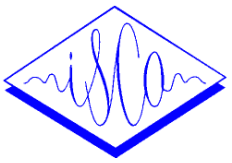

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A Robust Speech Understanding System Using Conceptual Relational Grammar

Jiping Sun, Roberto Togneri, Li Deng***
Department of Electrical and Computer Engineering
University of Waterloo, Waterloo, Canada
*on sabbatical from the University of Western Australia
** current address: Microsoft Research, Redmond, WA

ABSTRACT

We describe a robust speech understanding system based on our newly developed approach to spoken language processing. We show that a robust NLU system can be rapidly developed using a relatively simple speech recognizer to provide sufficient information for database retrieval by spoken language. Our experimental system consists of three components: a speech recognizer based on HMM, a natural language parser based on conceptual relational grammar and a data retrieval system based on the ATIS database. With the use of the robust parsing strategy, database query tasks can be successfully performed.

1. INTRODUCTION

Over the past few years, we have developed a robust speech understanding system based on our newly developed approaches to speech recognition and robust parsing. The purpose of developing this system is to investigate how well a robust NLU system can be rapidly created using a relatively simple continuous speech recognizer to provide sufficient information for applications such as database retrieval by spoken language. The ATIS spoken language database provides a suitable test platform for our purpose. It contains a large speech corpus and also information for a relational database. Therefore, we trained and tested both the speech recognition system and parser with the ATIS data.

Our basic assumption is that given a reasonably high recognition rate, a robust natural language parser should be able to identify meaningful structures and concepts to construct database queries. Our experimental results have shown that in order to build a functional speech understanding system, a robust NLU parser should be able to tolerate a certain degree of accuracy reduction on the part of the recognizer due to modeling errors, insufficient training data, confusable words or any other factors in coping with spontaneous speech. The close cooperation of speech recognition and robust parsing is needed to achieve reasonable speech understanding.

Our experimental system consists of three parts: a speech recognizer, a natural language parser, and a database simulator containing the ATIS database information. The ASR component is a continuous speech recognition system based on context-dependent hidden Markov models. The system was designed using the latest version of HTK. The ATIS-2 corpus was used for training and testing the acoustic models.

The NLU parser is based on a conceptual relational grammar developed in our group. The parser component of the system consists of an annotation module, a rule derivation module, a rule application module and a domain dependent semantic lexicon. The annotation module defines a convention of annotation: how to describe concepts and the relations between them. The rule derivation module derives probabilistic rules from the annotated corpus. The rule application module applies the rules to the input sentence. The lexicon defines semantic categories specific to a domain and a similarity relation between semantic categories assigned to words.

The simulated relational database used in our system development and experiments contains functions for retrieving information and performing database operations. A subset of SQL is implemented and functions as the interface between the parser and the database internal functions. The text files containing the ATIS air travel database information are transferred into relational database tables.

2. SPEECH RECOGNIZER

The Hidden Markov Toolkit (HTK) version 2.2 was used to implement the speech recognition front-end of the system. The reader is referred to [1] for a description of the capabilities and features provided.

2.1. Data Preparation

The ATIS-2 corpus was used for training and initial evaluation of the system. The data contains recordings of spontaneous speech from 453 speakers, collected at six different speech laboratories around the United States. All recordings contain the necessary transcription and annotation data used to collect the training data and define the HMM word network. Utterances containing mispronunciations, word fragments, truncated data or overlapping non-speech events were deleted from the initial training data as this data was deemed too unreliable. This yielded 12150 utterances for training the models.

The feature extraction consisted of generating a 39-dimensional feature vector for each 25 ms analysis window with 10 ms frame advance. The feature vectors included 13 static, 13 delta and 13 acceleration mel-spaced cepstral coefficients (including the energy term). Each utterance was further subject to cepstral mean normalization for enhanced robustness to convolutional noise.

2.2. Model Development

A 5-mixture, 3-state, left-right, no-skip, diagonal covariance topology was used for the base phone models and non-speech word models. There were four non-speech word models: silence,

short pause, non-speech-noise and filled-noise. The silence model was inserted at the beginning and end of each utterance whereas a special short-pause model was used for the pause between words.

Initially flat-start monophone models were trained. Word internal (no crossword) context-dependent triphones were cloned from the monophone models and re-trained. Due to the increased number of states decision tree-based clustering was used to designate candidate states to tie. One of the advantages of using decision-tree based clustering is that it allows previously unseen tripones to be synthesized. This feature was used to synthesize triphones which were not present in the training data but could be present in the testing and evaluation data. The final models consisted of both trained and synthesised triphone models based on the ATIS training data and dictionary respectively.

For decoding an unknown utterance a task grammar or language model word network is needed to constrain the Viterbi beam search. A backed-off word bigram language model was trained from the ATIS training data transcriptions. Although a class bigram language model would have been more appropriate, this feature is currently not yet implemented in the HTK software. A pruning threshold was also used to discard search paths with a low likelihood and thereby further reduce the search space and consequent recognition time. As the ATIS task is highly constrained, a relatively large weighting to the language model was found to improve performance.

The ATIS test data was used to evaluate the performance of the speech recognition component. A total of 1976 useable test utterances were available, including utterances with mispronunciations, truncated data, word fragments and overlapping non-speech events. For direct audio (live) recognition, speaker adaptation models were trained and evaluated on selected speakers. For adaptation, the enrolment data consisted of 20 sample utterance prompts from the ATIS training data set with a different set of 20 samples used to evaluate the performance. Table 1 shows the results where WER(CORR) measures the incorrect phone rate and WER(ACC) measures the incorrect phone plus phone insertion rate.

TEST DATA	%WER(CORR)	%WER(ACC)
ATIS-2 Test Data	14%	19%
1 native speaker no adaptation	18%	20%
1 native speaker MLLR adaptation	13%	15%
1 native speaker MLLR+MAP adaptation	14%	16%
2 non-native speakers no adaptation	58%	85%
2 non-native speakers MLLR adaptation	25%	38%
2 non-native speakers MLLR+MAP adaptation	22%	32%

Table 1. ASR test results.

3. ROBUST PARSING STRATEGY AND RESULTS

This parser uses a grammar formalism that follows the analysis style of dependency grammars and word grammar [2,3] with some modifications. First, the minimum unit of structural building is not the word but concept, which consists of one or more words with conceptual meanings. Second, the grammar rules are derived from an annotated corpus and organized as probabilistic functions with tolerance for ASR accuracy reduction and ungrammatical structures typical of spontaneous speech. We annotated the ATIS transcriptions as the basis for deriving grammar rules.

3.1. The Annotation of the ATIS Corpus

500 sentences in the ATIS training set were annotated for deriving statistical grammar rules. The annotation describes conceptual relations between words bearing conceptual contents. The conceptual relation structure follows similar analysis of case grammar and the NLCA [5,6].

The conceptual categories and relations have been considered to be specific to particular application domains, such as air travel. At the same time, we consider them to be on a hierarchy from general to specific. Thus, we envisage the possibility of generalizing domain specific regularities to those that can be used in other domains.

In the annotated ATIS corpus, 55 conceptual categories and 22 syntactic categories have been used. These categories are assigned to the words or word groups. Some of the more frequent conceptual categories are [airline], [time], [city], [person], [movement], [travel].

The annotation work was carried out in this way: First, words or word groups that carry conceptual meanings are identified and assigned a conceptual category. The other words are kept in their original places. Then conceptual relations are assigned to pairs of conceptual categories. Below is an annotated sentence.

```
I : sem#person,          'd like to : syn#aux,
find : sem#search,  the,  cheapest : sem#price_atb,
flight : sem#movement,  from,  Washington D C :
sem#city1,  to,  Atlanta : sem#city2
---------------------------------------------------------------------
agent(search,person),  patient(search,movement),
loc_source(movement,city1),loc_target(movement,city2),
qualifier(search,aux),  descriptor(movement,price_atb)
```

The first part is the categories and the second part is the relations. A relation takes the form of a predicate. The first argument is the head category that dominates the second or tail category. The relations used in our system are similar to those in case grammar [5], especially for verbs. Some abstract noun relations are used to avoid over-specification. Three such abstract relations occur frequently in our annotation work and deserve illustration here. They are **qualifiers**, **specifiers** and **descriptors**. A qualifier serves to identify the function of a word; a specifier adds concrete semantic content to a more abstract concept; a descriptor clarifies an internal aspect of a

concept. The following two annotated sentences from ATIS illustrate these concepts.

```
what : syn#wh,          type : sem#attribute,      of,
aircraft : sem#vehicle,              is : syn#be,
that : syn#index,              flight : sem#movement
-------------------------------------------------------------------
topic(be, attribute),          predicate(be, movement),
qualifier(attribute, wh),    specifier(attribute, vehicle),
qualifier(movement, index).
```

```
could : syn#aux,              you : sem#person1,
tell : sem#communicate,    me : sem#person2,      the,
fare : sem#money,     for,       flight : sem#movement,
number,                 thirty seven : sem#number
-------------------------------------------------------------------
agent(communicate, person1),
recipient(communicate, person2),
patient(communicate, money), purpose(money,
movement),      descriptor(movement, number)
```

3.2. Grammar Rule Derivation

Based on the annotated corpus a grammar derivation program was used to derive grammar rules. A grammar rule describes four items: the head category and its context, the dependent categories and their contexts, the relations and positions of the dependent categories relative to the head category. A context is the left and right three categories or words (if they are not assigned a category) of any category in the grammar rule. As an example, the verb *find* in the sentence *I'd like to find the cheapest flight from Washington D. C. to Atlanta* will give rise to a grammar rule as:

```
head = SEARCH  ,  context =
   $ + PERSON + AUX (*) 'the' + PRICE-ATB + MOVEMENT
agent = PERSON,  position = -2,  context =
   $ + $ + $ [ ] AUX + SEARCH + 'the'
patient = MOVEMENT,  position = +3,  context =
    SEARCH + 'the' + PRICE_ATB (*) 'from' CITY 'to'
qualifier = AUX,  position = -1,  context =
   $ + $ + PERSON (*)  SEARCH + 'the' + PRICE_ATB
```

For every head category that controls some dependent categories such a rule is derived. If two rules are same the frequency information is retained. A total of 45 conceptual relations were used in 1240 rules. The first ten most frequent relations are:

1. qualifier	334	6. loc-source	216
2. agent	291	7. topic	171
3. loc-targt	240	8. descriptor	143
4. patient	221	9. specifier	115
5. time	219	10. predicate	112

The grammar derivation program also keeps track of other statistics, such as word frequencies, category frequencies and category in context frequencies, and so on. The following list gives the ten most frequent conceptual categories and their frequencies.

1. sem#city	480	6. syn#wh	133
2. sem#movement	285	7. sem#airline	127
3. sem#person	243	8. syn#be	125
4. syn#aux	226	9. sem#move	119
5. sem#number	182	10. syn#index	69

The grammar rule derivation with the annotated corpus also produced a lexicon with words given conceptual categories, frequencies of words in a particular category and context. This lexicon was then combined with a big lexicon that contains words not in the annotated corpus. The words from the big lexicon do not have the category frequency and contextual information. They are only assigned conceptual categories.

3.3 The Parsing Process

The statistical parsing process is designed to work with the rules derived from the annotated lexicon. We assume that an annotated corpus, if big enough, will cover a large part of conceptual categories and relations in their typical contexts. Based on this assumption, when a new input sentence is given to the parser, it will compare the new sentence with the data in the derived grammar and lexicon in order to provide a best estimation of the word categories and their conceptual relations. This parser has the character of toleration for grammatically un-well-formed sentences. In particular, it is designed to have the potentiality of tolerating random insertions of conceptually irrelevant words between the correct ones. We assume that if a speech recognition n-best output word lattice is flattened into a situation as described above, the parser should be able to pick out the correct words in their appropriate conceptual relations, according to the corpus-derived grammar. The parsing process is described as follows.

- Lexicon Lookup: assign each input word or word group a conceptual category by using the lexicon.
- Scanning from left to right category by category. At each category:
  - Find all rules with this category as the head category.
  - Match the contexts between the rule and the category.
  - Select rules with context matching over some threshold.
  - Process each of the relations contained in a selected rule, looking for a tail category in the specified direction.
  - Match the contexts between the rule and the category. When more than one tail is found, select the one with the highest matching score.
  - For every selected relation, record **rel(head tail, acc-score, max-score)**. When the same relation has been selected before, increase the cumulative score and update the maximum matching score.
  - A matching score is a function of the head context matching, tail context matching and the relative position. Context matching is based on category similarity and weighting that favors immediate neighbors.

The parser depends on the frequency-based scores to identify the final category-to-category relations that will be used to form database queries. Using this parser to parse an ATIS test sentence *Find a flight from Boston to Dallas stopping in Denver*

we got the following result.

find [patient] flight	**16.86, 3.24**
flight [loc_boundary] boston	2.37, 1.27
flight [loc_path] dallas	3.18, 3.18
flight [loc_source] denver	1.97, 1.97
flight [loc_source] dallas	1.07, 1.07
flight [loc_target] denver	4.18, 2.08
flight [loc_path] denver	**6.34, 3.09**
flight [loc_target] boston	7.94, 3.21
flight [loc_source] boston	**304.49, 4.3**
flight [loc_target] dallas	**207.13, 3.18**
stopping [loc] boston	1.96, 1.9618
stopping [loc] denver	**7.27, 2.99**
stopping [agent] flight	**0.88, 0.88**

In this result the parser has identified all the correct relations between conceptual categories depending on the matching scores. We tested sentences that suffer distortions such as insertion, deletion and substitution of words as returned from speech recognition systems, and the results showed a good tolerance by the parser. Using the 1240-rule grammar, the average parsing time for a 10-word sentence is about 3 seconds. It should be noted that the parsing process for each word is independent of other words. This makes it possible to use parallel processing.

3.4 Initial Evaluation of the Parser

After the understanding system was created, we carried out initial evaluation of its performance. The initial tests showed that the two major components cooperated well. The parser rules based on conceptual relations could tolerate missing grammatical words, inserted irrelevant words and ungrammatical structures. As long as the words carrying key conceptual meanings appear among the recognition string, the NLU system was able to identify the conceptual structure and issue a database retrieval command. Current tests with the ATIS corpus have shown satisfactory performance. Expansion of the parsing system is being carried out for increased coverage of conceptual structures. The extension of the rule application system to word-lattices is also being investigated.

4 CONCLUSIONS AND DISCUSSION

Our robust parsing approach has the following characteristics: 1. The rules are derived from annotated corpus and based on conceptual relations; 2. Conceptual categories of words or word groups are used and these can be generalized from one domain to others; 3. The parsing algorithm is linear and highly parallel. Other robust parsing approaches are often syntactically based if corpus statistics are used. In other cases when semantic information is used, it is often the case that rules are hand-written and no statistical information can be used. Our approach combines the two together and has shown advantage when applied to a restricted domain.

Through our investigation of the integration of speech recognition and robust parsing, we have gained valuable experience on how to make the best use of each component to achieve the best possible understanding result. Our experience has shown that even with a relatively simple speech recognition system, insofar as the words carrying key conceptual meanings were recognized, a robust parser can identify the major information for completing application tasks. The inserted words, as long as they are not affecting the conceptual structure (i.e., not conceptually confusing), will not seriously affect the speech understanding process.

Our future work will concentrate on further investigations of: (1) the generalization power of the corpus based rules; (2) how well the system can tolerate irrelevant insertions and some deletions in a more systematic way; (3) the possibility of parsing n-best word lattices, in which more key words are expected to exist; and (4) various ways of representing context, conceptual meaning and their similarity measurement.

Acknowledgements: This work has been supported by NSERC, Canada. The work was performed while the second author was on leave at U. Waterloo from U. Western Australia.

5. REFERENCES

1. Young, S., "A Review of Large-Vocabulary Continuous-Speech Recognition", *IEEE Signal Processing Magazine*, Vol. 13, No. 5, pp. 45-57,1996.

2. Fraser, N.M., "Dependency Grammar", in K. Brown and J. Mille (eds) *Concise Encyclopedia of Syntactic Theories*, pp. 71-75. Pergamon Press, 1996.

3. Hudson R. and W. Wan Langendonck, "Word Grammar", in F. Droste and J. E. Joseph (eds) *Linguistic Theory and Grammatical Description*, pp. 307-336. Benjamins, 1991.

4. Hudson R., "Word Grammar", in R. Asher (ed) *The Encyclopedia of Language and Linguistics*, pp. 4990-3. Pergamon Press, 1994.

5. Fillmore, C., "The Case for Case", in Bach E. and Harms, R., (eds) *Universals in Linguistic Theory. New York*: Holt, Rhinehart & Winston, 1968.

6. Kamphuis V. and Sarbo, J. J., Natural Language Concept Analysis, in D. M. W. Power (ed) Proc of NeMLaP3/CoNLL98: International Conference on New Methods in Language Processing and Computational Natural Language Learning, ACL, pp. 205-14, 1998.