Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 357-358 recibido 1-05-2009; aceptado 5-06-2009

Baratinoo speaks Spanish as well

Baratinoo también habla español

Elixabete Murguía, Thierry Moudenc, Paul C. Bagshaw
Orange Labs R&D
2 ave. Pierre Marzin, Lannion 22307
{elixabete.murguia, thierry.moudenc, paul.bagshaw}@orange-ftgroup.com
tel. +33 296052353, 296051659, 299124154

Resumen: En esta demo presentaremos el conversor de texto a voz (CTV) desarrollado por Orange Labs R&D. Es un sistema multilingüe (francés, inglés, español y árabe) y explota una técnica de concatenación de unidades basada en corpus. Describiremos la arquitectura general del sistema, así como algunas de las características de los módulos adaptados para el español. Completaremos esta presentación con algunos ejemplos de frases sintetizadas, en particular, de nuestra voz femenina Marta.
Palabras clave: conversor de texto a voz, sistemas multilingües

Abstract: In this demo, the text-to-speech system (TTS) developed by Orange Labs R&D will be presented. It is a multilingual system (French, English, Spanish and Arabic) and exploits a corpus-based concatenative approach. We will describe the general architecture of the system, particularly focusing on some of the characteristics of the modules adapted for Spanish. We will complete this presentation with some examples of synthesized sentences, in particular, from the system's Spanish female voice, Marta.
Keywords: TTS systems, multilingual systems

1 Orange Labs R&D TTS system

Baratinoo is a multilingual concatenative TTS engine developed by Orange Labs R&D. This TTS system was initially developed for French and later adapted to other languages including English, Spanish and Arabic. The general architecture of the system, as well as the strategy adapted for speech synthesis, remains the same for all languages. Language particularities are dealt with by the adaptation of the linguistic analysis modules, together with the tailoring of the criteria and costs calculated for the Viterbi algorithm used in unit selection.

The system presented here is modular and, as in classical concatenative systems, it is divided in two parts: (i) a linguistic analysis part, and (ii) a synthesis part. Both parts are briefly described in the sections below.

1.1 Linguistic Analysis

The first step in processing text entering the system is to tokenise it, i.e. to splice it into minimal items that can be easily processed by subsequent modules. The linguistic analysis per se includes (i) a module for lexical and morphological analysis, (ii) a module for syntactic disambiguation, (iii) a module to insert pauses, and (iv) a module to adjust the pronunciation at a sentence-level.

The linguistic analysis starts with a lexical and morphological analysis of the tokens. Tokens may be regrouped into a unique lexical form (such as locutions) or read to form multiple lexical forms (such as numbers and dates). Each lexical form (commonly a word) is considered in isolation of the others, and is assigned one or more grammatical tags. This may be achieved by applying a set of morphological transformation rules (see 1 a rule that derives a feminine plural form from the masculine singular form) or by associating a set of unambiguous tags with a list of suffixes. The grammatical tags are then used to attribute one or more pronunciations (possibly applying grapheme-to-phoneme conversion rules) to each lexical form. The correspondences between tags and pronunciations are known for each word.

(1) o – as {NC.nsm – NC.npf ; AQ.sm – AQ.pf}.

Following the lexical/morphological analysis, a process of syntactic disambiguation

ISSN: 1135-5948

is needed, so that words assigned with multiple grammatical tags can finally be left with one tag and thus one pronunciation. Disambiguation is further required for the rules in charge of inserting pauses, which depend on a reliable description of the syntactic context of words. Disambiguation is carried out by applying rules that can examine a contextual window of words (up to 9 words to the left and/or right) surrounding the ambiguous item. Disambiguation is done in different steps: first rules that take care of disambiguating locutions and compound words are applied, followed by those which take care of closed-class words (e.g. determiners, pronouns, prepositions). Once closed-class words have been disambiguated, they serve as anchors to rules for open-class words that are more difficult to disambiguate, such as nouns, adjectives or verbs.

The next module takes care of locating points for pause insertion by the application of rules that describe the local syntactic context. This module uses a long list of rules for French. For Spanish, the number of rules has been significantly reduced. There are two types of pause placement rules. First, there are rules that place obligatory pauses (pauses signalled by orthographic marks). Depending on the type of mark, an ascending or descending intonation is assigned by default (e.g. ascending for comma), but rules that account for differing cases are also added. See 2 where intonation after the first comma is ascending.

(2) El Plan nuevo del gobierno, que podría estar listo en medio año, ha sido presentado.

Second, there are rules that place potential pauses. The most representative points of syntactic juncture (e.g. subordinates, between subject and verb or two objects, etc.) are identified and rules are applied at these points. These rules insert an index that indicates the type of intonation (ascending or descending) and the strength of the juncture. For a sentence like 3 two rules will apply inserting the indexes below:

(3) Los cuadros de la exposición organizada-130 son muy buenos-110 a pesar de que nadie se los haya comprado.

Amongst all the potential points for pause insertion, the final candidate(s) is chosen as a result of combining (i) the index (the higher the index, the smaller the probability of a pause), (ii) the number of syllables, and (iii) the position of preceding or following pauses, so that pauses are distributed in a balanced way in the sentence.

Finally, accents are assigned, words are divided into syllables and rules are applied to adjust word pronunciations as a function of their phonetic context in a sentence. For Spanish, there are rules for the allophonic variants of phonemes /b/, /d/, /g/, as well as for semivowels. See the sentence in example 4 where the result of this final stage is shown. Plosive allophones are marked with symbol B, D, G; fricative allophones with BB, DD, GG, accents with a double vowel; the semivowel with J; syllables with a short hyphen; word boundaries with a longer hyphen and pauses with #).

(4) Gana el gremio de siempre.
#GAA-NA—EL—GGREE-MJO—DDE—SJEEM-PRE#

1.2 Synthesis by concatenation

The output of the linguistic processing is a succession of target phonemes augmented with certain features such as their position in the sentence or word. When synthesis module generates a signal, it tries to match these phonemes and specifications. The signal is generated by the concatenation of acoustic units selected from a database of recorded sentences previously segmented into diphones.

The selection of the units is done by a standard dynamic algorithm (Viterbi) which chooses the best succession of units among all the possible candidates. This algorithm minimises a cost function for a succession of n phonemes. Two different costs are taken into account: (i) target cost, such as the position of the phoneme in the word or sentence and (ii) concatenation cost, such as the difference in F0 between two adjacent candidate units.

This TTS system makes no use of a model of prosodic prediction. Besides the prediction of pause insertion, other issues such as duration, energy and intonation are taken care of by concatenating the right units. Thus, a good unit selection is crucial, and the choice of specified selection criteria is of utmost importance. The speech engine interface can be accessed at http://tts.elibel.tm.fr/tts.