Maximizing SLU Performance with Minimal Training Data Using Hybrid RNN Plus Rule-based Approach

Takeshi Homma Adriano S. Arantes Maria Teresa Gonzalez Diaz Masahito Togami
Hitachi America, Ltd.
3315 Scott Boulevard, 4th Floor, Santa Clara, CA 95054, USA
takeshi.homma.ps@hitachi.com
{Adriano.Arantes, Teresa.GonzalezDiaz}@hal.hitachi.com

Abstract

Spoken language understanding (SLU) by using recurrent neural networks (RNN) achieves good performances for large training data sets, but collecting large training datasets is a challenge, especially for new voice applications. Therefore, the purpose of this study is to maximize SLU performances, especially for small training data sets. To this aim, we propose a novel CRF-based dialog act selector which chooses suitable dialog acts from outputs of RNN SLU and rule-based SLU. We evaluate the selector by using DSTC2 corpus when RNN SLU is trained by less than 1,000 training sentences. The evaluation demonstrates the selector achieves Micro F1 better than both RNN and rule-based SLUs. In addition, it shows the selector achieves better Macro F1 than RNN SLU and the same Macro F1 as rule-based SLU. Thus, we confirmed our method offers advantages in SLU performances for small training data sets.

1 Introduction

Spoken language understanding (SLU) was further researched by using rule-based methods (Bellegarda, 2013) and machine learning (ML) (Tur et al., 2010). ML achieves good SLU performances for large training data sets. However, ML-based SLU with small training data results in poor performances. Therefore, if we want to launch a new spoken dialog service as fast as possible, we cannot use ML-based SLUs as there is no time to prepare sufficient training data.

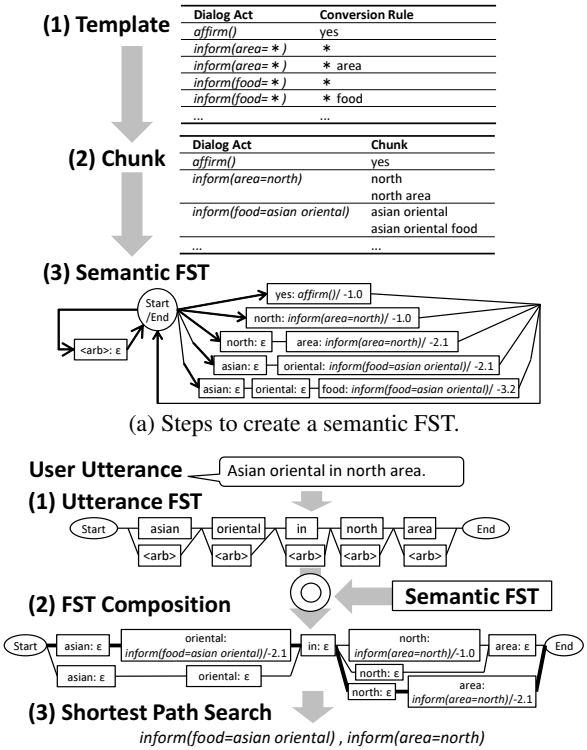
The goal of this study is to maximize SLU performances especially when the training data size is small. To achieve this objective, we propose a selection method which chooses a suitable SLU either from rule-based or ML-based SLUs depending on SLU output reliability. While researchers have studied selection methods to choose a suitable SLU result from plural SLUs by applying several algorithms (Hahn et al., 2008; Katsumaru et al., 2009; Karahan et al., 2003; Wang et al., 2002), most of them focused on selectors that improve SLU performances for large training data sets. However, their selection methods did not take into account the impact on performance for different training data sizes, specifically, how a selector would work on small training data.

Previous studies have evaluated SLU performances by metrics such as Micro F1. Nevertheless, performance evaluation by only Micro F1 is not suitable for practical dialog systems as these systems must recognize all dialog acts that users can say. In practical dialog systems, the distribution of dialog acts for actual user utterances is usually uneven. On this scenario, even if SLU completely fails to recognize some rare dialog acts, the Micro F1 remains almost unchanged and that is the main reason why systems cannot exclusively rely on this metric.

Macro F1 is another common major metric in SLU. Macro F1 computes an averaged Micro F1 of all dialog acts and decreases drastically when it fails to recognize rare dialog acts. Thus, we evaluate Macro F1 as a better metric to confirm that a selector can recognize all dialog acts.

This paper brings the following contributions to the SLU subject. First, we propose a conditional random fields (CRF) based selector which chooses suitable SLU outputs either from rule-based or ML-based SLUs. Second, we assess our selection method with different sizes of training data for recurrent neural network (RNN) based SLU. Finally, unlike most of previous studies, we evaluate SLU

Proceedings of the SIGDIAL 2018 Conference, pages 366–370,
Melbourne, Australia, 12-14 July 2018.



(a) Steps to create a semantic FST.

(b) Steps of SLU after user utterance using semantic FST.

Figure 1: Rule-based SLU by using semantic finite state transducers. <arb> is a symbol that accepts any word. ε means no dialog acts are output.

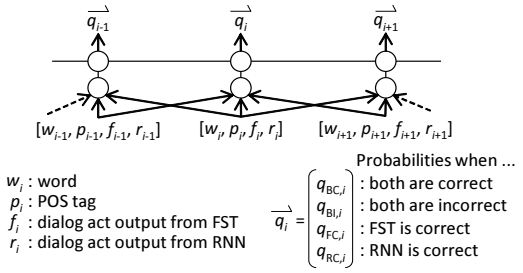


Figure 2: Model for dialog act selection.

performances by using not only Micro F1 but also Macro F1.

Experiments validate our novel approach and demonstrate that the proposed selector produces better SLU performances (up to 10.1% Micro F1 and 19.2% Macro F1) than ML-based for small training data sets and achieves "upper bound" of SLU performances regardless of training data size. This result confirms that our selector helps to improve ML-based SLU performance even if we utilize very limited training data.

2 SLU Algorithms

2.1 Rule-based Algorithm

Our rule-based SLU utilizes a SLU using finite state transducers (FST) modified from (Ferreira et al., 2015) (Figure 1). SLU developers prepare templates that convert each dialog act to chunks. Chunks are phrases that users may say when they intend to perform the dialog acts. The chunks are embedded to an FST which we call "semantic FST" (Figure 1.a). The user utterance is also converted to an utterance FST (Figure 1.b). Then, the method executes a FST composition operation (Mohri, 1997) between the utterance FST and the semantic FST. Finally, the method searches the shortest path within a composed FST. The SLU results are the dialog acts along the shortest path, i.e., a path with minimal summed weights. Based on heuristics, dialog acts generated from many words are more confident than the ones generated from just few words. Thus, the semantic FST weights are adjusted to prioritize dialog acts generated from many words.

2.2 RNN Algorithm

We used gated recurrent units (GRU) RNN cells for ML-based SLU (Mesnil et al., 2015; Zhang and Wang, 2016). Each GRU cell receives one word and POS (Part-Of-Speech) tag. We convert hidden states of a GRU to probabilities of dialog acts that the word belongs to. The algorithm selects the dialog acts with maximum probabilities from all words. The gathered dialog acts represent SLU results. In previous studies, each RNN cell outputs dialog acts with in/out/begin (IOB) tags. Our GRU cell, however, outputs dialog acts without IOB tags because this condition resulted in better accuracies in a preliminary experiment.

2.3 Selection Algorithm

Figure 2 shows a selection model that receives word and POS tag. In addition, it receives dialog acts obtained from FST and RNN generated for a corresponding word. Finally, the model outputs probabilities of 4-class judgements: both dialog acts are correct (BC), both dialog acts are incorrect (BI), FST outputs correct dialog act (FC), and RNN outputs correct dialog act (RC). We implement this model by using CRF.

Figure 3 shows a pipeline of the selection algorithm: (A) is for training of RNN SLU, (B) is for training of a selection model, and (C) is for evaluation. To obtain training data for a selection model, we first input RNN training data to FST to get FST SLU results. Besides, we do 10-fold cross validation for RNN SLU by using RNN training data to

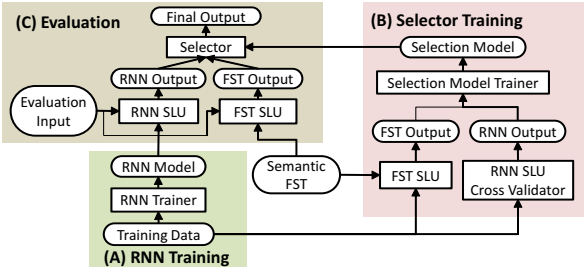


Figure 3: Pipeline for dialog act selection.

Table 1: Parameters of GRU RNN.

input	embedded word (100 dim.)
	POS tag (1-hot vector; 32 dim.)
output	dialog act probability (538 dim.)
hidden layer	bidirectional GRU
	(100 nodes, 1-layer)
context window	1
dropout rate	0.1
batch size	8

make training inputs. These SLU results are used for training of the selection model.

3 Evaluation

3.1 Dataset

We used a corpus from the Dialog State Tracking Challenge 2 (DSTC2), to evaluate our method (Henderson et al., 2014). This corpus contains transcribed sentences of user utterances for restaurant reservation dialogs. The sentences have sentence-level dialog acts. From the sentence-level dialog acts, we manually annotated word-level dialog acts. The DSTC2 corpus has a training set of 11,677 sentences, a development set of 3,934, and a test set of 9,890. From the training set, we randomly chose sentences to create training sets with various sentence sizes (100–10,000). Distribution of dialog acts in DSTC2 corpus is skewed; only 25% of dialog acts appeared in 90% of sentences for both training and test sets. The DSTC2 corpus has an "ontology" which defines all dialog acts that user may say. This ontology defines 659 dialog acts. 649 dialog acts are defined in forms of intent(slot=value), e.g., *inform(food=chinese)*, *deny(area=west)*, and *confirm(pricerange=cheap)*. Other 10 dialog acts are defined by only intent, e.g., *affirm()*, *negate()*, and *hello()*.

3.2 SLU Methods

RNN Table 1 shows the configuration of GRU for RNN SLU. The GRU receives an embedded word vector with 1-hot POS tag vector. The embedding weights are initialized with normally distributed random numbers. The hidden states of a GRU are converted to an output vector with dialog acts probabilities, by multiplying a linear matrix and softmax function. The dimension of an output vector is 538 (537 acts and "no act" class) because the largest training set (10k sentences) contains only 537 dialog acts. The hyper parameters for RNN is determined based on SLU performance in the development set. We terminate RNN training when Micro F1 on the development set is maximized.

FST We manually made 43 templates to convert dialog acts in DSTC2 ontology to 975 chunks. Figure 1.a step (1) shows template examples. When a dialog act has a value, we create chunks by embedding the value. Created chunks are converted to a semantic FST.

3.3 Selection Method

The CRF-based selector uses the following input features: word, POS tag, dialog act that FST SLU outputs, and dialog act that RNN SLU outputs. It also outputs a 4-class judgement (see Figure 2). The CRF model is trained to maximize probabilities that the selector outputs correct judgement classes. Features and hyper parameters for training CRF are determined based on selection accuracies of dialog acts in the development set. A window size for making features is set to 5. We use 3-gram features within the window. During evaluation, we choose dialog acts as follows. Assuming that the selection model outputs maximum probability in BI, we discard both dialog acts obtained from FST and RNN SLUs. Otherwise, we compare probabilities of FC and RC. For a larger FC, we adopt a dialog act output from FST SLU. In case RC is larger, we adopt a dialog act output from RNN SLU. We use CRF++ (Kudo, 2013) for training and evaluation of the selection model.

3.4 Training Data Expansion

Whitelaw et al. (2008) reported methods to increase small training data for named entity recognition by expanding them using entity dictionaries. We used the same method to increase training data for RNN by using the ontology in DSTC2. Figure 4 illustrates the method to increase training data. From one training sentence, we make additional training sentences by replacing the value of a dialog act and corresponding words with different ones. We added new sentences if the sen-

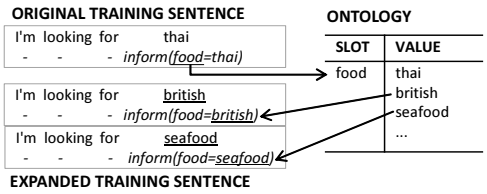
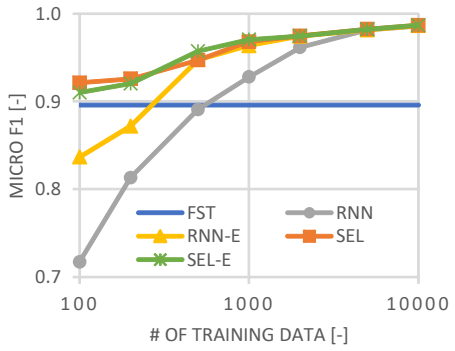
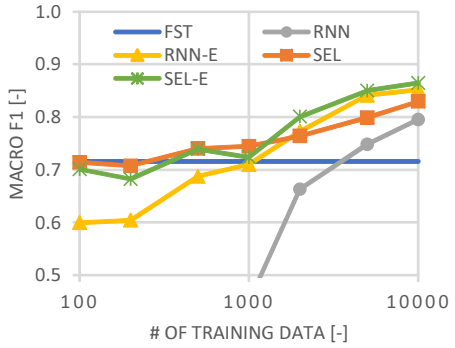


Figure 4: Expansion of training data for RNN training as baseline condition.



(a) Micro F1



(b) Macro F1

Figure 5: Evaluation results of SLU performances.

tences do not exist in the training data. By using this method, for example, we increase the training set with 100 sentences to 1.2k, and a set with 10k sentences to 67k.

Experimental conditions are as follows.

FST SLU by FST.

RNN SLU by RNN.

RNN-E SLU by RNN trained using expanded training data.

SEL Selection from FST and RNN.

SEL-E Selection from FST and RNN-E.

In SEL-E condition, RNN cross validation uses expanded sentences as training data, and non-expanded sentences as evaluation data.

3.5 Results

Figure 5 shows SLU performances. We first focus on results for small training data sets (<1k). SEL and SEL-E achieved better Micro F1 than others (Figure 5.a). Especially, when training sentences were less than 500, SEL achieved Micro F1 6.2–10.1% better than ML-based SLU (RNN-E), and 2.8–3.3% better than FST SLU. SEL also resulted in Macro F1 7.7–19.2% better than RNN-E (see Figure 5.b). Although SEL resulted in Macro F1 slightly lower than FST in some small-sized training data, the decreasing rate was at most 1.2% (FST 0.716, SEL 0.707 at 200 training sentences). SEL-E resulted in Macro F1 with the biggest decreasing rate compared to FST (4.7% at 200 training sentences). Therefore, our approach suggests that SEL is a suitable selection method to improve SLU accuracies for small training data.

Next, we focus on results for large training data sets (≥1k). SEL and SEL-E provided almost the same Micro F1 as RNN-E. Meanwhile, SEL-E achieved the best Macro F1 among all SLUs at 2k or larger training sentences. SEL-E improved Macro F1 with rates of 1.1–3.6% from RNN-E. Because SEL-E achieves the highest SLU performances, our approach suggests that SEL-E is the best selection method among the ones evaluated to improve SLU accuracies at large training data.

4 Conclusion

This work aims to improve SLU performance for small training data sets. We achieve this goal by proposing a novel CRF-based dialog act selector which chooses suitable SLU outputs either from rule-based or ML-based SLUs. Other main contributions are: novel selector method evaluation for different training data sizes; and, SLU performance assessment using Micro F1 and Macro F1. Experimental results show that our selection methods achieve up to 10.1% Micro F1 and 19.2% Macro F1 performance improvements compared to ML-based SLU for small training data. For large training data, our proposed methods outperform state-of-the-art RNN SLU methods for Macro F1 up to 3.6% while keeping Micro F1 equivalent to RNN SLU.

Consequently, our methods improve ML-based SLU performances for training data having scarce and abundant number of samples. This achievement opens up the possibility for fast launch of new spoken dialog services even with limited data available which was not possible before this work.

We also note that the best selection method is different depending on the training data size. As a follow-up paper, we will investigate selection algorithms that consistently achieve "upper bound" performances in all sizes of training data.

References

Jerome R. Bellegarda. 2013. Large-scale personal assistant technology deployment: the Siri experience. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2029–2033, Lyon, France.

Emmanuel Ferreira, Bassam Jabaian, and Fabrice Lefèvre. 2015. Zero-shot semantic parser for spoken language understanding. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1403–1407, Dresden, Germany.

Stefan Hahn, Patrick Lehnen, and Hermann Ney. 2008. System combination for spoken language understanding. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 236–239, Brisbane, Australia.

Matthew Henderson, Blaise Thomson, and Jason Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the SIGDIAL*, pages 263–272, Philadelphia, Pennsylvania, USA.

Mercan Karahan, Dilek Hakkani-Tür, Giuseppe Riccardi, and Gokhan Tur. 2003. Combining classifiers for spoken language understanding. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 589–594, St. Thomas, Virgin Islands, USA.

Masaki Katsumaru, Mikio Nakano, Kazunori Komatani, Kotaro Funakoshi, Tetsuya Ogata, and Hiroshi G. Okuno. 2009. Improving speech understanding accuracy with limited training data using multiple language models and multiple understanding models. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2735–2738, Brighton, United Kingdom.

Taku Kudo. 2013. CRF++: Yet another crf toolkit. https://taku910.github.io/crfpp. Accessed: Aug. 15, 2017.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tür, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, and Geoffrey Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.

Mehryar Mohri. 1997. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2):269–311.

Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. 2010. What is left to be understood in ATIS? In *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT)*, pages 19–24, Berkeley, California, USA.

Ye-Yi Wang, Alex Acero, Ciprian Chelba, Brendan Frey, and Leon Wong. 2002. Combination of statistical and rule-based approaches for spoken language understanding. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 609–612, Denver, Colorado, USA.

Casey Whitelaw, Alex Kehlenbeck, Nemanja Petrovic, and Lyle Ungar. 2008. Web-scale named entity recognition. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, pages 123–132, Napa, California, USA.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2993–2999, New York, USA.