); a test set (2000 lines); and the remaining lines comprise the seed domain data. The EMEA domain corpus gave rise to a seed with 238K lines, and the DGT was truncated to a similar size, 250K, to keep experiments comparable.

3.3 Neural Machine Translation

The aim of this work is to assess the impact of data selection techniques on NMT. For this purpose, we use the Marian framework[4] (Junczys-Dowmunt et al., 2018) to train models using the attention-based encoder–decoder architecture as described in Sennrich et al. (2017).

For all experiments a preprocessing routine similar to the one in Moses[5] (Koehn et al., 2007) is used. The preprocessing consists of the following steps: entity replacement (on numbers, emails, urls and alphanumeric entities), tokenisation, truecasing and Byte-Pair Encoding (BPE) (Sennrich et al., 2016) with 89,500 merge operations.

4 Experiments

We present MT results using the three data selection methods and then use the best of the three methods to conduct a series of experiments to assess the impact of data selection on NMT models. We present two evaluation scores, BLEU (Papineni et al., 2002) and Translation Error Rate (TER) (Snover et al., 2006), in the tables. These scores give an estimation of how good the translation is: For BLEU, higher scores indicate better translations, while for TER, as it measures an error rate, lower scores indicate better translation performance.

We performed three different experiments:

- A comparison of the three data selection methods introduced in this paper (Section 4.1).

[2]http://opus.nlpl.eu/EMEA.php
[3]http://opus.nlpl.eu/DGT.php
[4]https://marian-nmt.github.io/
[5]http://www.statmt.org/moses/

	TF-IDF		CED		FDA	
	BLEU ↑	TER ↓	BLEU ↑	TER ↓	BLEU ↑	TER ↓
Seed	.384	.535	.384	.535	.384	.535
+ 240K (1:1)	.417	.506	.409	.513	**.439**	**.487**
+ 480K (1:1)	.433	.497	.422	.497	**.441**	**.484**
+ 480K (2:1)	.453	.470	.443	.483	**.464**	**.467**
+ 1M (1:1)	.443	.477	.433	.493	**.445**	**.476**
+ 1M (4:1)	.466	.470	.456	.473	**.477**	**.457**
+ 2M (1:1)	.449	.479	.440	.483	**.452**	**.469**
+ 2M (8:1)	.488	**.445**	.479	.453	**.491**	.446

Table 1: Results of running three different data selection methods on different selection sizes for EMEA EN→FR. Both BLEU and TER are presented. The top result for each slice of selected data is presented in bold.

- A comparison of the data selection methods using different seeds (Section 4.2).
- The impact of the best data selection method in NMT (Section 4.3)

4.1 Comparison of methods

We start by comparing the three methods for the EMEA domain for English–French. Several experiments are run with different data selection sizes, between 250K and 2M lines, from the MultiUN corpus. We create different sizes of selected data in between these values, corresponding to a factor of 1, 2, 4 and 8 in relation to the size of the original seed. The comparison is not extended to larger selection sizes since a bigger slice, for example 4M, would already represent almost half of the total data available.

Table 1 shows the results of the three methods for models trained from scratch using seed data and different selected data. We present two approaches of combining the data. The first is a simple concatenation of the seed and the selected data. The second tries to balance the seed and the selected data in terms of the number of sentences used for training, by oversampling the seed a number of times such that there are approximately the same number of sentences in the selected data.

Two visible outcomes are shown in these experiments. The first is the overall gain of the Feature Decay Algorithm technique over its two counterparts. For every test (corresponding to a line in the table), the BLEU scores are better using the FDA method, followed by TF-IDF, with the CED method showing lower NMT performance. This result is interesting, since CED is one of the most common used methods for data selection and it has shown good results in several data selection experiments. However, these results are typically related to SMT, and in fact previous work in data selection has shown that these methods do not achieve the same performance for NMT.

The second result is that best performance was obtained when balancing the seed data with the selected data. We use this knowledge to guide the following experiments. Finally, in all experiments TER is also computed, and the results are consistent with those shown in BLEU scores.

4.2 Seed size variation

In previous experiments we used all the domain data available that passed our quality threshold, described in Section 3.2, and selected from the MultiUN corpus, which has little relation to the domain data. We conduct further experiments to analyse whether the previous results are dependent on the initial seed size and also to what extend the seed size impacts or limits the data selection gains.

We start with a seed of about 240K lines. To study the impact of the seed size we retrieve two subsets from the original seed with 50K lines and 100K lines. For each subset, we randomly sample the amount of lines from the original seed three different times and keep only the best subset, where the quality is evaluated by running a baseline MT experiment. Taking advantage of this preliminary experiment, we guarantee that the seed we choose from is not the worst to start with, increasing the reliability of these experiments.

Regarding our first goal, we can conclude that the previous results are not dependent on the initial seed size, from the results presented in Table 2, which consistently show that FDA performs best for all seeds. All experiments were run using balanced data since this showed enhanced perfor-

	TF-IDF		CED		FDA	
	BLEU ↑	TER ↓	BLEU ↑	TER ↓	BLEU ↑	TER ↓
240K	.384	.535	.384	.535	.384	.535
+ 240K	.417	.506	.409	.513	**.439**	**.487**
+ 480K	.453	.470	.443	.483	**.464**	**.467**
+ 1M	.466	.470	.456	.473	**.477**	**.457**
100K	.306	.613	.306	.613	.306	.613
+ 240K	.394	**.520**	.361	.550	**.396**	.523
+ 480K	.419	.507	.396	.533	**.425**	**.498**
+ 1M	.430	.498	.420	.505	**.432**	**.489**
50K	.219	.685	.219	.685	.219	.685
+ 240K	.368	.554	.296	.633	**.370**	**.552**
+ 480K	.379	.545	.339	.595	**.390**	**.537**
+ 1M	.391	**.531**	.384	.547	**.394**	.535

Table 2: Results of running different data selection methods on different seed sizes for EMEA EN→FR. The top result for each seed size and slice of data selected is presented in bold. The ratio in the parentheses indicate the number of times seed was oversampled

mance, as mentioned in the previous section.

For the impact of the seed size on the data selection gains, the results show that for similar selected data, the score decreases with the seed, which is visible from the seed score to the 1M data selection. This is an intuitive result, since the amount of information contained in the full size seed is obviously larger than its counterparts.

However, it also shows that the gains from the baseline to the data selection are actually bigger for smaller seeds, with around 5–9 BLEU points increase for the full seed, 9–13 for the 100K sample and 16–18 points for the smaller 50K sample. This is consistent with the fact that the amount of data used has a bigger impact in NMT, especially when compared with previous knowledge about these methods in SMT.

4.3 Impact of data selection in NMT

Using the previous results as starting points, we focus now only on the FDA method for data selection and use oversampling of the seed to obtain a balanced training set.

4.3.1 Full training

Several experiments are run for both domains, EMEA and DGT. To increase the confidence in our results, we repeat the experiment for English-Spanish, by selecting the corresponding Spanish sentences in both domain datasets.[6] All experiments for each language pair share the same seed data, oversampled to obtain a balanced corpus.

The results presented in Table 3 seem to support some of the previous conclusions that data selection does not yield as much gain for the NMT as it did for SMT. The best results are mostly data selection of 2M or 4M. However, the values are very close to the baseline obtained with the entire MultiUN data combined with the seed, which is balanced in the same way as the data selection methods. The results with 6M are also very close or slightly higher than the baseline, showing that more data helps almost as much as selected data.

4.3.2 Adaptation from generic models

To try and separate the impact of the huge amount of data the generic model represents, we ran the same experiments in a fine-tuning scenario. In this context, a model is firstly trained with all the generic data until convergence, without any added domain knowledge. Then, a new training pass is ran until convergence with the domain data, where we add the selected data to the seed as pseudo-domain data. We mean to compare these selections with a baseline using only the seed, since using the full data here is redundant.

The data selection performed in the fine-tuning scenario has a negative impact, as shown in Table 4, where most of the data selection sets used obtain scores lower than the original seed baseline. One possible factor is that the MultiUN data contains very little domain data. As mentioned in the previous section, this experiment would gain from

[6] Both the DGT and EMEA datasets are available in EN–FR, EN–ES, and ES–FR, where part of the lines are aligned across the three languages.

	$EMEA_{EN \rightarrow FR}$		$DGT_{EN \rightarrow FR}$		$EMEA_{EN \rightarrow ES}$		$DGT_{EN \rightarrow ES}$	
	BLEU ↑	TER ↓	BLEU ↑	TER ↓	BLEU ↑	TER ↓	BLEU ↑	TER ↓
Seed	.384	.535	.427	.469	.432	.485	.413	.453
+ 250K	.439	.487	.438	.436	.486	.434	.458	.410
+ 500K	.464	.467	.464	.417	.511	.418	.476	.397
+ 1M	.477	.457	.472	.409	.525	.403	.494	.382
+ 2M	.491	.446	.482	**.403**	.531	**.396**	**.496**	.383
+ 4M	**.492**	**.441**	.478	.404	**.535**	.398	.495	**.379**
+ 6M	.489	.448	.434	.453	.534	.399	.494	.385
+ all data (11M)	.487	.454	**.482**	.405	.495	.449	.493	.384

Table 3: BLEU and TER scores for NMT training with different slices of selected data, using FDA for data selection. The top two results for each column are shaded, with the top result presented in bold

gathering a larger and more diverse generic corpus.

Moreover, all fine-tuning results are below the fully trained models with all data from the previous section. The most important factor here seems to be the highly technical vocabulary the models can have access to. While the model trained with all data has access to both the generic and domain vocabulary, the fine-tuned models are built on top of the generic vocabulary only. Thus, the model's input vocabulary of the first contains the most relevant domain words, while in the second these are split into subwords, as would happen to rare words.

4.3.3 Human evaluation

We also conducted a human evaluation using Unbabel's quality control system. For each language pair, translation direction and domain, 150 sentences were chosen randomly for evaluation. We then shuffled the content and provided it to evaluators (professional linguists) for Fluency and Adequacy assessment. This assessment is done by rating each sentence from 1 to 5, and then computing the average for each model. The evaluators were not provided with the information as to which model was used to generate sentences. The definitions of Fluency and Adequacy, as used by the Unbabel Quality Team, are as follows.

Fluency addresses the linguistic well-formedness and naturalness of the text. Fluency errors include grammar, spelling or unintelligible text, sentence structure and word order issues, etc. In sum, these errors affect the reading and the comprehension of the text. The evaluation is done on the resulting translations without revealing their source sentences to the evaluators, to avoid biasing Fluency scores.

Adequacy addresses the relationship of the target text to the source text and can only be assessed by providing both translations and their source sentences to the editors. In other words, Adequacy addresses the extent to which a target text accurately renders the meaning of a source text. Adequacy errors include changes in intended meaning, addition and omission of content or any type of mistranslation, etc. In sum, Adequacy measures if the target text accurately reflect the meaning conveyed in the source text (Way, 2018).

The results of human evaluation on Fluency and Adequacy are presented in Table 5. The figures in the table correspond to the top scores in Tables 3 and 4. The results show that with fine-tuning of the training of models, Fluency is improved, especially for the EMEA models. Adequacy is also significantly improved in both EN-to-FR and EN-to-ES models. It shows very clear that data selection does improve the performance of all MT systems evaluated in this paper, in both Adequacy and Fluency.

It was also shown in Table 4 and Table 5 that for EN-to-FR, BLEU .452 of MT translated French sentences approximately corresponds to Fluency 4.25, and for EN-to-ES, BLEU .485 of MT translated Spanish sentences approximately corresponds to Fluency 4.50. In the future, we would like to make more comparisons between human evaluation metrics, e.g. Adequacy and Fluency as defined by Unbabel Quality Team, with commonly used MT performance metrics, e.g. BLEU and TER.

5 Conclusions

In this paper, we reviewed three commonly used data selection methods, i.e. TF-IDF, CED and FDA, for NMT. These methods improve the performance significantly for SMT. The results showed that FDA outperformed the other two methods. Although the gain in MT performance is not as much as

	$EMEA_{EN \rightarrow FR}$		$DGT_{EN \rightarrow FR}$		$EMEA_{EN \rightarrow ES}$		$DGT_{EN \rightarrow ES}$	
	BLEU ↑	TER ↓	BLEU ↑	TER ↓	BLEU ↑	TER ↓	BLEU ↑	TER ↓
MultiUN	.208	.699	.338	.528	.247	.657	.361	.495
Seed	.438	.481	**.476**	**.413**	**.486**	**.432**	**.487**	**.388**
+ 250K	.429	.485	.462	.418	.469	.442	.473	.399
+ 500K	**.439**	**.476**	.462	.416	.471	.438	.476	.396
+ 1M	.436	.478	.465	.414	.478	.440	.477	.397

Table 4: Fine-tuning approach for NMT training with data selection. The top two results for each column are shaded, with the top result presented in bold

Models trained		$EMEA_{EN \rightarrow FR}$		$DGT_{EN \rightarrow FR}$		$EMEA_{EN \rightarrow ES}$		$DGT_{EN \rightarrow ES}$	
		AD ↑	FL ↑	AD ↑	FL ↑	AD ↑	FL ↑	AD ↑	FL ↑
From Scratch	Seed	1.02	4.01	3.28	3.99	3.82	4.06	3.61	3.99
	+ best slice	4.18	3.95	3.87	4.39	4.25	4.42	4.22	4.50
	+ all data (11M)	4.1	3.95	3.78	4.29	3.99	4.33	4.19	4.47
With Fine-tuning	Seed	4.17	4.03	3.96	4.28	4.41	4.51	4.29	4.53
	+ best slice	4.22	4.05	4.12	4.45	4.43	4.50	4.30	4.52

Table 5: Human evaluation of Adequacy (AD) and Fluency (FL) for top scores in previous experiments in Tables 3 and 4

that in SMT systems, our experiments showed that using EMEA and MultiUN corpora, NMT systems trained with FDA-selected data still outperform systems trained with the whole corpus, in terms of both BLEU and TER.

In addition to using data selection, training with fine-tuning from pre-trained models is also employed to further improve MT performance. We conducted human evaluation by professional linguists, in which Adequacy and Fluency are assessed. The results show that models trained with selected data constantly outperformed those trained with the whole corpus, in both human evaluation measures. By employing fine-tuning on top of data selection, MT performance is further improved significantly in both Adequacy and Fluency.

Acknowledgements

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund. This project has partially received funding from the European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie Actions (Grant No. 734211). We would also like to thank Dr André Martins, Dr Julie Belião and Ms Marianna Buchicchio from Unbabel AI and Quality Teams for their help with the human evaluation.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Edinburgh, United Kingdom.

Ergun Biçici. 2013. Feature decay algorithms for fast deployment of accurate statistical machine translation systems. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 78–84, Sofia, Bulgaria.

Ergun Biçici and Deniz Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 272–283, Edinburgh, Scotland.

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*, pages 69–79, Vancouver, Canada.

Boxing Chen, Roland Kuhn, George Foster, Colin Cherry, and Fei Huang. 2016. Bilingual methods for adaptive training data selection for machine translation. In *The Twelfth Conference of The Association for Machine Translation in the Americas*, pages 93–106, Austin, Texas.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *The 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 678–683, Sofia, Bulgaria.

Sauleh Eetemadi, William Lewis, Kristina Toutanova, and Hayder Radha. 2015. Survey of data-selection methods in statistical machine translation. *Machine Translation*, 29(3-4):189–223.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 751–758, Berlin, Germany.

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++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for SMT. In *Proceedings of 45th annual meeting of the ACL on interactive poster & demonstration sessions*, pages 177–180, Prague, Czech Republic.

Will Lewis and Christian Federmann. 2015. Applying cross-entropy difference for selecting parallel training data from publicly available sources for conversational machine translation. In *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, pages 126–134, Da Nang, Vietnam.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden.

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*, pages 311–318, Philadelphia, Pennsylvania, USA.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2017. Applying n-gram alignment entropy to improve feature decay algorithms. *The Prague Bulletin of Mathematical Linguistics*, 108(1):245–256.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2018. Feature decay algorithms for neural machine translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 239–248, Alacant, Spain.

Alberto Poncelas, Andy Way, and Antonio Toral. 2016. Extending feature decay algorithms using alignment entropy. In *International Workshop on Future and Emerging Trends in Language Technology*, pages 170–182, Seville, Spain.

Gerard Salton and Chung-Shu Yang. 1973. On the specification of term values in automatic indexing. *Journal of documentation*, 29(4):351–372.

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*, pages 65–68, Valencia, Spain.

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

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Jörg Tiedemann. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248, Borovets, Bulgaria.

Andy Way. 2018. Quality expectations of machine translation. In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*, volume 1, pages 159–178. Springer.

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*, pages 1400–1410, Copenhagen, Denmark.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *International Conference on Language Resources and Evaluation LREC*, pages 3530–3534, Portorož, Slovenia.