BASQUE-SPANISH-ENGLISH MACHINE TRANSLATION USING FINITE-STATE TRANSDUCERS

Alicia Pérez, M. Inés Torres, V. Guijarrubia[*]

Dep. Electricity and Electronics
Faculty of Science and Technology
University of the Basque Country
manes@we.lc.ehu.es

Francisco Casacuberta

Dep. of Information Systems and Computation
Faculty of Computer Science
Technical University of Valencia
fcn@dsic.iti.es

ABSTRACT

The goals of this work are, on the one hand to develop a trilingual corpus suitable for example-based text and speech-input machine translation between English, Spanish and Basque; and on the other hand, to compare the same translation method for under two very different pair of languages: Spanish-Basque and Spanish-English.

The stochastic finite-state transducers have been trained from raw bilingual examples. As the experimental results show, they seem to be a good choice for both text and speech-input machine translation.

1. INTRODUCTION

Two main trends can be distinguished in machine translation (MT): rule-based and statistics-based. The first one is deductive and includes linguistic knowledge, whereas the second one is built on the basis of a collection of samples, so it is inductive. Nowadays some efforts are being made in order to take advantage of both information sources. For both rule-based and statistical frameworks, finite-state transducers (FST) have proved to offer a great versatility to compose with other finite-state models.

In this paper we focus on statistical models, however, we also offer some Spanish-Basque translation results obtained with *Opentrad* [1], an open-source deep-transfer MT toolkit developed by several Universities and enterprises in a Spanish challenging project. Specifically, we make use of FST, which are also the basis of *Opentrad*.

A stochastic FST (SFST) can be built for any pair of languages, whenever a representative amount of examples is available. As it is known, inductive approaches require great collection of bilingual data, that is, the same sentences translated into the languages we want to work with [2]. However, only bilingual corpora for some of the European languages are available. In this work we present a suitable trilingual corpus, in English, Spanish and Basque, completing the work initiated in [3], and letting us start with multilingual translation involving Basque language, as well as to face with speech-input MT.

1.1. Basque language

Basque is a minority language but shares official status, along with Spanish, in the Basque Country.

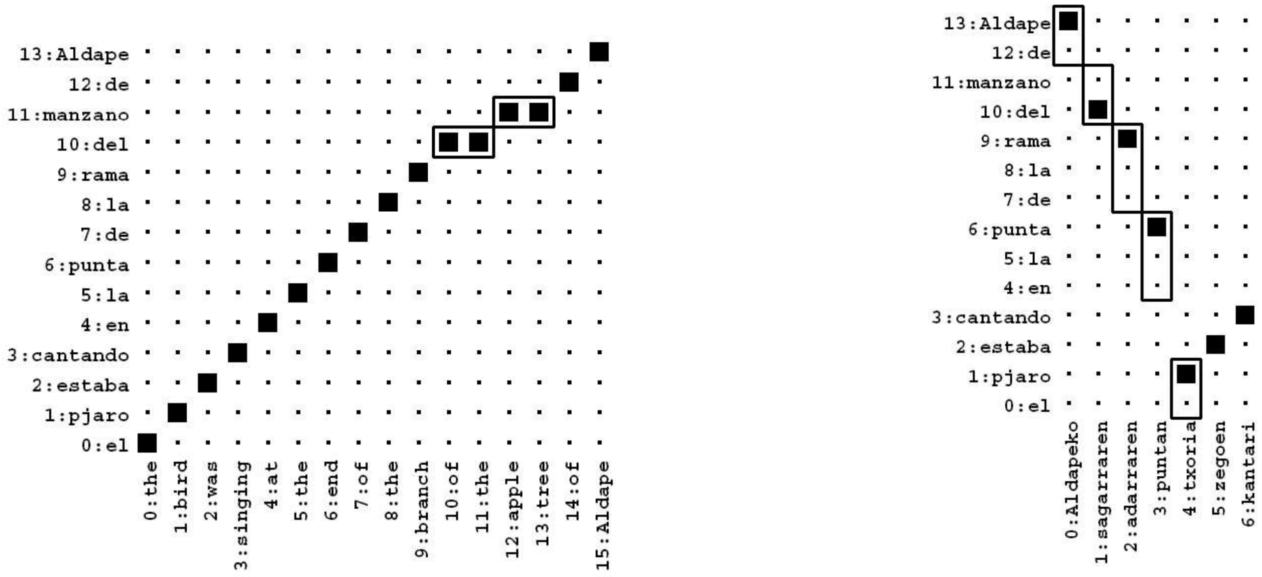
Basque is a pre-Indoeuropean language of unknown origin. Thus, with regard to etymology, Basque and Spanish are very different. On the other hand, Basque is an extremely inflected language, both in nouns and verbs. In addition, both languages present a different arrangement of the words within the sentence since, opposite to Spanish, Basque has left recursion. Figures 1(a) and 1(b) show the word arrangement through Spanish-English and Spanish-Basque alignment matrices respectively. The relationships provided by the statistical alignment model are shown by filled squares, while the linguistic ones are in hollow squares, e.g. the first Basque word is connected to the last two Spanish words, even the statistical alignment model provides only the first one.

2. STATISTICAL FRAMEWORK

Finite state transducers have proved to be useful in language processing and in automatic speech recognition systems. Recently they have also been proposed for SMT applications [4, 5]. Stochastic finite state transducers (SFST) can be automatically learnt from bilingual corpora by efficient algorithms, such as GIATI (Grammar Inference and Alignments for Transducers Inference).

The GIATI methodology is used in this work to build the translation models. Given a bilingual corpus, GIATI algorithm provides a probabilistic finite-state transducer [5].

[*] 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 TIN2005-08660-C04-03 respectively.



(a) Spanish-English

(b) Spanish-Basque

Figure 1. These alignment matrices show the word ordering for a sentence in the three languages under consideration.

GIATI algorithm works as follows:

1. Given a bilingual corpus, find a monotone segmentation, and thereby, assign an output sequence to each input word, leading to the so called *extended corpus*.

2. Infer a probabilistic finite state automaton from the extended corpus. We promote the use of a *k-testable in the strict sense* (k-TSS) language model [6], instead of n-gram models, since the k-TSS models keep the syntactic constraints of the language.

3. Split the output sequence from the input word, on each edge of the automaton, getting, in this way, the finite state transducer.

Once we have the transducer and an input sentence $\mathbf{s} \in \Sigma^+$, the translation process implies searching the most likely output string $\widehat{\mathbf{t}} \in \Delta^*$ through all the possible output strings as summarized in equation (1). Where $d(\mathbf{s},\mathbf{t})$, represents a path in the SFST, compatible both with the input sentence $\mathbf{s}$ and the output $\mathbf{t}$. Therefore, the searching criteria in the SFST deals with the joint probability of sentence pairs.

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

2.1. Two architectures for speech translation

The goal of the statistical speech translation (summarized in eq. (2)) is to find the likeliest target language string ($\mathbf{t}$) given the acoustic representation ($\mathbf{x}$) of a source language string ($\mathbf{s}$).

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

Two architectures can be used in order to build the speech translation system: the serial and the integrated one [5]. The serial architecture consists of a text to text translator after the speech decoder, whereas the integrated architecture works as a speech recognition system which makes use of a translation model instead of the usual language model. The translation model, in fact, involves two language models: the source and the target language models. Then, once the decoder has found the best path throughout the trellis, the output-language string related to the best sequence of states is provided. Both architectures are described in detail in the following subsections.

2.1.1. Serial architecture

The speech is decoded in a conventional speech recognition system. The output string provided by the speech recognizer is the input (source) string for the text to text translator (see Fig. 2). The full process can be described in two steps:

1. Word decoding of $\mathbf{x}$ (the acoustic representation):

$$\widehat{\mathbf{s}} \approx \arg\max_{\mathbf{s}} P(\mathbf{s})P(\mathbf{x}|\mathbf{s}) \quad (3)$$

where $P(\mathbf{s})$ is the source language model (k-TSS LM is used in this work).

2. Translation of $\widehat{\mathbf{s}}$ (the expected decodification of $\mathbf{x}$):

$$\widehat{\mathbf{t}} \approx \arg\max_{\mathbf{t}} P(\mathbf{t}|\widehat{\mathbf{s}}) = \arg\max_{\mathbf{t}} P(\mathbf{t},\widehat{\mathbf{s}}) \quad (4)$$

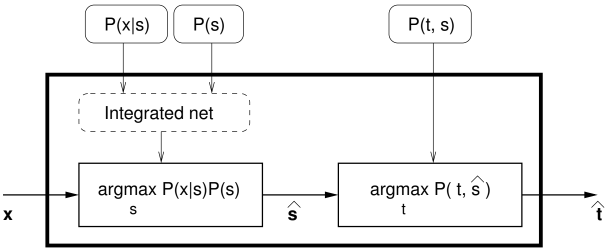


Figure 2. Serial architecture for speech translation. The output of the system is the text translation of an input speech signal. The system is supported on three knowledge sources: the acoustic model, the input language model and the translation model. The system consists of two main blocks: the speech decoder and the text to text translator. The acoustic model and the input language model are integrated into a finite state net whose output is supplied to the speech decoder to make decisions. The translator makes use of the translation model, which is also a finite state model.

In practice, $\widehat{\mathbf{s}}$ (the input sentence to the translator) is likely to be corrupted, since the speech recognition system is not ideal. Thus, we can not expect the output translation to be as close to the reference as it could be in case of a perfect input. Moreover, the weakest process is the translator, therefore we should preserve it from errors as much as possible. The recognition system alone introduces a $WER = 3.5$.

2.1.2. *Integrated architecture*

The strict way to deal with speech translation is:

$$\begin{aligned}\widehat{\mathbf{t}} &= \arg\max_{\mathbf{t}} \sum_{s} P(\mathbf{t},\mathbf{s}|\mathbf{x}) \\ &= \arg\max_{\mathbf{t}} \sum_{s} P(\mathbf{t},\mathbf{s}) P(\mathbf{x}|\mathbf{t},\mathbf{s}) \quad (5)\end{aligned}$$

Let's assume that the acoustic signal representation depends only on the source string, i.e. $P(\mathbf{x}|\mathbf{t},\mathbf{s})$ is independent of $\mathbf{t}$, then eq. (5) can be written as,

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} \sum_{s} P(\mathbf{t},\mathbf{s}) P(\mathbf{x}|\mathbf{s}) \quad (6)$$

In practice, the sum over all possible source strings can be approximated by the maximum term involved.

$$\widehat{\mathbf{t}} \approx \arg\max_{\mathbf{t}} \max_{s} P(\mathbf{t},\mathbf{s}) P(\mathbf{x}|\mathbf{s}) \quad (7)$$

In this approach, the acoustic knowledge is introduced in the whole FST. The main feature of this approach is it's ability to carry out both the recognition and the translation at the same time. The problem is solved as the equation (7) suggests, without any more assumptions. In the integrated architecture (shown in Fig. 3), the speech recognizer and the translation system have been coupled into a unique automaton, the finite-state transducer inferred by GIATI technique. The whole system works as a speech recognizer, where the output string is the text translation, instead of the text transcription, of the input speech.

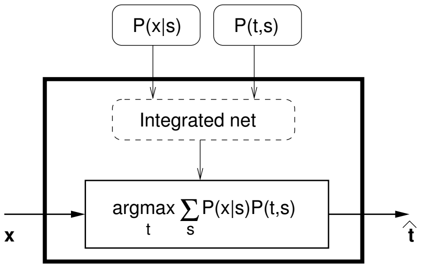


Figure 3. Integrated architecture for speech translation. The output of the system is the text translation of an input speech signal. The system is supported on two knowledge sources: the acoustic model and the translation model, both finite state models.

The strings of extended symbols are used to estimate the language model of such a recognizer, that is, the translation model of the integrated system. However, only the lexicon of the input language was transcribed into acoustic models and it was then integrated in the whole speech translation system.

The integrated automaton can be built on the basis of the text to text translator, just expanding the source word on one edge by its phonetic transcription, and each phoneme by the corresponding HMM.

3. EXPERIMENTAL RESULTS

In this section we are going to describe the corpus built during the last months, and then we show preliminary experimental results for text and speech input machine translation using the described techniques.

3.1. Task and corpus

METEUS is the weather forecast corpus composed from 28 months of daily weather forecast reports in the Spanish and Basque picked from those published in Internet. In those report pairs bilingual alignment was assured at paragraph level, but segmentation into sentences was solved by using statistical techniques, specifically *RECalign*, a greedy algorithm [7]. Afterwards each sentence was translated into English by experts, leading to a trilingual corpus aligned at sentence level.

The main features of METEUS corpus are shown in Table 1. The test set consists of 500 training independent pairs, all of them different. For speech input machine translation experiments, this test set was recorded by 36 bilingual speakers uttering 50 sentence-pairs each, resulting in around 3.25 hours of audio signal for each language. Thus, the test set is the same for text and speech translation in order to assess the model under the same conditions.

At this point, with the Basque-Spanish sentence pairs and their counterparts in English, we conclude the work [3], and we can start experimentation with example-based machine translation involving Basque language and including English for a matter of comparison with other works of reference.

		Spanish	Basque	English
Training	Sentences	14615		
	Different	7225	7523	6634
	Words	191156	187462	195627
	Vocabulary	702	1147	498
	Average Length	13.0	12.8	13.3
Test	Sentences	500		
	Words	8706	8274	9150
	Average Length	17.4	16.5	18.3
	Perplexity (3grams)	4.8	6.7	5.8

Table 1. METEUS corpus main features.

There is a great difference in terms of vocabulary size for the three languages taken into account within the same application (see Table 1). Basque language is a highly inflected language with many different running words, whereas English is the simplest one. 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 model involving the Basque language.

With regard to the practical issues, let us add that the underlaying translation model has been inferred with our own k-TSS language modeling toolkit, with Witten-Bell discount for smoothing purposes. With respect to the acoustic models, the speech signal database was parameterized into 12 Mel-frequency cepstral coefficients (MFCCs) with delta (ΔMFCC) and acceleration (Δ^2MFCC) coefficients, energy and delta-energy (E, ΔE), so four acoustic representations were defined. For the speech recognition system, a total of 24 context-independent acoustic-units were used. Each phone-like unit was modeled by a typical left to right non-skipping self-loop three-state continuous hidden Markov model, with 32 Gaussians per state and acoustic representation. To train these models, a phonetically balanced Spanish database, called Albayzin [8], was used. The decoding engine implements Viterbi algorithm with beam-search in an attempt to improve the searching time.

3.2. Text-input translation results

The SFST was learned from the training set described in Table 1. Then, the sentences from the test set were translated by that SFST. The translation given by the system was compared with the reference sentence. Two well-known evaluation measures were taken into account, WER (*Word Error Rate*) and PER (*Position-independent Error Rate*).

In Table 2 we present some preliminary translation experiments from Spanish into Basque and into English. Moreover, we compare our Spanish-Basque translation results with those provided by *Opentrad*, the first publicly available translator including this pair of languages.

Opentrad has been enriched with a bilingual vocabulary and the output has been post-processed in order to get closer results to the reference-style. Besides, all output words with an alarm symbol have been revised as a professional translator would do.

	Text-input			
	Spanish-Basque		Spanish-English	
	WER	PER	WER	PER
Transfer	84.1	75.1	–	–
Statistical	46.5	37.6	28.1	20.2

Table 2. Error rates for text-input machine translation for the 500 test sentences in Table 1 using different techniques: transfer and statistics.

As it has been mentioned at the beginning of this work, these are preliminary results, just to make measures about the recently harvested trilingual corpus. However, they have already shown the difficulty of the translation into Basque since there are much more errors under the same conditions (Table 2).

If we compare the transfer system versus the statistical one, the later seems to performs much better, anyway the difference is not that high if we compare the obtained translations. The point is that the statistical system can also learn the translation style, which is a good choice for simple tasks with restricted domain and construction.

Since only a single reference is available, when the translation given by the system does not match its expected translation or reference, it is punished even if it could be acceptable from a human point of view. Despite WER and PER are a bit pessimistic evaluation measures, at leas, they are automatic and therefore, objectives. Nevertheless, 50 randomly obtained translations were evaluated by 3 experts, within a ranking of 1-5 (1 the worst score). *Opentrad* was given 2 points by the three evaluators, and statistical FST was given 3 points by two evaluators and 4 points by the other one. Both, objective and subjective meassures show that a great effort has to be done in order to improve Spanish-Basque translation results for a real practical MT application.

Here we show Spanish into Basque translation examples obtained by the Transfer (T-sys) and Statistical (S-sys) systems for the first two test-sentences. In the first example, the statistical system (S-sys) performs well except for the last word (*da*), which should be *dira*, since *egonkor mantendu* could be taken as a synonym of *ez dira* *aldatuko*. However, the output of the transfer system (T-sys) is not well formed. In the second example, both systems offer an appropriate output even if they do not match the reference.

1. Las temperaturas máximas sin cambios o ligeramente más altas.

   ref Tenperatura maximoak ez dira aldatuko edo gutxi igoko dira.

   T-sys Aldaketarik gabe edo arinki areago garaiak tenperatura handienak.

   S-sys Tenperatura maximoak egonkor mantendu edo gutxi igoko da.

2. Brumas y bancos de niebla matinales.

   ref Goizean lanbroa eta lainoguneak izango dira.

   T-sys Lanbroak eta goizeko lainoguneak.

   S-sys Goizean lanbroa eta lainoguneak azalduko dira.

3.3. Speech-input machine translation results

Speech-input machine translation could be a cheap choice in order to translate TV weather forecasts. On Table 3 we show some translation results making use of the two architectures previously mentioned. Those architectures behave different for different language pairs. The integrated one works better for Spanish-English than for Spanish-Basque. Moreover, in terms of WER, there is not hardly any difference between text-input and speech-input machine translation for Spanish-English.

	Speech-input			
	Spanish-Basque		Spanish-English	
	WER	PER	WER	PER
Integrated	57.6	55.6	29.9	27.6
Serial	51.7	42.4	31.0	29.2

Table 3. Error rates for speech-input machine translation with both Integrated and Serial architectures.

4. CONCLUDING REMARKS AND FURTHER WORK

A great effort have been made to harvest a trilingual text and speech corpus in Basque, Spanish and English for example based machine translation. It has been successfully used for SFST training. Both text and speech-input preliminary translation results have been reported. At this point, we can compare the Basque language related translation results with other better exploited languages, such as Spanish or English.

Since multilingual corpora are too scarce, multi-target translation is still a vaguely studied field. With this trilingual corpus we aim to make an attempt in that area in future works.

Experimental results show that SFST models perform much better on similar languages and they work worse with morphologically rich languages, in special dealing with long-distance reorderings, as in the case of Spanish-Basque language pair (as shown in Figure 1). This being so, we are planing as further work, to make use of reordering models that have been proved to report good results in similar circumstances [9].

5. BIBLIOGRAPHY

[1] Antonio M. Corbí-Bellot, Mikel L. Forcada, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Iñaki Alegria, Aingeru Mayor, y Kepa Sarasola, “An open-source shallow-transfer machine translation engine for the romance languages of spain,” in *Proceedings of the Tenth Conference of the European Associtation for Machine Translation*, Budapest, Hungary, May 2005, pp. 79–86.

[2] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, y R. L. Mercer, “The mathematics of statistical machine translation: Parameter estimation,” *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[3] Alicia Pérez, Inés Torres, Francisco Casacuberta, y Víctor Guijarrubia, “A Spanish-Basque weather forecast corpus for probabilistic speech translation,” in *Proceedings of the 5th SALTMIL Workshop on Minority Languages*, Genoa, Italy, 2006.

[4] Enrique Vidal, “Finite-state speech-to-speech translation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, Apr. 1997, vol. 1, pp. 111–114.

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

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

[7] I. García-Varea, D. Ortiz, F. Nevado, P.A. Gómez, y F. Casacuberta, “Automatic segmentation of bilingual corpora: A comparison of different techniques,” in *Proceedings of the Second Iberian Conference on Pattern Recognition and Image Analysis*, vol. 3523 of *Lecture Notes in Computer Science*, pp. 614–621. Springer-Verlag, Estoril (Portugal), June 2005.

[8] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mariño, y C. Nadeu, “Albayzin speech database: Design of the phonetic corpus,” in *Proc.*

of the European Conference on Speech Communications and Technology (EUROSPEECH), Berlín, Germany, 1993.

[9] Stephan Kanthak, David Vilar, Evgeny Matusov, Richard Zens, y Hermann Ney, “Novel reordering approaches in phrase-based statistical machine translation,” in *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, Michigan, June 2005, pp. 167–174, Association for Computational Linguistics.