Filtering of Noisy Web-Crawled Parallel Corpus: the Japanese-Bulgarian Language Pair

Iglika Nikolova-Stoupak **Shuichiro Shimizu** **Chenhui Chu** **Sadao Kurohashi**
Kyoto University

{iglika, sshimizu, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

One of the main challenges within the rapidly developing field of neural machine translation is its application to low-resource languages. Recent attempts to provide large parallel corpora in rare language pairs include the generation of web-crawled corpora, which may be vast but are, unfortunately, excessively noisy. The corpus utilised to train machine translation models in the study is CCMatrix, provided by OPUS. Firstly, the corpus is cleaned based on a number of heuristic rules. Then, parts of it are selected in three discrete ways: at random, based on the "margin distance" metric that is native to the CCMatrix dataset, and based on scores derived through the application of a state-of-the-art classifier model (Acarcicek et al., 2020) utilised in a thematic WMT shared task. The performance of the issuing models is evaluated and compared. The classifier-based model does not reach high performance as compared with its margin-based counterpart, opening a discussion of ways for further improvement. Still, BLEU scores surpass those of Acarcicek et al.'s (2020) paper by over 15 points.

Keywords: neural machine translation, low-resource language pairs, Bulgarian language, Japanese language, corpus filtering, web-crawled corpora

1 Introduction

In recent years, web-crawled corpora have come as an attempt to tackle the problem of limited parallel corpora, notably when it comes to machine translation involving low-resource language pairs. They are the product of unsupervised covering of portions of the web based on a widely used metric, such as the cosine distance between sentence embeddings, and they tend to be produced in excess, leading to problems like redundancy and data of low quality (Schafer et al., 2014). Large web-crawled corpora are often associated with a lack of documentation and require further work before they can be used within the field of machine translation (Dodge et al., 2021).

In their study, Khayrallah and Koehn (2018) discuss the types of noise that tend to occur in web-crawled corpora, as well as their effect on potential machine translation systems. Notably, neural machine translation is affected by such noise to a considerably greater extent as compared with its statistical counterpart, derived BLEU scores decreasing dramatically at its experimental introduction (Khayrallah and Koehn, 2018).

Motivated by a desire to mitigate the described problems, associated with similarly derived parallel corpora, WMT has organised three shared tasks in 2018-2020, addressing their cleaning, the last two of which have specifically centred on low-resource language scenarios. Several excellent state-of-the-art models have been produced to handle the task. In this paper, a representative model (Acarcicek et al., 2020) is selected and applied to a particular, extremely under-resourced language pair: Japanese-Bulgarian. Acarcicek et al.'s model uses a classifier on top of RoBERTa in order to score sentence pairs according to the level of certainty that they are mutual translations.

The corpus discussed in this study is CCMatrix, the largest parallel dataset that is currently available in the addressed language pair. It is provided by the OPUS collection (Tiedemann, 2012) and contains over four million multi-domain web-crawled sentences, derived based on "margin distance." The last is an improved implementation of cosine distance that considers the ratio of the cosine distance between two candidate sentences' embeddings as compared with the average cosine distance that a sentence has with its nearest neighbours (Schwenk et al., 2019). Following preprocessing based on heuristic

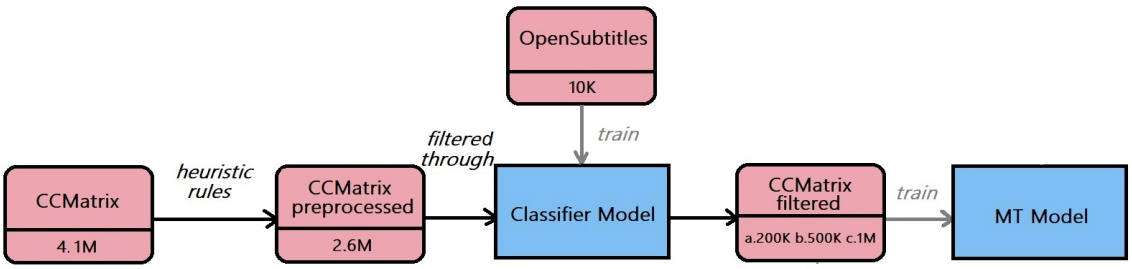

Figure 1: pipeline of the current study

rules that keep in mind the characteristics of the two languages in question, discrete subcorpora of three sizes (200K, 500K, and 1M) are selected based on margin distance and on the classifier-derived scores. They are compared to randomly selected subcorpora of the same size (see Figure 1). The margin-distance-based models show significantly improved performance. Conversely, the performance of the classifier models is largely non-optimal, showing the need for improvement of the selection techniques, such as through higher focus on the morphological and semantic specificities of the two languages. Importantly, the best derived model outperforms the one offered by Acarcicek et al. (2020) by over 15 BLEU points.

2 Related Work

2.1 WMT Corpus Filtering Shared Tasks

The particular languages addressed in this paper have not been involved in substantial research regarding the cleaning of noisy parallel corpora. This being said, the current study is highly inspired by the WMT corpus filtering shared tasks conducted in 2019 and 2020, which specifically targeted low-resource languages as an entity. Participants were prompted to provide a method of scoring the quality of each sentence within a provided noisy parallel corpus in order to then use the best scored pairs to train a translation model. In the process, they were allowed to use available clean parallel or monolingual data. The winning papers apply several distinct filtering techniques, including various uses of monolingual data, sentence embeddings, transfer learning, back translation, as well as the tool discussed in this paper, classifiers. In their highly successful model (which was consequently taken as a baseline within the shared task), Chaudhary et al. (2019) use only parallel data as they apply LASER sentence embeddings and calculate the cosine distance between sentences in order to obtain similarity scores. Lo and Joanis (2020) in turn utilise the semantic metric Yisi-2 in their scoring method, underlining the importance of vocabulary coverage. In their SMT system, Sen et al. (2019) come up with a fuzzy matching method akin to the one to be used in this paper, via which they calculate the Levenshtein distance between the corpus's English sentences and English translations of the additional language's sentences.

2.2 Use of Classifiers

A number of successful submissions to WMT's 2018-2020 shared tasks opt for a classifier model that differentiates between positive and negative examples of parallel sentences. In the 2018 edition of the shared task, Junczys-Dowmunt et al. (2018) assign cross-entropy scores to a noisy corpus's sentence pairs after first generating an inverse translation model trained on clean parallel data in the languages in question. Sánchez-Cartagena (2018) makes use of a classifier composed using the free open-source tool Bicleaner and enhanced with randomised trees and heuristic rules.

In fact, the use of classifier models in machine translation far predates the mentioned shared tasks as well as current state-of-the-art tools and recently assembled corpora. Munteanu and Marcu (2005) use a classifier to improve translation memory. Tyagi et al. (2015) apply support-vector machines and a Naive Bayes classifier in the ranking of translated sentences

into several categories ranging from "excellent" to "bad." Yogi et al. (2015) in turn rate the quality of produced machine translation with a Kneser-Ney smoothing language model that assigns probability scores to translated output. A year prior to the launching of WMT's shared tasks, Xu and Koehn (2017) come up with the data cleaning system Zipporah, which classifies the quality of translated sentences using bag-of-words.

3 Noise in the CCMatrix Corpus (Japanese-Bulgarian)

Akin to an experiment in Khayrallah and Koehn's (2018) study, a random 200-sentence sample from the described corpus is examined in an attempt to identify the nature of the different types of noise present.[1] The examined sentences demonstrate a large variety of domains and registers and feature a wide range of vocabulary, notably including a number of proper nouns. The main types of noise discovered include: non-corresponding numbers and dates, inappropriate punctuation, wrong use of abbreviations, presence of foreign languages, and machine-translated text.

As numbers and dates widely mismatch between the two languages within a sentence pair, they are regarded as noise. The next largest source of noise in the Bulgarian sentences comes in the face of problems with punctuation (for instance, a frequent use of "...") and capitalisation. What follows are instances of "non-standard language," including a large number of sentence fragments (for example, "Ако по някаква причина се преместят в друго училище,"). However, if one disregards the lack of final punctuation within these fragments, they read smoothly and match unproblematically between the two languages. In fact, the mandate for a sentence to contain a main verb, largely influenced by English grammar, is not intrinsic to either the Bulgarian or the Japanese language. While the Cambridge dictionary states that a sentence is "a group of words, *usually containing a verb*, that expresses a thought" ("Sentence", 2022; emphasis added), Bulgarian ("Изречение", 2022) and Japanese ("文", 2022) counterparts do not make a reference to the concept of "verb" in their definitions of a sentence.

One Bulgarian sentence contains the word "сори," a slang transliteration of the English "sorry" (the respective Japanese sentence does not demonstrate any parallelism). Seemingly machine-translated sentences come at as much as closely five per cent and are therefore placed in a separate category.

An example is the sentence "Как мога да защитавам моята PC?", which contains a gender mismatch and an unnatural English abbreviation. Other types of noise include abbreviations in both the Cyrillic and Latin alphabets, excessively large sentences and sentences written in (or partly in) a foreign language, predominantly English. Foreign language within a sentence ranges between a single word or phrase that can safely be regarded as a proper noun (e.g. "Google Assistant") and a full sentence written in a foreign language with a few seemingly mistakenly inserted Bulgarian words.

Similar patterns are observable when it comes to the noise in Japanese sentences: the use of numbers and dates, followed by abbreviations in the Latin alphabet and wrong punctuation. An additional problem related to punctuation is the fact that it differs significantly between the two languages; as a result, for instance, a Bulgarian "..." may be rendered as either "..." or "--" in the parallel Japanese sentence. Other examples of "non-standard language" come in the face of language attributable to "texting" (e.g. a "laughter" kanji in the end of a sentence) and supplementary hiragana renditions of kanji and katakana scripts, placed in brackets.

Some of the observed types of noise can be addressed directly during the preprocessing step (see Section 4.1). Such an issue as machine-translated language, however, is difficult to tackle using heuristic rules.

4 Methodology

4.1 Preprocessing

Like the majority of submissions for WMT's corpus filtering shared tasks, this study starts off with a preprocessing step that applies a series of heuristic rules to the noisy corpus. In concordance with observations described in Section 3, the

[1] See Appendix A for a detailed description of the types of noise found.

following preprocessing pipeline is applied: N/A entries and duplicates are removed; sentences in different languages are removed; Japanese sentences are tokenised; Japanese sentences with more than two pairs of brackets are removed (as they may indicate the use of multiple scripts); punctuation is removed; capitalisation is removed from Bulgarian sentences; sentences that show a large mismatch in size are removed; dates are replaced with the tag "DATE"; and numbers are replaced with the tag "NUM." The library *datefinder*[2] is utilised to locate dates written in a variety of formats. The tool used to identify sentences in languages other than the expected ones is *langdetect*[3]. Conveniently, in the case of short amounts of text in a foreign language, language is labelled in accordance with the large portion of text, thus allowing for sentences with words and phrases in English that take the role of proper nouns to remain in the corpus. Several patterns of wrong labelling are established and taken into consideration (e.g. Bulgarian text is occasionally mistakenly guessed to be in Russian or Macedonian).

Where applicable, the mentioned cleaning rules bear in mind the morphological and syntactic specificities of the two languages in question. For instance, the thresholds that are assumed to indicate unlikely proportions in sentence lengths are determined following observations of translation examples. Also, even though the later utilised neural models do not mandate prior tokenisation, a decision is made for Japanese text to be tokenised as part of preprocessing due to the language's notorious lack of space delimiters between words. The tool used for tokenisation is Juman++, developed in Kyoto University (Tolmachev et al., 2018).

4. 2 "Proxy Filter" Classifier

This study sought to apply a winning state-of-the-art model from the WMT corpus filtering shared tasks to the selected Japanese-Bulgarian corpus. Several criteria were considered within the choice of a model. Firstly, the focus was on 2019 and 2020 tasks, as they explicitly target low-resource language pairs (albeit in an English-centred setting). Simplicity, availability and reproducibility of research were also sought, thus dismissing for instance ensemble methods. Due to a strong recent shift toward NMT, SMT models were also disregarded, and so were models that involved not only corpus cleaning but also their own alignment of candidate parallel sentences (an option introduced in 2020's shared task). In the case of high similarity, newer models were preferred over older ones (for instance, Acarcicek et al. of 2020 was regarded as a better choice than Bernier-Colborne et al. of 2019). Finally, in the case of several experiments utilised within the same submission, only authors' best attempts were to be made use of.

Consequently, Acarcicek et al.'s 2020 model was selected. The authors enhance a multilingual RoBERTa-Large model (Liu et al., 2019) with a "proxy filter" i.e. a classifier that is trained to differentiate between positive and negative examples of parallel sentences. Specific attention is placed on the generation of challenging negative examples. The utilised technique is "fuzzy string matching," also known as "approximate string matching," which applies Levenshtein distance in the calculation of levels of similarity between texts.

A notable difference between Acarcicek's work and the one presented in this paper is the fact that the CCMatrix corpus already contains a metric pertaining to the level of parallelism of sentence pairs, the "margin distance." As a result, the study benefits from a comparison between a use of this native unsupervised metric and the newly derived classifier scores in the later translation model.

5 Experiments

5.1 Data

The parallel corpus whose cleaning is undertaken in this study is CCMatrix by OPUS (Japanese-Bulgarian). In its original form, the corpus contains 4.1M web crawled parallel Japanese-Bulgarian sentences. Following preprocessing based on heuristic rules, the corpus contains a little over 2.5M sentences.

The test and validation sets of the translation model comprise of 1,000 clean parallel sentence pairs each. The sentences are randomly taken from the top scoring 20K sentences following the

[2] datefinder.readthedocs.io

[3] https://pypi.org/project/langdetect/

classification task and are then removed from the training set. In order to guarantee quality and remove a bias toward sentences selected by the classification task, thorough manual editing and translation are applied.

The "proxy filter" classifier is trained on 10K parallel sentences from the OpenSubtitles (Japanese-Bulgarian) corpus. This corpus is significantly cleaner than CCMatrix, and it has notably been used by Koeva et al. (2012) in the construction of the "Bulgarian X-Language Parallel Corpus," the largest systematized Bulgarian bilingual corpus to date. Importantly, however, the OpenSubtitles corpus is more domain-specific as compared with CCMatrix, thus encouraging the extraction of a specific subtype of sentences from the latter.

All data is preprocessed following the same general pipeline as described in 4.1.

5.2 Classifier Model

The hyperparameters of the classifier model to be utilised were selected via grid searching: training epochs (0, 5), learning rate (2e-6, 2e-4, 2e-2, 0.2), negative random sampling[4] (2, 5, 8, 10), fuzzy ratio[5] (2, 1, 5), fuzzy max score[6] (30, 60, 100) and positive oversampling[7] (1, 2, 10). The models were trained on a single TITAN RTX GPU.

5.3 Translation Models

After the CCMatrix corpus was preprocessed, subcorpora were obtained through the application of three techniques: at random, based on margin distance and based on the classifier scores. Japanese-Bulgarian Transformer neural machine translation models [8] were trained as per the FAIRSEQ toolkit (Ott et al., 2019). In addition, three sizes of training data were introduced in an attempt to determine the optimal level of compromise between data size and data quality: 200K, 500K and 1M parallel sentences. The transformer models were trained on 8 TITAN X GPUS at a learning rate of 5e-4, using square root scheduler and a dropout of 0.3; early stopping was applied. The models were evaluated using BLEU scores.

6 Results

6. 1 Classifier Models

Over two thirds of the derived classifier models received an F1 score of 0 while at the same time showing high accuracy scores. An F1 of 0 implies that the value of either precision or recall is 0. A plausible reason is that such a model falsely identifies all examples as negative. While overall trends are difficult to pinpoint in relation to the models with highest F1 scores, all of them are trained for two epochs at a learning rate of 2e-6. Fuzzy matching scores and fuzzy ratios vary. When it comes to negative random sampling and positive oversampling, a general tendency is discernable for high values of the latter and slightly lower ones for the former (see Table 1).

Experiments with the application of several random seeds and a different amount of parallel data showed that, whilst a different random seed does not lead to significantly lower F1 scores for the best models, a different amount and organisation of parallel sentences often does reduce the score to 0. Training loss decreases smoothly with all models, the lowest score being associated with a model whose F1 score is 0.58.

	Fuzzy Ratio	Fuzzy Max Score	Positive Over-sampling	Negative Random Sampling	F1 Score
#1	1	100	2	10	0.72
#2	5	60	10	8	0.7
#3	5	30	10	8	0.7
#4	2	30	10	8	0.7
#5	1	30	10	8	0.7

Table 1: Varying hyperparameters among the top five classifier models according to F1 score

Due to the fact that the best scoring model (F1

[4] the ratio of negative examples in the classifier

[5] the number of similar sentences taken based on a sentence's fuzzy matching score

[6] a threshold (in percent) for the fuzzy matching similarity a sentence is allowed to exhibit; used in order to avoid the inclusion of duplicates or extremely similar sentences

[7] oversampling of the classifier's positive examples in order to maintain a given ratio with negative examples

[8] 6 layers, learning rate 5e-4, dropout 0.3, early stopping, vocabulary size 8,000

score of 0.72) demonstrates a slightly irregular pattern, such as the only negative ratio of 10 among the top five models and a fuzzy max score of 100 (a value that in fact negates the parameter's influence), the second best model (F1 score of 0.7) was selected as baseline.

6.2 Derived Scores

Following application of the classifier to the preprocessed CCMatrix corpus, each sentence pair received a score between 0 and 1, denoting its level of parallelism. The derived scores exhibit the following characteristics: their values range between 0.028 and 0.977, and their mean comes at 0.926.

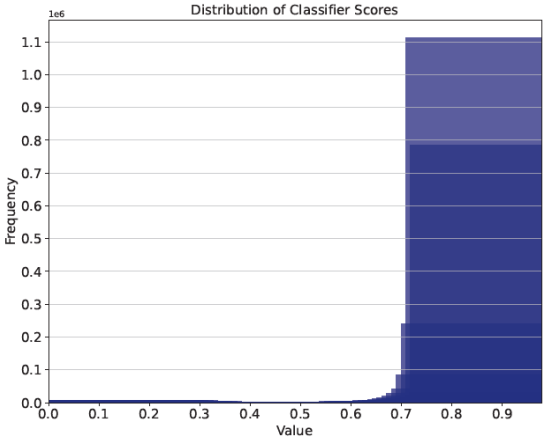

Figure 2a: Distribution of classifier scores.

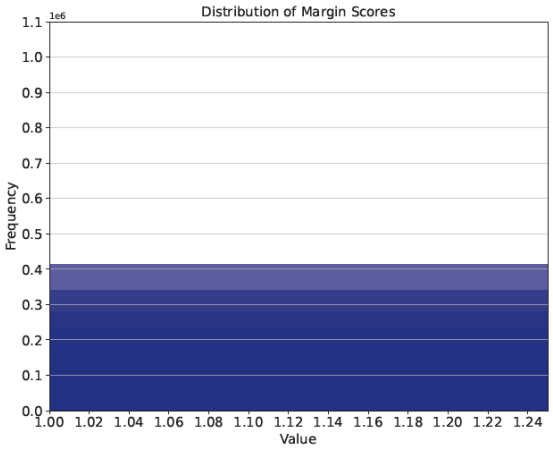

Figure 2b: Distribution of margin distance.

Figure 2 shows the distribution of classifier scores (a) as compared with the distribution of the native to CCMatrix margin scores (b). Whilst the latter demonstrates full uniformity at the given scale, the former exhibits high concentration as scores approach their maximum value.

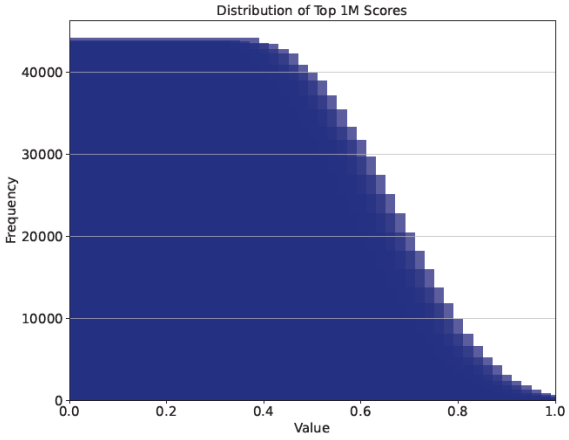

Figure 3: Distribution of the top 1M classifier scores.

In addition, Figure 3 provides a close-up overview of the distribution of the top 1M scores (that is to say, the scores corresponding to the sentences used in the study's translation model). These scores range between 0.967 and 0.977.

Manual evaluation of 20 scored sentence pairs (five with a score of over 0.9 and five with a score of under 0.9 from both the beginning and end of the corpus[9]) shows that classifier scores provide a discernibly better evaluation of sentence parallelism.

6. 3 Translation Models

With a BLEU score of 28.49, the highest scoring model is the one that is trained on 1M parallel sentences and uses the CCMatrix margin distance metric (Table 2). Its classifier-based counterparts score even lower than the randomly selected sample with a BLEU score of 22.28 vs 25.25. A possible reason for better performance of margin-based and randomly selected models as compared with classifier-based ones is the variety of domains and registers that is retained from the

[9] The CCMatrix corpus is ordered in descending order of margin distance scores.

original web-crawled corpus. In contrast, classifier scores, which are derived following training on a corpus of a narrower domain, encourage a focus on a specific type of sentences in addition to a higher level of cleanliness and are likely to have favored sentences "crawled" from the same or related sources.

Translation Model	Size	BLEU Score
Preprocessing + Random	200K	18.24
Preprocessing + Margin-Based	200K	19.85
Preprocessing + Classifier-Based	200K	17.02
Preprocessing + Random	500K	21.10
Preprocessing + Margin-Based	500K	23.91
Preprocessing + Classifier-Based	500K	20.52
Preprocessing + Random	1M	25.25
Preprocessing + Margin-Based	1M	**28.49**
Preprocessing + Classifier-Based	1M	22.28

Table 2: BLEU scores of the NMT models

In contrast, Acarcicek et al.'s (2020) best scoring classifier model increase the shared task's LASER-based baseline by 1.1 and 1.3 points for the two considered language pairs. It is worth noting, however, that overall BLEU scores are significantly lower, the highest results coming at 13.3 (Acarcicek, 2020). It is possible that this difference is partly explainable through the examined languages' characteristics combined with appropriate preprocessing.

7 Conclusion and Future Work

Although the exposed study exhibits high similarity to WMT's corpus filtering shared tasks, several crucial elements that distinguish it should be made note of. Firstly, the English language is not featured in either translation direction, and the examined language pair is not selected merely quantitatively based on its associated resources but is closely associated with the study and its goals. As a result, preprocessing is key within the filtering process. Part of the corpus's preprocessing is language-specific, and a suggested direction for future improvement of the utilised classifier model would involve further application of the two languages' morphological features (such as the use of an alternative, more morphologically-aware fuzzy search algorithm and the inclusion of Universal Dependencies annotations and relations).

Additionally, in this study a customised filtering model benefits from a comparison with one that uses margin scores, thus allowing for specific conclusions to be made, such as the effect of domain-specific data on the machine translation models. In order for this narrowing of the corpus to be avoided, clean multi-domain data could be attained if a manually cleaned portion of CCMatrix is used in training the classifier model.

Also, as performance increases steadily with subcorpora sizes, even larger models should be experimented with.

Importantly, the current work does not claim to propose a high quality translation system in the low-resource Japanese-Bulgarian language pair. Rather, it provides methods for improving the quality of noisy parallel sentences and for the selection of specific portions of higher-quality data. The study may be used as the starting point for further work toward an improved translation model in the described language pair as well as a general frame of reference in terms of a filtering pipeline that can be adapted to other corpora and language pairs.

References

Acarcicek, H. Colakoglu, T., Aktan, P. E., Huang, C., Peng, W. (2020). Filtering Noisy Parallel Corpus Using Transformers with Proxy Task Learning, *Proceedings of the Fifth Conference on Machine Translation (WMT)*.

Bernier-Colborne, G. and Lo, C. (2019). NRC Parallel Corpus Filtering System for WMT 2019, *Proceedings of the Fourth Conference on Machine Translation (WMT)*, Florence, Italy, pp.252-260.

"文" (2022) *Goo*. Available at:

https://dictionary.goo.ne.jp/thsrs/10547/meaning/m0u/文/ (Accessed 14 June 2022).

Chaudhary, V. Tang, Y., Guzmán, F., Schwenk, H., Koehn, P. (2019). Low-Resource Corpus Filtering using Multilingual Sentence Embeddings, *Proceedings of the Fourth Conference on Machine Translation.*

Dodge, J. Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groenveld, D., Mitchell, M., Gardner, M. (2021). Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. arXiv:1907.11692.

"Изречение" (2022) *OnlineRechnik.com*. Available at: m.onlinerechnik.com/duma/Изречение (Accessed 14 June 2022).

Junczys-Dowmunt, M. Grundkiewicz, R., Guha, S., Heafield, K. (2018). Approaching Neural Grammatical Error Correction as a Low-Resource Machine Translation Task, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, pp.595-606.

Khayrallah, H. and Koehn, P. (2018). On the Impact of Various Types of Noise on Neural Machine Translation, *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Melbourne, Australia, pp.74-83.

Koeva, S. Stoyanova, I., Dekova, R., Rizov, B., Genov, A. (2012). Bulgarian X-language Parallel Corpus, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, pp.2480–2486.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). Roberta: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Lo, C. and Joanis, E. (2020). Improving Parallel Data Identification using Iteratively Refined Sentence Alignments and Bilingual Mappings of Pre-Trained Language Models, *Proceedings of the Fifth Conference on Machine Translation*, pp.972–978

Munteanu, D. S. and Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora, *Computational Linguistics*, 31(4):477–504.

Ott, M. Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, Minnesota, pp. 48–53.

Sánchez-Cartagena, V. M. (2018). Prompsit's Submission to WMT 2018 Parallel Corpus Filtering Shared Task, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Belgium, Brussels, pp.955-962.

Schafer, R. et al. (2014). Focused Web Corpus Crawling, *Proceedings of the 9th Web as Corpus Workshop (WAC-9)*.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., Guzmán, F. (2021). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp.1351-1361.

Sen, S. Ekbal, A., Bhattacharyya, P. (2019). Parallel Corpus Filtering Based on Fuzzy String Matching. *Proceedings of the Fourth Conference on Machine Translation.*

"Sentence" (2022) *Cambridge Dictionary*. Available at: https://dictionary.cambridge.org/dictionary/english/sentence (Accessed 14 June 2022).

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS, *Proceedings of the Eighth International Conference on Language Resources.*

Tyagi, S. Chopra, D., Mathur, I., Joshi, N. (2015). Classifier-Based Text Simplification for Improved Machine Translation, *Proceedings of International Conference on Advances in Computer Engineering and Applications*.

Xu, H. and Koehn, P. (2017). Zipporah: a Fast and Scalable Data Cleaning System for Noisy Web-Crawled Parallel Corpora, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp.2945–2950.

Yogi, K. Jha, C. K., Dixit, S. (2015). Classification of Machine Translation Outputs Using NB Classifier and SVM for Post-Editing. doi:10.5121/mlaij.2015.2403

Appendix A Detailed Presentation of Noise in the CCMatrix Coprus (based on a 200-sentence sample)

Type of Noise	% of sentences
Punctuation and capitalisation	11.5
"..."	*5.5*
Capitalisation	*1*
Symbols	*4.5*
Misplaced Punctuation	*0.5*
Numbers/Dates	15
Numbers	*11.5*
Dates	*2*
Years	*1.5*
URLs	1
Long sentences	6
Abbreviations	7
In EN	*2.5*
In BG	*4.5*
Foreign language	5
EN	*4.5*
Other	*0.5*
Machine-translated	4.5
Non-standard language	8.5
Sentence fragments	*7.5*
Typoes	*0.5*
Slang	*0.5*

Table 3: Noise in Bulgarian sentences

Type of Noise	% of sentences
Punctuation	6.5
"..."	*2*
Symbols	*4.5*
Numbers/Dates	16.5
Numbers	*12.5*
Dates	*2.5*
Years	*1.5*
URLs	0.5
Long Sentences	1
Abbreviations (EN)	7.5
Foreign Language	5
EN	*4.5*
Other	*0.5*
Non-standard language	6
Hiragana + kanji/katakana	*2*
Sentence fragments	*3.5*
"Texting" language	*0.5*

Table 3: Noise in Japanese sentences

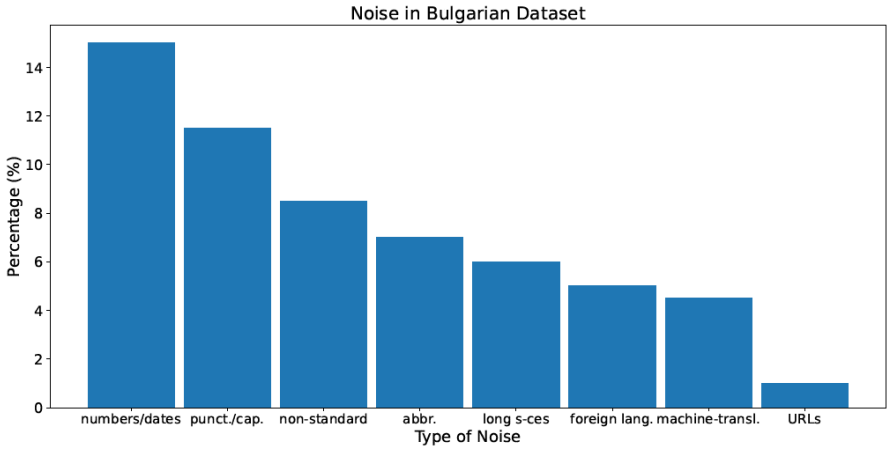

Figure 3: Noise in CCMatrix's Bulgarian sentences by type.

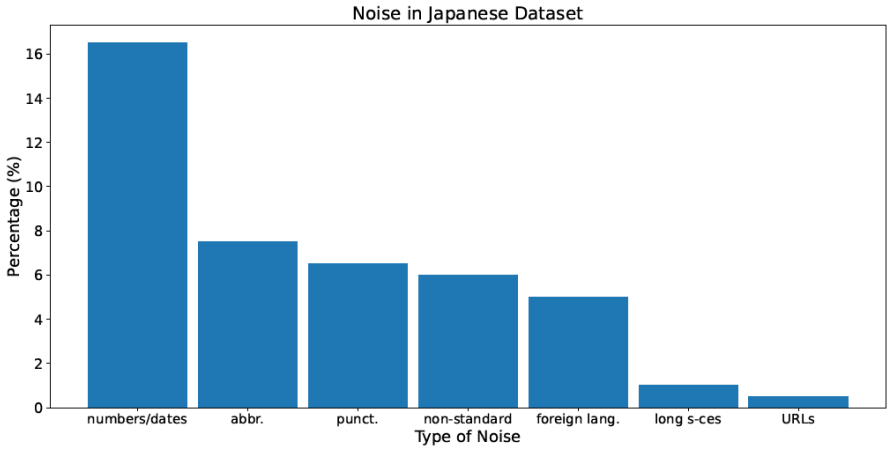

Figure 4: Noise in CCMatrix's Japanese sentences by type.