End-to-end Evaluation in JANUS: a Speech-to-speech Translation System

Donna Gates[1], **Alon Lavie**[1], **Lori Levin**[1],
Alex Waibel[2], **Marsal Gavaldà**[1],
Laura Mayfield[1], **Monika Woszczyna**[3]
and Puming Zhan[1]

Abstract. JANUS is a multi-lingual speech-to-speech translation system designed to facilitate communication between two parties engaged in a spontaneous conversation in a limited domain. In this paper we describe our methodology for evaluating translation performance. Our current focus is on *end-to-end evaluations* - the evaluation of the translation capabilities of the system as a whole. The main goal of our end-to-end evaluation procedure is to determine translation accuracy on a test set of previously unseen dialogues. Other goals include evaluating the effectiveness of the system in conveying the domain relevant information and in detecting and dealing appropriately with utterances (or portions of utterances) that are out-of-domain. End-to-end evaluations are performed in order to verify the general coverage of our knowledge sources, guide our development efforts, and to track our improvement over time. We discuss our evaluation procedures, the criteria used for assigning scores to translations produced by the system, and the tools developed for performing this task. Our most recent Spanish-to-English performance evaluation results are presented as an example.

Keywords: Speech Translation, Performance Evaluation, Spontaneous Speech.

1 Introduction

JANUS [8, 9] is a multi-lingual speech-to-speech translation system designed to facilitate communication between two parties engaged in a spontaneous conversation in a limited domain. In this paper we describe our methodology for evaluating the translation performance[1] of our system. Although we occasionally evaluate the performance of individual components of our system, our current focus is on *end-to-end evaluations* - the evaluation of the translation capabilities of the system as a whole. Translation in JANUS is performed on basic semantic dialogue units (SDUs). We thus evaluate translation performance on this level. SDUs generally correspond to a semantically coherent segmentation of an utterance into speech-acts.

The main goal of our end-to-end evaluation procedure is to determine the translation accuracy of each of the SDUs in a test set of unseen utterances. Because our system is designed for a limited domain, we are also interested in evaluating the effectiveness of the system in conveying the domain relevant information and in detecting and dealing appropriately with utterances (or portions of utterances) that are out-of-domain. Detection of out-of-domain material allows the system to recognize its own limitations and avoid conveying false or inaccurate information. End-to-end evaluations are used to verify the general coverage of our knowledge sources, guide our development efforts, and to track our improvement over time.

JANUS is evaluated on recordings and transcriptions of human-human dialogues in which two speakers are trying to schedule a meeting. Test sets for full evaluations are always taken from a set of completely "unseen" reserved dialogues. A test set is considered fully unseen only if the speakers of the dialogues have not been used for training the speech recognizer and the dialogues themselves have not been used for development of the translation components. The performance results reported in this paper reflect this type of evaluation. We strongly believe that this method of evaluation is the most meaningful register of performance of a speech translation system. We also believe that it reflects the performance of the system in a real situation.

The remainder of the paper is organized in the following way. Section 2 presents a general overview of our system and its components. Section 3 contains a detailed description of the evaluation procedure, the criteria used for assigning scores to translations, and the tools developed for performing this task. In Section 4, as an example of these procedures at work, we present our most recent Spanish-to-English performance evaluation results. Finally, a summary and conclusions are presented in Section 5.

2 System Overview

The JANUS system is composed of three main components: a speech recognizer, a machine translation (MT) module and a speech synthesis module. A diagram of the general architecture of the system is shown in Figure 1. The speech recognition component of the system is described elsewhere [11]. For speech synthesis, we use a commercially available speech synthesizer.

At the core of the system is the MT module. It is composed of two separate translation sub-modules which operate independently. The first is the Generalized LR (GLR) module [3, 4], designed to be more accurate. The second is the Phoenix module [6], designed to be more robust. Both modules follow an interlingua based approach. The source language input string is first analyzed by a parser. In the case of the GLR module, lexical analysis is provided by a morphological analyzer[5, 2]. Each parser produces a language independent interlingua content representation. The interlingua is then passed to

[1] Center for Machine Translation, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213 (USA), email: dmg+@cs.cmu.edu
[2] Carnegie Mellon University and University of Karlsruhe (Germany)
[3] University of Karlsruhe (Germany)

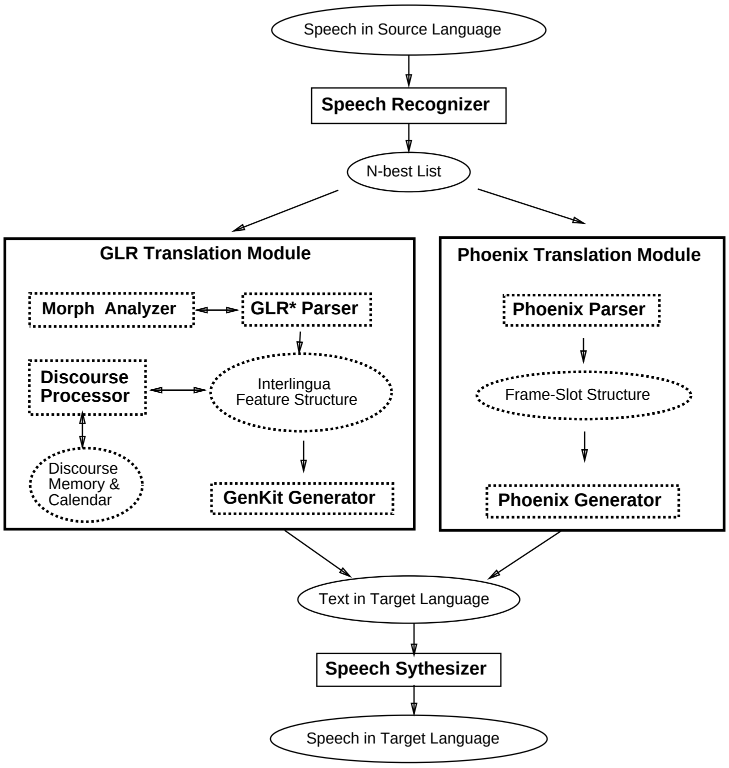

Figure 1. The JANUS System

a generation component, which produces a target language output string.

Both translation modules are equipped with procedures for detecting parts of utterances that are outside of the scheduling domain. Our goal here is to avoid partial translations of out-of-domain SDUs that force misleading interpretations of the input SDU. For example, we wish to avoid a situation in which *Tengo dos hijos. (I have two children.)* translated as "I have two o'clock free."

The discourse processor is a component of the GLR translation module. The discourse processor disambiguates the speech act of each sentence, normalizes temporal expressions, and incorporates the sentence into a discourse plan tree. The discourse processor also updates a calendar which keeps track of what the speakers have said about their schedules. The discourse processor is described in greater detail elsewhere [7].

3 The End-to-end Evaluation Procedure

In order to assess the overall effectiveness of our two translation modules in a consistent and cost effective manner, we developed a detailed end-to-end evaluation procedure. Prior evaluation procedures involved an expensive procedure for comparing the output of various components of the translation modules against hand-coded output. In the initial phase of development, this method was found to be very useful. However, it proved to be too labor intensive to hand-code output and evaluate each component of the system when frequent evaluations were needed. In addition, the two translation modules, required separate and very different procedures for evaluating their internal components. The solution was to use end-to-end translation output as the basis for the evaluation. This became possible only after the translation modules were fully integrated with the speech recognition component.

The development/evaluation cycle of our translation modules proceeds in the following way. System development and evaluation are performed on batches of data, each consisting of roughly 100 utterances. We refer to these batches as *test sets*. The test sets are chosen from a large pool of "unseen" data reserved for evaluations. System performance on each test set is first evaluated prior to any development based on the data. This allows us to isolate utterances (or parts of utterances) that are not translated adequately by the translation modules. Our translation development staff then spends some time on augmenting the analysis and generation knowledge sources, in order to improve their coverage. This is guided by the set of poorly translated examples that were isolated from the test set. Following development, the test sets are re-processed through the system using the updated knowledge sources. They are then re-scored, in order to measure the effect of the development on system performance. Once we are satisfied with the performance level on the current development test set, we proceed to a new "unseen" test set and begin the process anew. At the end of each evaluation-development cycle, we backup the current version of each translation module.

We believe evaluations should always be performed on data that is used neither for training speech recognition nor for developing the translation modules. All results reported in this paper are based on this type of evaluation. At the end of a development/evaluation cycle, we find that when we retest the data, we typically achieve over 90% correct translations. However, we believe that testing on unseen data represents a more valid test of how the system would perform under real conditions.

Evaluation is normally done in a "batch" mode, where an entire set of utterances is first recognized by the speech recognizer and then translated by the translation modules. Recorded speech for each utterance is processed through the speech recognizer, and an output file of the top recognition hypothesis for each utterance is produced. This file is then passed on to the translation modules, where the utterances are translated and the translation output is saved in an output file. Additionally, a file containing human transcribed versions of the input utterances is also processed through the machine translation modules. Both translation output files are then evaluated. The evaluation of transcribed input allows us to assess how well our translation modules would function with "perfect" speech recognition.

At least once a year we perform large scale system evaluations. The goals of the large evaluation are to measure the current performance of the system and to measure the progress made in development over a specific period of time. To measure progress over time, we take several backed up versions of the translators from significantly different points in time (at least 4 months apart) and run each of them over the same set of unseen test data. The translations are then scored and the end result is a series of scores that should increase from the oldest version to the most recent version.

3.1 Scoring Utterances

When scoring utterances we find that the most accurate results are derived by subdividing a spoken utterance into coherent semantically based chunks. Since an utterance in spontaneous speech may be very short or very long, we assign more than one grade to it based on the number of sentences or fragments it contains. We call these sentences or fragments "semantic dialogue units"(or SDUs). The utterance is broken down into its component SDUs in order to give more weight to longer utterances, and so that utterances containing both in- and out-of-domain SDUs can be judged more accurately. Each SDU translation is assigned one grade.

Transcribed data contains markers that represent the end of an SDU. These markers are encoded by hand with the symbol "{seos}" (which originally stood for "semantic end of sentence") as shown in the utterance: *sí* {*seos*} *está bien* {*seos*} *qué día te conviene más a ti* {*seos*} (English: "yes" "it's ok" "what day is more convenient for you"). Since this example has three SDUs, it will receive three grades.

Translations of speech recognizer output are scored by comparing them to the transcribed source language text. When scoring the translations from the output of speech recognition, the number of grades per utterance is determined by the number of SDUs in the transcribed source language dialogue. Since the output of speech recognition does not contain the SDU markings, the scorer is required to align the recognition output to the transcribed source language dialogue by hand. Then, for each SDU in the transcribed source language dialogue, the scorer must determine whether the translation of the speech recognition output is correct. This method also allows us to determine whether or not a mistranslation is due to an error in speech recognition.

3.2 Grading Criteria

When assigning grades to an utterance the scorer must make judgements as to the relevance of the SDU to the current domain, the acceptability of the translation given the task, and the quality and fluency of the translation when acceptable. Figure 2 illustrates the decision making process required to arrive at the various translation grades. Letters that appear in parentheses under the grade category correspond to the letter grade assigned by the scorer when using the grading assistant program described in the subsection 5.4.

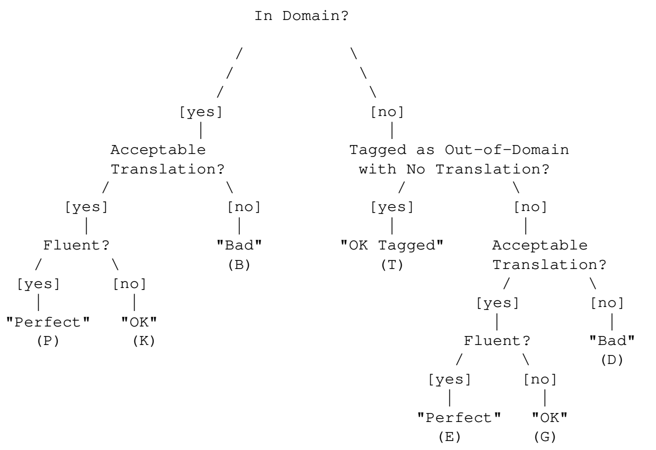

Figure 2. Decision Making Process for Grading an SDU

3.2.1 *Determining Domain Relevance*

The grader first classifies each utterance as either relevant to the scheduling domain (in-domain) or not relevant to the scheduling domain (out-of-domain). We established the following criteria for determining the relevance of each SDU to our domain. In-domain SDUs contain information that will be used specifically for scheduling a meeting such as suggesting a meeting or time, confirming a meeting or time, declining a suggested time or stating a scheduling constraint. If the SDU contains none of this information or does not imply a scheduling restriction, then the SDU is out-of-domain (e.g., discussing one's children). Out-of-domain also includes greetings and any conversation that falls outside of the scheduling task given the context in which it appears. This means that the SDU *Qué tal* which means "How are you?" or "How's that?" may be considered in-domain or out-of-domain. If it used in the context of a greeting "Hello... How are you?", then we consider it out-of-domain. If it is used in the context of suggesting a time "I can meet on Tuesday at two. How's that?", then we consider it in-domain.

3.2.2 *Determining Translation Quality*

After the domain of the SDU has been determined, the scorer then proceeds to assign one of the translation quality-accuracy grades listed in Figure 3. The grades "Perfect", "OK" and "Bad" are used for judging both in-domain and out-of-domain SDUs. "OK Tagged" is used when the translation system correctly recognizes the SDU as out-of-domain and does not translate it.

When a translation is judged as accurately conveying the meaning of the SDU, it is assigned the grade of "Perfect" or "OK". The grade "Perfect" is assigned to a high quality translation that contains all of the information of the input SDU and is presented in a fluent manner, provided the input was also fluent. For example *Tengo un almuerzo con Pedro a las dos.* receives the grade "Perfect" when translated as "I have a lunch date with Pedro at two." or "I am having lunch with Pedro at two." The grade "OK" is assigned when the translation is awkward or missing some non-essential information. For example *Tengo un almuerzo con Pedro a las dos.* receives the grade "OK" when translated as "I have a lunch date at two." In this example,

Perfect	Fluent translation with all information conveyed
OK	All important information translated correctly but some unimportant details missing or translation is awkward
OK tagged	The sentence or clause is out-of-domain and no translation is given.
Bad	Unacceptable translation

Figure 3. Evaluation Grade Categories

Pedro is not considered crucial for conveying that the speaker is busy at two o'clock. In reporting our results we use the category "acceptable" to represent the sum of the number of "Perfect" and "OK" translations. In addition to these two grades, out-of-domain SDUs may be assigned the grade "OK Tagged" as explained above. The "OK tagged" SDUs are included in the "acceptable" category for out-of-domain translations. A "Bad" translation is simply one that is not acceptable.

One of the drawbacks of relying on human judges to score translations is their subjectiveness. We find that scores from different judges may vary by as much as 10 percentage points. In addition, system developers do not make ideal judges because they are naturally biased. We believe that the most reliable results are derived from employing a panel of at least three judges who are not involved in system development to score the translations. Their scores are then averaged together to form the final result. When the only judges available are people who work on system development, it is absolutely necessary to cross grade the translations and average the results. When one judge is used, he or she cannot be affiliated directly with development.

3.3 The Grading Assistant Program

To assist the scorers in assigning the grades to the utterances, we have a simple Lisp program that displays dialogues and translations; prompts the scorer for grades; saves, tabulates and averages results; and displays these results in a table. When the same dialogue is translated by both translation modules, or by different versions of the same module, the grading program allows the scorer to compare the two translations, copying grades where the output is identical. When utterances are particularly long, the scorer may loop though the SDUs showing each transcribed SDU with its translation and assign it a grade. The program encourages the user to assign the same number of grades to an utterance as it has SDUs. Figure 4 shows an example of a translation displayed by the grading assistant with grades assigned to it by a scorer. The symbols in Figure 5 are used to assign both the quality grade and the domain relevance to an SDU at the same time.

```
11 (fbcg_04_11)

Transcribed:

   ((+s+ okay) (no) (yo tengo una reunioln de diez a once))

Generated:

   ("fine" "no" "i'm busy from ten o'clock to eleven o'clock")


(# SDUs = 3)

Grade: (p p p)
```

Figure 4. Example Display of a Translation with the Grading Assistant.

p	perfect in-domain
e	perfect out-of-domain
k	ok in-domain
g	ok out-of-domain
t	ok out-of-domain tag (not translated)
b	bad in-domain
d	bad out-of-domain

Figure 5. Grading Assistant Grades.

4 The Evaluation Results

Evaluations are performed periodically to assess our current translation performance and our progress over time. Because speech data often varies in style and content, we can only track progress over time by testing backed up versions of the translations system on the same set of unseen data. Comparing the results of evaluations performed on different test sets can often be misleading.

The following results were obtained from our Fall 1995 Spanish-to-English evaluation. We conduct similar evaluations with other source and target languages (English, German, Korean and Japanese). Recent performance results for these languages appear in [10].

Figure 6 shows a breakdown of the evaluation results for 6 unseen Spanish dialogues containing 120 utterances translated into English by the Phoenix translation module. Acceptable is the sum of "Perfect", "OK" and "OK tagged" sentences. For speech recognized input, we used the first-best hypotheses of the speech recognizer. The translations were scored by an independent scorer not involved in the development of the system.

Our last large scale evaluation of Spanish-to-English speech translation involved sixteen unseen dialogues that contained over 349 utterances with 1090 SDUs. The word accuracy for speech recognition was 63%. We chose three versions of both the GLR and Phoenix modules that coincided with the development periods ending in November 1994, April 1995 and September 1995. At the time of the evaluation, the September 1995 version represented the end of the most recent development period. Figure 7 shows the percent of acceptable translations for these time periods. Here we see a steady rise in acceptable scores over the various development periods.

5 Summary and Conclusions

The evaluation of a speech translation system must provide a meaningful and accurate measure of its effectiveness. In order to accomplish this, it is essential that the evaluation be conducted on sets of "unseen" data that reflect translation performance under real user conditions. The evaluation procedure must neutralize subjectivity in scoring, take into account utterance length and complexity, compensate for data which is not relevant to the domain being evaluated, and employ a consistent set of criteria for judging translation quality.

Our end-to-end evaluation procedure described in this paper allows us to consistently and inexpensively measure the overall performance of our speech translation system. Our experience has shown that conducting frequent end-to-end evaluations is an effective tool for the development of our system and measuring its performance over time.

In Domain (231 SDUs)	transcribed	speech 1st-best
Perfect	52	29
OK	21	23
Bad	27	48
Out of Domain (137 SDUs)		
Perfect	50	36
OK	8	10
OK tagged	26	29
Bad	16	25
Acceptable (Perfect + OK)		
In Dom	73	52
Out of Dom	84	75
All Dom	77	60

Figure 6. October 1995 evaluation of Phoenix translator on six dialogues.

Transcribed Input	GLR*	Phoenix
November-94	72	72
April-95	82	77
October-95	85	83
Output of the Speech Recognizer	**GLR***	**Phoenix**
November-94	49	57
April-95	55	62
October-95	58	66

Figure 7. Development of Spanish-to-English translation.

Acknowledgements

The work reported in this paper was funded in part by a grant from ATR - Interpreting Telecommunications Research Laboratories of Japan. We are grateful for ATR's continuing support of our project. We also thank Carol Van Ess-Dykema and the US Department of Defense for their support of our work.

We would like to thank all members of the JANUS teams at the University of Karlsruhe and Carnegie Mellon University for their dedicated work on our many evaluations.

REFERENCES

[1] R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen and V. Zue editors. Chapter 13: Evaluation *Survey of the State of the Art in Human Language Technology*, Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, 1996 *http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html*

[2] R. Hausser. Principles of Computational Morphology. Technical Report, Laboratory for Computational Linguistics, Carnegie Mellon University, Pittsburgh, PA, 1989.

[3] A. Lavie and M. Tomita. GLR* - An Efficient Noise Skipping Parsing Algorithm for Context Free Grammars, *Proceedings of the third International Workshop on Parsing Technologies (IWPT-93), Tiburg, The Netherlands, August 1993.*

[4] A. Lavie. An Integrated Heuristic Scheme for Partial Parse Evaluation, *Proceedings of the 32nd Annual Meeting of the ACL (ACL-94), Las Cruces, New Mexico, June 1994.*

[5] L. Levin, D. Evans, and D. Gates. *The* ALICE *System: A Workbench for Learning and Using Language. Computer Assisted Language Instruction Consortium (CALICO) Journal*, Autumn 1991, 27–56.

[6] L. Mayfield, M. Gavaldà, Y-H. Seo, B. Suhm, W. Ward, A. Waibel. "Parsing Real Input in JANUS: a Concept-Based Approach." In *Proceedings of TMI 95*.

[7] C. P. Rosé, B. Di Eugenio, L. S. Levin, and C. Van Ess-Dykema. Discourse processing of dialogues with multiple threads. In *Proceedings of ACL'95, Boston, MA*, 1995.

[8] B. Suhm, P. Geutner, T. Kemp, A. Lavie, L. J. Mayfield, A. E. McNair, I. Rogina, T. Sloboda, W. Ward, M. Woszczyna and A. Waibel. JANUS: Towards Multi-lingual Spoken Language Translation. In ARPA Workshop on Spoken Language Technology, 1995.

[9] A. Waibel. Translation of Conversational Speech. Submitted to IEEE Computer, 1996.

[10] A. Waibel et al. JANUS II: Advances in Spontaneous Speech Translation. Submitted to ICASSP, 1996.

[11] M. Woszczyna, N. Aoki-Waibel, F. D. Buo, N. Coccaro, K. Horiguchi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. P. Rosé, T. Schultz, B. Suhm, M. Tomita, and A. Waibel. JANUS-93: Towards Spontaneous Speech Translation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'94)*, 1994.