STATISTICAL METHODS FOR SPEECH TECHNOLOGIES IN BASQUE LANGUAGE

M. Inés Torres[1]*, Víctor Guijarrubia*[1]*, Raquel Justo*[1]*, Alicia Pérez*[1]*, Francisco Casacuberta*[2]

[1]Dep. Electricity and Electronics. University of the Basque Country
manes.torres@ehu.es
[2]Dep. of Information Systems and Computation. Technical University of Valencia
fcn@dsic.upv.es

ABSTRACT

The overall goal of this work is to build a speech-input decoder and translation application for Basque, Spanish and English, which allows to speak in whatever the aforementioned languages, and according to the identified language it proceeds to obtain its text transcription and translation into the other two languages. This application gathers different technologies such as language identification, recognition and translation, all of them developed on the basis of statistical methods. In addition, it entails a real challenge, as there are scarce resources developed for Basque language. After carrying out an analysis of different approaches, adapted methods for Basque language features have been developed and assessed.

1. INTRODUCTION

Current trends in both automatic speech recognition (ASR) and machine translation go towards the use of statistical methods as they have proved to offer a very competitive performance. In this work, we made use of them but we also proposed new approaches in order to obtain a better adaptation of the classical methods to the specific features of Basque language.

Basque is a pre-Indoeuropean language of unknown origin. It shares official status, along with Spanish, in the Basque Country, nevertheless it is spoken in a few more regions, such as in south west France and in small American communities. It is a minority language, and thus, great efforts are being made in order to enrich the currently scarce linguistic resources. Regarding the morphology, Basque is an extremely inflected language both in nouns and verbs.

Considering the syntactic structure, both languages present a different arrangement of the words within the sentence. Therefore, an appropriate *language model* capable of capturing the specific structure of the sentences in Basque has a great importance when speech recognition is carried out. In this work a category-based language model was used. Category-based models allow to capture relationships related to the structure of the sentences and in addition they have shown to be a good choice to face the issues derived from the resource scarcity, as it is the case.

From a phonetic point of view, the set of Basque phones does not differ much from the Spanish one. The two languages share the same vowel triangle (only five vowels). Nevertheless, Basque includes larger sets of fricative and affricate sounds [1].

In this work, several fields of the natural language processing have been explored and adapted for Basque language: language identification between Basque, Spanish and English; language modeling and categorization; speech translation. Needless to say, the main tool involved in all these fields is the automatic speech recognition system, which has been developed in this group and constitutes the state of the art in what continuous speech recognition systems concerns. The acoustic models employed are continuous Hidden Markov Models and regarding to the language model (LM), a *k-testable in the strict sense* [2] LM is used. This model integrates n-gram models (n ranging from 1 to k) and allows for back-off smoothing.

The following of this paper is organized as follows: section 2 is devoted to describe the task and corpus object of the study. Sections 3, 4 and 5 briefly describe statistical methods, experimental results and specific challenges that have to be faced when Basque language is involved in language identification, categorization and machine translation respectively. Finally, in section 6, a discussion of the overall results and proposals for future work are reported.

2. TASK AND CORPUS

METEUS is a trilingual text and speech corpus in Basque, Spanish and English. It consists of weather forecast reports picked up from the Internet in Spanish and Basque and later translated into English by a professional translator. The main features are shown in Table 1.

This is the first multilingual corpus that joins natural language text and speech in Basque. It seems to be a

This work has been partially supported by the University of the Basque Country under grant GIU07/57, by CYCIT under grant TIN2005-08660-C04-03 and by Consolider Ingenio-2010 program MIPRCV (CSD2007-00018).

suitable choice for comparison purposes between different languages, and above all, for statistical speech translation, a vaguely explored field in Basque language.

With regard to the speech test, it is a training-independent set that consists of 500 different sentences. Each sentence has been recorded for at least 3 speakers, getting as a result a total of 1800 utterances by 36 speakers for each language. Notice that since the speech sub-test is training independent (instead of being a randomly selected subset), it is suitable as a benchmark to evaluate the systems under the worst situation. Therefore, the results obtained with this speech test are pessimistic, and thus, appropriate in order to establish the lower threshold of the system.

		Basque	**Spanish**	**English**
Training	Sentences	14,615		
	Different sentences	7,523	7,198	6,634
	Words	187,195	191,156	195,575
	Vocabulary	1,135	702	498
	Average Length	12.8	13.0	13.3
Test	Sentences	500		
	Words	8,274	8,706	9,150
	Average Length	16.5	17.4	18.3
	Perplexity (3grams)	6.7	4.8	5.8

Table 1. Main features of METEUS corpus.

The figures of the Table 1 show that there is a great difference in terms of vocabulary for the three languages within the same application (see Table 1). Basque language is a highly inflected language with more than 25 declension cases, whereas English is morphologically simpler. The reliability of the statistics over a smaller number of words with the same amount of training sentences, is likely to be higher, therefore, we expect worse probability distributions to be estimated over the models involving the Basque language.

3. LANGUAGE IDENTIFICATION

Language identification (LID) is a classical problem that is strongly tied to multilingual speech recognition and dialogue systems. The ultimate goal of any LID system is to identify the language being used by an unknown speaker. It has been addressed in the past using a variety of tactics; for instance, those exploiting prosodic cues as rhythm or intonation. Nevertheless, most of them are based on speech recognition approximations: phone decoding approaches, which rely on phone sequences; or large-vocabulary continuous-speech recognition approaches, which operate based on full lexical sequences. A thorough analysis discussing the current state of the LID systems can be consulted here [3].

In general, a LID system is composed of three components: a speech tokenizer that converts the speech into a sequence of tokens; a statistical language model which captures the relationships between the tokens; and a classifier that hypothesizes a language from among the set of languages.

In this paper, we focus on phone decoding approaches. These techniques rely on acoustic phonetic decoders, which find the best sequence of phonetic units depending on the input signal. Some phonotactic models can then be used to analyze these sequences and assign some scores to each language. These phonotactic models can be applied after the decoding process, that is a phone recognition followed by *n*-gram Language Modeling (PRLM), or during the decoding process (PPR) [4]. The language of the utterance is selected to be that with the best score.

The results, in terms of LID accuracy, are summarized in Table 2. In this case, we opted for using a PPR approach, since this yielded the best results [5].

	Basque	**Spanish**	**English**
Accuracies(%)	99.67	99.89	95.33

Table 2. LID accuracies values.

As can be derived, for Spanish and Basque, accuracies of nearly 100% are achieved. For English, the accuracies are also competitive, but slightly lower than those for Spanish and Basque. The acoustic modeling could be the reason for this. Whereas the acoustic models for English are trained using a phonetical transcription based on a dictionary, for Basque and Spanish this transcription is performed using rules. So the HMM sets for Basque and Spanish are better estimated and the acoustic scores are higher. To improve the results, more accurate phonotactic models would be required. Another option could be incorporating different sources of information so that the classifier has more cues to hypothesize the uttered language.

4. CATEGORIZATION AND SPEECH RECOGNITION

Nowadays statistical language models (word n-gram LMs, k-tss models, etc.) are being used in *automatic speech recognition* (ASR) systems. Large amount of training data are required to get a robust estimation of the parameters defining such models. However, there are numerous ASR applications for which the amount of training material available is rather limited. One of the ways to deal with data sparseness is to cluster the vocabulary of the application into a smaller number of categories or classes. Thus, an alternative approach as a class n-gram LM (M_c) [6] could be used. Class n-gram LMs are more compact and generalizes better on unseen events. Nevertheless, relations among the categories of words are only captured, while it is assumed that the inter-word transition probability depends on the word classes. This fact degrades the performance of the ASR system.

In order to avoid the loss of information associated with the use of a class n-gram LM, alternative approaches might be used, e.g. [7]. We propose a different approach that takes advantage of two information sources: words and categories. This approach could be understood as a LM based on categories consisting of segments (or sequences of words) instead of being made up of isolated words. We have employed two different ways of dealing with sequences of words inside the classes. Thus, two different approaches to this kind of LM can be considered : M_{sl} and M_{sw} [8]. These models take into account the relations among the words that take part in the segments of a category.

In this work, we study whether different class-based LMs (M_c, M_{sl} and M_{sw}), integrated into an ASR system, can improve the ASR system performance when experiments over the Basque part of the METEUS corpus are carried out. Let us notice that Basque is a minority language and therefore, few training material is available. Due to this fact, this task is well-suited to study improvements derived from categorization within the language model.

The employed categories were automatically generated by the aid of *mkcls* [9], a free toolkit designed to train word classes based on a maximum-likelihood-criterion. On the other hand, the set of word sequences employed within M_{sl} and M_{sw} models were obtained using also a statistical criterion. Specifically the most frequent n-grams of the corpus were selected as segments. In this sense and in order to avoid rare or unimportant n-grams only segments exceeding a minimum number of occurrences were considered.

Different sets of 300, 400, 500 and 600 categories were generated and the correspoding class-based language models were inferred. The proposed language models were then integrated into the ASR system and evaluated in terms of *word error rate* (WER). The obtained WER results are shown in Table 3 along with that obtained under a classical word-based LM. As can be seen in Table 3, better results are obtained when using word segment based categories (in both M_{sl} and M_{sw} models) than when employing classical class n-gram models (M_c). On the other hand, regarding the experiments carried out with M_{sl} model, a significant drop of WER is observed compared to the M_{sw} model for all of the selected sets of categories. Furthermore, the result obtained with M_{sl} and 500 classes slightly improves the WER value obtained with the word-based LM M_w (a 1.7%).

WER (%)				
stat. cat.	M_{sl}	M_{sw}	M_c	M_w
300	6.17	6.68	6.66	5.91
400	6.01	6.59	6.53	
500	5.81	6.37	6.47	
600	6.02	6.51	6.62	

Table 3. WER results using METEUS, for a classical word n-gram LM (M_w), a classical class n-gram LM M_c and the two proposed category-based LMs containing segments of words (M_{sw} and M_{sl}). Different sets of 300, 400, 500 and 600 statistical classes were employed in all category-based LMs

5. SPEECH TRANSLATION

Stochastic finite state transducers (SFST), thoroughly described in [10], have proved to be useful for both text and speech input machine translation applications [11].

The SFST is characterized by both the topology and the probability distributions. These distinctive features can be automatically learnt from bilingual corpora by efficient algorithms, such as GIATI (Grammar Inference and Alignments for Transducers Inference) [12]. Regarding the topology, a 3-TSS with Witten-Bell smoothing has been selected for all the following experiments.

Once the SFST has been inferred, given a source sentence $\mathbf{s}$, the decoding process can be summarized in equation (1). This expression involves a searching for the most likely target string $\widehat{\mathbf{t}}$, being $d(\mathbf{s},\mathbf{t})$ a path in the SFST, compatible with both the input sentence $\mathbf{s}$ and the output $\mathbf{t}$. Therefore, the searching criterion in the SFST deals with the joint probability of sentence pairs.

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} P(\mathbf{s},\mathbf{t}) \approx \arg\max_{\mathbf{t}} \max_{d(\mathbf{s},\mathbf{t})} P(d(\mathbf{s},\mathbf{t})) \quad (1)$$

Furthermore, speech input translation aims at looking for the likeliest target language string ($\mathbf{t}$) given the acoustic representation ($\mathbf{x}$) of a source language hidden string ($\mathbf{s}$), as shown in equation (2). Once again, the most likely target is found in terms of Viterbi algorithm.

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{x}) = \arg\max_{\mathbf{t}} \sum_{s} P(\mathbf{s},\mathbf{t}|\mathbf{x}) \quad (2)$$

SFST allows to integrate acoustic models within the network so that speech translation can be carried out at a single decoding step [11]. This is an alternative to the commonly decoupled architecture that makes use of a text-to-text translation system in serial with a speech recognition system. The integrated architecture has shown to get better translation results in many tasks than the decoupled one, in addition, it is rather efficient in what time cost concerns.

In practice, the procedure to build such an integrated architecture from a common ASR is not other than making use of the SFST instead of the LM. In fact, the input projection of a transducer can be seen as an input LM. Regarding the lexical model of the ASR, both the input and the output substrings have to be stored in order to produce both the recognized sequence and its translation.

Experimental results of Table 4 show the performance of the SFST for text and speech input translation. The commonly used automatic evaluation measures have been selected in order to asses the performance of the system:

word error rate (WER) and *bilingual evaluation understudy* (BLEU).

	Basque→Spanish		Basque→English	
	WER	**BLEU**	**WER**	**BLEU**
Text	43.66	48.31	50.01	44.97
Speech	47.87	45.12	54.93	42.19

Table 4. Speech input machine translation results.

Analyzing the translations in detail, it has been shown that the most frequent error sources are the following ones: wrong order, either in local or long range; wrong lexical choices, related to either style or case; wrong punctuation.

There are some specific features of the Basque language to bear in mind in order to improve the translation models. On the one hand, the agglutinant nature, and on the other hand, the long distance reordering issues, since the usual construction for both Spanish and English is as Subject + Verb + Objects, while for Basque is as Subject + Objects + Verb. These difficulties have been proved to be more efficiently tackled in terms of phrases as a translation unit than in term of words [13], at least regarding Spanish to Basque translation.

6. CONCLUSIONS AND FUTURE WORK

Summing up, the following concluding remarks have been reached with regard to the different techniques explored in this work.

Good trilingual language identification accuracies are achieved. The Basque language is almost always properly identified even if it is acoustically similar to Spanish. A combination of different techniques should be explored to improve the LID accuracies.

Regarding category-based LMs for the *speech recognition system*, segment-based categories, taking advantage of two information sources, were needed to obtain improved results of WER. On the other hand, for further work, different criteria could be explored in the categorization and segmentation process, e.g. a linguistic one.

Better translation results could be obtained including more linguistic knowledge in the statistical model. Furthermore, alternative methodologies to exploit both statistical and linguistic knowledge sources should be explored. In future work we aim at studying *factored translation models* combining both running words, lemmas, POS and statistical Tags within a finite-state framework.

7. BIBLIOGRAPHY

[1] V. Guijarrubia, M. I. Torres, and L.J. Rodríguez, "Evaluation of a spoken phonetic database in basque language," in *Proceedings of LREC 2004*, Lisboa, 2004, vol. 6, pp. 2127–2130.

[2] M. I. Torres and A. Varona, "k-tss language models in speech recognition systems," *Computer Speech and Language*, vol. 15, no. 2, pp. 127–149, 2001.

[3] A. F. Martin and A. N. Le, "The current state of language recognition: NIST 2005 evaluation results," in *Proceedings of the IEEE Odyssey 2006, SLR Workshop*, Puerto Rico, 2006.

[4] M. A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and n-gram modelling," in *Proceedings of ICASSP-94*, Adelaide, Australia, 1994, vol. 1, pp. 305–308.

[5] V. G. Guijarrubia and M. I. Torres, "Phone-segments based language identification for spanish, basque and english," in *CIARP*, Luis Rueda, Domingo Mery, and Josef Kittler, Eds. 2007, vol. 4756 of *Lecture Notes in Computer Science*, pp. 106–114, Springer.

[6] T. R. Niesler and P. C. Woodland, "A variable-length category-based n-gram language model," in *IEEE ICASSP-96*, Atlanta, GA, 1996, IEEE, vol. I, pp. 164–167.

[7] I. Zitouni, "Backoff hierarchical class n-gram language models: effectiveness to model unseen events in speech recognition," *Computer Speech and Language*, vol. 21, no. 1, pp. 99–104, 2007.

[8] R. Justo and M. I. Torres, "Phrases in category-based language models for spanish and basque ASR," in *Proceedings of Interspeech 2007*, Antwerp, Belgium, August 27-31 2007, pp. 2377–2380.

[9] F. J. Och, "An efficient method for determining bilingual word classes," in *EACL '99*, Bergen, Norway, June 1999, pp. 71–76.

[10] E. Vidal, F. Thollard, F. Casacuberta C. de la Higuera, and R. Carrasco, "Probabilistic finite-state machines - part II," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 7, pp. 1025–1039, 2005.

[11] F. Casacuberta and E. Vidal, "Machine translation with inferred stochastic finite-state transducers," *Computational Linguistics*, vol. 30, no. 2, pp. 205–225, 2004.

[12] F. Casacuberta and E. Vidal, "Learning finite-state models for machine translation," *Machine Learning*, vol. 66, no. 1, pp. 69–91, 2007.

[13] A. Pérez, M. I. Torres, and F. Casacuberta, "Joining linguistic and statistical methods for Spanish-to-Basque speech translation," *Speech Communication*, 2008, doi:10.1016/j.specom.2008.05.016.