Lessons Learned from a Task-Based Evaluation of Speech-to-Speech Machine Translation

Lori Levin, Boris Bartlog, Ariadna Font Llitjos, Donna Gates, Alon Lavie, Dorcas Wallace, Taro Watanabe, Monika Woszczyna

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
lsl@cs.cmu.edu

Abstract

For several years we have been conducting Accuracy Based Evaluations (ABE) of the JANUS speech-to-speech MT system (Gates et al., 1997) which measure quality and fidelity of translation. Recently we have begun to design a Task Based Evaluation for JANUS (Thomas, 1999) which measures goal completion. This paper describes what we have learned by comparing the two types of evaluation. Both evaluations (ABE and TBE) were conducted on a common set of user studies in the semantic domain of travel planning.

1. Introduction

For several years we have been conducting Accuracy Based Evaluations (ABE) (Gates et al., 1997) of the JANUS speech-to-speech machine translation system (Waibel, 1996; Levin et al.,). Our ABE focuses on whether the meaning of a source language segment is totally and accurately conveyed in the target language, and also includes a separate measure of fluency. This type of evaluation was useful in the early stages of system development for tracking our improvement over time. The measure we used was percent of sentences that were accurate (we call these `acceptable`) and the percent that were both accurate and fluent (we call these `perfect`). However, when our system reached a level of coverage that allowed us begin user studies, we noticed that the ability of a user to complete a task (for example, getting a plane reservation) was higher than would be expected based on an ABE. For example, the ABE might be around 70% acceptable, but the users could almost always complete the task. Recently we have begun to design a Task Based Evaluation for JANUS (Thomas, 1999) which measures goal completion. This paper describes what we have learned by comparing the two types of evaluation.

2. Design Criteria

Most previous work on TBE has been conducted on human-machine dialogue (for example (Walker et al., 1997)). For machine translation, we need a TBE that is suitable for two humans each expressing communicative goals, but mediated by a machine. (Our coding scheme for communicative goals is described below.) In particular, we have to separate human clumsiness and error from machine error, because we are not evaluating the humans, but rather the translation of what they said. Additionally, we have to allow for a large and unpredictable number of communicative goals in each dialogue. For example, using the goal coding scheme described below, the dialogues we are evaluating each contain over one hundred communicative goals. After coding the communicative goals in a dialogue, we had to design a scoring function that takes into account whether the communicative goals ultimately succeed or fail and how many times each goal is attempted before succeeding (being understood by the interlocutor) or being abandoned.

3. The Data

The data used for this evaluation came from three user-study dialogues that were unseen by system developers. In each dialogue, the role of the traveller was played by a second-time user of our machine translation system and the role of the travel agent was played by one of the system developers. The traveller was told to book a trip to Kyoto. Input to the system was through a headset with microphone. The agent and traveller could not see or hear each other. The only communication was through the user interface, which included speech synthesis, written translations, and web pages showing itineraries and travel information. There is a total of 254 utterances in the three dialogues.

In these user studies, the source and target languages were both English. This *does* constitute a real translation in that it goes through all of the machine translation components: English sentences are parsed to produce interlingua representations (see below) and then new English sentences are generated from the interlingua. One could argue, however, that there may be some translation problems which do not appear in English-to-English translation. For this reason, we conducted an additional informal user study in which the travel agent was speaking German and the traveller was speaking English. This was not as carefully controlled as the original user studies; the two users could hear each other and the German speaker also understood English.

4. Coding Scheme for Communicative Goals

The most difficult issue in designing our TBE was defining what counts as a communicative goal. Because we need a definition that allows goals to be coded with high

Transcription (1)	Agent: WOHIN #6f REISEN SIE #7f
	Where are you travelling?
Ideal IF	a:request-information+features+trip (location=question)
Recognized as	WANN REISEN SIE
	When are you travelling
German Paraphrase	Wann reisen Sie ab?
	when are you leaving
English Translation	When will you leave?
Transcription (2)	Client: uh i'm leaving #8f next monday #9f
Ideal IF	c:give-information+temporal+departure (time=next monday)
Recognized as	MEETING ISN'T IT
English Paraphrase	Is that right?
German Translation	Stimmt das?
	Is that right?
Transcription (3)	Client: i'm leaving #8s on monday #9s
Ideal IF	c:give-information+temporal+departure (time=monday)
Recognized as	I'M LEAVING ON MONDAY
English Paraphrase	I will leave Monday
German Translation	Meinen Abreise ist an Montag.
	My departure is Monday.
Transcription (4)	Agent: Von wo #10f nach wo #6f reisen Sie ? #7f
	From where to where are you travelling?
Ideal IF	a:request-information+features+trip (origin=question, destination=question)
Recognized as	VON ROM NACH ROM REISEN SIE
	From Rome to Rome are you travelling?
German Paraphrase	Das ist in Rom. Die Reise ist nach Rom.
	That is in Rome. The trip is to Rome.
English Translation	That is in Rome. You will be travelling to Rome.
Transcription (6)	Client: i'm travelling #11s to Heidelberg #13s
Ideal IF	c:give-information+features+trip (destination=heidelberg)
Recognized as	I'M TRAVELLING TO HEIDELBERG
English Paraphrase	I will be arriving to Heidelberg.
German Translation	Meine Ankunft ist in Heidelberg.
	My destination is Heidelberg.

Figure 1: Example of a German-English dialogue tagged with successful and failed goals.

inter-coder reliability, we based the definition of communicative goals on our interlingua representation, IF (Levin et al., 1998; Levin et al.,). The IF representation for each sentence has two parts, a domain action and a list of arguments. A domain action is an extended speech act which includes some domain-specific concepts, for example, `give-information+availability+room`. The arguments of the domain actions include more specific concepts such as times, dates, names, flight numbers, etc. (We are working the the domain of travel planning.) Examples of IF representations are shown in Figure 1. Evaluation of the coverage of the interlingua for the travel domain is discussed in (Levin et al., 2000).

Figure 1 shows a portion of a dialogue between a German speaking travel agent and an English speaking travel customer. For each utterance, we show five things: (1) a human-generated transcription of an utterance annotated with our communicative goal coding scheme; (2) a human-generated interlingua representation; (3) the output of the JANUS speech recognizer; (4) a machine generated paraphrase-translation in the source language; and (5) a machine-generated translation in the target language. (Human-generated English translations of German are in italics.)

The coding scheme for communicative goals, as mentioned above, is based on the IF representation. Each domain action is counted as a goal and each argument is counted as a goal. Goal tags are inserted into the human-generated transcription. These tags are marked by #. Each tag is accompanied by a goal number followed by a mark of `s` or `f`, indicating whether the goal succeeded or failed. A goal is coded with `s` if the coder feels that the machine-generated translation correctly conveys it, and is coded with `f` otherwise. For example, the transcription for utterance (1) shows there are two goals (#6 and #7). Goal #6 is the domain action `request-information+features+trip`. Goal #7 is the argument `(location=question)`. Both goals are coded as failures.

The tagged transcription for utterance 2 indicates that the two goals in this utterance (#8 and #9) both failed, in this case due to speech recognition errors. These same two goals are then repeated in utterance 3 and both of them finally succeed.

In order to be effective for system evaluation, our coding scheme has to support reasonably consistent coding by human taggers. To evaluate its effectiveness, we conducted a preliminary inter-coder agreement experiment on one dialogue with three different coders. Results indicated that the average goal tagging pairwise agreement between the coders is about 79%. The goal tag of a pair of coders was considered to be in agreement if the tag: (1) covered the same part of the utterance/IF; (2) had the same success/fail tag; and (3) had the same new/old goal type of tag.

	Agent	Traveller	All
ABE	58.7%	44.7%	51.8%
TBE score	.75	.56	.65
TBE success	82.8%	64.7%	73.8%

Table 1: Results of Accuracy- and Task-Based Evaluations for English-English Paraphrase

5. The Scoring Function

Our TBE scoring scheme assigns each identified goal g_i in the dialogue a score $s(g_i)$, ranging between minus one and one. The score $s(g_i)$ is determined according to the formula below (Thomas, 1999). The formula takes into account whether the goal ultimately succeeds or fails and the number of times the goal was attempted before the user finally succeeded or gave up. The number of attempts is denoted by n.

$$s(g_i) = \begin{cases} \frac{1}{n} & \text{goal succeeds} \\ -(1 - \frac{1}{n}) & \text{goal fails} \end{cases}$$

The TBE score for a complete dialogue is calculated as the average of the score per goal, taken over all goals in the dialogue. The rationale behind the scoring formula is the following:

- A goal that succeeds in its first attempt receives the maximal score of one. Goals that succeed after further attempts should score less, with a penalty that decays as a function of the number of attempts.
- Goals that fail should be penalized more as a function of the number of attempts, since the number of attempts can be indicative of the importance of the goal. Thus, a goal that was attempted once and abandoned receives a score of zero, while a goal attempted ten failed times and then abandoned receives a score of -0.9. The penalty decays as a function of the number of attempts.

Our explicit goal in the design of the scoring function was to come up with a function that in fact followed the above rationale. Our formula is only one of a variety of functions which would have the above desired properties. We do not associate great significance to the specific function chosen , but rather to the desired properties themselves. While different functions would result in different absolute scores for individual goals as well as complete dialogues, it is the relative score of different dialogues that is ultimately of greater interest in a TBE.

6. Results

Table 1 shows the results of the ABE and TBE on English to English translation. There were four human coders. The ABE score is the percent of utterances whose translations preserved the original meaning. The TBE score was computed by the formula above, taking into account success/failure of goals in addition to the number of attempts for each goal. The row labeled *TBE success* shows the percentage of goals that ultimately succeeded (out of a total of approximately 460 goals in three dialogues). Each row breaks down into a score for the agent (who was an experienced user), a score for the traveller (a second-time user), and an overall score for agent and traveller.

The results for the less controlled English-German experiment are as follows. In one dialogue coded by one coder, there were 102 goals and a total of 133 attempts. 83% of the goals ultimately succeeded. The score returned by our scoring function is .73. The ABE showed 63% acceptable translations.

7. Discussion and Lessons Learned

There are a few things to notice about Table 1. For example, the users playing the travel agent role have more success in both ABE and TBE than users playing the traveller role. This is because the pretend travel agents were system developers and the travellers were second time users of our machine translation system.

Another notable point about Table 1 is that task success (73.8%) is higher than translation accuracy (51.8%). This confirms the need for TBE in addition to ABE. The reason for task success being higher than translation accuracy is that both experienced and inexperienced users accepted some bad translations as long as they can be understood in context. For example, in the context of the question *How much does it cost?*, users will accept the answer *128 hours*.

The percent of task success, however, does not provide a measure of user frustration (Walker et al., 1997). This is why we formulated the TBE scoring function to take into account success/failure of goals as well as the number of attempts at each goal. (In future work, we will give some thought to making the TBE score (on a minus one to one scale) more comparable to the ABE score (expressed as a percentage).) In sum, we find three kinds of measures useful — a measure of quality and fidelity, a measure of goal success/failure, and a measure of user effort combined with success/failure.

We will close by giving some examples that illustrate a peculiarity in our coding scheme: the utterance *two* is associated with the IF `give-information+numeral (numeral=2)`, which has a domain action and an argument. Therefore, it counts as two communicative goals. A slightly different problem is that the phrase *You'll be returning* in *You'll be returning on the twenty first* counts as two goals `give-information+reservation+temporal+transportation` and `trip-type=return`. Similarly, *is cheaper* in *The bus is*

cheaper counts as `give-information+price` and *price=cheaper* and *With a Mastercard* in the context of *How will you be paying?* counts as `giveinformation-+payment` and `method=mastercard`.

Acknowledgements

We would like to thank Alexandra Slavkovic for running the user studies and Kavita Thomas for her preliminary work on the design of the TBE.

8. References

Gates, Donna, Alon Lavie, Lori Levin, Marsal Gavaldà, Monika Woszczyna, and Puming Zhan, 1997. *End-to-End Evaluation in JANUS: a Speech-to-Speech Translation System*.

Levin, Lori, D. Gates, A. Lavie, F. Pianesi, Dorcas Wallace, Taro Watanabe, and Monika Woszczyna, 2000. Evaluation of a Practical Interlingua for Task-Oriented Dialogue. In *Workshop on Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP*. Seattle.

Levin, Lori, D. Gates, A. Lavie, and A. Waibel, 1998. An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'98)*. Sydney, Australia.

Levin, Lori, A. Lavie, M. Woszczyna, D. Gates, M. Gavaldà, D. Koll, and A. Waibel. The Janus-III Translation System. *Machine Translation*. To appear.

Thomas, Kavita, 1999. Designing a Task-Based Evaluation Methodology for a Spoken Machine Translation System. In *Proceedings of ACL-99 (Student Session)*. College Park, MD.

Waibel, Alex, 1996. Interactive Translation of Conversational Speech. *Computer*, 19(7):41–48.

Walker, Marilyn, D. Litman, C. Kamm, and A. Abella, 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL'97)*.