THE MSIIP SYSTEM FOR DIALOG STATE TRACKING CHALLENGE 5

Ying Su, Miao Li, Ji Wu

Multimedia Signal and Intelligent Information Processing Laboratory,
Department of Electronic Engineering,
Tsinghua University, Beijing , China
{suying16,miaoli-10}@mails.tsinghua.edu.cn
wuji_ee@mail.tsinghua.edu.cn

ABSTRACT

We present our work in Dialog State Tracking Challenge 5, the main task of which is to track dialog state on human-human conversations cross language. Firstly a probabilistic enhanced framework is used to represent sub-dialog, which consists of three parts, the input model for extracting features, the enhanced model for updating dialog state and the output model to give the tracking frame. Meanwhile, parallel language systems are proposed to overcome inaccuracy caused by machine translation for cross language testing. We also introduce a new iterative alignment method extended from our work in DSTC4. Furthermore, a slot-based score averaging method is introduced to build an ensemble by combining different trackers. Results of our DSTC5 system show that our method significantly improves tracking performance compared with baseline method.

Index Terms— dialog state tracking, probabilistic enhanced frame strcture, parallel language system

1. INTRODUCTION

In this paper, we focus on proposing an approach to build an effective system for dialog state tracking challenge 5. Tracking human-human dialog can be of great difficulty because human language is casual and meaningful, which is very hard for computer to understand. Recent researches have attached great attention to dialog state tracking. Conventional method like choosing the most possible hypothesis and discarding the rest may not successfully overcome uncertainty problem in natural language challenge, like Automatic Speech Recognition (ASR) and Spoken Language Understanding (SLU). Based on Partially Observable Markov Process (POMDP) framework, relevant systems are built [1, 2]. Recently, the application of deep learning has been used as well, like Deep Neural Network (DNN) [3], and Recurrent Neural Network (RNN) [4].

Previous DSTC4 task began to stress on dialog state tracking human-human dialogs. Furthermore, in DSTC5, different languages are used in the training, developing and testing corpus. The source language of training corpus is English, whereas the developing and testing corpus are in Chinese. To cope with this cross-language issue, machine translation results are provided. The domain of DSTC5 is the same as that in DSTC4, including food, traffic, accommodation, etc.

In our DSTC4 system [5], a probabilistic enhanced frame structure is proposed to represent sub-dialog states, which can handle ASR and NLU errors effectively. Whereas in DSTC5, we propose a new soft iterative alignment method based on soft attention mechanisms [6], which have been employed in deep learning recently and have got impressive results. We adopt an alignment method which holds all utterances in a sub-dialog with different weights. These weights imply how important one utterance in a sub-dialog is to match a specific frame label. To address the cross language issue, two dialog tracking systems in Chinese and English respectively are built based on same algorithm. With the parallel systems in different languages, the performance can benefit from information. However, there is one drawback due to machine translation. Despite alignment list for words in two languages given, semantic accuracy can not be perfectly guaranteed, which probably degrades system performance.

The rest of this paper is divided into 6 parts. In section 2 the details of main task will be introduced. Then our algorithms and trackers are described in section 3 and section 4. Section 5 introduce ensemble method. Section 6 presents the evaluation results. Finally section 7 concludes the paper.

2. TASK DESCRIPTION

The training dataset for main task consists of 35 dialog sessions, 4261 sub-dialogs, 31034 utterances in total on touristic information for Singapore, collected from Skype calls between tour guides and tourists. Each dialog session is divided into several sub-dialogs. Each sub-dialog is made up by several utterances. One specific topic and dialog state are assigned to every sub-dialog. Speech act and semantic labels have been manually annotated for every utterance. Dialog state is represented by slot-value pairs. A slot roughly describes more

detailed info of sub-dialog, like preference and place. Then a value is defined to clarify a slot more specifically, like market and hotel in type_of_place.

DSTC5 stresses a cross-language tracking problem, the goal of which is to build a tracker in the target language with existing resources in the source language and their translations generated automatically by machine translation technologies to the target language[7].

Topic	Frame/slot	Values
Attraction	INFO	Preference
	TYPE_OF_PLACE	Historic site
		Cultural site

Utterance	Source language English	Target language Chinese (top-1 translation)
Turn 1	Guide: Okay and ha-%uh okay, what do you and your friends like to do?	哈呃好的好的，你和你的朋友们喜欢做什么？
Turn 2	Tourist: %Um for me, I like to know about the culture, history and of course historical places in Singapore.	嗯，我想知道我的文化，当然历史和历史古迹在新加坡。
Turn 3	Guide: Uh sorry, I didn't get you.	呃对不起，我不明白你的意思。

Fig. 1. The above diagram shows an example of source language and its top-1 translation to target language on training dataset. The below diagram shows relationship between topic, slot and value of the about sub-dialog.

A baseline tracker is provided with two methods, both of which are based on fuzzy string matching, one in English and the other in Chinese. The source language on training dataset is English but Chinese on developing dataset. For both methods, 5-best translations for source language to target language is provided. Only the top-1 hypothesis of the 5 translations is used. For the rest of this paper, when the translation of source language context is mentioned, we refer to the top-1 hypothesis of its 5 translations.

The performances of trackers are evaluated on utterance level or sub-dialog level. Two sets of evaluation metrics are used for the main task. One is accuracy, which requires trackers output is exactly the same as the gold standard frame structure. The other one is precision, recall and F-measure. F-measure is harmonic mean of precision and recall.

3. PROBABILITY ENHANCED FRAME STRUCTURE

Previous research in DSTC4 [5] has come up with a probabilistic enhanced frame structure to represent a sub-dialog state. The frame structure is shown in Fig 2. u_t denotes the utterance at time t. h_t denotes a hidden state for u_t. s_t is a probability represented dialog state, related with its previous state and current hidden state. Based on s_t, we have output result o_t at time t. Two types of probabilities are calculated using this structure, one attached to slot and the other attached to a slot-value pair. The main structure can be divided into three parts: the input model $p(h_t|u_t)$, the update model $p(s_{t+1}|s_t, h_t)$ and the output model $p(o_t|s_t)$. Details of three models are described below.

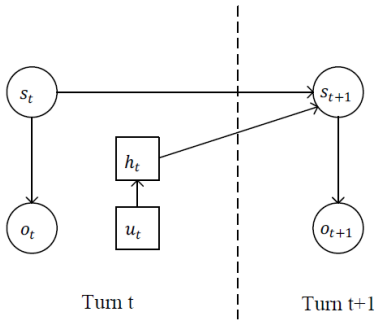

Fig. 2. The turn-taking algorithm from turn t to turn t+1.

3.1. input model

The input utterance is transformed into a feature vector. Three kinds of features are used: unigram feature, bigram feature and the slot-value pairs extracted by the baseline tracker. For different tracker, the feature used differs.

3.2. update model

The update method considering two types of probabilities is shown in (1),(2), which follows the idea in [5, 8].

$$p^t_{state}(s) = 1 - (1 - p^{(t-1)}_{state}(s))(1 - p^t_{turn}(s)) \quad (1)$$

$$q^t_{state}(s, v) = 1 - (1 - q^{(t-1)}_{state}(s, v))(1 - q^t_{turn}(s, v)) \quad (2)$$

Where $p^t_{state}(s)$ is the probability of slot s in sub-dialog state at time t, $p^t_{turn}(s)$ is the probability of slot s in turn t. A turn means one utterance from speakers, including guide and tourist. $q^t_{state}(s, v)$ is the probability of slot-value pairs s, v in sub-dialog at time t and $q^t_{state}(s, v)$ is the probability of slot-value pairs s, v in turn t. If a slot or slot-value pair if not in sub-dialog, then its probability is assigned 0.

3.3. output model

The probability of slot or slot-value pair increases with time forwarding. The more turns get involved, the more understanding results about slot or slot-value pairs will be acquired. A threshold T_s is set for probabilities of slots and a threshold T_v is set for probabilities of slot-value pairs. For both slot and slot-value, if the probability exceeds threshold, then it will be part of the output frame.

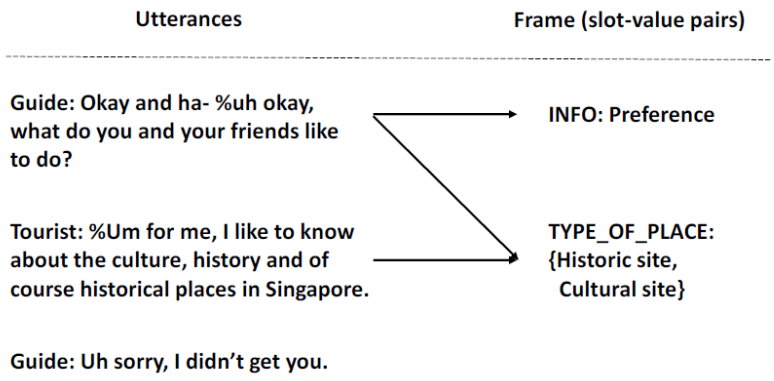

Fig. 3. The arrow represents the relation of utterance and slot-value pair. The utterance can be viewed as a positive sample of a slot-value pair if there is an arrow between them.

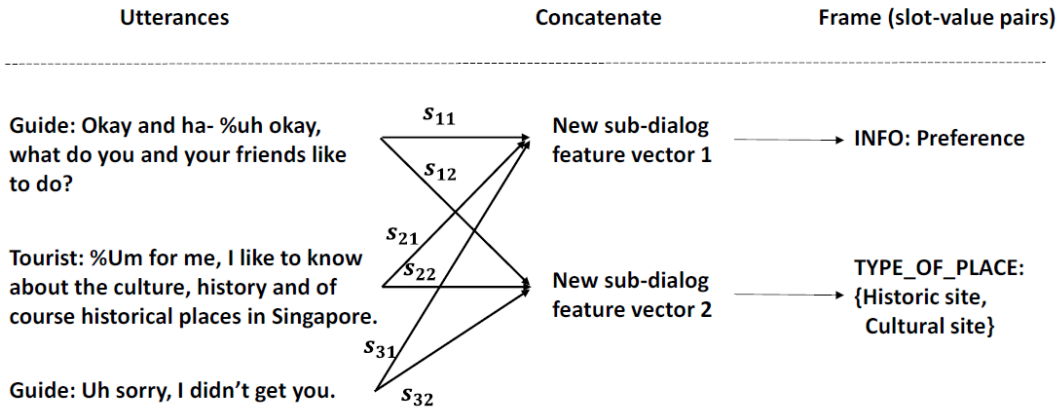

Fig. 4. A new feature vector is constructed by utterances concatenated of different weight. Then the feature vector is viewed as a new positive sample for slot-value pair. For different slot-value pairs, the set of weights is different from each other. s_{11},s_{12},s_{21},s_{22},s_{31},s_{32} are weights that utterances align to frame lables.

4. PARSERS WITH DIFFRENT SLU MODELS

This section first introduce related work inherited from D-STC4 [5]. Then the approach we propose to build new SLU models for DSTC5 are described.

4.1. Related Work

We have proposed a slot value classification method in [5]. A set of slot classifiers and slot-value pair classifiers are trained on sub-dialog level. Features are extracted from sub-dialog as samples, then the frame is decomposed into new slots and slot-value pairs as labels. Support Vector Machine (SVM) is used as basic classification model. We train models by using LibSVM package [9] with linear kernel. The probabilities of slot or slot-value pair of enumerable slot are based on output of SVM model. For non-enumerable slot, the probability of slot-pair is represented by fuzzy string matching score.

To eliminate mismatch between samples and labels, an iterative alignment method (IAM) for slot value classification (SVC) parser [5] is introduced. Fig 3 shows how IAM works. As we will introduce a new iterative alignment method later, we call the original IAM hard iterative alignment method (HIAM) and the new IAM soft iterative alignment method (SIAM), which will be described in 4.3.

4.2. Cross Language Parsers Building

The baseline tracker provides two methods of fuzzy string matching between the entries in the ontology and the transcription of an utterance. Method 1 is that translated utterances from Chinese to English are matched to English entries

in original ontology. Method 2 is that the original Chinese utterances are matched to translated entries in the ontology from English to Chinese.

For SVC parser, features are extracted from sub-dialog level. Unigram or bigram feature is statistical data of words or word groups. As there are differences in expressions cross language, we can build systems using the same algorithm but in different language. For training corpus, translations for English entries are extracted and reunited as a new corpus. Two parsers can be built on Chinese corpus(cn HIAM SVC parser) and English corpus(en HIAM SVC parser) respectively.

4.3. Slot Value Classification Method with Soft Iterative Alignment

Classification models are trained on sub-dialog level but are used on utterance level when parsing. The hard iterative alignment method for SVC in DSTC4 [5] is meant to solve the problem of mismatch between training and parsing. However, once an utterance is labeled as a positive sample or a negative sample for a slot-value pair, it will hold this state as iterative time increases. The performance of HIAM SVC are roughly the same after first iterative alignment. One possible drawback of HIAM SVC is that if a sample mistakenly labeled then it will not be corrected. To make alignment method more reasonable, we come up with a soft iterative alignment method for SVC tracker.

The SIAM SVC parser works as follows:

1. Base models are trained based on sub-dialog level.
2. For each sub-dialog and frame pairs. We decompose the frame into slot-value pairs. The weights for each utterance in sub-dialog is initialized as 1. Then weights will be updated by output of slot-value classifiers. A sample united by each utterance with specific weight in sub-dialog is labeled for slot-value pairs.
3. New models are trained based on the new samples.
4. Go to step 2.

Two more trackers are built here, named Chinese soft iterative alignment (cn SIAM) SVC parser and English soft alignment iterative (en SIAM) SVC parser.

4.4. SVM Weight Parameter

For a slot or slot-value pair, the number of negative samples is far more than that of positive samples. The weight parameter of SVM classifier is then set by the ratio of samples amount. We construct a new Chinese SVC tracker and a new English SVC tracker on basis of hard IAM method, named NHIAM(Normalized HIAM) cn tracker and NHIAM en tracker respectively.

5. SLOT-BASED SCORE AVERAGING ENSEMBLE

As mentioned above, a group of dialog state trackers are constructed. Like in [5], a slot-based score averaging ensemble is employed to build an ensemble tracker combining those trackers discussed before. Ensemble methods can be very useful to improve the performance in dialog state tracking tasks.

In DSTC5 task, as for the SLU models are quite different among the individual trackers, a set of ensemble trackers are constructed based on diverse combination of the above trackers. Following the main idea in [5], all the probabilities produced by different models are normalized to scale different thresholds (T_s, T_v) to a global one, which is set to 0.5 in all ensemble trackers.

6. EXPERIMENT

Based on different SLU models, we build several trackers. Baseline trackers, HIAM SVC parsers, SIAM SVC parsers and NHIAM SVC parsers are constructed on Chinese and English language respectively.

Our evaluation results of single method system is shown in Table 1 below. The bond figure is the best one of that column. Comparing Baseline 1 , Baseline 2 and Baseline cn/en, results of cross language systems merged outperforms single system is precision, recall and F-measure metrics. Then we merge results of cross language systems with same SLU model to get a single system. The hard iterative alignment method for SVC parser has been trained on two sets of parameters for SVM classification models. Trackers of NHIAM SVC parsers have adapted changing weight parameter. Result shows that the ratio of positive and negative samples is an important factor. The result of HIAM SVC parsers and SIAM SVC parsers are quite different. The former one has better performance of recall score, while the latter one has better performance of segment accuracy and precision.

From results of single systems, we can find that HIAM SVC parser has better performance of recall metric, which SIAM SVC parser is better on accuracy and precision metrics. For NHIAM SVC parser, it shows better balance between precision and recall rate, therefore, achieves improvement in F-measure metric.

We are team 1 in DSTC5 results. Three of our entries are listed in Table 2. The bond figure is the best one of that column. Result of one language systems merged is listed as English Tracker and Chinese Tracker. Entry 0, 3, 4 are our submit results. Different ensemble systems have quite different performance. With more systems merged, the recall score gets lower. In entry 3 and entry 4, SIAM SVC parser increases segment accuracy and precision metrics on schedule 2. In entry 4 , NHIAM SVC parser improves segment accuracy and F-measure on both schedules. The results show that different single system has its own advantage in some metric. Howev-

Table 1. Experimental Results of Single Systems

Model*	Schedule 1				Schedule 2			
	Accuracy	Precision	Recall	F	Accuracy	Precision	Recall	F
Baseline1	0.0250	0.1148	0.1102	0.1124	0.0321	0.1425	0.1500	0.1462
Baseline2	0.0161	0.1743	0.1279	0.1475	0.0222	0.1979	0.1774	0.1871
Baseline	0.0238	0.1661	0.1799	0.1727	0.0346	0.1718	0.2378	0.1995
cn/en HIAM	0.0272	0.2264	**0.3305**	0.2687	0.0360	0.2313	**0.4146**	0.2970
cn/en SIAM	**0.0343**	**0.3615**	0.1712	0.2323	**0.0584**	**0.3815**	0.2236	0.2819
cn/en NHIAM	0.0284	0.2744	0.3207	**0.2957**	0.0447	0.2824	0.4061	**0.3331**

* Baseline 1 is the result of method 1 in baseline. Baseline 1 is the result of method 2 in baseline. Baseline cn/en means result of method 1 and method 2 merged. HIAM cn/en means result of hard iterative alignment method SVC parser of two language systems merged. SIAM cn/en means result of soft iterative alignment method SVC parser of two language systems merged. NHIAM cn/en means result of hard iterative alignment method SVC parser with new SVM training parameter of two language systems merged.

Table 2. Experimental Results of Ensemble Systems

Model*	Schedule 1				Schedule 2			
	Accuracy	Precision	Recall	F	Accuracy	Precision	Recall	F
English	0.0287	0.2909	**0.3036**	0.2971	0.0440	0.2998	**0.3824**	0.3361
Chinese	0.0249	0.3050	0.2404	0.2689	0.0404	0.3206	0.3088	0.3146
Entry0	0.0397	0.3320	0.2934	0.3115	0.0551	0.3429	0.3712	0.3565
Entry3	0.0387	**0.4087**	0.2436	0.3052	0.0597	**0.4260**	0.3087	0.3580
Entry4	**0.0417**	0.3650	0.2795	**0.3166**	**0.0612**	0.3811	0.3548	**0.3675**

* English tracker is the result of all trackers in English language merged. Chinese tracker if the result of all trackers in Chinese language merged. Entry 0 is result of Baseline cn/en, HIAM SVC parser of two language systems merged. Entry 3 is result of Baseline cn/en, HIAM SVC parser and SIAM SVC parser, NHIAM of two language systems merged. Entry 4 is result of Baseline cn/en, HIAM SVC parser and NHIAM of two language systems merged.

er, ensemble method achieves best performance in F measure finally.

7. CONCLUSION

This paper describes our algorithm for building dialog state trackers in DSTC5. We construct a probabilistic enhanced frame structure to represent dialog state tracking. Then several kinds of SLU models are described, including HIAM SVC parser inherited from our work in DSTC4. Based on HIAM SVC parser, we propose SIAM SVC parser. To address cross language issue, parallel language systems can be great help to overcome inaccuracy caused by machine translation. Because the ratio between negative and positive samples in training process of the SVM classifiers greatly affect the system performances, we build NHIAM SVC parser to balance different influence caused by amount of negative and positive samples. Finally, a slot-based averaging method is applied to build ensemble systems.

The performance of single systems built by our new approach shows improvement in these metrics. The evaluation results of ensemble trackers show that different combination of single systems results in different performance, while all of them outperforms the single trackers in almost all metrics, especially accuracy, precision and F-measure.

8. REFERENCES

[1] Jason D. Williams and Steve Young, "Partially observable markov decision processes for spoken dialog systems," *Computer Speech & Language*, vol. 21, no. 2, pp. 393–422, 2007.

[2] Blaise Thomson and Steve Young, "Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems," *Computer Speech & Language*, vol. 24, no. 4, pp. 562–588, 2010.

[3] Kai Sun, Lu Chen, Su Zhu, and Kai Yu, "The sjtu system for dialog state tracking challenge 2," in *Meeting of the Special Interest Group on Discourse and Dialogue*, 2014.

[4] Matthew Henderson, Blaise Thomson, and Steve Young, "Word-based dialog state tracking with recurrent neural

networks," in *Meeting of the Special Interest Group on Discourse and Dialogue*, 2014, pp. 525–539.

[5] Li M and Wu J, "The MSIIP System for Dialog State Tracking Challenge 4," in *International Workshop on Spoken Dialog Systems (IWSDS)*, 2016.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *Computer Science*, 2014.

[7] Seokhwan Kim, Luis Fernando D'Haro, Rafael E. Banchs, Jason Williams, Matthew Henderson, and Koichiro Yoshino, "The Fifth Dialog State Tracking Challenge," in *Proceedings of the 2016 IEEE Workshop on Spoken Language Technology (SLT)*, 2016.

[8] Ji Wu, Miao Li, and Chin Hui Lee, "A probabilistic framework for representing dialog systems and entropy-based dialog management through dynamic stochastic state evolution," *IEEE/ACM Transactions on Audio Speech & Language Processing*, vol. 23, no. 11, pp. 2026–2035, 2015.

[9] Chang, ChihChung, Lin, and ChihJen, "Libsvm: A library for support vector machines," *Acm Transactions on Intelligent Systems & Technology*, vol. 2, no. 3, pp. 389–396, 2011.