TESTING GENERALITY IN JANUS: A MULTI-LINGUAL SPEECH TRANSLATION SYSTEM

Louise Osterholtz[1] *Charles Augustine*[2] *Arthur McNair*[3] *Ivica Rogina*[4]
Hiroaki Saito[5] *Tilo Sloboda*[4] *Joe Tebelskis*[3] *Alex Waibel* [3,4]

[1]Computational Linguistics Program, Carnegie Mellon University
[2]Computer Science Department, University of Pennsylvania [3]School of Computer Science, Carnegie Mellon University
[4]Fakultät für Informatik, Universität Karlsruhe, Germany [5]Keio University, Japan

ABSTRACT

For speech translation to be practical and useful, speech translation systems should be portable to multiple languages without substantial modification. We present the results of expanding the English-based JANUS speech translation system [1] to translate from spoken *German* sentences to *English* and *Japanese* utterances. We also report the results of implementing part of the LPNN speech recognition module on a massively parallel machine. The JANUS approach generalizes well, with overall system performance of 97%. This surpasses English-based JANUS performance.

1. INTRODUCTION

JANUS is a speech-to-speech translation system that incorporates connectionist and stochastic techniques for speech recognition, connectionist and knowledge-based parsing and knowledge-based text generation technologies [1]. The original JANUS system accepts spoken English sentences and translates them into German and Japanese utterances. Any practical speech translation system should be portable to multiple languages without substantial changes; we report the results of expanding JANUS to translate from German speech into Japanese and English.

German JANUS translates sentences from a database of 12 conference registration dialogs. The sentences are natural German expressions corresponding to the English dialogs. In the dialogs a caller is communicating with a conference secretary, trying to obtain information or register for an international conference. As in English JANUS, the dialog scripts were read and recorded in a quiet office environment. Testing is done on database recordings.

Translating German speech requires the same system architecture as English-based JANUS: a speech recognition system, text parsing and generation modules, and a speech synthesis component (see Fig. 1). The Linked Predictive Neural Networks speech recognition system (LPNN) [2] and the PARSEC connectionist parsing system [3] had previously been tested only on English. The Universal Parsing Architecture (UPA) developed at Carnegie Mellon [4, 5] is designed to be language-independent, but performance depends on hand-written grammars. The expansion of JANUS to German tests the portability of each module of JANUS. First we give an overview of each component of the JANUS system; then we discuss the application of each module to German input.

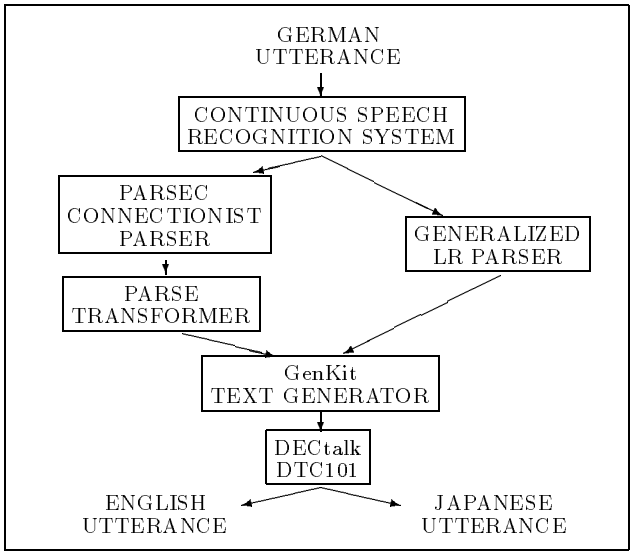

Figure 1: German JANUS System Components.

2. SPEECH RECOGNITION AND SPEECH SYNTHESIS

Two alternative speech recognition systems are currently used in JANUS: Linked Predictive Neural Networks (LPNNs) and Learned Vector Quantization networks (LVQ). They are both connectionist, continu-

ous speech recognition systems, and both have vocabularies of approximately 400 English and 400 German words. Both use statistical bigram or word-pair grammars derived from the conference registration database. The systems are based on canonical phoneme models (states) which can be logically concatenated in any order to create templates for different words. The need for training data with labeled phonemes can be reduced by first bootstrapping the networks on a small amount of speech with forced phoneme boundaries, then training on the whole database using only forced word boundaries.

In the LPNN system each phoneme model is implemented by a predictive neural network. Each network is trained to accurately predict the next frame of speech within segments of speech corresponding to its phoneme model. Continuous scores (prediction errors) are accumulated for various word candidates. The LPNN module produces either a single hypothesized sentence or the first N best hypotheses using a modified dynamic-programming beam-search algorithm. [6] The LPNN system has speaker-dependent word accuracy rates of 93% with first-best recognition and sentence accuracy of 69%.

LVQ is a vector clustering technique based on neural networks. [7] We have used LVQ to automatically cluster speech frames into a set of acoustic features; these features are fed into a set of output units which compute the emission probability for HMM states. This technique gives speaker-dependent word accuracy rates of 98%, 86%, and 82% for English conference registration tasks of perplexity 7, 61, and 111, respectively. The sentence recognition rate at perplexity 7 is 80%.

The recognition systems' text output serves as input to the alternative parsing modules of JANUS. Synthesis of English and Japanese speech is provided by the commercial Digital DECtalk DTC01 system. The system accepts English text and Japanese phonetic representations and produces sounds through an audio speaker.

3. KNOWLEDGE BASED TRANSLATION

The Universal Parser Architecture (UPA) consists of a parsing and a generation module and is capable of efficient multi-lingual translation. The parser uses Tomita's generalized LR parsing algorithm [8]; grammars are pre-compiled and parsing is reduced to fast table-lookup operations. Sentence generation is performed using GenKit, which compiles a generation grammar into LISP functions.

Hand-written parsing and generation grammars are required for each language to be translated. The grammars are based on a Lexical Functional Grammar formalism [10]. Both syntactic and semantic information is encoded in the rules and lexicon. The parser's ability to handle novel sentences depends on the scope of the handwritten grammar. Given a text sentence in the source language the parser uses the precompiled grammar, to produce a frame-based representation of the meaning of the sentence without language-dependent syntactic details.[9] This interlingua structure is the input to the generation module which uses the compiled generation grammar to produce the corresponding sentence in the target language. Both parsing and generation with the UPA approach real-time.

4. CONNECTIONIST PARSING

PARSEC is a connectionist parsing system developed by Jain[3]. The parser is constructed from separate connectionist modules arranged in a hierarchical fashion; each module must be trained separately to learn to parse words sequences into words, phrases and finally clauses. PARSEC's output is not designed for processing by the text generation module, so a separate parse transformation module makes the appropriate modifications. It uses simple match rules to instantiate case frames and their slots, which are filled in using further match rules. The parse transformer opportunistically tries to create an interlingua structure from any PARSEC output.

5. APPLICATION TO GERMAN JANUS

5.1 Speech Recognition

No modifications to the English-based LPNN system were required in order to recognize German speech, aside from using German phonemes, training on German acoustic data and using German text for stochastic language modeling. The neural networks used for acoustic modeling were bootstrapped on labelled spectrograms of 100 German sentences, spoken by a female speaker, which were unrelated to the conference registration task. [11] The 46 phonemes in this data were mapped into 40 phoneme models for German JANUS. The networks were further trained on unlabelled spectrograms of a male speaker saying the 204 sentences in the conference registration database. The data used for bootstrapping was recorded with a different microphone under different conditions from the unlabelled conference registration data. Special problems for recognition included many words ending in the suffixes "-en" and "-er"; these endings were pronounced weakly and said differently in different contexts. No specific action has been taken to improve their recognition at present.

Experiment		English	German
Sentence Recognition	N-best	87%	98%
	First-best	69%	N/A
Correct Translation	N-best	87%	97%
	First-best	70%	86%

Table 1: Comparison of performance of English and German JANUS.

LR Parser	English 1	5%
	English 2	38%
	German 1	37%
	German 2	48%
PARSEC	English	77%

Table 2: Performance of LR parsing grammars and PARSEC on novel text input

5.2 Knowledge Based Parsing

In English-based JANUS, the first hand-written parsing grammar was designed to parse exactly the sentences in the conference registration database without attempting to parse novel sentences. Performance of the LR parser in English JANUS is 87% in N-best recognition mode.[1]. The first English parsing grammar could correctly parse only 5% of a database of 117 novel text sentences using the same vocabulary as the original sentences. A second version of the grammar achieved 38% parsing accuracy on the same set of novel sentences.[3] The test sentences were not available to the grammar writer. In contrast, English PARSEC correctly parses 77% of the 117 novel sentences.

Making the grammar flexible enough to handle novel sentences was a major design goal for the German parsing grammar. Linguistic principles from Lexical Functional Grammar theory [10] guided grammar design more fully than in the English JANUS system. For example, a verb's lexical entry explicitly lists the types of arguments (subject, object, indirect object) it can and must take, which constrains the parser in selecting the appropriate verb phrase rules. Verb subcategorization information was not present in the original English parsing grammar.

We report the results of testing the full German JANUS system, with the LPNN speech recognition component trained on the first 9 dialogs of the conference registration database. We tested overall translation of the 63 sentences in the last three dialogs.[1] These three dialogs were not used in training the speech recognizer but had been seen by the grammar writers working on the translation module. Testing in N-best recognition mode, the UPA system correctly translated 97% of the 63 test sentences. One sentence was incorrectly translated and one sentence was unparsable due to recognition errors. In first-best mode, performance degraded to 86% accurate translation.

To test the German parsing grammar on novel sentences, German students were asked to write typical conference registration dialogs; these dialogs were collected after the German grammar was completed. The parser correctly parsed 37% of the 161 new text sentences.[2] After the addition of 17 grammar rules, most of which added flexibility in accepting sentence fragments and conversation openings, performance increased to 48% correct output.

5.3 German PARSEC

No changes in the basic architecture of the English PARSEC system were required in order to retrain it on German sentences. A German lexicon and a corpus of target parses were defined for training the neural networks; this is required whenever the PARSEC architecture is applied to a new language.[3]

German PARSEC currently parses the 53 sentences in the first three conference registration dialogs with only one error. The problem is in determining whether a sentence is a question or a polite command; in the database such requests take the syntactic form of questions. Even for the two sentences where this error occurred, all other aspects of parsing, such as constituent analysis and thematic role assignment, are correct. Only one "English-only" assumption about had to be eliminated to make PARSEC general enough to handle German text. English PARSEC assumed that any sequence of numbers should be parsed as one large number. In a typical German address, the house number is followed immediately by a city code number. Combining the two numbers results in the wrong representation of the address.

5.4 Parallel Implementation

Neural net forward passes for the speech recognizer were programmed on two general purpose parallel processing machines, a MasPar computer at the University of Karlsruhe, Germany and an Intel Iwarp

[1] Testing in German JANUS has not been completed on the full database. For English JANUS the LPNN system was trained on all 12 dialogs, and testing was carried out using a separate recording of the entire database.

[2] After adding new lexical entries.

[3] The parse transformation module is being re-written to conform to the cleaner interlingua model developed for German JANUS.

at Carnegie Mellon. The MasPar used is a parallel SIMD (single-instruction, multiple-data) machine with 4096 processing elements. The Iwarp is a parallel MIMD (multiple-instruction, multiple-data) machine; a 16MHZ, 64 cell experimental version was used for testing.

The use of parallel hardware and algorithms has significantly decreased JANUS processing time. Compared to the forward pass calculations performed by a DecStation 5000, the Iwarp is 9 times faster (15.6 million connections per second), and the MasPar 24 times faster (41.4 million connections per second). The MasPar does the forward pass calculations for a two second utterance in less than 500 milliseconds. Both the Iwarp and MasPar implementations are scalable, and should provide proportional increases in speed with increased numbers of processing elements or cells. Currently, both the Fast-Fourier Transform calculations and the N-best search for the speech recognizer are being ported to a parallel machine and should lead to close to real-time speech translation on the conference registration task.

6. CONCLUSIONS

The JANUS approach to speech translation is cross-linguistically general: neither the speech recognition system nor the text translation components needed fundamental modification in order to translate from German speech into English and Japanese utterances. Overall system performance is 97% accurate translation; performance of the knowledge based parser has improved with more principled grammar design. The connectionist parsing system, PARSEC, is general enough to parse different languages with no major modifications. An implementation of the LPNN system on the parallel MasPar machine gives a 24-fold speedup in recognition. Ongoing research is aimed at upgrading to speaker-independent speech recognition and, using an interlingua approach which separates domain-specific and general linguistic information, real-time speech translation performance.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of ATR Interpreting Telephony Research Laboratories and Siemens Corporation.

REFERENCES

References

[1] A. H. Waibel, A. N. Jain, A. E. McNair, H. Saito, A. G. Hauptmann, and J. Tebelskis. JANUS: A speech-to-speech translation system using connectionist and symbolic processing strategies. In *Proceedings of the 1991 IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1991.

[2] J. Tebelskis, A. H. Waibel, B. Petek, and O. Schmidbauer. Continuous speech recognition using linked predictive neural networks. In *Proceedings of the 1991 IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1991.

[3] A. N. Jain, A. H. Waibel, and D. S. Touretzky. PARSEC: A structured connectionist parsing system for spoken language. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1992.

[4] M. Tomita and J. Carbonell. The universal parser architecture for knowledge-based machine translation. Technical Report CMU-CMT-87-101, Center for Machine Translation, Carnegie Mellon University, 1987.

[5] M. Tomita and E. Nyberg. Generation kit and transformation kit. Technical Report CMU-CMT-88-MEMO, Center for Machine Translation, Carnegie Mellon University, 1988.

[6] V. Steinbiss. Sentence-hypotheses generation in a continuous-speech recognition system. In *Proceedings of the 1989 European conference on Speech Communication and Technology*, volume 2, pages 51-54, 1989.

[7] O. Schmidbauer and J. Tebelskis. An LVQ Based Reference Model for Speaker-Adaptive Speech Recognition. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing,*. 1992.

[8] M. Tomita. *Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems.* Kluwer Academic Publishers, Boston, MA, 1987.

[9] T. Miramura, E. H. Nyberg, and J. G. Carbonell. An Efficient Interlingua Translation System for Multilingual Document Production. In *Proceedings of Machine Translation Summit III,* 1991.

[10] J. Bresnan, ed. *The Mental Representations of Grammatical Relations.* MIT Press, Cambridge, MA, 1982.

[11] Die Sprachdatesammlung im Projekt SPICOS. Siemens Corp. Internal Report ZFE IS KOM 31. April 8, 1991.