A Multiple Classifier-based Concept-Spotting Approach for Robust Spoken Language Understanding

Jihyun Eun, Minwoo Jeong, Gary Geunbae Lee

Department of Computer Science and Engineering
Pohang University of Science and Technology, Korea (South)
{tigger,stardust,gblee}@postech.ac.kr

Abstract

In this paper, we present a concept spotting approach using manifold machine learning techniques for robust spoken language understanding. The goal of this approach is to find proper values for pre-defined slots of given meaning representation. Especially we propose a voting-based selection using multiple classifiers for robust spoken language understanding. This approach proposes no full level of language understanding but partial understanding because the method is only interested in the pre-defined meaning representation slots. In spite of this partial understanding, we can acquire necessary information to make interesting applications from the slot values because the slots are properly designed for specific domain-oriented understanding tasks. In several experimental results, the SLU (Spoken Language Understanding) performance degradation of spoken inputs compared with textual inputs are only F-measure 10.72, 11.43 and 11.51 for speech act, main goal and component slot extraction task respectively although the WER of spoken inputs is as high as 18.71%. That is, the evaluation results show that our concept spotting approach for SLU system is especially robust for spoken language input which has large recognition errors.

1. Introduction

Understanding spoken language has been long studied extensively [1] [2] [3] [4] [5]. Language understanding systems that use a large set of rules to explain the syntactic and semantic possibilities for spoken sentences suffer from a lack of robustness when faced with the wide variety of ill-formed spoken sentences that people really use. Moreover, this approach is much more complicated because of abundant ambiguities in natural languages, grammatically incorrect sentences, and very noisy speaking environments. The output from the speech recognizer is likely to contain misrecognized words as well as features of spontaneous speech such as filled pauses, restarts, repetitions and repairs.

To overcome these speech recognition limitations, we attempt to understand spoken languages by concept spotting approach which is aimed to extract only essential factors for predefined meaning representations. This approach for spoken language understanding proposes no full level of language understanding but partial understanding because the method is only interested in the pre-defined meaning representation slots. In spite of this partial understanding, we can acquire necessary information to make many interesting applications from the slot values because the slots are properly designed for domain-oriented language understanding tasks. Basically, we use manifold machine learning techniques to map directly from sentences to the intended meaning structures. Especially, we propose a voting-based selection from multiple classifiers for robust spoken language understanding.

The remainder of this paper is organized as follows. Section 2 reviews some related works for spoken language understanding to compare with our new concept spotting approach. Section 3 describes the detail methodology used in our system. We describe our experimental results in section 4. Finally, we present our conclusions in section 5.

2. Related Works

The handling methods for spoken language understanding are largely divided into rule-based methods and statistical methods. As previously stated, language understanding systems that use a large set of rules can suffer from a lack of robustness especially when faced with a wide variety of spoken sentences. One reason for this frailty is that, for most limited domains, a traditional syntactic explanation of a sentence is often much out focused than the direct explanation of the meaning of the sentences in terms of the words spoken and the relations between the words. So, these systems have typically been implemented via hand-crafted semantic level grammar rules and some form of robust parsers [1] [2]. However, this semantic grammar approach carries a high development cost and it 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 map directly from word strings to the intended meaning structures. In this approach, hand-crafted grammars and rules are replaced by statistical models that are automatically learned from some training data. Statistical methods are attractive because they can be adapted to new conditions (tasks or languages) if an appropriate training data is available. Statistical methods for spoken language understanding have already been investigated in the AT&T's CHRONUS [3], the BBN's Hidden Understanding Model (HUM) [4] systems and the Cambridge Univ.'s Hidden Vector State (HVS) Model [5]. These models were primarily motivated by the single technique that has been extremely successful in speech recognition and natural language processing: especially Hidden Markov Models (HMMs).

In this paper, we propose to use manifold machine learning techniques such as the maximum entropy (ME) model, support vector machine (SVM), neural network (NN), and conditional random fields (CRF) which have been successfully applied to pattern classification and sequence labeling tasks for our own concept spotting approach in spoken language understanding.

3. Concept Spotting Approach for Spoken Language Understanding

The goal of the language understanding components is to analyze the output of the speech recognition components and to assign a meaning representation that can be used by the dialogue manager [6]. In this section, first we describe the meaning representation defined in our SLU system and second, how to extract the meaning representation from given utterances through classifiers and a slot extractor. Figure 1 illustrates the structure of the concept spotting SLU system.

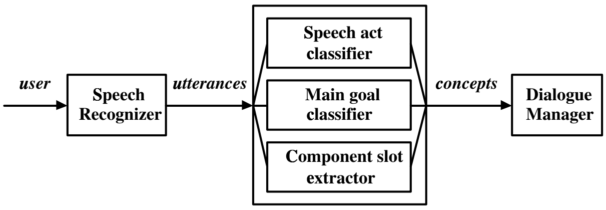

Figure 1: The structure of the concept spotting SLU system.

3.1. Meaning Representation

In the concept spotting approach, essential factors for spoken language understanding are extracted based on the predefined meaning representation. So, the representation should be solidly defined to express the intentions of users as fully as possible but not too much complex to be managed in SLU system.

Basically, our concept spotting approach is originated from the idea of information extraction (IE). The goal of IE task is to build systems that find and link relevant information while ignoring extraneous and irrelevant information. The desired knowledge in IE task can be described by a relatively simple and fixed template with slots that need to be filled [7] [8]. These characters of IE are well suited for natural language processing tasks such as POS tagging and partial parsing.

We design our concept spotting approach which attempts to understand spoken languages by extracting essential factors for pre-defined meaning representations in specific domain-oriented SLU tasks. This approach combines SLU which is one of NLP tasks with the IE concept.

Considering these aspects, we construct the meaning representation, such as the example shown in Table 1, consisting of a speech act slot, a main goal slot, and several component slots for each main goal for a sentence. We treat the tasks about speech act and main goal slots as classification tasks. The value of a speech act slot is assigned from one of the classes which designate the surface-level speech acts, such as, *yn_question*, *wh_question*, *request*, etc. per single sentence. Similarly to the speech act slot, the value of the main goal slot is assigned from one of the classes which classify the main application actions[1] in a specific domain, for instance, in a telebanking service domain, such as, *confirm_qualified*, *search_info*, *confirm_info*, etc. The tasks for the component slots, such as, *info*, *period*, *rate* in the telebanking service domain are viewed as sequence labeling tasks.

input	*What are the popular savings nowadays ?*
speech act slot	wh_question
main goal slot	search_info
component slots	info (popular savings) period (nowadays)

Table 1: An example of a meaning representation.

3.2. Speech Act and Main Goal Classifiers

As previously stated, the tasks about speech act and main goal slots are treated as classification tasks. We separately adjust several classifiers based on ME, SVM, NN to these tasks. The performances of the classifiers are dropped when erroneously recognized utterances are given as input for classification. So we propose to combine these classifiers using a voting method for better performance and robust spoken language understanding. Now, each classifier and how to vote among multiple classifiers are described.

3.2.1. *Maximum Entropy Classifier*

ME classifier is based on the idea that the most uniform distribution among the probability distributions that satisfy the given constraints is the ideal. This classifier offers a clean way to combine diverse pieces of contextual evidence in order to estimate the probability of a certain class occurring with a certain linguistic context [9]. Also, ME classifier is good for integrating information from many heterogeneous information sources.

3.2.2. *Support Vector Machine Classifier*

SVM offers the possibility to train generalizable, nonlinear classifiers in high dimensional spaces using a small training set. SVM generalization error is not related to the input dimensionality of the problem but to the margin with which it separates the data [10]. Although SVM is generally a binary classifier, we use an SVM version which is modified for multiple classes to apply to speech act and main goal classification.

3.2.3. *Neural Network Classifier*

Neural network learning provides a practical method for learning real-valued and vector-valued functions in a way that is robust to noise in training data. In a rough analogy, neural networks are built out of a densely interconnected set of simple units, where each unit takes a number of real-valued inputs and produces a single real-valued output [11].

3.2.4. *Voting Classifiers*

Many of the simple techniques that aim to combine multiple evidences into a single prediction are based on voting. One of the several voting methods is simple voting. That is, based on the predictions of different base classifiers, a final prediction is chosen as the classification with a plurality of votes [12]. In our case, if the three classifiers disagreed with each prediction, the prediction of the classifier that has the best performance individually should receive the vote. The structure of voting classifiers is shown in Figure 2.

3.3. Component Slot Extractor

The tasks handling component slots are treated as sequence labeling tasks and we apply CRF to these tasks. We used a

[1] Currently, we assume that a single utterance only contains a single main action. If we want to treat multiple actions in a single utterance, we have to perform the simple sentence segmentation first before applying a concept spotting SLU approach.

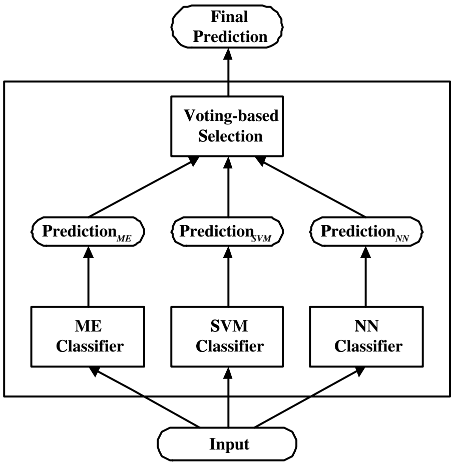

Figure 2: The structure of voting classifiers.

linear-chain CRF model, which is a model that assigns a joint probability distribution over labels conditional on the input sequences, where the distribution respects the independence relations encoded in a graph [13]. An example of sequence labeling for the sentence in the telebanking service domain is shown in Table 2. We make use of BIO representation [14] in which INFO-B, INFO-I, O stand for begin, inside, and outside of a label.

input	*What are the popular savings nowadays ?*
labeling	[O *What*] [O *are*] [O *the*] [INFO-B *popular*] [INFO-I *savings*] [PERIOD-B *nowadays*]

Table 2: An example of a sequence labeling.

4. Experiments

We have implemented a spoken language understanding system based on the concept spotting approach and have conducted extensive experiments using a telebanking service domain corpus as experimental data.

4.1. Speech Act and Main Goal Classifiers

For evaluating speech act and main goal classifiers, we used 10-fold cross validation using the telebanking service domain corpus that consists of 2239 fully-annotated Korean sentences. There are two types of this corpus: the textual sentences and the spoken sentences (speech recognizer output) by an HTK-based Korean speech recognizer (morpheme-based recognition system). Its performance is Word Error Rate (WER) of 18.71% in this domain. Our concept spotting SLU system was trained on textual sentences and tested on both textual sentences and spoken sentences. ME, SVM, and NN classifiers are created using Zhang's Maxent toolkit [15], LIBSVM [16], and the nnet toolbox of MATLAB® [17] respectively.

The number of classes for the speech act slot is 5 and the performances of each classifier are F-measure 97.63, 98.66, 97.95 for the textual input and 85.75, 85.41, 86.61 for the spoken input in Table 3. The results in this table show that the performances of ME, SVM and NN classifiers for speech act classification decrease as F-measure 11.88, 13.25, 11.34 when testing with the spoken input (WER 18.71%). The performance of the combined classifier more endures by decreasing only 10.72 in F-measure.

	Textual input (WER 0%)	*Spoken input (WER 18.71%)*
ME classifier	97.63	85.75
SVM classifier	98.66	85.41
NN classifier	97.95	86.61
Combined classifier	98.30	87.58

Table 3: Performances of the speech act classifiers.

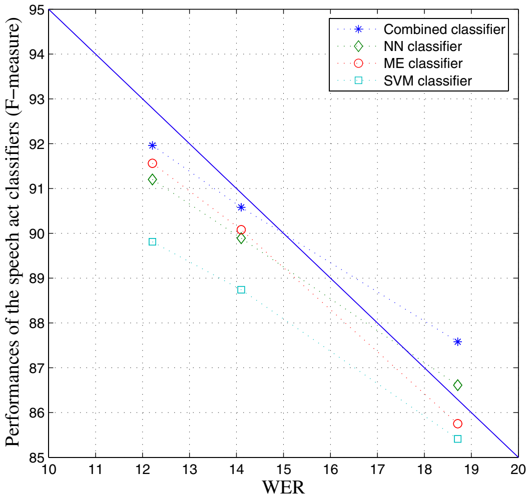

Figure 3: The performance trend of the speech act classifiers.

The similar performance trend can be confirmed for the main goal recognition task. In this case, the number of classes for the main goal slot is 25 and the performances of each classifier are F-measure 95.80, 96.38, 94.16 for the textual input and 82.76, 81.24, 79.33 for the spoken input in Table 4. The results in this table show that the performances of the classifiers for main goal classification decrease as F-measure 13.04, 15.14, 14.83 when testing with the spoken input. As in the previous case, the performance of the combined classifier endures by decreasing only 11.43.

	Textual input (WER 0%)	*Spoken input (WER 18.71%)*
ME classifier	95.80	82.76
SVM classifier	96.38	81.24
NN classifier	94.16	79.33
Combined classifier	96.56	85.13

Table 4: Performances of the main goal classifiers.

As the results of speech act and main goal recognition tasks showed, the decreased understanding performances when testing with the several different error level spoken inputs are generally less than the increased rate of WER. Figure 3 and 4 show

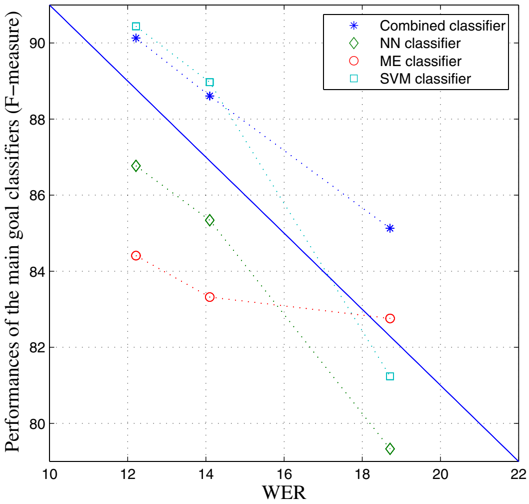

Figure 4: The performance trend of the main goal classifiers.

these trends by comparing the slow-decreased performances of multiple classifiers (especially, combined classifiers) with the steep-decreased speech recognition performances[2]. So, the experimental results show that our concept spotting SLU system is robust for spoken language input that has large errors.

4.2. Component Slot Extractor

We used the previous telebanking service domain corpus for evaluating our CRF-based component slot extractor with 10-fold cross validation. The results in Table 5 show that the concept spotting SLU performance difference of two inputs is only F-measure 11.51 although WER of spoken input is 18.71%. In the same manner with the speech act and main goal recognition tasks, the decreased rate of the performance of the component slot extraction task with the spoken input is less than the increased WER rate.

WER of Input	*Precision*	*Recall*	*F-measure*
0%(textual input)	93.55	92.05	92.79
12.21%	87.72	83.60	85.61
14.10%	85.57	81.16	83.31
18.71%	84.23	78.54	81.28

Table 5: Performances of the component slot extractor.

5. Conclusions

This paper proposed a concept spotting approach as a new spoken language understanding method using manifold machine learning techniques especially for highly erroneous speech recognition environments. We applied several techniques such as ME, SVM, NN, and CRF to the concept spotting SLU as a method of pattern classification and sequence labeling. We showed that the concept spotting approach can be successfully used for language understanding by robustly finding the values of the pre-defined meaning representation of the sentence templates.

In several experimental results, the SLU performance degradation of spoken inputs compared with textual inputs are only F-measure 10.72, 11.43 and 11.51 for speech act, main goal and component slot extraction task respectively although the WER of spoken inputs is as high as 18.71%. So, the evaluation results on the telebanking service domain corpus show that our concept spotting approach for SLU is robust for the spoken language input that is prone to large errors.

Currently, our system is based on the standard word-based features like lexical and part-of-speech tag features in Korean understanding. We are now planning to generate more effective features through robust high-level processing of natural languages such as syntactic and semantic structures.

6. Acknowledgements

This research was supported as a Brain Neuroinformatics Research Program sponsored by Ministry of Commerce, Industry and Energy.

7. References

[1] Seneff, S., ”TINA: A Natural Language System for Spoken Language Applications”, Computational Linguistics, 1992, 18.1, pp. 61-86.

[2] Ward, W. and Pellom, B., ”The CU Communicator System”, IEEE Workshop on ASRU Proc., Keystone, Colorado. 1999.

[3] Levin, E. and Pieraccini, R., ”CHRONUS, the next generation”, The DARPA Speech and Natural Language Workshop Proc., 1995, pp. 269-271.

[4] Miller, S., Bates, M., Bobrow, R. Ingria, R., Makhoul, J., and Schwartz, R., ”Recent progress in hidden understanding models”, The DARPA Speech and Natural Language Workshop Proc., 1995, pp. 276-280.

[5] He, Y. and Young, S., ”Semantic Processing using the Hidden Vector State Model”, Computer Speech and Language, Vol. 19, No. 1, pp. 85-106, 2005.

[6] McTear, M. F., Spoken Dialogue Technology, Springer, 2004.

[7] Cowie, J. and Lehnert, W., ”Information extraction”, Communications of the ACM, Vol. 39, Issue 1, pp. 80-91,1996.

[8] Jurafsky, D. and Martin, J. H., Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall, 2000.

[9] Ratnaparkhi, A., ”Maximum Entropy Models for Natural Language Ambiguity Resolution”, Ph.D. Dissertation, University of Pennsylvania, 1998.

[10] Burges, C. J. C., ”A tutorial on Support Vector Machines for pattern recognition”, Data Mining and Knowledge Discovery, 2(2): pp. 121-167, 1998.

[11] Mitchell, T., Machine Learning, McGraw Hill, 1997.

[12] Chan, P. K. and Stolfo, S. J., ”A Comparative Evaluation of Voting and Meta-learning on Partitioned Data”, ICML Proc., pp. 90-98, 1995.

[13] Lafferty, J., McCallum, A. and Pereira, F., ”Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, ICML Proc., 2001.

[14] Ramshaw, L. A. and Marcus, M. P., ”Text chunking using transformation-based learning”, The Third Workshop on Very Large Corpora Proc., pp. 82-94, 1995.

[15] Maxent Toolkit, `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html`

[16] LIBSVM, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[17] MATLAB®, `http://www.mathworks.com`

[2] The solid line designates linear decrease (45 ° slope) of F-measure along with WER increase.