Steps taken in Spanish-Basque speech translation using stochastic finite-state transducers

Alicia Pérez, M. Inés Torres
Electricity and Electronics
University of the Basque Country
manes.torres@ehu.es

Francisco Casacuberta
Information Systems and Computation
Polytechnic University of Valencia
fcn@iti.upv.es

Abstract

The goal of this paper is to summarise a work in progress focused on speech translation making use of stochastic finite-state transducers (SFSTs). The aim of these devices lays on their versatility to integrate acoustic models within translation models.

Our interest lays on Spanish and Basque, official languages in the 600.000 inhabitant Basque Autonomous Community. These two languages, show remarkable differences in both syntax and morphology, as a result they represent a challenge for SFSTs. In addition, we deal with little linguistic resources due to the fact that Basque is a minority language.

1 On the use of SFSTs for ST

Finite-state models constitute an elementary framework not only in syntactic pattern recognition but also in language processing. Particularly, stochastic finite-state transducers (SFSTs) have proved to be of use in machine translation of restricted domains. As it is known, SFSTs can be inferred from positive bilingual samples following GIATI algorithm (Casacuberta and Vidal, 2004; González and Casacuberta, 2009). In a few words, given the bilingual training set, GIATI looks for a monotonic segmentation and next generates a stochastic regular grammar made up of bilingual symbols (source words together with target phrases), yielding an SFST.

What make SFSTs interesting for speech translation (ST), besides of the fast decoding algorithms they rely on, is their versatility to get them integrated with other finite-state models.

Decoupled architectures tackle speech translation in two consecutive decoding steps. Basically, the first step converts speech utterances into text transcription, and the second step consists on text to text translation of the recognised string. Admittedly, there are different approaches that differ on the amount and type of information rendered from the first stage to the second one. As integration is concerned, one of the challenges of speech translation consists on exploring different ways of integrating both acoustic and translation knowledge sources and translation in an attempt to make them collaborate. Intuitively, cooperative models might yield more accurate estimated hypotheses than those making decisions in an isolate manner. the most of both knowledge sources. In (Pérez et al., 2010) it was proved that integrated architectures deal with significantly more accurate hypotheses than decoupled architectures.

Apart from the ability to integrate acoustic and translation models, SFSTs have also proved to allow the integration of multiple languages in order to carry out multi-target speech translation (Pérez et al., 2007a). As a result, speech was translated simultaneously into several languages.

2 Overcoming challenges of Basque

Basque language is a minority language of unknown origin by contrast to Spanish, which is a Romance language. While both languages co-exist in the Basque Autonomous Community, they differ in both morphology and syntax.

As for morphology, Basque (by contrast to Spanish) is very productive in both noun and verbs, with more than 17 declension cases that can be recur-

sively appended to a lemma. As a result, a Basque word tends to be translated into Spanish by more than one word.

Moreover, the morphology of a word might include syntactic features, e.g. the Basque word *irakasleek* means *the teachers* as the subject of a transitive clause (Fig. 1).

In order to tackle the rich morphology of Basque, in (Pérez et al., 2007b) phrase-based SFSTs (PB-SFST) were proposed within GIATI framework. Those PB-SFSTs represented a step ahead with respect to previous SFSTs since the monotonic bilingual segmentation groups not only words in the target language but also words in the source language. In this line, both statistically and linguistically motivated phrases were explored. Amongst the linguistically motivated phrases morphologically and syntactically motivated ones were distinguished (Pérez et al., 2008). Syntactically motivated phrases improved the performance of the system significantly.

As a consequence of the rich morphology of Basque, inflected words show a little repetition ration within the corpus. As an alternative mechanism to deal with sparsity of data, categorisation was used yielding a hierarchically arranged SFST in (Justo et al., 2010) . This approach allowed to categorise the bilanguage and infer specialised SFSTs for each category, which, in addition, allowed to integrate the acoustic models in the same network. While the formulation of the models is neat, experimentally did not offer much benefits in terms of performance. Nevertheless, it might had to do with the small dimensions of the corpus with which the experiments were carried out.

Regarding the syntax, Spanish tend to follow SVO arrangement while Basque would follow SOV arrangement. As a result very long distance alignments are frequent, and this is a big deal for SFSTs under GIATI approach. The SFSTs which we are dealing with have shown a limited ability to cope with reordering.

preffix (*make it*)	verb stem (*learn*)	suffix (*collective*)	det. plural (*the*)	ergative case (*subject*)
ira-	- kas -	- le -	- e -	- k

Figure 1: Analysis of a Basque word.

3 Concluding remarks and further work

Speech translation making use of SFSTs offers a versatile framework. Recognised utterance and its translation can be obtained in a single-pass decoding strategy.

PB-SFSTs under GIATI approach have shown to be useful to tackle speech translation between Spanish and Basque. PB approach deals with more accurate alignments than word-based one. In addition, gathering words into phrases helps alignments not happen at so long distance, and thus overcome one of the weakness of regular SFSTs. Reordering still represents an open problem for SFSTs facing this pair of languages.

Since Basque is a minority language, linguistic resources are limited. The aforementioned methods were explored with a restricted domain corpus. Admittedly, in order to draw solid conclusions we should experiment with ample-domain corpora. On this account, we are currently making efforts to collect a EuroParl-like corpus for Basque with text and speech.

References

F. Casacuberta and E. Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205–225.

J. González and F. Casacuberta. 2009. GREAT: a finite-state machine translation toolkit implementing a Grammatical Inference Approach for Transducer Inference. In *EACL workshop on Computational Linguistics Aspects of Grammatical Inference*, 24–32

R. Justo, A. Pérez, M. Inés Torres, and F. Casacuberta. 2010. Hierarchical finite-state models for speech translation using categorization of phrases. In *11th International Conference on Intelligent Text Processing and Computational Linguistics*

A. Pérez, M. I. Torres, M. T. González, and F. Casacuberta. 2007a. An integrated architecture for speech-input multi-target machine translation. In *Proc. NAACL-HLT*, 133–136

A. Pérez, M. I. Torres, and F. Casacuberta. 2007b. Speech translation with phrase based stochastic finite-state transducers. In *Proc. IEEE ICASSP*

A. Pérez, M.I. Torres, and F. Casacuberta. 2008. Joining linguistic and statistical methods for Spanish-to-Basque speech translation. *Speech Communication*.

A. Pérez, M.I. Torres, and F. Casacuberta. 2010. Potential scope of a fully-integrated architecture for speech translation. In *Proc. EAMT10*