SPEECH-TO-SPEECH TRANSLATION SYSTEM EVALUATION: RESULTS FOR FRENCH FOR THE NESPOLE! PROJECT FIRST SHOWCASE

Solange ROSSATO, Hervé BLANCHON, Laurent BESACIER

CLIPS-IMAG Lab.
BP 53, 38041 Grenoble cedex 9, France
Prénom.Nom@imag.fr

ABSTRACT

In this paper we give the results of a set of evaluations conducted in the context of a speech to speech translation project (NESPOLE!). The chosen situation involves a client (French, German, American) talking to an Italian travel agent (both using their own language) to organize a stay in Italy. Fives series of evaluation were conducted on the same data set. The first series concerned the Automatic Speech Recognition alone. Two other series were about monolingual (back-) translation from ASR outputs on the data set and form transcriptions of the data set. The last ones were about bilingual translation from both the ASR outputs and the transcriptions. The goal of the evaluation was to check the performances of the system at the end of the second year of the project. The fives sets of results concerning the French modules are given and commented.

INTRODUCTION

The NESPOLE![1] Project is a common EU NSF funded project exploring future applications of automatic speech to speech translation in e-commerce and e-service sectors [1]. The languages involved in this project are Italian, German, English and French. Partners of the project are ITC/IRST from Trento (Italy), ISL Labs. from UKA (Karlsruhe, Germany) and CMU (Pittsburgh, USA), Aethra (an italian company specialized in videoconferencing software), APT : a tourism agency in the Trentino area and finally CLIPS laboratory (Grenoble, France).

The scenario for the first showcase of NESPOLE! involves an Italian speaking agent, located in a tourism agency in Italy and a client located anywhere (English, German or French speaking) using a simple terminal (PC, sound and video cards, H323 videoconferencing software like NetMeeting™). This choice is related to present available technology, in the near future the third generation cellular can be also used as terminal.

The client wants to organize a trip in the Trentino area, and refers to APT web pages in order to get information. If the client wants to know more about a particular topic or prefers to have a more direct contact, a speech to speech translation service allows him to interact in his own language with an APT Italian agent. A videoconferencing session can then be opened between client and agent and the dialog starts between them.

This paper particularly address the speech to speech translation evaluation campaign that was conducted at the end of the second year of the project. *Section 1* describes the speech to speech translation modules that were used for the French language. The evaluation methodology is then described in *section 2* of this paper. *Section 3* describes the evaluation conducted during the Nespole! Project. Finally conclusions and perspectives are drawn at the end of this paper.

[1] see http://nespole.itc.it/

1. SPEECH TO SPEECH TRANSLATION MODULES

1.1 IF-based translation architecture

1.1.1 Approach

An Interchange Format (IF) based approach was adopted in the project. An important advantage of this approach concerns portability to a new language; given the described configuration, a lower effort is necessary to make an existing system capable of dealing with a new language. The major drawback is the hardness of defining the pivot itself and what has to be covered or not as far as the syntax and semantics are concerned, even within a task-based approach.

1.1.2 The Interchange Format

The IF [2] is defined by a Dialogue Act, and a list, possibly empty, of arguments. Dialogue Acts describe speaker's intention, goal, and need. They are made of a speech acts, a possible attitude, a possible main predication and possible predication participants. Arguments are the values of the discourse variables. Their presence is constrained by the DA participants. An IF is encoding a SDU (semantic Dialogue Unit), thus a turn may have to be described with several IFs.

1.1.3 Architecture

The global architecture for speech translation using the IF approach is thus the one described in *Figure 1*:

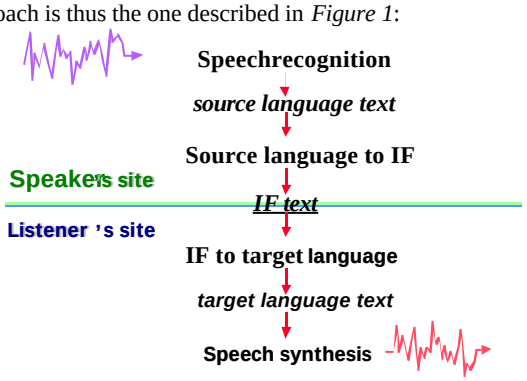

Figure 1: Overall components interaction

1.2 Automatic Speech Recognition

The first step of a speech-to-speech translation system is automatic speech recognition (ASR). Our continuous French speech recognition system RAPHAEL uses Janus-III toolkit [3] from CMU. The context dependent acoustic model was learned on a corpus that contains 12 hours of continuous speech of 72 speakers extracted from Bref80 [4] corpus.

The vocabulary contains nearly 20000 lexical forms: some lexical forms are specific to the reservation or the tourist information domains whereas the other words are the most frequent words that can be encountered in the French language.

The trigram language model that we used for our experimentation was computed on a large corpus extracted from Internet documents because it was shown that they give very large amount of training data for spoken language modeling. More details on the French ASR used in Nespole! can be found in [5].

1.3 French-to-IF analyzer

The current French-to-IF module is developed using a pattern-based translation approach. It is basically working in four steps. The turns are first split into SDU (one SDU is associated with one IF).

A topic is then associated with each SDU (activity, accommodation, attraction, …). According to the topic, the possible arguments are then instantiated if present. Finally, a dialogue act is built according to the instantiated arguments and some other markers (attitudes, questions, negations, …).

Patterns are used in several steps of the process, but mainly to describe the possible realizations of the arguments.

1.4 IF-to-French generator

For the generation two tracks are pursued. The first one is developed under Ariane-G5 and a rule-based approach [5]. The second one implements a template-based approach.

With Ariane-G5, we are using the specification files of the IF to produce automatically parts of the dictionaries and the grammars. The IF is parsed into a French linguistic tree passed to a general-purpose French generation module.

With the template-based approach, DA families are associated with a template sentence (a fill in the blanks sentence). The blanks are filled, if possible, with the French phrase associated with the right argument value.

1.5 Speech synthesis

The last part of automatic translation process is the text-to-speech synthesis (TTS). The Euler TTS[2] system of the Polytechnical Faculty of Mons is used. The quality of the synthetic speech is acceptable for our application.

2. EVALUATION METHODOLOGY

2.1 Single Component Evaluations

Single component evaluations were focused on the different speech-to-speech translation (STST) modules developed independently at each of the participating partner sites. The main STST components that are to be evaluated and the relevant procedures are described follow.

2.1.1 Speech Recognition evaluation

Accuracy-based evaluations are performed on the individual speech recognition modules using standard word-error-rate (WER) criteria. However, since standard WER evaluations do not take into account that certain mis-recognitions are far more harmful to translation than others, we have also considered performing human evaluations that evaluate the output of the recognizer as if it were a paraphrase translation of the original (transcribed) input. This was done using the same grading methodology used for evaluating actual translation quality (further described in *section 2.3*).

2.1.2 Back-translation evaluation

The analysis module from language X to IF and the generation module form IF to language X modules were considered a one unique back box. Evaluation of the whole back-translation was performed as follows: performing analysis from original (transcribed) input in language X into the IF followed by generation back into the language X. Then, the output of the translation is evaluated by human graders as a paraphrase translation of the original input.

2.2 Whole speech to speech translation Evaluations

2.2.1 Monolingual evaluation

In this monolingual evaluation, the ASR system is combined with the back-translation system. The input signal utterance in language X is first recognized by the ASR system and the hypothesis string is then analyzed into IF and generated back to language X. The output text string is then compared to the original (transcribed) input.

Such evaluation is meaningful since it truly combines the STST modules for a particular language into an end-to-end path. Because these are monolingual evaluations, native speakers of the language involved can easily perform them independently at each of the partner's sites.

2.2.2 Bilingual evaluation

End-to-end evaluations across sites were also conducted. These were performed in a batch-mode mode: the evaluation set was analyzed at one site, producing a corpus annotated with IF. This corpus was then sent to the second site, which applied the generation chain to the IFs in the corpus, producing a corpus of translations. Human graders then assessed end-to-end translation performance.

2.3 Grading

Three graders performed each evaluation. The grading was made at the SDU (Semantic Dialog Unit) level. For this, all the evaluation data was segmented manually into SDUs. For each SDU and its translated version, the grader had to evaluate the quality of the translation with one grade. Three choices were proposed for the grade:

p for *perfect* if the grader decided that the quality of translation was good

k for *okay* if the grader decided that the quality of translation was acceptable

[2] http://tcts.fpms.ac.be/synthesis/euler

b for *bad* if the grader decided that the quality of translation was not acceptable

3. EVALUATION RESULTS

3.1 Evaluation data

Four dialogs were extracted from the NESPOLE database [6]. Two of them were related to a client / agent discussion for organizing winter holidays in Val di Fiemme in Italy; the two others were related to summer vacations in the same region. Speech signals were then re-recorded from client turn transcriptions of these 4 dialogs (8kHz sampling rate). This data represents 235 signals related to 235 speaker turns of two different speakers (1 male, 1 female). Finally, these 235 speaker turns were segmented manually into 427 SDUs for translation evaluation.

After applying the recognition and/or translation modules on this data, grading was then performed at the SDU level by the graders. In all the tables of sections 3.2 to 3.6, the percentage values of the grade scores are given for each grader and for each dialog. In each cell, the first number represents the percentage of SDUs evaluated as correct (*p+k* cumulated); the second number (into braces) represents the percentage of SDUs evaluated as perfect (*p* only).

A majority vote was also applied for each SDU evaluated by 3 graders. In that case, grade scores were kept only if at least two graders among three gave the same mark for the SDU considered. Thus, SDUs that lead to confusion between graders were removed in this case. Results of this "majority vote" are given on the last line of each table; the total number of SDUs kept (i.e. which lead to unanimity graders' responses) is also given into braces.

3.2 ASR evaluation

We first evaluated our ASR system on the 235 client turn signals with conventional Word Error Rate (WER) criterion. The WER obtained is 28.8% (which represents 71.2% of Word Correct Rate).

Human evaluations that evaluate the output of the recognizer as if it were a paraphrase translation of the original (transcribed) input were also performed and are given in *Table 1*.

About 65% of the SDUs were judged correct by the human graders. This evaluation of the ASR system is more informative than evaluation with the WER% since 65% of correct SDUs means that after the ASR phase, we already know that 35% of the SDUs will not be correctly translated. This evaluation step is also a good way to evaluate the graders and check if they give approximately the same marks (which is the case here).

3.3 Back-translation evaluation

Evaluation of the translation (Analysis + Generation) was performed by the human graders. The evaluation results are given in *Table 2*.

About 55% of the SDUs were judged correctly translated. The «majority » vote line shows good consistency between graders. After this evaluation phase, we know for sure that 45% of the SDUs will not be correctly translated anyway ; it is thus important to know if this percentage includes the SDUs badly recognized by the ASR system or not. That is shown in the next section.

3.4 Monolingual speech to speech evaluation

In this monolingual evaluation, the French ASR system is combined with the French Translation (Analysis + Generation) modules. The evaluation results are given in *Table 3*.

We see here, that the whole speech to speech translation chain allows to correctly translate about 40% of the SDUs. This result alone would have been very difficult to interpret, but the evaluations conducted in *sections 3.1* and *3.2* show the respective contribution of ASR and Translation to this performance. This is an important information that will be used to further improve the current system.

3.5 Bilingual Translation Evaluation

Translation evaluation across sites (French to Italian) is presented in *Table 4.* It seems that for bilingual evaluation, the graders behavior is not as consistent as for mono-lingual evaluation. This can be due to different levels of expertise of the graders in the languages considered. Moreover, the average number of SDUs correctly translated is lower than for the monolingual experiment of *section 3.2*. This could lead to the conclusion that the Italian generator is slightly less optimized to the evaluation data than the French one.

3.5 Bilingual speech to speech evaluation

End-to-end path speech to speech translation evaluation across sites is presented in *Table 5.* Here again, the average number of SDUs correctly translated is slightly lower than for the monolingual experiment of *section 3.3*. We can see that about 1/3 of the SDUs are correctly translated from French to Italian.

CONCLUSION

This paper was dedicated to the evaluation problem in speech-to-speech translation domain. We have proposed a step-by-step evaluation process that allows to clearly identify the errors due to the different modules of a complex system: ASR, Translation, Cross-language interface modules. To our knowledge, very few papers have already addressed this problem.

Our evaluation was conducted during the Nespole! Project on a large and significant quantity of data. Results show that 30 to 40 percent of single dialog units can be correctly translated with our system.

ACKNOWLEDGEMENTS

The authors thank the graders (A.C. Descalle, F. Tajariol, D. Vaufreydaz, R. Lamy, S. Mazenot) and the IRST translation team of the Nespole! project (who generated italian from our IFs), for their contribution to this paper.

REFERENCES

[1] Lazzari G., Spoken translation: challenges and opportunities, ICSLP'2000, Beijing, China. Oct. 16-20, 2000, vol 4/4 : pp. 430-435

[2] Levin L. & al. An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues. Proc. ICSLP'98, 30th November - 4th December 1998, Sydney, Australia, vol.4/7, pp.1155- 1158.

[3] Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C., Sloboda, T., Tomita,

M., Tsutsumi, J., Aoki-Waibel, N., Waibel, A., and Ward, W. "Recent Advances in JANUS: A Speech Translation System". Eurospeech, 1993, volume 2, pages 1295-1298.

[4] Lamel, L.F., Gauvain, J.L., Eskénazi, M. "BREF, a Large Vocabulary Spoken Corpus for French", Eurospeech, Gênes, Italy, Vol 2, pp. 505-508, 24-26 September 1991.

[5] L. Besacier, H. Blanchon, Y. Fouquet, J.P. Guilbaud, S. Helme, S. Mazenot, D. Moraru, D. Vaufreydaz "Speech Translation for French in the NESPOLE! European Project", Eurospeech 2001, Aalborg, Danemark, September 2001

[6] Burger, S., Besacier, L. Metze, F., Morel, C., Coletti, P., The NESPOLE! VoIP dialog database, Eurospeech 2001, Aalborg, Danemark, September 2001

TABLES

Table 1 ASR evaluation by graders

%acceptable *(%perfect)*	Dial. A1 60 utt., 109 SDUs	Dial. A2 74 utt., 139 SDUs	Dial. C3 64 utt., 101 SDUs	Dial. C4 37 utt., 78 SDUs	All dialogues 235 utt., 427 SDUs
Grader 1 (lb)	74.5 (70.0)	49.3 (45.0)	68.0 (61.2)	81.0 (78.5)	66.0 (61.3)
Grader 2 (sr)	70.9 (68.2)	46.8 (40.4)	64.1 (59.2)	80.0 (78.8)	63.1 (59.0)
Grader 3 (rl)	73.6 (70.0)	47.1 (42.1)	69.9 (60.2)	82.5 (77.5)	65.8 (60.0)
Average	73 (69.4)	47.7 (42.5)	67.3 (60.2)	81.2 (78.3)	**65.0 (60.1)**
Vote_maj	72.5 (69.7)	46.0 (41.0)	68.3 (61.4)	76.9 (75.6)	**63.7 (59.5)**

Table 2 Back-translation evaluation by graders

%acceptable *(%perfect)*	Dial. A1 60 utt., 109 SDUs	Dial. A2 74 utt., 139 SDUs	Dial. C3 64 utt., 101 SDUs	Dial. C4 37 utt., 78 SDUs	All dialogues 235 utt., 427 SDUs
Grader 1 (lb)	67.7 (48.6)	44.7 (34.0)	49.5 (36.9)	64.6 (49.4)	55.3 (41.2)
Grader 2 (sr)	63.6 (50.9)	44.7 (38.3)	45.6 (40.8)	56.4 (52.6)	51.8 (44.7)
Grader 3 (rl)	67.6 (55.9)	47.9 (36.6)	49.5 (41.7)	62.0 (51.9)	55.9 (45.5)
Average	66.3 (51.8)	45.8 (36.3)	48.2 (39.8)	61.0 (51.3)	**54.3 (43.8)**
Vote_maj (395)	61.5 (52.3)	40.3 (36.0)	48.5 (40.6)	53.8 (47.4)	**50.7 (43.3)**

Table 3 Monolingual speech to speech translation evaluation by graders

%acceptable *(%perfect)*	Dial. A1 60 utt., 109 SDUs	Dial. A2 74 utt., 139 SDUs	Dial. C3 64 utt., 101 SDUs	Dial. C4 37 utt., 78 SDUs	All dialogues 235 utt., 427 SDUs
Grader 1 (lb)	49.1 (31.3)	26.4 (15.0)	37.3 (24.5)	52.5 (32.5)	39.6 (24.7)
Grader 2 (sr)	54.6 (33.6)	28.2 (18.3)	40.8 (29.1)	57.7 (42.3)	43.2 (29.1)
Grader 3 (rl)	51.8 (34.5)	27.0 (19.9)	35.9 (30.1)	51.3 (46.3)	39.9 (30.9)
Average	51.8 (33.1)	27.2 (17.7)	38.0 (27.9)	53.8 (40.4)	**40.9 (28.2)**
Vote_maj (382)	49.5 (31.2)	24.5 (14.4)	36.6 (27.7)	46.2 (37.2)	**37.7 (26.0)**

Table 4 Bilingual translation (French to Italian) evaluation by graders

%acceptable *(%perfect)*	Dial. A1 60 utt., 109 SDUs	Dial. A2 74 utt., 139 SDUs	Dial. C3 64 utt., 101 SDUs	Dial. C4 37 utt., 78 SDUs	All dialogues 235 utt., 427 SDUs
Grader 1 (an)	56.0 (40.4)	34.8 (24.1)	40.8 (31.1)	50.0 (36.3)	44.3 (32.1)
Grader 2 (fe)	51.4 (38.5)	25.0 (21.3)	34.3 (30.4)	50.7 (42.5)	38.6 (31.7)
Grader 3 (sy)	59.8 (47.7)	37.9 (30.7)	43.1 (37.3)	56.3 (46.3)	48.0 (39.4)
Average	55.7 (42.2)	32.6 (25.4)	39.4 (32.9)	52.3 (41.7)	**43.6 (34.4)**
Vote maj (370)	48.6 (37.6)	27.3 (23.7)	39.6 (34.7)	41.0 (34.6)	**38.2 (31.9)**

Table 5 Bilingual speech to speech translation (French to Italian) evaluation by graders

%acceptable *(%perfect)*	Dial. A1 60 utt., 109 SDUs	Dial. A2 74 utt., 139 SDUs	Dial. C3 64 utt., 101 SDUs	Dial. C4 37 utt., 78 SDUs	All dialogues 235 utt., 427 SDUs
Grader 1 (an)	39.1 (27.3)	23.4 (15.6)	32.0 (24.3)	46.3 (36.3)	33.6 (24.4)
Grader 2 (fe)	43.5 (35.2)	20.1 (11.5)	30.7 (23.8)	46.8 (36.7)	33.5 (25.1)
Grader 3 (sy)	39.1 (29.1)	22.9 (19.3)	33.0 (27.2)	51.3 (41.3)	34.6 (27.7)
Average	40.6 (30.5)	22.1 (15.5)	31.9 (25.1)	48.1 (38.1)	**33.9 (25.7)**
Vote_maj	37.6 (28.4)	18.7 (15.1)	30.7 (25.7)	42.3 (35.9)	**30.7 (24.8)**