Procesamiento del lenguaje Natural, nº 41 (2008), pp. 319-320 recibido 7-05-2008; aceptado 16-06-2008

TECNOPARLA - Speech technologies for Catalan and its application to Speech-to-speech Translation *

TECNOPARLA - Tecnologías del habla en catalán y su aplicación a la traducción oral automática

Henrik Schulz hschulz@gps.tsc.upc.edu
Marta R. Costa-Jussà mruiz@gps.tsc.upc.edu
José A. R. Fonollosa adrian@gps.tsc.upc.edu
TALP Research Center, Universitat Politècnica de Catalunya, Tel. +34 934016439

Resumen: Este artículo describe los objetivos y las principales tareas del proyecto TECNOPARLA dedicado al desarrollo de tecnologías avanzadas del habla en Catalán
Palabras clave: tecnología del habla, reconocimiento del habla, conversión texto a voz, traducción automática, traducción oral

Abstract: The paper reports on objectives and activities of the project TECNOPARLA that aims to develop state-of-the-art Speech Technologies in Catalan
Keywords: Speech-to-speech translation, automatic speech recognition, statistical machine translation

1 Overview

Speech-to-speech translation offers a range of applications for interpersonal communication for people not sharing a common language as well as in information exchange and access across languages. It becomes even more eminent and desireable as societies across the globe are moving together, regional languages and cultures more strengthened, and multi-lingual and multi-cultural societies more present. The TECNOPARLA project aims to improve Catalan language technology, and its application to speech-to-speech translation between Catalan, English and Spanish. Members of the TALP group and RWTH Aachen university contribute to the project. Broadcast radio and television can be considered as one of the most widely used sources of information, and therefore may suggest task and domain to be disposed within the project. Translating broadcast debates implies several technologies involved, and may most simplified be regarded as a sequential process of automatic speech recognition (ASR), machine translation (SMT), and speech synthesis (TTS). Besides those - acoustic events such as music or noise are detected beforehand, speaker diarization keeps track on involved speakers in debates facilitating the use of speaker adapted acoustic and language models in speech recognition and the selection of dedicated voices in speech synthesis. Language detection maintains the use of monolingual models in multilingual debates. System components are integrated using the Unstructured Information Management Architecture (UIMA).

2 Speech Recognition

The development of the speech recognition subsystem is carried out in collaboration with RWTH Aachen University (Germany) using the RWTH open source speech recognition framework[1].

An initial Catalan acoustic model was derived from a Spanish acoustic model developed during the project TC-STAR (Lööf et al., 2007). The feature space comprises Mel frequency cepstral coefficiants (MFCC) extended by a voicedness feature.

A training phase is carried out by several steps: prior to the acoustic model estimation, a linear discriminative analysis estimates a feature space projection matrix. Furthermore a new phonetic classification and regression tree (CART) is grown tying the HMM states to generalized tri-phone states, finally the model estimation, that iteratively splits and refines the gaussian mixture models.

* funded by Generalitat de Catalunya

[1] http://www-i6.informatik.rwth-aachen.de/rwth-asr/

ISSN: 1135-5948

The acoustic model provides context dependent semi-tied continuous density HMM using a 6-state topology for each tri-phone. Their emission probabilities are modeled with Gaussian mixtures sharing a common diagonal covariance matrix.

Both, the lexicon encompassing the 50k most frequent words, and the 4-gram back-off language model comprising about 10.1 M multi-grams have been derived from the 'El Periodico' corpus. The latter achieves minimal perplexity with a linear discounting and modified Kneser-Ney smoothing methodology.

The recognition follows a multi-pass approach, i.e. a first pass using a speaker independent acoustic model, followed by segmentation and clustering of segments, a second pass using speaker cluster adapted acoustic models.

3 Statistical Machine Translation

Our machine translation system follows an statistical corpus-based approach, based on an n-grams, that offers state-of-the-art results. A source sequence is translated by searching the most probable target sequence given by a bilingual Ngram-based model (Mariño et al., 2006).

Spanish and Catalan are high inflected languages that generate an enormous variability of genre and number agreement among other linguistic challenges. However, most of them could be solved by introducing morphological information and good quality bilingual corpora. First, the morphological information is introduced in the system by means of:

Monolingual-expert rules. Lately, three matters have been addressed: the Catalan apostrophe and clitics, and the Spanish conjunctions.

Categorization. Hours are context-independent and are categorized in order to produce the right forms. Furthermore, the numbers written in letters do not appear in the training corpora, so they are categorized at a previous stage and translated in their arabic form.

Part-of-Speech (POS) information. Genre and number agreement is improved by using additional statistical information provided by the FreeLing tagger. A POS language model trained on the target language helps in the decoding decision.

4 Speech Synthesis

The quality of speech synthesis systems corpus based has improved during the last years considerably, being the new goals to achieve expressive speech synthesis imitating the human voice in several styles (i.e. reading and talking). A breakthrough in speech synthesis requires the development of new models for prosody, emotions and for expressive speech in general.

The main goal of the project concerning speech synthesis is the production of a state-of-the-art Catalan text-to-speech (TTS) system and its integration in a speech-to-speech translation application. This system must be capable of expressing the speaking styles, accents, and voice quality parameters specified in an input message or text.

In order to achieve this general goal, the following tasks are considered: adequate interpretation of the input message or text, production of expressive speech in several speaking styles, voices and languages, new speech generation and voice adaptation algorithms, new voices based on Hidden Markov Models (HMM) and the development of Catalan-Spanish bilingual voices.

5 Achievements

Throughout the project, resources facilitating the development of models and lexica for the technologies involved have been acquired, initial acoustic and language models for speech recognition and translation models for statistical machine translation have been developed and are to be refined. An overall system architecture has been established integrating the described components with UIMA.

Bibliography

Lööf, J., C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, and H. Ney. 2007. The RWTH 2007 TC-STAR Evaluation System for European English and Spanish. En *Interspeech*, páginas 2145–2148, Antwerp, Belgium.

Mariño, J.B., R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, and M.R. Costa-jussà. 2006. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549, December.