TextLearner at SemEval-2020 Task 10: A Contextualized Ranking System in Solving Emphasis Selection in Text

Zhishen Yang [*]
Tokyo Institute
of Technology
zhishen.yang
@nlp.c.titech.ac.jp

Lars Wolfsteller [*]
Technical University
of Munich (TUM)
lars.wolfsteller
@tum.de

Naoaki Okazaki
Tokyo Institute
of Technology
okazaki
@c.titech.ac.jp

Abstract

This paper describes the emphasis selection system of the team TextLearner for SemEval 2020 Task 10: *Emphasis Selection For Written Text in Visual Media*. The system aims to learn the emphasis selection distribution using contextual representations extracted from pre-trained language models and a two-staged ranking model. The experimental results demonstrate the strong contextual representation power of the recent advanced transformer-based language model RoBERTa, which can be exploited using a simple but effective architecture on top.

1 Introduction

Visual communication aims at conveying intended information from the author to the audience with the help of visual elements. One of its essential design principles is to reduce the ambiguity and increase the effectiveness of communication. Visual communication usually consists of two elements: text and image. For text, the author often emphasizes some parts of the text by changing the visual representation of that part to better convey the intention. Visual communication design tools often fail in understanding the meaning of text and user intention. They rely solely on visual attributes when suggesting emphasized textual parts and their visual representations, which often leads to the wrong text emphasis (Shirani et al., 2020). Therefore, a system that could help design tools to understand the semantic meaning and common human interpretation of the text would better assist the user to design effective visual communication.

SemEval 2020 Task 10: *Emphasis Selection For Written Text in Visual Media* is a shared task that invites participants to propose methods that can model human emphasis selection for short written English text (Shirani et al., 2020). The task contains several challenges. First, emphasis selection is highly subjective; different annotator may emphasize different parts of the text. Second, there is a lack of additional context information; the model therefore has to rely only on the text. Third, there is a relatively small amount of training data; the training set provided by the organizers contains only 2742 instances of short written text.

Our strategy to tackle these three challenges is to model the emphasis selection distribution of annotators by a combination of contextual representations of text extracted from a pre-trained language model and a two-staged ranking model to suggest possible tokens for emphasis. From experimental results, we demonstrate that our simple two-staged ranking model built on top of RoBERTa (large model, fine-tuned on MNLI) (Liu et al., 2019) outperforms the LSTM-based baseline system (Shirani et al., 2020), which indicates transformer-based pre-trained language models such as RoBERTa have a strong contextual representation power and the combination of such language models with a simple ranking model is beneficial to the problem of emphasis selection.

1.1 Task Definition

The task organizers formalized this task as to determine an emphasis subset S of a sequence of tokens $C = [x_1, x_2, ..., x_n]$. Our goal is to propose a system that can generate this subset S with emphasis scores assigned to each token within it, with which we can sort the top K emphasized tokens.

[*] Equal contribution

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1691–1697
Barcelona, Spain (Online), December 12, 2020.

1.2 Dataset

The SemEval-2020 Task 10 dataset is a collection of short English texts crawled from *Adobe Spark* and *Wisdom Quotes* (Shirani et al., 2019). It contains 3877 instances with total of 44977 tokens. The task organizer randomly split 70% of the data into the training set, 10% into the development set, and the remaining 20% into the test set. The sentences in the training set contain an average of 11.8 words.

Based on the dataset statistics, we concluded the following points: First, the amount of training data is relatively small to drive training of deep neural networks. Additionally, the short text and lack of additional information make it difficult for a model to capture the contextual information, but reduce the problem of long-range dependency. Therefore, we would like to utilize a pre-trained language model to obtain the contextual representations of the tokens and use those for a ranking model to approximate the emphasis selection distribution given in the training dataset.

2 System

We propose a two-staged emphasis selection system. The first stage of this system functions as the initial top K word selection to select candidate words for emphasizing. The second stage then re-ranks these K words by re-assigning an emphasis score to each of them. Figure 1 depicts the overall system architecture.

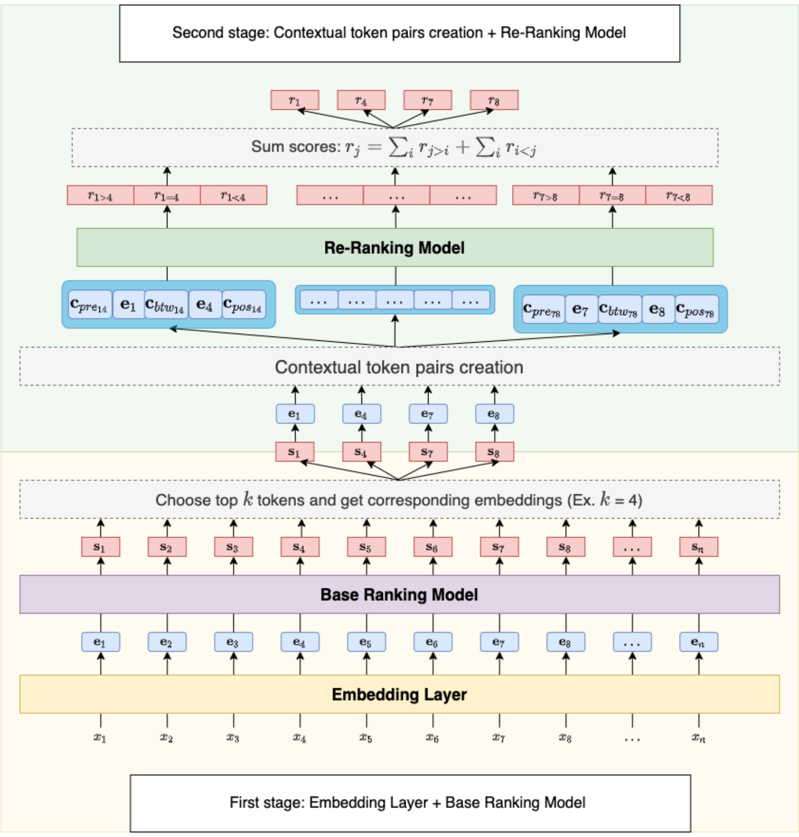

Second stage: Contextual token pairs creation + Re-Ranking Model
r_1
r_4
r_7
r_8
Sum scores: $r_j = \sum_i r_{j>i} + \sum_i r_{i<j}$
$r_{1>4}$
$r_{1=4}$
$r_{1<4}$
...
$r_{7>8}$
$r_{7=8}$
$r_{7<8}$
Re-Ranking Model
$\mathbf{c}_{pre_{14}}$ $\mathbf{e}_1$ $\mathbf{c}_{btw_{14}}$ $\mathbf{e}_4$ $\mathbf{c}_{pos_{14}}$
$\mathbf{c}_{pre_{78}}$ $\mathbf{e}_7$ $\mathbf{c}_{btw_{78}}$ $\mathbf{e}_8$ $\mathbf{c}_{pos_{78}}$
Contextual token pairs creation
$\mathbf{e}_1$ $\mathbf{e}_4$ $\mathbf{e}_7$ $\mathbf{e}_8$
$\mathbf{s}_1$ $\mathbf{s}_4$ $\mathbf{s}_7$ $\mathbf{s}_8$
Choose top k tokens and get corresponding embeddings (Ex. k = 4)
$\mathbf{s}_1$ $\mathbf{s}_2$ $\mathbf{s}_3$ $\mathbf{s}_4$ $\mathbf{s}_5$ $\mathbf{s}_6$ $\mathbf{s}_7$ $\mathbf{s}_8$ ... $\mathbf{s}_n$
Base Ranking Model
$\mathbf{e}_1$ $\mathbf{e}_2$ $\mathbf{e}_3$ $\mathbf{e}_4$ $\mathbf{e}_5$ $\mathbf{e}_6$ $\mathbf{e}_7$ $\mathbf{e}_8$ ... $\mathbf{e}_n$
Embedding Layer
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 ... x_n
First stage: Embedding Layer + Base Ranking Model

Figure 1: The figure of the proposed two-staged emphasis ranking system. The first stage (yellow box) performs the initial selection of top K words. The second stage (green box) creates contextual token pairs for the Re-Ranking Model to generate the final emphasis scores for the top K words.

First stage The first stage consists of an embedding layer and a base ranking model. Facing the limited amount of training data and lack of additional contextual information, the design of the first stage originates from the idea of utilizing the pre-trained language model to extract contextual rich representations. The base ranking model built on top of the embedding layer is meant to use those representations to further learn the emphasis selection distribution from the training data.

Given a sentence $C = [x_1, x_2, ..., x_n]$, the system first inputs this sentence into the embedding layer to obtain a list of word embedding vectors as $E = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n]$. Each of these is then fed into the base ranking model, which consists of two fully connected layers:

$$\mathbf{h}_b = \alpha(\mathbf{W}_{b,1}\mathbf{e}_i + \mathbf{b}_{b,1})$$
$$s_i = \alpha(\mathbf{W}_{b,2}\mathbf{h}_b + \mathbf{b}_{b,2}),$$

where $\mathbf{e}_i \in R^d$ is the d-dimensional embedding vector of a token x_i, $\mathbf{h}_b \in R^m$ is the hidden vector, $\mathbf{W}_{b,1} \in R^{m\times d}$ and $\mathbf{W}_{b,2} \in R^{1\times m}$ are the weight matrices and $\mathbf{b}_{b,1} \in R^m$ and $\mathbf{b}_{b,2} \in R^1$ the biases. α is an activation function. The base ranking model predicts an emphasis probability score s_i for each word embedding. Based on this, the top k words are selected as set $S_k \subseteq C$ of candidate tokens for emphasizing.

Second Stage The second stage performs a re-ranking of the candidate set S_k by performing a pairwise comparison of all tokens $x_i \in S_k$. It is inspired by Eberts and Ulges (2019), in particular by their relation classifier, which is used for joint entity and relation extraction.

The system first retrieves the embedding vectors of the top words of S_k to compose a set E_k. As a second step, it creates contextual word pairs. For each pair $(\mathbf{e}_i, \mathbf{e}_j)$ of word embeddings from E_k, we define three types of context vectors: pre-context, btw-context, and post-context:

$$\text{pre-context: } \mathbf{c}_{\text{pre}_{ij}} = \text{MaxPooling}([\mathbf{e}_1, ..., \mathbf{e}_{i-1}])$$
$$\text{btw-context: } \mathbf{c}_{\text{btw}_{ij}} = \text{MaxPooling}([\mathbf{e}_{i+1}, ..., \mathbf{e}_{j-1}])$$
$$\text{post-context: } \mathbf{c}_{\text{pos}_{ij}} = \text{MaxPooling}([\mathbf{e}_{j+1}, ..., \mathbf{e}_n])$$
$$\mathbf{c}_{\text{pre}_{ij}}, \mathbf{c}_{\text{btw}_{\text{ij}}}, \mathbf{c}_{\text{pos}_{ij}} \in R^{\text{dim}(\mathbf{e}_\text{i})}$$

Concatenating the pair $(\mathbf{e}_i, \mathbf{e}_j)$ and its context vectors gives a new training sample $\mathbf{e}_{\text{new},ij} = [\mathbf{c}_{\text{pre}_{ij}}, \mathbf{e}_i, \mathbf{c}_{\text{btw}_{ij}}, \mathbf{e}_j, \mathbf{c}_{\text{pos}_{ij}}]$ for the re-ranking model. The re-ranking model consists of N fully connected layers followed by a softmax layer and the emphasis score calculation:

$$\mathbf{h}_1 = \alpha(\mathbf{W}_{r,1}\mathbf{e}_{\text{new},ij} + \mathbf{b}_{r,1})$$
$$....$$
$$\mathbf{h}_N = \alpha(\mathbf{W}_{r,N}\mathbf{h}_N + \mathbf{b}_{r,N})$$
$$\mathbf{r} = [r_{i>j}, r_{i=j}, r_{i<j}] = \text{Softmax}(\mathbf{h}_N)$$

where $\mathbf{e}_{\text{new},ij} \in R^c,\ c \in [2 * \text{dim}(\mathbf{e}_i), 5 * \text{dim}(\mathbf{e}_i)]$ (depending on which context vectors are used, see subsection 3.3), $\mathbf{h}_l$ are the hidden vectors, $\mathbf{W}_{r,l}$ the weight matrices and $\mathbf{b}_{r,l}$ the biases for $l \in [1, N]$. $\mathbf{r} = [r_{i>j}, r_{i=j}, r_{i<j}]$ is the result vector, predicting if a word x_i has a higher emphasis probability $(r_{i>j})$, the same emphasis probability $(r_{i=j})$ or a lower emphasis probability $(r_{i<j})$ than word x_j. Finally, for a word $x_j \in S_k$, the final emphasis score is calculated as following, where top_k is the set of indices corresponding to the candidate set S_k:

$$r_j = \sum_{i \in \text{top}_k} r_{j>i} + \sum_{i \in \text{top}_k} r_{i<j}$$

3 Experiments

This section describes the experiments on selecting the embedding layer for the first stage and examining the different compositions of contextual word pairs in improving the performance of the re-ranking model at the second stage.

3.1 Evaluation Metric

The task organizer provided task-specific evaluation metric match_m (Shirani et al., 2020) as following:

$$\text{match}_m := \frac{\sum_{x \in D_{test}} |S_m^{(x)} \cap \hat{S}_m^{(x)}|/(\min(m, |x|))}{|D_{\text{test}}|} s, m \in \{1...4\}$$

In the test set D_{test}, for each sentence x, the words with top m emphasis probabilities from ground truth set forms $S_m^{(x)}$, while $\hat{S}_m^{(x)}$ consists of words with top m predicted probabilities from the prediction set. We used this metric as the only evaluation criterion in our experiments.

3.2 Embedding Layer Selection

This experiment is to select the best performing pre-trained conventional word embeddings or pre-trained language models as the embedding layer in the first stage.

For pre-trained conventional word embeddings, we selected GloVe (Pennington et al., 2014) pre-trained on Common Crawl (840B tokens). For the pre-trained language models, we selected Flair embeddings (Akbik et al., 2018) pre-trained with 1-billion word corpus (Chelba et al., 2013) and RoBERTa (large model) (Liu et al., 2019) finetuned on the Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2018) (RoBERTa-large-mnli). Since each layer in RoBERTa preserves different linguistic information (Tenney et al., 2019), we experimented with concatenating the features extracted from 17th-24th layer and concatenating 21st-24th layer to get richer contextual representation.

We trained the base ranking models with each candidate as embedding layer and compared their performances based on match_4 scores, since we will select at least 4 words for the second stage.

From Table 1, the base ranking model using GloVe pre-trained on Common Crawl performs the worst across match_m and $\text{match}_{average}$ scores. For the base ranking model using RoBERTa-large-mnli, both concatenated 17th-24th and concatenated 21st-24th layer outperform the baseline. Among those, concatenated 17th-24th layer yields the best match_4 among all candidate pre-trained embeddings.

Compared to GloVe, RoBERTa-large-mnli can embed a word with rich contextual as well with syntactic and semantic information, which provides comprehensive representations for the base ranking model to learn emphasis selection. Therefore, we selected RoBERTa-large-mnli (17th-24th layer) as embedding layer in the first stage of our proposed emphasis selection system.

Base Ranking Model (Embeddings)	match_1	match_2	match_3	match_4	$\text{match}_{average}$
Base Ranking Model (GloVe)	0.500	0.655	0.753	0.798	0.676
Base Ranking Model (Flair embedding (Akbik et al., 2018))	0.559	0.728	0.786	0.831	0.726
Base Ranking Model (RoBERTa-large-mnli, 17th-24th layer)	0.625	**0.773**	**0.825**	**0.858**	**0.770**
Base Ranking Model (RoBERTa-large-mnli, 21st-24th layer)	**0.628**	0.741	0.812	0.848	0.757
Task Baseline (Shirani et al., 2020)	0.597	0.756	0.809	0.829	0.748

Table 1: match_m scores for the base ranking model using candidate embeddings trained on training set, tested on development set. The base ranking model using RoBERTa-large-mnli (17th-24th layer) yields the best $\text{match}_{2,3,4}$ and $\text{match}_{average}$ scores.

3.3 Contextual Word Pairs

In the second stage, different compositions of contextual word pairs determine the amount of contextual information given to the re-ranking model for re-ranking the candidate words from the first stage. This experiment aims at searching for the optimal contextual word pair composition. We defined the parameters of the experiment as top words $K = 5$ (top K words from first stage to be re-ranked at the second stage) and the contextual word pair composition (whether to include surrounding context and in-between context) as Context = [[pre-context= False, btw-context= True / False, post-context= False], [pre-context= True, btw-context= True/False, post-context= True]].

Table 2 shows the experimental results. The proposed emphasis selection system with contextual word pair setting: [pre-context= True, btw-context= True, post-context= True] and [pre-context= False, btw-context= True, post-context= False] , achieves the best $\text{match}_{average} = 0.786$. Compared to the best performing base ranking model (roberta-large-mnli, 17th-24th layer) at the first stage, the best performing re-ranking model at second stage improves the match_1 score by 0.056 and the $\text{match}_{average}$ score by 0.016.

Regarding the contribution of the context information, the re-ranking model can improve the scores even when comparing word pairs without any types of context vectors involved. Including the surrounding

System	pre-context	between-context	post-context	match_1	match_2	match_3	match_4	$\text{match}_{average}$
Proposed System	True	True	True	**0.684**	**0.780**	0.832	0.850	**0.786**
Proposed System	True	False	True	0.597	0.751	0.805	0.837	0.747
Proposed System	False	True	False	**0.684**	0.772	**0.834**	**0.854**	**0.786**
Proposed System	False	False	False	0.673	0.772	0.833	0.848	0.782
Task Baseline (Shirani et al., 2020)				0.597	0.756	0.809	0.829	0.748

Table 2: match_m scores for our proposed system (base ranking model + re-ranking model) using RoBERTa-large-mnli (17th-24th layer) trained on training set, tested on development set.

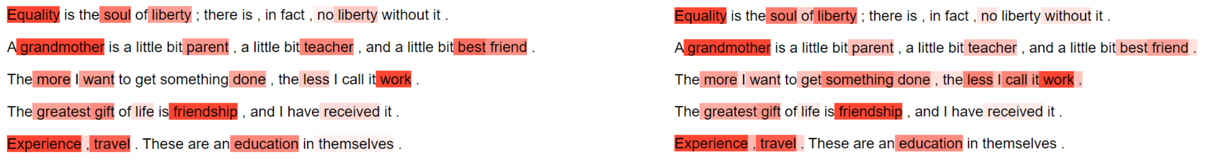
System	pre-context	between-context	post-context	match_1	match_2	match_3	match_4	$\text{match}_{average}$
Proposed System (Post evaluation phase)	True	True	True	0.634	0.774	0.830	0.859	0.774
Proposed System (Post evaluation phase)	True	False	True	0.567	0.729	0.794	0.838	0.732
Proposed System (Post evaluation phase)	False	True	False	**0.642**	**0.780**	**0.831**	**0.861**	**0.778**
Proposed System (Post evaluation phase)	False	False	False	0.634	0.775	0.827	0.858	0.773
Task Baseline (Shirani et al., 2020)				0.608	0.737	0.807	0.849	0.750
Proposed System (Evaluation Phase)				0.627	0.769	0.823	0.850	0.767

Table 3: match_m scores for our proposed system (base ranking model + re-ranking model) using RoBERTa-large-mnli (17th-24th layer) trained on training set, tested on test set.

context vectors and in-between context vectors or only in-between context vectors in the contextual word pairs slightly improves the match_m scores. Adding only the surrounding context vectors worsens the performance of the re-ranking model.

In the post-evaluation phase, we tested our proposed emphasis selection system on the test set with the same experimental settings. As Table 3 shows, the proposed emphasis selection system with contextual word pair setting [pre-context= False, between-context= True, post-context= False] outperforms the baseline system across all match_m and $\text{match}_{average}$ scores, with 0.034 improvement over match_1 score and 0.028 over $\text{match}_{average}$.

With our proposed system, we rank 21st on the final leaderboard of the evalution phase. See Table 3 for the respective match_m scores. The match_m scores in the post evaluation phase are slightly different from the submitted system in the evaluation phase due to different random initialization. Figure 2 shows some example predictions of our proposed system compared to the ground truth of the development set.



Equality is the soul of liberty ; there is , in fact , no liberty without it .

A grandmother is a little bit parent , a little bit teacher , and a little bit best friend .

The more I want to get something done , the less I call it work .

The greatest gift of life is friendship , and I have received it .

Experience , travel . These are an education in themselves .

(a) Proposed system

Equality is the soul of liberty ; there is , in fact , no liberty without it .

A grandmother is a little bit parent , a little bit teacher , and a little bit best friend .

The more I want to get something done , the less I call it work .

The greatest gift of life is friendship , and I have received it .

Experience , travel . These are an education in themselves .

(b) Ground Truth

Figure 2: Heatmap of emphasis scores from the proposed system (with contextual word pair setting [pre-context= False, between-context= True, post-context= False]) compared to ground truth emphasis probabilities of example sentences from the development set.

4 Conclusion

In this paper, we presented our emphasis selection system for SemEval 2020 Task 10: *Emphasis Selection For Written Text in Visual Media*. The system is based on an embedding layer to extract contextual representation from the input sentence, followed by a two-staged ranking model to learn the emphasis selection distribution. The experimental results show that the transformer-based language model RoBERTa provides rich contextual representation to support the proposed two-staged ranking system in outperforming the LSTM-based baseline system (Shirani et al., 2020) and successfully coping with this task.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

François Chollet et al. 2015. Keras. `https://keras.io`.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. *arXiv preprint arXiv:1909.07755*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yinhan Liu, 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. *arXiv preprint arXiv:1907.11692*.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Amirreza Shirani, Franck Dernoncourt, Paul Asente, Nedim Lipka, Seokhwan Kim, Jose Echevarria, and Thamar Solorio. 2019. Learning emphasis selection for written text in visual media from crowd-sourced label distributions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1167–1172.

Amirreza Shirani, Franck Dernoncourt, Nedim Lipka, Paul Asente, Jose Echevarria, and Thamar Solorio. 2020. Semeval-2020 task 10: Emphasis selection for written text in visual media. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

A Implementation Details

Following our design in section 2, we defined the base ranking model with two fully connected layers. The first layer has the input dimension corresponding to the dimension of the word vector and output dimension $m = 8$. The second layer has input dimension $m = 8$, and output dimension $r = 1$.

The re-ranking model at the second stage consists of four fully connected layers. The first layer has an input dimension corresponding to the contextual word pairs and output dimension of 1024. The second layer has an input dimension of 1024 and output dimension of 64. The third layer has an input dimension of 64 and output dimension of 4. The input dimension of the last layer is 4, and the output dimension is 3.

The activation function for both base ranking model and re-ranking model is LeakyReLU (Maas et al., 2013). When training the base ranking model, the optimizer is Adam (Kingma and Ba, 2014) with learning rate of 0.0005, and loss function is mean squared error. For the re-ranking model, we selected stochastic gradient descent (learning rate = 0.001, decay=0.0, momentum=0.9) as optimizer and categorical cross entropy as loss function. We used the training dataset provided by the task organizers to train the models and the development set for the initial evaluation.

We implemented our system using Keras (2.3.1) (Chollet and others, 2015) and Flair (0.4.4) (Akbik et al., 2019).