

JANUS

A System for Translation of Conversational Speech

Alex Waibel, Alon Lavie, Lori Levin
Carnegie Mellon University
Pittsburgh, USA

Abstract

JANUS is a large scale system for interactive spoken language translation that has been developed at Carnegie Mellon University and the University of Karlsruhe in the course of the last seven years. The system currently accepts spontaneous conversational speech in a limited domain in English, German or Spanish and produces output in German, English, Spanish, Japanese and Korean. In this overview article of the JANUS system we describe how the system has evolved over the years and developed its current architecture. We briefly describe the current system components, summarize our system development and evaluation methods, and present some of our most recent performance evaluation results. Finally, we discuss our current efforts to significantly expand the domain of coverage of the system, and the future directions we intend to explore within the context of this project.

Introduction

JANUS is a large scale system effort aimed at interactive spoken language translation. JANUS accepts spontaneous conversational speech in a limited domain in English, German or Spanish and produces output in German, English, Spanish, Japanese and Korean. The challenges of co-articulated, disfluent, ill-formed speech are manifold, and have required advances in acoustic modeling, dictionary learning, language modeling, semantic parsing and generation, to achieve acceptable performance. The intended meaning of an input sentence is represented by a semantic "interlingua" that facilitates the generation of culturally and contextually appropriate translation in the presence of irrelevant or erroneous information. Application of statistical, contextual, prosodic and discourse constraints permits a progressively narrowing search for the most plausible interpretation. During translation, JANUS produces paraphrases that are used for interactive correction of translation errors. Beyond our continuing efforts to improve robustness and accuracy, we have also begun to study possible forms of deployment. Several system prototypes have been implemented to explore translation in different settings: speech translation in one-on-one

video conferencing, portable mobile interpretation, and passive simultaneous conversation translation.

Historical Context

Current speech translation research evolved from systems of the late eighties and early nineties whose main goal was to demonstrate feasibility of the concept. In addition to domain constraints, these early systems had fixed speaking style, grammatical coverage and vocabulary size. Early systems include independent research prototypes developed by ATR, AT&T, Carnegie Mellon University and the University of Karlsruhe, NEC, and Siemens AG. Most were developed through international collaborations that provided the cross-linguistic expertise. Among these international cooperations, the Consortium for Speech Translation Advanced Research, or C-STAR, was formed as a voluntary group of institutions committed to building speech translation systems. It arose from a partnership among ATR Interpreting Telephony Laboratories in Kyoto, Japan, Carnegie Mellon University (CMU) in Pittsburgh, USA, Siemens AG in Munich, Germany, and University of Karlsruhe (UKA) in Karlsruhe, Germany. Additional members joined forces as partners or affiliates: ETRI (Korea), IRST (Italy), LIMSI (France), SRI (UK), IIT (India), Lincoln Labs (USA), DFKI (Germany), MIT (USA), and AT&T (USA). C-STAR continues to grow and to operate in a fairly loose and informal organizational style with each of its partners building complete systems or component technologies, thereby maximizing the technical exchange and minimizing costly software/hardware interfacing work between partners. In addition to the activity of consortia such as C-STAR, and the industrial research described above, there are government sponsored initiatives in several countries. One of the largest is Verbmobil, an eight year effort sponsored by the BMFT, the German Ministry for Science and Technology that involves 32 research groups.

JANUS was one of the early systems designed for speech translation. The first version of the system, JANUS-I, was developed at Carnegie Mellon University and University of Karlsruhe in the late '80s and early '90s in partnership with ATR (Japan) and Siemens AG

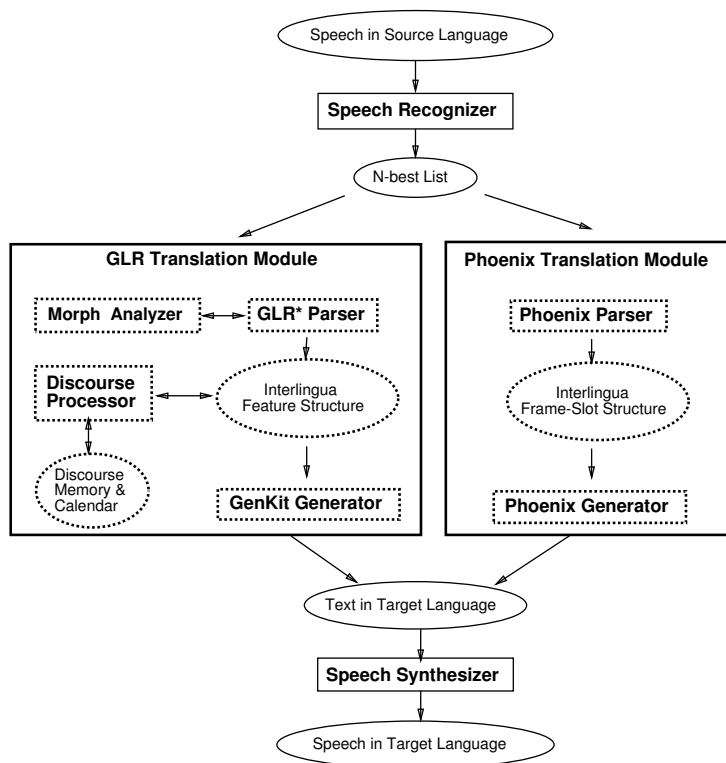


Figure 1: The JANUS System

(Germany). Since then it has been extended to more advanced tasks. While JANUS-I processed only syntactically well-formed (read) speech over a rather small (500 word) vocabulary, JANUS-II operates on spontaneous conversational human-human dialogues in limited domains with vocabularies of around 3000+ words. At present, it accepts English, German, Spanish, Japanese and Korean input and delivers translations into German, English, Spanish, Japanese or Korean.

The most recent version of our system is JANUS-III, in which substantial advances in our speech recognition engine have been incorporated. On the Spontaneous Scheduling Task our JANUS-III recognition system now achieves word error rates (WERs) below 10% on Japanese, 23% on English, 14% on German, and 17% on Spanish. On the broad domain telephone quality spontaneous speech task of the Switchboard corpus, our system performs at a WER of 36% [1]. This is a state-of-the-art performance result which illustrates the difficulty inherent in spontaneous speech tasks.

JANUS System Overview

A component diagram of our system for the Scheduling domain can be seen in Figure 1. The main modules of the current JANUS system are speech recognition, parsing, discourse processing, and generation. Each module is language independent in the sense that it consists of a general processor that can be loaded with language spe-

cific knowledge sources. The translation system follows an interlingua-based approach. The source language input string is first analyzed by a parser, which produces a language-independent interlingua content representation. The interlingua is then passed to a generation component, which produces an output string in the target language, which can then be synthesized, if desired, into speech (using commercially available speech synthesizers). In an attempt to achieve both robustness and translation accuracy when faced with speech disfluencies and recognition errors, we use two different parsing strategies: a GLR parser designed to be more accurate, and a Phoenix parser designed to be more robust. Detailed descriptions of the system components appear in our previous publications [2] [3] [4] [5] [6] [7].

The Speech Recognition Engine

In our effort of enhancing the overall system performance, we continue to improve the underlying speech and translation strategies. Particularly, in the light of our need to rearrange and redeploy our recognizer for different languages and different tasks, our new JANUS-III recognition engine automates many aspects of the system design that might otherwise be predetermined once. The JANUS-III recognizer is based on the Janus Recognition Toolkit (JRTk) [8], a flexible architecture for experimenting with language specific phenomena. The general configuration uses one or more streams of input features derived from Mel-scale, cepstral or PLP filters processed using Linear Discriminant Analysis (LDA). The acoustic units are context dependent, modeled via continuous density HMMs. Explicit noise models are added to help the system cope with breathing, lip-smack, and other human and non-human noises inherent in a spontaneous speech task.

Improved results have recently been achieved through the use of several advanced techniques. These include the following:

- **Speaker Normalization** - One major source of inter-speaker variability is the variation in their vocal tract shape. In order to normalize for the vocal tract length, a maximum likelihood scaling in the frequency axis of the speech signal is performed for each speaker.
- **Polyphonic Modeling** - We allow questions in the allophonic decision tree to not only refer to the immediate neighboring phones but also to phones further away. This increases the degree of context-dependency.
- **MLLR Model Adaptation** - Based on the first pass recognition, we allow our models to adapt to specific speakers. The more data is available for a speaker, the more specific the models can become.
- **Dictionary Learning** - Due to the variability, dialect variations, and coarticulation phenomena found in spontaneous speech, pronunciation dictionaries

have to be modified and fine-tuned for each language. To eliminate costly manual labor and for better modeling, we resort to data-driven ways of discovering such variants.

- **Morpheme Based Language Models** - For languages characterized by a richer morphology, which make wider use of inflections and compounding (compared to English), units more suitable than a “word” are used for dictionaries and language models [9].
- **Phrase Based and Class Based Language Models** - Words that belong to word classes (such as days of the week), or frequently occurring phrases (e.g., *out-of-town*, *I’m-gonna-be*, *sometime-in-the-next*) are discovered automatically by clustering techniques and added to a dictionary as special words, phrases or mini-grammars.

Speech Translation in JANUS

Speech translation in the JANUS system is guided by the general principle that spoken utterances can be analyzed and translated as a sequential collection of *semantic dialogue units* (SDUs), each of which roughly corresponds to a speech-act. SDUs are semantically coherent pieces of information. The interlingua representation in our system was designed to capture meaning at the level of such SDUs. Each semantic dialogue unit is analyzed into an interlingua representation.

As mentioned earlier, in an attempt to achieve both robustness and translation accuracy when faced with speech disfluencies and recognition errors, we use two different parsing strategies: a GLR parser (GLR*) designed to be more accurate, and a Phoenix parser designed to be more robust. Although both GLR* and Phoenix were specifically designed to deal with spontaneous speech, each of the approaches has some clear strengths and weaknesses. Because each of the two translation methods appears to perform better on different types of utterances, they may be combined in a way that takes advantage of the strengths of each of them. An example of a Spanish-to-English Phoenix translation of an utterance is shown in Figure 2.

For both parsers, segmentation of an input utterance into SDUs is achieved in a two-stage process, partly prior to and partly during parsing. Pre-parsing segmentation relies on acoustic, lexical, syntactic, semantic, and statistical knowledge sources. We use a statistical measure that attempts to capture the likelihood of an SDU boundary between any two words of an utterance. The measure is trained on hand-segmented transcriptions of dialogues. Pre-parsing segmentation substantially reduces parsing time, increases parse accuracy, and reduces ambiguity. Final segmentation into SDUs is done during parse time, guided by the grammar rules. The same statistical measure used to find the most likely SDU boundaries during pre-parsing segmentation is used to filter out unlikely segmentations during parse time.

Original utterance:

SÍ QUÉ TE PARECE TENGO EL MARTES DIECIOCHO Y EL MIÉRCOLES DIECINUEVE LIBRES TODO EL DÍA PODRÍAMOS IR DE MATINÉ O SEA EN LA TARDE VER EL LA PELÍCULA (Roughly “*Yes what do you think I have Tuesday the eighteenth and Wednesday the nineteenth free all day we could go see the matinee so in the afternoon see the the movie.*”)

As decoded by the recognizer:

```
%NOISE% S11 QUE1 TE PARECE %NOISE% TENGO EL
MARTES DIECIOCHO Y EL MIEIRCOLES DIECINUEVE LI-
BRES TODO EL D11A PODRI1AMOS IR DE MATINE1 %NOISE%
O SEA LA TARDE A VER LA
```

Parsed:

```
%<S> s11 quel te parece tengo el martes dieciocho y el miercoles
diecinueve libres todo el d11a podri1amos *IR *DE -MATINE1 o sea
la tarde a ver LA %</S>
```

Parse Tree (≡ Semantic Representation):

```
[respond] ( [yes] ( S11 ))
```

```
[your'turn] ( QUE1 TE PARECE )
```

```
[give'info] ( [my'availability] ( TENGO [temp'loc]
( [temporal] ( [point] ( [date] ( EL [d'o'w] ( MARTES ))
[date] ( [day'ord] ( DIECIOCHO ) [conj] ( Y ) EL [d'o'w]
( MIE1RCOLES )) [date] ( [day'ord] ( DIECINUEVE )))))
LIBRES ))
```

```
[give'info] ( [my'availability] ( [temp'loc]
( [temporal] ( [range] ( [entire] ( TODO )EL [unit]
( [t'unit] ( D11A )))))PODRI1AMOS ))
```

```
[suggest] ( [suggest'meeting] ( [temp'loc] ( [temporal]
( O SEA [point] ( LA [t'o'd] ( TARDE ))))A VER ))
```

Generated:

English = <Yes what do you think? I could meet Tuesday eighteenth and Wednesday the nineteenth I could meet the whole day do you want to try to get together in the afternoon>

Figure 2: A Phoenix Spanish-to-English Translation Example

For the scheduling domain, we have been using semantic grammars, in which the grammar rules define semantic categories such as **busy-free-phrase** and **schedule-meeting** in addition to syntactic categories such as NP and VP. There were several reasons for choosing semantic grammars. First, the domain lends itself well to semantic grammars because there are many fixed expressions and common expressions that are almost formulaic. Breaking these down syntactically would be an unnecessary complication. Additionally, spontaneous spoken language is often syntactically ill formed, yet semantically coherent. Semantic grammars allow our robust parsers to extract the key concepts being conveyed, even when the input is not completely grammatical in a syntactic sense. Furthermore, we wanted to achieve reasonable coverage of the domain in as short a time as possible. Our experience has been that, for limited domains, 60% to 80% coverage can be achieved in a few months with semantic grammars.

Evaluation Methods

In order to assess the overall effectiveness of the translation system, we developed a detailed end-to-end evalu-

	Transcription	Output of Speech-recognition
GLR*	82.9%	54.0%
Phoenix	76.3%	48.6%
Combined	83.3%	63.6%

Figure 3: End-to-end Translation Performance Results

ation procedure [10]. We evaluate the translation modules on both transcribed and speech recognized input. The evaluation of transcribed input allows us to assess how well our translation modules would function with “perfect” speech recognition. Testing is performed on a set of unseen dialogues that were not used for developing the translation modules or training the speech recognizer.

The translation of an utterance is manually evaluated by assigning it a grade or a set of grades based on the number of SDUs in the utterance. Each SDU is classified first as either relevant to the scheduling domain (in-domain) or not relevant to the scheduling domain (out-of-domain). Each SDU is then assigned one of four grades for translation quality: (1) Perfect - a fluent translation with all information conveyed; (2) OK - all important information translated correctly but some unimportant details missing, or the translation is awkward; (3) Bad - unacceptable translation; (4) Recognition Error - unacceptable translation due to a speech recognition error. These grades are used for both in-domain and out-of-domain sentences. However, if an out-of-domain sentence is automatically detected as such by the parser and is not translated at all, it is given an “OK” grade. The evaluations are performed by one or more independent graders. When more than one grader is used, the results are averaged together.

Translation Performance

To date, our translation system for the scheduling domain has achieved performance levels on unseen data of over 80% acceptable translations on transcribed input, and over 70% acceptable translations on speech input recognized with a 75-90% word accuracy, depending on the language.

The results in Figure 3 show the performance of the GLR and the Phoenix Spanish-English translation modules on a recent test set of 3 dialogues (103 utterances) recorded in a cross-talk setting (see following subsection). The results shown are for in-domain SDUs only and reflect the percent of acceptable translations. The speech recognition average word accuracy on this test set was 66.8%. The results in the last row of Figure 3 reflect the combination of the GLR* and Phoenix systems. As can be seen, the combination of the two parsers results in a significant improvement in translation performance on speech recognized input. On transcribed input the improvement is much less significant.

System Prototype Applications

We have constructed several system prototypes in order to explore translation needs in different settings. The three main prototype implementations we have built are a speech translation one-on-one video conferencing station, a portable mobile interpretation system, and a passive simultaneous conversation translation system.

The speech translation video conferencing station integrates our speech translation system within a video conferencing setting. Each user has two displays, one facing the user and another, touch sensitive display, embedded in the desk. The user operates his own station by way of the desk screen. A record button activates speech acquisition and displays both the recognition result and a paraphrase of the analyzed utterance. This is accomplished by performing a generation from the (language independent) interlingua back into the user’s language. The user can now verify if the paraphrase reflects the intended meaning of the input utterance. If so, he presses a send button, which replaces the paraphrase by the translation into the selected output language and sends it on to the other video conferencing site. At the other site, the translation appears as subtitles under the transmitted video image of our user. It is also synthesized in the target language for speech output. The translation display can also be used to run collaborative virtual environments such as joint whiteboards or applications that the conversants make reference to. Translation can be delivered in about two times real time.

JANETTE is a down-sized version of the JANUS system that runs on a Laptop PC (a 75 MHz Pentium) with 32 MB of memory. In this configuration the system currently still takes about twice as long per utterance to translate than on our video stations. The system can be carried in a knapsack or a carrying bag. Translation is presented either by an acoustic earpiece, or by a wearable heads-up display. The wearable heads-up display displays the translation in text form on see-through goggles, thereby allowing the user to see subtitles under the face of the person he/she is talking to. This alternate presentation of translation results allows for greater throughput, as the translation can be viewed without interrupting the speaker. While acoustic output may allow for feedback with the system, a simultaneously displayed translation may provide greater communication speed. The human factors of such new devices still await further study in actual field use.

The language interpreting systems described so far offer the opportunity for feedback, verification and cor-

rection of translation between two conversants who want to cooperate with each other. Not every situation affords this possibility, however. Many situations (such as N-party conferences, foreign TV or radio broadcasts) require a passive un-cooperative simultaneous translation of speeches or conversations. The rapid succession of sometimes overlapping turns makes the cognitive planning of a translation particularly difficult for humans attempting to translate conversational dialogue. Our experiments with cross-talk and push-to-talk dialogues, however, suggest that the same cognitive limitations experienced by human translators do not hold for machines - two separate speech translation processes can easily process separate channels of a dialogue and produce translations that keep up with the conversants. In our lab, a conversational translator has been installed that slices turns at major breaking points and sends the corresponding speech signals to an array of 5 processors, that incrementally generate translations during the course of a human conversation. Despite the disfluent nature of such an interactive and rapid conversation, translation of conversational dialogues within the scheduling domain can still be performed accurately more than 70% of the time.

Current and Future Plans

We are currently in the preliminary stages of extending the JANUS translation system from the appointment scheduling domain to a broader domain, travel planning, which has a rich sub-domain structure. Our preliminary experiments with English travel domain data indicate that it is characterized by higher out-of-vocabulary rates and greater levels of semantic complexity, compared with English scheduling domain data. In order to effectively deal with the significantly greater levels of ambiguity, we plan to use a collection of sub-domain grammars, which will in sum cover the entire travel planning domain. Our system design will be modified to facilitate working with multiple sub-domain grammars in parallel. The collection of appropriate sub-domains will be determined empirically. Automatic pruning methods will be used to derive each of the sub-domain grammars from a manually constructed comprehensive grammar. We expect to complete an initial implementation of the above methods and a preliminary evaluation of their effectiveness by late 1997.

Acknowledgements

The work reported in this paper was funded in part by grants from ATR - Interpreting Telecommunications Research Laboratories of Japan, the US Department of Defense, and the Verbmobil Project of the Federal Republic of Germany.

We wish to thank all members of the JANUS project at Carnegie Mellon and the University of Karlsruhe for their valuable contributions to the project. Their team efforts and devotion have made the JANUS system a state-of-the-art speech translation system.

References

- [1] T. Zeppenfeld, M. Finke, K. Ries, M. Westphal and A. Waibel. *Recognition of Conversational Telephone Speech using the JANUS Speech Engine*, in Proceedings of ICASSP-97, Munich, Germany
- [2] A. Lavie, D. Gates, M. Gavalda, L. Mayfield, A. Waibel, and L. Levin. Multi-lingual translation of spontaneously spoken language in a limited domain. In *Proceedings of COLING*, 1996.
- [3] Alon Lavie, Alex Waibel, Lori Levin, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, Puming Zhan and Oren Glickman. Translation of Conversational Speech with JANUS-II, In Proceedings of ICSLP-96, Philadelphia, USA, October 1996.
- [4] P. Zhan, K. Ries, M. Gavalda, D. Gates, A. Lavie and A. Waibel. *JANUS-II: Towards Spontaneous Spanish Speech Recognition*, Proceedings of ICSLP-96, Philadelphia, PA, October 1996
- [5] A. Lavie. *A Grammar Based Robust Parser For Spontaneous Speech*. PhD thesis, School of Computer Science, Carnegie Mellon University, 1995.
- [6] L. Mayfield, M. Gavalda, Y-H. Seo, B. Suhm, W. Ward, A. Waibel. *Parsing Real Input in JANUS: a Concept-Based Approach*, In Proceedings of TMI 95.
- [7] Y. Qu, C. P. Rose, B. Di Eugenio. Using Discourse Predictions for Ambiguity Resolution, In Proceedings of COLING-96, Copenhagen, Denmark, August 1996.
- [8] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries and M. Westphal. *The Karlsruhe-Verbmobil Speech Recognition Engine*, in Proceedings of ICASSP-97, Munich, Germany
- [9] P. Geutner. *Using Morphology towards better Large-Vocabulary Speech Recognition Systems*, in Proceedings of ICASSP-95, Detroit, Michigan, 1995
- [10] D. Gates, A. Lavie, L. Levin, A. Waibel, M. Gavalda, L. Mayfield, M. Woszczyna and P. Zhan. *End-to-end Evaluation in JANUS: a Speech-to-speech Translation System*, in Proceedings of ECAI Workshop on Dialogue Processing in Spoken Language Systems, Budapest, Hungary, August 1996.