JANUS II — ADVANCES IN SPONTANEOUS SPEECH TRANSLATION

M.Woszczyna *M.Finke* *T.Kemp* *A.McNair* *A.Lavie* *L.Mayfield* *M.Maier* *I.Rogina*
T.Sloboda *A.Waibel* *P.Zahn* *T.Zeppenfeld*

INTERACTIVE SYSTEMS LABORATORIES
at Carnegie Mellon University, USA
and University of Karlsruhe, Germany

ABSTRACT

JANUS II is a research system to design and test components of speech to speech translation systems as well as a research prototype for such a system. We will focus on two aspects of the system: 1) new features and recognition performance of the speech recognition component JANUS-SR and 2) the end-to-end performance of JANUS II, including a comparison of two machine translation strategies used for JANUS-MT (PHOENIX and GLR*).

1. INTRODUCTION

Currently JANUS II components for English, German, Korean, Japanese, and Spanish speech input and translation are under development; though not all language pairs can always be kept at the same performance level, multilinguality is required to ensure generality in the recognition and translation approaches. A multitude of smaller and larger scale research projects contribute to the JANUS II system[1], including robust speech recognition[2], noise modeling[3], speaker and channel adaptation, strategies for porting recognition and translation to new languages[4], language identification, language modeling [5], user interfaces, repair strategies [6], interfaces between speech recognition and speech translation [7], machine translation issues[8], discourse modeling and software enginering. Explaining all of them would go beyond the scope of a conference paper. We can therefore only focus on some selected aspects of the system. For general descriptions of other parts of the recognizer and the GLR* and PHOENIX parsers refer to the list of references at the end of the paper.

2. THE SCHEDULING TASK DATABASE

We are collecting a large database of human-to-human dialogs centered around the scenario of appointment scheduling. Data is collected and transcribed for five languages, English, German, Korean, Japanese and Latin-American Spanish. The collection sites are Carnegie Mellon University, the University of Pittsburgh and Multicom (USA), Karlsruhe University (Germany)[1], ETRI (Korea), UEC and ATR (Japan); in each recording session, two subjects are each given a calendar and asked to schedule a meeting with the dialog partner. For most recordings the recording setup allows only one person to speak at a time by way of a push-to-talk switch and close-speaking microphones to avoid crosstalk.

		dialogs	utterances	words
English	ESST	1000	5618	147898
German	GSST	390	4000	73613
Japanese	JSST	200	—	—
Korean	KSST	150	1808	19352
Spanish	SSST	300	5365	90901

Table 1. The Spontaneous Scheduling Task Database

On average, the resulting dialogs cover about 8-12 utterances, each up to 60 seconds long. Many utterances are over 20 words long (the average is 25 for ESST and 35 for SSST) and cover several concepts. This makes the task more difficult for speech translation as the number of ambiguities increases with the length of the sentence.

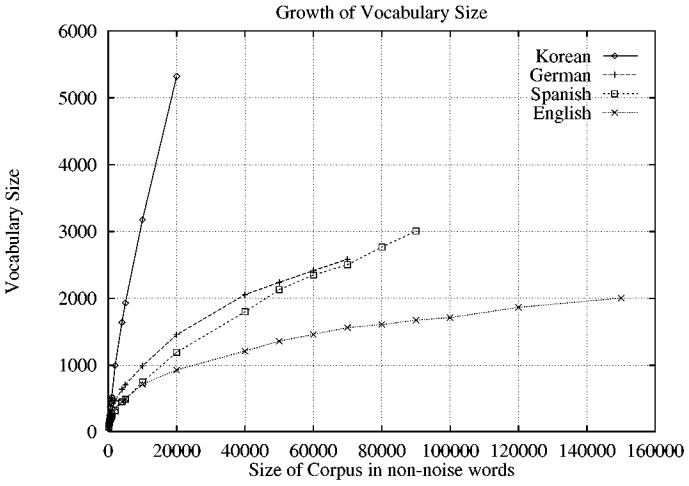

Figure 1. Development of Vocabulary Size

The steep vocabulary growth for Korean is due to the definition of a word unit: for Korean and Japanese the natural unit is similar to a phrase and it cannot be split to smaller units in streightforward ways. To make Korean and Japanese accessible to speech recognition and translation, the definition of smaller units will be necessary.

[1] 5000 additional German utterances from other sites are available from the *VERBMOBIL* project sponsored by the BMBF

3. RECOGNITION ENGINE

The recognizer used in the current JANUS II prototype system is a CDHMM based recognizer. The exact configuration varies from tast to task. For the last *VERBMOBIL* evaluation on German scheduling dialogs we used a preprocessing on a 7 frame melscale window, LDA, two 16 coefficient streams, CDHMM's using 70 codebook vectors per codebook and top-all evaluation. Important new features of the recognizer are the flexible decoder and an utterance level speaker adaptation technique.

3.1. The Decoder

The decoder has been substantially expanded to fulfill the growing needs of both international large vocabulary speech recognition evaluations and real time performance for the JANUS II prototype system. All search passes involving acoustic scoring are now forward oriented, avoiding time delays and model inversion.

T-pass: A first tree structured pass without tree copies selects probable words for each starting point. This pass uses only approximate bigrams and trigrams. As the tree is built on an allophone vocabulary, it is not very dense and will only improve the overall speed for large vocabularies.

F-pass: The second pass uses a flat, linear structured vocabulary allowing full bigrams and a better trigram computation. Trigrams are still only approximated to avoid overhead. As the F-pass only works on a subspace of the T-pass it is about 10 times faster. With better language models the word error on GSST after the F-pass is reduced by about 10% to 12% relative compared to the T-pass.

L-pass: By pruning the back pointer table of the second path a word-lattice is computed using full trigram information. Extracting only the best hypothesis gives full trigram information. As the L-pass does not access the scoring module, it is typically 100 times faster than even the F-pass, but it requires the full utterance for pruning. Using trigrams instead of bigrams in all passes of the search yields a 4% error reduction for GSST. Half of this is already achieved by using approximate trigrams in the F-pass, the other half requires running the L-pass.

3.2. Speaker Adaptation

For the hypothesis H_1 of an initial recognition the viterbi path $S = (s_{i_1}, s_{i_2}, \ldots s_{i_T})$ is computed. We're now looking for a transformation $\mu \rightarrow A\mu + b$ for all codebook vectors that increases the probability of observing the acoustic of the current sample given H_1. After the transformation this probability is given by

$$\prod_{t=1}^{T} p_{(A,b)}(x_t|s_{i_t}) =$$

$$\prod_{t=1}^{T} \frac{1}{\sqrt{(2\pi)^d|\Sigma_{i_t}|}} e^{(x_t-(A\mu_{i_t}+b))^T\Sigma_{i_t}^{-1}(x_t-(A\mu_{i_t}+b))}$$

therefore we need to find

$$(\bar{A}, \bar{b}) = \text{argmax}_{(A,b)} \prod_{t=1}^{T} p_{(A,b)}(x_t|s_{i_t})$$

and then replace each codebook vector μ by $\bar{A}\mu + \bar{b}$; With the modified coodebooks a new hypothesis is computed.

For adaptation on longer sequences, groups of codebooks are clustered together and each cluster is adapted individually.

On the June 1995 *VERBMOBIL* evaluation this adaptation on the utterance level yielded a relative error reduction of 2-3%; On SWB, where adaptation on whole dialogs is possible relative error reductions of 3-5% can be achieved.

3.3. Recognition Results

The results presented for *VERBMOBIL* in table 2 are the results from the Karlsruhe University System in the official 1995 *VERBMOBIL* evaluation on German scheduling dialogs. The system was the best among the five competing systems in this evaluation. The GSST, ESST and SSST results for German, English and Spanish scheduling are obtained on internal evaluations on unseen data from our own databases.

The results presented for the Switchboard task are the official results from the international SWB evaluation. The acoustic quality of the recording for SWB (telephone speech including crosstalk) is much lower than for GSST/ESST (close speaking microphone, no crosstalk). To allow JANUS II to be used over telephone lines, improving the performance on telephone speech will be an important research issue for JANUS-SR. As we have only recently started to build Japanese and Korean recognition components there are no results for these systems yet.

	Word Error
VERBMOBIL	30.0 %
GSST	28.0 %
ESST	30.2 %
SSST	29.8 %
Switchboard	61.9 %
NAB (WSJ)	22.8 %

Table 2. Performance of JANUS II

4. THE MACHINE TRANSLATION ENGINE

For a description of the parsing strategies developed in the JANUS project refer to [7, 8, 9]. In this section we will compare the performance on transcribed and spoken dialogs using two translation approaches: the GLR* skipping parser and the PHOENIX concept-based paser.

The evaluations comparing GLR* with PHOENIX on both the Spanish and English test sets indicate that the portion of acceptable translations produced with each of the parsers is very similar. On the Spanish transcribed test set, GLR* is slightly better (see table 3), while on English transcribed data, PHOENIX is slightly better. On speech recognized data, PHOENIX performed slightly better than GLR* in both Spanish and English. These slight performance differences should not, however, be regarded as statistically significant. The translation quality evaluations are necessarily very subjective. Although the grading is cross validated, differences in judgement between the graders have repeatedly amounted to 5% or more.

The two parsers have clear strengths and weaknesses. GLR* tries to match input utterances to an interlingua specification, so although words can be skipped with a penalty, the parser is less robust over disfluent input. Input that is parsed, though, is generated in the target language using syntactic constraints; this means that translations through GLR* are more likely to be complete grammatical sentences than those translated through PHOENIX, which parses and generates only at the speech act level.

GLR* tends to break down when parsing long utterances that are highly disfluent, or that significantly deviate from the grammar. In many such cases, GLR* succeeds in parsing only a small fragment of the entire utterance, and important input segments end up being skipped. PHOENIX is significantly better in analyzing such utterances. Because PHOENIX is a chart parser that is capable of skipping over input segments that do not correspond to any top level semantic concept, it can recover from out of domain segments in the input, and "restart" itself on the in-domain segment that follows. However, pre-breaking input to GLR* based on occurrences of human noise and parsing the shorter subutterances separately significantly reduced this problem. Pre-breaking benefits PHOENIX only slightly, mainly in better resolution of time expression attachment ambiguities. At the current time, PHOENIX uses only very simple disambiguation heuristics, whereas a parse quality mechanism helps to decide between possible parses in GLR*.

Computational requirements of GLR*, which is implemented in lisp, are far greater than those of PHOENIX, implemented in C. PHOENIX is also much faster, averaging 16 ms per parse compared to GLR*'s 1-2 minutes.

Because the two parsing architectures perform better on different types of utterances, they may be combined in a way that takes advantage of the strengths of each.

5. SPEECH TRANSLATION RESULTS

As the goal of the translation in JANUS is to preserve the content of an utterance, the recognition (SR), translation (MT) and end-to-end quality need to be assessed in terms of how well the meaning is preserved. Three grades were choosen for evaluation, **good**, **ok**, and **bad**;

`Transcription:  tuesday morning I have a meeting`

if an important semantic concept of an utterance is lost during recognition or translation, the whole recognition or translation is judged as **bad**;

`bad (SR):  you say morning I have a meeting`
`bad (MT):  tuesday  morning works for me`

if the meaning is preserved but the sentence comes out somehow funny, it is judged as **ok**. For an **ok** recognition there is still a chance of getting a good translation.

`ok (SR):  tuesday the morning I I have a meeting it`
`ok (MT):  tuesday morning won't for me work`

a 100% correct recognition or a translation that maintains the meaning and sounds correct, it is judged as **good**.

`good (SR): tuesday morning I have a meeting`
`good (MT): tuesday morning won't work for me`

	bad	ok or good
Recognition Quality		
English	—	—
German	—	—
Spanish	—	—
PHOENIX on transcriptions		
English — English	15%	85%
English — German	29%	71%
English — Spanish	—	—
German — German	—	—
German — English	—	—
Spanish — English	25%	75%
PHOENIX on speech data		
English — English	54%	46%
English — German	—	—
English — Spanish	—	—
German — German	—	—
German — English	—	—
GLR * on transcriptions		
English — English	16%	84%
English — Spanish	—	—
Spanish — English	21%	79%
GLR * on speech data		
English — English	56%	44%
English — Spanish	—	—

Table 3. Performance of JANUS II

REFERENCES

[1] M.Woszczyna, N.Aoki-Waibel, F.D.Buø, N.Coccaro, K.Horiguchi, T.Kemp, A.Lavie, A.McNair, T.Polzin, I.Rogina, C.P.Rose, T.Schultz, B.Suhm, M.Tomita, A.Waibel *JANUS 94: Towards Spontaneous Speech Translation* ICASSP94, V1-345;

[2] I.Rogina, A.Waibel. *Learning state-dependent Stream Weights for multi-codebook HMM Speech Recognition Systems* ICASSP94, V1-217;

[3] T.Schultz, I.Rogina *Acoustic and Language Modeling of Human and Nonhuman Noises* ICASSP95, V1-293;

[4] T.Sloboda *Dictionary Learning: Performance Through Consistency* ICASSP95, V1-453;

[5] B.Suhm, A.Waibel *Towards better Language Models for Spontaneous Speech* ICSLP94;

[6] A.E.McNair, A.Waibel *Improving Recognizer Acceptance through Robust, Natural Speech Repair* ICSLP94;

[7] B.Suhm, L.Levin, N.Coccaro, J.Carbonell, K.Horiguchi, R.Isotani, A.Lavie, L.Mayfield, C.P.Rose, C.Van Ess-Dykema, A.Waibel *Speech-Language Integration in a multi-lingual speech translation system* AAAI94;

[8] L.J. Mayfield, M. Gavaldà, W. Ward, A. Waibel *Concept-based Speech Translation* ICASSP95, V1-197;

[9] F.D.Buø, T.Polzin, A.Waibel *Learning Complex Output Representations in Connectionist Parsing of Spoken Language* ICASSP94, V1-365;