Comparison of Parsing and Spotting Approaches for Spoken Dialogue Understanding

Tatsuya Kawahara*, Masahiro Araki* and Shuji Doshita†

Abstract

We have studied the optimal strategies for both LR parsing and spotting. In this report, several parsing and spotting approaches for spoken dialogue understanding are compared and evaluated. Here, a novel phrase spotting approach based on progressive search is proposed for robust understanding. The experimental results show that (1) sentence-level parsing is most powerful but not robust, (2) phrase spotting approach is robust against ill-formed utterances, (3) simple word bigram and word spotting get good word accuracy but do not lead to sentence-level understanding. Furthermore, we explore a hybrid approach where the sentence-level parsing is tried and, if it fails, the phrase spotting is performed.

1 Introduction

There have been studied several approaches for speech understanding in spoken dialogue systems. In this paper, we focus on the approaches to model and utilize linguistic knowledge, which interface between an acoustic model and a semantic analyzer. It seems obvious that more constraint brings better accuracy. However, use of strong constraint loses robustness against variety of utterances that do not satisfy it, as in spoken dialogue systems. Therefore, we make evaluation and comparison with respect to both accuracy and robustness, using both grammatical and ill-formed utterances.

2 Progressive Search Strategies

The basic concept that underlies our approaches is progressive search strategy[1]. It applies multiple-level constraints sequentially in the order of their strength. By constraining the search space, admissible and efficient search is realized.

In any cases, word-pair constraint or word bigram is used as the baseline heuristics. In order to decode a speech input to a word sequence, this statistical language model is effective and robust. To lead higher-level understanding of the speech, a grammar of structural and

*Tatsuya Kawahara (河原 達也), Masahiro Araki (荒木 雅弘) : Research Associates, Department of Information Science, Faculty of Engineering, Kyoto University

†Shuji Doshita (堂下 修司) : Professor, Department of Information Science, Faculty of Engineering, Kyoto University

symbolic knowledge is incorporated. Specifically speaking, we use a sentence-level grammar and phrase-level syntax, which realize two different approaches called parsing and spotting, respectively.

The integration of the word bigram and the grammars is realized in the framework of heuristic search. It is formulated as follows. A partial sentence or phrase hypothesis n is evaluated by the sum of the score given by the both language models.

$$f(n) = g(n) + h(n) \tag{1}$$

$g(n)$: score of the parts parsed by a grammar
$h(n)$: score of the rests constrained by word bigram

Here, the heuristics given by the word bigram $h(n)$ provides perspective of the hypothesis, which guides the symbolic parser. The progressive search first applies word bigram constraint, and then performs symbolic parsing using the result of the first pass as heuristics. The significant point is we do not make critical decision of candidate selection nor segmentation at the first pass, but we store all the scores of possible words at every time-point as a trellis for further analysis. This feature avoids any loss of information and realizes admissible search that gets the globally optimal result.

The concrete algorithms of the sentence parsing and the phrase spotting is described in the following subsections.

2.1 Optimal LR Parsing

At first, we describe the optimal search strategy for LR parsing with sentence-level syntax and a lexicon.

It is realized as a two-pass search strategy[2][3]. It first performs rough decoding with word bigram, and then applies the syntactic constraint of an LR grammar. The second pass is realized as A* search using the result of the first pass as admissible heuristics. The best-first search uses the grammar for predicting the next possible words, whose evaluation scores are provided by the heuristics.

In order to guarantee the admissibility, we derive the word bigram that is subset of the probabilistic LR grammar. First, the word-pair constraint is derived by choosing the possible connections of the words that are allowed by the grammar. Then, the word transition probability is computed as the maximum of the probabilities attached to the associated grammar rules[4].

2.2 Optimal Phrase Spotting

In a spoken dialogue system, it is significant to deal with ill-formed sentences. The ill-formedness such as fillers, hesitations and unknown words is of variety, and it is hard to model and describe all of them. When the task domain is specified, we can mostly make senses of utterances by putting attention to keywords. It is also possible to make clear the unrecognized parts through the following dialogue. Therefore, spotting-based approach that extracts only recognizable parts and skips the rests is attractive.

The basis of this approach and the key to its success is keyword spotting. Although it has been studied a lot, the spotting strategy has not succeeded so well especially at a large-vocabulary task. The reason is spotting itself fails to obtain enough detection accuracy, and causes too many false alarms to deal with in the following processing.

To overcome the above defect, we propose phrase spotting based on the progressive search strategy. It applies multiple-level constraints sequentially in the order of their strength as follows.

1. Simple language model such as word/syllable bigram for the whole utterance
2. Local phrase syntax at the spotting stage
3. Inter-phrase semantic constraint for sentence understanding

As the unit or target of the spotting, phrase is advantageous because it can incorporate more distinguishing information and linguistic knowledge. A phrase consists of a few keywords and functional words, for example, 'from Tokyo', or 'at three o'clock in the afternoon'. Even in spontaneous speech, a phrase is uttered at a moment without break, and its syntactic structure is rarely violated. Moreover, a phrase makes a semantic case and is directly mapped into a semantic representation.

The phrase spotting[5] is realized as a heuristic search as formulated in equation (1). Here, $h(n)$ represents not only the prospect for the incompleted part of the phrase but also the score of the rest parts that are not covered by the phrase. While word bigram is applied for constraining the whole utterances, the syntactic analysis is performed on only those parts that are recognizable with the available knowledge. The unrecognizable parts including ill-formed parts are left approximated with the statistical model. The best-first search extends plausible phrase hypotheses.

The point is we store the intermediate results of the HMM trellises at each step. The trellises are compacted by reserving only the initial and final states of the words/phrases. In the conventional robust parsing approaches[6][7], the phrases or fragments are collected only from the word sequences of the first trial, and the overall optimization is not performed with Viterbi algorithm. Our strategy progressively constructs sentence hypotheses. By retaining the trellises, the globally optimal hypothesis is obtained with the correct Viterbi score. By choosing such constraint at every step as subset of that for the following steps, the search gets A*-admissible.

Concretely speaking, if the constraint for $h(n)$ is subset of that for the phrase syntax, the spotting algorithm is A*-admissible and outputs the phrase candidates correctly in the order of their scores. The word bigram that is derived from the probabilistic phrase grammar as in the previous subsection satisfies the condition.

The sentence hypotheses are also searched in the same way. It is constructed by combining the spotted phrases and evaluated by concatenating their compacted trellises with Viterbi algorithm.

3 Comparison of Approaches

Then, we make comparison of the several parsing and spotting approaches. Specifically, the four typical approaches are investigated.

(1) word bigram only

The simplest approach applies word bigram constraint to decode an input speech into (N-best) word sequences, which are passed to a semantic analyzer.

(2) word spotting

As a robustness oriented approach, a set of plausible words are picked up and the semantic analysis is performed island-driven. In the spotting stage, we incorporate word bigram constraint and choose the word candidates with better scores. So, the difference from (1) is whether the output is word sequences or a word lattice.

(3) sentence LR parsing

The approach, described in subsection 2.1, uses more powerful linguistic constraint, that is sentence-level grammar. It imposes constraint of longer distance than (1) or (2) that focuses on only neighboring words. The (N-best) sentence candidates that satisfy the grammar are passed to a semantic analyzer. The method, on the other hand, restricts the patterns of user utterances more strictly and may lose robustness.

(4) phrase spotting

The approach, described in subsection 2.2, is regarded as a compromise of (2) and (3). It just describes syntax of phrases that correspond to semantic cases, and searches for plausible phrases. A set of phrases are passed to a semantic analyzer.

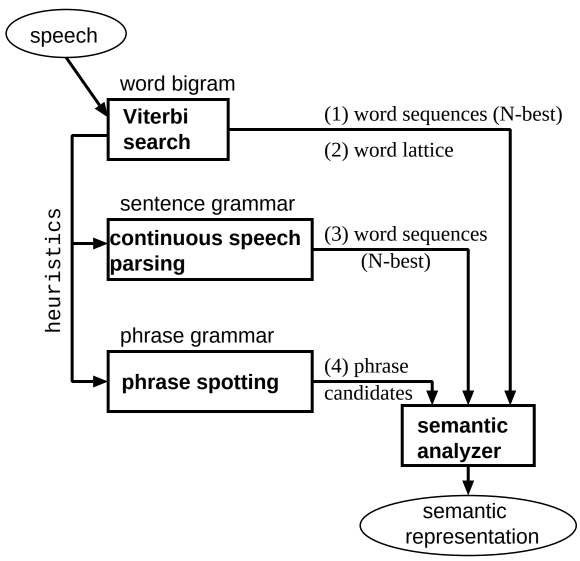

Figure 1: Several parsing and spotting approaches

Table 1: Comparison of the approaches in sentence understanding

approach	grammatical word	grammatical sent.	ill-formed word	ill-formed sent.
(1) word bigram only	84.9	56.3	69.3	27.0
(N-best interface)	(94.6)	61.5	(78.3)	29.0
(2) word spotting	80.9	44.0	69.6	27.5
(3) sentence LR parsing	86.2	66.3	59.9	36.5
(N-best interface)	(93.9)	68.3	(67.2)	36.5
(4) phrase spotting	81.5	66.8	72.7	43.0
hybrid: (3) $\rightarrow$ (4)	87.1	69.0	73.2	40.5

The procedure and relations of above methods are illustrated in Figure 1. The same word bigram is used for (1) and (2), and the two methods are different in their output forms. Its results are utilized in (3) and (4) in much the same way. The two methods are different in the constraint of knowledge sources.

We have evaluated the above approaches on recognition and understanding of speaker-independent continuous speech. The task domain of our spoken dialogue system is the personal schedule management. We use two sets of sample sentences: grammatical utterances and ill-formed ones. While the grammatical sentences are accepted by the LR grammar, the ill-formed ones contain out-of-syntax phrases and filled pauses unknown to the system, and can hardly be dealt with by conventional speech recognizers. In all the approaches, either Viterbi search or A* search is adopted so that the optimal word sequences or lattices are obtained. Moreover, they interface with the same semantic analyzer that can handle both word sequences and a lattice. As for parsing, both 1-best and N-best interfaces are evaluated. In the N-best interface, the 10-best word sequences are passed to the semantic analyzer in turn until a semantic representation is obtained. The word perplexity of the word bigram is 51.7. The word perplexity of the sentence LR grammar and the phrase grammar is 17.2 and 56.1, respectively. Table 1 lists the word accuracy (word) and the sentence understanding rate (sent.), the ratio of the samples whose correct semantic representations are obtained.

For the grammatical utterances, the accuracy is clearly ordered according to the strength of the constraint. The sentence LR parser with the most constraining grammar gets the best accuracy. In this case, the N-best interface is effective.

In recognition of the ill-formed utterances that violate the prepared constraint, different tendency is found. In this case, the LR parser gets the worst word accuracy. It cannot cope with some of the ill-formedness at all, namely lacks of robustness. The other approaches that use local constraint are more robust. In the sentence-level understanding, however, the word bigram and the word spotting do not achieve good accuracy. The phrase spotting that directly leads to semantic representations is the best. Notice that the N-best interface does not improve the sentence rate. Robustness is not realized simply with the use of the N-best candidates.

Although the word accuracy is the worst, the sentence LR parser can get correct semantic representations in spite of some syntax errors, while it often fails to output any candidates, namely rejects out-of-syntax utterances. This property would be advantageous, if coupled with a good dialogue manager that guides the user to remake an adequate utterance when unrecognized.

In general, the simple constraint such as word bigram is effective only for utterances that are matched well with the acoustic model. In our experiments, about 80% of the inputs are classified into this category, as shown by the sentence recognition rate. For acoustically unclear utterances, the use of the linguistic constraint is significant. It is confirmed that the sentence-level grammar is most effective for grammatical sentences, while the phrase spotting approach achieves the best accuracy for ill-formed ones.

4 Hybrid Approach

From these results, an optimal approach should be taken, according to the expected ratio of grammatical utterances and ill-formed ones. When we can generally predict and cover the patterns of utterances as in a restricted task, use of a sentence-level grammar will be the best. If we allow variety of utterances that cannot be expected beforehand, spotting strategy is desirable.

It is also possible to take a hybrid approach that first applies sentence parsing and then spotting if necessary.

Since not a few part of the utterances are well-formed and parsing is more efficient, parsing should be performed at first. And if it fails, then spotting should be applied. Since the basis of the two approaches are common in our methods, namely the common heuristics is shared by the both, it is possible to switch them without overhead computation.

In this kind of constraint relaxation method, the problem is how to judge the failure of the first parsing, or how to reject out-of-syntax utterances. In parsing the word sequences[7][8], the judgement is simple. But it is hard in the speech parsing that assumes all the possible sentence hypotheses. Here we use the property of A* search that it generates too many hypotheses and fails to get any candidates for out-of-syntax utterances. So parsing is aborted at some threshold on the number of generated hypotheses, and switched to the spotting mode. It is still an incomplete criterion, but almost sound.

Actually, in our experiments, the hybrid strategy can realize robustness for ill-formed utterances, while keeping the accuracy for grammatical ones (see the last entry of the Table 1).

5 Discussions

We have compared and evaluated several parsing and spotting approaches for spoken dialogue systems.

The conditions are significantly affected by the dialogue management. If the system takes an initiative to proceed with a dialogue session, it will be easier to make user utterances follow the prepared patterns. Actually, we demonstrated the effectiveness of dialogue-level prediction for speech recognition in [9]. However, the system-driven dialogue might hamper user-friendliness at some task. Therefore, the optimal approach depends on the design of a task and a dialogue manager.

The results in the paper will contribute as a quantitative hint in any cases.

References

[1] H.Murveit, J.Butzberger, V.Digalakis, and M.Weintraub. Large-vocabulary dictation using SRI's DECIPHER speech recognition system : Progressive search techniques. In *Proc. of IEEE-ICASSP*, volume 2, pages 319–322, 1993.

[2] D.Goddeau and V.Zue. Integrating probabilistic LR parsing into speech understanding systems. In *Proc. of IEEE-ICASSP*, volume 1, pages 181–184, 1992.

[3] T.Kawahara, S.Matsumoto, and S.Doshita. A*-admissible context-free parsing on HMM trellis for speech understanding. In *Proc. Pacific Rim Int'l Conf. on Artificial Intelligence*, volume 2, pages 1203–1208, 1992.

[4] T.Kawahara, M.Araki, and S.Doshita. Heuristic search integrating syntactic, semantic and dialog-level constraints. In *Proc. IEEE Int'l Conf. Acoust., Speech & Signal Process.*, volume 2, pages 25–28, 1994.

[5] T.Kawahara, T.Munetsugu, N.Kitaoka, and S.Doshita. Keyword and phrase spotting with heuristic language model. In *Proc. Int'l Conf. on Spoken Language Processing*, volume 2, pages 815–818, 1994.

[6] W.Ward. Understanding spontaneous speech: The PHOENIX system. In *Proc. of IEEE-ICASSP*, pages 365–367, 1991.

[7] S.Seneff. Robust parsing for spoken language systems. In *Proc. of IEEE-ICASSP*, volume 1, pages 189–192, 1992.

[8] D.Stallard and R.Bobrow. Fragment processing in the DELPHI system. In *Proc. of DARPA Speech & Natural Language Workshop*, pages 305–310, 1992.

[9] T.Kawahara, M.Araki, and S.Doshita. Reducing syntactic perplexity of user utterances with automaton dialogue model. In *Int'l Sympo. on Spoken Dialogue*, pages 65–68, 1993.