Multi-Source Text Classification for Multilingual Sentence Encoder with Machine Translation

Reon Kajikawa[†] **Keiichiro Yamada**[‡] **Tomoyuki Kajiwara**[†] **Takashi Ninomiya**[†]
[†] Graduate School of Science and Engineering, Ehime University, Japan
[‡] Undergraduate Program of Design and Architecture, Kyoto Institute of Technology, Japan
{reon@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp b2161513@edu.kit.ac.jp

Abstract

To reduce the cost of training models for each language for developers of natural language processing applications, pre-trained multilingual sentence encoders are promising. However, since training corpora for such multilingual sentence encoders contain only a small amount of text in languages other than English, they suffer from performance degradation for non-English languages. To improve the performance of pre-trained multilingual sentence encoders for non-English languages, we propose a method of automatic translating a source sentence into English and then inputting it together with the source sentence in a multi-source manner. Experimental results on sentiment analysis and topic classification tasks in Japanese revealed the effectiveness of the proposed method.

1 Introduction

Fine-tuning of pre-trained sentence encoders (Devlin et al., 2019; Liu et al., 2019) has been remarkably successful in a variety of NLP (natural language processing) application tasks (Wang et al., 2018). Pre-trained sentence encoders have developed in the direction not only of higher accuracy (Clark et al., 2020; He et al., 2021) and efficiency (Sanh et al., 2019; Zafrir et al., 2019), but also of multilingualization, with pre-trained multilingual sentence encoders such as mBERT and XLM-R (Lample and Conneau, 2019; Conneau et al., 2020), which are capable of handling 100 languages, being widely used (Liang et al., 2020). Since it is costly for developers to train models for each language, these multilingual sentence encoders are promising for the efficient multilingual deployment of NLP applications.

However, the training data for existing multilingual sentence encoders is dominated by English texts, with only a few percent in other languages. Table 1 shows a breakdown of the languages in the Common Crawl corpus[1] used to train XLM-R, one of the popular multilingual sentence encoders. This table shows that about half of the training data for the multilingual sentence encoder is English text, even Russian and German, which are the next largest languages, account for only about 6%, and other languages, such as Japanese, account for very little, less than 5%. Therefore, in languages such as Japanese, where training data is scarce and the grammatical structure differs significantly from that of English, the performance of the multilingual sentence encoder is degraded (Pires et al., 2019; Ahuja et al., 2023).

Language	Number of web pages	%
English	1,440 M	46.2
Russian	182 M	5.8
German	180 M	5.8
French	146 M	4.7
Chinese	144 M	4.6
Spanish	142 M	4.5
Japanese	138 M	4.4

Table 1: Most 7 languages in Common Crawl corpus.

To address this issue, we propose a method to improve the performance of multilingual sentence encoders in non-English languages by exploiting English, which is rich in the training data of multilingual sentence encoders. Our proposed method combines a non-English source sentence and its machine translation into English in a multi-source manner for input to a multilingual sentence encoder. We expect that utilizing English translations gives multilingual sentence encoders the benefit of large-scale pre-training. Although machine translation may contain translation errors, multi-source modeling with the source sentence can mitigate the negative effects of semantic changes.

[1] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 226–232

Experimental results on sentiment polarity classification of Japanese SNS (social networking service) posts[2] (Kajiwara et al., 2021; Suzuki et al., 2022) and topic classification of Japanese news titles[3] reveal the effectiveness of our multi-source modeling. In addition, our detailed analysis revealed that the performance of the proposed method is insensitive to differences in machine translation quality, that the proposed method is also effective for non-English languages other than Japanese, and that the proposed method is effective independent of the training corpus size.

2 Related Work

2.1 Multilingual Sentence Encoders

Following the success of masked language modeling (Devlin et al., 2019; Liu et al., 2019) in monolingual pre-training, its multilingual versions are being developed. Widely used multilingual sentence encoders include mBERT[4] (Devlin et al., 2019), pre-trained on Wikipedia in 104 languages, DistilmBERT (Sanh et al., 2019), its knowledge-distilled version, and XLM-R[5] (Lample and Conneau, 2019; Conneau et al., 2020), pre-trained on Common Crawl in 100 languages.

These multilingual sentence encoders are Transformer encoders (Vaswani et al., 2017) pre-trained with masked language modeling on corpora in multiple languages. Training corpus sizes for these models significantly vary between languages. As shown in Table 1, each of the non-English languages contains less than a few percent of the entire training corpus. This leads to degraded performance of multilingual sentence encoders in non-English languages (Pires et al., 2019; Ahuja et al., 2023).

2.2 Multi-Source Modeling

Previous research has improved the performance of NLP models by combining multiple input sentences. Zoph and Knight (2016) proposed a method of multi-source machine translation that utilizes a multilingual parallel corpus consisting of sentences in three or more languages that express the same meaning. For example, machine translation into English by inputting two sentences, one in French and the other in German, can improve translation quality compared to a single-sentence input. Instead of an encoder for each language, a simplified version of multi-source machine translation based on a single multilingual encoder (Dabre et al., 2017) is also being studied. While these previous studies on multi-source machine translation require the special resource of a multilingual parallel corpus consisting of three or more languages, this study, by contrast, addresses multi-source modeling in a generic way that can be applied in any NLP task.

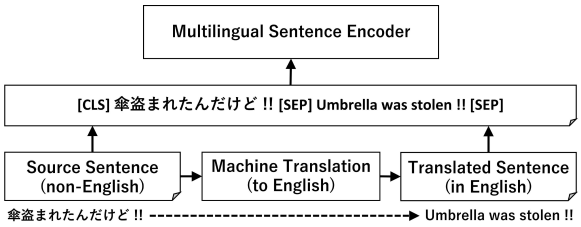

Figure 1: Overview of the proposed method

3 Proposed Method

To improve the performance of pre-trained multilingual sentence encoders, we propose a method to transform source texts into synonymous expressions that perform well for the model. In this study, we assume that expressions that occur frequently in the pre-trained corpus are easy for the model to process, and machine translate lower-frequency non-English sentences into higher-frequency English sentences, which are then input to a multilingual sentence encoder. To reduce the effect of noise during machine translation, we employ multi-source modeling (Dabre et al., 2017), in which sentences before and after the machine translation are concatenated and input.

An overview of the proposed method is shown in Figure1. We target non-English languages and assume situations where machine translation models from the target language to English are available.

First, the given source sentence is machine-translated, and the sequence "[CLS] source sentence [SEP] English translation [SEP]" is input to the multilingual sentence encoder. As shown in Table 1, since many English expressions are included in the training corpus of the multilingual sentence encoder, multi-source modeling via machine translation can be expected to improve performance by adding high-frequency expressions that are easier for the multilingual sentence encoder.

[2] https://github.com/ids-cv/wrime

[3] https://www.rondhuit.com/download.html#ldcc

[4] https://github.com/google-research/bert/blob/master/multilingual.md

[5] https://github.com/facebookresearch/XLM

	WRIME (QWK)			Livedoor (Accuracy)		
	A. Source	B. English	A+B (Proposed)	A. Source	B. English	A+B (Proposed)
DistillmBERT	0.446	**0.453**	**0.465**	0.851	0.778	**0.855**
mBERT	0.473	0.449	**0.476**	0.862	0.790	**0.864**
XLM-R (base)	0.555	0.503	**0.561**	0.859	0.791	**0.867**
XLM-R (large)	0.587	0.495	**0.598**	0.870	0.796	**0.873**

Table 2: Experimental results on the Japanese text classification tasks. Scores that improve over the baseline (A), which uses only the source sentence, are highlighted in **bold**.

4 Experiments

We evaluate the effectiveness of the proposed method on Japanese text classification tasks for four pre-trained multilingual sentence encoders.

4.1 Experimental Setup

4.1.1 Task

We experimented with two Japanese text classification tasks: sentiment polarity classification of SNS posts and topic classification of news titles.

For Japanese sentiment polarity classification, we used the WRIME corpus[2] (Kajiwara et al., 2021; Suzuki et al., 2022). This is a corpus of Japanese SNS posts annotated by the writers with their own sentiment polarity on a 5-point scale of [-2, -1, 0, +1, +2]. As shown in Table 3, we used the corpus split into training and evaluation corpora according to the official settings. For evaluation, we used Quadratic Weighted Kappa (QWK) (Cohen, 1968).

For Japanese topic classification, we used the Livedoor news corpus.[3] This is a corpus of Japanese news articles annotated with nine topics. Although this corpus contains both the main text and headlines of articles, only the headlines were used in this experiment. As shown in Table 3, we used the corpus split into training and evaluation corpora according to the official settings. For evaluation, we used Accuracy.

4.1.2 Model

For Japanese to English machine translation model, we used a 6-layer, 512-dimensional 8-attention-head Transformer (Vaswani et al., 2017) implemented using the fairseq toolkit[6] (Ott et al., 2019). This machine translation model was trained on JParaCrawl[7] (Morishita et al., 2020, 2022), an English-Japanese parallel corpus. A beam search with a beam width of 5 was applied for decoding.

[6] https://github.com/facebookresearch/fairseq

[7] https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

	Train	Valid	Test
WRIME	30,000	2,500	2,500
Livedoor	5,894	737	736

Table 3: Number of sentences for each corpus.

Four pre-trained multilingual sentence encoders were evaluated using HuggingFace Transformers (Wolf et al., 2020): DistilmBERT[8] (Sanh et al., 2019), mBERT[9] (Devlin et al., 2019), XLM-R (base[10] and large[11]) (Lample and Conneau, 2019; Conneau et al., 2020). They are multilingual sentence encoders that have been pre-trained with masked language modeling and are capable of handling approximately 100 languages, including English and Japanese.

For fine-tuning pre-trained multilingual sentence encoders, we used AdamW (Loshchilov and Hutter, 2019) for optimization, with a maximum learning rate of 2×10^{-5} and a batch size of 64 tokens, and training was stopped by early stopping with 3 epochs of patience on evaluation metric in the validation dataset. We report the average of three evaluations, except for the maximum and minimum values, of five evaluations with different random seed values.

4.2 Results

Table 2 shows the experimental results. All multilingual sentence encoders consistently achieve higher performance with the multi-source input (A+B) compared to the baseline with only the source sentence, revealing the effectiveness of the proposed method.

[8] https://huggingface.co/distilbert-base-multilingual-cased

[9] https://huggingface.co/bert-base-multilingual-cased

[10] https://huggingface.co/xlm-roberta-base

[11] https://huggingface.co/xlm-roberta-large

	WRIME (QWK)					Livedoor (Accuracy)				
		Multi-Source					Multi-Source			
	Baseline	Big	Base	Small	M2M100	Baseline	Big	Base	Small	M2M100
DistillmBERT	0.446	**0.464**	**0.465**	**0.458**	**0.467**	0.851	**0.857**	**0.855**	**0.858**	0.846
mBERT	0.473	**0.481**	**0.476**	**0.483**	**0.479**	0.862	0.861	**0.864**	**0.863**	**0.864**
XLM-R (base)	0.555	0.543	**0.561**	**0.562**	**0.562**	0.859	**0.872**	**0.867**	**0.860**	**0.860**
XLM-R (large)	0.587	0.580	**0.598**	**0.589**	0.584	0.870	**0.880**	**0.873**	**0.876**	**0.879**

Table 4: Experimental results of multi-source text classification based on machine translation with different translation quality. As shown in Table 5, the more left model is a higher quality machine translation.

When the English translation of the source sentence (B) is used as a stand-alone, the performance is often worse than the baseline with only the source sentence. This is assumed to be due to the effect of translation errors caused by machine translation. However, since the English translation contains translation errors but also includes useful expressions that are easy for multilingual sentence encoders, the multi-source input in combination with the source sentence improves the text classification performance.

4.3 Impact of Translation Quality

To analyze the impact of machine translation performance on the multi-source input of the proposed method, we conducted the same experiments using four Japanese to English machine translation models with different translation quality. For Japanese to English machine translation models, we used pre-trained models (Big, Base, Small) on JParaCrawl (Morishita et al., 2020, 2022) and M2M100[12](Fan et al., 2021), which can translate between 100 languages. The model structure of each machine translation model and the translation quality BLEU (Papineni et al., 2002) on the WMT20 news translation task (Barrault et al., 2020) are shown in Table 5.

Table 4 shows the experimental results. In 27 of the 32 experimental settings, the proposed method improved the text classification performance. In particular, the multi-source input was always effective when using the medium-quality machine translation models of Base and Small. The text classification performance sometimes worsened when using the Big model with the highest translation quality and the M2M100 model with the lowest translation quality. Poor translation quality may cause translation errors to mislead text classification, but understanding the cause of the negative impact of the high-quality machine translation model remains our future work.

	BLEU	# Layers	# Heads	# Dimension
Big	24.0	6	16	1,024
Base	21.3	6	8	512
Small	20.0	6	4	512
M2M100	16.2	12	16	1,024

Table 5: Model structure and translation quality for each machine translation model. The translation quality here is BLEU score on the Ja → En news translation task.

	A. Source	B. English	A+B (Proposed)
French	0.941	0.894	**0.942**
Korean	0.842	0.766	**0.844**

Table 6: Experimental results in non-English languages other than Japanese. Accuracy of the two-class sentiment polarity classification task in French and Korean.

4.4 Experiments in Other Languages

To evaluate the effectiveness of the proposed method in languages other than Japanese, we experimented with sentiment polarity classification in French and Korean. Note that the statistics on the amount of training data (number of Web pages) for the multilingual sentence encoder shown in Table 1 show that Japanese accounts for 138M pages or 4.4% of the total, French for 146M pages or 4.7% of the total, and Korean for 21M pages or 0.7% of the total.

For sentiment polarity classification, we used Allociné[13] for French and NSMC[14] for Korean. Both are binary classification tasks that annotate movie review texts with positive or negative sentiment polarity. We selected 30,000 sentences for training and 2,500 sentences for each validation

[12] https://huggingface.co/facebook/m2m100_418M

[13] https://huggingface.co/datasets/allocine

[14] https://github.com/e9t/nsmc

Label	Gold: +2　Pred: -2（Only Source）　Pred: +1（Multi-Source）
Source	初めてのエストニアで、日本食に救われなう。美味しい…
English	It was my first time in Estonia, and I was saved by Japanese food.
Label	Gold: +1　Pred: +1（Only Source）　Pred: -1（Multi-Source）
Source	う～あかん。気持ちがどんより。珈琲淹れるだす。
English	Wow, I feel like I'm going to brew coffee.

Table 7: Examples of sentiment polarity classification in Japanese (Upper row: successful examples of the proposed method, lower row: unsuccessful examples of the proposed method)

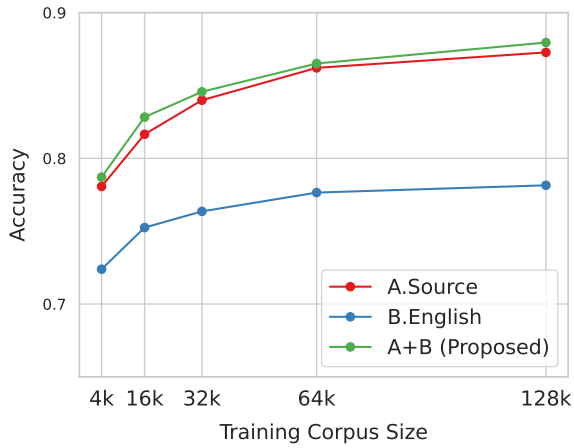

Figure 2: Performance by corpus size in Korean.

and evaluation, aligned to the Japanese WRIME, respectively. Each corpus was randomly selected to have equal proportions of positive and negative labels. M2M100[12](Fan et al., 2021) was used for the machine translation and mBERT[9] was used for the multilingual sentence encoder to evaluate the Accuracy. Other settings are the same as in Section 4.1.

Table 6 shows the experimental results. In French and Korean, the classification performance was slightly improved by the proposed method. However, because of the high baseline performance of the source text only, no significant changes were observed in either language compared to Japanese.

We analyzed the change in performance of the proposed method when the training corpus size was changed. Figure2 shows the results of the experiment in Korean. Consistently improved performance was confirmed, regardless of the size of the training corpus.

4.5 Qualitative Analysis

An example of sentiment analysis in Japanese is shown in Table 7. In the successful example in the upper row, the broken expression "救われなう", which is peculiar to SNS, may have affected the classification performance. The English translation does not include the broken expression, so it is thought that the writer's positive sentiment can be read from words such as "saved". In the bottom example, the English translation does not include negative expressions such as "あかん" and "どんより" caused by mistranslation of the machine translation. These expressions are considered to be difficult for a multilingual sentence encoder because they are dialects and low-frequency expressions in Japanese, therefore, the proposed method could not improve the results.

5 Conclusion

This study proposed a multi-source input method that uses machine translation of source texts and a combination of source and English translations to improve the performance of pre-trained multilingual sentence encoders in languages other than English, aiming at the efficient deployment of natural language processing services in multiple languages. The proposed method benefits from the large amount of pre-trained data in English and is expected to improve the performance of multilingual sentence encoders.

Evaluation experiments on sentiment polarity classification of SNS posts and topic classification tasks of news articles in Japanese showed that the proposed method with English translations can improve the classification performance compared to the baseline method with only the source sentences. The proposed method with both the source and target sentences consistently improves the performance of the multilingual sentence encoder, while the performance of the method with only the target sentences deteriorates because the machine-translated sentences can contain translation errors.

Acknowledgements

These research results were obtained from the commissioned research (No.22501) by National Institute of Information and Communications Technology (NICT), Japan.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the Eighth International Conference on Learning Representations*.

Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, 70(4):213–220.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Raj Dabre, Fabien Cromieres, and Sadao Kurohashi. 2017. Enabling Multi-Source Neural Machine Translation By Concatenating Source Sentences In Multiple Languages. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 96–107.

Jacob Devlin, Ming-Wei Chang, Kenon Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1–48.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *Proceedings of the Ninth International Conference on Learning Representations*.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. In *Proceedings of the Thirty-third Conference on Neural Information Processing Systems*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6008–6018.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Joshi. Mandar, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the Seventh International Conference on Learning Representations*.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3603–3609.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael

Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. In *Proceedings of the Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Haruya Suzuki, Yuto Miyauchi, Kazuki Akiyama, Tomoyuki Kajiwara, Takashi Ninomiya, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2022. A Japanese Dataset for Subjective and Objective Sentiment Polarity Classification in Micro Blog Domain. In *Proceedings of the 13th International Conference on Language Resources and Evaluation*, pages 7022–7028.

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 *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8BERT: Quantized 8Bit BERT. In *Proceedings of the Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Barret Zoph and Kevin Knight. 2016. Multi-Source Neural Translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34.