CONCEPT-BASED SPEECH TRANSLATION

L. Mayfield, M. Gavaldà, W. Ward, A. Waibel

Interactive Systems Laboratories
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213-3890
USA

ABSTRACT

As part of the JANUS speech-to-speech translation project, we have developed a robust translation system based on the information structures inherent to the task being performed. The basic premise is that the structure of the information to be transmitted is largely independent of the language used to encode it. Our system performs no syntactic analysis; speaker utterances are parsed into semantic chunks, which can be strung together without grammatical rules, and passed through a simple template-based translation module. We have achieved encouraging coverage rates on English, German and Spanish input with English, German and Spanish output.

1. INTRODUCTION

If all that a speech translation system were required to work with were perfectly formed and pronounced sentences, consisting of only words familiar to all processing components, it could reliably employ elegant syntactic parsing schemes which key on short function words and produce an interlingua-level representation which can be accurately translated into the target language. Unfortunately, spontaneous speech is seldom grammatically perfectly formed, often not even expressing a complete thought; poorly articulated and often containing incorrect function words if any. These short function words are also those most easily misrecognized, so the decoded utterance that the parser must process may bear little resemblance to the kind of sentence a syntax-based parser is prepared to handle.

Our system, an extension of the Phoenix Spoken Language System [4], tries to model the information structures in a scheduling task and the way these structures are realized in words in various languages. Grammatical constraints are introduced at the phrase level and regulate the semantic rather than the syntactic category. This method allows the ungrammaticalities that often occur between phrases to be ignored and reflects the fact that syntactically incorrect spontaneous speech is often semantically well-formed.

2. SYSTEM OVERVIEW

The translation component of the JANUS system [1] consists of parsing and generation modules. Decoded speech data is sent to the parser, which identifies the key concepts and variables in each utterance; the generation module reëncodes the resultant parse in the specified target language.

Translation of English, German and Spanish as source and target languages is currently operational. We have also implemented Japanese and Korean as additional target languages.

2.1. Parsing

Starting from the assumption that semantic units used in a task domain are, unlike individual words, not language specific, we have designed a set of tokens, representing the different concepts a speaker would use, as the fundamental units in our parser. The set of semantic tokens for the appointment scheduling task was developed from a set of 45 example English dialogues.

Top-level tokens, also called slots, represent speech acts, such as suggestion or agreement; lower-level tokens capture the specifics of the utterance, such as days of the week.

The parsing grammar specifies patterns which represent concepts in the domain. The patterns are composed of words and other tokens for constituent concepts. Elements (words or tokens) in a pattern may be specified as optional or repeating (as in a Kleene star mechanism). Each concept, irrespective of its level in the hierarchy, is represented by a separate grammar file. These grammars are compiled into Recursive Transition Networks.

This general approach has been described in earlier papers [2, 3]. A typical *temporal* token could have as a subtoken a *date*, which could in turn consist of *month* and *day* subtokens. The *temporal* could be used in a statement of unavailability, in which case a second slot suggesting an alternate time might follow.

The parser matches as much of the input utterance as possible to the patterns specified by the RTNs. Out-of-lexicon words are ignored. Words in the system lexicon, but not fitting the pattern being matched, will cause the concept pattern not to match. This does not cause the entire parse to fail, simply the concept slot being matched. The parser can ignore words between slot-level concepts, but cannot ignore words interior to a concept pattern. A version of the parser is under development which allows substitution, deletion and insertions in a pattern with a penalty.

The parser may string slots together in any order, but in cases in which slot boundaries are not clear-cut it must decide how to segment the utterance. First, it looks for the interpretation with the most words matched. If there is no single best interpretation in this sense, it searches for the interpretation with the fewest number of slots. This is equivalent to finding the least fragmented version. If the interpretation is still ambiguous, it picks the one which has a fewer number of tokens at a higher level in the parse tree. Thus, an interpretation in which two tokens are nested is preferable to one in which they are sequential.

Figure 1 shows an example of a speaker utterance and the parse that was produced using this system. The recognizer output, which is the text sent to the parser, is shown with unknown (-) and unexpected (*) words marked. Here we see the disfluencies common in spontaneous speech; this compounded with misrecognitions presents a syntactic parsing challenge. Relevant concepts, however, are easily extracted, and strung together they provide an accurate representation of what the speaker actually said.

The system is significantly different from conventional ones in that the goal is not to reproduce in the target language precisely what the speaker said, but rather to elicit the desired response from the listener. Therefore concepts with very different linguistic realizations may be mapped onto the same token. The expressions "what do you think" and "let me know" serve the same discourse function, namely, to indicate that the speaker is turning over the floor to his conversation partner. These word strings appear as possible matches for the slot *your_turn*.

```
Original utterance:
  THAT SATURDAY I'M NOT SURE ABOUT BUT YOU SAID
    YOU MAY BE BACK IF YOU THINK YOU'LL BE BACK
    THE THIS SUNDAY THE TWENTY EIGHTH I COULD SEE
    YOU AFTER ELEVEN AM ON THAT IF YOU'RE BACK

As decoded by the recognizer:
  *that saturday i'm not sure about but *you -said
    *you *maybe -back *into *think *to *be *back
    the sunday the twenty eighth i could see you
    after eleven am on *that *if *you -back

Parsed:
  [temporal] ( [point] ( [d_o_w] ( SATURDAY )))
  [give_info] ( [my_reluctance]
                ( I'M NOT SURE ABOUT ))
  [interject] ( [conj] ( BUT ))
  [give_info] ( [my_availability]
                ( [temporal] ( [point] ( THE
                [date] ( [d_o_w] ( SUNDAY ) THE
                          [day_num] TWENTY EIGHTH )))
                 I COULD SEE YOU ))
  [temporal] ( [range]
        ( [after] ( AFTER ) [time]
                  ( [hour] ( ELEVEN AM )) ON ))
```

Figure 1: A Typical Utterance

2.2. Generation

With the input string reduced to the concept level, target language generation is easily accomplished. The generation segment of the system is a simple left-to-right processing of the parsed text. The translation grammar consists of a set of target-language phrasings of each token, including lookup tables for such variables as numbers and days of the week. When a lowest-level token is reached in tracing through the parse, the process reverses itself and a target-language representation is created by inserting the translation for each subtoken into the template from the translation grammar for the parent token which it fits. The process then continues with the next concept. The result is a meaningful, if somewhat telegraphic, translation:

Saturday that's not so good for me Sunday the twenty eighth works for me after eleven a.m.

El sábado no me va demasiado bien pero el domingo veintiocho me va bien después de las once de la mañana.

Samstag könnte ich nur zur Not aber Sonntag der Achtundzwanzigste geht bei mir ganz gut nach elf Uhr morgens.

PARSER PERFORMANCE

	Transcribed		Speech	
	token	utterance	token	utterance
English	87.5%	76.0%	70.0%	49.8%
German	85.0	76.0	56.0	34.0

Figure 2: *Coverage of transcribed vs. recognizer-decoded speech. Recognizer word accuracy is 61% for English and 70% for German.*

END-TO-END EVALUATION

	Parsed from		Translated into
	token	utterance	utterance
English	95.6%	90.0%	90.2%
German	92.4	89.6	87.3
Spanish	88.8	58.3	82.2

Figure 3: *Evaluation of full translation of transcribed data. Figures represent percent of correct translations.*

3. ANALYSIS AND RESULTS

The results in this section represent evaluations at two different stages of system development. Because the transcriptions corresponding to the available speech data were used for training after the initial test, coverage rates shown in Figure 2 are those of the parser at that point in its development. With further training, however, parse accuracy has improved to its current level, shown in Figure 3. No Spanish speech data is available at this time. Evaluations were done on seven unseen dialogues of approximately ten utterances each.

Figure 2 compares the performance of the parser on transcribed and spoken input. Parse evaluations were performed at both the token (concept) and full utterance level. In token analysis, the tokens and variables identified by the parser were compared to a hand-coded set of tokens designated acceptable for each utterance. Recall coverage was then calculated.

While token analysis provides a framework for understanding how well individual concepts are being extracted, utterance analysis shows how often a response consistent with the intention of the speaker will be elicited. In utterance level evaluation only parses with no missing or incorrect key tokens were counted as correct. Analysis was performed on transcribed data in all evaluations and speech data where available. Coverage of speech data input does not reflect the word accuracy of the input.

In order to evaluate the generation component, native speakers of the target language fluent in the source language were asked to make subjective judgements as to whether the sense and key details of the source utterances were conveyed in the target language translations. This was done only at utterance level; when working with speech input the judges saw only the original speaker utterance and the final translation.

Figure 3 shows coverage rates in the three fully implemented languages. This reflects full system performance. Independent evaluation of the generation module on only well-formed input would show a much higher accuracy rate.

4. DISCUSSION

This system has several strengths which allow it to handle spontaneous speech in a very natural way. By focusing on the phrase as the fundamental unit, it can extract meaningful chunks from a grammatically fragmented sentence. This same capability allows it to process run-on sentences easily. Without an explicit notion of a sentence, the parser simply continues to extract and string together concepts until the end of an utterance is reached — it has no need for syntactic boundary markers. In early evaluations utterances that had been segmented manually were used; we found that coverage actually improved when all boundaries were removed.

Although some accuracy is lost when small function words are ignored, the ability to do so is of enormous benefit when working with recognizer output in which such words are often mistaken. By keying on high-confidence words this system takes advantage of the strengths of the speech decoder.

This method of parsing, and response-oriented translation philosophy, makes target-language generation simple. Translation grammars can be written and integrated very quickly, and while stringing translated phrases together at first seems unlikely to produce a meaningful target-language sentence, in languages with similar phrase order conventions, any gaps produced by missegmentation in parsing simply disappear. What happens between more dissimilar languages is a topic for further research and is currently under investigation.

Most of the errors that occur in both parsing and generation are due to inadequate lexical coverage and out-of-domain input. Recognition errors are still typically responsible for 70% of errors in end-to-end translations; coverage problems are the cause of approximately 25% more with the remaining 5% due to a variety of factors, including global ambiguity.

One disadvantage of this approach is the telegraphic and repetitive nature of the translations. A more detailed set of tokens would help to overcome this nuisance; however, the advantages gained by striving for expressive accuracy in this way are outweighed by the

problems that might arise were acceptable input expressions to be limited. Rather than expand the token framework to distinguish between different expressions with the same discourse function, in order to produce a more varied generation, the target-language module can provide multiple translation options for individual tokens.

5. CONCLUSION

The concept-based approach to speech parsing and translation described in this paper is especially well-suited to processing of spontaneous speech, which is often ungrammatical and subject to recognition errors. We feel that this approach is more robust than those requiring well-formed input and relying upon markers and syntactic cues provided by short function words such as articles and prepositions. This system is still in the beginning stages; however, the facility with which system improvements (increased coverage, additional source and target languages; porting to other domains by redesigning the token set) could be accomplished causes us to be confident about its potential.

6. ACKNOWLEDGEMENTS

This research has been supported in part by a grant from the Advanced Research Project Agency and the Department of the Navy. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. We also gratefully acknowledge partial support by the Advanced Telecommunication Research Laboratories and by NEC Research Laboratories.

7. REFERENCES

[1] M. Woszczyna *et al.* Recent advances in JANUS: A speech translation system. In *Eurospeech'93, 3rd European Conference on Speech, Communication and Technology, Berlin, Germany*, pages 1295–1298, 1993.

[2] W. Ward. Understanding spontaneous speech. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 137, 141, 1989.

[3] W. Ward. The CMU Air Travel Information Service: Understanding spontaneous speech. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 127, 129, 1990.

[4] W. Ward. Extracting information in spontaneous speech. In *Proceedings of International Conference on Spoken Language Processing*, 1994.