Feature Decay Algorithms for Neural Machine Translation

Alberto Poncelas, Gideon Maillette de Buy Wenniger, Andy Way
ADAPT Centre, School of Computing,
Dublin City University, Dublin, Ireland
{firstname.lastname}@adaptcentre.ie

Abstract

Neural Machine Translation (NMT) systems require a lot of data to be competitive. For this reason, data selection techniques are used only for fine-tuning systems that have been trained with larger amounts of data. In this work we aim to use Feature Decay Algorithms (FDA) data selection techniques not only to fine-tune a system but also to build a complete system with less data. Our findings reveal that it is possible to find a subset of sentence pairs, that outperforms by 1.11 BLEU points the full training corpus, when used for training a German-English NMT system .

1 Introduction

In Statistical Machine Translation (SMT) it has been shown that having more data does not always lead to better results [Ozdowska and Way, 2009]. In fact, performance can increase by limiting the training data to a smaller but more relevant set [Eetemadi et al., 2015]. Neural Machine Translation (NMT) models in contrast are data-hungry, and perform better only with large amounts of available training data, in some cases underperforming SMT when low amounts of data are available [Östling and Tiedemann, 2017; Dowling et al., 2018]. However, the amount of training data required to make NMT work really well depends a lot on the domain of the training data and test set, and possibly also how similar they are. For certain training domains such as TED talks [Bentivogli et al., 2016] it has already been shown that even with fairly limited training sizes NMT can already outperform SMT by a large margin.

Larger training sets also introduce noise and require models to cover a larger number of words, whereas for practical reasons the vocabulary cannot be arbitrarily increased to facilitate these extra words. Consequently, training material that is not relevant for the test set risks wasting limited entries in the vocabulary on source words that are not relevant to the test domain. This is why domain adaptation has proven to be useful in NMT [Chu et al., 2017] by tailoring a model towards in-domain data.

While traditional Machine Translation (MT) approaches perform an inductive learning (i.e. learn a model from translated sentences in order to predict unseen examples), transductive learning approaches

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 239–248
Alacant, Spain, May 2018.

aim to identify the best training instance to predict the test set [Bianchini et al., 2016]. Models trained with sentences retrieved by transductive learning methods are tailored towards the test set. This is similar to the way many domain-adaptation methods adapt the training material to be suitable for a specific domain.

In this work, we use Feature Decay Algorithms (FDA), a transductive data selection method that has achieved good results in SMT and apply it to NMT. Our purpose is twofold: i) to question the widely held assumption that in Neural Machine Translation it is always better to use more data; and ii) to explore how a transductive data-selection technique like FDA should be applied in order to build models that outperform those built with all training data.

2 Related Work

Feature Decay Algorithms In our work, in order to extract a subset of the data, we use Feature Decay Algorithms [Biçici and Yuret, 2011; Biçici et al., 2015; Biçici and Yuret, 2015]. This is a method that uses the source side of the test set to select sentences that provide translation examples that are most relevant for this set. Furthermore, FDA aims to maximize the variability of these selected relevant *n*-grams in the training set by decreasing their value as they are being selected.

In order to do that, the features (*n*-grams extracted from the test set) are assigned an initial value, and each sentence of the training data is scored as the normalized (by dividing by the number of words) sum of the values of its *n*-grams. Then, the method iteratively selects the sentence with the highest score and adds it to the set of selected data (which initially is empty). After selecting a sentence, the values of the features contained in it are decreased according to the decay function. By default, the value of a feature f is defined as in (1):

$$decay(f) = init(f)0.5^{C_L(f)} \quad (1)$$

where $init(f)$ is the initial value and $C_L(f)$ is the count of the feature f in selected data.

The score of a sentence s at a particular iteration is the sum of the values of $C_L(f)$ of the features present in s, normalized by the length of s. The score of a sentence, using the default configuration in Equation (1), computed as in (2):

$$score(s) = \frac{\sum_{f \in F_s} 0.5^{C_L(f)}}{\text{\# words in s}} \quad (2)$$

where F_s is the set of features present in s.

FDA has proven to be useful in Statistical Machine Translation (SMT) Biçici [2013]; Poncelas et al. [2016, 2017]. Selecting a small subset of sentences from a parallel corpus using FDA is enough to train SMT systems that perform better than systems trained using the whole parallel corpus.

Neural Machine Translation We use neural machine translation [Kalchbrenner and Blunsom, 2013; Cho et al., 2014] in the form of sequence-to-sequence models [Sutskever et al.] based on recurrent neural networks [Bahdanau et al., 2014; Luong et al., 2015].

Fine-tuning A method of domain adaptation that has been used in NMT is "fine-tuning", which involves using a pre-built NMT system and training it further for several epochs with smaller amounts of in-domain data.

Most works Luong and Manning [2015]; Freitag and Al-Onaizan [2016] first use general domain data for training a system, and then a different in-domain data set for fine-tuning. Chu et al. [2017] train a system using a resource-rich domain corpus, and then use a small domain corpus to fine-tune the system.

The approach of Li et al. [2016] is the closest to our work, as they use the information of the test set to retrieve the data for tuning. Li et al. [2016] use string similarity measures, such as Levenshtein [Levenshtein, 1966] or the cosine similarity of the average of the word embedding [Mikolov et al., 2013] in order to find sentences that are close to a given sentence of the test set.

An alternative technique of performing fine-tuning is proposed by van der Wees et al. [2017]. They train the model with a dataset that is varied for each epoch, instead of training a model with a fixed training set, and then tuning it with a subset or another dataset for the last epoch. The size of the data is decreased gradually, keeping the sentences that are more in-domain, weighted using Cross-Entropy Difference [Axelrod et al., 2011]. The size of the subset of a training data S at each epoch e is defined as Equation (3):

$$n(e) = \alpha \cdot |S| \cdot \beta^{\lfloor (e-1)/\eta \rfloor} \quad (3)$$

where α is the relative start size, the fraction of training data used for the first epochs (relative start size), β is the fraction of training data kept in the new selection (retention rate), and η is the number of epochs for which the same subset is used.

3 Data selection using the source-side of the test set

Using the source side of test examples is central to machine translation. For example, SMT effectively uses only those phrase-pairs that match the source side of a test sentence. Matching can be done implicitly, inside the decoder and during translation, or explicitly, by filtering the phrase-table with the source-side of the test set before passing it to the decoder.

The usage of the test set source side by FDA is conceptually not different from well-established lazy supervised learning methods such as K-nearest neighbors, and is also not fundamentally different from the usage of source information for matching phrase-pair selection by SMT grammar extractors/decoders [Lopez, 2008].

4 Research questions

Due to the good performance achieved by FDA in SMT, we want to explore whether the improvements also maintain in NMT. Accordingly, the first question we want to answer is:

- Is FDA also useful for selecting a subset of training data to train NMT models that perform better than models trained with the larger (full) training data without any selection?

In NMT there are several possible configurations for applying a data-selection techniques. One method is to build a complete model from scratch using just a subset of the data. Another way is to use fine-tuning to specialize an existing model.

On the top of that, there are several possibilities of how to tune a model: (i) Performing fine-tuning (and even in this option, there are several possibilities as we can choose different epochs of the model to tune); and (ii) perform a gradual tuning, where at each epoch the model is trained in using gradually smaller in-domain subsets

Due to the different configurations available, our second research question is:

- What configuration should be applied so that NMT model benefits the most from FDA techniques?

The test set may not always be accessible when building the NMT model. However, a system tuned for a document of one domain using FDA may be be useful for translating a different one if they share the domain.

- Can a model biased towards one test set using FDA be useful for translating a different test set in the same domain?

5 Experiments

In this work we have constructed a German-to-English NMT system using the Pytorch port [1] of OpenNMT [Klein et al., 2017] to train the models. According to the creators of OpenNMT [2] a good baseline for German-to-English WMT 2015 data is the one built with default parameters (2-layer LSTM with 500 hidden units, vocabulary size of 50002 and 50004 for source and target language, respectively) executed for 13 Epochs. The words in the output that are not in the vocabulary are replaced with the word in the source with the highest attention.

The data sets used in the experiments are based on the ones used in the work of Biçici [2013]:

- *Training data*: The training data provided in the WMT 2015 [Bojar et al., 2015][3] translation task setting a maximum sentence length of 126 words (4.5M sentence pairs, 225M words).
- *Validation data*: 5K randomly sampled sentences from development sets from previous years.

We extract subset of different sizes (100K, 200K, 500K, 1M and 2M sentences) from the training data with FDA using the test set from the WMT 2015 Translation Task. We use the default configuration of FDA (i.e. 3-grams as features, 0.5 as decay factor and 0 as decay exponent of 0). We perform several experiments building different NMT models using the training data and the data extracted with FDA. In order to answer the research questions in Section 4, models are built following different configurations:

- *FDA* experiments: Build NMT models from scratch, using only the output of FDA as training data.
- *BASE12+FDA* experiments: Fine-tune the last epoch of the baseline model with the output of FDA. Since the baseline is run for 13 epochs, we use the model of the 12th epoch.
- *BASE8+FDA* experiments: Fine-tune the the baseline model starting from the 8th epoch. We choose the 8th epoch not only because it is close to the middle stage of the training, but also because it is the point where fine-tuning and convergence of the model is initiated by starting the decay of the learning rate.
- *Gradual fine-tuning* experiments: Perform a gradual fine-tuning where the complete training data is used on the first epochs but gradually smaller sizes of training data are used thereafter. The sentences that are kept for the next iteration are the top sentences retrieved by FDA (being smaller at each epoch). The experiments are performed with the same configuration in the original work of van der Wees et al. [2017], using $\alpha = 0.5$, $\beta = 0.7$ and $\eta = 2$ in Equation (3).

In addition, we are interested in exploring whether the model trained on data retrieved by FDA using one document could also be useful for translating different documents in the same domain. We use the same models (trained with data using the test set of WMT 2015 as seed) for translating a different test set, the namely WMT 2014 [Bojar et al., 2014] news test set, which is in the same domain.

[1] https://github.com/OpenNMT/OpenNMT-py
[2] http://opennmt.net/Models/
[3] http://www.statmt.org/wmt15/translation-task.html

6 Results

In Table 1 and 2 we show several evaluation metrics: BLEU [Papineni et al., 2002], TER [Snover et al., 2006], METEOR [Banerjee and Lavie, 2005] and CHRF3 [Popovic, 2015]. These scores give an estimation of the quality of the output of the experiment when compared to a translated reference. In Table 2 we have also marked in bold the scores that outperform the baseline (Table 1) and computed the statistical significance (marked with an asterisk) with multeval [Clark et al., 2011] for BLEU, TER and METEOR when compared to the baseline at level p=0.01 using Bootstrap Resampling [Koehn, 2004].

	baseline
BLEU	0.2474
TER	0.5525
METEOR	0.2798
CHRF3	48.9473

Table 1: Results of the model trained with all available training data.

In the *baseline* column of Table 2 we see the scores of the translation of the test set (WMT 2015 document) using all training data. In the column *FDA* we present the results of the models built from scratch on different sizes of data retrieved by FDA (different subtables). As expected, an NMT model trained with small sets of data achieves worse results than the baseline. However, we discover that after selecting enough data, the system trained with less data outperforms the baseline. Using just 11% of sentences is enough to obtain better results (500K subtable) that are statistically significant for more than one evaluation metric. We observe the best results when selecting 2 million sentences, which is just 44.6% of the total number of sentences.

If we compare the models which have been fine-tuned (columns *BASE8+FDA*

	FDA	BASE8 +FDA	BASE12 +FDA
100K lines (2%)			
BLEU	0.1951	0.244	0.2458
TER	0.6243	0.5567	0.553
METEOR	0.245	0.2771	0.2793
CHRF3	42.9756	48.5617	48.7841
200K lines (4%)			
BLEU	0.2304	0.2445	**0.2479**
TER	0.5788	0.5562	**0.5523**
METEOR	0.2722	0.2773	**0.2804**
CHRF3	47.2747	48.5487	**49.0209**
500K lines (11%)			
BLEU	**0.2517***	**0.2478**	**0.2487**
TER	0.5601	0.5536	**0.5518**
METEOR	**0.2886***	0.2797	**0.2805**
CHRF3	**49.8314**	48.8575	**49.0866**
1M lines (22.3%)			
BLEU	**0.2560***	**0.2480**	**0.2475**
TER	**0.5497**	0.5533	**0.5524**
METEOR	**0.2886***	0.279	**0.2801**
CHRF3	**50.0932**	48.8372	48.9158
2M lines (44.6%)			
BLEU	**0.2585***	**0.2484**	0.2472
TER	**0.5454***	0.5543	**0.5522**
METEOR	**0.2894***	0.2795	**0.2802**
CHRF3	**50.0950**	48.8752	48.9247
Gradual fine-tuning			
BLEU	**0.2478**	-	-
TER	0.5588	-	-
METEOR	0.2798	-	-
CHRF3	48.8834	-	-

Table 2: Comparison of results of system trained in different sizes of training data retrieved by FDA

and *BASE12+FDA*), the scores obtained in *BASE12+FDA* experiments are better than *BASE8+FDA*. Almost all the evaluation metrics (the only exception is the BLEU score in subtable of 1M lines) are better when the fune-tuning is applied in the last epoch rather than in earlier stages. The *BASE12+FDA* experiment performs better than the baseline when using subsets of more than 200K sentences (we see in column *BASE12+FDA* that most of the scores are in bold). However none of them are statistically significant better than the baseline.

In the last subtable of Table 2 we show the performance of the model built using gradual fine-tuning. Even if it obtains a higher BLEU score the output is not statistically significantly better than the baseline at level p=0.01 for any of the metric.

We have seen that models trained with a subset of data perform better than those trained with all the data. As models built from scratch are not required to extract the words from the whole training data but only from the subset of sentences pairs relevant to the test set source, these are able to focus the limited vocabulary space more on those words that are relevant for the test set source. Tuning approaches in contrast preserve the initial vocabulary, which means they do not benefit from the more focused vocabulary training from scratch using FDA allows, which is one of the principles behind the working of FDA.

6.1 Further analysis: generalisation to additional test sets within the same domain

In order to explore whether the models built are also useful for translating another test set, we present Table 4. Here we see that the only scores that are statistically significantly better (marked with an asterisk) than the baseline (Table 3) are those of the *FDA* experiment

	baseline
BLEU	0.2502
TER	0.5558
METEOR	0.2824
CHRF3	49.5967

Table 3: Results of the model trained with all available training data using a different test set (WMT 2014 test set).

when 2M sentences are selected. These results are consistent with those observed in Table 2.

Training models with smaller in-domain data sets achieves better results. In addition, fine-tuning applied in the last epoch causes the results to improve, as in Table 2. However, while we can still see improvements over the baseline for *BASE12+FDA* (numbers in bold in Table 4, column *BASE12+FDA* when 500K sentences or more are selected), none of these improvements are observed for the *BASE8+FDA* configuration/column in Table 4. Furthermore none of these improvements are statistically significant.

The main difference with Table 2 is that more training data is necessary to achieve results that are better than the baseline. This is because in this set of experiments, the vocabulary is not directly obtained from the test set but from a document in the same domain.

Note that in this set of experiments the seed used to extract in-domain data is the WMT 2015 test set which contains only 2169 lines. In future work, we want to explore whether the results can improve if we use documents with more sentences as seed.

As we argued in the introduction, using the source side of the test set is used even implicitly for fragment selection by all data-oriented (fragment based) methods, including SMT, though this may not be widely realized by practitioners in the field. But these results show that FDA can give improvements even if

	FDA	BASE8 +FDA	BASE12 +FDA
100K lines (2%)			
BLEU	0.1625	0.2419	0.2489
TER	0.6633	0.5623	0.5563
METEOR	0.2185	0.279	0.282
CHRF3	39.7603	48.9277	49.4858
200K lines (4%)			
BLEU	0.1982	0.2432	0.2501
TER	0.6157	0.5625	0.5566
METEOR	0.2483	0.2786	0.282
CHRF3	44.1265	48.7811	49.4807
500K lines (11%)			
BLEU	0.2307	0.2478	**0.2502**
TER	0.5759	0.5582	**0.5555**
METEOR	0.2711	0.2813	**0.2830**
CHRF3	47.752	49.2136	**49.6680**
1M lines (22.3%)			
BLEU	0.2458	0.2484	**0.2504**
TER	0.5662	0.558	0.5559
METEOR	0.2797	0.2814	**0.2828**
CHRF3	48.8866	49.2997	49.5829
2M lines (44.6%)			
BLEU	**0.2530***	0.2491	0.2501
TER	**0.5553**	0.556	0.5549
METEOR	**0.2849***	0.282	**0.2826**
CHRF3	**49.8117**	49.3921	49.5804
Gradual fine-tuning			
BLEU	0.245	-	-
TER	0.5644	-	-
METEOR	0.2787	-	-
CHRF3	48.8506	-	-

Table 4: Comparison of results of system trained in different sizes of training data retrieved by FDA using a different test set (WMT 2014 test set).

we omit the direct use of the source side of the test set, as is normally done by FDA.

7 Conclusion and Future Work

In this work we have discovered that using FDA, it is possible to find a subset of data that can be used to train an NMT model that achieves better results than a model trained with all data. In particular, our best model, trained on 44.6% of the data improves over the baseline trained on the full training set, while also giving significant improvements on other metrics. Besides the significant improvement in translation quality, this also implies (in the chosen training regime, with 13 epochs and FDA after 8 epochs) a linear reduction in training time compared to the baseline system. For example, by reducing the training data by half for the last 8 epochs we use only 81% of the original training time[4].

In future work, we want to study the impact of the differences in vocabulary in each experiment. We also want to compare these results to different data-selection techniques or different variants of FDA (either using different values in the parameters, or different variants of the algorithm such us the one proposed in Poncelas et al. [2016, 2017]).

Acknowledgements

This research has been supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

This work has also received funding from the European Union's Horizon 2020 research and innovation programme under the

[4]If the last 5 epochs are trained using 50% of the data, the training time is $((8 * 1) + (5 * 0.5))/13 = 0.81$

Marie Skłodowska-Curie grant agreement No 713567.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., July 2011.

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

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, Ann Arbor, Michigan, 2005.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, 2016.

Monica Bianchini, Anas Belahcen, and Franco Scarselli. A comparative study of inductive and transductive learning with feedforward neural networks. In *Conference of the Italian Association for Artificial Intelligence*, pages 283–293, Genova, Italy, 2016. Springer.

Ergun Biçici. 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, 2013.

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

Ergun Biçici and Deniz Yuret. Optimizing instance selection for statistical machine translation with feature decay algorithms. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2):339–350, 2015.

Ergun Biçici, Qun Liu, and Andy Way. ParFDA for fast deployment of accurate statistical machine translation systems, benchmarks, and statistics. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 74–78, Lisbon, Portugal, 2015.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, 2014.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisboa, Portugal, 2015.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua

Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, 2014.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 385–391, Vancouver, Canada, 2017.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, page 176–181, Portland, Oregon, 2011.

Meghan Dowling, Teresa Lynn, Alberto Poncelas, and Andy Way. Smt versus nmt: Preliminary comparisons for irish. *Technologies for MT of Low Resource Languages (LoResMT 2018)*, page 12, 2018.

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

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

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, October 2013.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 67–72, Vancouver, Canada, 2017.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, 2004.

Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710, 1966.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. One sentence one model for neural machine translation. *arXiv preprint arXiv:1609.06490*, 2016.

Adam David Lopez. *Machine Translation by Pattern Matching*. PhD thesis, College Park, MD, USA, 2008.

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

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Dis-

tributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 3111–3119, Daegu, South Korea, 2013.

Robert Östling and Jörg Tiedemann. Neural machine translation for low-resource languages. *arXiv preprint arXiv:1708.05729*, 2017.

Sylwia Ozdowska and Andy Way. Optimal Bilingual Data for French-English PB-SMT. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation*, pages 96–103, Barcelona, Spain, 2009.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

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

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

Maja Popovic. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, 2015.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 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, 2006.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 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, 2017.