A Speech-to-Speech Translation System for Catalan, Spanish and English

Victoria Arranz[1], Elisabet Comelles[1], David Farwell[2 1], Climent Nadeu[1], Jaume Padrell[1], Albert Febrer[3], Dorcas Alexander[4] and Kay Peterson[4]

[1] TALP Research Centre, Universitat Politècnica de Catalunya, Barcelona, Spain
{varranz,comelles,farwell,climent,jaume}@talp.upc.es
[2] Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain
[3] Applied Technologies on Language and Speech, Barcelona, Spain
{afebrer}@verbio.com
[4] Language Technologies Institute/Interactive Systems Lab, Carnegie Mellon University, Pittsburgh, USA
{dorcas,kay}@cs.cmu.edu

Abstract. This paper describes the FAME interlingual speech-to-speech translation system for Spanish, Catalan and English, which is intended to assist users in the reservation of a hotel room when calling or visiting abroad. The system has been developed as an extension of the existing NESPOLE! translation system[4] which translates between English, German, Italian and French.

1 Introduction

This paper describes an interlingual speech-to-speech translation system for Spanish, Catalan and English which is under development as part of the European Union-funded FAME project[5]. The system is an extension of the existing NESPOLE! translation system [4] to Spanish and Catalan in the domain of hotel reservations. At its core is a robust, scalable, interlingual speech-to-speech translation system having cross-domain portability which allows for effective translingual communication in a multimodal setting. Although the system architecture was initially based on the NESPOLE! platform, now the general architecture integrating all modules is based on an Open Agent Architecture (OAA)[6] [2]. This type of multi-agent framework offers a number of technical features that are highly advantageous for the system developer and user.

Our system consists of an analyzer that maps spoken language transcriptions into interlingua representation and a generator that maps from interlingua into natural language text. The central advantage of this interlingua-based architecture is that in adding further languages to the system, it is only necessary to develop analysis and generation components for the new languages.

[5] FAME stands for *Facilitating Agent for Multicultural Exchange* and focuses on the development of multimodal technologies to support multilingual interactions. http://isl.ira.uka.de/fame/

[6] http://www.ai.sri.com/~oaa

The Interchange Format (IF) [3], the interlingua currently used in the C-STAR Consortium, is being adapted for this effort. Its central advantage for representing dialogue interactions such as those typical of speech-to-speech translation systems is that it focuses on identifying the speech acts and the various types of requests and responses typical of a given domain. Thus, rather than capturing the detailed semantic and stylistic distinctions, it characterizes the intended conversational goal of the interlocutor. Even so, in mapping to IF it is necessary to take into account a wide range of structural and lexical properties related to Spanish and Catalan.

2 Automatic Speech Recognition and Language Parsing

For both Spanish and Catalan speech recognition, we used the JANUS Recognition toolkit (JRTk) developed by UKA and CMU [6]. Both Spanish and Catalan language models (LM) are trigram models. They were trained on text of the same topic used to build the text-to-text translation system from the C-STAR and LC-STAR[7] corpora. Both Spanish and Catalan acoustic models were trained on a 30-hour DB of speech, respectively.

The parsing side utilizes the top-down, chart-based SOUP parser[1]. It was developed specifically to handle spontaneous speech. Typically one would not expect a single parse tree to cover an entire utterance of spontaneous speech, because such utterances frequently contain multiple segments of meaning. These segments are called Semantic Dialogue Units (SDUs) and correspond to a single domain action (DA). Thus, one of the SOUP features most relevant to the FAME system is the capability to produce a sequence of parse trees for each utterance, effectively segmenting the input into SDUs (corresponding to DAs) at parse time. An analysis mapper completes the analysis chain by performing formatting functions on the parser output to produce standard IF.

The Spanish and Catalan parsing grammars are context-free based and have been developed using a corpus of dialogues obtained from the C-Star-II database[8]. Among the main differences between English and both Spanish and Catalan, there are the following: a) The high inflection in the latter; b) NP word-order (adjectives usually preceed the noun in English whereas they usually follow it in Spanish and Catalan); c) Constituent order within the sentence: English word-order is much more fixed while Spanish and Catalan are free-word order.

3 Language Generation and Text-to-Speech Synthesis

The generation module of the translation part of our interlingua-based MT system includes the NESPOLE! generation mapper and the GenKit generator. The mapper converts a given interchange format representation into a feature structure. This feature structure then serves as input to the GenKit generator [5], a pseudo-unification-based generation system.

[7] http://www.c-star.org & http://www.lc-star.com
[8] http://www.is.cs.cmu.edu/nespole/db/current/cstar-examples.db

The generator uses hybrid syntactic/semantic grammars for generating a sentence from an IF feature structure. Generation knowledge employed with GenKit consists of grammatical, lexical, and morphological knowledge. Words are associated with semantic IF concepts and values through the lexical entries. These lexical entries contain not only the root forms of these words, but are also enriched with, for example, lexical information pertinent to morphology generation (such as gender information in the case of nouns) and, in the case of verbs, subcategorization requirements.

Generation of the correct morphological form is performed via inflectional grammar rules that draw on additional information stored in the lexical entries. In the more complex case of verb morphology, the correct form is then retrieved from an additional morphological form look-up table. The output is then post-processed before reaching the speech synthesizer.

For Spanish and Catalan, several specific linguistic phenomena had to be dealt with that had not been implemented in GenKit grammars before. These include language-specific phenomena such as *que+subjunctive* constructions, pronominal verbs, clitics, pronouns in ditransitive constructions, etc. Some of these phenomena are reflected in the lexica, which have been developed from scratch and are enriched with a considerable amount of information vital for generation.

For both Spanish and Catalan, we use the Text-to-Speech (TTS) system developed at the UPC. It is a unit-selection based concatenative speech synthesis system, an approach that gives high levels of intelligibility and naturalness.

4 Evaluation of the Translation Component

A preliminary evaluation of the translation component has been carried out on text input (from Spanish into English). The evaluation data set has been obtained from the unseen data in TALP-MT.db, a database containing nine new dialogues recorded at TALP for evaluation purposes.

Prior to the revision of the translation output, a set of evaluation criteria was defined. Evaluation was separately performed on the grounds of *form* and *content*. The evaluation of *content* took into account both the Spanish input and the English output. Accordingly, the meaning of the evaluation metrics varies if they are being used to judge either *form* or *content*, as shown below:

- **Perfect**: well-formed output (*form*) or full communication of speakers' information (*content*).
- **Ok+/Ok/Ok-**: acceptable output, grading from only some minor *form* error (e.g., missing determiner) or some minor non-communicated information (*Ok+*) to some more serious *form* or *content* problems (*Ok-*).
- **Unacceptable**: unacceptable output, either essentially unintelligible or simply totally unrelated to the input.

On the basis of the test dialogues used, the results presented in table 1 were obtained. Result figures are shown in percentages.

Table 1. Evaluation Results for the Translation Component

Scores	Form	Content
Perfect	65,95	70,21
Ok+	12,23	4,25
Ok	7,44	1,59
Ok-	3,19	3,72
Unacceptable	11,17	20,21

5 Conclusions and Future Work

With the initial stage of system development nearing successful completion, our efforts will turn to system evaluation, to confronting the more serious technical problems which have arisen thus far and to extending the systems both within the reservations domain and to further travel-related domains. In regard to difficulties, perhaps the most serious problem we confront is the need to enhance the capacity of the systems to deal with the often degraded transcriptions provided by the speech recognition components. We are looking into strategies for dealing with this issue which we hope to implement in the coming year.

Acknowledgments This research is partly supported by the FAME (IST-2001-28323) and ALIADO (TIC2002-04447-C02) projects.

References

1. Gavaldà, M.: SOUP: A Parser for Real-world Spontaneous Speech. In Proceedings of the 6th International Workshop on Parsing Technologies, Trento, Italy (2000)
2. Holzapfel, H., Rogina, I., Wölfel, M., Kluge, T.: FAME Deliverable D3.1: Testbed Software, Middleware and Communication Architecture (2003)
3. Levin, L., Gates, D., Wallace, D., Peterson, K., Lavie, A., Pianesi, F., Pianta, E., Cattoni, R., Mana, N.: Balancing Expressiveness and Simplicity in an Interlingua for Task based Dialogue. In Proceedings of ACL-2002 workshop on Speech-to-speech Translation: Algorithms and Systems, Philadelphia, PA, U.S.(2002)
4. Metze, F., McDonough, J., Soltau, J., Langley, C., Lavie, A., Levin, L., Schultz, T., Waibel, A., Cattoni, L., Lazzari, G., Mana, N., Pianesi, F., Pianta, E.: The NESPOLE! Speech-to-Speech Translation System. In Proceedings of HLT-2002, San Diego, California, U.S., (2002)
5. Tomita, M., Nyberg, E.H.: Generation Kit and Transformation Kit, Version 3.2, User's Manual. Technical Report CMU-CMT-88-MEMO. Pittsburgh, PA: Carnegie Mellon, Center for Machine Translation (1988)
6. 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., Ward, W.: Recent Advances in JANUS: A Speech Translation System. In Proceedings of Eurospeech-93 (1993) 1295–1298