Zero-Shot Translation Quality Estimation with Explicit Cross-Lingual Patterns

Lei Zhou[§] **Liang Ding**[†] **Koichi Takeda**[§]

[§]FVCRC, Graduate School of Informatics, Nagoya University
{zhou.lei@a.mbox,takedasu@i}.nagoya-u.ac.jp
[†]UBTECH Sydney AI Centre, School of Computer Science
Faculty of Engineering, The University of Sydney
ldin3097@uni.sydney.edu.au

Abstract

This paper describes our submission of the WMT 2020 Shared Task on Sentence Level Direct Assessment, Quality Estimation (QE). In this study, we empirically reveal the *mismatching issue* when directly adopting BERTScore (Zhang et al., 2020) to QE. Specifically, there exist lots of mismatching errors between source sentence and translated candidate sentence with token pairwise similarity. In response to this issue, we propose to expose explicit cross lingual patterns, *e.g.* word alignments and generation score, to our proposed zero-shot models. Experiments show that our proposed QE model with explicit cross-lingual patterns could alleviate the mismatching issue, thereby improving the performance. Encouragingly, our zero-shot QE method could achieve comparable performance with supervised QE method, and even outperforms the supervised counterpart on 2 out of 6 directions. We expect our work could shed light on the zero-shot QE model improvement.

1 Introduction

Translation quality estimation (QE) (Blatz et al., 2004; Specia et al., 2018, 2020) aims to predict the quality of translation hypothesis without golden-standard human references, setting it apart from reference-based translation metrics. Existing reference-based evaluation metrics, e.g. BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), NIST (Doddington, 2002), ROUGE (Lin, 2004), TER (Snover et al., 2006), are commonly used in language generation tasks including translation, summarization, and captioning but all heavily rely on the quality of given references.

Recently, (Edunov et al., 2020) show that reference-based automatic evaluation metrics, e.g., BLEU, are not always reliable because the human translated references are translationese (Koppel and
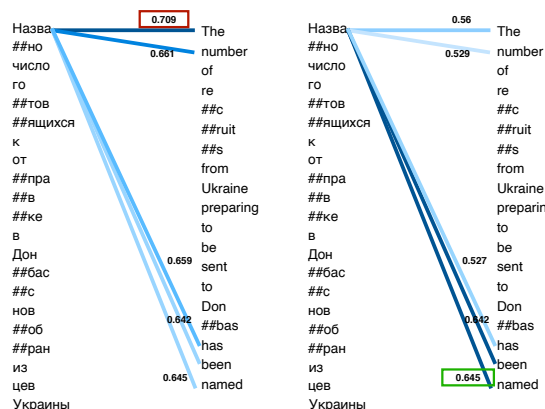

Figure 1: Example of mismatching error, Russian→English. On the left, token "Назва" is mismatched to "The" with the maximal probability (within the red rectangle) only. On the right, guided by our proposed cross-lingual patterns, "Назва" is correctly matched to the token "named" with the maximal probability (within the green rectangle.)

Ordan, 2011; Graham et al., 2019). Thus, an automatic method with no access to any references, i.e., QE, is highly appreciated.

In this paper, we mainly focus on sentence level QE metrics, where existing studies categorize it into two classes: 1) supervised QE with human assessment as supervision signal: a feature extractor stacked with an estimator (Yankovskaya et al., 2019; Wang et al., 2016b; Fan et al., 2019); 2) unsupervised QE without human assessment, which normally based on the pre-trained word embeddings, for example, YISI (Lo, 2019) and BERTScore (Zhang et al., 2020). Our work follows the latter, where we adopt BERTScore (Zhang et al., 2020) without extra fine-tuning. In particular, we implement our approach upon the pre-trained multilingual BERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019).

We first empirically reveal the *mismatching issue* when directly adopting BERTScore (Zhang

et al., 2020) to QE task. Specifically, there exist lots of mismatching errors between source tokens and translated candidate tokens when performing greedy matching with pairwise similarity. Figure 1 shows an example of the mismatching error, where the Russian token "Назва" is mismatched to the English token "The" due to lacking of proper guidance.

To alleviate this issue, we design two explicit cross-lingual patterns to augment the BERTScore as a QE metric:

- CROSS-LINGUAL ALIGNMENT MASKING: we design an alignment masking strategy to provide the pairwise similarity matrix with extra guidance. The alignment is derived from GIZA++ (Och and Ney, 2003).
- CROSS-LINGUAL GENERATION SCORE: we obtain the perplexity, dubbed *ppl*, of each target side token by force decoding with a pre-trained cross-lingual model, e.g. multilingual BERT and XLM. This generation score is weighted added on the similarity score.

2 Methods

2.1 BERTScore as Backbone

A pre-trained multilingual model generates contextual embeddings of both source sentence and translated candidate sentence, such that this pair of sentences in different language can be mapped to the same continuous feature space. Given a source sentence $x = \langle x_1, \ldots, x_k \rangle$, the model generates a sequence of vectors $\langle \mathbf{x_1}, \ldots, \mathbf{x_k} \rangle$ while the candidate $\hat{y} = \langle \hat{y}_1, \ldots, \hat{y}_l \rangle$ is mapped to $\langle \mathbf{\hat{y}_1}, \ldots, \mathbf{\hat{y}_l} \rangle$. Different from the reference-based BERTScore, where they compute the pairwise similarity between reference sentence and translated candidate sentence, we calculate the pairwise similarity between the source and translated candidate with dot-production, i.e., $\mathbf{x_i}^\top \mathbf{\hat{y}_j}$. We adopt greedy matching to force each source token to be matched to the most similar target token in the translated candidate sentence. The QE function based on BERTScore backbone therefore can be formulated as:

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{y}_j \in \hat{y}} \mathbf{x_i}^\top \mathbf{\hat{y}_j},$$

$$P_{\text{BERT}} = \frac{1}{|\hat{y}|} \sum_{\hat{y}_j \in \hat{y}} \max_{x_i \in x} \mathbf{x_i}^\top \mathbf{\hat{y}_j}, \qquad (1)$$

$$F_{\text{BERT}} = 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}.$$

where R_{BERT}, P_{BERT} and F_{BERT} are inherited from Zhang et al. (2020), representing Recall rate, Precision rate and F-score, respectively.

2.2 Alignment Masking Strategy

With aforementioned QE function, we can follow Zhang et al. (2020) to obtain the distance between the source sentence and translated candidate sentence via directly adding up the maximum similarity score of each token pair. However, because there exist lots of mismatching errors (as shown in Figure 1), above sentence-level similarity calculation may be sub-optimal. Moreover, Zhang et al. (2020)'s calculation is suitable for monolingual scenario, which may be insensitive for cross-lingual computation. Thus, we propose to augment our QE metric with more cross-lingual signals.

Inspired by Ding et al. (2020), where they show it's possible to augment cross-lingual modeling by leveraging cross-lingual explicit knowledge. we therefore employ word alignment knowledge from external models, e.g., GIZA++[1], as additional information.

Alignment masking Both BERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019) utilize BPE tokenization (Sennrich et al., 2016). It should be noted that in this paper, by word alignment we mean alignment of BPE tokenized word and subword units. Given a tokenized source sentence x and candidate sentence $\hat{y}$, alignment (Och and Ney, 2003) is defined as a subset of the Cartesian product of position, $\mathcal{A} \subseteq \{(i,j) : i = 1, \ldots, k; j = 1, \ldots, l\}$. Alignment results represented by $\mathcal{M}$ is defined as:

$$\mathcal{M} = \begin{cases} 1 & (i,j) \in \mathcal{A} \\ 0 \leq a \leq 1 & \text{otherwise} \end{cases} \qquad (2)$$

$\mathcal{M}$ is a penalty function over the similarity of unaligned tokens. It's a mask like matrix to assign a penalty weight a[2] to the similarity of unaligned tokens while keeping that of aligned ones unchanged, as illustrated in Figure 2. Thus, greedy matching is performed on a renewed similarity matrix, which is defined as the average of $\mathbf{x_i}^\top \mathbf{\hat{y}_j}$ and masked $\mathbf{x_i}^\top \mathbf{\hat{y}_j}$ by word alignment. For example, R_{BERT}

---

[1] https://github.com/moses-smt/giza-pp
[2] In our preliminary studies, $a = 0.8$ picking from $\{0.0, 0.2, 0.4, 0.8, 1.0\}$ performs best, which then leaves as the default setting in the following experiments.

#	Metrics	en-de	en-zh	ro-en	et-en	ne-en	si-en	ru-en
1	Baseline (test)	0.146	0.190	0.685	0.477	0.386	0.374	0.548
2	BERT	**0.120**	0.167	0.650	0.306	0.475	-	**0.354**
3	BERT (align)	0.091	0.170	0.672	0.307	**0.478**	-	0.340
4	BERT (ppl)	0.068	0.187	0.671	0.321	0.468	-	0.311
5	BERT (align+ppl)	0.099	**0.189**	**0.677**	**0.324**	0.477	-	0.332

Table 1: Pearson correlations with sentence-level Direct Assessment (DA) scores. The results of supervised baseline (Kepler et al., 2019), provided by the organizer, show it's agreement with DA scores on the test set of WMT20 QE. As DA scores on test set aren't available at this point, we report our experiment results on valid set.

#	Metrics	en-de	en-zh	ro-en	et-en	ne-en	si-en	ru-en
1	BERT	**0.143**	0.131	0.389	0.217	0.318	-	**0.259**
2	BERT (align)	0.122	0.133	0.422	0.219	**0.322**	-	0.251
3	BERT (ppl)	0.105	0.145	0.416	0.225	0.315	-	0.240
4	BERT (align+ppl)	0.132	**0.152**	**0.439**	**0.228**	0.320	-	0.247

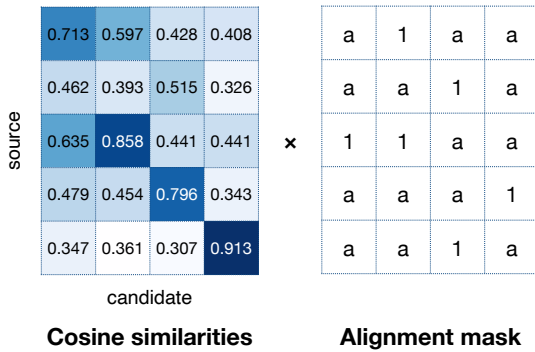Table 2: Kendall correlations with sentence-level Direct Assessment (DA) scores.

Figure 2: Word alignment as a mask matrix

is changed into:

$$R_{\text{BERT(align)}} = \frac{1}{2|x|} \sum_{x_i \in x} \max_{\hat{y}_j \in \hat{y}} (\mathbf{x_i}^\top \hat{\mathbf{y}}_\mathbf{j} + \mathcal{M} \cdot \mathbf{x_i}^\top \hat{\mathbf{y}}_\mathbf{j}) \tag{3}$$

which can be characterized as balancing our proposed extra explicit cross-lingual patterns, i.e., word alignment.

2.3 Generation Score

In additional to token similarity score, we introduce force-decoding perplexity of each target token as a cross-lingual generation score. For better coordination and considering our cross-lingual setting, we use the same pre-trained cross-lingual model, e.g. multilingual BERT, for both token embedding extraction and masked language model (MLM) perplexity generation. This cross-lingual generation

score is added as:

$$F_{\text{BERT(ppl)}} = (1 - \lambda) * F_{\text{BERT}} + \lambda * ppl_{\text{MLM}} \tag{4}$$

where the λ can be seen as a variable that regulates the interpolation ratio between F_{BERT} and our proposed ppl_{MLM}, making the generation score after combination more wisely. The effect of λ will be discussed in the experiments.

3 Experimental Results

3.1 Data

Main experiments were conducted on the WMT20 QE Shared Task, Sentence-level Direct Assessment language pairs. The task contains 7 directions, including:

- English→German (**en-de**)
- English→Chinese (**en-zh**)
- Romanian→English (**ro-en**)
- Estonian→English (**et-en**)
- Nepalese→English (**ne-en**)
- Sinhala→English (**si-en**)
- Russian→English (**ru-en**)

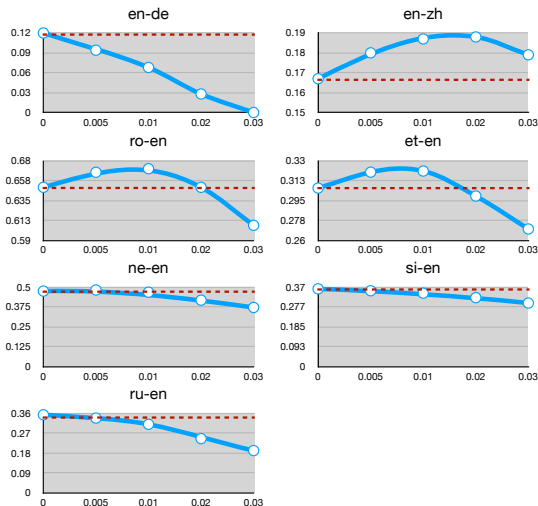Each of them consists of 7K training data, 1K validation data and 1K test data.

Figure 3: Pearson correlations with Direct Assessment (DA) scores when $\lambda \in [0, 0.03]$.

3.2 Setup

Based on our proposed QE metric in Section 2.1, we conduct the validataion and main experiments with two pre-trained cross-lingual model: bert-base-multilingual-cased[3] (12-layer, 768-hidden, 12-heads, trained on 104 languages) and xlm-mlm-100-1280[4] (16-layer, 1280-hidden, 16-heads, trained on 100 languages) for both contextual embedding representation and generation score. The 9th layer of multilingual BERT and the 11th of XLM are used to generate contextual embedding representations. Furthermore, we obtain bidirectional word alignment of all the training, validation and test dataset with GIZA++. Notably, this work is a zero-shot approach that doesn't involve training on Direct Assessment (DA) scores, which makes our method suitable for real industry scenarios.

3.3 Ablation Study

In order to maximize the advantages of our proposed method for zero-shot translation QE, we conducted extensive ablation studies. We report the results of ablation studies on the validation dataset.

Effect of λ We conduct ablation studies to empirically decide the value of of λ in Equation 4 when introducing generation scores. We observe positive effect of proper weighted additional generation score on **en-zh**, **ro-en**, **et-en**, **ne-en**, **si-en**.

---

[3]https://huggingface.co/
bert-base-multilingual-cased
[4]https://huggingface.co/
xlm-mlm-100-1280

	mBERT	XLM
en-de	0.120	0.056
en-zh	0.167	0.008
ro-en	0.650	0.568
et-en	0.306	0.254
ne-en	0.475	0.398
si-en	-	0.362
ru-en	0.354	0.228

Table 3: This is a comparison between multilingual BERT ("mBERT") and XLM in terms of the Pearson correlations with Direct Assessment (DA) scores. Multilingual BERT performs better than XLM.

As illustrated in Figure 3, considering the average performance, we pick $\lambda = 0.01$ from $[0, 0.03]$.

Effect of different pretrained models We also investigated the effect to deploy our proposed fixed cross-lingual patterns on different state-of-the-art large scale pre-trained models, e.g., XLM (Conneau and Lample, 2019) (xlm-mlm-100-1280), BERT (Zhang et al., 2020) (bert-base-multilingual-cased). Table 3 lists a comparison of multilingual BERT and XLM in terms of the Pearson correlations with Direct Assessment (DA) scores. As seen, multilingual BERT outperforms XLM on almost all language pairs, excepting for **si-en**. One possible reason is that multilingual BERT is not pre-trained on Sinhala corpus while XLM does. In this end, we generate our final submission with XLM in **si-en** direction, and with multilingual BERT in other directions.

3.4 Main Results

In the main experiments, we evaluate the agreement of our approach with Direct Assessment (DA) scores on validation dataset, as DA scores of the test set are not available at this point. Baseline results, which are evaluated on test set though, are also listed for general comparison.

As shown in Table 1, our method could achieve improvements on 4 out of 6 directions, including **en-zh**, **ro-en**, **et-en** and **ne-en**. Particularly, combination of two strategies, i.e., CROSS-LINGUAL ALIGNMENT and CROSS-LINGUAL GENERATION SCORE, could achieve better performance on **en-zh**, **ro-en** and **et-en** directions.

Besides Pearson correlations, we also calculated Kendall correlations for all language pairs. As seen in Table 2, the trends of Kendall correlations

	Ours	Kepler et al. (2019)
en-de	0.111	0.146
en-zh	0.085	0.190
ro-en	0.650	0.685
et-en	-	-
ne-en	**0.488**	0.386
si-en	**0.388**	0.374
ru-en	0.400	0.548

Table 4: Comparison of our submission and supervised baseline (Kepler et al., 2019) on WMT20 sentence-level QE official test set, in terms of Pearson correlations.

are same as Pearson correlations, validating the effectiveness of our proposed methods.

3.5 Official Evaluations

The official automatic evaluation results of our submissions for WMT 2020 are presented in Table 4. We participated QE (Sentence-Level Direct Assessment) in following language pairs: **en-de**, **en-zh**, **ro-en**, **ne-en**, **si-en**, **ru-en**, except for **et-en**. From the official evaluation results (Specia et al., 2020) in terms of absolute Pearson Correlation, our submission achieves higher performance than supervised baseline (Kepler et al., 2019) in **ne-en** and **si-en** (As shown in Table 4).

Encouragingly, our proposed zero-shot QE metric could achieve comparable performance with supervised QE method, and even outperforms the supervised counterpart on 2 out of 6 directions.

4 Related Work

MT evaluation Taking sentence-level evaluation as an example, reference-based metrics describe to which extend a candidate sentence is similar to a reference one (Sellam et al., 2020). BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), NIST (Doddington, 2002), ROUGE (Lin, 2004) measure such similarity through n-gram matching, which is restricted to the exact form of sentences. TER (Snover et al., 2006) and CHARACTER (Wang et al., 2016b) use edit distance at word or character level to indicate the distance between candidate and reference. Different from these metrics that are restricted to the exact form of sentences, recent dominated neural model metrics *learn* to evaluate with human assessment as supervision signal, such as BEER (Stanojević and Sima'an, 2014) and RUSE (Shimanaka et al., 2018), or oth-

ers as YiSi (Lo, 2019) and BERTScore (Zhang et al., 2020) , evaluate with pre-trained word embedding, without using human assessment.

Incorporating Explicit Knowledge Several approaches have incorporated pre-defined or learned features into neural networks. Tai et al. (2015) demonstrate that incorporating structured semantic information could enhance the representations. Sennrich and Haddow (2016) feed the encoder cell combined embeddings of linguistic features including lemmas, subword tags, etc. Ding et al. (2017) leverage the domain knowledge to perform data selection to improve the machine translation models. Ding and Tao (2019) incorporate the structure patterns of sentences, i.e., syntax, into the Transformer network to enhance seq2seq modeling performance. Raganato et al. (2020) utilize the pre-defined fixed patterns to replace the attention weights and show promising results. Inspired by above works, we propose to augment zero-shot QE model with cross-lingual patterns.

5 Conclusion and Future Work

In this work, we revealed a mismatching issue in zero-shot QE modeling. To alleviate it, we introduced two explicit cross-lingual patterns based on BERTScore backbone. Extensive experiments indicated that our proposed patterns, without fine-tuning, the QE model can be improved marginally. Notably, our zero-shot QE method outperforms supervised QE model on 2 out of 6 directions, shedding light on zero-shot QE researches.

In the future, we plan to explore more strategies for incorporating various auxiliary information and better in-domain fine-tuning (Gururangan et al., 2020) or introduce an non-autoregressive refiner (Wu et al., 2020) to address our revealed *mismatching issue*. Also, it will be interesting to apply QE metrics on document-level machine translations with considering the dropped pronoun (Wang et al., 2016a, 2018).

Acknowledgments

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *COLING*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *NIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Liang Ding, Yanqing He, Lei Zhou, and Qingmin Liu. 2017. Combining domain knowledge and deep learning makes nmt more adaptive. In *CWMT*.

Liang Ding and Dacheng Tao. 2019. Recurrent graph syntax encoder for neural machine translation. *arXiv*.

Liang Ding, Longyue Wang, and Dacheng Tao. 2020. Self-attention with cross-lingual position representation. In *ACL*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *HLT*.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *ACL*.

Kai Fan, Jiayi Wang, Bo Li, Fengming Zhou, Boxing Chen, and Luo Si. 2019. "Bilingual Expert" Can Find Translation Errors. *AAAI*.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *arXiv*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *ACL*.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *ACL*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Chi-kiu Lo. 2019. Yisi-a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In *WMT*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020. Fixed encoder self-attention patterns in transformer-based machine translation. In *arXiv*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *ACL*.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *WMT*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *WMT*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *AMTA*.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André FT Martins. 2020. Findings of the wmt 2020 shared task on quality estimation. In *WMT*.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André FT Martins. 2018. Findings of the wmt 2018 shared task on quality estimation. In *WMT*.

Miloš Stanojević and Khalil Sima'an. 2014. Beer: Better evaluation as ranking. In *WMT*.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *ACL*.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018. Translating pro-drop languages with reconstruction models. In *AAAI*.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016a. A novel approach to dropped pronoun translation. In *NAACL*.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016b. Character: Translation edit rate on character level. In *WMT*.

Di Wu, Liang Ding, Fan Lu, and J. Xie. 2020. Slotrefine: A fast non-autoregressive model for joint intent detection and slot filling. In *EMNLP*.

E Yankovskaya, A Tättar, M Fishel Volume 3 Shared Task Papers, Day, and 2019. 2019. Quality estimation and translation metrics via pre-trained word and sentence embeddings. In *WMT*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.