Short-text intention recognition based on multi-dimensional dynamic word vectors

Mei Xiao¹,a, Miao Yao²*, Ying Li¹

¹Tourism and Cultural management college, Sichuan Tourism University, Chengdu, Sichuan Province, 610000, China
²Department of Software engineering, Xinjiang University, Urumqi, Xinjiang Province, 830000, China
a0000965@sctu.edu.cn
*Corresponding author’s e-mail: 845308358@qq.com

Abstract: For short-texts with few words and sparse features, it becomes very important to use text context semantic information to enrich its text representation. At present, the text representation method of static word vectors is the same for the same word in different sentences. From a certain perspective, contextual semantic information is not fully utilized. Therefore, this paper proposes to use the combination of BERT and ELMo to mine more comprehensive contextual semantic information from different dimensions to enrich the text representation. The BERT model fully describes the characteristics of word level, sentence level and relationship between sentences. The ELMo model can solve the problem of polysemy. LSTM and Attention Mechanism are used to extract high-dimensional text features, and CRF is used to recognize intentions. Experiments show that the multi-dimensional dynamic word vector intention recognition model (MD-Intention) proposed in this paper has achieved good performance on both the ATIS-2 dataset and the SNIPS dataset.

1. Introduction
With the rapid development of information technology and the widespread of the Internet, the number of Internet users in various countries is growing rapidly, and the network is gradually integrated into people's daily life in all directions, which makes it an inevitable trend of social development for the whole people to access the Internet. At the same time, with the popularity of social media software, people have a better way to express their feelings and intentions. Because more and more people begin to use social software to express their intention, such as buying and selling houses, traveling, entertainment recommendation, mining the deep semantic information of social text information, analyzing the real scene of the text, and identifying the user's intention have a very positive effect on the network public opinion and accurate marketing.

1.1. Related Work
At present, scholars define user intention recognition tasks as three kinds: sequence segmentation task, classification task or clustering task. Luong T et al. regard intention extraction task as sequence segmentation problem, put forward to combine Bi-LSTM and CRF model, use Bi-LSTM to integrate long-term dependent information, use CRFs to segment sequence data, use real estate, business and beauty in social media posts in Vietnam to identify online user intention in specific fields[1]. Vaswani
A et al. put forward a Transformer model which is composed of Attention Mechanism, and verified the effectiveness of Attention Mechanism in Machine Translation tasks[2]. Bing L et al. associate intention recognition with slot filling task, established RNN model based on Attention Mechanism, which improved the accuracy of intention recognition in oral comprehension[3]. Zhu S et al. propose a novel coding and decoding framework of focus mechanism to solve the problem of sequence tagging, in which input and output sequences are placed according to words followed by words, while attention mechanism can not provide accurate placement [4]. Goo C et al. propose a recurrent neural network (RNN) model based on Attention Mechanism, which uses the correlation between slot filling and intention recognition tasks to improve the performance of the two tasks by continuously learning the correlation[5].

Agarwal S et al. design a cascaded ensemble learning classifier to recognize racist and hateful radical intentions in posts by mining training models of semantic, emotional and linguistic features in texts[6]. Sakar C et al. say that multi-layer perceptron and Long Short-Term Memory network are combined to identify real-time online users' purchase intention[7]. Agarwal S et al. propose to classify the motivation and intention of users' comments based on user characteristics: document emotion, emotion, writing characteristics, social mood and semantic characteristics[8]. Dulan D S et al. propose to use text semantic analysis to identify racial comments, and build 2-class SVM based on Microsoft Azure Machine Learning Studio Platform to classify short text comments[9]. Hemant P et al. make use of mixed text features, including declarative features, social behavior features and contrast pattern features, to enrich text representation and improve the accuracy of intention classification[10]. Devlin J et al. establish the Bert model and achieve the best performance in 11 natural language understanding tasks[11]. Qian C et al. use the Bert pre-training model to train text representation, get rich semantic information, and achieve the purpose of improving the accuracy of intention classification[12]. Xia C et al. use the Capsule Neural Network to carry out migration learning, use the trained user intention classifier and the relationship between the new and old intention to distinguish the new user intention[13].

Cai R et al. use CNN-LSTM model to capture global semantic expression and local phrase level information from original medical text questions, and use part of speech tagging and nomenclature recognition technology to enrich semantic feature information. First, the user intention classification method is established by DBSCAN clustering, and then the user intention is identified from the real medical text request[14]. Jayarathna S et al. use LDA theme model and user comments to build user interest profile. Through the calculation of probability distribution similarity, it can measure the similarity between the user generated comments and the document generated comments, so as to mine the user interest[15]. White R W et al. use comprehensive user information to predict users' short-term interest intention based on user behavior and open catalog topic classification[16]. Wang J et al. propose a semi-supervised method to identify intention, manually annotated a small number of posts with intention, marked other posts by calculating similarity between texts, and finally judged intention category by confidence score[17].

It is very important to effectively use the context semantic information to improve the performance of natural language processing tasks. Although static word vector can capture the context semantic information, the expression of the same word in different sentences is the same, which shows that this way does not effectively use the context semantic information. Ethayarah K et al. prove that in the high-dimensional representation of Bert model, the similarity between words in the same sentence is lower and in the high-dimensional representation of Elmo model, the similarity between words in the same sentence is higher[18]. It shows that Bert and Elmo focus on different levels of semantic information. This paper combines these two models to truly capture the semantic information of context and get more rich text representation.
2. Material and Methods

The input of the model is a document $D$, which is a $W_1, W_2, \ldots, W_n$ sequence of $n$ words. Input documents $D$ as the input of Bert model and Elmo model respectively train the word vectors. The word vectors of the two models are spliced by concat function to get the final word vector representation $e_1, e_2, \ldots, e_n$. Using the combination of LSTM and Attention Mechanism to extract high-dimensional text features, and finally using CRF for intention recognition.

2.1. Multi-dimensional dynamic word vector representation

Bidirectional Encoder Representation from Transformers (Bert) model further increases the generalization ability of word vector model, and fully describes the features of word level, sentence level and even sentence relationship. Bert makes use of transformer to make a real bidirectional encoder, which has deeper layers and better parallelism, so as to realize context dependence[11].

Embedding from Language Models (Elmo) is a context independent character level language model proposed by Allen NLP[19]. It uses Bi-LSTM to obtain deep context word representation. It has the ability to handle the complex features in word usage, including syntax and semantics, and the dynamic word vector, including word diversity, which can change according to the difference between the above and the following.

Because the training corpus of this paper is English, we do not choose the Chinese or multilingual Bert Large pre-training model. According to the actual application, this chapter selects the pre-training model of Bert Base-uncased, which has 12 layers, 768 hidden layers, 12 heads in the multi-head Attention Mechanism and 110M parameter size.

In the Bert model and Elmo model, the representation vectors of context scene words are very different. In the Elmo model, in the high-dimensional space, the context uniqueness increases, and the word representation vector in the same sentence are more similar, but they are still more similar on average than the randomly selected words.

For an input sentence $x=(x_1, x_2, \ldots, x_s)$, through the two results of Bert model and Elmo output...
respectively, the concat function is used to splice the word vectors obtained by Bert model and Elmo model, and the final word vector representation \( e_e, \cdot \cdot \cdot e_e \) is obtained.

### 2.2. Feature representation

After the pre-training of Bert model and Elmo model, the final word vector representation is obtained, which is used as the input of LSTM model. LSTM is suitable for processing and predicting important events with relatively long interval and delay in time series. It can delete or add information to the state of neural unit through gating mechanism to capture the long-term dependence of intention text. Attention Mechanism selects the more critical information of current intention recognition target from a lot of information, which has the advantage of capturing local key information. Therefore, this chapter uses the advantages of both to express and extract the features of the text.

LSTM recurrent neural network is composed of three parts: input gate with corresponding weight matrix \( W_i, W_i, W_i, b_i \), forget gate with corresponding weight matrix \( W_f, W_f, W_f, b_f \), and output gate with corresponding weight matrix \( W_o, W_o, W_o, b_o \). All gate control mechanisms use the current input \( x_i \), the state generated by the previous layer \( h_{t-1} \), and the activation value of the current unit \( c_{t-1} \) to decide whether to accept the input, whether to forget the previously stored memory, and whether to output the last state. The following is the equations for updating each gate control mechanism and unit status:

\[
i_i = \sigma(W_i x_i + W_i h_{t-1} + W_i c_{t-1} + b_i) \\
f_i = \sigma(W_f x_i + W_f h_{t-1} + W_f c_{t-1} + b_f) \\
g_i = \tanh(W_i x_i + W_i h_{t-1} + W_i c_{t-1} + b_i) \\
c_t = i_t g_t + f_t c_{t-1} \\
o_t = \sigma(W_o x_i + W_o h_{t-1} + W_o c_{t-1} + b_o) \\
h_t = o_t \tanh(c_t)
\]

In 2017, Vaswani, the Google machine translation team proposed the necessity of Attention Mechanism in natural language processing[2]. The input of Attention Mechanism \( H \) is a matrix \([h_1, h_2, \cdots, h_n]\), where \( n \) is the sentence length. The expression \( r \) of the sentence is composed of the weight of the output vector. The specific equations is as follows:

\[
M = \tanh(H) \\
\alpha = \text{softmax}(W^\alpha M) \\
r = H^\top \alpha \\
h^* = \tanh(r)
\]

Where \( M \in \mathbb{R}^{d_w \times d_w}, \alpha \in \mathbb{R}^{d_w \times d_w}, d_w \), is the dimension of word vector, \( w^\alpha \) is the parameter vector obtained by training, and \( W^\alpha \) is the transposition of parameter vector. The dimensions of are \( W, \alpha, r \). A tanh activation function is used to get the final sentence representation \( h^* \) by Equation (10).

The multi-dimensional dynamic word vector, that is, the final text representation, \( e_e, e_e, \cdot \cdot \cdot e_e \) is used as the input of the LSTM model, and the results are calculated by Equations (1) to (6). The results are used as the input of the attention mechanism, and the text feature representation is obtained by Equations (7) to (10).

### 2.3. Intention Recognition

Conditional Random Field(CRF ) is the Markov random field of random variables \( Y \) under given conditions \( X \). In this chapter, CRF is used for intention recognition, and conditional probability \( P(Y|X) \). The observation sequence composed of each word in the sentence as the observation object \( X \), \( Y \) is the output.
3. Results

3.1. Data sets

**SNIPS data set**: In snips data set, there are 14484 sentences, which are divided into training set of 13084 sentences, verification set of 700 sentences and test set of 700 sentences. Snips data set includes 7 intentions: addtoplaylist, bookrestaurant, getweather, playmusic, ratebook, searchcreativework and searchscreeningevent.

**ATIS-2 data set**: ATIS (airline travel information system) contains the recordings of the flight reservation personnel. In atis-2 data set, there are 5871 sentences, which are divided into 4478 sentences in training set, 500 sentences in verification set and 893 sentences in test set. Atis-2 data set includes 18 kinds of intentions: abbreviation, aircraft, airport, city, cheapest, flight, day_name, mean, etc.

3.2. Model parameters

| Parameters     | Value |
|----------------|-------|
| BERT_layer     | 1     |
| ELMo_layer     | 1     |
| Word_embedding | 100   |
| Learning_rate  | 0.001 |
| Num_size       | 100   |
| Batch_size     | 64    |

Bert_Layer represents the number of layers of the Bert model, ELMo_Layer represents the number of layers of the Elmo model, Word Embedding represents the dimension of word vector, Learning_Rate represents the learning rate of the model, Num_Size represents the number of LSTM hidden layer nodes, and Batch_Size represents the batch size.

3.2.1. The dimensions of Word_embedding

| Word_embedding | ATIS-2 (F1 value) | SNIPS (F1 value) |
|----------------|-------------------|-----------------|
| 50             | 96.19%            | 95.88%          |
| 100            | **98.93%**        | **98.86%**      |
| 150            | 97.37%            | 97.48%          |
| 200            | 97.39%            | 97.54%          |

Different dimensions of word embedding contain different semantic information. The paper makes a comparative experiment on the word vector dimension settings to select the optimal word vector dimension. In ATIS-2 data set, it can be seen that when the word vector dimension is 100, the F1 value of MD-Intention model is 98.93%. In SNIPS data set, when the word vector dimension is 100, the highest F1 value of MD-Intention model is 98.86%.

3.2.2. The value of Learning_rate

| Learning_rate | ATIS-2 (F1 value) | SNIPS (F1 value) |
|---------------|-------------------|-----------------|
| 0.1           | 91.67%            | 90.87%          |
| 0.03          | 97.91%            | 95.35%          |
| 0.01          | 97.33%            | 97.01%          |
| 0.003         | 98.21%            | 98.17%          |
| 0.001         | **98.93%**        | **98.86%**      |
The learning rate controls the learning progress of the model, which directly affects the optimal effect of model training. According to the different value of learning rate, the convergence effect of the model to achieve the optimal performance is different. The comparison experiment is carried out on the commonly used learning rate value. From the experimental results, it can be seen that when the learning rate is 0.001, on the ATIS-2 and SNIPS data sets, the F1 value of the MD-Intention model is the highest, which is 1.6% and 1.85% higher than that when the learning rate is 0.01, respectively.

4. Discussion

4.1. Verify the validity of BERT + ELMo representation word vector

By verifying the validity of the word vector of BERT + ELMo combination, this paper compares it with the following three combination models on ATIS-2 data set and SNIPS data set. The comparison model is numbered are shown in Table 4. In the first group of experiments, the Bert and Elmo models in the MD-Intention model were replaced by static word embedding to represent word vector, which verified the validity of the combination of Bert model and Elmo model to represent multi-dimensional dynamic word vector. In the second group of experiments, the Bert model in MD-Intention model was removed to verify the validity of the Bert model. In the third group of experiments, Elmo model in MD-Intention model was removed to verify the effectiveness of Elmo model. On the ATIS-2 data set and SNIPS data set, use the F1 value under the micro-average as the evaluation index of the experiment, and carry out the comparative experiment. The experimental results are shown in Figure 2 and Figure 3 below.

| Serial number | Model                                      |
|---------------|--------------------------------------------|
| 1             | LSTM+ATTENTION+CRF                         |
| 2             | ELMo+LSTM+ATTENTION+CRF                    |
| 3             | BERT+LSTM+ATTENTION+CRF                    |
| 4             | BERT+ELMo+ATTENTION+CRF (MD-Intention)     |

![Figure 2. The results on ATIS-2.](image1)

![Figure 3. The results on SNIPS.](image2)

Experimental analysis:

In ATIS-2 data set and SNIPS data set, the F1 value of model 3 pre-trained by Bert model is 98.30% and 98.14% respectively, which is 4% and 3.97% higher than that of model 1 directly embedded by static words. Because of the Bert model, the word vector representation includes three parts: static word vector representation, sentence word relation word vector representation and word position word vector representation, which greatly captures the context information of words. The F1 value of model 2 is 98.21% and 97.47% respectively, which is improved by 3.91% and 3.3% compared with model 1. Because [19] proves that the first layer LSTM model in Elmo model can learn more syntactic information and the second layer LSTM model solves the polysemy problem well, that is, the more able to capture the meaning information of words. In model 1, the use of a single static...
word vector representation only considers the compactness in words, and fails to capture the context information. The F1 value of MD-Intention model reached 98.93% and 98.86% respectively, which was 4.63% and 4.96% higher than that of the first group. In a word, in ATIS-2 data set and SNIPS data set, although the meaning of a word can be inferred from the context, the expression of the same word in different contexts is always the same, which fully reflects that the static word vector fails to effectively capture the context information. And Bert model and Elmo model, from the different dimension, complete the enrichment of intention text word vector, get a more complete expression of intention text word vector.

4.2. Verify the efficiency of Attention Mechanism in capturing key local information

CNN and the Maximum Pooling layer can capture the local key information. By comparing with them, we can verify the effectiveness of the Attention Mechanism in capturing the local key information. The comparison model is numbered are shown in Table 5. In ATIS-2 data set and SNIPS data set, F1 value under micro-average is used as evaluation index to carry out comparative experiment, and the experimental results are shown in Figure 4 and Figure 5.

| Serial number | Model                                      |
|---------------|--------------------------------------------|
| 1             | BERT+ELMo+LSTM+CNN+CRF                     |
| 2             | BERT+ELMo+LSTM+MAXPOOLING+CRF              |
| 3             | BERT+ELMo+LSTM+CRF                        |
| 4             | BERT+ELMo+LSTM+ATTENTION+CRF (MD-Intention)|

![Figure 4. The results on ATIS-2.](image)

![Figure 5. The results on SNIPS.](image)

Experimental analysis:

In ATIS-2 data set and SNIPS data set, compared with model 1, the F1 value of MD-Intention model is increased by 1.19% and 1.76%, respectively. Because CNN model uses the idea of local convolution, although it can capture the local characteristics of text sequence, it lacks consideration of global context information. Compared with model 2, the F1 value of MD-Intention model is increased by 2.38% and 2.93%, respectively. Because using a maximum pooling layer to find the most critical semantic information in the intention text has limitations and does not consider increasing the attention value of key information. The Attention Mechanism can not only express the text well, but also increase the attention value of key information. Compared with model 3, the F1 value of MD-Intention model is increased by 3.09% and 4.62%, respectively, which proves that the rich intention text word vector representation and the capture of partial key information of text are very important for the F1 value of intention recognition.

5. Conclusion

Aiming at the problem that the static word vector does not fully mine the context scene information
and make use of the local key information in the text, this paper uses the combination of the dynamic word vector Bert pre-training model and the Elmo pre-training model to solve the problem that the static word vector cannot recognize the polysemy of a word and does not make full use of the context scene information, and to use the Attention Mechanism to capture the local key information. In ATIS-2 data set and SNIPS data set, experiments show that MD-Intention model is superior to other contrast models in performance.

Acknowledgments
The authors wish to acknowledge assistance or encouragement from colleagues, and thank Goo et al. and Coucke et al. for their help in data acquisition.

References
[1] Luong T, Cao M, Le D, et al. Intent extraction from social media texts using sequential segmentation and deep learning models[C]. knowledge and systems engineering, 2017: 215-220.
[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. neural information processing systems, 2017: 6000-6010.
[3] Liu B, Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling[J]. arXiv: Computation and Language, 2016.
[4] Zhu S, Yu K. Encoder-decoder with Focus-mechanism for sequence labelling based spoken language understanding[J]. arXiv: Computation and Language, 2016.
[5] Goo C, Gao G, Hsu Y, et al. Slot-gated modeling for joint slot filling and intent prediction[C]. north american chapter of the association for computational linguistics, 2018: 753-757.
[6] Agarwal S, Sureka A. But I did not mean it!—intent classification of racist posts on Tumblr[C]. european intelligence and security informatics conference, 2016: 124-127.
[7] Sakar C O, Polat S O, Katircioglu M, et al. Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks[J]. Neural Computing and Applications, 2019, 31(10): 6893-6908.
[8] Agarwal S, Sureka A. Role of Author Personality Traits for Identifying Intent Based Racist Posts[C]. european intelligence and security informatics conference, 2016: 197-197.
[9] Dias D S, Welikala M D, Dias N G J. Identifying racist social media comments in Sinhala language using text analytics models with machine learning[C]. International conference on advances in ICT for emerging regions, 2018.
[10] Purohit H, Dong G, Shalin V L, et al. Intent classification of short-text on social media[C]. IEEE International Conference on smart city socialcom sustaincom, 2015: 222-228.
[11] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: Computation and Language, 2018.
[12] Chen Q, Zhuo Z, Wang W. BERT for joint intent classification and slot filling[J]. arXiv: Computation and Language, 2019.
[13] Xia C, Zhang C, Yan X, et al. Zero-shot user intent detection via Capsule Neural Networks[J]. arXiv: Computation and Language, 2018.
[14] Cai R, Zhu B, Lei J, et al. An CNN-LSTM attention approach to understanding user query intent from online health communities[C]. 2017 IEEE International Conference on Data Mining Workshops (ICDMW), 2017.
[15] Jayarathna S, Patra A, Shipman F M. Mining user interest from search tasks and annotations[C]. conference on information and knowledge management, 2013: 1849-1852.
[16] White R W, Bennett P N, Dumas S T. Predicting short-term interests using activity-based search context[C]. conference on information and knowledge management, 2010: 1009-1018.
[17] Wang J, Cong G, Zhao W X, et al. Mining user intents in twitter: a semi-supervised approach to inferring intent categories for tweets[C]. national conference on artificial intelligence, 2015: 318-324.
[18] Ethayarajh K. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 Embeddings[J]. arXiv: Computation and Language, 2019.
[19] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]. North American chapter of the association for computational linguistics, 2018: 2227-2237.
[20] Coucke A , Saade A , Ball A , et al. Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces[J]. 2018.