YNU-HPCC at SemEval-2022 Task 8: Transformer-based Ensemble Model for Multilingual News Article Similarity

Zihan Nai, Jin Wang and **Xuejie Zhang**
School of Information Science and Engineering
Yunnan University
Kunming, China
{wangjin,xjzhang}@ynu.edu.cn

Abstract

This paper describes the system submitted by our team (YNU-HPCC) to SemEval-2022 Task 8: Multilingual news article similarity. This task requires participants to develop a system which could evaluate the similarity between multilingual news article pairs. We propose an approach that relies on Transformers to compute the similarity between pairs of news. We tried different models namely BERT, ALBERT, ELECTRA, RoBERTa, M-BERT and Compared their results. At last, we chose M-BERT as our System, which has achieved the best Pearson Correlation Coefficient score of 0.738.

1 Introduction

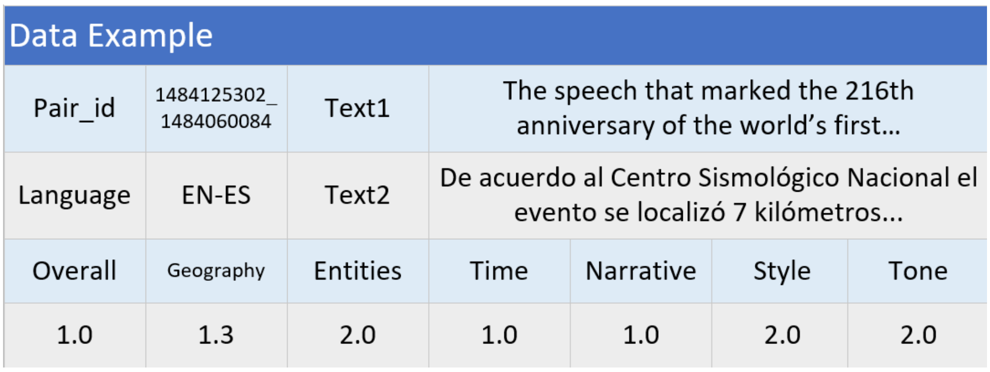
In the field of Natural Language Processing (NLP), how to measure the similarity of two texts has been a topic of interest among researchers for decades. Text similarity measures play an increasingly important role in text related research and applications in tasks such as information retrieval, text classification, document clustering, topic detection, topic tracking, questions generation, question answering, essay scoring, short answer scoring, machine translation, text summarization and others(Gomaa, Fahmy, et al. 2013). It is widely known that text is a high-dimensional semantic space, hence how to abstractly decompose it so that we can quantify its similarity from a mathematical point of view has become the focus for many researchers. There are three methods to measure text similarity: one is the traditional method based on keyword matching, such as N-gram similarity; the second is to map the text to the vector space, and then use the cosine similarity and other methods; the third is the method of deep learning, such as the deep learning semantic matching model DSSM based on user click data, ConvNet based on convolutional neural network, and the current state-of-art Siamese LSTM and other methods. However, since the introduction of bidirectional encoder representations from transformers (BERT)(Devlin, Chang, Lee, and Toutanova 2018), the accuracy and training efficiency in both text classification and sequence labeling have reached new heights.

SemEval 2022 Task 8 is a multilingual news article similarity task(X. Chen, Zeynali, Camargo, Flöck, Gaffney, Grabowicz, Hale, Jurgens, and Samory 2022). There are mainly 3 difficulties in the task compared with regular text similarity task: (1) the task is interested in the real world-happenings covered in the news articles, not their style of writing, political spin, tone, or any other more subjective design. Therefore, the system built by participant should neglect the subjective part of the text and focus on objective part only;(2) there were over six different languages in both training and test dataset, and some of test data set were composed of multilingual text pair to test the multilingual ability of participating system; (3) the test dataset contains new languages which have never appeared in training data. This situation disguisedly reduces the training data required to train the model.

In this paper, we primarily present a deep learning system for the SemEval-2021 Task 8: Multilingual News Article Similarity. Since there are not any subtasks in the SemEval 2022 Task 8, our system will focus on calculating the overall similarity only. Our approach is based on Transformers, which is a classic NLP model proposed by Google's team in 2017(Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin 2017). We fine-tuned pre-trained masked language models namely BERT, ALBERT, ELECTRA, RoBERTa and M-BERT, which are all based on Transformers. We compared their performance at the task, then picked best of them as our system.

Experimental results show that most of the Transformers based model are valid in the field of text similarity(Mittal and Modi 2021). However, when it comes to multilingual text, there is a clear drop

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1215 - 1220
July 14-15, 2022 ©2022 Association for Computational Linguistics



Data Example						
Pair_id	1484125302_ 1484060084	Text1	The speech that marked the 216th anniversary of the world's first...			
Language	EN-ES	Text2	De acuerdo al Centro Sismológico Nacional el evento se localizó 7 kilómetros...			
Overall	Geography	Entities	Time	Narrative	Style	Tone
1.0	1.3	2.0	1.0	1.0	2.0	2.0

Figure 1: the example of the given training data

in results for all models. Among these models, M-BERT has achieved the best score with the Pearson Correlation Coefficient of 0.738. That is why we chose M-BERT as our system at last. The implementation for our system is made available via Github [1].

The rest of the paper is organized as follow. Section 2 describes the specific requirements of the task and dataset. Section 3 describes the details of the Transformers model used in our system. Section 4 presents the experimental results. Finally, the results and conclusions are presented in Section 4 and 5.

2 Task Overview

The task organizers have provided training and test dataset for the task. The details of this task are given below, and some examples are shown in the Figure 1.

2.1 Problem Description

Given a pair of news articles, are they covering the same news story? Based on the same event, different journalist could write completely different news article, because of their political stance. The kernel of this task is to ignore the subjective part of news article, and calculate the similarity score according to objective facts such as time, geography, entities, etc. A pair of news article will be rate pairwise on a 4-point scale from most to least similar. Systems will be evaluated on their ability to estimate the Overall Similarity between two pairs of news stories, not any of the other scores. The similarity ratings will be compared with the gold standard ratings using Pearson's correlation.

2.2 Data Description

The task organizers have provided training and test dataset. The training data consists of 4,964 pairs of news articles, and every pair of news articles have their unique pair_id, the counts of language-pairs is following: en-en: 1800, de-de: 857, de-en: 577, es-es: 570, tr-tr: 465, pl-pl: 349, ar-ar: 274, fr-fr: 72. Apart from the overall score, the score of "Geography", "Entities", "Time", "Narrative", "Style", and "Tone" are also given to contestants. However, they are all for reference only and will not be evaluated as final score. The test data consists of 4954 pairs of news articles. It is worth mentioning that the test dataset is contained more forms of pairs of multilingual news article which are not existed in training data.

3 System Description

We use the transformer based pre-trained model as solution to accomplish the task. As shown in Figure 2, the system we built contains a tokenizer, a model layer, fully connected layer and mean squared error function. The mean squared error is loss function of our system. The model layer represent a Transfomers based pre-trained model, it will be replaced by BERT or any model mentioned above to compare their effect on task 8. The rest of the section will describe the details of every part of the system and their mechanics.

3.1 Tokenizer

Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens. Since the requirement of the task is to evaluate the similarity between two news articles, both the articles will be entered into the tokenizer at the same time. To

[1] https://github.com/151140043/Sem2022task8.git

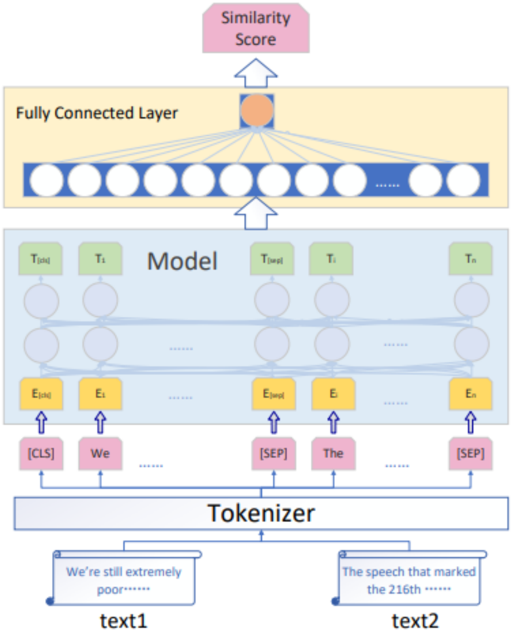


Figure 2: The structure of system

distinguish them, the tokenizer will add a special token named "[SEP]" between them and the last of second article. in addition to "[SEP]", the tokenizer will also add a special token named "[CLS]" at the begin of the first article which is necessary for subsequent step.

3.2 Model

Model layer is an abstraction of a pre-trained Transformer model, it can be a BERT model or any models mentioned below in practical application. In this step, the tokens we got from tokenizer will be passed to the layer, and layer will generate 768-dimensional word embeddings for each word in the news article(Xinge Ma and Zhang 2021) (Most of the models we used generate 768-dimensional word embeddings, a few models generate other parameters, which we will mention later). Then, the model will take the word embeddings of the first token of each article (i.e., '[CLS]') to evaluate the similarity between two news articles, because it integrated the semantic information of the whole sentence. Below we will introduce the specific model we used in the task.

BERT. BERT is a pretrained language representation model, which stands for Bidirectional Encoder Representations from Transformers (Devlin, Chang, Lee, and Toutanova 2018). BERT builds two pre-training tasks, Masked Language Model (MLM) and Next Sentence Prediction (NSP). Unlike traditional left-to-right language model pre-training, BERT is using a MLM pre-training objective, which make BERT to generate deep Bidirectional Linguistic Representation (Devlin J, Chang M W, Lee K, et al. 2018). We used the "bert-base-uncased" in our task. The size of BERT model we use in the task: Layers=12, Hidden Dimension=768, self-attention head =12, Word Piece embedding size =768, Total Parameters=110M.

ALBERT. ALBERT,which stands for A Lite Bert, is proposed to solve the problem that the parameters of the current pre-training model are too large (Lan, M. Chen, Goodman, Gimpel, Sharma, and Soricut 2019). In the classic BERT model, the size of Word Piece embedding (E) is always the same as the hidden layer size(H), i.e., E=H. ALBERT break the binding relationship between E and H, thereby reducing the number of parameters of the model and improving the performance of the model. Another method for ALBERT to reduce the amount of parameters is parameter sharing between layers, which mean, multiple layers could use the same parameters. There are three ways to share parameters: (1) Only share the parameters of the feed-forward network. (2) Only share the parameters of the attention. (3) Share all the parameters. Through these methods, ALBERT could greatly reduce the total parameters. We chose "albert-base-v2" as the model, and the size of model we use in the task: L=12, H=768, A=12, E=128, Total Parameters=12M.

ELECTRA. ELECTRA is a model that share some ideas with BERT, but the main structure is still different. It also can be named as "Efficiently Learning an Encoder that Classifies Token Replacements Accurately" (Clark, Luong, Le, and Manning 2020). The pre-training of ELECTRA can be divided into two parts, which are generator and discriminator. The generator is still MLM, the structure is similar to BERT, but the model will be much smaller than BERT. The output of generator is the input of discriminator. The role of discriminator is to distinguish whether each token input is original or replaced. For each token, the discriminator will perform a binary classification on it, and get the loss. The approach above is called replaced token detection. We chose the "google/electra-base-discriminator" as our model, whose size is: L=12, H=768, A=12, E=768, Total Parameters=110M.

RoBERTa. The full name of RoBERTa is "Ro-

model name	L	H	E	A	P
BERT	12	768	768	12	110M
ALBERT	12	768	128	12	12M
ELECTRA	12	768	768	12	110M
RoBERTa	24	1024	1024	16	355M
M-BERT	12	768	768	12	110M

Table 1: Model structure
L represents L Layers,H represents Hidden Dimension H, E represents WordPiece embedding size E, A represents A self-attention head, P represents Total Parameters

bustly optimized BERT approach" (Liu, Ott, Goyal, Du, Joshi, D. Chen, Levy, Lewis, Zettlemoyer, and Stoyanov 2019). From the perspective of the model, there are not novel innovation in RoBERTa. There are only some adjustment made on the basis of BERT: 1) The training time is longer, the batch size is larger, and the training data is more; 2) The next predict loss is removed; 3) The training sequence is longer; 4) The Masking mechanism is dynamically adjusted. The model we used in the task is "roberta-base", and the architecture of it is: L = 24, H = 1024, E=1024, A = 16, Total Parameters =355M.

M-BERT. The structure of Multilingual-BERT(M-BERT) is exactly the same with the common BERT model. The biggest difference between M-BERT and BERT is that M-BERT is pre-trained on the top 104 languages with the largest Wikipedia using a masked language modeling (MLM) objective. While the common BERT is pre-trained on English Corpus.

3.3 Transformers

Transformers is the base of all the model we mentioned above. Like many neural sequence transduction models, Transformers also have an encoder-decoder structure.

Encoder. The encoder is composed of a stack of N = 6 identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a fully connected feed-forward network (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin 2017). The output of the sub-layer can be expressed as:

$$sub\_layer\_output = LayerNorm(x + Sublayer(x))$$

Decoder. The decoder is also composed of a stack of N = 6 identical layers. In addition to the two sub-layers in each encoder layer, the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack.

Multi-Head Attention

Given a set of vector set values, and a vector query, the attention mechanism is a mechanism that computes a weighted sum of values based on the query. In Transformers, they compute attention as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

The multi-head attention allows the model to concatenate different attention results, and it can be represent as:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O$$

$$where \quad head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

3.4 Fully Connected Layer

In the fully connected layer, word embeddings acquired from the previous step will be converted into 1-dimensional numerical values. then, the fully connected layer will output the similarity score depend on the 1-dimensional numerical values.

4 Experiments

4.1 Data Preprocessing

As shown in Figure 2, tokenizer is the data processing structure of our system. After feeding data into tokenizer, it will return a dictionary of 3 lists of ints, which are input_ids, attention_mask and token_type_ids. The input ids are token indices, numerical representations of tokens building the sequences that will be used as input by the model. The attention mask is an optional argument used when batching sequences together. This argument indicates to the model which tokens should be attended to, and which should not. The token_type_ids allow some models to understand where one sequence ends and where another begins. But RoBERTa is an exception, RoBERTa removes NSP, so RoBERTa do not need the token_type_ids as input. We will split ten percent of the training data as validation data to prevent overfitting.

4.2 Evaluation Metrics

Systems will be evaluated on their ability to estimate the Overall Similarity between two pairs of news stories, not any of the other scores. The similarity ratings will be compared with the gold standard ratings using Pearson's correlation.

Epoch	1	2	3	4	5
score	0.681	0.712	0.738	0.721	0.694

Table 2: The relationship between epoch and Pearson coefficient

4.3 Implementation Details

The 5 models that mentioned above are all applicated in the task to evaluate the overall similarity between the given pairs of news articles. The datasets we used were all provided by the competition, with no other external corpus. For all the models, we set learning rate = $5e^{-6}$, epsilon=$1e^{-8}$, loss function as mean squared error, and a batch size of 8 for three epochs.

4.4 Hyper-parameters Fine-tuning

In the experiment, we have tried to change specific parameters while controlling other parameters unchanged, to see if we can get better results. The following will introduce our attempts on fine-tuning parameters.

Loss function. At the beginning of the experiment, we had to choose loss function from mean squared error and cosine similarity. Therefore, we trained an ALBERT model with the loss function of mean squared error based on the news article provided by official, and get the Pearson's correlation of 0.543. Then, we trained another ALBERT model with the loss function of cosine similarity, but only got Pearson's correlation score of 0.055. As a result, we chose mean squared error as loss function.

Batch size. Appropriate batch size is important for the optimization of model. If batch size is too small, the result may be poor. If the batch is too large, it will cause memory overflow. So we chose a batch size of 8.

Epoch. Take M-BERT as an example, the test data provided by task organizers are feed into the model to exam the effect of different epochs. The relationship between epoch and Pearson coefficient score is shown in table 2. It is obvious that the Pearson coefficient score come to the highest when epoch=3. So, we set the epoch equals to 3 in the experiment.

4.5 Comparative Results and Discussion

The results are evaluated by Pearson Correlation Coefficient with the test data provided by official, which is shown in table 3. BERT reach an accuracy of 0.464, ALBERT of 0.543, ELECTRA of 0.474, RoBERTa of 0.475, and M-BERT of 0.738. Our best individual score is 0.738 for M-BERT. As can be seen from the results, BERT, ALBERT, ELECTRA and RoBERTa have similar scores which greater than 0.45 and less than 0.5. M-BERT is the highest among them, whose score is over 0.7. Our results show that M-BERT is able to perform cross-lingual generalization surprisingly well. We believe that the reason why M-BERT outperforms other models is that M-BERT is pre-trained on the Corpus contained 104 languages while other models are pre-trained on a Corpus contained English only. Our conjectures are not groundless. A research (Papadimitriou, Chi, Futrell, and Mahowald 2021) demonstrate that mBERT representations are influenced by high-level grammatical features that are not manifested in any one input sentence, and that this is robust across languages. And mBERT does not encode subjecthood purely syntactically, but that subjecthood embedding is continuous and dependent on semantic and discourse factors, as is proposed in much of the functional linguistics literature. But there is a defect in M-BERT which is while M-BERT's multilingual representation is able to map learned structures onto new vocabularies, it does not seem to learn systematic transformations of those structures to accommodate a target language with different word order (Pires, Schlinger, and Garrette 2019). For example, cross-script transfer is less accurate for pairs like English and Japanese, which have a different order of subject. Therefore, our experiments still have many areas for improvement.

model	Pearson Correlation Coefficient
BERT	0.464
ALBERT	0.543
ELECTRA	0.474
RoBERTa	0.475
M-BERT	0.738

Table 3: Comparable results of experiments

5 Conclusion

In this paper, we described our deep learning models for the multilingual text similarity task SemEval-2022 shared Task 8. The best Pearson's correlation score we got was 0.738. We showed that the Transformer based approaches is valid in the field of multilingual text similarity. However, our system is far from perfect, lots of possible improvement can be implemented in the current model. We would like to further explore how to improve it, and employ more interesting methods in the task.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No. 61966038. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Chen, Xi, Ali Zeynali, Chico Q. Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw A. Grabowicz, Scott A. Hale, David Jurgens, and Mattia Samory (2022). “SemEval-2022 Task 8: Multilingual news article similarity”. In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Clark, Kevin, Minh-Thang Luong, Quoc V Le, and Christopher D Manning (2020). “Electra: Pre-training text encoders as discriminators rather than generators”. In: *arXiv preprint arXiv:2003.10555*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.

Gomaa, Wael H, Aly A Fahmy, et al. (2013). “A survey of text similarity approaches”. In: *international journal of Computer Applications* 68.13, pp. 13–18.

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut (2019). “Albert: A lite bert for self-supervised learning of language representations”. In: *arXiv preprint arXiv:1909.11942*.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692*.

Mittal, Abhishek and Ashutosh Modi (2021). “ReCAM@IITK at SemEval-2021 Task 4:BERT and ALBERT based Ensemble for Abstract Word Prediction”. In: *In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*.

Papadimitriou, Isabel, Ethan A Chi, Richard Futrell, and Kyle Mahowald (2021). “Deep subjecthood: Higher-order grammatical features in multilingual BERT”. In: *arXiv preprint arXiv:2101.11043*.

Pires, Telmo, Eva Schlinger, and Dan Garrette (2019). “How multilingual is multilingual BERT?” In: *arXiv preprint arXiv:1906.01502*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30.

Xinge Ma, Jin Wang and Xuejie Zhang (2021). “YNU-HPCC at SemEval-2021 Task 11: Using a BERT Model to Extract Contributions from NLP Scholarly Articles”. In: *In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 478–484,*