A Simultaneous Recognition Framework for the Spoken Language Understanding Module of Intelligent Personal Assistant Software on Smart Phones

Changsu Lee and Youngjoong Ko
Computer Engineering, Dong-A University
840 Hadan 2-dong, Saha-gu,
Busan 604-714 Korea
{blue772001,youngjoong.ko}@gmail.com

Jungyun Seo
Computer Science, Sogang University
Sinsu-dong 1, Mapo-gu
Seoul, 121-742, Korea
seojy@ccs.sogang.ac.kr

Abstract

The intelligent personal assistant software such as the *Apple's Siri* and *Samsung's S-Voice* has been issued these days. This paper introduces a novel Spoken Language Understanding (SLU) module to predict user's intention for determining system actions of the intelligent personal assistant software. The SLU module usually consists of several connected recognition tasks on a pipeline framework, whereas the proposed SLU module simultaneously recognizes four recognition tasks on a recognition framework using Conditional Random Fields (CRF). The four tasks include named entity, speech-act, target and operation recognition. In the experiments, the new simultaneous recognition method achieves the higher performance of 4% and faster speed of about 25% than other method using a pipeline framework. By a significance test, this improvement is considered to be statistically significant as a *p-value* of smaller than 0.05.

1 Introduction

Currently, one of the most issued and promising software is the intelligent personal assistant software such as *Apple's Siri* (Wikipedia, 2011) and *Samsung's S-Voice* (Wikipedia, 2012). This kind of software provides users a natural language user interface to answer questions, make recommendations and perform actions. One of the core modules to develop this software is the Spoken Language Understanding (SLU) module. The SLU module predicts the user's intention of user utterance, and one of the various software actions is selected to provide appropriate information to a user (Wang et al., 2005).

The SLU model of the intelligent personal assistant software has several different aspects from the previous other SLU modules, such as ones of ATIS (Automatic Terminal Information Service) and DARPA (Defense Advanced Research Project Agency) projects, which are based on rule-based methods (Ward et al. 1994; Wang et al. 2001) and statistical methods (Wang et al. 2006; Raymond et al. 2007). Because the SLU module is operated for various applications (Apps) of mobile devices such as weather, transportation, etc., it has to be able to deal with more heterogeneous domains than the ATIS and DARPA projects and it does more detailed analysis for each domain in order to offer users accurate information. In addition, since the SLU module in the previous dialogue systems has a complicated architecture that is composes of many sub-modules, it is difficult for them to be directly applied into the SLU module of intelligent personal assistant software with those many domains for mobile devices. That is, building up a complicated architecture for each domain can make a heavy system and this kind of system is not proper to mobile devices.

In this paper, we propose a new SLU module with a simultaneous recognition framework for the intelligent personal assistant software. The proposed SLU module consists of four components: named entity (NE), speech-act, target and operator recognition. Each component of the proposed SLU module has different recognition unit, e.g. the named entity recognition is based on a morpheme/phrase unit, whereas the target, operator and speech-act are on an utterance unit. To integrate these recognition units into the same unit, we develop a new tag addition approach that represents a user utterance as a tag sequence for an input to CRF (Lafferty et al. 2001).

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 818–822,
Beijing, China, July 26-31, 2015.

In the experiments, the proposed simultaneous recognition module showed the better performance of 4% than a pipeline module. And it has an additional benefit that it is composed of a simple architecture with only one recognition module so it can be more efficient than other methods with respect to processing time, etc. As a result, the processing time of our system was reduced about 25% when compared to the pipeline system.

The remainder of the paper is organized as the follows. Section 2 describes related work. In the section 3, we define four components of our SLU module for the intelligent personal assistant software. Section 4 introduces our simultaneous recognition framework in detail. Section 5 explains our experimental settings and results. Finally, section 6 draws conclusions.

2 Related Work

The approaches for developing the SLU modules are largely divided into the rule-based methods and the statistical methods. The rule-based modules have typically been implemented via hand-crafted semantic level grammar rules and some robust parsers (Seneff. 1992; Ward et al. 1999). However, these semantic grammar approaches carry a high development cost and they can also lead to fragile operations since users do not typically know what grammatical constructions are supported by the system. An alternative approach is to use some statistical methods to directly map from word strings to the intended meaning structures. Statistical methods are attractive because they can be easily adapted to new conditions using only annotated training data. Statistical methods for SLU have been studied in a Hidden Vector State (HVS) Model (He et al., 2005) and a data-driven statistical models (Miller et al. 1994; Pieraccini et al. 1992; Wang et al. 2006). In addition, Jeong and Lee (2008) proposed a unified probabilistic model (triangular-chain CRF) combining the named entity and dialog-act of SLU. This method achieved the high performance for SLU. But the triangular-chain CRF has a complicated architecture with a modified CRF. And this method was built only to combine the named entity and dialog-act, whereas we need to combine four components. In practical, the triangular-chain CRF showed low performance when combining four components in the experiments. As a result, the proposed SLU module achieved high performance in spite of its simple architecture.

3 Components of the Proposed SLU Module for the Intelligent Personal Assistant Software

Since the SLU module of intelligent personal assistant software needs to determine the actions of Apps of smart phone according to user needs, they require more elaborate user intent analysis. Thus we define four components of the SLU module. An analysis result of our SLU module is shown in Figure 1 as follows:

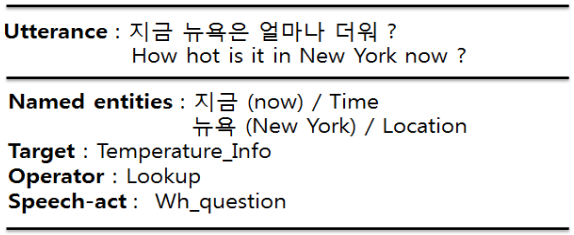

Figure 1: Example of analysis results

Named Entity (NE) recognition: NE recognition extracts keywords from user utterances, such as person, time, location, etc.

Target recognition: target describes the object of system action. In Figure 1, the target is "Temperature_Information." By this recognized target, the software can offer users accurate information.

Operator recognition: operator is to detect one of the various software actions (Lookup, Set, Delete, etc.). In Figure 1, the operator is identified as "Lookup".

Speech-act recognition: speech-act tries to designate a surface level speech-act. "Wh_Question" as speech-act in Figure 1 provides the user's intention of surface level to dialogue systems.

4 Simultaneous Recognition Framework

We assume that four components of our SLU module are correlated with each other. In order to improve the performance and speed of the SLU module, we propose a new framework to simultaneously recognize the four components. But these components have different recognition units; NE has a morpheme/phrase unit and target, operator and speech-act have an utterance unit. A new tag addition method is proposed to solve this problem. Using this method, we can construct a novel simultaneous recognition framework for SLU.

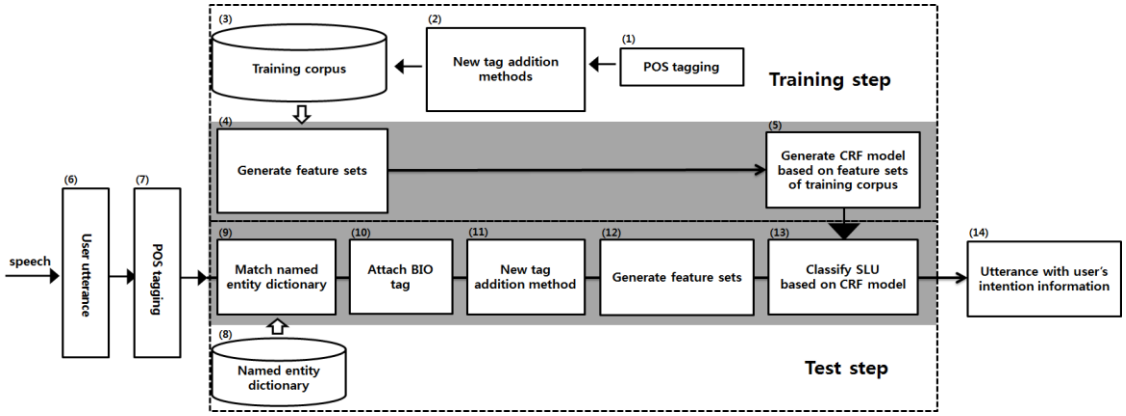

Figure 3: Architecture for the simultaneous recognition framework

4.1 New tag addition method

Target, operator and speech-act are based on an utterance unit. In order to construct a simultaneous recognition framework, we attach pseudo morphemes with target, operator and speech-act tags in front of each user utterance. Using these pseudo morphemes, target, operator and speech-act can utilize the features of NE, and NE can also do target, operator and speech-act information as additional features. Figure 2 shows an example of the new tag addition method.

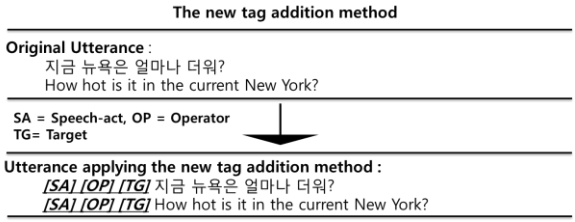

Figure 2: Example of the new tag addition method

4.2 Simultaneous recognition framework

On the simultaneous recognition framework with the new tag addition method, an input utterance is analyzed by a sequential labeling classifier, CRF. It is possible to use all of component labels as additional features in this classification method. We think that this is a main reason why the proposed method improves recognition performances.

Our framework needs only NE dictionary and BIO annotated training corpus; BIO tags were used in (Ramshaw and Marcus. 1995). It is very simple and fast because it can output all of four different SLU results in one classification execution. The architecture of our framework is shown in Figure 3. Our SLU module is widely divided into a training step and a test step.

4.3 Feature Sets

The three feature sets are extracted for SLU: basis features (Lee et al. 2010), NE dictionary features and target/operator/speech-act features. All the basic and NE dictionary features are analyzed based on the morpheme unit.

- **Basis features**

Current lexicon/POS tag information
Based on the position of the current lexicon, lexicon contextual information. window size : -2~2
Based on the position of the current POS tag, POS tag contextual information. window size : -2~2
The words of Korean language can consist of one or more morphemes; - current morpheme position information in a word - current morpheme POS tag/word length information

- **NE dictionary features**

Based on current morpheme, NE tag information matched from NE dictionary

- **Target/operator/speech-act features**

Verb information in the utterance
Lexicon unigram information in the utterance
Lexicon & POS tag bigram information in the utterance
Lexicon & POS tag trigram information in the utterance

5 Experiments

5.1 Experimental settings

The MADS data set (Multi-Applications Dialogues for Smart phones) was constructed and used to develop the SLU modules for the intelligent personal assistant software. The MADS data set was annotated by 8 NEs, 28 targets, 5 operators and 6 speech-act tags. In addition, The MADS data set consists of 1,925 user utterance in 6 domains: *weather, clock, alarm, schedule, exchange and traffic*. The Mallet toolkit was chosen for our CRF model (McCallum. 2002).

All experiments were evaluated by accuracy in the utterance level. When the proposed SLU module generates all the correct labels of NE, target, operator and speech-act of an input utterance, the utterance is considered as correct. The performance of the SLU module is averaged on 5-fold cross validation. In addition, we used the paired *t*-test and Wilcoxon singed rank test to

verify statistically significant between our framework and compared baseline framework. The pipeline framework (Moreira et al., 2011) is used a baseline system in our experiments because it is the most common method for multi-domains SLU module.

5.2 Experimental results

Each component of the SLU module is first evaluated by comparison of accuracies between the proposed and baseline frameworks. Figure 4 illustrates the accuracies of each component.

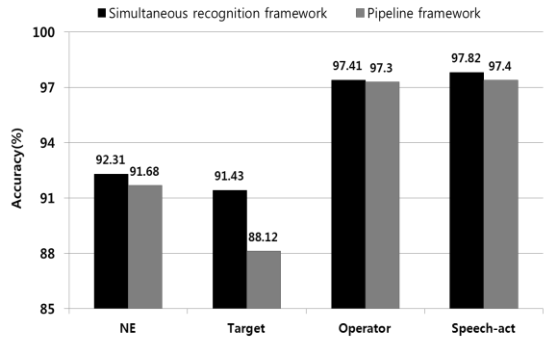

Figure 4: Comparison of the accuracies of each component for SLU

A pipeline framework commonly has some disadvantage that the errors of previous component are propagated to the next components. It can cause a cascade of performance degradation.

Figure 5 shows the accuracies of entire SLU modules in an utterance level.

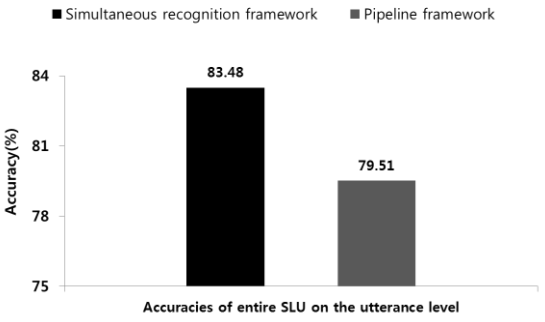

Figure 5: Comparison of accuracies of entire SLU modules on the utterance level

The proposed framework achieved significantly better performance than the baseline framework.

To verify statistically significant on accuracy difference between the proposed and baseline frameworks, we performed significant test using the *t*-test and Wilcoxon singed rank test (Demsar. 2006). Table 1 shows the results of significant test.

p-value < 0.05 (95%)	Our framework *vs.* Pipeline framework
paired t-test	0.00001
Wilcox signed rank test	0.021

Table 1: Results of significant tests

In both of two significance tests, our framework was statistically significantly better than the pipeline framework ($p<0.05$).

In the comparison of processing time, our framework obtained faster processing speed than pipeline framework with about 25% reduction.

Test user utterance (388 utterances)	
Our framework	15 sec.
Pipeline framework	19 sec.

Table 2: Results of processing time comparison

In addition, we tried to compare our module and the triangular-chain CRF (Jeong and Lee, 2008). Table 3 shows the performances when NE and speech-act recognition tasks are combined and all four recognition tasks are combined. As a result, our module outperformed the triangular-chain CRF in both of cases.

	NE+Speech-act	All (four tasks)
Our framework	90.61	83.48
Triangular-chain CRF	87.07	16.4

Table 3: comparison of our module and triangular-chain CRF

6 Conclusions

In this paper, we have presented a novel SLU framework to predict user's intention for determining system actions of the intelligent personal assistant software. The proposed SLU module with a simultaneous recognition framework achieved higher performance and faster processing speed than the existing pipeline system. In addition, our module outperformed other method, the triangular-chain CRF, especially when four components were all analyzed.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. NRF-2013R1A1A2009937)

References

Janez Demsar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, Vol. 7. pp.1–30.*

Yulan He and Steve Young. 2005. Semantic Processing using the Hidden Vector State Model. *Computer Speech and Language, Vol. 19, No. 1, pp. 85-106.*

Minwoo Jeong and Gary-Geunbae Lee, 2008. Triangular-chain conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 16, pp. 1287-1302.

John Lafferty, Andrew McCallum and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *In Proceedings of the Eighteenth International Conference on Machine Learning." Morgan Kaufmann Publishers Inc., SanFrancisco, CA, USA, pp. 282-289.*

Changki Lee and Myung-Gil Jang. 2010. Named Entity Recognition with Structural SVMs and Pegasos algorithm. *Korean Journal of Cognitive Science*. Vol. 21. No. 4, 655-667.

Andrew McCallum. 2002. Mallet: A machine learning for language kit, http://mallet.cs.umass.edu.

Scott Miller, Revert Bobrow, Robert Ingria, and Robert Schwartz. 1994. Hidden understanding models of natural language. *In Proceedings of the ACL, Association for Computational Linguistics, pp. 25–32.*

Catarina Moreira, Ana Cristina Mendes, Lu´ısa Coheur and Bruno Martins, 2011. Towards the rapid development of a natural language understanding module. *In Proceedings of 10th international conference on Intelligent virtual agents, pp. 309–315.*

Roberto Pieraccini, Evelyne Tzoukermann, Zakhar Gorelov, Jean-Luc Gauvain, Esther Levin, Chine-Hui Lee and Jay G. Wilpon. 1992. A speech understanding system based on statistical representation of semantics. *In Proceedings of the ICASSP, San Francisco, CA.*

Launce A. Ramshaw and Mitchell P. Marcus. 1995. Text Chunking using Transformation-Based Learning. *In Proceedings of the Third Workshop on Very Large Corpora*, pp. 82-94.

Christian Raymond and Giuseppe Riccardi. 2007. Generative and discriminative algorithms for spoken language understanding. *In Proceedings of the Interspeech, Antwerp, Belgium.*

Stephanie Seneff. 1992. TINA: A Natural Language System for Spoken Language Applications. *Computational Linguistics*.

Ye-Yi Wang. 2001. Robust Spoken Language Understanding in MiPad. *In proceedings of Eurospeech, Aalborg, Denmark.*

Ye-Yi Wang and Alex Acero. 2006. Discriminative models for spoken language understanding. *In Proceedings of the ICSLP, Pittsburgh, PA.*

Ye-Yi Wang, Li Deng and Alex Acero. 2005. Spoken language understanding : an introduction to the statistical framework. *IEEE Signal Processing Magazine 22(5): 16-31.*

Wayne Ward, Bryan Pellom, and Sameer Pradhan. 1999. The CU Communicator System, *IEEE Workshop on ASRU Proc., Keystone, Colorado.*

Wayne Ward and Sunil lssar. 1994. Recent Improvements in the CMU Spoken language Understanding System. *in Human Language Technology Workshop, Plainsboro, New Jersey.*

Wikipedia Contributors. 2011. Siri, Wikipedia, the Free Encyclopedia.

Wikipedia Contributors. 2012. S-Voice, Wikipedia, the Free Encyclopedia.