Linguistic Resources and Tools for Automatic Speech Recognition in Basque

Bordel G.[1], Ezeiza A.[2], Lopez de Ipina K.[3], Méndez M.[3],

Peñagarikano M.[1],Rico T.[3], Tovar C.[3], Zulueta E [3].

University of the Basque Country

[1]Elektrizitate eta Elektronika Saila, Leioa. {german,mpenagar}@we.lc.ehu.es
[2]Ixa taldea. Sistemen Ingeniaritza eta Automatika Saila, Donostia. ispezraa@sp.ehu.es
[3] Sistemen Ingeniaritza eta Automatika Saila, Gasteiz. {isplopek,jtzmebej, iszripat,vcbtomoc,iepzugue}@we.lc.ehu.es

Abstract

The development of Automatic Speech Recognition systems requires of appropriated digital resources and tools. The shortage of digital resources for minority languages slows down the development of ASR systems. Furthermore, the development of the system in Basque represents even a bigger challenge, since it has a very uncommon morphology.

Nevertheless, recent advances have been achieved thanks to the Basque mass media, particularly the Basque Public Television (EITB) and the only daily newspaper in Basque (Egunkaria). These media have provided digital multimedia resources for the development of a digital library for both Basque and Spanish (Bordel, et al., 2004), and these rich resources have been employed to develop new tools for Speech Recognition.

Introduction

There is considerable current interest in development of digital resources and tools for Automatic Speech Recognition systems. Minority languages constitute a bigger exertion because of the lack of resources in general, and Basque is not an exception.

Recent works address this shortfall processing digital resources from Basque mass media. The interest is mutual, since mass media aim to employ Human Language Technology based applications to search and index multimedia information.

The purpose of the experiments reported in this paper is to develop appropriated resources for Automatic Speech Recognition in Basque. Both Basque and Spanish are official in the Basque Autonomous Community, and they are used in the Basque Public Radio and Television *EITB* (EITB) and in most of the mass media of the Basque Country (radios and newspapers).

Basque is a Pre-Indo-European language of unknown origin and it has about 1.000.000 speakers in the Basque Country. It presents a wide dialectal distribution, being six the main dialects. This dialectal variety entails phonetic, phonological, and morphological differences.

Moreover, since 1968 the Royal Academy of the Basque Language, *Euskaltzaindia* (Euskaltzaindia) has been involved in a standardisation process of Basque. At present, morphology, which is very rich in Basque, is completely standardised in the unified standard Basque, but the lexical standardization process is still going on.

The standard Basque, called “Batua”, has nowadays a great importance in the Basque community, since the public institutions and most of the mass media use it. Furthermore, all the digital resources developed for this job are in this standard version of Basque, and thus, it has been used to develop all the new tools for Speech Recognition that we present in this report.

Besides, these new tools are classified in two sides. In the one hand, an automatic tool to generate appropriated Lexicons in Basque has been generated, in order to widely use the contents of the digital libraries.

In the other hand, the information extracted from the broadcast news videos and the newspaper texts has been used to develop a prototype of CSR system based on the HTK tool (Young, et al. 1997). The HTK tool-based system produced interesting results, but a more sophisticated approach was tried analysing the research developed for Japanese, a language that has a similar phonetic concerns to Basque. Thus, a prototype of simple tasks was developed based on Julius, a Multipurpose Large Vocabulary CSR Engine (Lee et al., 2001). This tool is based on word N-grams and context-dependent Hidden Markov Models, and the prototype produced significant results.

The following section describes the resources developed from the data provided by the aiding mass media. The third section presents the tools developed during this work, the fourth section describes the processing of the data, and finally, conclusions are summarised in the last section.

Multimedia Resources

In order to develop new tools for ASR systems, this project has demanded a previous work involving the development of richer digital resources. As it has been mentioned before, the main Basque media have collaborated in this development, providing videos from daily programs of broadcast news and newspaper texts.

Next follows a relation of the resources provided and created within the scope of this project:

- 6 hours of video in MPEG4 (WMV 9) format of "Gaur Egun" program, the daily program of broadcast news in Basque directly provided by the Basque Public Radio and Television (EITB).
- 6 hours of audio (WAV format) extracted from the video (MPEG4) files.
- 6 hours of audio transcription in XML format containing information about speaker changes, noises and music fragments, and each word's phonetic and orthographical transcription including word's lemma and Part-Of-Speech disambiguated tags.
- 1 year of scripts, in text format, of the "Gaur Egun" program.
- 1 year of local newspapers in Basque, Euskaldunon Egunkaria (Egunkaria), in text format.
- Lexicon extracted from the XML transcription files, including phonological, orthographical, and morphological information.

Automatic Speech Processing tools

The Automatic Speech Processing tools developed have been based on existing tools for Basque. The improvements included in this late part of continuous development are specially linked with the new resources adverted in the previous section. As it has been brought up in the introduction, two are the main lines of work: Lexicon development tools and Continuous Speech Recognition engines.

Lexicon development tools

Lexicon development has critical dependency with morphological features of a given language. Namely, Basque is an agglutinative language with a special morpho-syntactic structure inside the words (Alegria et al., 1996) that may lead to intractable vocabularies of words for a CSR when the size of task is large.

The lexicon extracting tool for Basque *AHOZATI* (Lopez de Ipina et al., 2002) tackles this problem using morphemes instead of words in order to define the system vocabulary. This approach has been evaluated over three textual samples analysing both the coverage and the Out of Vocabulary rate, using words and pseudo-morphemes obtained by the automatic morphological segmentation tool. Fig. 1 and Fig. 2 illustrate the analysis of Coverage and Out of Vocabulary rate over the textual sample from the broadcast news scripts. When pseudo-morphemes are used, the coverage in texts is better and complete coverage is easily achieved. OOV rate is higher in this sample.

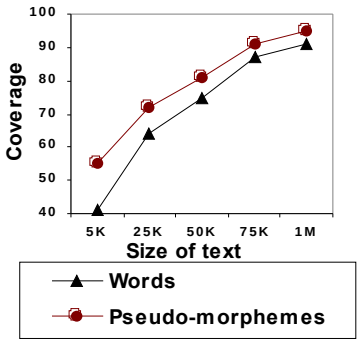

Fig. 1: Coverage for the textual sample.

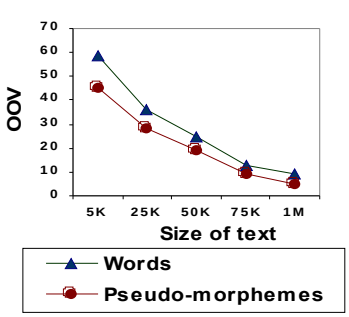

Fig. 2: OOV rate for the textual sample.

In addition, AHOZATI has been improved during this project extracting the lexicon from the XML transcription files mentioned in the *Multimedia Resources* section.

CSR systems

Hidden Markov Models are widely used in Continuous Speech Recognition systems, and toolkits like HTK (Young, et al. 1997) are well known and tested. Moreover, the prototypes we have developed yet have been principally adapted to the HTK standards. Thus the first use given to the recently obtained enriched lexicon was to be adapted to the HTK toolkit.

The HTK tool-based system produced interesting results, but due to its limitations both in licensing terms and language modeling, a more refined approach was tried utilising the open source tools developed for Japanese, a language that has a similar phonetic concerns to Basque.

Thus, a prototype of simple tasks was developed based on Julius, a Multipurpose Large Vocabulary CSR Engine (Lee et al., 2001). This tool is based on word N-grams and context-dependent Hidden Markov Models, and uses similar input files to HTK. The use of pseudomorphemes and N-grams fits in Basque much more than the word approach of HTK. This conclusion has appeared in our first experiments, although further evaluation has to be done.

Processing Methodology

Processing of the video data

The video data used in this work has been provided directly by the Basque Public Radio and Television. The format used to store the broadcast contents is MPEG4 (WMV 9), and the Basque Public Radio and Television has been very kind offering us all these resources.

Processing of the audio data

The audio data has been extracted out from the MPEG4 video files, using FFmpeg free software[1]. The audio files have been stored in WAV format (8 KHz, 16 KHz, linear, 16 bits).

When the audio data was ready, the XML label files were created manually, using the Transcriber free tool (Barras et al., 1998). The XML files include information of distinct speakers, noises, and paragraphs of the broadcast news. The files also contain phonetic and orthographic information of each of the words. Basque XML files include morphological information such as each word's lemma and Part-Of-Speech tag.

The Lexicon aforementioned has been extracted using this transcribed information. The Lexicon stores information of each different word that appears in the transcription.

Processing of the textual data

There are two independent types of textual resources: The text extracted from the newspaper Euskaldunon Egunkaria (Egunkaria), and the scripts of the "Gaur Egun" program.

All of the texts were processed to include morphologic information such as each word's lemma and Part-Of-Speech tag. Using all the information, a Lexicon for each language has been extracted taken into account the context of the word in order to eliminate the ambiguity. This Lexicon differs from the Lexicon extracted from the transcription files, and it is been developed to be used in testing and evaluating scenarios for Machine Learning techniques.

Concluding Remarks

This work deals with the development of appropriated resources and tools for an automatic index system of broadcast news in Basque. Since Basque is an agglutinative language, analysis of coverage and words OOV has been carried out in order to develop appropriated Lexicon. New resources were developed during this work. Subsequently, the Lexicon Extraction tool AHOZATI was improved with the new resources and finally two CSR prototypes were developed based on Hidden Markov Models. First, a HTK toolkit-based word prototype, and second, a Julius toolkit-based N-gram prototype.

Acknowledgments

We would like to thank all the people and entities that have collaborated in the development of this work, specially: EITB, Gara and Euskaldunon Egunkaria.

References

EITB Basque Public Radio and Television, *http://www.eitb.com/*

Euskaltzaindia, *http://www.euskaltzaindia.net/*

Peñagarikano M., Bordel G., Varona A., Lopez de Ipina: "Using non-word Lexical Units in Automatic Speech Understanding", Proceedings of IEEE, ICASSP99, Phoenix, Arizona.

[1] Available on-line at http://ffmpeg.sourceforge.net

Lopez de Ipiña K., Graña M., Ezeiza N., Hernández M., Zulueta E., Ezeiza A., Tovar C.: " Selection of Lexical Units for Continuous Speech Recognition of Basque", Progress in Pattern Recognition, pp 244-250. Speech and Image Analysis, Springer. Berlin. 2003.

Lopez de Ipina K., Ezeiza N., Bordel. N., Graña M.: "Automatic Morphological Segmentation for Speech Processing in Basque" IEEE TTS Workshop. Santa Monica USA. 2002.

Egunkaria, Euskaldunon Egunkaria, the only newspaper in Basque, which has been recently replaced by Berria, on-line at *http://www.berria.info/*

Barras C., Geoffrois E., Wu Z., and Liberman M.: "Transcriber: a Free Tool for Segmenting, Labelling and Transcribing Speech" First International Conference on Language Resources and Evaluation (LREC-1998).

Alegria I., Artola X., Sarasola K., Urkia M.: "Automatic morphological analysis of Basque", Literary & Linguistic Computing Vol,11, No, 4, 193-203, Oxford Univ Press, 1996.

Young S., J. Odell, D. Ollason, V. Valtchev, P. Woodland, *The HTK BOOK, HTK 2.1 Manual*, 1997

A. Lee, T. Kawahara and K. Shikano. "Julius --- an open source real-time large vocabulary recognition engine."In Proc. European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1691--1694, 2001.

Bordel G., Ezeiza A. , Lopez de Ipina K., Méndez M. ,Peñagarikano M., Rico T., Tovar C., Zulueta E."Development of Resources for a Bilingual Automatic Index System of Broadcast News in Basque and Spanish", Fourth International Conference on Language Resources and Evaluation (LREC-2004).