Tightly Integrated Spoken Language Understanding using Word-to-Concept Translation

Katsuhito Sudoh, Hajime Tsukada

NTT Communication Science Labs. NTT Corporation
2-4 Hikaridai, Keihanna Science City, Kyoto 619-0237, Japan
{sudoh,tsukada}@cslab.kecl.ntt.co.jp

Abstract

This paper discusses an integrated spoken language understanding method using a statistical translation model from words to semantic concepts. The translation model is an N-gram-based model that can easily be integrated with speech recognition. It can be trained using annotated corpora where only sentence-level alignments between words and concepts are available, by automatic alignment based on cooccurrence between words and concepts. It can reduce the effort for explicitly aligning words to the corresponding concept. The method determines the confidence of understanding hypotheses for rejection in a similar manner to word-posterior-based confidence scoring in speech recognition. Experimental results show the advantages of integration over a cascaded method of speech recognition and word-to-concept translation in spoken language understanding with confidence-based rejection.

1. Introduction

The progress of continuous speech recognition technology has led to the development of various speech interfaces, such as spoken dialogue systems. Spoken language understanding (SLU), which is composed of automatic speech recognition (ASR) and natural language understanding (NLU), plays an important role in spoken dialogue systems.

These systems often employ frame-based semantic representation that consists of slot-value pairs [1, 2]. We define a concept as such a slot-value pair. By this definition, the semantic representation of a user utterance is regarded as a set of concepts. In this paper, we consider SLU as mapping from spoken utterances to sets of concepts. In previous works, SLU was modeled by a combination of ASR and NLU mediated by 1-best or N-best word sequences, and many rule-based (e.g., [3, 4]), corpus-based (e.g., [1, 5, 6]), and hybrid methods (e.g., [7, 8]) were proposed. These methods assume that explicit correspondences between words and concepts are available. They represent which words correspond to each concept and are given by hand-written rules or annotated corpora with such information. However, it is expensive to consistently determine the correspondence because aligning words to concepts in a sentence is ambiguous, which is complicated by spontaneous speech phenomena. Therefore, we propose a statistical NLU model that can be trained using loose correspondence between pairs of a word sequence and a set of concepts associated at the sentence level.

In addition, NLU in spoken dialogue systems have to reject erroneous concept hypotheses as well as finding the most likely ones. For rejection, a confidence measure of concepts is necessary. We can choose confident concepts whose confidence exceeds a certain threshold. To define the confidence measure of the concept, it is necessary to consider both its ASR- and NLU-level appropriateness as dominance among a number of hypotheses in a integrated search of ASR and NLU [9]. However, there are few works considering both the alignment problem and the concept-level confidence.

Motivated by these issues, this paper proposes a tightly integrated SLU method using a statistical translation model from words to concepts. The translation model can be trained using the corpus with simplified concept annotation, where each sentence is aligned to a set of concepts but each concept is not aligned explicitly to the corresponding words in the sentence. Here, alignments between words and concepts can be automatically obtained based on those cooccurrence, so that we can reduce the effort for alignment. The model is an N-gram-based joint probability model that can easily be integrated with existing N-gram-based ASR engines. The confidence of SLU hypotheses can be obtained as posterior probabilities in integrated searches, in a similar manner to word-level ASR confidence scoring [10].

The remainder of this paper is organized as follows. Section 2 reviews related studies. Section 3 describes our definition of concepts. Section 4 formulates the SLU problem. Section 5 explains our translation model and concept-level confidence scoring. Section 6 describes automatic alignment between words and concepts in the corpus. Section 7 presents experimental results. Concluding remarks are made in section 8.

2. Related Work

Statistical NLU models that can be trained with sentence-level aligned corpora based on statistical machine translation (SMT) have been proposed [11,12]. By these methods we can simplify corpus annotation. They assume sentence-level alignment of parallel corpora, and SMT models are trained using the alignment between words and concepts that can be obtained automatically. We extend this approach to SLU and propose a tightly integrated method to define SLU-level confidence by considering overall scores for ASR and NLU across all possible hypotheses.

There are other approaches for integrated SLU with classifier-based NLU models trained using sentence-level aligned corpora [13–15]. These methods employ text classification models, such as semantic classification trees and boosting-based classifiers. Here, the confidence of a SLU hypothesis is obtained as the output score of the classifier, incorporating word-level confidence in ASR. Compared to these methods, our method enables efficient decoding for SLU using existing ASR engines. Furthermore, these methods do not explicitly model the cooccurrence of concepts in user utterances. In contrast, our
method has the advantage of being able to model such cooccurrence by incorporating concept-level contexts in the form of an N-gram.

3. Definition of Concepts

We employ Hierarchical Concept Keys [2] as the representation of concepts. Each concept is represented by a slot-value pair connected by ‘=’ (e.g., “hour=(10)”), or a pair of slot and a set of concepts (e.g., “arrivetime={hour=(10), minute=(30)}”). For example, the semantics of user utterance, “I’d like to go to Tokyo station at ten thirty,” are represented by “to=(station=(Tokyo)), arrivetime=(hour=(10), minute=(30))”. We also define several concepts required by dialogue systems, such as for ‘Yes’ and ‘No’. These concept representations are considered sufficient to interpret user utterances as a set for use in frame-based spoken dialogue systems.

4. Statistical Spoken Language Understanding

Naive SLU modeling considers ASR and NLU separately. Given speech input \( X \), the ASR process finds word sequence \( \hat{W} \) that maximizes posterior probability:

\[
\hat{W} = \underset{W}{\operatorname{argmax}} \, P(W|X) = \underset{W}{\operatorname{argmax}} \, P(X|W)P(W).
\]

In this equation, \( P(X|W) \) represents the acoustic model, and \( P(W) \) represents the language model. Language model probability is approximated by N-gram probabilities as follows:

\[
P(W) = \prod_{i} P(w_i|w_{i-1}, ..., w_{i-N+1}) \\
\approx \prod_{i} P(w_i|w_{i-N+1}, ..., w_{i-1}).
\]

After finding word sequence \( \hat{W} \) by (1), the NLU process finds a set of concepts \( \hat{C} \) that maximizes the following posterior probability:

\[
\hat{C} = \underset{C}{\operatorname{argmax}} \, P(C|\hat{W}) = \underset{C}{\operatorname{argmax}} \, P(\hat{W}|C)P(C).
\]

In contrast to a SLU using the two cascaded processes denoted by (1) and (3), the proposed method directly finds \( \hat{C} \) from \( X \) by integrating ASR (1) and NLU (3):

\[
\hat{C} = \underset{C}{\operatorname{argmax}} \, P(C|X) = \underset{C}{\operatorname{argmax}} \, P(X|C)P(C) \\
= \underset{C}{\operatorname{argmax}} \sum_{W} P(X|W,C)P(W|C)P(C) \\
\approx \underset{C}{\operatorname{argmax}} \, \frac{P(X|W,C)P(W|C)P(C)}{P(W)} \\
= \underset{C}{\operatorname{argmax}} \, \frac{P(X|W)P(W|C)}{P(W)}. \quad (X \text{ is independent of } C + \text{Viterbi approximation for } W)
\]

5. Word-to-Concept Translation based on N-gram

To model joint probability \( P(W,C) \) in equation (4), we propose a method inspired by SMT. We model \( P(W,C) \) by N-gram probability assuming alignments between words and concepts, as the modeling in [16].
get the corresponding ASR hypothesis by extracting the word sequence from $S$.

For determining the confidence of the SLU hypotheses, we use the confidence of $s_1, \ldots, s_K$ obtained in a similar manner to word level ASR confidence scoring [10]. The confidence of $s_k = (w_k, c_k)$ $(0 \leq k \leq K)$, which appeared in time interval $[t_{k-1}, t_k]$, $C([s_k, t_{k-1}, t_k])$ is denoted as follows:

$$ C([s_k, t_{k-1}, t_k]) = \sum_{S' \in S_{Sat}, S' \text{includes } [s_k, t_{k-1}, t_k]} \frac{p(X|S') p(S')}{p(X)}, $$

where $X$ is the input speech signal and $S_{Sat}$ is the set of all possible hypotheses for $S$. Although the confidence of $c_k$ appeared in $[t_{k-1}, t_k]$ $C([c_k, t_{k-2}, t_k])$ is the sum of $C([s_k', t_{k-1}, t_k])$ for all $s_k'$ including $c_k$, we assume $C([c_k, t_{k-1}, t_k])$ to be equal to $C([s_k, t_{k-1}, t_k])$ for simplification.

6. Automatic Word-Concept Alignment

To determine the $N$-gram probabilities of the word-to-concept translation model, alignments between words and concepts over a training corpus are required. The training corpus is assumed to be sentence-level aligned where each word sequence is aligned to a set of concepts obtained through transcription and annotation to spoken utterances. It is also assumed that each concept is not aligned explicitly to the corresponding subsequence in the word sequence.

We employ automatic alignment between word sequence $W$ and a set of concepts $C$. We define the likelihood of alignment as the geometric mean of a cooccurrence measure between the aligned word and concept. We use $\phi^2$ [18] as the cooccurrence measure between word $w$ and concept $c$, denoted as follows:

$$ \phi^2(w, c) = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)} $$

$$ a = freq(w, c) $$

$$ b = freq(w) - freq(w, c) $$

$$ c = freq(c) - freq(w, c) $$

$$ d = N - a - b - c, $$

where $freq()$ denotes the number of sentences in the corpus that contain word $w$, concept $c$, or both, and $N$ is the total number of sentences in the corpus, $\phi^2(w, eps)$ is defined as 1. Other cooccurrence measures such as mutual information can be considered, however, $\phi^2$ showed the best results in our pilot test.

Therefore, using $\phi^2$, the most likely alignment $\hat{a}$ can be determined as follows:

$$ \hat{a} = \text{argmax}_a \left[ \prod_{i=1}^{t} \phi^2(w_i, c_{a_i}) \right] $$

$$ = \text{argmax}_a \left[ \prod_{i=1}^{t} \phi^2(w_i, c_{a_i}) \right]. $$

$\hat{a}$, which satisfies equation (8), is obtained by considering all possible permutations of the concepts and the word sequence using dynamic programing.

7. Experiments

To investigate the effect of the tight integration of ASR and NLU in the proposed method, we compared the following four methods in the SLU experiments.

Baseline Naive cascaded SLU using ASR 1-best hypotheses as inputs to NLU.

WordReject Cascaded SLU using ASR 1-best with word hypothesis rejection, where word hypotheses whose word level ASR confidence are below the threshold are rejected and replaced with unknown word symbols.

Proposed Tightly integrated SLU proposed in this paper.

Transcription (for reference) Cascaded SLU using transcriptions as inputs to NLU.

In Baseline, WordReject, and Transcription, NLU finds the most likely concepts that maximize joint probability $P(W,C)$ for 1-best ASR hypothesis $\hat{W}$ by exhaustive search, based on word-to-concept translation models.

The speech data were from a train timetable information domain collected with our telephone-based, Japanese spoken dialogue system. The system handled 17 types of concepts, station, time (hour:minute), from (superclass of station), to (superclass of station), departtime (superclass of time), arrivetime (superclass of time), their negation (6 types: such as not=(station=(STATIONNAME)) and not=(hour=(9))), acknowledgement, denial, goodbye, once again, and reset. These speech data were manually transcribed and annotated. In the concept annotation, a non-expert annotator selected appropriate concept types and filled their values for each transcribed utterance. After transcription and annotation, the corpus had 11,555 utterances from 129 different subjects that contained 63,719 words (616 different words, an average of 5.51 words per utterance) and 16,412 concepts (441 different concepts, an average of 1.42 concepts per utterance).

The experiment was held by 4-fold cross validation, dividing the corpus into four groupings of about 2,900 utterances in each grouping. We used Julius [19] as the ASR engine in the experiments. It can calculate confidence scores based on word posterior probabilities during 2-pass decoding [20]. The acoustic model was a domain-independent one trained on phoneme balanced sentences uttered by 517 speakers (50 sentences per speaker). Word-to-concept translation models and word N-gram models ($N=3$) were trained on the corpus using the CMU-Cambridge SLM Toolkit [21] and its default options. The average vocabulary size of the word-to-concept translation models in each grouping was about 930 and that of the word N-gram models was about 560. The number of concepts used in the word-to-concept translation models was averaged about 410. The average word accuracy by Baseline was 81.5%.

The evaluation measurements were SLU accuracies by precision and recall in the concept level, as well as word accuracy in ASR. Precision is the fraction of correctly recognized concepts over all recognized concepts, and recall is the fraction of correctly recognized concepts over the reference concepts.

Figure 2 shows the precision and recall of the four methods varying the confidence threshold for hypothesis rejection. The thresholds were for concept-level rejection in Baseline, Proposed, and Transcription, and for word-level rejection in WordReject. In Baseline and Transcription, the confidence of a SLU
8. Concluding Remarks

This paper presented an SLU method that tightly integrates ASR and NLU by word-to-concept translation models. They are based on N-gram and can be trained using a corpus with sentence-level alignments between words and concepts, which helps us reduce the effort for explicitly aligning words to concepts in corpus annotation. The proposed SLU method can be easily implemented with existing N-gram-based ASR, and the confidence of an SLU hypothesis can be obtained in a similar manner to posterior-based word-level confidence scoring. Experimental results showed that the proposed method was effective in SLU with confidence-based rejection.

Acknowledgment

We would like to thank Drs. Naonori Ueda and Eisaku Maeda for discussion and support. We also thank Dr. Mikio Nakano for valuable discussion about SLU and spoken dialogue systems as well as the members of the knowledge processing research group for their helpful comments.

9. References