BERT-CRF: Chinese Spoken Language Understanding

Jiude Wei*, **Letian Yang***
{wjd_kznwl, moekid101}@sjtu.edu.cn
(*equal contribution)

Abstract

Spoken language understanding (SLU) is a crucial area of research for natural language comprehension and generation. The emergence of large language models (LLM) has raised its accuracy to another peak, aiming to achieve human-like ability of language understanding. BERT [2], one of the most popular LLM, has wonderful performance across spoken language understanding area. Considering the availability and performance, we introduced BERT into our Chinese spoken language understanding framework, as well as introduced modules for activating conversation history, reducing the negative inference from noises and modeling unseen slot values. Through comprehensive experiment on the given Chinese SLU dataset, we observed great performance of our model.

1 Introduction

Spoken Language understanding is a fundamental but challenging task in natural language processing with wide-spread application, including dialogue system, automatic summarization and multi-language translation, etc. BERT [2] has been one of the most successful LLM since its born, and its feature is publicly recognizable for natural language understanding tasks. Among its downstream applications, bert-base-chinese has the advantage of utilizing complicated Chinese sentence representation. Conditional random field (CRF) has proved itself with ability of sequence labeling. Based on BERT and CRF, we proposed BERT-CRF for Chinese spoken language understanding, specified in semantic triple analysis task. Our **contributions** are: 1) We introduced BERT into the baseline model as encoder, as well as LSTM-CRF as decoder for sequence labeling, which results in a better performance. 2) We introduced multiple mechanics in our model to activate conversation history, reduce negative influence from noise in data, and deal with unseen slot values.

2 Related Works

Spoken language understanding typically consists of two tasks: intent detection and slot filling [8]. Intent detection is a sentence classification problem, hence neural networks based on classification methods have emerged, including CNN [9] and RNN [5]. Slot filling is a sequence labeling task. Typical methods to sequence labeling includes conditional random field (CRF) [6], RNN based model [9] and LSTM [5]. With the popularity of deep-learning and the emergence of pretrained large language models, SLU bas made significant progress.

Chinese spoken language understanding faces difficulties in nature while compared with SLU in English, because word segmentation based on semantics is essential for understanding sentences in Chinese. Take the utterance "*Navigate me to Beijing*" in Chinese (Fig.1) as an example, the output include an intent class label (i.e. *Navigate*) and a slot label sequence (i.e. *O, O, B-(Inform-Action), I-(Inform-Action), O, B-(Inform-Action), I-(Inform-Action)*).

The main challenges of Chinese SLU are concentrated in word information corporation and prior precise word segmentation.

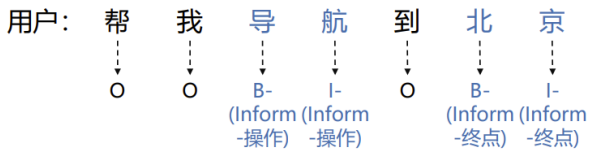

Figure 1: **Utterance of Chinese SLU.** The above row shows input sentence "*Navigate me to Beijing*" in Chinese. The row at bottom show the respective intent detection and slot filling.

In this project, we adopted BERT for word segmentation and intent detection, as well as LSTM-CRF decoder for slot filling task in order to solve Chinese semantic triple analysis problem.

3 Method

3.1 Baseline Model

Embedding. The baseline model embeds words with a certain file that maps words to vectors. Then it extracts features of word vectors by *torch.nn.Embedding* module.

Encoder & Decoder. A typical structure to solve SLU is separating the intent detection mechanism and slot filling mechanism. Here the baseline model proposes a RNN-based encoder to classify the labels of embedded word vectors. The encoder is selected among bidirectional RNN, GRU and LSTM. A linear layer with subsequent softmax is proposed by baseline model as a simple decoder to solve slot filling task and extract the semantic triples from each given input sentence.

Semantic triple decoding. A decoding module is essential for transforming the predicted labels of each word back to human recognizable semantic triples. A label vocabulary is introduced to convert the indexes to word tags, which indicate the role of a word is 1) *B*: the first word of a key phrase. 2) *I*: word in a key phrase despite the first word. 3) *O*: word not in key phrase. The slots and values of semantic triplets are then arranged according to the tags.

3.2 BERT-CRF

The overall pipeline is shown in Fig.2. Here we introduce each stage of our model in detail.

Embedding. In addition to the embedding method given by baseline model, we use the tokenizer of BERT to segment the sentence into words and then map them into vectors.

Specifically, the mapping contains 74 slot labels in original. We map all **unseen slots** into the 75-th label to cope with the embedding of unseen slots. Meanwhile, we dynamically pad all sentence within the same batch into the maximum sentence length among them, to enable training with batched input.

Encoder. We use a pretrained BERT-base-chinese model as the encoder to extract features from the vectors mapped from words. The input ids, token type ids and attention masks given by the tokenizer are taken as input of the BERT encoder.

We introduced a memory pool in encoder module to make use of conversation history, inspired by [1]. Our main idea is to compare the similarity between the present sentence and sentences across conversation history. We impose the historical features on the present feature with balanced weight and a parameterized factor according to the similarities. Specifically, we restore the latest 128 features extracted from the encoder. For each feature f_i, we compute the inner product between the present feature f_p and f_i as similarity. We cast softmax function on the inner product to generate its weight w_i. Then we add each historical feature multiplied with its corresponding weight and a factor λ to the f_p as the feature influenced with conversation history f'_p.

$$
\begin{aligned}
w &= softmax(\langle f, f_p \rangle) \\
f'_p &= f_p + \lambda \sum_i w_i f_i
\end{aligned}
\tag{1}
$$

Decoder. We believe CRF has capacity of sequence labeling, therefore we implemented bidirectional LSTM-CRF as our decoder. Specifically in this task, the features of corresponding words forms the observation sequence, and the label of words to be labeled forms the status sequence. The status can be divide into three modes *B*, *I* and *O* as we mentioned in Sec.2.

Semantic triple decoding. Here we inherited the decoding method of baseline model to our BERT-CRF.

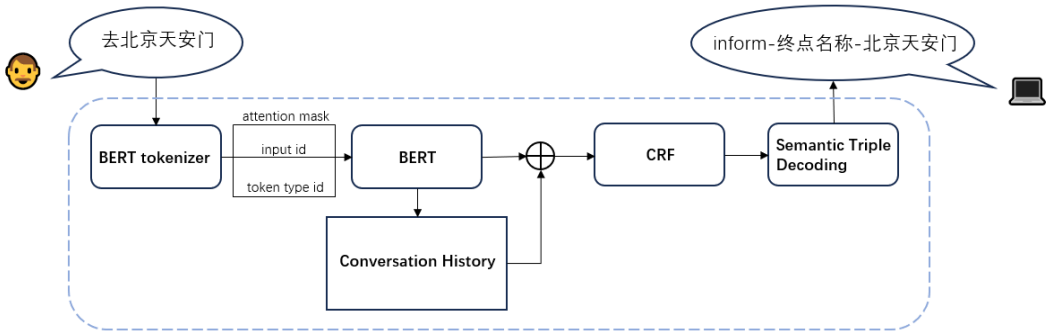

Figure 2: **Overall pipeline of BERT-CRF.** We first embedding the input with BERT tokenizer into vectors and corresponding token ids as well as attention masks. The pretrained BERT model takes them as input and extracts features. The features are sent into a memory pool for activating conversation history in subsequent training round. Current feature is compared with latest 128 historical features and is added by them with weight. The new feature is sent to a CRF decoder for sequence labeling. A semantic triple decoder is introduced to translate the labels of word into semantic triples.

4 Experiments

4.1 Spoken Language Understanding

To thoroughly evaluate the effectiveness of our approach, we compare results of models with cross combination between different encoders and decoders on a unified dataset.

Metrics. We adopt the commonly used metrics for SLU, namely accuracy, precision, recall and f1-score.

Implementation. For all experiments in this section, we use a series of parameters where learning rate is 0.001, batch size is 32, training epoch is 100, dropout rate is 0.2, embedding size is 768, hidden layer size (baseline encoder) is 512, layer number (baseline encoder) is 2, random seed is 999. We use bert-base-chinese as the pretrained model. The memory weighing factor is $\lambda = 0.4$ for BERT-CRF. The sentence labeled *asr1_best* are selected as input sentences.

Main Results. The results are reported in Tab.1. Remarkable performance improvements over the baselines are achieved for our proposals under all metrics, with prominent improvements observed in the more rigorous metric f1-score. The results shows that our BERT-CRF has great performance in Chinese spoken language understanding task. There is a nearly 10% improvement on f1-score against baseline models, which shrines the capacity of spoken language understanding of our BERT-CRF model.

4.2 Ablation Study

Conversation History. We have introduced a memory pool along the BERT encoder in our pipeline as stated in Sec.3.2, where we use a hyperparameter λ to balance the ratio between current feature and historical features. In this section, we will show the result of BERT-CRF model on SLU over different λ, detailed in Fig.3. Notice the result of $\lambda = 0$ is namely the model without memory mechanism. The outperformance of models with memory mechanism over the model without memory mechanism demonstrates the availability of our proposal. We can observe a slight drop over small $\lambda \leq 0.4$, and

Method	Encoder	Decoder	Accuracy(↑)	Precision(↑)	Recall(↑)	F1-score(↑)
baseline	LSTM	softmax	71.0615	81.6492	73.3055	77.2527
	RNN		70.1732	80.2691	72.3670	74.6237
	GRU		71.1732	80.2691	74.6611	77.3636
Impr. Encoder	BERT	softmax	75.6944	83.5052	87.2705	79.7985
Impr. Decoder	LSTM	CRF	72.1788	83.5088	74.4526	78.7211
	RNN		71.7318	77.5244	74.4526	75.9574
	GRU		72.1788	82.9128	75.3910	78.9732
BERT-CRF	BERT	CRF	**81.7130**	**84.2286**	**90.2081**	**87.1158**

Table 1: **Experiment results in detail.** There are four major categories of models across our experiment: 1) Baseline model can be divided to encoders of LSTM, RNN and GRU and a simple linear-softmax decoder. 2) We only replaced the encoder with BERT for improved encoder. 3) We replaced the decoder with LSTM-CRF, constituting model with three baseline encoders for improved decoder. 4) BERT-CRF is the model we proposed, which outperformances other models.

a subsequent bounce of precision, recall and f1-score over $\lambda = 0.5$. The performance falls when λ grows too large, where the current feature could not dominate the new feature.

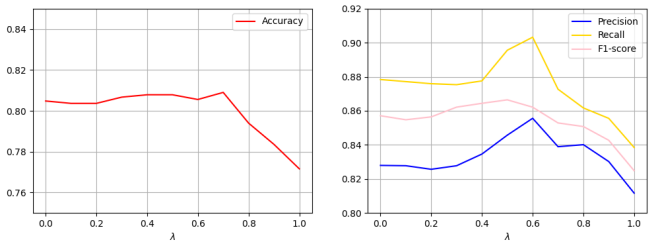

Figure 3: Result of BERT-CRF model with various factor λ on conversation history mechanism. There is slight rise for accuracy rate over λ. A huge bounce for precision and recall is observed with $\lambda = 0.5, 0.6$. The comparison with the result of $\lambda = 0$ shows the availability of the conversation history mechanism.

Noise in data. Based on BERT-CRF model, we would like to ease the negative influence from the noises in *asr1_best* input. We compared the performance with noisy input *asr1_best* and noise-less input *manual_transcript* on BERT-CRF model (See Tab.2 Part 1& 2). We attempted to improve the robustness of our model in order to confront agaist noise. We proposed two simple ways: 1) We introduce dropout layers across the encoder, a typical way to improve robustness. 2) We finetune the pretrained BERT model. Both results are reported in Tab.2 Part 3& 4. However there is no statistic improvement with our proposals based on BERT-CRF.

For the first method of adding dropout layers, we compared the output and features through the neural network between BERT-CRF and the one with dropout. We found that there is no huge difference between the features, which shows the introduce of dropout here is useless. For the second method of finetuning the BERT, we found that the output always includes words outside input sentence, which is translated by the vocabulary from the output vector. This indicates that we need further approaches to tame BERT.

Input	Method	Accuracy(↑)	Precision(↑)	Recall(↑)	F1-score(↑)
noise-less	BERT-CRF	92.9398	94.2857	93.1151	93.6968
noisy	BERT-CRF	81.7130	84.2286	90.2081	87.1158
	B'C+Dropout	81.7870	87.5609	90.3305	87.7601
	B'C+finetune	73.6111	72.0588	83.9657	77.5579

Table 2: **Experiment results for methods against noise.** No obvious performance improvement is observed with our proposals against noises in data. None of our method with noisy input is able to achieve the performance with noise-less input.

5 Conclusion

In this paper, we propose BERT-CRF, a model for solving Chinese spoken language understanding task. We also introduce some mechanics to activate conversation history, reduce the negative inference from noises and model unseen slot values. We discussed the former two approaches in detail. We comprehensively evaluate the effectiveness of BERT-CRF on Chinese spoken language understanding task and the experiments suggest the superiority of our approach.

Despite the great improvement of our model against baseline models on multiple metrics, our model is unable to cope with noises in data. This is a possivle direction where we can improve our model in further way.

References

[1] Yun-Nung Chen, Dilek Hakkani-Tür, Gökhan Tür, Jianfeng Gao, and Li Deng. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Interspeech*, pages 3245–3249, 2016.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Yijin Liu, Fandong Meng, Jinchao Zhang, Jie Zhou, Yufeng Chen, and Jinan Xu. Cm-net: A novel collaborative memory network for spoken language understanding. *arXiv preprint arXiv:1909.06937*, 2019.

[4] Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. A survey on spoken language understanding: Recent advances and new frontiers. *arXiv preprint arXiv:2103.03095*, 2021.

[5] Suman Ravuri and Andreas Stolcke. Recurrent neural network and lstm models for lexical utterance classification. In *Sixteenth annual conference of the international speech communication association*, 2015.

[6] Christian Raymond and Giuseppe Riccardi. Generative and discriminative algorithms for spoken language understanding. In *Interspeech 2007-8th Annual Conference of the International Speech Communication Association*, 2007.

[7] Dechuan Teng, Libo Qin, Wanxiang Che, Sendong Zhao, and Ting Liu. Injecting word information with multi-level word adapter for chinese spoken language understanding. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8188–8192. IEEE, 2021.

[8] Gokhan Tur and Renato De Mori. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.

[9] Puyang Xu and Ruhi Sarikaya. Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 ieee workshop on automatic speech recognition and understanding*, pages 78–83. IEEE, 2013.