STACC, OOV Density and N-gram Saturation: Vicomtech's Participation in the WMT 2018 Shared Task on Parallel Corpus Filtering

Andoni Azpeitia and **Thierry Etchegoyhen** and **Eva Martínez Garcia**
Vicomtech
Mikeletegi Pasalekua, 57
Donostia / San Sebastián, Gipuzkoa, Spain
{tetchegoyhen, aazpeitia, emartinez}@vicomtech.org

Abstract

We describe Vicomtech's participation in the WMT 2018 Shared Task on parallel corpus filtering. We aimed to evaluate a simple approach to the task, which can efficiently process large volumes of data and can be easily deployed for new datasets in different language pairs and domains. We based our approach on STACC, an efficient and portable method for parallel sentence identification in comparable corpora. To address the specifics of the corpus filtering task, which features significant volumes of noisy data, the core method was expanded with a penalty based on the amount of unknown words in sentence pairs. Additionally, we experimented with a complementary data saturation method based on source sentence n-grams, with the goal of demoting parallel sentence pairs that do not contribute significant amounts of yet unobserved n-grams. Our approach requires no prior training and is highly efficient on the type of large datasets featured in the corpus filtering task. We achieved competitive results with this simple and portable method, ranking in the top half among competing systems overall.

1 Introduction

Data-driven approaches to Machine Translation (MT) have been the dominant paradigm in the last two decades, with the development of Statistical Machine Translation (SMT) (Brown et al., 1990), and, more recently, of Neural Machine Translation (NMT) (Bahdanau et al., 2015). These approaches require large volumes of parallel sentences to properly model translation in a given language pair. However, large quality parallel corpora based on human translations are scarce across language pairs, and there is a strong need to build clean corpora from different sources.

The World Wide Web is a rich source of multilingual data, from which parallel corpora can be automatically created under appropriate conditions of use (Forcada et al., 2016). However, corpora created via crawling, with automated document and sentence alignment, tend to exhibit significant volumes of noisy data, which can be detrimental to the training of MT systems (Khadivi and Ney, 2005; Khayrallah and Koehn, 2018a).

The task of cleaning noisy data from parallel corpora has been tackled by various researchers over the years. In (Munteanu and Marcu, 2005), noise removal is performed via a maximum entropy model trained on observations of clean and noisy data. Esplá-Gomis and Forcada (2009) include sentence alignment scores in BiTextor, a tool that performs the complete chain of corpus creation from web data, to filter dubious sentence pairs. In (Khadivi and Ney, 2005), two approaches are evaluated, based on length and on lexical translation likelihood, showing statistically significant improvements in translation quality using the filtered corpus. An unsupervised filtering method based on outlier detection is proposed in (Taghipour et al., 2011), who also report improvements in translation quality from their filtered corpus. In (Cui et al., 2013), the approach to data filtering is based on graph-based random walks, with improvements observed for Chine-English machine translation. Recently, Xu and Koehn (2017) introduced Zipporah, a fast data selection system for noisy parallel corpora, which is shown to result in improved SMT system quality.

The WMT 2018 task on parallel corpus filtering offers the possibility to compare different approaches to the task, evaluating their impact on both SMT and NMT systems on several test sets in different domains. Our participation in the task aimed to evaluate a simple and portable approach, based on the efficient STACC system for parallel sentence extraction from comparable corpora (Etchegoyhen and Azpeitia, 2016). We extended

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 860–866

https://doi.org/10.18653/v1/W18-64100

the original approach with a simple method based on the number of unknown words, to tackle the significant amounts of noise featured in the corpus filtering task. Additionally, we experimented with a simple approach to data redundancy, based on n-gram saturation. Our contribution centred on providing a sound method that can be easily deployed, does not require prior training, and can efficiently process large volumes of data.

2 Approach

Our approach to the task is based on STACC, a portable and efficient method for the identification of parallel sentences in comparable corpora (Etchegoyhen and Azpeitia, 2016) which obtained the best results for all language pairs in the BUCC shared tasks (Azpeitia et al., 2017, 2018). As the method assigns an alignment score to source and target sentence pairs, it can be directly applied to parallel corpus filtering as well, with a simple extension for this specific task. We describe the components of our approach in the next sub-sections.

2.1 STACC

The STACC approach has been described and explored in detail in (Etchegoyhen and Azpeitia, 2016), and we briefly summarise below how similarity is computed with this method.

Let s_i and s_j be two tokenised and truecased sentences in languages l_1 and l_2, respectively, S_i the set of tokens in s_i, S_j the set of tokens in s_j, T_{ij} the set of lexical translations into l_2 for all tokens in S_i, and T_{ji} the set of lexical translations into l_1 for all tokens in S_j.[1]

Lexical translations are initially computed from sentences s_i and s_j by retaining the k-best translations for each word, if any, as determined by the ranking obtained from the lexical translation probabilities computed with IBM word alignment models (Brown et al., 1990). The sets T_{ij} and T_{ji} that comprise these k-best lexical translations are then expanded by means of two operations:

1. For each element x in the set difference $T'_{ij} = T_{ij} - S_j$ (respectively $T'_{ji} = T_{ji} - S_i$), and each element y in S_j (respectively S_i), if x and y share a common prefix of more than n characters, the prefix is added to both T_{ij} and S_j (respectively T_{ji} and S_i). This longest common prefix matching strategy is meant to capture morphological variation via minimal computation.

2. Numbers and capitalised truecased tokens not found in the translation tables are added to the expanded translation sets T_{ij} and T_{ji}. This operation addresses named entities, which are strong indicators of potential alignment given their low relative frequency and are likely to be missing from translation tables trained on different domains.

With source and target sets as defined here, the STACC similarity score is then computed as in Equation 1:

$$stacc(s_i, s_j) = \frac{\frac{|T_{ij} \cap S_j|}{|T_{ij} \cup S_j|} + \frac{|T_{ji} \cap S_i|}{|T_{ji} \cup S_i|}}{2} \quad (1)$$

Similarity for the core metric is thus defined as the average of the Jaccard similarity coefficients obtained between sentence token sets and expanded lexical translations in both directions.

The STACC approach has been extended in (Azpeitia et al., 2017, 2018), notably via a word weighting scheme that led to significant improvements in the parallel sentence extraction task. In this work, we used the original weightless approach, as it performed slightly better in preliminary experiments on the noisy web data of the WMT 2018 task.

2.2 OOV Density

The corpus for the WMT 2018 shared task on parallel corpus filtering features significant volumes of noise, as is typical with parallel corpora gathered via web crawling that targets recall. (Khayrallah and Koehn, 2018b) manually examined a sample of data generated by the Paracrawl project,[2] of the type used in this shared task, and identified as noise misaligned sentences, content in the wrong languages, untranslated sentences, random byte or HTML markup sequences. The latter four types can be notably characterised as displaying significant percentages of out-of-vocabulary (OOV) words, assuming a vocabulary extracted from a separate parallel corpus with limited amounts of noisy data.

As previously described, the STACC approach, which constitutes the core of our method, is geared

[1] As in the original approach, we use sets rather than multisets, i.e. without repeated elements. The term *tokens* refers to the components of the tokenised sentences, and repeated tokens are thus only represented once in the sets.

[2] https://paracrawl.eu/

towards computing alignment scores in comparable corpora, with lower volumes of noise, notably allowing OOV words to contribute to the score if they are capitalised words in truecased sentences or numbers. This enables the capture of surface-defined named entities, which are a decisive factor for parallel sentence identification in comparable datasets (Azpeitia et al., 2018). However, this approach can be weaker in highly noisy datasets, where, for instance, random sequences of numbers may lead to an unwarranted high alignment score.

Since we aimed to avoid adding task-specific cleanup heuristics, such as performing time-consuming language identification or filtering sequences in an ad-hoc manner, we experimented with a penalty based on the number of unknown words in the corpus to be filtered, determined from the separate parallel corpus used to extract lexical translations. The penalty is computed as follows for each sentence s, source or target, where $|oov|$ is the number of unknown words in the sentence and $|s|$ is the sentence length, in number of words:

$$p(s) = 1 - \frac{|oov|}{|s|} \tag{2}$$

The STACC.OOV alignment score for each sentence pair (s_i, s_j) is then computed as follows:

$$stacc.oov(s_i, s_j) = stacc(s_i, s_j) \cdot \frac{p(s_i) + p(s_j)}{2} \tag{3}$$

Thus, sentences with a small amount of OOV words, of interest to extend MT coverage, will be assigned a score close to the original STACC score, whereas the score for dubious sentences with large numbers of unknown words will tend to zero. Our primary submission was based on the metric in Equation 3, as the initial goal of the task was to assign an absolute alignment quality score.

2.3 N-gram Saturation

The organisers of the shared task had allowed the use of metrics that did not score sentences in isolation. That is, sentence pairs could be scored by considering their redundancy with regards to higher scoring pairs. This aspect enables the design of methods that select the n-best sentence pairs to train machine translation models.

To experiment with data redundancy, we implemented a simple method based on n-gram coverage, similar in spirit to the n-gram coverage and saturation methods of Eck et al. (2005) and Lewis and Eetemadi (2013). The method can also be related to the Feature Decay approach proposed in (Biçici and Yuret, 2011), originally applied to SMT models and recently evaluated on NMT as well (Poncelas et al., 2018).

We compute n-gram saturation by first sorting the corpus according to the STACC.OOV scores, from high to low scores. We then process the sorted corpus by extracting n-grams (up to a specific order n) from each source sentence, storing the collected n-grams in a Patricia trie T (Morrison, 1968) for fast retrieval, and computing the amount of new n-grams for each sentence. The steps for a given sentence s are described below:

1. Retrieve all n-grams in s.
2. Determine all new n-grams from step 1, i.e. n-grams not found in the trie T.
3. Compute the ratio of new to existing n-grams in s as in Equation 4, for each n-gram ng up to order k:

$$ngsat(s) = \frac{\sum_{n=1}^{k} ng_n \notin T}{\sum_{n=1}^{k} ng_n \in T} \tag{4}$$

4. Add all new n-grams to the trie T.

Finally, we compute the score of the STACC.OOV.NGSAT variant for each sentence pair by multiplying the pair's existing score in the sorted corpus, computed as in Equation 3, by its *ngsat* score. Thus, pairs that provide no new n-grams would get an overall score of zero, while pairs with a large amount of new n-grams would get a score close to the existing score.

This simple method differs from the one in (Eck et al., 2005) in two ways: we do not pre-compute nor use n-gram frequency, and our normalisation factor is the total number of n-grams for the sentence instead of sentence length. Our approach also has linear complexity instead of quadratic, since, contrary to their different scenario focussed on data selection, we do not need to recalculate costs for all sentence pairs after processing one pair. Our method also differs from that of (Lewis and Eetemadi, 2013), as we do not use a threshold of n-gram counts but the percentage of new n-grams contributed by a given sentence,

MT	SYSTEM	AVG	RANK	NEWS	IWSLT	ACQUIS	EMEA	GLOBAL	KDE
SMT 10M	BEST	24.58	1/48	29.59	22.16	21.45	28.28	22.67	25.51
SMT 10M	STACC.OOV	23.25	16/48	27,48	20.42	19.33	26.51	21.20	24.55
SMT 10M	STACC.OOV.NGSAT	23.29	13/48	27,52	19.80	19.33	26.84	21.12	25.14
SMT 100M	BEST	26.50	1/48	31.35	23.17	22.51	31.45	24.00	26.93
SMT 100M	STACC.OOV	25.91	24/48	30.47	22.47	22.16	30.30	23.43	26.63
SMT 100M	STACC.OOV.NGSAT	25.80	29/48	30.17	22.39	22.12	30.03	23.36	26.70
NMT 10M	BEST	28.62	1/48	36.04	25.23	25.30	32.72	26.72	28.25
NMT 10M	STACC.OOV	26.35	13/48	32.33	22.57	22.55	28.96	24.28	27.39
NMT 10M	STACC.OOV.NGSAT	25.64	17/48	31.25	21.81	20.67	29.09	23.48	27.56
NMT 100M	BEST	32.06	1/48	39.85	27.43	28.36	36.70	29.26	30.79
NMT 100M	STACC.OOV	30.40	27/48	37.08	26.35	26.81	34.54	27.74	29.89
NMT 100M	STACC.OOV.NGSAT	24.91	40/48	27.23	22.44	23.15	26.92	22.94	26.76

Table 1: Results on the WMT 2018 test sets

MT	SYSTEM	Δ_{MEAN}	Δ_{MEDIAN}	Δ_{BEST}
SMT 10M	STACC.OOV	+1.83	+0.74	-1.33
SMT 10M	STACC.OOV.NGSAT	+1.87	+0.79	-1.29
SMT 100M	STACC.OOV	+1.03	+0.03	-0.59
SMT 100M	STACC.OOV.NGSAT	+0.92	-0.08	-0.71
NMT 10M	STACC.OOV	+4.51	+1.79	-2.27
NMT 10M	STACC.OOV.NGSAT	+3.80	+1.09	-2.98
NMT 100M	STACC.OOV	+2.47	-0.27	-1.65
NMT 100M	STACC.OOV.NGSAT	-3.03	-5.77	-7.15
ALL	STACC.OOV	+2.46	+0.57	-1.46
ALL	STACC.OOV.NGSAT	+0.89	-0.99	-3.03

Table 2: Scoring differences on core statistics

and also assume the initial ordering provided by the STACC.OOV scores. Finally, our approach differs from the Feature Decay method in (Biçici and Yuret, 2011) on several aspects, as it is not based on rate of decay and n-gram saturation scores are computed in a single pass on the corpus to be filtered, without referring to source test features.

Our goal in experimenting with n-gram saturation was mainly to include a low complexity method that could account for data redundancy in a simple way. The scope of the experiments was also reduced to only cover n-grams on the source side, as we meant to evaluate the impact of data redundancy in terms of source context coverage. This evidently excludes cases where a saturated source context can be translated differently in the target language, which can impact the number of learned translation options and subsequently affect evaluation scores. We leave further evaluations of such cases for future research. In the next sections, we evaluate the STACC.OOV.NGSAT variant as our secondary submission to the WMT 2018 task.

3 Experimental Setup

Our approach implies only minimal deployment settings. We ran STACC with the following two hyper-parameters: minimal prefix length was set to 4 and k-best translation lists limited to 5 candidates. For the STACC.OOV.NGSAT variant, the n-gram order was set to 3.

For the lexical translation tables needed by the STACC algorithm, we trained IBM2 models with the FASTALIGN toolkit (Dyer et al., 2013), on corpora made available for the WMT 2018 news translation task. The corpora thus included *Europarl v7*, *Common Crawl*, *NewsCommentary*, and the *Rapid corpus of EU press releases*. The *Paracrawl* corpus was excluded from the training data in order to extract reliable lexical translation tables from less noisy bilingual corpora. After duplicates removal, the training corpus amounted to $5,623,721$ parallel sentences.

The corpus was processed on an in-house server, using 64 threads. The total processing time for the 104 million sentence pairs of the corpus was around 57 minutes with the STACC.OOV variant, consuming a maximum of 11.3GB of RAM. With the STACC.OOV.NGSAT variant, processing time was approximately 5 times slower, with an order of magnitude larger consumption of RAM, mainly due to our online trie computation.

Given our stated objectives of evaluating a simple and portable method for the task, our preliminary experiments were all based on variants of the STACC approach, evaluated on the development set provided by the organisers. We no-

tably experimented with the variant in (Azpeitia et al., 2017), where the STACC score is computed via frequency-based lexical weighting that favours content words, and the variant in (Azpeitia et al., 2018), which features a scoring penalty that promotes named-entity matching. Although the differences were minor, the original STACC approach performed better overall and was thus selected as the core of the metric for our final submissions.

4 Results

The results of our approach on the WMT 2018 test sets are shown in Table 1.[3] Overall, our primary submission, STACC.OOV performed well on the task, ranking in the top third for SMT 10M and NMT 10M, and as a mid-performing system in the other two scenarios. Given the simplicity and efficiency of our approach, and the relatively minor differences with the top performing systems, we view these results as quite satisfactory.

The ranking was relatively uniform between test sets, with the notable exception of the KDE test set for which our approach was among the top 10 submissions in 3 out of 4 scenarios, and ranked 20th in the fourth case. This may be due to the fact that our scores are assigned purely in terms of alignment and not geared towards selecting sentence pairs that may be more informative for the news domain or similar, for instance. Thus, short sentence pairs with technical content that are correct translations will receive high scores although they may not be the most relevant pairs for the other test sets that feature less technical language.

Both systems performed similarly for SMT and NMT in terms of rankings obtained on the 10M and 100M versions. The variant of our approach that includes n-gram saturation performed similarly to our primary submission overall for SMT, but worse for NMT. Given these results and the fact that computing n-gram saturation is more resource-consuming, our primary submission was the optimal option of the two. A more detailed analysis would be needed to evaluate the causes for the drop caused by n-gram saturation for NMT. It could be conjectured that NMT training is optimal with the largest number of contextual variants in the training data, variants which would tend to be demoted via n-gram saturation. Phrase-based SMT can be considered less sensitive to contextual variants, given its core phrase-independence translation assumption. We leave a more precise analysis of these aspects for future research.[4]

To further compare our systems to the other submissions, we computed the core statistics on the average scores of all systems. In Table 2 we indicate the differences between our submission scores and the mean (Δ_{MEAN}), median (Δ_{MEDIAN}) and best (Δ_{BEST}) scores. In the last two rows of the table, we indicate the average differences for all scenarios in each category.

Our system performed better than the mean, in particular for NMT, with improvements of 4.51 and 2.47 for the primary submission. The one exception is the n-gram saturation variant, whose performance dropped significantly for NMT 100M, which may be explained under the aforementioned conjecture. The results in terms of the median are in line with the rather similar results obtained by a large number of participating systems.

Another notable aspect illustrated by this view of the results is the relatively higher differences with respect to the best performing system when considering NMT results, with a 1 BLEU point difference on average. Determining whether this difference reflects a systematic tendency would require a larger set of experiments with different corpora and language pairs. On average, our primary submission was 1.46 BLEU points below the best system and 2.46 points above the mean. Considering also the high efficiency of the approach, which

[3] For ease of presentation, we only indicate the official results in terms of C-BLEU scores, as provided by the shared task organisers. Along with the results of our systems, we also indicate the scores of the best system for each test set. The column AVG indicates the average score across all test sets and RANK denotes the ranking of the system among all participants according to the average score.

[4] As pointed out by one of the reviewers, an alternative explanation could be formulated. SMT systems are more sensitive to missing n-grams, contrary to NMT models, which rely on word embeddings and are thus less sensitive to specific words or n-grams. Thus, rather than NMT needing more contextual variants, the results could reflect that SMT benefits more from the additional n-grams provided via the saturation method, whereas NMT suffers from the imbalance in the training data that results from n-gram saturation filtering. Although this explanation has its merits, it would also warrant further examination. First, our results did not actually improve when using n-gram saturation for SMT, overall, which would tend to show that the additional n-grams collected via saturation did not have a significant impact in these experiments; only NMT systems were negatively impacted by the saturation method. Secondly, the suggested data imbalance could actually be viewed as a reduction in contextual variants, as we hypothesised, which could impact the computation of both embeddings and context vectors in attention-based NMT. Whether data imbalance could be viewed differently from contextual variants reduction is an interesting topic to be further explored.

can process the 104M parallel sentences in under an hour without the need for language-dependent tools nor any prior training, we view our method as a practical and reliable alternative to filter large noisy parallel corpora.

5 Conclusions

We have described our participation in the WMT 2018 shared task on parallel corpus filtering. Our approach was based on the STACC system, which only requires lexical translation tables to assign alignment quality scores. For this task, the core system was augmented with a simple penalty based on the number of unknown words in the sentences, to account for the significant volumes of noise in the corpus. Additionally, we experimented with a simple n-gram saturation scheme to evaluate the impact of demoting redundant data.

The results were satisfactory for such a simple and computationally efficient approach, which does not require prior training, sophisticated setups, language-dependent analysers, complex feature sets or extensive computational resources. In fact, our approach only requires pre-trained IBM2 lexical translation tables, which can be efficiently computed with generic off-the-shelf tools. We achieved competitive results overall, ranking in the top half among competing systems overall, with scores above the mean and less than 1.5 BLEU points below the top performing systems on average. The n-gram saturation variant did not provide significant improvements and actually performed significantly worse in one scenario, while also consuming more computational resources. The simpler primary variant of the system thus proved optimal for the task and more research would be needed to better account for data redundancy within our core approach.

The system we submitted is also quite efficient, being able to process the 104M sentence pairs in the task corpus in under an hour. Overall, we view our approach as a portable and efficient method to filter noisy data from parallel corpora. In future work, we will evaluate variants of the approach, exploring in particular the specifics of the data featured in different domains.

Acknowledgements We wish to thank the anonymous WMT 2018 reviewers for their detailed and helpful reviews of this work. Opinions, errors and omissions are our own.

References

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. 2017. Weighted Set-Theoretic Alignment of Comparable Sentences. In *Proceedings of the Tenth Workshop on Building and Using Comparable Corpora*, pages 41–45.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. 2018. Extracting Parallel Sentences from Comparable Corpora with STACC Variants. In *Proceedings of the Eleventh Workshop on Building and Using Comparable Corpora*, pages 48–52.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

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.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A Statistical Approach to Machine Translation. *Computational linguistics*, 16(2):79–85.

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, pages 340–345.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Low cost portability for statistical machine translation based on n-gram coverage. In *Proceedings of MTSummit X*.

Miquel Esplá-Gomis and Mikel L Forcada. 2009. Bitextor, a free/open-source software to harvest translation memories from multilingual websites. *Proceedings of MT Summit XII*.

Thierry Etchegoyhen and Andoni Azpeitia. 2016. Set-Theoretic Alignment for Comparable Corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 2009–2018.

Mikel L Forcada, Miquel Espla-Gomis, and Juan Antonio Perez-Ortiz. 2016. Stand-off annotation of web content as a legally safer alternative to bitext crawling for distribution. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 152–164.

Shahram Khadivi and Hermann Ney. 2005. Automatic filtering of bilingual corpora for statistical machine translation. In *International Conference on Application of Natural Language to Information Systems*, pages 263–274. Springer.

Huda Khayrallah and Philipp Koehn. 2018a. On the impact of various types of noise on neural machine translation. *CoRR*, abs/1805.12282.

Huda Khayrallah and Philipp Koehn. 2018b. On the impact of various types of noise on neural machine translation. *arXiv preprint arXiv:1805.12282*.

William Lewis and Sauleh Eetemadi. 2013. Dramatically reducing training data size through vocabulary saturation. In *Proceedings of the Eight Workshop on Statistical Machine Translation*, pages 281–291.

Donald R Morrison. 1968. Patricia practical algorithm to retrieve information coded in alphanumeric. *Journal of the ACM (JACM)*, 15(4):514–534.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2018. Feature decay algorithms for neural machine translation.

Kaveh Taghipour, Shahram Khadivi, and Jia Xu. 2011. Parallel corpus refinement as an outlier detection algorithm. *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, pages 414–421.

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*, pages 2945–2950.