Adam Mickiewicz University at WMT 2022: NER-Assisted and Quality-Aware Neural Machine Translation

Artur Nowakowski * 1,2 and **Gabriela Pałka** * 1,3 and **Kamil Guttmann** † 1,2 and **Mikołaj Pokrywka** † 1,2
[1] Adam Mickiewicz University, Poznań, Poland
[2] Poleng, Poznań, Poland
[3] Applica.ai, Warsaw, Poland
{artur.nowakowski,gabriela.palka}@amu.edu.pl, {kamgut,mikpok1}@st.amu.edu.pl

Abstract

This paper presents Adam Mickiewicz University's (AMU) submissions to the constrained track of the WMT 2022 General MT Task. We participated in the Ukrainian ↔ Czech translation directions. The systems are a weighted ensemble of four models based on the Transformer (big) architecture. The models use source factors to utilize the information about named entities present in the input. Each of the models in the ensemble was trained using only the data provided by the shared task organizers. A noisy back-translation technique was used to augment the training corpora. One of the models in the ensemble is a document-level model, trained on parallel and synthetic longer sequences. During the sentence-level decoding process, the ensemble generated the n-best list. The n-best list was merged with the n-best list generated by a single document-level model which translated multiple sentences at a time. Finally, existing quality estimation models and minimum Bayes risk decoding were used to rerank the n-best list so that the best hypothesis was chosen according to the COMET evaluation metric. According to the automatic evaluation results, our systems rank first in both translation directions.

1 Introduction

We describe Adam Mickiewicz University's submissions to the constrained track of the WMT 2022 General MT Task. We participated in the Ukrainian ↔ Czech translation directions – a low-resource translation scenario between closely related languages.

The data provided by the shared task organizers was thoroughly cleaned and filtered, as described in section 2.

The approach described in section 3 is based on combining various MT enhancement methods, including transfer learning from a high-resource language pair (Aji et al., 2020; Zoph et al., 2016), noisy back-translation (Edunov et al., 2018), NER-assisted translation (Modrzejewski et al., 2020), document-level translation, model ensembling, quality-aware decoding (Fernandes et al., 2022), and on-the-fly domain adaptation (Farajian et al., 2017).

The results leading to the final submissions are presented in section 4. Additionally, we performed a statistical significance test with paired bootstrap resampling (Koehn, 2004), comparing the baseline solution with the final submission on the test set reference translations released by the shared task organizers. According to the automatic evaluation results based on COMET (Rei et al., 2020) scores, our systems rank first in both translation directions.

2 Data

In the initial stage of system preparation, the sentence-level data was cleaned and filtered using the OpusFilter (Aulamo et al., 2020) toolkit. With the use of the toolkit, language detection filtering based on fastText (Joulin et al., 2016) was performed, duplicates were removed, and heuristics based on sentence length were applied. In particular, we removed sentence pairs with a length ratio over 3 and long sentences (> 200 words). Then, using Moses (Koehn et al., 2007) pre-processing scripts, punctuation was normalized and non-printing characters removed. Finally, the text was tokenized into subword units using SentencePiece (Kudo and Richardson, 2018) with the unigram language model algorithm (Kudo, 2018). For Ukrainian→Czech and Czech→Ukrainian models trained from scratch, we used separate vocabularies for the source and the target language. Each vocabulary consisted of 32,000 units.

We used concatenated data from the Flores-101 (Goyal et al., 2022) benchmark (flores101-dev, flores101-devtest) for our development set, as pro-

[*] AN and GP contributed equally.
[†] KG and MP contributed equally.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 326 - 334
December 7-8, 2022 ©2022 Association for Computational Linguistics

Data type		Sentences	Corpora
Monolingual cs	available	448,528,116	News crawl, Europarl v10, News Commentary, Common Crawl, Extended Common Crawl, Leipzig Corpora
	used	59,999,553	
Monolingual uk	available	70,526,415	News crawl, UberText Corpus, Leipzig Corpora, Legal Ukrainian
	used	59,152,329	
Parallel cs-uk	available	12,630,806	OPUS, WikiMatrix, ELRC – EU acts in Ukrainian
	used	8,623,440	

Table 1: Statistics of the total available corpora and the corpora used for system training after filtering.

vided by the task organizers.

Table 1 shows statistics for the total available corpora in the constrained track and the corpora used for system training after filtering.

3 Approach

We used the Marian (Junczys-Dowmunt et al., 2018) toolkit for all of our experiments. Our model architecture follows the Transformer (big) (Vaswani et al., 2017) settings. For all model training, we used 4x NVIDIA A100 80GB GPUs.

3.1 Transfer Learning

For our initial experiments, we used transfer learning (Aji et al., 2020; Zoph et al., 2016) from the high-resource Czech→English language pair. We used only the parallel data provided by the organizers to train the model in this direction. In this case, we created a single joint vocabulary for three languages (Czech, English, Ukrainian), consisting of 32,000 units. The Czech→English model was fine-tuned for the Ukrainian→Czech and Czech→Ukrainian language directions. Our later experiments showed that there were no gains in translation quality compared with models trained from scratch using separate vocabularies for source and target languages – the upside was that the models took less time to converge during training.

3.2 Noisy Back-Translation

We used models created by the transfer learning approach to produce synthetic training data through noisy back-translation (Edunov et al., 2018). Specifically, we applied Gumbel noise to the output layer and sampled from the full model distribution. We used monolingual data available in the constrained track, which included all ~59M Ukrainian sentences after filtering and ~60M randomly selected Czech sentences.

After training the model with concatenated parallel and back-translated corpora, we replaced the training data with filtered parallel data and further fine-tuned the model. We kept the same settings as in the first training pass, training the model until it converged on the development set.

3.3 NER-Assisted Translation

Translation in domains such as news, social or conversational texts, and e-commerce is a specialized task, involving such challenges as localization, product names, and mentions of people or events in the content of documents. In such a case, it proved helpful to use off-the-shelf solutions for recognizing named entities. For Czech, the Slavic BERT model (Arkhipov et al., 2019) was used, with which entities such as persons (PER), locations (LOC), organizations (ORG), products (PRO), and events (EVT) were tagged. Due to the lack of support for the Ukrainian language in the Slavic BERT model, the Stanza Named Entity Recognition module (Qi et al., 2020) was used to detect entities in the Ukrainian text, recognizing persons (PER), locations (LOC), organizations (ORG), and miscellaneous items (MISC). With these ready-made solutions, the parallel and back-translated corpora were tagged. The named entity categories were then numbered to assign appropriate source factors to words in the text, supporting the translation process. The source factors were later transferred to subwords in a trivial way.

Source factors (Sennrich and Haddow, 2016) have previously been used to take into account various characteristics of words during the translation process. For example, morphological information, part-of-speech tags, and syntactic dependencies have been added as input to neural machine translation systems to improve the translation quality.

In the same way, it is possible to add information about named entities found in the text (Modrzejewski et al., 2020), making it easier for the model to translate them correctly. However, the AMU machine translation system does not dis-

```
Hlavní|p0 inspektor|p0 organizace|p0 RSPCA|p3 pro|p0 Nový|p2 Jižní|p2 Wales|p2
David|p1 O'Shannessy|p1 televizi|p0 ABC|p5 sdělil|p0 ,|p0 že|p0 dohled|p0 nad|p0
jatky|p0 a|p0 jejich|p0 kontroly|p0 by|p0 měly|p0 být|p0 v|p0 Austrálii|p2
samozřejmostí|p0 .|p0

_Hlavní|p0 _inspektor|p0 _organizace|p0 _R|p3 SP|p3 CA|p3 _pro|p0 _Nový|p2 _Jižní|p2
_Wales|p2 _David|p1 _O|p1 '|p1 S|p1 han|p1 ness|p1 y|p1 _televizi|p0 _A|p5 BC|p5
_sdělil|p0 ,|p0 _že|p0 _dohled|p0 _nad|p0 _ja|p0 tky|p0 _a|p0 _jejich|p0 _kontroly|p0
_by|p0 _měly|p0 _být|p0 _v|p0 _Austrálii|p2 _samozřejmost|p0 í|p0 .|p0
```

Figure 1: An example of a sentence tagged with NER source factors before and after subword encoding.

	cs				**uk**			
Category	**train-bt**	**train-parallel**	**dev**	**test**	**train-bt**	**train-parallel**	**dev**	**test**
PER	33,633,602	1,545,658	747	306	30,778,893	1,623,370	827	478
LOC	24,552,404	1,954,319	1,191	454	18,178,736	1,912,604	1,197	771
ORG	29,380,436	1,997,685	566	314	24,117,485	2,221,371	544	606
MISC	-	-	-	-	4,140,394	893,867	168	76
PRO	5,452,326	1,104,860	172	59	-	-	-	-
EVT	1,150,301	111,563	83	10	-	-	-	-

Table 2: The number of recognized named entity categories in the training, development and test data. The training data statistics are split into *train-bt*, which was created by noisy back-translation, and *train-parallel*, which is the filtered parallel training data.

tinguish between inside-outside-beginning (IOB) tags (Ramshaw and Marcus, 1995), treating the named entity tag names as a whole. Specifically, we introduce the following source factors:

- p0: source factor denoting a normal token,
- p1: source factor denoting the PER category,
- p2: source factor denoting the LOC category,
- p3: source factor denoting the ORG category,
- p4: source factor denoting the MISC category,
- p5: source factor denoting the PRO category,
- p6: source factor denoting the EVT category.

An example of a tagged sentence is shown in Figure 1.

Models were trained in two settings: concatenation and sum. In the first setting, the factor embedding had a size of 16 and was concatenated with the token embedding. In the second setting, the factor embedding was equal to the size of the token embedding (1024) and was summed with it.

As shown in Table 4, we observe an increase in the string-based evaluation metrics (chrF and BLEU) while COMET scores remain about the same. This is in accordance with Amrhein and Sennrich (2022), who show that COMET models are not sufficiently sensitive to discrepancies in named entities.

Table 2 presents the numbers of recognized named entity categories in the training, development and test data.

3.4 Document-Level Translation

Our work on document-level translation is based on a simple data concatenation method, similar to Junczys-Dowmunt (2019) and Scherrer et al. (2019).

As our training data, we use parallel document-level datasets (GNOME, KDE4, TED2020, QED), as well as synthetically created data, concatenating random sentences to match the desired input length. Specifically, we merge datasets created in the following ways as a single, large dataset:

- Curr → Curr: sentence-level parallel data,
- Prev + Curr → Prev + Curr: previous sentence given as a context,
- 50T → 50T: a fixed window of 50 tokens after subword encoding,

```
Netvrdím, že bakteriální celulóza jednou nahradí bavlnu, kůži, nebo jiné látky.
<SEP> Ale myslím, že by to mohl být chytrý a udržitelný přírůstek k našim stále
vzácnějším přírodním zdrojům. <SEP> Možná že se nakonec tyto bakterie neuplatní
v módě, ale jinde. <SEP> Zkuste si třeba představit, že si vypěstujeme lampu,
židli, auto, nebo třeba dům. <SEP> Má otázka tedy zní: Co byste si v budoucnu
nejraději vypěstovali vy?
```

Figure 2: An example document consisting of five sentences separated with <SEP> tags.

- 100T → 100T: a fixed window of 100 tokens after subword encoding,
- 250T → 250T: a fixed window of 250 tokens after subword encoding,
- 500T → 500T: a fixed window of 500 tokens after subword encoding.

By concatenating such datasets, we allow the model to gradually learn how to translate longer input sequences. It is also capable of sentence-level translation. To separate sentences from each other, we introduced a <SEP> tag. An example of a document-level input sequence is shown in Figure 2. All data used to train the document-level model were tagged with NER source factors, including the back-translated data.

3.5 Weighted Ensemble

We created a weighted ensemble of four best-performing models. It consisted of the following model types:

- (A) sentence-level models trained with NER source factors (concat 16),
- (B) sentence-level model trained with NER source factors (sum),
- (C) document-level model trained with NER source factors (concat 16).

In this case, the document-level model was used only for the sentence-level translation. The optimal weights for each model were selected using a grid search method. For the specific language pairs, we used the following model and weight combinations:

- Czech → Ukrainian: $1.0 \cdot (2{\times}A) + 0.8 \cdot (B) + 0.6 \cdot (C)$,
- Ukrainian → Czech: $1.0 \cdot (2{\times}A) + 0.8 \cdot (B) + 0.4 \cdot (C)$.

3.6 Quality-Aware Decoding

Having the final model ensemble, we created an n-best list containing 200 translations for each sentence with beam search. Then we merged it with a second n-best list containing 50 translations for each sentence, created by a single document-level model with document-level decoding. The idea behind it was that the hypotheses produced by the document-level decoding take into account the context of surrounding sentences, which is not the case with the ensemble. This enabled the use of quality-aware decoding (Fernandes et al., 2022).

We applied a two-stage quality-aware decoding mechanism: pruning hypotheses using a tuned reranker (T-RR) and minimum Bayes risk (MBR) decoding (Kumar and Byrne, 2002, 2004), as shown in Figure 3.

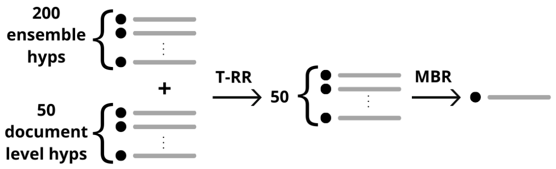

Figure 3: A two-stage (T-RR → MBR) quality-aware decoding process. 200 hypotheses generated by the ensemble are merged with 50 hypotheses generated by the document-level model. A tuned reranker is used to prune the total number of hypotheses to 50, and these are then used as input for minimum Bayes risk decoding.

First, we tuned a reranker on the development set, using as features NMT model scores, as well as existing QE models based on TransQuest (Ranasinghe et al., 2020) and COMET (Rei et al., 2020), which are based on Direct Assessment (DA) (Graham et al., 2013) scores or MQM (Lommel et al., 2014) scores. Specifically, we used:

- model ensemble log-likelihood $\log p_\theta(y|x)$ scores,

- TransQuest QE model trained on DA scores (`monotransquest-da-multilingual`),
- COMET QE model trained on MQM scores (`wmt21-comet-qe-mqm`),
- COMET QE model trained on DA scores (`wmt21-comet-qe-da`).

We tuned the feature weights to maximize the COMET reference-based evaluation metric value using MERT (Och, 2003).

After tuning the reranker, we used it to prune the n-best list from 250 to 50 hypotheses per input sentence. The resulting n-best list was used for minimum Bayes risk decoding, using the COMET reference-based metric as the utility function. Minimum Bayes risk decoding seeks, from the set of hypotheses, the hypothesis with the highest expected utility.

$$\hat{y}_{\text{MBR}} = \arg\max_{y \in \bar{\mathcal{Y}}} \underbrace{\mathbb{E}_{Y \sim p_\theta(y|x)}[u(Y, y)]}_{\approx \frac{1}{M}\sum_{j=1}^{M} u(y^{(j)}, y)} \tag{1}$$

Equation 1 shows that the expectation can be approximated as a Monte Carlo sum using model samples $y^{(1)}, \ldots, y^{(M)} \sim p_\theta(y|x)$. In practice, the translation with the highest expected utility can be chosen by comparing each hypothesis $y \in \bar{\mathcal{Y}}$ with all other hypotheses in the set.

The described two-stage quality-aware decoding process allowed us to further optimize our system for the COMET evaluation metric, which has been shown to have a high correlation with human judgements (Kocmi et al., 2021).

3.7 Post-Processing

The final step involved post-processing. We applied the following post-processing steps for each best obtained translation:

- transfer of emojis from the source to the translation using word alignment based on SimAlign (Jalili Sabet et al., 2020),
- restoration of quotation marks appropriate for a given language,
- restoration of capitalization (e.g. if the source sentence was fully uppercased),
- restoration of punctuation, exclamation and question marks (if a source sentence ends with such a mark, we make the translation do likewise),
- replacement of three consecutive dots with an ellipsis,
- restoration of bullet points and enumeration (e.g. if the source sentence starts with a number or a bullet point),
- deletion of consecutively repeated words.

Approach	**Sim. score**	**COMET**	**chrF**
Baseline	-	0.8322	0.5263
Default	0.4	0.8316	0.5260
Best-334	0.19	0.8322	0.5259
Best-133	0.25	0.8323	0.5262

Table 3: Results of the on-the-fly adaptation method on the development set. The *default* approach is based on Farajian et al. (2017). However, only 11 sentence pairs were found in this scenario. The experiments denoted as *best-334* and *best-133* used the learning rate values of 0.002 and 10 epochs. In our development set containing 2009 sentence pairs, 334 matching sentences were found in *best-334* and 133 in *best-133*.

3.8 On-The-Fly Domain Adaptation

The General MT Task tests the MT system's performance on multiple domains. Therefore, we investigated the possibility of improving our translation system with the on-the-fly domain adaptation method.

This experiment was based on Farajian et al. (2017). Our idea was to retrieve similar sentences from the training data for each input sentence and to fine-tune the model on their translations. After the translation of a single sentence is complete, the model is reset to the original parameters. We used Apache Lucene (McCandless et al., 2010) as our translation memory to search for similar sentences. We indexed all of the training data and used the Marian dynamic adaptation feature. We compared the translation quality with and without the retrieved context. The experiments were carried out with a different similarity score used to choose similar sentence pairs for the fine-tuning process. We empirically modified the learning rate and the number of epochs to find optimal values that improved the translation quality.

Table 3 shows the results of the aforementioned experiments on the full development set. We found

System		uk→cs			cs→uk		
		COMET	chrF	BLEU	COMET	chrF	BLEU
Baseline (transformer-big)		0.8622	0.5229	24.29	0.7818	0.5175	22.64
+back-translation		0.9053	0.5309	25.41	0.8356	0.5280	23.14
+ner	concat 16	0.9003	0.5314	25.62	0.8362	0.5309	24.28
	sum	0.8991	0.5323	25.87	0.8421	0.5302	23.91
+fine-tune	concat 16	0.9021	0.5344	25.94	0.8387	0.5330	24.51
	sum	0.8990	0.5357	25.99	0.8456	0.5321	24.24
+ensemble		0.9066	0.5376	**26.36**	0.8522	0.5373	**24.85**
+quality-aware		0.9874	0.5376	25.42	0.9238	0.5384	24.50
+post-processing		**0.9883**	**0.5392**	25.89	**0.9240**	**0.5388**	24.63
Document-level	sent-level dec.	0.8942	0.5326	25.47	0.8350	0.5289	23.92
	doc-level dec.	0.8920	0.5324	25.44	0.8356	0.5297	23.78

Table 4: Results of COMET, chrF and BLEU automatic evaluation metrics on the concatenated datasets flores101-dev and flores-101-devtest. ChrF and BLEU metrics were computed with sacreBLEU. Document-level model evaluation includes added back-translation, NER source factors (concat 16) and fine-tuning.

System	uk→cs			cs→uk		
	COMET	chrF	BLEU	COMET	chrF	BLEU
Baseline (transformer-big)	0.8315	0.5627	31.79	0.8008	0.5849	31.43
Final submission	**1.0488**	**0.6066**	**37.03**	**0.9944**	**0.6153**	**34.74**

Table 5: Results of COMET, chrF and BLEU automatic evaluation metrics on the test set. ChrF and BLEU metrics were computed with sacreBLEU. The final submission results are statistically significant ($p < 0.05$).

that only a small number of sentences in the training data were similar to those present in the development set. The results showed that tuning the model on similar sentences from the training data did not significantly improve translation quality. In the end, we decided not to use this method in our WMT 2022 submission.

4 Results

The results of our experiments are presented in Table 4. We evaluated our models with the COMET[1] (Rei et al., 2020), chrF (Popović, 2015) and BLEU (Papineni et al., 2002) automatic evaluation metrics. ChrF and BLEU scores were computed with the sacreBLEU[2,3] (Post, 2018) tool. We also include scores for the document-level model. In this case, the scores include improvements added by back-translation, NER source factors and fine-tuning. The document-level evaluation was split into sentence-level decoding and document-level decoding. In the first scenario, the model translates a single sentence at a time, which is not different from a sentence-level model. In the second scenario, the model translates concatenated chunks of at most 250 subword tokens at a time.

We found that the largest gain in the COMET value was achieved due to the quality-aware decoding method, at the cost of BLEU value. The chrF value remained the same in the Ukrainian→Czech translation direction, while it increased slightly in the Czech→Ukrainian direction. As discussed in section 3.3, the inclusion of NER source factors helped the model with the translation of named entities, which is not well reflected in the COMET value, as this metric is not sufficiently sensitive to discrepancies in named entities (Amrhein and Sennrich, 2022).

Table 5 shows results for our final submissions compared with the baseline. We performed a statistical significance test with paired bootstrap resampling (Koehn, 2004), running 1000 resampling trials to confirm that our submissions are statistically significant ($p < 0.05$).

5 Conclusions

We describe Adam Mickiewicz University's (AMU) submissions to the WMT 2022 General

[1] COMET scores were computed with the `wmt20-comet-da` model.

[2] BLEU signature: nrefs:1|case:mixed|eff:no|tok:13a |smooth:exp|version:2.0.0

[3] chrF signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:0 |space:no|version:2.0.0

MT Task in the Ukrainian ↔ Czech translation directions. Our experiments cover a range of MT enhancement methods, including transfer learning, back-translation, NER-assisted translation, document-level translation, weighted ensembling, quality-aware decoding, and on-the-fly domain adaptation. We found that using a combination of these methods on the test set leads to a +0.22 (26.13%) increase in COMET scores in the Ukrainian→Czech translation direction and a +0.19 (24.18%) increase in the Czech→Ukrainian direction, compared with the baseline model. According to the COMET automatic evaluation results, our systems rank first in both translation directions.

References

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710, Online. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. *arXiv preprint arXiv:2202.05148*.

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137, Copenhagen, Denmark. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José De Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. 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. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2002. Minimum Bayes-risk word alignments of bilingual texts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 140–147. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*, (12):455–463.

M. McCandless, E. Hatcher, and O. Gospodnetić. 2010. *Lucene in Action*. Manning Pubs Co Series. Manning.

Maciej Modrzejewski, Miriam Exel, Bianka Buschbeck, Thanh-Le Ha, and Alexander Waibel. 2020. Incorporating external annotation to improve named entity translation in NMT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 45–51, Lisboa, Portugal. European Association for Machine Translation.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics – Volume 1*, ACL '03, page 160–167, USA. Association for Computational Linguistics.

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

Maja Popović. 2015. 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. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. 2019. Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine*

Translation: Volume 1, Research Papers, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.