HLT Modules Scalability within the NESPOLE! Project

Hervé Blanchon

GETA, CLIPS, IMAG-campus
385 rue de la Bibliothèque, BP 53
38041 Grenoble cedex 9, France
Herve.Blanchon@imag.fr

Abstract

The spoken dialogue translation project NESPOLE! proposed two showcases in order to focus on two important issues: *scalability* — namely, the capability of a system to progressively handle larger portions of a given domain — and *cross-domain portability*. Those concerns were rather new when the project was proposed. ShowCase-1 dealt with limited tourism, while ShowCase-2 consisted of ShowCase-2a on extended tourism (thus focusing on scalability) and ShowCase-2b on a medical domain (thus focusing on cross-domain portability). In this article, we address the issue of scalability for the French analysis and generation modules in the tourism domain. We discuss ShC-1 and ShC-2a evaluations results and evaluate our progress.

1. Introduction

The NESPOLE! system [3] provides English, German, French clients and Italian tourist agents with simultaneous dialog interpretation services for the tourism domain over Internet.

For the translation lingware architecture, we adopted the interlingua-based approach we successfully experimented within the C-STAR II project. The NESPOLE! interlingua, which we call IF (Interchange Format), is based on representing the speaker's intention rather than the literal meaning of the utterance.

The tackled scenarios were inspired by actual APT data (mails, phone calls) we analyzed and divided into 5 class: simple summer and winter vacations – used for ShowCase-1 –, and information enquiries about lakes, castle and packages – used for ShowCase-2a. Those two showcases were designed in order to measure lingware scalability.

We first discuss the overall results of the NESPOLE! project from the perspective of scalability, and study in more detail the scalability of the French analysis and generation modules. Section 2 and 3 give an overview of the HLT modules developed for each showcase. We explain how we tried to deal with scalability from the lingware design point of view. The showcase evaluation results are presented in section 4. We try to analyze how we succeeded as far as scalability is concerned.

2. First showcase modules for French

2.1. Introduction to the IF

The components of the IF are represented in the example below, which corresponds to the question *Does the flight leave at 2*. The first element is a speaker tag, c: for client or a: for agent. The reason for the speaker tag is that some Domain Actions (DA) correspond to different sentences depending on who the speaker is. For example, the DA of requesting information about payment methods corresponds to *How can I pay* if the client is speaking or *How would you like to pay* if the agent is speaking.

```
c:request-information+departure+transportation
(transportation-spec=(flight, identifiability=yes),
time=(clock=(hours=2)))
```

The second component of an IF is the Speech Act (SA). In this case, the speech act is request-information. Then come one or more concepts. Our example IF contains two concepts, +departure and +transportation. The domain DA is the combination of the speech act and concepts, in this case:

```
request-information+departure+transportation.
```

Following the domain action is a list of arguments. In this example, there are two main arguments, transportation-spec= and time=. These arguments contain sub-arguments (identifiability=, clock=, and hours=).

A SA is defined by two slots: (1) a continuation slot pointing to a list of concepts that can directly follow the SA; (2) an arguments slot pointing to a list of arguments that are licensed by the SA. Concept definitions follow the same schema as SA definitions.

A definition of an argument is made through three slots: (1) a values slot pointing to the list of values which are legal for the argument; (2) a relations slot, and (3) an attributes slots, the last 2 introducing legal sub-arguments. Arguments may be defined with one or several definitions.. Thus, restrictions on the co-occurrence of values and sub-arguments can be defined as well as domain specific constraints.

Values are often defined as value groups that can recursively point to the definitions of other groups.

2.2. ShC-1 analyzer

For building the ShC-1 analyzer, we mainly focused on arguments. Collected data [2] showed a poor variety of SAs. The analysis process [1] is divided in 3 major steps. The input text is first split into semantic dialogue units (SDU[1]). The topic of each SDU is then searched out. For each topic, most probable arguments are instantiated, and then the Dialogue Act is built using the instantiated arguments and some other features of the SDU.

The analyzer implement a pattern-based approach that is well suited to handle the kind of "ill-formed" input it has to handle. This approach is quite similar to the one used in island-based parsing. In fact, the construction of the Dialogue Act is fairly weak, it does not take Rhetorical Relations, Attitudes and Actions into account.

[1] An SDU is a portion of an utterance roughly corresponding to a sentence or a sentence fragment.

2.3. ShC-1 generator

As the set of DAs used in the first Showcase data was fairly small, we chose to use a "blank filling" approach. Each different speech act is associated with a set of sentences with blanks. The number and position of blanks depend on the instantiated arguments. Blanks are filled with generated arguments values.

3. Second Showcase modules

The first Showcase evaluation showed that, with the French analyzer used with either the French or Italian generators, applied on both speech recognition outputs and reference transcriptions, we reached a level of performance comparable to that of other such schemes applied on English and German.

On the other hand, the Italian analyzer and French generator pair, applied on both speech recognition outputs and reference transcriptions, reached a lower level of performance than the other bilingual combinations. The generator was not able to handle a large proportion of the IFs produced by the Italian analyzer.

3.1. New choices

The DAs observed in the second Showcase data [4] were more numerous and complex than those of the first Showcase data. Thus it was necessary to scale up both the analyzer and the generator.

For the analyzer, we moved towards a better coverage of the potential input within the same pattern-based approach. We put the emphasis on a better turn splitting and on the construction of the speech act. We also concentrated on argument embedding.

As far as the generator is concerned, we departed from our first approach in order to ground the generator on the IF specification and not on the observed IF-transcribed data available.

3.2. ShC-2a analyzer

The new analysis module uses the same pattern-based approach as the analyzer developed for ShC-1. The overall architecture has not really changed, but all the SAs and almost all the concepts are handled. It is also the same for the arguments.

Each speech turn is first split into SDUs using a more fine-grained approach allowing better segmentation. We use simple sentences, coordination and subordination patterns. We added patterns to handle sentences juxtaposition (several sentences following one another without any marker).

A domain is then associated to each SDU. The defined domains cover all the terminal Speech Acts (i.e. the SAs with no continuation), and all the focus concepts (i.e. concepts without continuation).

The arguments of the terminal SA are instantiated (e.g.: manner for the SA thank) in a way specific to each SA.

Focus concepts are all handled the same way. The DA is first built by finding out the SA, the Rhetorical Relations, the Attitudes and the Actions. Actions are described by verbal constructions that may realize either a concept (i.e. +clarify, +click, +confirmation, …, +view, +explain, +write), or an action in the IF *actions* value set. The IF values for nearly all actions are represented by WordNet synset entries (e.g., set of synonyms) for ease of disambiguation and for specifying exact definitions. For example, the IF value e-drink-1 represents the synset and meaning that corresponds to the verb drink and specifically the first WordNet lexical entry for drink.

During this process, each concept instantiates its arguments. This step delivers a prefix of the final DA and a list of arguments instantiated for the members of the DA.

The Arguments of the focus concept are then built by trying to instantiate the potential arguments of the concept. During this process, the DA may be completed (e.g., if the SDU is about the price of the focus concept, +price is added to the DA). Finally, the IF is produced by concatenating the current DA with the focus concept and adding the arguments IF representations.

3.3. ShC-2a generator

The new generation module does not use the fill-in the blanks approach used for the first Showcase, because it lacks too much of generality.

For the non-terminal SAs, the generation is performed by traversing the DA, whatever it may consist of. Each Dialog Act is thus covered, giving a far better coverage. This process is carried out in 5 steps.

The rhetoricals are first generated, which may give a prefix formula for the sentence to be produced (e.g.: "can you recommend" for the request-suggestion DA). Next, if present, the attitudes are generated.

Then, we remove the SA and the +attitude and traverse the remaining concepts present in the DA. For each concept, the "essential" arguments are generated (e.g.: accommodation-spec= for +accommodation, price-spec= and price= for +price). During the traversal, some look ahead may be necessary when the current concept applies to the following one (e.g., the +price concept applies to the following argument if there is one).

When the traversal is finished, the remaining arguments are finally generated in the order they appear in the instantiated arguments list.

The generation of the arguments has been revised and now gives better results.

4. Scalability assessment

4.1. 2001's Showcase on "restricted tourism"

Results of the first evaluation have been presented in [5]. We present here a summary.

4.1.1. Data and protocol

Four dialogues (2 for summer vacations and 2 for winter vacations) were randomly picked up from the first NESPOLE! data collection [2] for each language. For Italian, tourist agent turns were used. For English, French and German, client turns were used.

Evaluation was done at two levels: (1) *Hypos*, which are the automatic transcriptions produced by the Automatic Speech Recognition modules, and (2) *Refs*, which are the manual transcriptions of the same speech signals. Each turn was also manually split into Semantic Dialogue Units[1] (SDU) in order to get a SDU-based (and not a turn-based) evaluation of the translation quality.

Speech recognition was evaluated using the Word Accuracy Rate (WAR) score. However, WAR does not allow

[1] An SDU is a portion of an utterance roughly corresponding to a sentence or a sentence fragment.

to measure precisely how speech recognition errors influence translation quality. We also graded the *Hypos* as paraphrases of the *Refs*, at the SDU level, to measure the of loss of semantic information due to recognition errors.

We performed monolingual evaluation (where the generated output language was the same as the input language), as well as crosslingual evaluations. For crosslingual 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.

For each set, we used three human graders with bilingual ability. Each SDU was graded as either "Perfect" (meaning is translated correctly and output is fluent), "OK" (meaning is translated almost correctly but output may be disfluent), or "Bad" (meaning is not properly translated). We calculated the percentage of SDUs in each of these three categories. "Perfect" and "OK" were also merged into a larger category of "Acceptable".

Average percentages were calculated for each dialogue, each grader, and separately for client and agent utterances. Combined averages for all graders and for all dialogues were then computed for each language pair.

4.1.2. Results

The following table combines all the results (in %) for acceptable translations using average or majority vote when computed.

Table 1: results of the NESPOLE! first Showcase evaluation

ASR *WAR*	71	62	64	77
Hypos as paraphrases	64	66	68	70
Mono-lingual trans.	*F-F c*	*E-E c*	*G-G c*	*I-I a*
on Refs/Hypos	51/38	48/45	46/40	60/44
Cross-lingual trans.	*F-I c*	*E-I c*	*G-I c*	
on Refs/Hypos	38/31	55/43	32/27	
	I-F a	*I-E a*	*I-G a*	
on Refs/Hypos	38/26	46/35	45/20	

4.1.3. Comments

The performance of the different speech recognizers, in producing *Hypos as paraphrases*, is almost the same what ever the WAR of the speech recognized may be (from 62% up to 77%).

The results indicate acceptable monolingual translations (on clients and agent turns) in a range of 40-48%. of SDUs on *Hypos*. On *Refs*, the scores are, not surprisingly, better (46-61%). For crosslingual translation towards Italian (on clients turns only), there is a performance drop (higher on *Refs* than on *Hypos*) compared with the monolingual systems. It shows that either the Italian generator does not handle properly some IFs produced by the French, English and German analyzers (problem of coverage) or that there in an intercoder agreement problem across sites. The same problem occurs for crosslingual translation from Italian (on agent turns only). The performance drop is higher than on client turns. The same reasons explain the phenomenon. However, the problem of coverage is probably dominant in this case. For the French generator, we could indeed check that the latter is true.

As references simulate a 100% speech recognition success rate, the translation scores on *Refs* for the four monolingual end-to-end systems must be considered as upper bounds for the scores on *Hypos*. However, we found that the behavior of the systems is not a linear function of the *Hypos as paraphrase* rate. If it had been the case, for French, the percentage of acceptable translations on *Hypos* would have been 35% for 65% of *Hypos as paraphrases*. The actual score (41%) is 6 points higher than "expectation". Figures are almost the same for all four monolingual systems.

For the three crosslingual systems towards Italian (used on client turns), the situation is the same. We observed a 5 points increase. The analyzers in the monolingual and crosslingual systems towards Italian are the same for each source language. Thus, we may say that those scores are better than expected thanks to the analyzers "robustness".

When checking the scores for the three crosslingual systems from Italian (1 Italian analyzer and 3 generators), we get unclear results: the French and German generators do not reach expectation on hypothesis by 1 or 2 points, but the English generator scores over expectation by 5 points. We can only conclude that the generators of French and German are not as robust as those of English and Italian. Maybe this is due to the fact that the IF is, in a way, based on English and mostly defined by the CMU and IRST team.

4.2. 2002's Showcase "Extended Tourism"

The second Showcase evaluation methodology has been designed to overcome some problems of our first evaluation.

4.2.1. Data and protocol

For each language, two unseen dialogues were picked up from the second NESPOLE! data collection [4]. The dialogues focused on additional scenarios such as tours of castles and lakes. The evaluation data sets were of the same kind as those used in the first Showcase evaluation.

This evaluation also includes a comparison of the ShowCase-1 components and the ShowCase-2a components. ShowCase-1 components were frozen and saved after the ShowCase-1 evaluation. ShowCase-1 components were then run on the ShowCase-2a evaluation data in order to compare the two systems on the same data. The ShowCase-1 system was only run on transcribed input.

In this evaluation, we departed from our previous grading methodology in several ways. First, the 3-point scale (perfect, OK, bad) was replaced with a 4-point scale, based only on meaning preservation, taking neither fluency nor grammatical accuracy into account. Second, whereas we previously reported average scores across graders for each SDU, we calculated majority scores as well as averages. The majority votes are generally close to the averages, except where there is an outlier (a grader who was exceptionally harsh or lenient), but this problem did not occur. Third, the graders for this evaluation were last year students in a school for translators. Previously, graders had no special training in translation, and the groups were less homogeneous in terms of education and of second language knowledge than this year.

4.2.2. Results

The following table combines all percentage results for acceptable translations using majority vote.

Table 2: results of the NESPOLE! second Showcase evaluation

ASR *WAR*	58	56	51	76
Hypos as paraphrase	60	67	62	76
Mono-lingual trans.*	*F-F c*	*E-E c*	*G-G c*	*I-I a*
on Refs (01)| Refs/Hypos (02)	*69*| 77/58	*68*| 68/50	*45*| 61/51	*36*†| 51/42‡
Cross-lingual trans.	*F-I c*	*E- I c*	*G-I c*	
on Refs (01)| Refs/Hypos (02)	*72*| 77/58	*64*| 70/50	*44* x/x	
	I-F a	*I-E a*	*I-G a*	
on Refs (01)| Refs/Hypos (02)	*19*| 37/33	*33*| 33/30	*38*| 45/38	

4.2.3. Comments

The results indicate acceptable monolingual translations (on clients and agent turns) in a range of 42-48% of SDUs on *Hypos*. On *Refs*, the scores are, not surprisingly, better (51-77%). For crosslingual translation towards Italian (on clients turns only), there is no performance drop compared with the monolingual systems. In this second Showcase, the Italian generator handles correctly all IFs produced by the French, English and German analyzers. There no problem of IF coverage or intercoder disagreement across sites. Those problems still occur for crosslingual translation from Italian (on agent turns only). The problem of coverage is dominant in this case.

4.3. Discussion on scalability

One noticeable observation is that translation performance from Italian is significantly lower than translation into Italian. This is mainly due to the characteristics of the evaluation data: agent utterances (translated from Italian) are more complex, and in some cases are actually out of domain, while client sentences (translated into Italian) are on average shorter, easier and in domain.

In order to quantify this difference and assess its effect on our results, we asked system developers to manually classify the SDUs in the test data into three categories: (1) falls within the domain of coverage of ShowCase-1; (2) falls within the domain of coverage of ShowCase-2a; and (3) out of domain. We then calculated the performance results for the three groups of SDUs separately.

Interestingly, for English, German and French input (client data), we discovered that only a very small number of SDUs were classified in either group-2 or group-3 (less than 5 SDUs for each). Thus, the data is overwhelmingly within the domain of the ShowCase-1 system. For the Italian input (agent data), however, 13% of SDUs were classified within the domain of ShowCase-2a (group-2), and 25% of SDUs were out of domain (group-3).

The difference in system performance on these three separate categories is also quite insightful.

On the group-1 data, we see an improvement in performance between the results of the ShowCase-1 system and the ShowCase-2a system: from 56.6% to 63.2% acceptable translation on transcribed input. This demonstrates improvements in domain coverage in the ShowCase-2a system (within the domain of the first Showcase).

On the group-2 data, the difference is much more pronounced. The ShowCase-1 system achieves only 14.2% acceptable translations, while the ShowCase-2a system achieves 38.4%. The ShowCase-1 system was not designed to cover this type of data, so this is not surprising. While the ShowCase-2 system performs much better on this data, it did not reach the same level of performance as for the ShowCase-1 domain.

The Italian system can not handle out of domain SDUs. When excluding these SDUs from consideration, the performance figures are 49% (instead of 36%) for the ShowCase-1 system and 59% (instead of 42%) for the ShowCase-2a system – more similar to the results we find for the French, English and German monolingual systems (***** in Table 2). For the system from Italian to French, English and German, we observe a 9% increase on *Hypos(02)*. Namely we reached 42% for French, 39% for English, and 47% for German.

5. Conclusions

In this paper we presented the design of the French to IF analyzer and the IF to French generator in the framework of the NESPOLE! project. We exposed design improvement geared towards scalability improvement between the 2 tourism showcases. We have shown that for ShowCase-2a, our analyzer reached expectation: with 60% of *Hypos as paraphrase* we get 58% acceptable monolingual and crosslingual translation (only 2% loss). For the generator, our ShowCase-2a generator reached 33% acceptable translation on the whole set of Italian IFs. This number may be compared with the 19% score reached by our ShowCase-1 generator on the same clean *Refs(01)* data. When discarding out-of-domain SDUs, our generator reached 42% acceptable translation on *Hypos(02)*.

6. Acknowledgements

The author would like to thank all the NESPOLE! partners for 3 fruitful years of cooperation.

7. References

[1] Blanchon, H. (2002). *A Pattern-Based Analyzer for French in the Context of Spoken Language Translation: First Prototype and Evaluation. Proc.* COLING. Taipei, Taiwan. 24 August - 1 September, 2002. vol. **1/2:** pp. 92-98.

[2] Burger, S., Besacier, L., Metze, F., Morel, C., et al. (2001). *The NESPOLE! VoIP Dialog Database. Proc.* Eurospeech. Aalborg, Denmark. September 3-7, 2001. 4 p.

[3] Lazzari, G. (2000). *Spoken Translation: Challenges and Opportunities. Proc.* ICSLP 2000. Beijing, China. Oct. 16-20, 2000. vol. **4/4:** pp. 430-435.

[4] Mana, N., Burger, S., Cattoni, R., Besacier, L., et al. (2003). *The Nespole! VoIP Corpora in Tourism and Medical Domains. Proc.* EUROSPEECH 2003. Geneva, Switzerland. September 1-4, 2003. 4 p.

[5] Rossato, S., Blanchon, H. & Besacier, L. (2002). *Speech-to-Speech Translation System Evaluation: Results for French for the NESPOLE! Project First Showcase. Proc.* ICSLP. Denver, USA. 16-20 September, 2002. 4p.