End-to-End Evaluation of a Speech-to-Speech Translation System
The Case Study of NESPOLE!

Alon Lavie, Lori Levin, Donna Gates, Dorcas Wallace, Kay Peterson and Andrew Tavares
Language Technologies Institute
Carnegie Mellon University

1 Introduction

The C-STAR consortium partner groups have been engaged in active research on Speech-to-Speech Translation for over a decade. In the course of our joint work, we have developed several interlingua-based prototype systems. The issue of how to evaluate the performance of such systems has itself been an active area of investigation, and our methodologies have evolved significantly over the years. Much of the focus has been devoted to "end-to-end" evaluation, where the goal is to assess the performance of the system as a whole, rather than it's individual components. Our latest most active collaboration has been within the NESPOLE! project.

NESPOLE![1] is a speech-to-speech machine translation project designed to provide fully functional speech-to-speech capabilities within real-world settings of common users involved in e-commerce applications. The project is a collaboration between three European research groups (IRST in Trento, Italy; ISL at Universität Karlsruhe (TH); and CLIPS at Université Joseph Fourier in Grenoble, France), one US research group (ISL at Carnegie Mellon University in Pittsburgh, PA) and two industrial partners (APT; Trento, Italy – the Trentino provincial tourism board, and AETHRA; Ancona, Italy – a tele-communications company). The project is funded jointly by the European Commission and the US NSF. Over the past year, we have developed a fully functional showcase of the NESPOLE! system within the domain of travel and tourism, and have significantly improved system performance and usability based on a series of studies and evaluations with real users. A full multi-site, multi-lingual end-to-end evaluation of the system was conducted in the last two months of 2001. In this paper, we describe our evaluation methodology and report the results and our experience from this full scale evaluation of the NESPOLE! system, as a case study for large multi-lingual interlingu-based speech-to-speech translation systems.

2 The NESPOLE! System

The design principles of the NESPOLE! [2] system are described in [1]. The system uses a client-server architecture to allow a common user, who is initially browsing through the web pages of a service provider on the Internet, to connect seamlessly to a human agent of the service provider who speaks another language, and provides speech-to-speech translation service between the two

[1] NESPOLE! – NEgotiation through SPOken Language in E-commerce. See the project web-site at `http://nespole.itc.it` for further details.

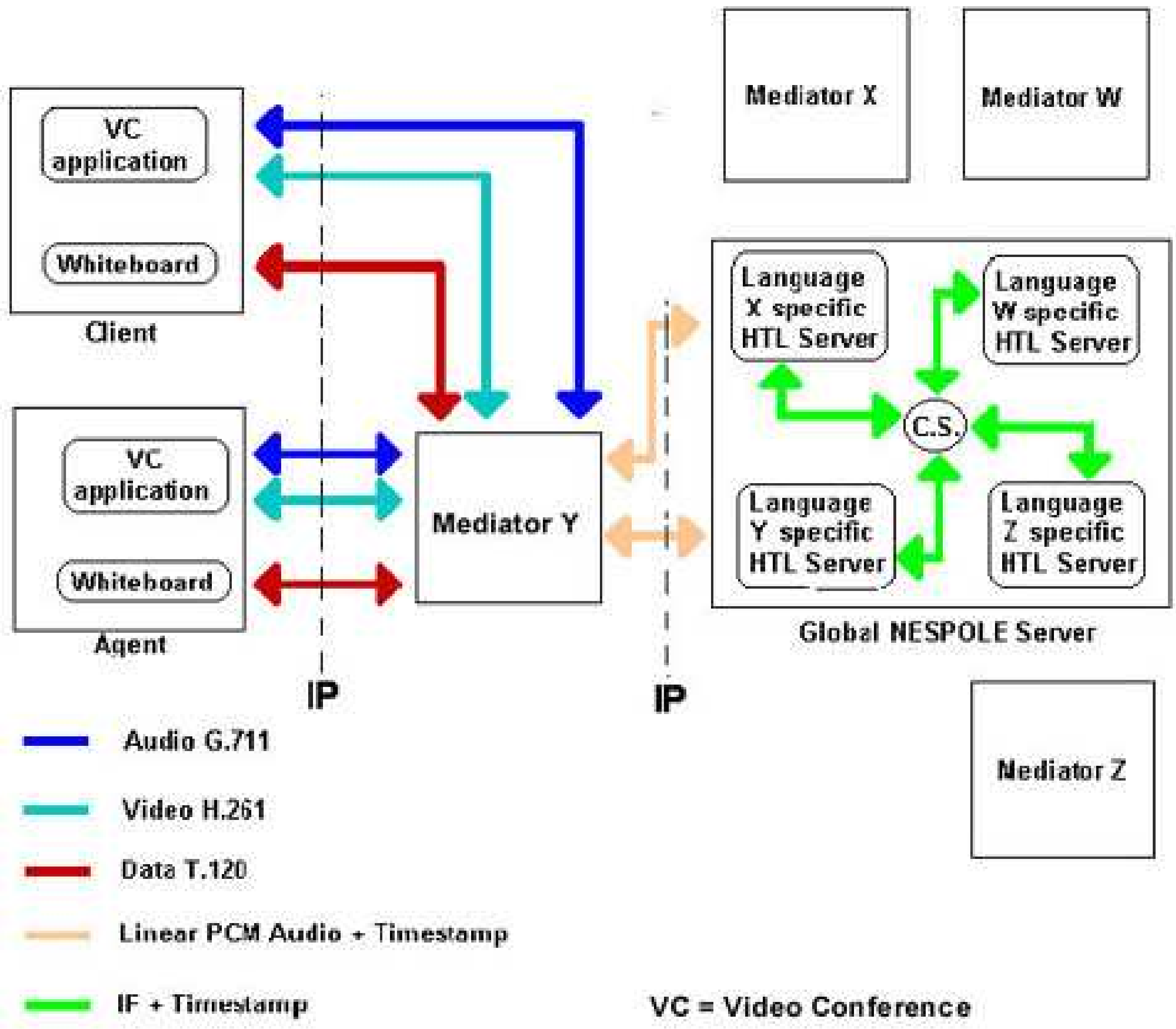

Figure 1: *The Nespole! System Architecture*

parties. Standard commercially available PC video-conferencing technology such as Microsoft's NetMeeting® is used to connect between the two parties in real-time.

In the first showcase described here, the scenario is the following: a client user is browsing through the web-pages of APT – the tourism bureau of the province of Trentino in Italy – in search of winter-sport tour-packages in the Trentino region. If more detailed information is desired, the client can click on a dedicated "button" within the web-page in order to establish a video-conferencing connection to a human agent located at APT. The client is then presented with an interface consisting primarily of a standard video-conferencing application window and a shared whiteboard application. Using this interface, the client can carry on a conversation with the agent, where the NESPOLE! server provides two-way speech-to-speech translation between the parties. In the current setup, the agent speaks Italian, while the client can speak English, French or German.

2.1 System Architecture

The basic system design is shown in Figure 1. A key component in the NESPOLE! system is the "Mediator" module, which is responsible for mediating the communication channel between the two parties as well as interfacing with the appropriate Human Language Technology (HLT) speech-translation servers. The HLT servers provide the actual speech recognition and translation capabilities. This system design allows for a very flexible and distributed architecture: Mediators and HLT-servers can be run in various physical locations, so that the optimal configuration, given the locations of the client and the agent and anticipated network traffic, can be taken into account at any time. A well-defined API allows the HLT servers to communicate with each other and with the Mediator, while the HLT modules within the servers for the different languages are implemented using very different software packages.

For example, let us suppose an English-speaking client in the US is connecting to an APT agent in Italy. A connection request from the client's PC (in the US) would be made to the Mediator, which can be physically located anywhere on the net (in practice, located at the agent site in Italy). The Mediator establishes a connection over the internet with both an English HLT server and the Italian HLT server (also physically located anywhere on the internet), before calling the agent in Trento. The English HLT server provides English speech recognition, translation from English text into our Interlingua "IF" (Interchange Format) as well as English generation and speech synthesis from IF. The Italian HLT server provides similar functionalities to and from Italian.

The computationally intensive part of speech recognition and translation is done on dedicated server machines, whose nature and location is of no concern to the user. A wide range of client-machines, even portable devices or public information kiosks, are therefore able to run the client software, so that the service can be made available nearly everywhere. The main technical difficulty for VoIP ("Voice over Internet Protocol") applications is coping with adverse internet bandwidth conditions. In order to guarantee real-time communication under insufficient bandwidth conditions, video-conferencing software often drops short segments of speech that were delayed in transport. This, however, can be very detrimental to the performances of speech recognizers [4]. In order to reduce bandwidth requirements, it is possible to use the NESPOLE! system without video transmission as we currently do not transmit critical information over this channel.

Speech-packets transmitted over the H323 protocol can be processed as soon as they arrive without waiting for subsequent packets that contain the remaining speech. The inclusion of "Run-On" processing in the speech recognizers and the optimization of inter-process communication within the HLT servers reduced the answering times of the overall system to approximately three times real-time in recent demonstrations.[2] Our experience indicates that users appreciate the fact that the system provides some feedback on its progress in the form of text messages (i.e. recognition results), even before paraphrasing and translation are complete.

2.2 User interface

The user interface display is designed for Windows®and consists of four windows: (1) a Microsoft®Internet Explorer web browser; (2) a Microsoft®Windows NetMeeting video-conferencing application; (3) the AeWhiteboard; and (4) the Nespole Monitor. Using Internet Explorer, the client initiates the audio and video call with an agent of the service provider, by a simple click of a button on the browser page. Microsoft Windows Netmeeting is automatically opened and the audio and video

[2]For comparison, German Speech recognition runs in about real-time on a standard 1GHz Pentium-III machine using a 12k vocabulary.

connection is established. The two additional displays – the AeWhiteboard and the Nespole Monitor are also launched at the same time. Client and agent can then proceed in carrying out a dialogue with the help of the speech translation system. For a screen snapshot of these four displays, see [5].

We found it important to visually present aspects of the speech-translation process to the end users. This is accomplished via the Nespole Monitor display. Three textual representations are displayed in clearly identified fields: (1) a transcript of their spoken input (the output from the speech recognizer); (2) a paraphrase of their input – the result of translating the recognized input back into their own language; and (3) the translated textual output of the utterance spoken by the other party. These textual representations provide the users with the capability to identify mis-translations and indicate errors to the other party. A bad paraphrase is often a good indicator of a significant error in the translation process. When a mis-translation is detected, the user can press a dedicated button that informs the other party to ignore the translation being displayed, by highlighting the textual translation in red on the monitor display of the other party. The user can then repeat the turn. The current system also allows the participants to correct speech recognition and translation errors via keyboard input, a feature which is very effective when bandwidth limitations degrade the system performance.

3 Evaluation Methodology

In December 2001, we conducted a large scale multi-lingual end-to-end translation evaluation of the Nespole! first-showcase system. For each of the three language pairs (English-Italian, German-Italian and French-Italian), four previously unseen test dialogues were used to evaluate the performance of the translation system. The dialogues included two scenarios: one covering winter ski vacations, the other about summer resorts. One or two of the dialogues for each language contained multi-modal expressions. The test data included a mixture of dialogues that were collected mono-lingually prior to system development (both client and agent spoke the same language), and data collected bilingually (during the July 2001 MM experiment), using the actual translation system. This mixture of data conditions was intended primarily for comprehensiveness and not for comparison of the different conditions.

We performed an extensive suite of evaluations on the above data. The evaluations were all end-to-end, from input to output, not assessing individual modules or components. We performed both mono-lingual evaluation (where generated output language was the same as the input language), as well as cross-lingual evaluation. For cross-lingual evaluations, translation from English German and French to Italian was evaluated on client utterances, and translation from Italian to each of the three languages was evaluated on agent utterances. We evaluated on both manually transcribed input as well as on actual speech-recognition of the original audio. We also graded the speech recognized output as a "paraphrase" of the transcriptions, to measure the levels of semantic loss of information due to recognition errors. Speech recognition word accuracies and the results of speech graded as a paraphrase appear in Table 1. Translations were graded by multiple human graders at the level of Semantic Dialogue Units (SDUs). For each data set, one grader first manually segmented each utterance into SDUs. All graders then used this segmentation in order to assign scores for each SDU present in the utterance. We followed the three-point grading scheme previously developed for the C-STAR consortium, as described in [3]. Each SDU is graded as either "Perfect" (meaning translated correctly and output is fluent), "OK" (meaning is translated reasonably correct but output may be disfluent), or "Bad" (meaning not properly translated). We calculate the percent of SDUs that are graded with each of the above categories. "Perfect" and "OK" percentages are also summed together into a category of "Acceptable" translations. Average percentages are calculated

Language	WARs	SR Graded (% Acc)
English	61.9%	66.0%
German	63.5%	68.0%
French	71.2%	65.0%
Italian	76.5%	N/A

Table 1: *Speech Recognition Word Accuracy Rates and Results of Human Grading (Percent Acceptable) of Recognition Output as a Paraphrase*

Language	Transcribed	Speech Rec.
English-to-English	58%	45%
German-to-German	46%	40%
French-to-French	54%	41%
Italian-to-Italian	61%	48%

Table 2: *Monolingual End-to-End Translation Results (Percent Acceptable) on Transcribed and Speech Recognized Input*

for each dialogue, each grader, and separately for client and agent utterances. We then calculated combined averages for all graders and for all dialogues for each language pair. Additionally, a majority vote scoring was also calculated. For each SDU, we assign a grade category based on majority (when a majority exists). In cases where there was no majority, or where there was at least one vote for both "Perfect" and "Bad", the SDU was excluded from the scoring.

3.1 Performance Results

Table 2 shows the results of the monolingual end-to-end translation for the four languages, and Table 3 shows the results of the cross-lingual evaluations. The results indicate acceptable translations in the range of 27–43% of SDUs (interlingua units) with speech recognized inputs. While this level of translation accuracy cannot be considered impressive, our user studies and system demonstrations indicate that it is already sufficient for achieving effective communication with real users. We expect performance levels to reach a range of 60–70% within the next year of the project.

4 Analysis of the Evaluation Scheme

We are currently in the process of several advanced investigations into issues related to our evaluation methodolgy, based on our experience from the NESPOLE! evaluation reported earlier. The current evaluations are described below, along with some preliminary results. Full results will be reported in the final version of the paper. Three main issues are under investigation:

Three-way Versus Binary Grading: The current evaluation scheme requires graders to assign one of three possible grades (Perfect/OK/Bad) in a single pass. The intention behind the three categories, however, is that we are applying two types of judgements: (1) is the meaning of the original SDU preserved in translation? (2) is the translation output fluent and grammatical? If the answer to the first judgement is negative, the grade assigned is "Bad" (regardless of the answer to

Language	Transcribed	Speech Rec.
English-to-Italian	55%	43%
German-to-Italian	32%	27%
French-to-Italian	44%	34%
Italian-to-English	47%	37%
Italian-to-German	47%	31%
Italian-to-French	40%	27%

Table 3: *Cross-lingual End-to-End Translation Results (Percent Acceptable) on Transcribed and Speech Recognized Input*

the second question). The second question distinguishes between the case of a "Perfect" versus and "OK" grade. An alternative way to assign the same judgements is therefore to sequentially pose two binary questions to the graders. The current experiment is intended to assess whether there are any significant differences between these two methods of scoring.

Averaging Scores Versus Majority Votes: Our custom in the past, when grading with multiple human subjects, has been to calculate percentages for each of the grade categories (Perect/OK/Bad) for each subject separately, and then to compute a simple arithmetic average of all the human subjects. In the current NESPOLE! evaluation, we experimented for the first time with also calculating majority votes. Under this scheme, for each SDU, we assign a grade category based on majority (when a majority exists). In cases where there was no majority, or where there was at least one vote for both "Perfect" and "Bad", the SDU was excluded from the scoring.

Intercoder and Intracoder Agreement: To establish the stability and coherence of our evaluation scheme, it is importnat to have a good measure of how well different human graders agree on scoring the same output, and also how consistent the graders are over time. The experiment we are currently conducting is collecting the necessary data to evaluate both of these questions. We will use standard mesures from the literature to quantify agreement: the Kappa coefficient, and confusion matrices.

4.1 Evaluation Experiment

We are currently conducting an experiment designed to evaluate the three issues identified above. The experiment involves multiple subjects grading sets of SDUs translated by our NESPOLE! system, using the two different grading schemes. Four groups of three subjects each are participating. Each group grades two different SDU sets (set1 and set2), two weeks apart. The two sets are different, but contain a common core subset of 200 SDUs. The use of two different sets is intended to minimize any memory effects of subjects evaluating the same data two weeks apart. The groups differ on the grading scheme used at each session (binary versus three-way), with one group assigned to each combination (i.e. group-1 uses the binary scheme in both sessions, group two uses binary in session-1 and three-way in session-2, etc.).

The data collected during this experiment will support the following evaluations:

- **Intercoder Agreement:** we compare the results of all eight subjects using the same evaluation scheme in a particular session, and test for intercoder agreement between them.

Group (Procedures)	W1 Acc	W1 Bad	W2 Acc	W2 Bad
1 (binary - 3way)	50.2	49.8	48.7	51.3
2 (3way - binary)	52.4	47.6	48.8	51.2
3 (3way - 3way)	53.8	46.2	54.9	45.1
4 (binary - binary)	49.0	51.0	50.0	50.0

Table 4: **Average Acceptable and Bad Grades Across All Groups for Overlapping Data**

- **Intracoder Agreement:** we compare the results of each of the 8 subjects on the common subset of 200 SDUs, which they graded twice (two weeks apart) using the same grading scheme, and test for coder agreement.
- **Inter-scheme Differences:** we compare the results of eight subjects using the binary scheme with the results of eight subjects using the three-way scheme (in each of the sessions), and test for statistical significance of any differences between them.
- **Average versus Majority Scores:** we will calcualte the final performance results using both the scheme that averages results from different subjects and the scheme that computes majority votes. We will comapre the resulting figures.

The grading experiment has just been completed, and the analysis of the results, as described above will be reported in the final version of the paper. Preliminary results appear in Table 4. The results reported are calculated on the common (overlapping) set of SDUs between the sets graded in week-1 and week-2. For each subject group, we report the average across subjects of `Acceptable` translations (sum of Perfect+OK) and `Bad` translations.

Acknowledgements

The research work reported here was supported by the National Science Foundation under Grant number 9982227 and the European Union under Grant number IST 1999-11562 as part of the joint EU/NSF MLIAM research initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the EU or the NSF.

References

[1] Alon Lavie, Fabio Pianesi, and al. Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-Commerce Applications. In *Proc. of the HLT2001*, San Diego, CA, 2001. ACM.

[2] Gianni Lazzari. Spoken translation: challenges and opportunities. In *Proc. ICSLP 2001*, Beijing, China, 10 2001.

[3] Lori Levin, Donna Gates, Fabio Pianesi, Donna Wallace, Takeshi Watanabe, and Monika Woszczyna. Evaluation of a Practical Interlingua for Task-Oriented Dialogues. In *Proceedings NAACL-2000 Workshop On Interlinguas and Interlingual Approaches*, Seattle, WA, 2000. AMTA.

[4] Florian Metze, John McDonough, and Hagen Soltau. Speech Recognition over NetMeeting Connections. In *Proc. EuroSpeech 2001*, Aalborg, Denmark, 2001. ISCA.

[5] Florian Metze, John McDonough, Hagen Soltau, Alex Waibel, Alon Lavie, Susan Burger, Chad Langley, Kornel Laskowski, Lori Levin, Tanja Schultz, Fabio Pianesi, Roldano Cattoni, Gianni Lazzari, Nadia Mana, Emanuele Pianta, Laurent Besacier, Herve Blanchon, Dominique Vaufreydaz, and Loredana Taddei. The NESPOLE! Speech-to-Speech Translation System. In *Proc. HLT 2002*, San Diego, CA, 3 2002.