ENCODER-DECODER WITH FOCUS-MECHANISM FOR SEQUENCE LABELLING BASED SPOKEN LANGUAGE UNDERSTANDING

Su Zhu and Kai Yu

Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
SpeechLab, Department of Computer Science and Engineering
Brain Science and Technology Research Center
Shanghai Jiao Tong University, Shanghai, China
{paul2204,kai.yu}@sjtu.edu.cn

ABSTRACT

This paper investigates the framework of encoder-decoder with attention for sequence labelling based spoken language understanding. We introduce Bidirectional Long Short Term Memory - Long Short Term Memory networks (BLSTM-LSTM) as the encoder-decoder model to fully utilize the power of deep learning. In the sequence labelling task, the input and output sequences are aligned word by word, while the attention mechanism cannot provide the exact alignment. To address this limitation, we propose a novel *focus mechanism* for encoder-decoder framework. Experiments on the standard ATIS dataset showed that BLSTM-LSTM with focus mechanism defined the new state-of-the-art by outperforming standard BLSTM and attention based encoder-decoder. Further experiments also show that the proposed model is more robust to speech recognition errors.

Index Terms— Spoken language understanding, encoder-decoder, focus-mechanism, robustness.

1. INTRODUCTION

In a spoken dialogue system, the Spoken Language Understanding (SLU) is a key component that parses user utterances into corresponding semantic concepts. The semantic parsing of input utterances in sequence labelling typically consists of three tasks: domain detection, intent determination and slot filling. In this paper, we focus on the sequence labelling based slot filling task which assigns a semantic slot tag for each word in the sentence. The main challenges of SLU are the performance improvement and its robustness to ASR errors.

Slot filling is a main task of SLU to obtain semantic slots and the associated values. Typically, slot filling would be treated as a sequence labelling (SL) problem to predict the slot tag for each word in the utterance. As a typical alignment task, one example of slot filling is illustrated in Figure 1. The goal is to label the word "*Boston*" as the departure city, "*New York*" as the arrival city, and "*today*" as the date.

Sentence	show	flights	from	Boston	to	New	York	today
Slots	O	O	O	B-FromCity	O	B-ToCity	I-ToCity	B-Date

Fig. 1. An example of ATIS sentence and the annotated slots.

Standard approaches to solve this problem include generative models, such as HMM/CFG composite models [1] , hidden vector state (HVS) model [2], and discriminative or conditional models such as conditional random fields (CRFs) [3], and support vector machines (SVMs) [4]. Recently, motivated by a number of very successful continuous-space, neural network and deep learning approaches [5, 6], many neural network architectures have been applied to this task, such as simple recurrent neural networks (RNNs) [7, 8, 9], convolutional neural networks (CNNs) [10], long short-term memory (LSTM) [11] and the variations of different training criterions [12, 13]. The most recent papers use variations on LSTM based sequence models, including encoder-decoder, external memory [14, 15].

Inspired by the success of the attention mechanism [16] in Natural Language Processing (NLP) field, we first applied an attention-based encoder-decoder [17] to treat the sequence labelling based SLU as a language translation problem. In order to consider the previous and the future information, we modelled the encoder with a bidirectional LSTM (BLSTM), and the decoder with an unidirectional LSTM. The attention mechanism takes the weighted average of scores provided by the matches between inputs around position A and output at position B. There are two main limitations of attention model in sequence labelling task:

- Input and output in the sequence labelling are aligned while the attention model scores the overall input words.

This work was supported by the China NSFC project No. 61573241 and the Interdisciplinary Program (14JCZ03) of Shanghai Jiao Tong University in China.

- The alignment could be learned by the attention model, but is difficult to approach with limited annotated data in sequence labelling task (unlike Machine Translation in which paired data is easier obtained).

To address the limitations of the attention mechanism in sequence labelling, we propose the focus mechanism which is emphasizing the aligned encoder's hidden states.

The remainder of the paper is organized as follows. Section 2 discusses related research. Section 3 describes the BLSTM-LSTM based the encoder-decoder, the attention and focus mechanisms. Section 4 reports the experiment results. Finally, Section 5 draws conclusions.

2. RELATED WORKS

Recent research regarding slot filling has been focused on RNN and its extensions. At first, [7] used RNN to beat CRF in the ATIS dataset. [8] tried bi-directional and hybrid RNN to investigate using RNN for slot filling. [11] introduced LSTM and deep LSTM architecture for this task and obtained a marginal improvement over RNN. [14] proposed RNN-EM which used an external memory architecture to improve the memory capability of RNN. [13] proposed to use the ranking loss function to train a bi-directional RNN.

Except for the architectures of neural networks, many studies have been conducted to model the label dependencies. [10] proposed to combine CNN and CRF for sentence-level optimization. [8, 18] combined Elman-type and Jordan-type RNNs to consider the dependency on the last output label.

Following the success of attention based models in the NLP field, [19] applied the attention-based encoder-decoder to the slot filling task, but without LSTM cells. [15] proposed encoder-labeler architecture with two LSTMs which are encoder LSTM and labeler LSTM. The encoder-labeler model got the best performance of 95.66% F_1-score in the ATIS dataset.

In order to achieve a full investigation, we combine BLSTM which considers the past and future information within the powerful encoder-decoder model to introduce the BLSTM-LSTM based encoder-decoder in sequence labelling task.

3. PROPOSED MODELS

By considering the past inputs only, unidirectional LSTM cannot solve long distance dependencies of future inputs. BLSTM addressed this shortcoming with two unidirectional LSTMs: a forward pass which processes the original input word sequence; a backward pass which processes the reversed input word sequence. To learn the advantages of these models, we are going to introduce a BLSTM-LSTM based encoder-decoder architecture.

3.1. BLSTM-LSTM + Attention

We followed the encoder-decoder from [16] which is based on RNN. To consider both the previous history and the future history, we use BLSTM as the encoder and LSTM as the decoder.

An important extension of encoder-decoder is to add an attention mechanism. We adopted the attention model from [17]. The only difference is that we use BLSTM as encoder in advance. The encoder reads the input sentence $\mathbf{x} = (x_1, x_2, ..., x_{T_x})$ and generates T_x hidden states by BLSTM:

$$\begin{aligned} h_i &= [\overleftarrow{h_i}, \overrightarrow{h_i}] \\ \overleftarrow{h_i} &= f_l(\overleftarrow{h_{i+1}}, x_i) \\ \overrightarrow{h_i} &= f_r(\overrightarrow{h_{i-1}}, x_i) \end{aligned}$$

where $\overleftarrow{h_i}$ is the hidden state of backward pass in BLSTM and $\overrightarrow{h_i}$ is the hidden state of forward pass in BLSTM at time i.

The decoder is trained to predict the next semantic label y_t given the all input words and all the previously predicted semantic labels $\{y_1, ..., y_{t-1}\}$:

$$\begin{aligned} P(y_t|y_1, ..., y_{t-1}; \mathbf{x}) &= g(s_t) \\ s_t &= f_d(s_{t-1}, y_{t-1}, c_t) \\ c_t &= q(s_{t-1}, h_1, ..., h_{T_x}) \end{aligned}$$

where g refers to the output layer (often with softmax) and s_t is the hidden state of decoder LSTM at time t, with f_d set as LSTM unit function. c_t denotes the contextual information for generating label y_t according to different encoder hidden states, which is typically implemented by an attention mechanism [16], e.g.

$$\begin{aligned} c_t &= \sum_{i=1}^{T_x} \alpha_{ti} h_i \\ \alpha_{ti} &= \frac{\exp(a(s_{t-1}, h_i))}{\sum_{j=1}^{T_x} \exp(a(s_{t-1}, h_j))} \end{aligned}$$

where a is a feed-forward neural network. s_0 is initialized with $\overleftarrow{h_1}$. In order to apply this model for sequence labelling task, we enforce the output sequence generated by the decoder to get the same length of the input word sequence.

3.2. Focus mechanism

As referenced in the introduction, the attention mechanism is facing with two limitations in sequence labelling based SLU task. To address these problems, we propose the focus mechanism that only considers the aligned encoder hidden state, i.e. $\alpha_{ti} = 0, \text{ if } t \neq i; \alpha_{ti} = 1, \text{ if } t = i$. Thus,

$$c_t = h_t$$

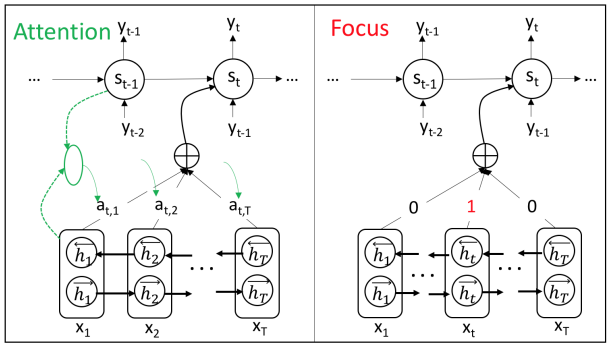

Fig. 2. Illustration of the attention and focus mechanism.

Therefore, there is no necessity to learn the alignment by utilizing the attention model. The encoder-decoder with attention and focus mechanisms are illustrated as figure 2.

4. EXPERIMENTS

4.1. Experimental Setup

We use the ATIS corpus which has been widely used as a benchmark by the SLU community. In ATIS, the sentence and its semantic slot labels are in the popular in/out/begin (IOB) representation. An example sentence is provided in figure 1. The training data consists of 4978 sentences and 56590 words. Test data consists of 893 sentences and 9198 words. We randomly selected 80% of the training data for model training and the remaining 20% for validation [9].

In addition to ATIS, we also apply our models for a custom Chinese dataset from the car navigation domain which contains 8000 utterances for training, 2000 utterances for validation and 1944 utterances for testing. Each word has been manually assigned a slot using IOB schema. Not only the natural sentence, the top hypothesis of each utterance produced from the automatic speech recognition (ASR) is also evaluated. These ASR top outputs have a word error rate (WER) of 4.75% and a sentence error rate (SER) of 23.42%.

We report the F_1-score on the test set with parameters that achieved the best F_1-score on the validation data. We deal with unseen words in the test set by marking any words with only one single occurrence in the training set as $< unk >$.

Our implemented LSTM neural networks are identical to the ones in [20]. As described earlier, the encoder-decoder model utilized a BLSTM for encoding and a LSTM for decoding. For training, the network parameters are randomly initialized in accordance with the uniform distribution (-0.2, 0.2). We used the stochastic gradient descent (SGD) for updating parameters. In order to enhance the generalization capability of our proposed models, we applied *dropout* with a probability of 0.5 during the training stage.

For encoder-decoder, we used left-to-right beam searching for decoding with beam size of two empirically. We tried different learning rates, ranging from 0.004 to 0.04 similar to grid-search. We kept the learning rate for 100 epochs and saved the parameters that gave the best performance on the validation set, which is measured after each training epoch.

4.2. Results on the ATIS Dataset

Table 1 shows the results on ATIS dataset. For all architectures, we set the dimension of word embeddings to 100 and the number of hidden units to 100. We only use the current word as input without any context words. BLSTM, which considers both the past and the future history, outperforms LSTM (+2.03%). The *attention based BLSTM-LSTM* model got lower F1-score than *BLSTM* (-2.7%). We think the reason is that the sequence labelling problem is a task, whose input and output sequences are aligned.

Having only limited data, it is difficult to learn the alignment accurately by using the attention mechanism. We try to expand the training data of ATIS by randomly replacing the value of each specific slot within sentences to 10 times that of the original scale. For example, "Flights from Boston" can be expanded to "Flights from New York", "Flights from Los Angeles", etc. The BLSTM-LSTM with attention achieves a 95.19% F_1-score, while other methods did not benefit from the expanded training set.

Model	Mechanism	F_1-score (%)
LSTM	-	93.40
BLSTM	-	95.43
BLSTM-LSTM	Attention	92.73
	Focus	**95.79**

Table 1. *Experimental results on ATIS dataset.*

By considering the alignment of the sequence labelling task, the *BLSTM-LSTM with focus* increased the F1-score from 92.73% to 95.79% and achieved an 0.36% improvement (significant level 10%) in comparison to *BLSTM*. We think the *BLSTM-LSTM with focus* has two advantages over the *BLSTM*: 1) the initialization of hidden state of decoder LSTM with $s_0 = \overleftarrow{h_1}$ provides sentence leveraging features; 2) it enables label dependency within the decoder.

Compared with the published results on the ATIS dataset, our method outperforms the previously published F1-score, illustrated in Table 2. Table 2 summarizes the recently published results on the ATIS slot filling task and compares them with the results of our proposed methods. Our proposed model achieves state-of-the-art performance [1] but not statistically significant.

[1]There are other published results that achieved better performance by using Name Entity features, e.g. [8] achieved 96.24% F_1-score. The NE features are annotated and really strong. If only using NE features, BLSTM obtained 97.00% F_1-score. So it would be more meaningful to use only lexicon features.

Model	F_1-score
CRF [8]	92.94
simple RNN [7]	94.11
CNN-CRF [10]	94.35
LSTM [11]	94.85
RNN-SOP [18]	94.89
Deep LSTM [11]	95.08
RNN-EM [14]	95.25
Bi-RNN with Ranking Loss [13]	95.47
Encoder-labeler Deep LSTM [15]	95.66
BLSTM-LSTM (focus)	**95.79**

Table 2. *Comparison with published results on ATIS.*

4.3. Results on Chinese Navigation Dataset

To investigate the robustness of the BLSTM-LSTM architectures with the attention or focus mechanism, we conduct additional experiments on the Chinese navigation dataset described in the experimental setup. For the neural network architectures, we also set the dimension of word embeddings to 100 and the number of hidden units to 100. Additionally, only the current word is used as LSTM input, in comparison to CRF which used a context window size of 5. We train the model on natural text sentences (without any speech recognition errors) and test it on not only manual transcriptions (correct text sentences), but also top hypotheses from speech recognition systems (including recognition errors).

Model	Mechanism	Manual Trans.	ASR Hyp.
CRF	-	94.55	91.51
LSTM	-	79.90	74.25
BLSTM	-	95.33	91.23
BLSTM-LSTM	`Attention`	95.65	91.76
	`Focus`	**96.60**	**93.08**

Table 3. *F_1-scores of manual transcription and top hypothesis from ASR on Navigation dataset.*

Table 3 shows the results. CRF baseline seems competitive to BLSTM, due to the sentence-level optimization of the output. In comparison, the LSTM does not meet our expectations. Because the main challenge in this dataset is detecting longer phrases like location name (the length varies from 1 to 24 words). It suffers from long distant dependencies on past and future inputs. Subsequently, BLSTM solves this problem.

BLSTM-LSTM with focus-mechanism outperforms BLSTM on both natural sentences and top hypotheses from ASR significantly (significant level 5%). It seems BLSTM-LSTM encoder-decoder with focus mechanism is more robust to ASR errors. A possible reason is, that the label dependency in the decoder helps omit the error transformed from the encoder. CRF also models label dependency and outperforms BLSTM by parsing ASR outputs.

5. CONCLUSIONS

In our study, we have applied multiple BLSTM-LSTM encoder-decoders with attention and focus mechanism to SLU slot filling task. The BLSTM-LSTM architecture with focus mechanism achieved a state-of-the-art result on the ATIS dataset and shows to be robust to the ASR errors on a custom dataset. We also revealed that the attention mechanism needs more data to learn the alignment, while the focus mechanism has considered the alignment property of the sequence labelling problem. In future, we want to investigate BLSTM-LSTM with focus mechanism to other sequence labelling tasks (e.g. part-of-speech tagging, named entity recognition). Furthermore, we plan to use attention based BLSTM-LSTM for solving the SLU task in cases data is only provided unaligned.

6. REFERENCES

[1] Ye-Yi Wang, Li Deng, and Alex Acero, "Spoken language understanding," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 16–31, 2005.

[2] Yulan He and Steve Young, "A data-driven spoken language understanding system," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2003, pp. 583–588.

[3] John Lafferty, Andrew McCallum, and Fernando CN Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001.

[4] K Taku and M Yuji, "Chunking with support vector machine," in *Proceedings of North American chapter of the association for computational linguistics*, 2001, pp. 192–199.

[5] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur, "Recurrent neural network based language model.," in *INTERSPEECH*, 2010, vol. 2, p. 3.

[6] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig, "Linguistic regularities in continuous space word representations.," in *HLT-NAACL*, 2013, pp. 746–751.

[7] Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu, "Recurrent neural networks for language understanding.," in *INTERSPEECH*, 2013, pp. 2524–2528.

[8] Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding.," in *INTERSPEECH*, 2013, pp. 3771–3775.

[9] Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al., "Using recurrent neural networks for slot filling in spoken language understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 530–539, 2015.

[10] 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 (ASRU)*. IEEE, 2013, pp. 78–83.

[11] Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Geoffrey Zweig, and Yangyang Shi, "Spoken language understanding using long short-term memory neural networks," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 189–194.

[12] Kaisheng Yao, Baolin Peng, Geoffrey Zweig, Dong Yu, Xiaolong Li, and Feng Gao, "Recurrent conditional random field for language understanding," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4077–4081.

[13] Ngoc Thang Vu, Pankaj Gupta, Heike Adel, and Hinrich Schütze, "Bi-directional recurrent neural network with ranking loss for spoken language understanding," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[14] Baolin Peng, Kaisheng Yao, Li Jing, and Kam-Fai Wong, "Recurrent neural networks with external memory for spoken language understanding," in *Natural Language Processing and Chinese Computing*, pp. 25–35. Springer, 2015.

[15] Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu, "Leveraging sentence-level information with encoder lstm for natural language understanding," *arXiv preprint arXiv:1601.01530*, 2016.

[16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[17] Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton, "Grammar as a foreign language," in *Advances in Neural Information Processing Systems*, 2015, pp. 2755–2763.

[18] Bing Liu and Ian Lane, "Recurrent neural network structured output prediction for spoken language understanding," in *Proc. NIPS Workshop on Machine Learning for Spoken Language Understanding and Interactions*, 2015.

[19] Edwin Simonnet, Nathalie Camelin, Paul Delglise, and Yannick Estve, "Exploring the use of attention-based recurrent neural networks for spoken language understanding," in *Machine Learning for Spoken Language Understanding and Interaction NIPS 2015 workshop (SLUNIPS 2015)*, Montreal (Canada), 11 dec. 2015.

[20] Alex Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.