

TeacherSim: Cross-lingual Machine Translation Evaluation with Monolingual Embedding as Teacher

1st Hao Yang
2012 Labs

Huawei Technologies CO., LTD
yanghao30@huawei.com

2nd Min Zhang
2012 Labs

Huawei Technologies CO., LTD
zhangmin186@huawei.com

3rd Shimin Tao
2012 Labs

Huawei Technologies CO., LTD
taoshimin@huawei.com

4th Miaomiao Ma
2012 Labs

Huawei Technologies CO., LTD
mamiaomiao@huawei.com

5th Ying Qin
2012 Labs

Huawei Technologies CO., LTD
qinying@huawei.com

6th Daimeng Wei
2012 Labs

Huawei Technologies CO., LTD
weidaimeng@huawei.com

Index Terms—Machine Translation Evaluation, Knowledge Distillation, Monolingual Embedding

Abstract—Cross-lingual machine translation (MT) evaluation or quality estimation is usually performed either from cross-lingual transfer in supervised tasks or via unsupervised multilingual textual similarity. XMoverScore [1] investigates a range of state-of-the-art cross-lingual evaluations by feature extraction on token-level similarity and multilingual pre-trained models mBERT [2] and LASER [3]. In this paper, monolingual and sentence-level embedding is used as a teacher encoder (TeacherSim) to align cross-lingual MT sentences. It is found that the monolingual teacher encoder can enhance both quality and usability in that: (a) A monolingual embedding teacher assures cross-lingual semantic alignment and enhances evaluation performance. (b) The last layer always delivers the best performance based on the Sentence Transformer architecture, while layer selection is sensitive for different models with BERTScore [4]. Experiments like segment-level MT evaluation show that TeacherSim surpasses most strong baselines for all language pairs.¹

I. INTRODUCTION

Automatic MT evaluation or quality estimation has been studied for years. It has progressed significantly in recent years as the pre-trained contextual word embeddings improve. Different evaluation metrics are designed to estimate the semantic equivalence between the input text (e.g. a source sentence) and output text that is produced or revised (e.g. a machine translation), and to assess the translation quality (e.g. fluency and accuracy). Therefore, it is clearly that evaluation metrics are supposed to make input-output comparisons.

There are generally two types of generation evaluation methods: token-level evaluation and semantic-level evaluation. Token-level evaluation methods include the famous BLEU [5], METEOR [6], and ROUGE [7]. Based on n-gram matching, these methods break reference and target sentences down to

¹Code will be released upon publication.

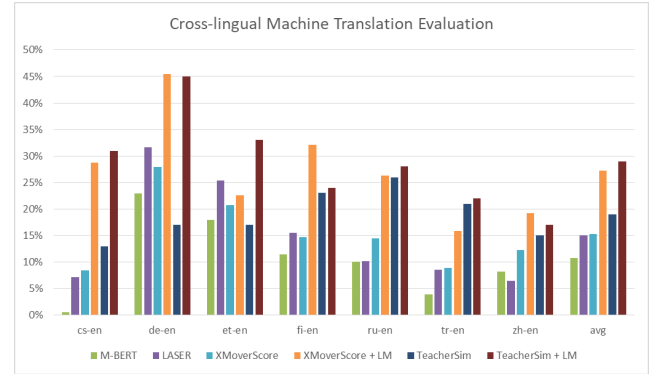


Fig. 1. Kendall correlation of cross-lingual MT evaluation of WMT18 Segment-Level Metrics

unigram, bigram, trigram, and four-gram words, and then calculate the proportion of repeated words to all words to get the similarity. The more similar words, the higher the similarity. Evaluation methods based on BLEU and others are widely used in various research domains, such as MT evaluation, dialog generation, and summarization. This evaluation has two typical problems: (1) It can only achieve token-level evaluation. The similarity of synonyms is not necessarily high and of antonyms is not necessarily low. (2) Cross-lingual or multilingual evaluation cannot be implemented because different languages have different vocabularies.

In order to solve the semantic similarity problem that token-level matching cannot solve, many studies tried to use semantic similarity for evaluation. [4] proposed the BERTScore method, which decomposes both the source and target sentences into words, obtains the semantic representation of each word, and then calculate the count of source-sentence words in the target sentence as well as the count of target-sentence words in the source sentence. The cross-lingual similarity is obtained by

calculating the corresponding F1 value. [8] used the Word Mover’s Distance (WMD), not cosine similarity, to obtain the semantic similarity of each word, thus getting the sentence-level semantic similarity. In addition, this method implements one-to-many matching. That is, one word in the source sentence may match multiple target words, and vice versa. In this case, a higher accuracy rate is obtained. [9] obtains the sentence-level similarity by adding an average pooling layer to the word-level representation, and then linearly combines the sentence-level similarity and word-level similarity for higher accuracy.

For multilingual sentence embedding, [3] used the LASER approach, which is based on zero-shot transfer (zero-shot learning is about leveraging deep learning networks already trained by supervised learning in other ways, without additional supervised learning) of NLP models from several languages, for example, English and Chinese, to scores of other languages, including those with limited training data. LASER uses one single model or embedding to handle all varieties of languages, including very low-resource languages, and it is not best embedding of language-pair sentences alignment and evaluation.

Based on knowledge distillation [10], this paper proposes a cross-lingual representation model and a multilingual semantic evaluation model. This approach is implemented by using a monolingual teacher model and a multilingual student model. The teacher model is designed as non-training, while the student model can be trained and its parameters can be changed. In the training process, the input is semantically aligned sentences. This process tries to make the outputs of student model semantically aligned with the output of the teacher model. It is ideal to make the representations of same-meaning sentences in different languages 100% similar. Through this knowledge distillation process, different sentences with the same semantic meaning can be effectively aligned, and sentences with similar semantic meaning can be brought closer while sentences with different semantic meanings can be pushed farther. In the inference process, the input is semantically different sentences, and different semantic representations are obtained through the student model. By calculating the cosine similarity, the cross-lingual similarity can be evaluated. This approach is called TeacherSim. Experiments show that TeacherSim outperforms the strongest baselines of most language pairs.

In summary, TeacherSim makes the following contributions:

- 1) A Knowledge-Distillation Cross-lingual Machine Translation method is provided, which is evaluated by multilingual embedding model with monolingual embedding model as Teacher.
- 2) Language models have been shown useful for evaluation [1], and also work with TeachSim to get SOTA performance.
- 3) Two visualization analyses are provided, showing that (1) TeacherSim is not sensitive to layer selection like SBERT [9], and its last layer is most accurate for semantic sentence presentation; while for BERTScore, layer selection is sensitive for different models. (2)

TeacherSim better distinguishes token alignment than other pre-trained models.

II. RELATED WORK

a) MT evaluation based on token-level matching: Token-level evaluation methods include the famous BLEU or Sent-BLEU [5], NIST [11], chrF++ [12], and METEOR++ [13]. Based on the n-gram matching mechanism, they decompose sentences into unigram, bigram, trigram, or four-gram words or phrases, and then calculates the proportion of the same phrases to obtain the translation similarity. As mentioned earlier, token-level matching can only achieve word-level evaluation, where synonyms may get a low similarity and antonyms may get a high similarity. Worse still, this evaluation cannot be used across languages [14] [15].

b) MT evaluation based on semantic-level matching: With the development of deep learning and pre-trained language models, it has become mainstream to evaluate MT quality by comparing semantic similarity at the sentence level. [16] and [17] convert the sentence representation into a vector embedding by using the pre-trained model, and use the vector similarity to obtain the sentence similarity without considering the position of the words. [8] uses an approach that is based on the word position and WMD, and realizes that the closer the words, the higher the similarity. Also, HW-TSC [18] provides a model based on Seq2Seq translation model and HW-TSC2 [19] provides only an encoder model that achieves best WMT 2020/2021 QE performance.

c) MT evaluation for Cross-lingual matching: With the development of multilingual pre-trained models, including mBERT [20], XLM-RoBERTa [21], and LASER [22], they are introduced to MT quality evaluation. These models use aligned multilingual corpora for semantic similarity alignment, and use the same model to output similarity between different languages. There are two ways to calculate the similarity. [10] adds an average pooling layer to the word representations, and calculates the cosine similarity of the pooling layer to obtain the sentence-level similarity. [2] and [21] use the word distance to optimize the similarity calculation. They calculate the shortest moving distance of similar words in different languages to obtain cross-lingual similarity.

III. PROPOSED APPROACH

A. TeacherSim Knowledge Distillation Framework

Like [10], the knowledge-distillation style framework for cross-lingual MT evaluation is based on a student model and a teacher model. For the training process, parallel data of language pairs are provided to the student model and a source sentence is provided to the teacher model. The teacher model is frozen and only the student model is optimized. Two student embeddings are aligned with the teacher embedding with two Mean squared error (MSE) loss. After training, Student can align cross-lingual sentences. When in evaluation, source and target sentences are fed to the student model, and the

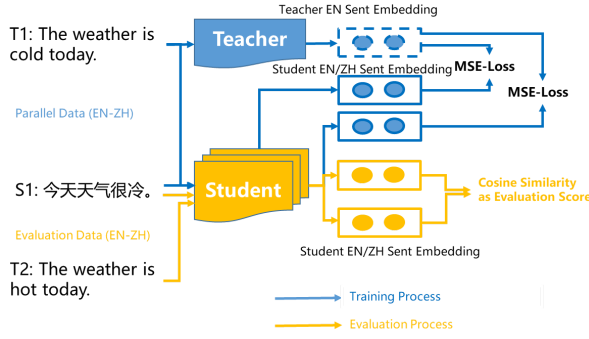


Fig. 2. TeacherSim framework: Cross-lingual evaluation based on multilingual sentence embedding model with monolingual embedding model as Teacher

evaluation score is calculated by cosine similarity between two student sentence embeddings.

$$Loss = MSEloss(Tea(t1), Stu(t1)) + MSEloss(Tea(t1), Stu(s1)) \quad (1)$$

$$Score = CosSim(Stu(s1), Stu(t2)), \quad (2)$$

where Tea is the teacher model, Stu is the student model, $(s1, t1)$ is a parallel sentence pair, $(s1, t2)$ is an evaluation sentence pair, training loss is two $MSELoss$ summarization, and the predicted score is $CosSim$ between two presentations.

B. Sentence-level Similarity

The idea is borrowed from Computer Vision research, which uses Siamese and Triplet network structures to derive a fixed-sized sentence embedding vector and then uses a similarity measure like cosine similarity or Manhattan/Euclidean distance to compute semantically similar sentences. Just like SBERT and other sentence embedding pre-trained models, some specific layer or just the last layer is chosen as the representation layer. Cross-entropy loss of sentence embeddings is chosen for Classification scenarios, MSE loss is chosen for Regression scenarios, and Cosine Similarity is chosen for Predicted Score.

$$CLS_Loss := Softmax(s, t, s - t) \quad (3)$$

$$REG_Loss := MSE(s, t, s - t) \quad (4)$$

$$TeaSim_Score := cos(s, t) \quad (5)$$

where s, t are source and target sentences embeddings, $|st|$ is embedding difference, $Softmax$, MSE are training losses, and $TeaSim_Score$ is TeacherSim predicted score.

C. Language Model

As language models (LMs) can effectively reflect the fluency of the language, LM of the target language can be used to evaluate the translation quality. LM scores can be used to improve the performance of our metrics based on fluency. In this paper, the authors use the logarithmic form of the

geometric mean perplexity (PPL) of all sentences calculated by using XLM-R as a metric of LM evaluation.

TeacherSim Score with LM is defined below.

$$TeaSim_Score_LM = Score(s, t) + \lambda * PPL_t \quad (6)$$

where λ is a hyper-parameter to combine two scores. $TeaSim_Score_LM$ is TeacherSim predicted score with language model enhancement. According to experiments, the best λ is near 0.15.

IV. EXPERIMENTS

In this work, the proposed method is tested with the WMT18 metrics segment-level task data. For multilingual reference-free evaluation, (source, candidate) is chosen like QE as a metric.

Experiments are contrasted with two types of systems, including:

- M-BERT and LASER as multilingual pretrained model
- XMoverScore and XMoverScore+LM as strong baselines

A. Datasets

Our experiments use the WMT18 news dataset as the evaluation dataset, which includes 3000 source sentences. The evaluation dataset also includes the translated sentences of all teams participating in WMT18. These target sentences are classified into 14 language pairs, such as DE-EN, FI-EN, and ZH-EN. The DE-EN sentences are the largest in number (177,811) while the CS-EN are the least in number (5,110). Pairwise comparison is conducted on the target sentences by professional multilingual experts before these sentences are used as the test data.

B. Models

From [9] and [23], we selected SBERT-nli-stsb, the model with the highest quality of aligned multilingual sentences, as the teacher model. It is an optimization of the SBERT model and outputs sentence-level representations through a pre-trained model and an average pooling layer. Based on the Siamese architecture, the model is fine-tuned based on multiple Natural Language Inference (NLI) datasets and semantic textual similarity (STS) datasets. In the language of English, this model enables the best sentence-level discrimination. The cosine similarity of similar sentences' representations is higher, of dissimilar sentences is lower, and of identical sentences is closer to 1. If the sentences are random, the cosine similarity is approaching 0.5 or even lower.

According to [24] and [25], our student model of knowledge distillation is fine-tuned based on XLM-R or mBERT. Many papers have shown that XLM-R offers better quality than mBERT because XLM-R does not need to take into account the context of the sentence.

The knowledge distillation dataset is a multilingual STS 2017 dataset that contains DE-EN, ZH-EN, EN-AR, EN-ES, EN-FI, and EN-TR aligned corpus. Some of the data is augmented by an MT engine. The data is cleaned and filtered by professional human translators before being used.

Language-Pair	de-en	cs-en	et-en	fi-en	ru-en	zh-en	tr-en
RR-Systems	16	5	14	9	8	5	14
Dev-Datasize	177811	5110	56712	15648	10404	33357	5525

TABLE I

WMT18 SEGMENT-LEVEL DATASETS THAT INCLUDES TO-EN LANGUAGE-PAIR MACHINE TRANSLATION TO ENGLISH. RR-SYSTEMS IS SYSTEM NUMBER OF HUMAN-LABELED TRANSLATED SENTENCES. DEV-DATASIZE IS THE NUMBER OF EVALUATION SENTENCE PAIRS OF EACH LANGUAGES.

C. Training Parameters and Hyperparameters

In the training, the warm-up steps are 10,000, epoches are 15, learning rate is $2e-5$, with the batch size being 64 for training and 32 for evaluation. MSE loss is measured and optimized on sentences pairs. For language model score parameter λ , grid search is experimented from 0.1 to 0.9 with step 0.05, and best results achieved with 0.15.

D. Main Results

Based on multiple datasets such as the WMT18 to-EN datasets, we compare the similarity of the source and translated sentences and make an estimation by using the Kendall rank correlation coefficient. The result reveals that TeacherSim (19%) outperforms mBERT (11%), LASER (15%), and XMoverScore (15%). And TeacherSim+LM (29%) surpasses TeacherSim (19%) and XMoverScore+LM (27%).

V. ANALYSIS

A. Visualization for Token-level Alignment

Like BERTScore [4], computing a similarity score for each token in the candidate sentence with each token in the reference sentence. This paper compares token-level alignment visualization with BERTScore both on multilingual pre-trained model XLM-R and TeacherSim. Typical case in Fig3 of ZH-EN shows that the similarity score of all words in the multilingual pre-trained model XLM-R exceeds 0.9, making it difficult to make a discrimination. However, the token-level similarity distribution based on TeacherSim is quite uneven, which indicates that semantic similarity of multiple language tokens is much more accurate than pre-trained models in Fig4.

VI. RESULTS

In this work, The TeacherSim evaluation method is proposed, with a monolingual teacher model and a multilingual student model. With the multilingual knowledge distillation strategy and LM enhancement, TeacherSim surpasses many other strong baselines both in sentence-level alignment and token-level alignment. In particular, layer selection is not sensitive for different models with TeacherSim, making TeacherSim much easier to use.

In future work, the authors will aim to improve fine-tuning performance of TeacherSim for a wider adoption in more languages and domains, like BLEURT.

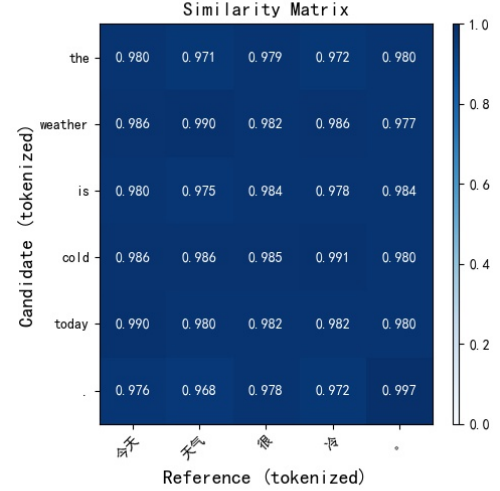


Fig. 3. Token level semantic similarity comparison for BertScore over XLM-R

REFERENCES

- [1] W. Zhao, G. Glavaš, M. Peyrard, Y. Gao, R. West, and S. Eger, "On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 1656–1671. [Online]. Available: <https://aclanthology.org/2020.acl-main.151>
- [2] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung *et al.*, "Multilingual universal sentence encoder for semantic retrieval," *arXiv preprint arXiv:1907.04307*, 2019.
- [3] M. Artetxe and H. Schwenk, "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 597–610, 2019.
- [4] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," *CoRR*, vol. abs/1904.09675, 2019. [Online]. Available: <http://arxiv.org/abs/1904.09675>
- [5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318. [Online]. Available: <https://doi.org/10.3115/1073083.1073135>
- [6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations*, 2015.
- [7] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003, pp. 150–157.
- [8] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, "Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance," in *Proceedings of the 2019 Conference*

model	de-en	cs-en	et-en	fi-en	ru-en	zh-en	tr-en	avg
M-BERT	23%	1%	18%	12%	10%	8%	4%	11%
LASER	32%	7%	25%	16%	10%	6%	9%	15%
XMoverScore	28%	8%	21%	15%	15%	12%	9%	15%
XMoverScore-LM	46%	29%	23%	32%	16%	19%	16%	27%
TeacherSim	17%	13%	17%	23%	26%	15%	21%	19%
TeacherSim-LM	45%	31%	33%	24%	28%	17%	22%	29%

TABLE II

WMT18 SEGMENT-LEVEL RESULTS, TO-ENGLISH. BOLD DENOTES BEST PERFORMANCE, AND HIGHER IS BETTER

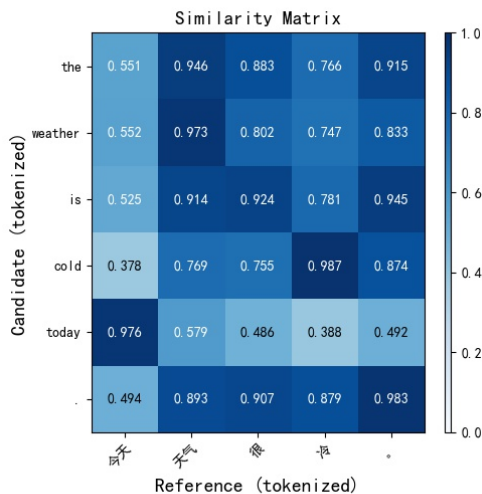


Fig. 4. Token level semantic similarity comparison for BertScore over TeacherSim

on Empirical Methods in Natural Language Processing. Hong Kong, China: Association for Computational Linguistics, August 2019.

- [9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. [Online]. Available: <https://aclanthology.org/D19-1410>
- [10] —, "Making monolingual sentence embeddings multilingual using knowledge distillation," *arXiv preprint arXiv:2004.09813*, 2020.
- [11] G. Doddington, "Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics," in *Proceedings of the Second International Conference on Human Language Technology Research*, ser. HLT '02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 138–145. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1289189.1289273>
- [12] M. Popović, "chrF++: Words Helping Character N-grams," in *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, 2017, pp. 612–618. [Online]. Available: <https://www.aclweb.org/anthology/W17-4770>
- [13] Y. Guo and J. Hu, "Meteor++ 2.0: Adopt Syntactic Level Paraphrase Knowledge into Machine Translation Evaluation," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 501–506. [Online]. Available: <https://www.aclweb.org/anthology/W19-5357>
- [14] O. Bojar, Y. Graham, and A. Kamran, "Results of the WMT17 metrics shared task," in *Proceedings of the Conference on Machine Translation (WMT)*, 2017. [Online]. Available: <https://www.aclweb.org/anthology/W17-4755>
- [15] —, "Results of the WMT17 metrics shared task," in *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 489–513. [Online]. Available: <https://www.aclweb.org/anthology/W17-4755>
- [16] L. Specia, K. Shah, J. G. de Souza, and T. Cohn, "QuEst - a translation quality estimation framework," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 79–84. [Online]. Available: <https://www.aclweb.org/anthology/P13-4014>
- [17] L. Specia, D. Raj, and M. Turchi, "Machine translation evaluation versus quality estimation," *Machine translation*, vol. 24, no. 1, pp. 39–50, 2010.
- [18] M. Wang, H. Yang, Y. Deng, Y. Qin, L. Lei, D. Wei, H. Shang, N. Xie, X. Li, and J. Guo, "The HW-TSC Video Speech Translation system at IWSLT 2020," in *Proceedings of the 17th International Conference on Spoken Language Translation*, 2020, pp. 187–190.
- [19] Y. Chen, C. Su, Y. Zhang, Y. Wang, X. Geng, H. Yang, S. Tao, G. Jiaxin, W. Minghan, M. Zhang, Y. Liu, and S. Huang, "HW-TSC's participation at WMT 2021 quality estimation shared task," in *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2021, pp. 890–896. [Online]. Available: <https://aclanthology.org/2021.wmt-1.92>
- [20] S. Dev and J. M. Phillips, "Attenuating bias in word vectors," in *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, 2019, pp. 879–887. [Online]. Available: <http://proceedings.mlr.press/v89/dev19a.html>
- [21] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of ACL*, 2020. [Online]. Available: <https://arxiv.org/pdf/1911.02116.pdf>
- [22] Y. Chen and T. Avgustinova, "Are language-agnostic sentence representations actually language-agnostic?" in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Held Online: INCOMA Ltd., Sep. 2021, pp. 274–280. [Online]. Available: <https://aclanthology.org/2021.ranlp-1.32>
- [23] Y. Song, J. Zhao, and L. Specia, "SentSim: Crosslingual semantic evaluation of machine translation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 3143–3156. [Online]. Available: <https://aclanthology.org/2021.naacl-main.252>
- [24] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, aug 2017, pp. 1–14. [Online]. Available: <https://www.aclweb.org/anthology/S17-2001>
- [25] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4512–4525. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.365>