Towards Informed Pre-Training for Critical Error Detection in English-German

Lisa Pucknat[1,2], Maren Pielka[1] and Rafet Sifa[1]

[1]Fraunhofer IAIS
[2]Universität Bonn

Abstract

This paper presents two data augmentation methods for pre-training, to find critical errors in machine translations. This includes an alignment approach used in traditional machine translation and an imitation method, mimicking the structure of the data. Both methods are adapted to a binary classification. Our approach achieves competitive results on the WMT'21 critical error detection (CED) dataset while only using 0.06% of datapoints in comparison to the first placement.

Keywords

Machine Translation, Quality Estimation, Critical Error Detection, Informed Machine Learning

1. Introduction

Finding critical errors for machine translations in the scope of quality estimation (QE) is a new research field, introduced during the WMT'21 shared task: Quality Estimation[1]. It aims to create a supervisory system for critical translation errors independent of the translation model. The organizers of the shared task define a critical error to fall into five categories, which are deviation in toxicity, health- or safety risks, named entities, sentiment polarity or negation and deviation in units/time/date/numbers. The necessity of the task is motivated by health, safety, legal, reputation, religious or financial concerns. Generally, the task is a binary classification on whether a sentence and its machine translation contain at least one critical error, without a respective gold translation. In contrast to other QE tasks, translation errors are tolerated if they are not critical. Due to the novelty of the task and the newly introduced dataset for benchmarking, there is only limited research available [1]. Therefore, we present two new approaches with data augmented pre-training. We utilize pre-training methods which were successfully applied to machine translation and natural language inference and adapt them to the CED task [2][3]. This includes aligning the languages and mimicking the structure of the CED dataset. For this, we rely only on a parallel corpus and lists of synonyms and antonyms, eliminating the need for human annotators, which can be costly.

LWDA'22: Lernen, Wissen, Daten, Analysen. October 05–07, 2022, Hildesheim, Germany
lisa.pucknat@iais.fraunhofer.de (L. Pucknat); maren.pielka@iais.fraunhofer.de (M. Pielka); rafet.sifa@iais.fraunhofer.de (R. Sifa)

[1]https://www.statmt.org/wmt21/quality-estimation-task.html

2. Related Work

Pre-training large language models, including multilingual settings, has led to many state-of-the-art results in natural language processing tasks [4, 5, 6, 7]. Different QE tasks also benefited from further pre-training with artificial data that approximated the structure of the given training dataset, starting with the common component of an initial parallel corpus [8, 9, 10, 11]. Explicitly for the CED task, the best performing approach on English-German sentences [1], is a multimetric-multilingual pre-training proposed by [12]. Here, large amounts of parallel sentences in different languages were gathered, totaling to about 72.3 million examples. By generating new machine translations for each sentence pair and automatically calculating common QE metrics, a large dataset for pre-training was created. Other approaches included feature extraction [13], incorporation of high quality machine translations of source sentences [14] and incorporation of uncertainty features into the fine-tuning process [15]. Regarding informed machine learning, prior information assists in the learning process. So-called prototypes, which are representative for the dataset, yield a beneficial effect [16, 17]. Recreating a dataset with simpler structure, could be seen as a prototype approach.

3. Data

For the downstream task, we use the English-German split of the CED WMT'21 competition dataset[2], which is composed out of sentences from the Jigsaw Toxic Comment Classification Challenge[3] and Wikipedia comments. Translations are created by the ML50 multilingual translation model by FAIR [18]. The dataset consists of a total of 12,000 sentences, splitting into 10 000 from the train dataset and respectively 1000 for each valid and test set. Exemplary sentences and translations of the dataset are displayed in table 1.

For data-augmentation and pre-training, we decided to use a parallel corpus composed out of subtitles from TED Talks, which are short presentations with innovative, semi-scientific topics. They resemble the conversational, less complex nature of the sentences in the CED dataset. We extract them from OPUS[4], which is an open collection of parallel corpora. In its raw state, it totals to approximately 189 000 sentence pairs. We want to incorporate as much information as possible from the CED dataset into the creation of the new dataset, for which reason we only select sentences with a similar sentence length. This corresponds to sentences with a maximum length of 20 and 24 for English and German, respectively, determined by the 75% percentile of sentence lengths from the CED dataset. We obtain structurally similar sentences by tolerating a maximum deviation of 10% between the source and the translated sentence lengths, since English and German sentence lengths are highly correlated [19].

[2] https://github.com/sheffieldnlp/mlqe-pe/tree/master/data/catastrophic_errors/
[3] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
[4] https://opus.nlpl.eu/

Source	Translation	Label
So what you're saying is the internet police AREN'T coming to my door???	Also was du sagst ist , dass die Internet-polizei nicht zu meiner Tür kommen ? ? ?	NOT
What the fuck? Ur a dick, urv never even been 2 helensville ya mommas boi!!!	Was ist das ? Ur ein dick , urv noch nie 2 helensville ya mommas boi ! ! !	ERR
This is a fact for which reliable, published sources do exist!	Das ist eine Tatsache , für die zuverlässige , veröffentlichte Quellen existieren !	NOT

Table 1
Exemplary sentence pairs from the WMT21's Critical Error Detection dataset.

4. Methodology

For the **alignment** approach, an adaption of the pre-training method for machine translation is utilized [2]. Aligning languages in the feature space might be beneficial for finding semantic differences [14]. Words of the same type within the parallel corpus (TED dataset) are exchanged across languages, creating new sentence pairs at the same time. Only words with a 1-1 correspondence are exchanged, as we have no explicit alignment information. Pairs with a critical error (ERR) are created by exchanging a randomly sampled translated word that does not match the actual translation. Non-error pairs (NOT) are generated by simply exchanging the relevant word for its translation. Augmented data pairs would be as follows:

Take the autonomous vehicle. – Nehmen Sie das autonome Fahrzeug.
NOT: Take the autonomous Fahrzeug. – Nehmen Sie das autonome vehicle
ERR: Take the autonomous Fahrzeug. – Nehmen Sie das autonome voice

We also experimented with not simply choosing a random word as the critical error, but exchanging for a word with opposite meaning.

For the **imitation** approach, we again exchange words to augment the data but deviate from the multilingual pre-training approach and stick only to the approach of exchanging synonyms and antonyms in the same language. This is due to the definition of the task, which specifies the following as the reason for a critical error.: "Mistranslation: critical content is translated incorrectly into a different meaning, **or not translated**". An exemplary pair with exchanged synonyms and antonyms could look like this:

The teacher was fascinated. – Die Lehrerin war fasziniert.
NOT: The lecturer was fascinated. – Die Lehrerin war fasziniert.
ERR: The pupil was fascinated. – Die Lehrerin war fasziniert.

For both approaches, we substitute only selected word types (e.g. nouns, verbs, adjectives), as there are no reasonable antonyms especially for fill and stop words. Further, we reason that critical errors appear in connection with major word types in a sentence. As there are sentences with and without errors in the dataset, we hope that the model can determine correct corresponding words and shift the focus away from unimportant words.

5. Experiments and Results

Experimental Setup First, the CED dataset was cleaned by deleting special characters and lowercasing all caps words. For training, we use the `xlm-roberta-large` checkpoint[5] as a starting point, which is an XLM-RoBERTa model pre-trained on 2.5 TB filtered CommonCrawl[6] data in 100 languages. The classification layer is composed out of two feed-forward layers

$$y_{pred} = softmax(tanh(h_0 W_{c1} + b_{c1}) W_{c2} + b_{c2}) \tag{1}$$

on top, where $y_{pred} \in \mathbb{R}^2$ is the model's binary prediction, h_0 is the last hidden state of the special token of XLM-RoBERTa and $W_{c1} \in \mathbb{R}^{d \times d}, b_{c1} \in \mathbb{R}^d, W_{c1} \in \mathbb{R}^{d \times 2}$ and $b_{c1} \in \mathbb{R}^2$ are parameters of the feed-forward network with $d = 1024$. For both pre-training and finetuning, we make use of the AdamW optimizer [20] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-6}$. Also, a linear warm-up was used for both experiments for 10% of the total training steps, reaching a maximum value of 5×10^{-6}. Batch sizes of 68 and 16 are respectively set for pre-training and finetuning. A weighted sampling was used during finetuning due to class imbalance. Lastly, we performed pre-training with merely 50,000 examples for one epoch, which took around 15 minutes on an NVIDIA V100.

In order to exchange antonyms and synonyms, word types noun, adjective and verb were extracted from the sentences and counted according to their occurrence. For each category, we took the 500 most frequently occurring words and collected automatically up to five synonyms and antonyms[7] and randomly substituted them accordingly.

Results and Evaluation In the following, we evaluate our approaches against the WMT'21 baseline and a XLM-R model without further pre-training. Special attention is kept on the Matthews correlation coefficient (MCC), as it is the main metric used in the competition. Table 2 shows our results for the CED test set. Both approaches almost always boost the performance of the model. We found that the best performing model incorporated both the word type adjective and the imitation approach. In this combination, we would rank 2nd in the competition[8] with a correlation of 0.5117.[9] The alignment approach is inferior to the imitation approach, which makes it clear that pre-training on an identical task is noticeably more effective. We also did not find performance differences between the approaches of aligning with random words and aligning with antonyms, and therefore decided to omit the results.

6. Conclusion

In this paper, two informed approaches for pre-training with syntactic data for finding critical errors in machine translations were proposed. We rank second in the English-German task 3, with only 0.06% of data points used in comparison to the first placement. Therefore, we assume

[5] https://huggingface.co/xlm-roberta-large
[6] https://commoncrawl.org/
[7] From https://www.thesaurus.com/
[8] https://www.statmt.org/wmt21/quality-estimation-task_results.html#task3_results
[9] Because of time constraints, we did not take part in the actual challenge.

		Acc	MCC	F1-ERR	F1-NOT	F1-Multi
WMT'21	**baseline**	-	0.3974	0.5317	0.8484	0.4511
XLM-R	**baseline**	0.786	0.4504	0.5810	0.8581	0.4986
Alignment Simple	**verb**	0.795	0.4748	0.6065	0.8614	0.5224
	adj.	0.799	0.4845	0.6127	0.8643	0.5296
	noun	*0.804*	0.4945	0.6157	*0.8685*	0.5347
	mix	0.793	0.4755	0.6131	0.8587	0.5265
Imitation	**verb**	0.798	*0.5137*	*0.6565*	0.8569	*0.5626*
	adj.	0.800	**0.5196**	**0.6610**	0.8582	**0.5673**
	noun	*0.804*	0.4980	0.6231	0.8676	0.5406
	mix	0.793	0.4979	0.6437	0.8541	0.5498

Table 2
Performance comparison for the pre-training tasks evaluated on the CED test set. *F1-ERR* and *F1-NOT* denote the F1-scores obtained on the error class and on the non-error class and *F1-Multi* is the multiplication of *F1-ERR* and *F1-NOT*. A MCC result of -1 means inverse, 0 not, and 1 highly correlated. The leftmost column specifies the data-augmentation method and the next column the word type exchanged. *mix* is created by concatenation of the different datasets.

that an informed approach using augmented data following the structure of the downstream task i.e., training on prototypical examples as proposed by [17], can be more efficient than training with a lot of data. Future work could include maximizing the prior information available to the model to further reduce the amount of data needed. In addition, alignment in general seems to be a good starting point for follow-up research in this domain. Other options for future work include additional linguistically informed pre-training tasks, as described by [21] and [22].

The insights gathered in this paper can likely be applied to other areas of text mining research as well, e.g. natural language inference [23, 24, 25]. In the context of financial document analysis [26, 27], our methods can be applied e.g. when comparing different versions of a document to find critical errors or contradictions. It can be worthwhile to design similar data augmentation strategies that are tailored to the domain, e.g. replacing specific financial terms with their opposite meaning.

Acknowledgement

In parts, the authors of this work were supported by the Competence Center for Machine Learning Rhine Ruhr (ML2R) which is funded by the Federal Ministry of Education and Research of Germany (grant no. 01|*S*18038*B*).

References

[1] L. Specia, F. Blain, M. Fomicheva, C. Zerva, Z. Li, V. Chaudhary, A. F. Martins, Findings of the wmt 2021 shared task on quality estimation, in: Proceedings of the Sixth Conference on Machine Translation, 2021, pp. 684–725.

[2] Z. Lin, X. Pan, M. Wang, X. Qiu, J. Feng, H. Zhou, L. Li, Pre-training multilingual neural machine translation by leveraging alignment information, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2649–2663.

[3] S. Wang, H. Fang, M. Khabsa, H. Mao, H. Ma, Entailment as few-shot learner, 2021. arXiv:2104.14690.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[6] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[7] N. Goyal, J. Du, M. Ott, G. Anantharaman, A. Conneau, Larger-scale transformers for multilingual masked language modeling, arXiv preprint arXiv:2105.00572 (2021).

[8] M. Negri, M. Turchi, R. Chatterjee, N. Bertoldi, Escape: a large-scale synthetic corpus for automatic post-editing, arXiv preprint arXiv:1803.07274 (2018).

[9] D. Lee, Two-phase cross-lingual language model fine-tuning for machine translation quality estimation, in: Proceedings of the Fifth Conference on Machine Translation, 2020, pp. 1024–1028.

[10] S. Eo, C. Park, H. Moon, J. Seo, H. Lim, Word-level quality estimation for korean-english neural machine translation, IEEE Access 10 (2022) 44964–44973.

[11] R. Rubino, E. Sumita, Intermediate self-supervised learning for machine translation quality estimation, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4355–4360.

[12] R. Rubino, A. Fujita, B. Marie, Nict kyoto submission for the wmt'21 quality estimation task: Multimetric multilingual pretraining for critical error detection, in: Proceedings of the Sixth Conference on Machine Translation, 2021, pp. 941–947.

[13] G. Jiang, Z. Li, L. Specia, Icl's submission to the wmt21 critical error detection shared task, in: Proceedings of the Sixth Conference on Machine Translation, 2021, pp. 928–934.

[14] Y. Chen, C. Su, Y. Zhang, Y. Wang, X. Geng, H. Yang, S. Tao, G. Jiaxin, W. Minghan, M. Zhang, et al., Hw-tsc's participation at wmt 2021 quality estimation shared task, in: Proceedings of the Sixth Conference on Machine Translation, 2021, pp. 890–896.

[15] J. Wang, K. Wang, B. Chen, Y. Zhao, W. Luo, Y. Zhang, Qemind: Alibaba's submission to the wmt21 quality estimation shared task, arXiv preprint arXiv:2112.14890 (2021).

[16] L. Von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy, et al., Informed machine learning–a taxonomy and survey of integrating knowledge into learning systems, arXiv preprint arXiv:1903.12394

(2019).
[17] L. von Rueden, S. Houben, K. Cvejoski, C. Bauckhage, N. Piatkowski, Informed pre-training on prior knowledge, arXiv preprint arXiv:2205.11433 (2022).
[18] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, A. Fan, Multilingual translation with extensible multilingual pretraining and finetuning, arXiv preprint arXiv:2008.00401 (2020).
[19] W. A. Gale, K. W. Church, et al., A program for aligning sentences in bilingual corpora, Computational linguistics 19 (1994) 75–102.
[20] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).
[21] J. Zhou, Z. Zhang, H. Zhao, LIMIT-BERT : Linguistic informed multi-task BERT, CoRR abs/1910.14296 (2019).
[22] A. Wahab, R. Sifa, Dibert: Dependency injected bidirectional encoder representations from transformers, in: Proc. of IEEE SSCI 2021, 2021.
[23] R. Sifa, M. Pielka, R. Ramamurthy, A. Ladi, L. Hillebrand, C. Bauckhage, Towards contradiction detection in german: A translation-driven approach, in: Proc. of IEEE SSCI 2019, 2019.
[24] M. Pielka, R. Sifa, L. P. Hillebrand, D. Biesner, R. Ramamurthy, A. Ladi, C. Bauckhage, Tackling contradiction detection in german using machine translation and end-to-end recurrent neural networks, in: Proc. of ICPR 2020, 2021.
[25] L. Pucknat, M. Pielka, R. Sifa, Detecting contradictions in german text: A comparative study, in: 2021 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2021, pp. 01–07.
[26] R. Sifa, A. Ladi, M. Pielka, R. Ramamurthy, L. Hillebrand, B. Kirsch, D. Biesner, R. Stenzel, T. Bell, M. Lübbering, U. Nütten, C. Bauckhage, U. Warning, B. Fürst, T. Dilmaghani Khameneh, D. Thom, I. Huseynov, J. Kahlert, R. amd Schlums, H. Ismail, B. Kliem, R. Loitz, Towards automated auditing with machine learning, in: Proceedings of the ACM Symposium on Document Engineering 2019, 2019, pp. 1–4.
[27] L. Hillebrand, T. Deußer, C. Bauckhage, T. Dilmaghani, B. Kliem, R. Loitz, R. Sifa, Kpi-bert: A joint named entity recognition and relation extraction model for financial reports, in: Proc. ICPR (to be published), 2022.