Human-Level Semantic Score Prediction from Dialogue

JaeYeon Bae **Chaeri Kim**
Department of Artifical Intelligence
Ulsan National Institute of Science and Technology(UNIST)
Ulsan 44919, Republic of Korea
{qowodussla, chaerikim}@unist.ac.kr

Abstract

Sentiment analysis is one of the famous topic in natural language processing(NLP), used to determine whether data is positive, negative or neutral. Sentiment analysis researches are actively conducted in various fields such as movie reviews and product reviews. However, previous studies on semantic analysis are based on a single sentence rather than a dialogue. In this paper, we would like to analyze emotions between people through sentiment analysis based on conversation. So we propose a human-level Semantic Score Prediction (SSP) model which utilize "turn embeddings" that allows us to know which speaker is speaking during a conversation. Additionally, we can get a "relation score" so that we can know the degree of affinity between two people. We evaluated our proposed model in synthetic dataset and real-world messenger dataset. Our model significantly outperforms baseline model.

1 Introduction

Sentiment analysis is used to determine whether given text or sentence contains negative, positive, or neutral emotions. There are many types of sentiment analysis. First one is fine-grained sentiment analysis. This provides a more precise level of polarity by taking apart it into more categories, usually very positive to very negative. Second one is emotion detection. This determines specific sentiment rather than positivity and negativity. Examples could include happiness, frustration and so on. Third one is intent-based analysis. This recognizes actions inherent in a text, along with the opinion. For example, an online comment expressing frustration about changing a battery could prompt customer service to reach out to resolve that specific issue. Last one is aspect-based analysis. This gathers the specific information being positively or negatively mentioned. For example, a customer might leave a review on a product saying the battery life was too short. Then, the system will return that the negative sentiment is not about the product as a whole, but about the battery life.

We are focusing on the first one, fine-grained sentiment analysis. There are many applications of in this field. Prior researches about this field are usually based on a single sentence. For example, when customer buy a product and write a review, sentiment analysis is conducted with a single review written by the customer. There are few researches about sentiment analysis based on a dialogue. In the work from Zou, Yicheng, et al. [1], they proposes sentiment classification model in dialogues. However, this model also classifies sentiments based on a single sentence.

In this paper, we aim to develop a model that proceeds with semantic analysis based on conversations between two people, not on a single text. So we propose human-level Semantic Score Prediction (SSP) model. This model adds information about which speaker speaks for each dialogue while embedding in addition to word embedding and positional embedding. As a result, it is possible to

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

extract relation score between two people by learning not only the semantic information of a sentence but also the emotional information of the individual in the context of the conversation.

We evaluate our model on synthetic dataset and real-world messenger dataset. Experiment on datasets shows that our proposed model outperforms baseline model. This proves that information about each speaker's turn is helpful in analyzing sentiment in dialogue.

To summarize, we can say our contribution in three aspects.

- Unlike previous researches that analyze sentiment in a sentence, we would like to analyze sentiment between people through sentiment analysis based on dialogue.
- For effective dialogue sentiment analysis, we propose a human-level Semantic Score Prediction (SSP) model which uses **turn embedding** that enables to know which speaker is speaking during conversation.
- In addition to the sentence's semantic information spoken by each speaker, the context of the conversation and the emotional information of each person can be learned to obtain a **relation score** between two people. By looking at the relation score, we can know the degree of likability.

2 Related Work

2.1 Natural Language Processing

Natural language processing (NLP) is one of the major research areas of deep learning. Because the languages are most likely to have similarities with sequential data, sequential models like Recurrent Neural Network from ROBINSON, Tony et al. [2], Long-Short Term Memory from GRAVES, Alex. [3], and Bidirectional RNN from SCHUSTER, Mike et. al [4] are used to deal with NLP tasks.

Attention mechanism from BAHDANAU, Dzmitry et al. [5] consists of RNN encoder-decoder architecture in which RNN encoder contains the information of the source sentence to be translated and RNN decoder generates the translated sentence based on the information saved above. By using the attention score which shows how meaningful the word is in translation, this could outperform the previous research in both short and long sentences.

Transformer model from VASWANI, Ashish, et al. [6], most commonly used these days, solves the essential problem in the sequential model by using only the attention mechanism in encoder-decoder modules. It outperformed in all the NLP tasks and changed the research paradigm. Bidirectional Encoder Representation of Transformers(BERT) from DEVLIN, Jacob, et al. [7], following transformer, is a pre-trained model composed of several encoders from transformer so that it could have powerful representation ability in language processing.

2.2 Semantic Analysis

By providing the result in detail, fine-grained semantics analysis is useful in predicting human emotion. Many researchers tried to predict it using BERT. ALBERT from LAN, Zhenzhong, et al.[8] presents two parameter-reduction techniques to lower memory consumption and increase the training speed of BERT. And also, they use a self-supervised loss that focuses on modeling inter-sentence coherence and shows it consistently helps downstream tasks with multi-sentence. AGHAJANYAN, Armen, et al. [9] showed that pre-finetuning consistently improves performance for pre-trained discriminators(e.g. RoBERTa) and generation models(e.g. BART) on a wide range of tasks while also significantly improving sample efficiency during fine-tuning. They also show that they can effectively learn more robust representations through multi-task learning (MTL) at scale.

There are other researches to overperform previous research in other ways. XLNet from YANG, Zhilin, et al.[10] tried to generalize the autoregressive pre-training method that enables learning bidirectional context and overcomes the limitation of BERT from its autoregressive formula. Instead of using BERT directly, it could achieve substantial improvement over previous pretraining objectives on various tasks. Another method called ELECTRA from CLARK, K. Luong, et al. [11] tried to reduce the amount of computation with a more sample-efficient pre-training task called replaced token detection. This approach performs comparably to RoBERTa and XLNet while using less than 1/4 of their compute and outperforms them when using the same amount of computing. There is

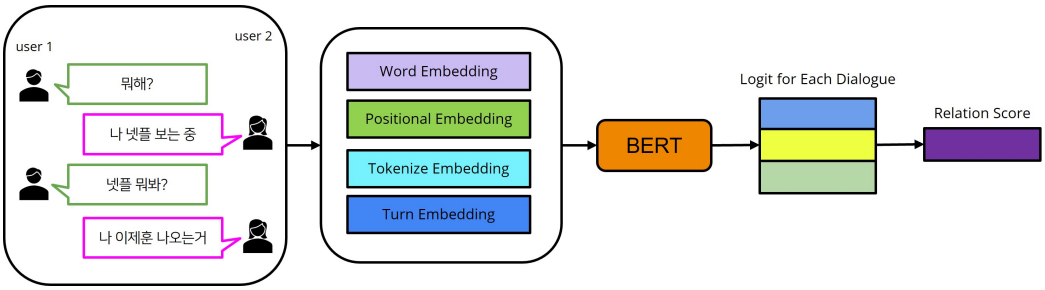

Figure 1: The overview of our human-level Semantic Score Prediction (SSP) model

another model called T5 from RAFFEL, Colin, et al. [12]. This model is first pre-trained on a data-rich task before being fine-tuned on downstream tasks. Unlike the BERT and GPT series, it has both encoder and decoder to deal with NLP tasks. They explore the landscape of transfer learning techniques for NLP by introducing a unified framework and it overperformed all the pre-trained models significantly in most of the downstream tasks.

Research about semantic analysis is keep going on these days. But in our knowledge, most of the semantic tasks and models based on language uses only a single sentence, and even in the dataset with multi-speaker, no speaker information is given to the model. In our research, we try to provide information about who the speaker of the sentence is and let the model train the semantics of each speaker. We hypothesize that it will lead the model could predict the relationship between the speakers.

2.3 Dialogue State Tracking

State tracking in dialogue is to monitor the state of the conversation from dialogue history. It is important in generating dialogue because the state of the conversation is one of the key factor for the speakers to select the words and composition of the next sentence. The simple BERT model is shown to be effective in this state tracking by LAI, Tuan Manh, et al. [13]. This work found that the number of parameters does not grow with the ontology size and the model can operate in situations where the domain ontology may change dynamically. This model also could be compressed with the knowledge distillation method. ZHAO, Jeffrey, et al. [14] showed that the choice of pre-training objective makes a significant difference to the state tracking quality. This paper found that masked span prediction is more effective than auto-regressive language modeling. The authors found that pre-training for the seemingly distant summarization task works surprisingly well for dialogue state tracking and context representation works well while recurrent state. FENG, Yue et al. [15] proposes a new approach to dialogue state tracking, referred to as Seq2Seq-DU, which formalizes dialogue state tracking as a sequence-to-sequence problem. It employs two BERT-based encoders to respectively encode the utterances in the dialogue and the descriptions of schemes. It could leverage the rich representations of utterances and schemas based on BERT.

Even though there are several models to deal with state tracking in dialogue, most of them are based on how to use and modify BERT based model efficiently or how it affects the results. But these works also do not consider the speakers of the dialogue. Based on these, we try to add turn-based embedding before adapting it into the model and make the model could refer to the human-level conversation.

3 Semantic Score Prediction(SSP) model

Figure 1 is an overview of the model we proposed. In contrast to the basic semantic analysis models, Semantic Score Prediction (SSP) model tries to deal with the semantic analysis tasks as regression problems, not the classification problem. From pre-trained BERT, it is fine-tuned with the dialogue from multiple speakers. As the model can know which speaker said the input sentence, it can learn the context of the conversation and predict the relationship between the speakers.

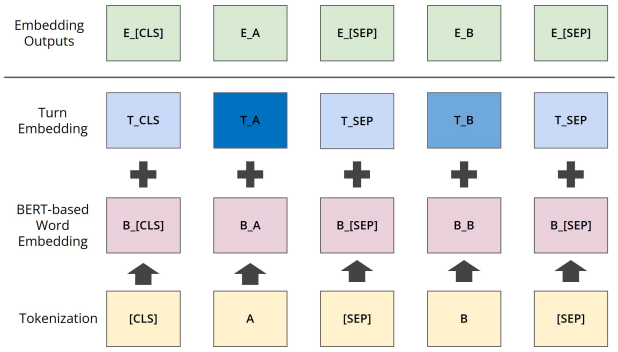

Figure 2: The process of word embeddings

3.1 Word Embeddings

In the embedding section, we used BERT based multilingual cased pre-trained tokenizer consisting of 104 languages with the largest Wikipedia using a masked language modeling(MLM) objective to tokenize the dialogue made in Korean. We also used BERT-based word embedding to embed the tokenized words as you can see in Figure 2. This base embedding consists of word embedding, positional embedding, and tokenize embedding. To make the model recognize the speaker of the sentence, we added turn embedding on the base BERT embedding. We set the special tokens like [CLS] and [SEP] as 0, the first speaker started the conversation as 1, and the second speaker who is going to talk with the first speaker as 2. This turn embedding is commonly used in the chatbot system but to our knowledge, there was no prior research that adapt it to the semantic analysis. By making the model can recognize the subject of the sentences, it can figure out the semantic differences of each speaker and further can predict the relationship between them.

3.2 Model Architecture

Figure 3 is the main structure of our SSP model. It consists of BERT layer, some fully connected layers, and a scoring matrix. We bring the base pre-trained BERT model for our baseline and modify it to have an ability to deal with turn-based knowledge. And then, we fine-tuned it with many dialogue data which know the concept of turn and information about semantics between love and breakup. From the fine-tuned BERT, we can get the logits of the breakup, daily conversation, and love each. The meaning of the logits are the scale of each characteristic. These results pass through two fully connected layers so that we can get the yardstick of break-up conversation, daily conversation, and love conversation between two speakers. Finally, we defined the scoring matrix to predict the relation score. The scoring matrix is defined as:

$$P = (-1 * \mathbb{S}_{\mathbb{B}}) + \mathbb{S}_{\mathbb{D}} + \mathbb{S}_{\mathbb{L}} \tag{1}$$

where P represents the predicted relation score and $\mathbb{S}_{\mathbb{B}}$, $\mathbb{S}_{\mathbb{D}}$ and $\mathbb{S}_{\mathbb{L}}$ represent the yardstick of break-up conversation, daily conversation, and love conversation each. Because the breakup conversation and love conversation have an opposite characteristic, we can say that minus yardstick of breakup can serve as a basis for the love. From these processes, we can get the relation score between two speakers.

4 Experiments

4.1 Dataset and Preprocessing

Preprocessing In the case of the BERT classification model, each sentence is recognized by adding [CLS] token to the front, and the end of the sentence is recognized by adding [SEP] token. By recognizing [CLS] token, it is possible to know that it is the beginning of a sentence, and by recognizing [SEP] token, it is possible to know the end of the sentence. we put [CLS] in front of the dialogue and [SEP] not only in the end of the sentence but also when speaker changes.

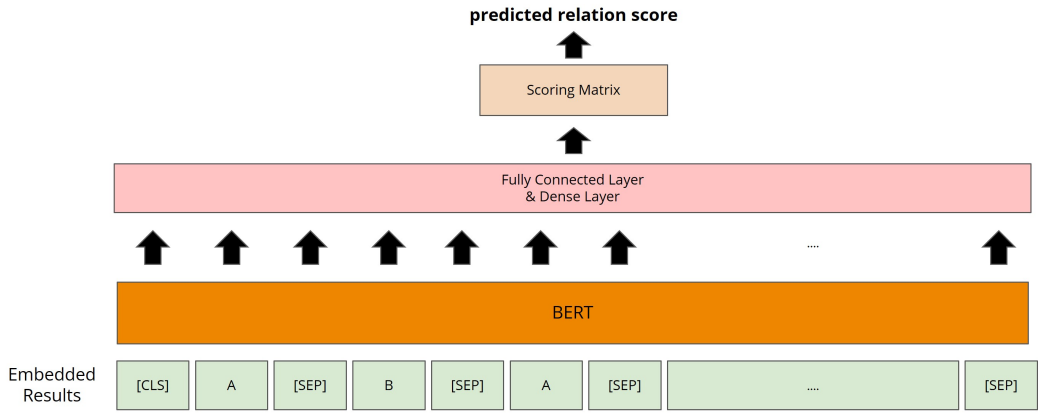

Figure 3: Architecture of Semantic Score Prediction model(SSP)

Synthetic dataset We used chatbot dataset, it is synthetic data and there are 11,876 pairs of questions and answers. It was produced by referring to the stories frequently found in Daum Cafe's "Better than Love" in some breakup-related questions. In original data, daily conversations are labeled 0, break-up conversations are labeled 1, and love conversations are labeled 2. We re-labeled the break-up dialogue as -1, the daily dialogue as 0, and the love dialogue as 1.

Real-world dataset We used real-world messenger dataset. Everyone's corpus is data released by the National Institute of Korean Language. There are 47,421 training examples. This data is not labeled, so we labeled 5,000 dialogues personally. In addition to the each speaker's utterance, this data includes various information such as the speaker's gender and relationship. Based on this information, we labeled some data in rule-based.

4.2 Evaluation method

Confusion matrix contains all the details of how precise the model is when evaluating the classification model, how practical the classification is, and how accurate the classification is. We can use this matrix to calculate accuracy, precision, recall and F1-score. We used this four metrics. Additionally, we used Root Mean Square Error(RMSE) because is no evaluation method to evaluate the relation score.

4.3 Experimental details

We divided the synthetic chatbot dataset into train, validation, and test at a ratio of 6:2:2, and real-world messenger dataset at a ratio of 8:1:1. We used AdamW optimizer from LOSHCHILOV, Ilya; HUTTER, Frank.[16] with a learning rate $\mathbf{2^{e-5}}$ and eps $\mathbf{1^{e-8}}$. Eps is a term added to the denominator to improve numerical stability. During each train process, epoch was set to 30. We use BERT as our baseline model.

4.4 Results

As you can see in Table 1, our SSP model outperforms the baseline in the semantic analysis task with dialogue system. Even though our purpose is to predict the relation score, we used argmax to the output to calculate the accuracy because it is hard to label the dialogue data as score. But even we see this task as classification problem, the performance outperforms the baseline in most of metrics.

The actual conversation and its integer encoding result, the maximum value of output(logits) and relation score between people can be seen in Figure 4.

Table 1: Result

	Accuracy	Precision	Recall	F1 Score	RMSE
Baseline	79.38%	**57.56%**	59.61%	58.26%	0.45
SSP	**83.51%**	57.52%	**67.86%**	**58.99%**	**0.41**

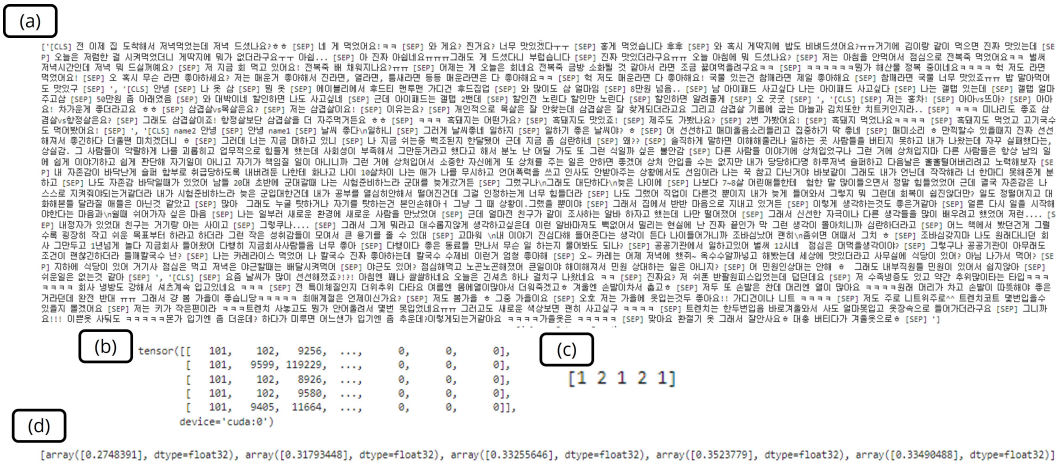

Figure 4: (a): Some examples of dialogue from real-world messenger dataset. (b): Integer encoding result of dialogues. (c): Result of using argmax of the output(logits). (d): Realtion score extracted from dialogues.

5 Analysis

There must be several speakers in a conversation who have different thoughts, different feelings, and different minds. Because of this, models should try to calculate the representations of speakers independently based on their characteristics. Our SSP model successfully learns the sentences in the dialogue separately according to speakers. The main difference between the baseline model and SSP model is the existence of turn embedding and the number of fully connected layers. Turn embedding let the model know which speaker is saying so that it can learn the feature of each speakers separately. This can lead the model to predict the relationship between them more precisely and can have better performance in the semantic analysis tasks.

Previous research for semantic analysis tried to classify the semantics by analyzing the sentences one by one. Some researchers tried to deal with multiple sentences at once, but they also do not consider the subject of the sentence. But to predict the semantics and relationship between two speakers, the semantics results from two speakers must be calculated separately and mixed properly because of the semantics of each speaker matter rather than the number or strength of positive sentences. To achieve this, we add two fully connected layers which get the results of each speaker output and get the relation score as an output. The lower the RMSE score is, the better performance is in relation score. Even though the relation score is not labeled in this experiment, it is enough to let the discrete labels as ground truth because even if the labeled values seem extreme, it is the right value if the score is close to the labeled result.

6 Conclusion

We propose human-level Semantic Score Prediction (SSP) model that learns turn embedding which contains information about which speaker is speaking during conversation. Also, semantic analysis is conducted based on dialogue. Specifically, we extracted relation score between two people based on a dialogue. Through turn embedding method, our model outperforms existing baseline model when evaluated on synthetic dataset and real-world messenger dataset.

However there are some limitations in our work. First one is quality of data. Because of the ambiguity in natural language, labeling is not trustworthy. In conversation, there are many cases where you don't follow the grammar perfectly. So when tokenizing dialouges, vocab size can be significantly large. In this case, word embedding is not efficient. Second limitation is data imbalance. Most of dialogues from real-world dataset are labeled as 0. It means, daily conversation. Because of this imbalance, the model cannot be trained properly. Third limitation is that there is no prior researches. In our knowledge, it is the first work to try to predict the semantics in human-level using dialogue so that there is no evaluation metrics for the relation score and hard to compare with other models.

In future work, We could try to use multimodal model so that the model not only learns the language but also emotion and accent. It will predict much better. We can also use pre-trained knowledge based model for this manner.

7 Contribution of each team member

JaeYeon Bae Investigate the prior works, Define evaluation metrics, Model Design

Chaeri Kim Data collection, Data pre-processing, Data analysis, Model Implementation

Collaboration Baseline model analysis & tuning, Debugging, Result analysis & Model development, Code optimization

References

[1] Zou, Yicheng, et al. "Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 16. 2021.

[2] ROBINSON, Tony; HOCHBERG, Mike; RENALS, Steve. The use of recurrent neural networks in continuous speech recognition. In: Automatic speech and speaker recognition. Springer, Boston, MA, 1996. p. 233-258.

[3] GRAVES, Alex. Long short-term memory. Supervised sequence labelling with recurrent neural networks, 2012, 37-45.

[4] SCHUSTER, Mike; PALIWAL, Kuldip K. Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 1997, 45.11: 2673-2681.

[5] BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[6] VASWANI, Ashish, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

[7] DEVLIN, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[8] LAN, Zhenzhong, et al. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942, 2019.

[9] AGHAJANYAN, Armen, et al. Muppet: Massive multi-task representations with pre-finetuning. arXiv preprint arXiv:2101.11038, 2021.

[10] YANG, Zhilin, et al. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 2019, 32.

[11] CLARK, K. Luong, et al. Pre-training Text Encoders as Discriminators Rather Than Generators. Preprint at https://arxiv. org/abs/2003.10555, 2020.

[12] RAFFEL, Colin, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 2020, 21.140: 1-67.

[13] LAI, Tuan Manh, et al. A simple but effective bert model for dialog state tracking on resource-limited systems. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020. p. 8034-8038.

[14] ZHAO, Jeffrey, et al. Effective sequence-to-sequence dialogue state tracking. arXiv preprint arXiv:2108.13990, 2021.

[15] FENG, Yue; WANG, Yang; LI, Hang. A Sequence-to-Sequence Approach to Dialogue State Tracking. arXiv preprint arXiv:2011.09553, 2020.

[16] LOSHCHILOV, Ilya; HUTTER, Frank. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.