CS224N - Natural Language Processing with Deep Learning - Stanford - Winter 2021

Sadegh Aryanpanah

Assessment of Neural Machine Translation Performance Based on a new Sentence Embedding Cosine Similarity Metric.

1. Key Project Information

- Project Type: Custom
- Mentor: John Hewitt

2. Abstract

The deficiencies in the assessment of machine translation outputs by BLEU metric has been known for a long time. The main drawback of BLEU score arises from the fact that it works based on matching of the sequence of words between a reference human translation and a machine translation output. In the core, this method is in contrast with the way human experts asses translations. In the human assessment of two different translations of a single text, the emphasis is on similarity of the meaning and semantics rather than degree of imitation between two translations. Considering this fact, in this study, I evaluated the idea of constructing of a NMT assessment metric based on sentence's semantics. This project was divided into 3 phases. In the first phase, I selected and prepared necessary data for training and testing of a NMT model from ParsiNLU Persian NLP bench mark. Selection of an appropriate NMT model for Persian-English translation and generation of a NMT output was done in the second phase. Initially, homework 4 LSTM - RNN based code was selected as the NMT model. After different trials and errors to get the model worked on the selected dataset, the model was finally trained for 14 hours. In the testing, however, the trained model did not generate any meaningful output and the translation BLEU score was close to zero. Therefore, the NMT model was chanced to transformer based mT5 model. After different trials and errors, the model parameters were selected and the mT5 model was trained on a selected Mizan Persian-English dataset from ParsiNLU suit. The dataset included 996,371 pairs of Persian-English sentences and training took around 2.5 days. For testing of the model, the Quran dataset from ParsiNLU suite was selected. 9 different high quality human translations were available for this dataset. I introduced a new metrics based on average cosine similarities of sentence embedding generated by Siamese Bert model. The comparison between this new metric and BLEU score has been done and results has been discussed in this report.

3. Introduction

There has been tremendous progress in the field of neural machine translation (NMT) in recent years through the use of better algorithms, wealth of available data and more powerful hardware. One of the challenges in the field of NMT is to come up with a metric that can access the performance of NMT. Human evaluation of NMT is time consuming and expensive and cannot be automated and reused. In 2002, Papineni and other researchers at IBM Watson research center introduced an automatic evaluation score with name of Bilingual Evaluation Understudy or BLEU for short [1]. Since the introduction of BLEU score, it has served as a defacto standard in the evaluation machine translation outputs and substituted the skilled human machine translation evaluation.

There are, however, some deficiencies in BLEU score metrics. The BLEU score is based on a simplistic text string matching between different sources of translations. The more a NMT output imitates the human translation, the higher the BLEU score is. Fomicheva et al [2] found that there can be up to 6-point difference in BLEU scores of a machine translation output depending on the reference that was used. They showed that BLEU score can strongly penalize a good translation that happen to use the word that are different from the provided reference. The NMT goal is to generate a text that delivers the true sentence meaning with highest degree of accuracy in the output and not just copying a human generated translation.

One way to enhance the BLEU score accuracy is to compare a NMT output against multiple human generated translation [3]. This method is however costly and labor intensive and cannot address the root problem of BLEU score. Fomincheva et al [4] explored the idea of generating different reference translations using paraphrasing and synonyms through the use machine translation algorithms. Echizen et al [5] proposed a new metrics for evaluation of NMT using word embeddings and their position information. Yankovskaya et al [6] used a pretrained BERT model and LASER sentence level embedding and feed them through a feed-forward neural network to predict NMT performance. It is logical to assume that accessing NMT performance based on sentence embedding, generated by methods like BERT, offer a better performance metrics than BLEU. This metrics will consider sentence semantics and true meanings and not n-gram similarities.

There are, however, some problems associated with BERT sentence embeddings. Li et al [7], pointed out that sentence embeddings generated by pre-trained language models like BERT poorly capture semantic meaning of sentences. New models has been introduced by researcher to address the deficiencies of BERT sentence embedding in the preservation of semantics. The main models include BERT-flow and BERT-sentence models that are discussed in following sections.

3.1. Models for preservation of semantics in BERT sentence embedding

Li et al [7] found out that sentence embeddings generated by BERT without any fine tuning is not a suitable tool to find semantic textual similarities among different sentences. They argued that BERT embeddings without any post processing under performs even simpler models like GLOVE. Ethayarajh et al [8] discovered that BERT sentence embedding space suffers from anisotropy. Reimers et al [9] demonstrated that BERT sentence embeddings lag behind the state-of-the-art sentence embeddings in terms of semantic similarity. Li et al [7] observed that BERT sentence embedding space is non-smooth

and poorly defined in some areas. They argued that this non-smoothness of BERT embedding space makes the embeddings hard to be used by simple semantic similarity methods such dot product or cosine similarity. Following two models are the main models that have been introduced to address issues in the preservation semantic in BERT sentence semantics.

3.1.1. BERT Flow model

Li et al, introduced Bert-flow method which is constructed by transformation of BERT senetnce embeddings into isotropic Gaussian latent space [7]. Li et al showed that, in word embedding space, the high frequency words are close to origin whereas the low frequency words are far from the origin. This finding highlights the effect of word frequency on the anisotropy of word embedding space. Li et al, also, showed that low frequency word embeddings tend to disperse sparsely in embedding space. Because of this dispersity, many holes can be induced around the sparse low frequency word embeddings. These holes violate the convexity of embedding space and causes the semantics to poorly defined around those areas. These poorly defined semantic areas cause distortion in similarities among words. To overcome these problem, Li et al [8] used an invertible mapping from BERT embedding space to standard Gaussian latent space as illustrated in picture 1.

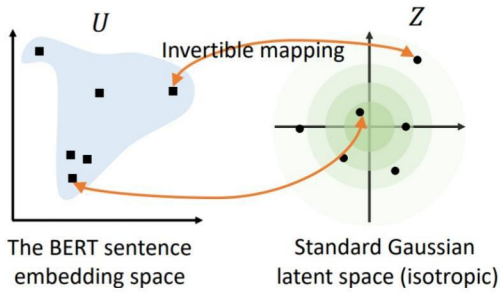

Figure 1. Illustration of the invertible transformation from BERT sentence embedding space to standard Gaussian latent space.

Li et all introduced following formulation to transform embedding vectors from embbeding space u to Gaussian space z.

$$\max_{\phi} \mathbb{E}_{\mathbf{u}=\text{BERT(sentence)},\text{sentence}\sim\mathcal{D}}$$
$$\log p_{\mathcal{Z}}(f_{\phi}^{-1}(\mathbf{u})) + \log \left|\det \frac{\partial f_{\phi}^{-1}(\mathbf{u})}{\partial \mathbf{u}}\right|,$$

Where, $u = f_{\phi}(z)$ is the invertible transformation function from observed embedding space, u, to Gaussian space z, p_z is the probability density function that need to be maximized through learning process and det function is the determinant operator. During the training process, the parameters of inverse of $f_{\phi}(z)$ or $z = f_{\phi}^{-1}(u)$ will be learned so we can have a mapping from embedding space u, to Gaussian space, z. Li et al argued that the standard Gaussian latent space satisfies isotropy. Therefore, the anisotropy of BERT embeddings space diminishes by transformation to latent Gaussian space. Li et al [8] showed that the BERT flow model improve the preservation of semantics in BERT sentence embedding. The results of their study will be discussed in section 3.1.3.

3.1.2. Sentence-BERT Model

Reimers et al introduced Sentence-BERT model based on Siamese and triplet networks[9]. Sentence-BERT model (SBERT) fine-tunes BERT model parameters in a Siamese / triplet network architecture. Reimers et al evaluated the quality of SBERT algorithm on various common Semantic Textual Similarity (STS) tasks and showed that SBERT model could achieve a significant improvement over state-of-the-art sentence embeddings methods. They also showed that Replacing BERT with RoBERTa did not yield a significant improvement in our experiments.

3.1.3. Comparison of Performance of Bert flow and Sentence flow models

For evaluation of the performance of different embedding methods, it is common to use Semantic Textual Similarity (STS) tasks. The usual metric to assess the performance of an embedding generation method in STS benchmarks is Spearman's rank correlation or ρ. The Spearman's metric ranks the correlation between the cosine similarity of sentence representations and the gold labels for various Textual Similarity (STS) tasks. The comparison of between the performance of SBERT model with other embedding models based on Spearman's metric is shown in table 1.

Table 1. Comparison of the performance Sentence – BERT (SBERT) model with other embedding methods using STS tasks and Spearman's metric [9]

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
InferSent - Glove	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder	64.49	67.80	64.61	76.83	73.18	74.92	**76.69**	71.22
SBERT-NLI-base	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT-NLI-large	72.27	**78.46**	**74.90**	80.99	76.25	**79.23**	73.75	76.55
SRoBERTa-NLI-base	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa-NLI-large	**74.53**	77.00	73.18	**81.85**	**76.82**	79.10	74.29	**76.68**

The same assessment of the performance of BERT-flow model in preservation of semantics in the sentence embedding has bee done by Li et al [7]. They used same STS benchmarks and Spearman's factor to access the

Table 2. Comparison of the performance of BERT-flow model with other embedding methods using STS tasks and Spearman's metric[9]

Dataset	STS-B	SICK-R	STS-12	STS-13	STS-14	STS-15	STS-16
Published in (Reimers and Gurevych, 2019)							
Avg. GloVe embeddings	58.02	53.76	55.14	70.66	59.73	68.25	63.66
Avg. BERT embeddings	46.35	58.40	38.78	57.98	57.98	63.15	61.06
BERT CLS-vector	16.50	42.63	20.16	30.01	20.09	36.88	38.03
Our Implementation							
$BERT_{base}$	47.29	58.21	49.07	55.92	54.75	62.75	65.19
$BERT_{base}$-last2avg	59.04	63.75	57.84	61.95	62.48	70.95	69.81
$BERT_{base}$-flow (NLI*)	58.56 (↓)	**65.44** (↑)	59.54 (↑)	64.69 (↑)	64.66 (↑)	72.92 (↑)	71.84 (↑)
$BERT_{base}$-flow (target)	70.72 (↑)	63.11(↓)	63.48 (↑)	72.14 (↑)	68.42 (↑)	73.77 (↑)	75.37 (↑)
$BERT_{large}$	46.99	53.74	46.89	53.32	49.27	56.54	61.63
$BERT_{large}$-last2avg	59.56	60.22	57.68	61.37	61.02	68.04	70.32
$BERT_{large}$-flow (NLI*)	68.09 (↑)	64.62 (↑)	61.72 (↑)	66.05 (↑)	66.34 (↑)	74.87 (↑)	74.47 (↑)
$BERT_{large}$-flow (target)	**72.26** (↑)	62.50 (↑)	**65.20** (↑)	**73.39** (↑)	**69.42** (↑)	**74.92** (↑)	**77.63** (↑)

Based on the above tables, the calculated average of Spearman's metric for BERT-flow base and BERT-flow large models are 69.57 and 70.76, respectively. Also, as shown in table 1, the average Spearman's scores for SBER method is 74.89 for and 76.55 for SBERT base and SBERT large, respectively. Therefore, the averages Spearman's factors across different STS benchmarks for BERT flow model are lower than those of Sentence-BERT method. Additionally, sentence-BERT model code is readily available to use whereas BERT flow model code lacks the availability for public use. Based on this facts, the SBERT model was used to construct the new metric for assessment of NMT performance.

4. Approach

The motivation of this project is to address the current drawbacks of BLEU metric and try to come up with a metrics based on semantics. In this project, I evaluated the idea of construction of a NMT evaluation metric based on average cosine similarity of sentence embeddings. The overall steps of the project are:

1. Preparation of training and testing datasets from ParsiNLU Persian NLP benchmark for Persian – English translation.
2. Finding a NMT model for Persian – English translation.
3. Training and testing of the NMT model and generating NMT output.
4. Assessment of the performance of the NMT model based on BLEU score and with comparison to 9 sets of high quality human English translation of a unique Persian text.
5. Calculation of word embeddings of NMT English output and embeddings of 9 sets of Persian-English test human translations based on sentence-BERT method.
6. Calculation of BLEU scores and cosine similarity values between human translation 1 with respect to other 8 reference translations.
7. Calculating the cosine similarity between NMT sentences embeddings and the sentence embeddings of 9 different sets of human generated English translations.
8. Calculation of the average of cosine similarity values of different pairs of translations and report it as the new NMT metrics.
9. Analyzing the results and assessing the merits of the project idea.

5. Experiments

5.1. Data

5.1.1. Preparation of Training dataset

The PasriNLU datasets (Kashabi et al, 2020, [10]) was selected for training of a NMT model for Persian-English translation. ParsiNLU is a comprehensive suit of high-level NLP tasks for Persian language. One of the data set within the ParsiNLU suit is the Persian-English NMT dataset which includes 1,617,788 pairs of Farsi-English sentences. This set includes different subsets of data from different sources. I tested

different ideas for preparation of dataset like random shuffling of rows of dataset or random selection of pairs of English -Persian dataset to prevent bias toward a dataset. However, I finally decided to select a dataset without random shuffling or selection to prevent disruption of flow of semantics among neighbor sentences. The final subset that was selected from ParsiNLU NMT datasets for training of the NMT model was Mizan dataset which includes 996,371 pairs of Persian-English sentences.

5.1.2. Preparation of Testing Dataset

The intention of this section was to prepare several authentic human translations for a single Persian text. A section of the ParsiNLU NMT test datasets that contains multiple English translations of a single Persian translation of Quran text was selected. The datasets contains 6,228 pairs of Persian – English sentences. There are 9 different human translations among the English section of the dataset. The Persian and 9 English sections were separated and saved into 9 test datasets.

5.2. Evaluation Method

5.2.1. NMT Model

Efforts were made to find a NMT model for Persian – English translation and generating necessary output for this project. For this part, I, first, tried to use my homework 4 RNN-LSTM NMT code. The homework 4 NMT code was slightly modified to allow long training before an early termination. The model was trained using the train datasets that was described section 5.1. After some trial and errors, the code was trained on the dataset. Here are the summary of training outcomes:

- Number of Epochs: 15
- Number of Iterations: 43,800
- Average loss at the code termination: 55.40
- Training Elapsed Time: 51,556 sec equal to around 14 hours

However, during the testing, it was noticed that the model did not generate any meaningful results and the output BLEU score was close to zero. The homework 4 code is mostly geared toward polysynthetic languages like Cherokee and the model doesn't work well on Persian language which is not a polysynthetic language. The sub-word tokenization in the code has probably disturbed the correspondence between Persian and English tokens and caused the deficiency in translation. So, I changed the NMT model to transformer based mT5 model. Different trials were made to find the mT5 model parameters to successfully train the model. The plot of loss versus the number of iterations for two final trails is shown in the figure 2.

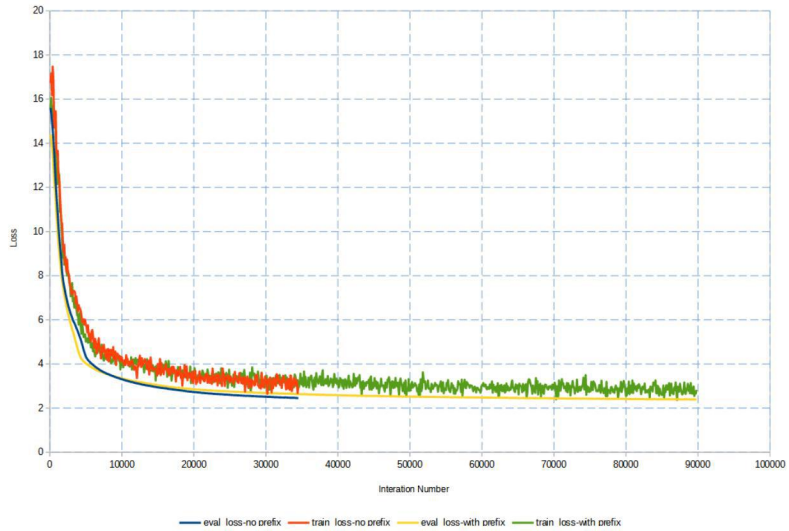

Figure 2. The plot of loss versus iteration for two final training trials of mT5 model

Two epochs were used in the final training and it took almost 2.5 days to finish the training. After training, the model was tested on the dataset, described in the section 5.1, and the necessary output for construction of the cosine similarity based NMT metric was generated.

5.2.2. Calculation of Semantic Base Metric:

The human translation reference texts and the mT5 translation output were fed into sentence-BERT model to calculate the sentence embeddings. The dimension of sentence embedding vector was 768 and the cosine similarity was calculated using following formulation.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

The average values of the cosine similarities over the entire datasets for each pairs of translations were calculated and reported as the new NMT performance metric. Similar to BLEU score, I multiplied the new score by 100 to have a metric between 0 and 100 in which the higher scores means more semantic similarity to reference and therefore higher quality translation.

6. Results

6.1. Compression of Reference Human Translations

As described in the data section, 9 high quality human translations of a single Persian text were used to evaluate the idea of the new cosine similarity based metric. These 9 different texts are semantically the same since they are the high quality human translations of a single Persian text. Therefore, a correct translation metric needs show a high score in comparison of one reference with others. For evaluation of this idea, I compared translation 1 with other 8 translations based of their BLEU score the new cosine similarity score. The results are shown in figure 3.

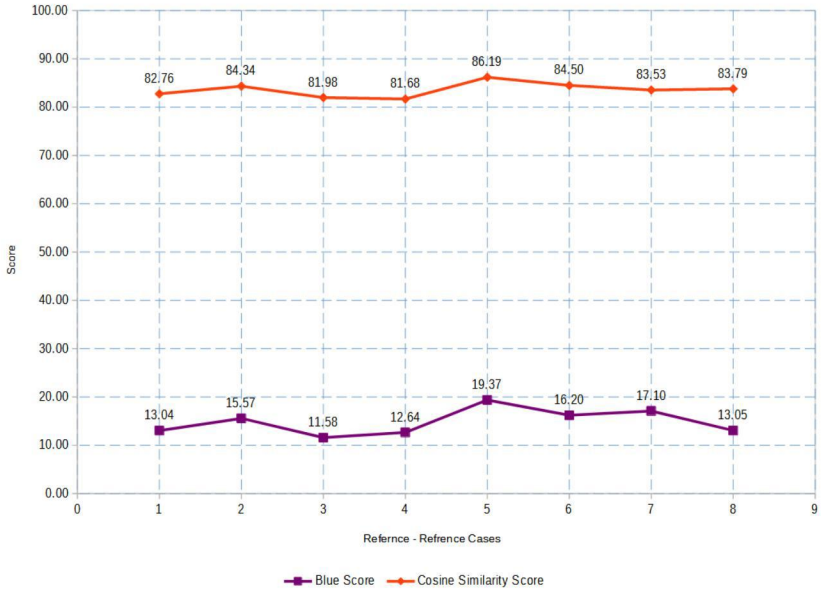

Figure 3. The comparison of reference translation 1 with other reference translations based on both BLEU score and cosine similarity score.

6.1. Compression of NMT output with different Human Translations

The comparison of the output of NMT translation with 9 different human translation references based on BLEU score and cosine similarity scores was done and the results are shown in figure 4.

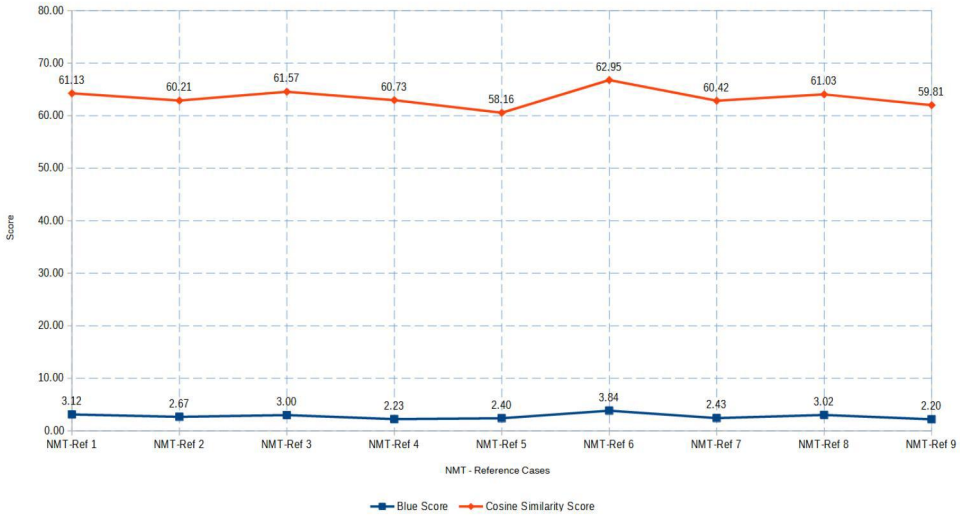

Figure 4. Comparison of NMT output with different human translation references based on BLEU score and cosine similarity scores.

7. Analysis

As I discussed in previous sections, human comparing different translations based on their semantic similarities and not the degree of repetition of words between two texts. In the case of comparison of different translations of a same input text, we expect that a good metric, that mimics the human judgment, will have a high score. As shown in figure 3, the BLEU score between reference is quit low with the average of 14.8 for comparison of reference 1 translation with other translations. This is only because of different words or sequences of words used in different translations. Therefore, we have a low degree of imitation between different texts and, consequently, low BLEU scores among them. In contrast, the average cosine similarity score, which is based on cosine similarity score between reference translation 1 and other translation is high with average of 8.36 as shown in table 3.

Table 3. The average and standard deviation of BLEU and cosine similarity scores for comparison between reference translations

	Blue Score	Cosine Similarity Score
Average	14.82	83.60
standard Deviation	2.50	1.37

This fact gives us a good confidence that the new cosine similarity score is close to human judgment and can replace BLEU score for the assessment of NMT or other machine translation performance. The averages and standard deviations of BLEU and cosine similarity scores in the case of comparison of NMT output with reference translations are mentioned in table 3.

Table 3. The average and standard deviation of BLEU and cosine similarity scores for comparison between NMT and different reference translations

	Blue Score	Cosine Similarity Score
Average	2.77	60.67
standard Deviation	0.5	1.2

Again as we see here, the average BLEU score is very low whereas the average cosine similarity score is relatively high which shows a translation with a moderate quality. As a native Persian speaker, I can confirm that the output of the NMT translation has a moderate quality.

8. Conclusion

We introduced a NMT assessment metric based on the semantic similarity between different translations. This metric is constructed based average cosine similarity of sentence embeddings of two different translations. I showed that this new score is close to human judgment in the assessment of the semantically identical translations and can replace BLEU metric for the assessment of machine translations.

4. References

1. Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, Bleu: A Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July, 2002, Pennsylvania, USA

2. Marina Fomicheva, Lucia Specia, Reference Bias in Monolingual Machine Translation Evaluation, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistic, August 2016, Berlin, Germany

3. Markus Dreyer and Daniel Marcu. 2012. Hyter, Meaning-equivalent semantics for translation evaluation. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2012 pages 162–171.

4. Marina Fomicheva, Lucia Specia, Francisco Guzmán, Multi-Hypothesis Machine Translation Evaluation, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July, 2020,

5. Hiroshi Echizen'ya, Kenji Araki,Eduard Hovy, Word Embedding-Based Automatic MT Evaluation Metric, using Word Position Information, Proceedings of NAACL-HLT 2019, pages 1874–1883, Minneapolis, Minnesota, June 2 - June 7, 2019

6. Lisa Yankovskaya, Andre Tattar, Mark Fishel, Quality Estimation and Translation Metrics via Pre-trained Word and Sentence Embeddings, Proceedings of the Fourth Conference on Machine Translation (WMT), Vol. 3, pages 101–105, August, 2019, Florence, Italy

7. Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, Lei Li, On the Sentence Embeddings from Pre-trained Language Models, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 9119–9130, November, 2020

8. Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In Proceedings of EMNLP-IJCNLP.

9. Nils Reimers and Iryna Gurevych. 2019. Sentence BERT: Sentence embeddings using Siamese BERTnetworks. In Proceedings of EMNLP-IJCNLP.

10. Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabadi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, Yadollah Yaghoobzadeh, ParsiNLU: A Suite of Language Understanding Challenges for Persian, Decmeber 2020, arXiv:2012.06154 [cs.CL], https://arxiv.org/abs/2012.06154