Streamlined Decoder for Chinese Spoken Language Understanding

Puhai Yang
Beijing Institute of Technology
Beijing, China
yangpuhai@126.com

Heyan Huang
Beijing Institute of Technology
Beijing, China
hhy63@bit.edu.cn

Xian-Ling Mao
Beijing Institute of Technology
Beijing, China
maoxl@bit.edu.cn

ABSTRACT

As a critical component of Spoken Dialog System (SDS), spoken language understanding (SLU) attracts a lot of attention, especially for methods based on unaligned data. Recently, a new approach has been proposed that utilizes the hierarchical relationship between act-slot-value triples. However, it ignores the transfer of internal information which may record the intermediate information of the upper level and contribute to the prediction of the lower level. So, we propose a novel streamlined decoding structure with attention mechanism, which uses three successively connected RNN to decode act, slot and value respectively. On the first Chinese Audio-Textual Spoken Language Understanding Challenge (CATSLU), our model exceeds state-of-the-art model on an unaligned multi-turn task-oriented Chinese spoken dialogue dataset provided by the contest.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**;

KEYWORDS

spoken language understanding, spoken dialog system, streamlined decoder, pointer network, attention mechanisms, long short term memory networks

ACM Reference Format:
Puhai Yang, Heyan Huang, and Xian-Ling Mao. 2019. Streamlined Decoder for Chinese Spoken Language Understanding. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3340555.3356097

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICMI '19, October 14–18, 2019, Suzhou, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3356097

1 INTRODUCTION

At present, task-oriented SDS is being widely used, which puts forward extreme requirements for SLU as the core of the system. Typically, SLU is treated as a sequence annotation task and solved using methods such as conditional random fields (CRFs) [7], convolutional neural networks (CNN) [5] and Long Short Term Memory networks (LSTM) [13] . Although these methods achieve satisfactory results, they all require word-level annotations for training, which is difficult to obtain or requires a large amount of manual labor. e.g., the utterance "I want to fly to Boston tomorrow" will be annotate as "I want to fly to Boston (Dest) tomorrow (ArDay)" [10].

Recently, increasing attention has been paid to the method based on sentence-level annotations, which are unaligned to parse out specific tuples from utterance, e.g., (Dest= Boston) and (ArDay=tomorrow) can be parsed from above utterance. In SDS, SLU is usually a follow-up to Automatic Speech Recognition (ASR), and unaligned annotations are more robust to ASR errors. In this paper, we focus on unaligned methods which extract a set of act-slot-value triples from sentence.

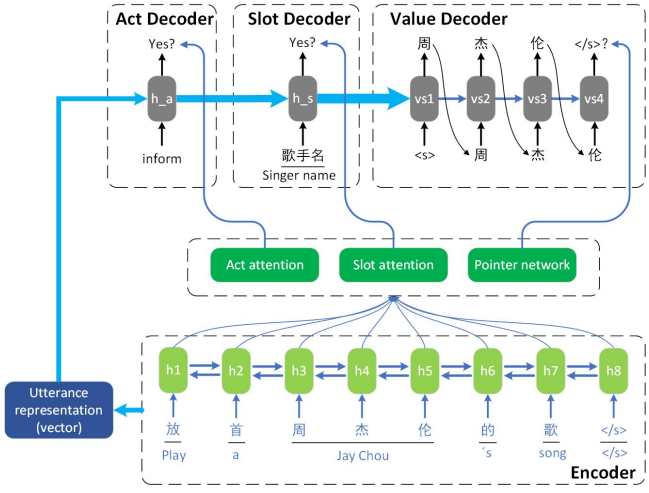

Figure 1: Our streamlined decoding model for SLU

Many different methods from unaligned annotations have been proposed, Henderson et al. [3] explored a SVM model

for semantic classifier on n-gram features. Barahona et al. [1] designed a combination of CNN for the sentence representation and LSTM for the context representation which joint predicts acts and slots. And Zhao et al. [11] proposed using pointer network to solve the out-of-vocabulary (OOV) problem. In a recent study, Zhao et al. [12] proposed a hierarchical model based on the pointer network to solve SLU. However, they all ignore the fluidity of internal information when predicting act-slot-value triples, which is particularly important in a hierarchical structure.

In this paper, we present a streamlined decoding structure for SLU on unaligned utterances. Based on Seq2Seq learning, the model uses bidirectional LSTM (BLSTM) [2] as encoder to obtain the representation of input utterance, and uses three successively connected LSTM to form a streamlined joint decoder, predicting act, slot and value respectively. In act decoder and slot decoder, attention mechanisms are used to give more reasonable attention to utterance representation. In value decoder, due to the lack of abundant morphological changes of words in Chinese as in English [6], pointer network is used to directly extract value from utterance, and the OOV problem is addressed accordingly.

In the following part of this paper, section 2 introduces the previous related work, and section 3 elaborates on the streamlined decoding model, section 4 is the part of the experiment and analysis, and the last section summarizes the work.

2 RELATED WORK

Our work is an improvement and extension of the research of Zhao et al. [12], who proposed solving SLU by using a hierarchical structure. The hierarchical model use BLSTM to encode the input text, which served as a shared text representation. Next, the slot multi-category prediction module uses predicted act and the shared text representation, while the value generation module applies a context-aware attention mechanism within the pointer network by incorporating the shared text representation and the predicted act and slot. However, internal information in the prediction, such as the structure of input text inferred from previous modules, is ignored and only the superior results are used. In this paper, we take the text representation as the source and stream it over a streamlined joint decoder consisting of multiple LSTMs, thus ensuring the correct transmission of needed internal information. The experimental comparison on CAT-SLU dataset [9] also proves the superiority of this model compared with the hierarchical model.

3 STREAMLINED DECODING MODEL

Our model, as shown in figure 1, consists of four parts: Encoder, Act Decoder, Slot Decoder and Value Decoder. Encoder is used to convert the input utterance into its vector representation, Act Decoder and Slot Decoder judge whether an act or slot exists in the input utterance, and finally, Value Decoder is used to generate the value of the slot. Visually, the vector representation of the utterance is reinforced as it flows through three decoders, what we call enhanced vector representation (EVR). Next, each part of the model is described in detail.

Utterance Encoder

In the encoder, we use BLSTM [4][8] to process the input utterance $U = (x_1, x_2, \ldots, x_n)$ and generate n hidden states by BLSTM as follows:

$$h_k = (\overleftarrow{h_k}, \overrightarrow{h_k})$$
$$\overleftarrow{h_k} = f_{left}(\overleftarrow{h_{k+1}}, x_k)$$
$$\overrightarrow{h_k} = f_{right}(\overrightarrow{h_{k-1}}, x_k)$$

where $\overleftarrow{h_k}$ is the hidden state of backward pass in BLSTM and $\overrightarrow{h_k}$ is the hidden state of forward pass in BLSTM at time k. Finally, the vector representation of the utterance is defined as:

$$\hat{h} = \overleftarrow{h_1} \oplus \overrightarrow{h_n}$$

which will be the input for act decoder. Meanwhile, use all of the hidden states $h_1, h_2, \ldots, h_n$ for the following attention and pointer network mechanism of streamlined decoder.

Act Decoder and Slot Decoder

Act decoder and slot decoder have similar structures, which are used to determine whether an act or slot is contained in the utterance, and their output are the result of prediction and utterance EVR. In the paper, the LSTM [4] is used to implement this process.

In act decoder, based on the research of Zhao et al. [12], we add act vector to the input, so as to improve the accuracy of prediction by increasing the input information of the model. In addition, act information can be integrated into the utterance representation vector to obtain EVR, so as to promote the prediction of the next slot. The output of the act decoder are an EVR and a binary classification result that characterize whether act is contained in the text, as follows:

$$h_{act\_evr} = f_{act\_lstm}(x_{act}, \hat{h})$$
$$P_{act} = f_{act\_bi\_class}(h_{act\_evr} \oplus a\_s)$$

where x_{act} is the vector of act, f_{act_lstm} is the LSTM unit, $f_{act\_bi\_class}$ is a binary classifier containing drop out and a full connection layer, and a_s is the context vector calculated by the attention mechanism using h_{act_evr}, as follows:

$$a\_s = \sum_{k=1}^{n} a_k * h_k$$

$$a_k = \frac{exp({h_k}^T * h_{act\_evr})}{\sum_{j=1}^{n} exp({h_j}^T * h_{act\_evr})}$$

For the slot decoder, the structure is exactly the same as act decoder, except that the EVR h_{act_evr} output by the act decoder is used as input, and the EVR it outputs will be used as the input for the value decoder, as follows:

$$h_{slot\_evr} = f_{slot\_lstm}(x_{slot}, h_{act\_evr})$$

Due to the differences in the CATSLU dataset (acts in English while slots in Chinese), we create dictionaries for act and slot respectively and take one-hot vector as their input.

Value Decoder

After the utterance representation vector flows through act decoder and slot decoder, we obtain the vector enhanced with act and slot information, namely EVR. In the value decoder, based on seq2seq learning, LSTM is used to generate value. Different from Zhao et al. [11], who use attention and pointer network to build value decoder, value can be extracted directly from the input Chinese utterance since there is no need for word morphology transformation in Chinese. Therefore, we simplify the value decoder structure of Zhao et al. [11], remove attention and retain pointer network for selecting the position of words in value directly from input.

At timestep t, the probability of selecting the word of position i in the input utterance can be calculated as follows:

$$P_i^t = \frac{exp({h_i}^T * vs_t)}{\sum_{j=1}^{n} exp({h_j}^T * vs_t)}$$

where vs_t is the output of the LSTM unit at timestep t, as follows:

$$vs_t = f_{value\_lstm}(y_{t-1}, vs_{t-1})$$

where f_{value_lstm} is the LSTM unit, y_{t-1} is the word output at timestep $t-1$. At timestep $t = 1$, y_{t-1} is "<s>" what we defined as a start word, and vs_{t-1} is h_{slot_evr} comes from the output of slot decoder.

4 EXPERIMENTS

In this section, we present the experiments of the proposed model on CATSLU dataset, including SLU results in the experiments and analysis of the attention mechanism. In the experiment, we use PyTorch to deploy the model, and set word embeddings dimension to 200, hidden states dimension to 256, drop out to 0.5 and batch size to 10. Instead of pre-trained word embedding, one-hot vectors of words are implanted into the model to learn word embedding. Moreover, 50 epochs are carried out when training the model.

Data

The dataset is provided by the CATSLU challenge [9]. The CATSLU is a competition hosted by the 21st ACM International Conference on Multimodal Interaction (ICMI 2019) and consists of two sub-challenges: SLU in a single domain and Domain adaptation of SLU. The single domain task provides a large dataset for both the map and music domains, while the domain adaptation task requires adapt the SLU models of map and music domains to the weather and video domains. The dataset includes text and audio of multiple rounds of Chinese dialogues. And the statistical information of the dataset is given in table 1.

Table 1: Statistics of CATSLU dataset

Domain	Slots	Train	Development	Test
Map	24	5093	921	1578
Music	20	2189	381	676
Weather	22	341	378	2660
Video	28	205	195	1641

In our experiment, the experiment on each domain is treated as a single domain task and SLU model is trained independently on the data of each domain.

Baseline

The CATSLU challenge provides two baselines for experimental comparison, a rules-based model that uses ontology information for simple string matching and a neural network-based model adopts the newly proposed hierarchical decoding structure [12]. In order to compare the impact of model structure improvement on the results, the neural network-based model uses exactly the same settings as our proposed model in the experiment.

SLU Result

Two types of input utterance text are provided in the CATSLU dataset: manual transcriptions and ASR 1-best hypothesis, and the utterance text in the test datasets contains only ASR 1-best hypothesis, the results on the test datasets returned by the challenge organizer demonstrate the superiority of our model as shown in table 2. We can see that our proposed streamlined decoding model always achieve the best F1-score and accuracy in map and music domains with large amount of training data. For weather and video domains with few training data, our model always performs best except F1-score in the video domain. Overall, our model always outperforms the hierarchical decoding model which also based on neural network on the test datasets.

Table 2: Results on test datasets

Domain		Map	Music	Weather	Video
Rule-based	F1	37.92	77.39	85.52	**78.25**
Model	Acc	40.43	49.26	75.38	45.28
Hierarchical	F1	77.61	81.57	85.25	75.18
Model	Acc	74.65	71.15	78.16	57.53
Our Model	F1	**78.40**	**82.80**	**87.39**	77.84
	Acc	**74.71**	**72.19**	**80.86**	**62.10**

Next, we will analyze the key role that enhanced vector representation of utterance plays in our model by visualizing the attention mechanism.

Analysis

We randomly select four samples in the music domain and drew a profile of the manually labeled results, as shown in figure 2. For a more detailed analysis of the role of the proposed streamlined decoder structure, we visualize the attention mechanisms in act decoder and slot decoder respectively, as shown in figure 3 and figure 4.

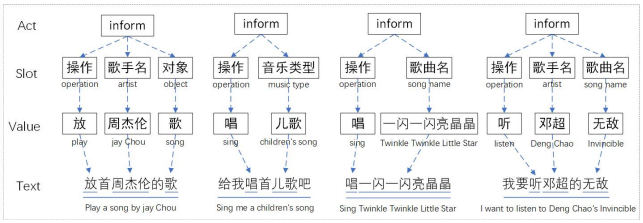

Figure 2: Results of manual labeling of four samples

In figure 3, the weight corresponding to each pixel is not only how important each word in the input text is when predicting act, but also how much attention EVR enhanced by act information pays to each word in the input text. It can be found that for acts that exist in the text, some information about potentially useful words is added to EVR, while for acts that do not exist, only useless end-sign information to predicting slots is added to EVR.

The role played by EVR in streamlined decoder is more clearly shown in figure 4, it can be observed that slot decoder adds information about the starting position of value in the input text to EVR for existing slot. We can reasonably assume that, after integrating act and slot information with streamlined decoder structure, some other internal information, such as text structure information, is added to EVR in addition to the starting position information, and then the streamlined structure transmits the information to value decoder through EVR to help value prediction.

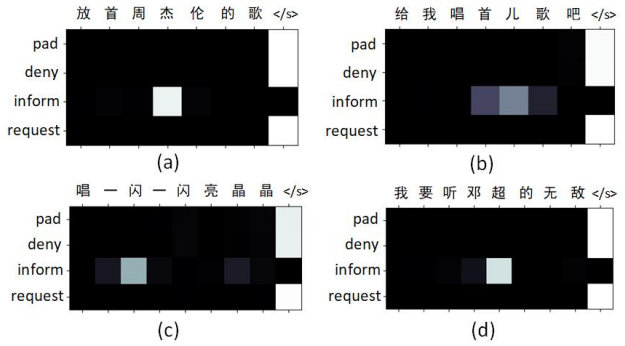

Figure 3: Visualization of attention mechanism in act decoder for four samples, the x-axis and y-axis of each plot correspond to the words in the input text and the act in dictionary, respectively. Each pixel shows the weight of the annotation of the word in input text for the act, in grayscale (0: black, 1: white).

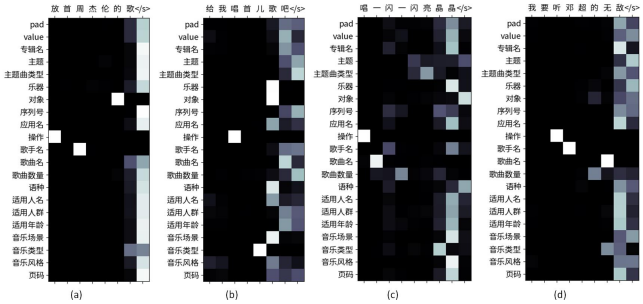

Figure 4: Visualization of attention mechanism in slot decoder for four samples, the x-axis and y-axis of each plot correspond to the words in the input text and the slot in dictionary, respectively. Each pixel shows the weight of the annotation of the word in input text for the slot, in grayscale (0: black, 1: white).

5 CONCLUSION

We proposed a streamlined decoding model for Chinese spoken language understanding, utilizing three sequentially connected RNN to decode act, slot. And pointer network is adopted in value decoder to extract value directly from input text, which also avoids OOV problem. On the Chinese multi-round dialogue dataset provided by the CATSLU challenge, our proposed streamlined decoding model outperforms the state-of-the-art hierarchical decoding model.

ACKNOWLEDGMENTS

The work is supported by SFSMBRP (2018YFB1005100), BIGKE (No. 20160754021), NSFC (No. 61772076 and 61751201), NSFB (No. Z181100008918002), Major Project of Zhijiang Lab (No. 2019DH0ZX01), and CETC (No. w-2018018).

REFERENCES

[1] Lina M Rojas Barahona, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, Stefan Ultes, Tsung-Hsien Wen, and Steve Young. 2016. Exploiting sentence and context representations in deep neural models for spoken language understanding. *arXiv preprint arXiv:1610.04120* (2016).

[2] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5-6 (2005), 602–610.

[3] Matthew Henderson, Milica Gašić, Blaise Thomson, Pirros Tsiakoulis, Kai Yu, and Steve Young. 2012. Discriminative spoken language understanding using word confusion networks. In *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 176–181.

[4] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[5] Chunxi Liu, Puyang Xu, and Ruhi Sarikaya. 2015. Deep contextual language understanding in spoken dialogue systems. In *Sixteenth annual conference of the international speech communication association.*

[6] James H Martin and Daniel Jurafsky. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Pearson/Prentice Hall Upper Saddle River.

[7] Christian Raymond and Giuseppe Riccardi. 2007. Generative and discriminative algorithms for spoken language understanding. In *Eighth Annual Conference of the International Speech Communication Association.*

[8] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[9] Tiejun Zhao Chengqing Zong Su Zhu, Zijian Zhao and Kai Yu. 2019. CATSLU: The 1st Chinese Audio-Textual Spoken Language Understanding Challenge. In *21st ACM International Conference on Multimodal Interaction.*

[10] Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu. 2013. Recurrent neural networks for language understanding.. In *Interspeech*. 2524–2528.

[11] Lin Zhao and Zhe Feng. 2018. Improving Slot Filling in Spoken Language Understanding with Joint Pointer and Attention. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 426–431.

[12] Zijian Zhao, Su Zhu, and Kai Yu. 2019. A Hierarchical Decoding Model for Spoken Language Understanding from Unaligned Data. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7305–7309.

[13] Su Zhu and Kai Yu. 2017. Encoder-decoder with focus-mechanism for sequence labelling based spoken language understanding. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5675–5679.