An integrated architecture for speech-input multi-target machine translation

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

M. Teresa González, Francisco Casacuberta
Dep. of Information Systems and Computation
Technical University of Valencia
fcn@dsic.upv.es

Abstract

The aim of this work is to show the ability of finite-state transducers to simultaneously translate speech into multiple languages. Our proposal deals with an extension of stochastic finite-state transducers that can produce more than one output at the same time. These kind of devices offer great versatility for the integration with other finite-state devices such as acoustic models in order to produce a speech translation system. This proposal has been evaluated in a practical situation, and its results have been compared with those obtained using a standard mono-target speech transducer.

1 Introduction

Finite-state models constitute an important framework both in syntactic pattern recognition and in language processing. Specifically, stochastic finite-state transducers (SFSTs) have proved to be useful for machine translation tasks within restricted domains; they usually offer high speed during the decoding step and they provide competitive results in terms of error rates (Mohri et al., 2002). Moreover, SFSTs have proved to be versatile models, which can be easily integrated with other finite-state models (Pereira and Riley, 1997).

The article (Casacuberta and Vidal, 2004) explored an automatic method to learn an SFST from a bilingual set of samples for machine translation purposes, the so-called GIATI (*Grammar Inference and Alignments for Transducers Inference*). It described how to learn both the structural and the probabilistic components of an SFST making use of underlying alignment models.

A multi-target SFST is a generalization of standard SFSTs, in such a way that every input string in the source language results in a tuple of output strings each being associated to a different target language. An extension of GIATI that allowed to infer a multi-target SFST from a multilingual corpus was proposed in (González and Casacuberta, 2006). A syntactic variant of this method (denoted as GIAMTI) has been used in this work in order to infer the models from training samples as it is summarized in section 3.

On the other hand, speech translation has been already carried out by integrating acoustic models into a SFST (Casacuberta et al., 2004). Our main goal in this work is to extend and assess these methodologies to accomplish spoken language multi-target translation. Section 2 deals with this proposal by presenting a new integrated architecture for speech-input multi-target translation. Under this approach spoken language can be simultaneously decoded and translated into m languages using a unique network. In section 4, the performance of the system has been experimentally evaluated over a trilingual task which aims to translate TV weather forecast into two languages at the same time.

2 An integrated architecture for speech-input multi-target translation

The classical architecture for spoken language multi-target translation involves a speech recogni-

Proceedings of NAACL HLT 2007, Companion Volume, pages 133–136,
Rochester, NY, April 2007.

tion system in a serial architecture with m decoupled text-to-text translators. Thus, the whole process involves $m+1$ searching stages, a first one for the speech signal transcription into the source language text string, and further m for the source language translation into the m target languages. If we replaced the m translators by the multi-target SFST, the problem would be reduced to 2 searching stages. Nevertheless, in this paper we propose a natural way for acoustic models to be integrated in the same network. As a result, the input speech-signal can be simultaneously decoded and translated into m target languages just in a single searching stage.

Given the acoustic representation ($\mathbf{x}$) of a speech signal, the goal of multi-target speech translation is to find the most likely m target strings ($\mathbf{t}^m$); that is, one string ($\mathbf{t}_i$) per target language involved ($i \in \{1, \ldots, m\}$). This approach is summarized in eq. (1), where the hidden variable $\mathbf{s}$ can be interpreted as the transcription of the speech signal:

$$\widehat{\mathbf{t}^m} = \arg\max_{\mathbf{t}^m} P(\mathbf{t}^m|\mathbf{x}) = \arg\max_{\mathbf{t}^m} \sum_{\mathbf{s}} P(\mathbf{t}^m, \mathbf{s}|\mathbf{x}) \tag{1}$$

Making use of Bayes' rule, the former expression turns into:

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

Empirically, there is no loss of generality if we assume that the acoustic signal representation depends only on the source string: i.e., that $P(\mathbf{x}|\mathbf{t}^m, \mathbf{s})$ is independent of $\mathbf{t}^m$. In this sense, eq. (2) can be rewritten as:

$$\widehat{\mathbf{t}^m} = \arg\max_{\mathbf{t}^m} \sum_{s} P(\mathbf{t}^m, \mathbf{s}) P(\mathbf{x}|\mathbf{s}) \tag{3}$$

Equation (3) combines a standard acoustic model, $P(\mathbf{x}|\mathbf{s})$, and a multi-target translation model, $P(\mathbf{t}^m, \mathbf{s})$, both of whom can be integrated on the fly during the searching routine. Nevertheless, the outer maximization is computationally very expensive to search for the optimal tuple of target strings $\mathbf{t}^m$ in an effective way. Thus we make use of the so called Viterbi approximation, which finds the best path.

3 Inference

Given a multilingual corpus, that is, a finite set of multilingual samples $(\mathbf{s}, \mathbf{t_1}, \ldots, \mathbf{t_m}) \in \Sigma^* \times \Delta_1^* \times \cdots \times \Delta_m^*$, where $\mathbf{t_i}$ denotes the translation of the source sentence $\mathbf{s}$ (formed by words of the input vocabulary Σ) into the i-th target language, which, in its turn, has a vocabulary Δ_i, the GIAMTI method can be outlined as follows:

1. Each multilingual sample is transformed into a single string from an *extended vocabulary* ($\Gamma \subseteq \Sigma \times \Delta_1^* \times \cdots \times \Delta_m^*$) using a *labelling function* ($\mathcal{L}^m$). This transformation searches an adequate monotonous segmentation for each of the m source-target language pairs. A monotonous segmentation copes with monotonous alignments, that is, $j < k \Rightarrow a_j < a_k$ following the notation of (Brown et al., 1993). Each source word is then joined with a target phrase of each language as the corresponding segmentation suggests. Each *extended symbol* consists of a word from the source language plus zero or more words from each target language.

2. Once the set of multilingual samples has been converted into a set of single extended strings ($\mathbf{z} \in \Gamma^*$), a stochastic regular grammar can be inferred.

3. The extended symbols associated with the transitions of the automaton are transformed into one input word and m output phrases ($w/\tilde{p}_1/\ldots/\tilde{p}_m$) by the inverse labeling function ($\mathcal{L}^{-m}$), leading to the required transducer.

In this work, the first step of the algorithm (as described above), which is the one that handles the alignment and segmentation routines, relies on statistical alignments obtained with GIZA++ (Och, 2000). The second step was implemented using our own language modeling toolkit, which learns stochastic k-testable in the string-sense grammars (Torres and Varona, 2001), and allows for back-off smoothing.

4 Experimental results

4.1 Task and corpus

We have implemented a highly practical application that could be used to translate on-line TV weather forecasts into several languages, taking the speech of the presenter as the input and producing as output text-strings, or sub-titles, in several languages. For

this purpose, we used the corpus METEUS (see Table 1) which consists of a set of trilingual sentences, in English, Spanish and Basque, as extracted from weather forecast reports that had been published on the Internet. Basque language is a minority language, spoken in a small area of Europe and also within some small American communities (such as that in Boise, Idaho). In the Basque Country it has an official status along with Spanish. However both languages differs greatly in syntax and in semantics. The differences in the size of the vocabulary (see Table 1), for instance, are due to the agglutinative nature of the Basque language.

With regard to the speech test, the input consisted of the speech signal recorded by 36 speakers, each one reading out 50 sentences from the test-set in Table 1. That is, each sentence was read out by at least three speakers. The input speech resulted in approximately 3.50 hours of audio signal. Needless to say, the application that we envisage has to be speaker-independent if it is to be realistic.

		Spanish	Basque	English
Training	Sentences	14,615		
	Different Sent.	7,225	7,523	6,634
	Words	191,156	187,462	195,627
	Vocabulary	702	1,147	498
	Average Length	13.0	12.8	13.3
Test	Different Sent.	500		
	Words	8,706	8,274	9,150
	Average Length	17.4	16.5	18.3
	Perplexity (3grams)	4.8	6.7	5.8

Table 1: Main features of the METEUS corpus.

4.2 System evaluation

The experimental setup was as follows: the multi-target SFST was learned from the training set in Table 1 using the GIAMTI algorithm described in section 1; then, the speech test was translated, and the output provided by the system in each language was compared to the corresponding reference sentence. Additionally, two mono-target SFST were inferred from the same training set with their outputs for the aforementioned test to be taken as baseline.

4.2.1 Computational cost

The expected searching time and the amount of memory that needs to be allocated for a given model are two key parameters to bear in mind in speech-input machine translation applications. These values can be objectively measured based on the size and on the average branching factor of the model displayed in Table 2.

	multi-target	mono-target	
		S2B	S2E
Nodes	52,074	35,034	20,148
Edges	163,146	115,526	69,690
Braching factor	3.30	3.13	3.46

Table 2: Features of multi-target model and the two decoupled mono-target models (one for Spanish to Basque translation, referred to as S2B, and the second for Spanish to English, S2E).

Adding the states and the edges up for the two mono-target SFSTs that take part in the decoupled architecture (see Table 2), we conclude that the decoupled model needs a total of $185,216$ edges to be allocated in memory, which represents an increment of 13% in memory-space with respect to the multi-target model.

On the other hand, the multi-target approach offers a slightly smaller branching factor than each mono-target approach. As a result, fewer paths have to be explored with the multi-target approach than with the decoupled one, which means that searching for a translation can be faster. In fact, experimental results in Table 3 show that the mono-target architecture works $\%11$ more slowly than the multi-target one.

	multi-target	mono-target		
		S2B	S2E	S2B+S2E
Time (s)	30,514	24,398	9,501	33,899

Table 3: Time needed to translate the speech-test into two languages.

Summarizing, in terms of computational cost (space and time), a multi-target SFST performs better than the mono-target decoupled system.

4.2.2 Performance

So far, the capability of the systems have been assessed in terms of time and spatial costs. However, the quality of the translations they provide is, doubtless, the most relevant evaluation criterion. In order to assess the performance of the system in a quantitative manner, the following evaluation parameters

were computed for each scenario: *bilingual evaluation under study* (BLEU), *position independent error rate* (PER) and *word error rate* (WER).

As can be derived from the Speech-input translation results shown in Table 4, slightly better results are obtained with the classical mono-target SFSTs, compared with the multi-target approach. From Spanish into English the improvement is around 3.4% but from Spanish into Basque, multi-target approach works better with an improvement of a 0.8%.

	multi-target		**mono-target**	
	S2B	**S2E**	**S2B**	**S2E**
BLEU	39.5	59.0	39.2	61.1
PER	42.2	25.3	41.5	23.6
WER	51.5	33.9	50.5	31.9

Table 4: Speech-input translation results for Spanish into Basque (S2B) and Spanish into English (S2E) using a multi-target SFST or two mono-target SFSTs.

The process of speech signal decoding is itself introducing some errors. In an attempt to measure these errors, the text transcription of the recognized input signal was extracted and compared to the input reference in terms of WER as shown in Table 5.

	multi-target	**mono-target**	
		S2B	**S2E**
WER	10.7	9.3	9.1

Table 5: Spanish speech decoding results for the multi-target SFST and the two mono target SFSTs.

5 Concluding remarks and further work

A fully embedded architecture that integrates the acoustic model into the multi-target translation model for multiple speech translation has been proposed. Due to the finite-state nature of this model, the speech translation engine is based on a Viterbi-like algorithm. The most significant feature of this approach is its ability to carry out both the recognition and the translation into multiple languages integrated in a unique model.

In contrast to the classical decoupled systems, multi-target SFSTs enable the translation from one source language simultaneously into several target languages with lower computational costs (in terms of space and time) and comparable qualitative results.

In future work we intend to make a deeper study on the performance of the multi-target system as the amount of targets increase, since the amount of parameters to be estimated also increases.

Acknowledgements

This work has been partially supported by the University of the Basque Country and by the Spanish CICYT under grants 9/UPV 00224.310-15900/2004 and TIC2003-08681-C02-02 respectively.

References

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

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

F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. García-Varea, D. Llorens, C. Martínez, S. Molau, F. Nevado, M. Pastor, D. Picó, A. Sanchis, and C. Tillmann. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language*, 18:25–47, January.

M. Teresa González and Francisco Casacuberta. 2006. Multi-Target Machine Translation using Finite-State Transducers. In *Proceedings of TC-Star Speech to Speech Translation Workshop*, pages 105–110.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer, Speech and Language*, 16(1):69–88, January.

Franz J. Och. 2000. GIZA++: Training of statistical translation models.

Fernando C.N. Pereira and Michael D. Riley. 1997. Speech Recognition by Composition of Weighted Finite Automata. In Emmanuel Roche and Yves Schabes, editors, *Finite-State Language Processing*, Language, Speech and Communication series, pages 431–453. The MIT Press, Cambridge, Massachusetts.

M. Inés Torres and Amparo Varona. 2001. k-tss language models in speech recognition systems. *Computer Speech and Language*, 15(2):127–149.