Lexicon and Corpora for Speech to Speech Translation (LC-STAR)

Maximilian Bisani
RWTH

Antonio Bonafonte
UPC

Nuria Castell
UPC

Elviira Hartikainen
Nokia

Giulio Maltese
IBM

Asunción Moreno
UPC
asuncion@gps.tsc.upc.es

Shaunie Shammass
NSC

Ute Ziegenhain
Siemens

Resumen: El objetivo del proyecto europeo LC-STAR (Lexica and Corpora for Speech-to-Speech Translation Components) is la realización de córpora y lexicones para reconocimiento automático del habla y conversión texto a voz necesarios para un sistema de traducción automática voz-voz. Durante el proyecto (2002-2005) los lexicones serán especificados, realizados y validados. Los lexicones serán realizados en trece idiomas distintos y contienen información fonética, prosódica y morfosintáctica [1]. En este artículo se presenta una breve descripción del proyecto.
Palabras clave: Corpora, Lexica, Recursos lingüísticos, multilingüe.

Abstract: The objective of the EU-project LC-STAR (Lexica and Corpora for Speech-to-Speech Translation Components) is corpora collection and lexica creation for the purposes of Automatic Speech Recognition (ASR) and Text-to-speech (TTS) that are needed in speech-to-speech translation (SST). During the lifetime of the project (2002-2005) these lexica will be specified, built and validated. Large lexica consisting of phonetic, prosodic and morpho-syntactic content will be provided with well-documented specifications for at least 13 languages [1]. This project description provides a short overview of the speech-to-speech translation lexica in general as well as a summary of the LC-STAR project itself
Keywords: Corpora, Lexica, Language Resources, Multilingual

1 Introduction

Current approaches to the development of speech recognition, text to speech synthesis and speech centered translation necessitate the development of a wide range of Language Resources (LR).

Great advances have been made in the development of annotated speech databases needed for building speech recognition systems for many languages and for many applications in specific acoustical environments. Less attention has been given to the linguistic oriented language resources needed for the language transfer of SST components that include: Flexible vocabulary speech recognition, high quality text-to-speech synthesis, speech centered translation

Such linguistic oriented resources include suitable text corpora for producing lexica that are enriched with phonetic, prosodic and morpho-syntactic information. These generic LRs are needed to build SST components covering a wide range of application domains in different languages. Currently, such large-scale LR are not publicly or commercially available and industrial standards are lacking. The main objective of the LC-STAR project is to make large lexica available for many languages that cover a wide range of domains along with the development of standards relating to content and quality. Pioneering work is being done for defining standards with respect to content and format issues.

For speech-centered translation, the project will focus on statistical approaches allowing an efficient transfer to other languages using suitable LR. The LR needed for this purpose are aligned bilingual text corpora and monolingual lexica with morpho-syntactic information.

2 Overview of the LC-STAR Project

The LC-STAR consortium consists of 4 industrial companies, namely IBM, Nokia, NSC (Natural Speech Communication) and Siemens and 3 universities, RWTH-Aachen (Rheinisch-Westfälische Technische Hochschule Aachen) UM (University of Maribor) and UPC (Universitat Politècnica de Catalunya). SPEX (Speech Processing Expertise) and CST (Center for Sprogteknologi) are responsible for validating the lexica. Project partners have wide experience either in previous speech database projects (e.g. Speech-Dat family, Speecon etc.) or projects relating to machine translation (e.g. Verbmobil).

Currently, the project covers 13 languages from various parts of the world, whereby each partner is responsible for creating lexica for two languages. It

is possible that more languages will eventually be covered (more lexica created) since the project is still open for new external partners. The list of all languages covered are: Italian, Greek, Finnish, Mandarin, Hebrew, US-English, German, Classical Arabic, Turkish, Russian, Spanish, Catalan and Slovenian. The languages that are covered include main languages of the world as well as ones that are less common. In addition, they represent a wide spectrum of language types, including languages constructed with minimal morphological information (e.g. Mandarin) as well as highly inflective languages (e.g. Turkish). The range of language types raises interesting specification and standardization issues, which are addressed within the project.

Each language-specific lexicon consists of three parts: 1) at least 50,000 inflected common word entries covering six major domains, 2) 45,000 proper names covering three major domains and 3) at least 5,000 entries for special voice-driven applications.

For common words, corpora were collected in six major domains: sports and games, finance, news, culture, consumer information and personal communications. These domains were further divided into subdomains. Each corpus was required to be at least 10 million tokens in total. A minimum of 1 million tokens were required in each domain with no upper limit, with the exception of the personal communications domain, where the minimum amount of data was limited to 500,000 tokens. Sources for the data collection included recent electronically available text corpora. To optimize the coverage criterion, the word lists were required to achieve a self-coverage of at least 95% in each domain and at least 95% over all domains. Furthermore, the final wordlist had to contain the most frequent 50,000 entries without singletons, abbreviations and proper names.

For proper names, three major domains were chosen: personal names (first and last names), place names and organizations.

The special application word list consists of numbers, letters, abbreviations and seven major semantic domains related to voice-driven applications. For voice-driven application words, a reference word list of 5,700 entries in US-English was collected and translated into the other languages. This entailed providing example sentences for translation, since there were language-specific considerations involved in the translation process (e.g. case in Russian, morphological considerations in Turkish, etc.)

In order to fulfil the needs of ASR/TTS components of Speech-to-Speech translation applications, the lexica has to contain detailed grammatical, morphological, and phonetic information for each language. The grammatical information needed per language within the scope of the lexica was specified at the beginning of the project [3]. The available information in all languages was merged into a unique list of POS tags (part-of-speech tags). Most of the POS have an internal structure with attributes common to several languages (e.g. number or gender), but some POS relate only to a subset of languages (e.g. case), or are relevant only for specific individual languages (e.g. polarity for Turkish verbs). The advantage of such an approach is that a single description of grammatical features can cover the set of twelve languages.

In addition to grammatical features, morphological (including lemma) and phonetic information is specified for each word in a given lexicon. In agglutinative languages, such as Turkish or Hebrew, some morphological boundary information is also provided. Phonetic transcriptions use SAMPA symbols, which are specified for each lexica; foreign words are phonetized according to the SAMPA set. Syllable boundaries and primary stress are also provided in the phonetic transcriptions. For each word, it is possible to specify more than one POS, and/or more than one lemma, and/or more than one phonetic transcription.

The formal representation of lexica is implemented via an XML-based mark-up language that meets the requirements of representing the linguistic information in a formal, unambiguous manner. Such a representation is both easy to read and easily processed by generic applications. A formally specified grammar (Document Type Definition or DTD) containing all the described linguistic information allows for automatic validation of the XML-based lexica.

Later phases of the project have a goal to create special speech-to-speech translation lexica. At RWTH, speech-to-speech translation experiments using different methods are being carried out.

The main purpose is to find out if machine translation can be improved if more linguistic features are added to translation lexica. Based on the results of these experiments, lexica for translating into seven languages (Catalan, Finnish, German, Hebrew, Italian, Russian and Spanish) will be specified and created. The reference word lists in US-English for these ‘translation lexica’ will be created from aligned corpora covering a tourist domain. A demonstrator showing the language transfer within 3 language pairs (Catalan, Spanish, US-English) will be built.

3 References

[1] Project homepage: http://www.lc-star.com

[2] Ziegenhain, U. et al. Specification of corpora and word lists in 12 languages. Public Project Deliverable D1.1.

[3] Maltese, G. Montecchio, C. et al. General and language specific specification of contents of lexica. Public Project Deliverables D2. 2003.