A Spanish-Basque weather forecast corpus for probabilistic speech translation

Alicia Pérez[1], Inés Torres[1], Francisco Casacuberta[2], Víctor Guijarrubia[1]

[1] Elektrika eta elektronika saila
Euskal Herriko Unibertsitatea
{manes@we.lc.ehu.es}
[2] Departament de Sistemes Informàtics y Computació
Universitat Politècnica de València
{fcn@dsic.upv.es}

Abstract

The main goal of this work is to develop a bilingual corpus suitable for example based machine translation between Spanish and Basque. Spanish to Basque machine translation has proved to be a difficult task for statistical machine translation. Apart from the great linguistic differences between these two languages, the low performance of the few presently available statistical machine translation systems is likely due to the lack of adequate parallel corpora. Here we present the methodology involved to pick up and process bilingual data about weather forecast reports. We also present preliminary experiments carried out for speech and text input (statistical) machine translation.

1. Introduction

In recent years, statistical methods have been successfully applied to machine translation. In this framework: stochastic translation models can be obtained for any pair of languages, whenever a representative number of examples is available. Thus, inductive approaches require a great deal of collection of bilingual data, that is, sentences from a (source) language and the corresponding translation from another (target) language (Brown et al., 1993). However, bilingual corpora for only some European languages are available.

The goals of this study are, on the one hand, to create a suitable bilingual corpus for a specific translation task from Spanish into Basque, for both text and speech-input purposes, and on the other hand to present the preliminary (speech and text) translation results using statistical machine translation (SMT) techniques. Another practical extension of the corpus is being created for translation from Basque into English.

1.1. The Basque language

Basque language is a minority but official language in the Basque Country. It is also spoken in some adjoining areas, such as Navarre, in Spain, and Atlantic Pyrenees, in France, as shown in Table 1.

Area	Total inhabitants	Basque speakers
Navarre	236,963	33.20%
Atlantic Pyrenees	515,989	10.15%
Basque Country	2,089,995	24.58%

Table 1: The percentage of Basque speakers in different areas (according to data published by the Basque Government).

Basque is a pre-Indoeuropean language of unknown origin. Thus, the etymology of words in Basque and Spanish is usually different. It also presents a different arrangement of the words within phrases, since, unlike Spanish, Basque has left recursion. These features are shown through the example in Figure 1. On the other hand, Basque is a highly inflected language, in both nouns and verbs.

1.2. The state of the art

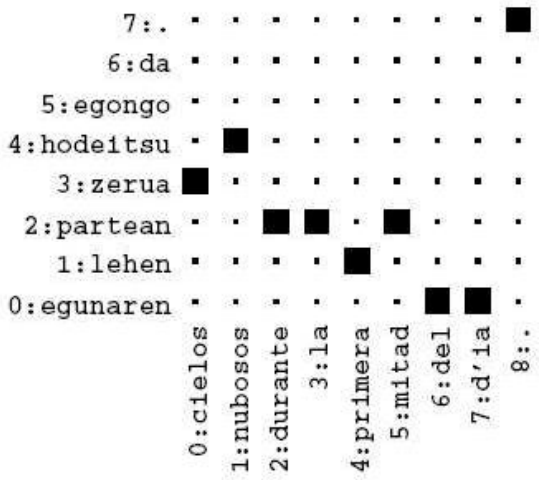
There are some papers related to transfer-based translation tools from Spanish to Basque (Alegria et al., 2005). Morpho-syntactic parsing has also been broadly studied (Arriola, 2004). However, there are few papers related to SMT between Spanish and Basque. Among them, we may highlight (Ortiz et al., 2003; González et al., 2004), where a corpus from the Basque Country's official government records was harvested. Nevertheless, low translation results were obtained. This low performance is mainly due to the scant bilingual corpora available. The lack of samples results in poorly trained statistical alignment models. We realize that those alignments were unable to capture long distance relationships (see Figure 1) between Spanish and Basque.

2. Stochastic finite-state transducers

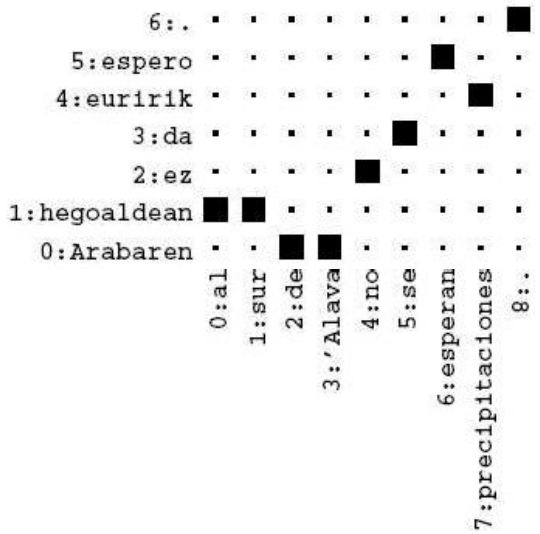
Finite state transducers have proved to be useful in language processing and in automatic speech recognition (ASR) systems. In recent years they have also been proposed for SMT applications (Vidal, 1997; Casacuberta and Vidal, 2004). Moreover, FST can be easily integrated into an ASR system for speech translation application (Casacuberta et al., 2004).

Stochastic finite-state transducers (SFST) can be automatically learned from bilingual corpora by efficient grammar inference algorithms, such as GIATI (Grammar Inference and Alignments for Transducers Inference). Given a bilingual corpus, the GIATI algorithm provides a probabilistic finite-state transducer (Casacuberta and Vidal, 2004). This algorithm works as follows:

1. Given a bilingual corpus, find a monotone segmentation, and thereby, assign an output sequence to each



(a)



(b)

Figure 1: These two alignment matrices, extracted from the corpus, show the natural relationships between the words of a sentence in Spanish and their counterparts in Basque. Those relationships are not monotone.

input word, leading to the so called *extended corpus*.

2. Infer a probabilistic finite state automaton from the extended corpus. In this paper we propose the use of a *k-testable in the strict sense* (k-TSS) language model (Torres and Varona, 2001), rather than an n-gram model, since 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 to obtain the finite state transducer.

The *stochastic translation* $\widehat{t} \in \Delta^*$, of an input sequence $\mathbf{s} \in \Sigma^+$, is the string which maximizes the joint probability described by the equation (1):

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

where, $d(\mathbf{s},\mathbf{t})$, is a path in the SFST that deals with $\mathbf{s}$ and produces $\mathbf{t}$.

The resolution of the eq. (1) has proved to be a hard computational problem (Casacuberta and de la Higuera, 2000), but it can be efficiently computed by the *maximum approximation*, which replaces the sum by the maximum:

$$\widehat{t} \approx \arg\max_{\mathbf{t}} \max_{d(\mathbf{s},\mathbf{t})} P(d(\mathbf{s},\mathbf{t})) \quad (2)$$

3. A weather forecast corpus

METEUS is the weather forecast corpus that we present here. It was composed from 28 months of daily weather forecast reports in the Spanish and Basque languages. These reports were picked from those published in Internet by the Basque Institute of Meteorology[1]. We obtained a first bilingual corpus where each report in Spanish was the translation of a report in Basque. Thus, *bilingual alignment* was assured at paragraph level.

At the end of the first corpus acquisition, there were 3,865 paragraph pairs, consisting of many sentences, with around 54 words per paragraph on average. Segmentation into sentences was solved by using statistical techniques, specifically *RECalign*, a greedy algorithm (Nevado and Casacuberta, 2004; García-Varea et al., 2005). A hundred paragraphs, randomly chosen, were checked and validated by experts, which assures the success of the algorithm.

After the segmentation process, the corpus was divided into training and testing sets (Table 2). Notice that the Basque language vocabulary size for this task is 1.6 times higher than the Spanish one. This is not unusual given the Basque language inflection mentioned in section 1.1.. In order to deal with this problem, a morphological analysis was carried out. As a result, we can work with word-forms or stems. The vocabulary size, in terms of stems, has been decreased to 462 units in Spanish and 578 in Basque.

The text-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.

4. Experimental results

In this section we present a preliminary evaluation of this corpus. We learned an SFST from the training set, and the test set was translated with the inferred models. Finally, the translations provided by the system were compared to the reference sentences.

Apart from text-to-text translation, we performed speech-input translation. There are many ways of building a speech input translation system, see (Vidal, 1997; Casacuberta et al., 2004). For these preliminary experiments, we simply chose the so called serial architecture, which consists of two steps. In the first the speech signal is decoded into a source

[1] http://www.euskalmet.net

		Spanish	Basque
Training	Pair of sentences	14,615	
	Different pairs	8,462	
	Different sentences	7,226	7,523
	Words	191,156	187,462
	Vocabulary	720	1147
	Average length	13.0	12.8
Test	Pair of sentences	500	
	Different sentences	500	500
	Words	8,706	8,274
	Average length	17.4	16.5
	Perplexity (3-grams)	4.8	6.7

Table 2: Features of the training and test sets.

sentence. In the second the decoded sentence is translated into a target sentence.

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 (Rodríguez and Torres, 2003). 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 (Moreno et al., 1993), was used. With regard to the language model, a 3-TSS was used, learned using the training corpus in Table 2.

In Table 3 we summarize some text and speech-input translation results for Spanish-to-Basque SMT. We have dealt with the following automatic evaluation measures:

WER: *Word Error Rate* is the string edit distance between the reference sentence and the system's output.

PER: *Position independent Error Rate* is similar to the WER but without taking into account the words-order inside the sentence.

BLEU: *BiLingual Evaluation Understudy* is based on the n-grams of the hypothesized translation that occur in the reference translations. The BLEU metric ranges from 0.0 (worst score) to 1.0 (best score) (Papineni et al., 2002).

	WER	PER	BLEU
Text input	46.5	37.6	0.46
Speech input	51.7	42.4	0.40

Table 3: Spanish to Basque speech-input and text-input machine translation scores.

Notice that the test set is completely independent of the training set (see Table 2). Therefore, our system is subjected to the worse case that could appear in practice.

4.1. Examples

Some translation examples are shown below, where **input** means "the input sentence", **reference** "the reference", and **system** "the system's output". Taking into account that an input sentence can be translated in more than one way, even though the output does not exactly match the reference, it could be translated correctly. Since we only have a single reference for each input, some outputs are unfairly penalized, therefore the reported error rates are quite pessimistic. In the following examples we emphasize the real errors in italics (in the third example the errors are related to the absence of some words).

- Test sentence #1.

  input las temperaturas máximas sin cambios o ligeramente más altas .

  reference tenperatura maximoak ez dira aldatuko edo gutxi igoko dira .

  system tenperatura maximoak egonkor mantendu edo gutxi igoko *da* .

- Test sentence #2.

  input por la tarde - noche cielos muy nubosos con precipitaciones débiles .

  reference arratsalde - gau aldera zerua oso hodeitsu egongo da eta euri arina egingo du .

  system arratsalde - gauean zerua oso hodeitsu egongo da eta euri arina egingo du .

- Test sentence #3.

  input por la tarde , los vientos girarán a componente sur , flojos a moderados , con intervalos fuertes en el litoral y zonas de montaña .

  reference arratsaldean , haizeak hegoaldera egin eta ahul - bizia ibiliko da , tarte gogorrekin kostaldean eta mendi inguruetan .

  system haizeak hegoaldera egin eta hegoaldeko haize ahul - bizia ibiliko da , kostaldean eta mendi inguruetan .

5. Concluding remarks and further work

The overall result of this work is a Spanish-Basque text and speech corpus, appropriate for building example based translation tools. It has been obtained using statistical techniques, which seem to be suitable from the practical point of view. Apart from that it has been enriched with morphological information. Professional translators are now translating this corpus into English, and it is also being recorded in the way mentioned in Section 3. This will lead us to a trilingual corpus in English-Spanish-Basque, along with speech resources. From this point onwards, we will complete the current work, evaluating the selected translation method under the same conditions for different language pairs. Due to the difficulties of the translation from Spanish to Basque, morpho-syntactic and phrase-based tags are being added to the corpus that will help the construction

of SMT systems. Further work is needed in order to improve the statistical translation models used in this work. We suggest taking advantage of linguistic information such as word stems and declinations to enrich the corpus and, at the same time, to get more accurate translation models

6. Acknowledgements

This work has been partially supported by the Industry Department of the Basque Government and by The University of the Basque Country under grants INTEK CN02AD02 and 9/UPV 00224.310-15900/2004 respectively.
We would like to thank the *Ametzagaiña* group (`http://www.ametza.com`), and Josu Landa, in particular, for providing us with the morpho-syntactic tags.

7. References

I. Alegria, A. Diaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor, K. Sarasola, M. L. Forcada, S. Ortiz-Rojas, and L. Padr. 2005. An open architecture for transfer-based machine translation between spanish and basque. In *Proceedings of OSMaTran: Open-Source Machine Translation, A workshop at Machine Translation Summit X (Phuket, Thailand, September 12–16)*.

A. Aranzabe Arriola. 2004. A cascaded syntactic analyser for basque.

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.

F. Casacuberta and C. de la Higuera. 2000. Computational complexity of problems on probabilistic grammars and transducers. In Arlindo L. Oliveira, editor, *ICGI*, volume 1891 of *Lecture Notes in Computer Science*, pages 15–24. Springer-Verlag.

F. Casacuberta and E. 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.

I. García-Varea, D. Ortiz, F. Nevado, P. A. Gómez, and F. Casacuberta. 2005. Automatic segmentation of bilingual corpora: A comparison of different techniques. In *Iberian Conference on Pattern Recognition and Image Analysis*, volume 3523 of *Lecture Notes in Computer Science*, pages 614–621. Springer-Verlag, Estoril (Portugal), June.

J. González, D. Ortiz, J. Tomás, and F. Casacuberta. 2004. A comparison of different statistical machine translation approaches for spanish-to-basque translation. In *Actas de las III Jornadas de Tecnología del Habla*, Valencia, Spain, November.

A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mario, and C. Nadeu. 1993. Albayzin speech database: Design of the phonetic corpus. In *Proc. of the European Conference on Speech Communications and Technology (EUROSPEECH)*, Berlín, Germany.

F. Nevado and F. Casacuberta. 2004. Bilingual corpora segmentation using bilingual recursive alignments. In *Actas de las III Jornadas en Tecnologías del Habla*, Valencia, Spain, November.

D. Ortiz, I. García-Varea, F. Casacuberta, A. Lagarda, and J. González. 2003. On the use of statistical machine translation techniques within a memory-based translation system (AMETRA). In *Proc. of Machine Translation Summit IX*, pages 115–120, New Orleans, USA, September.

K. Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association Computational Linguistics (ACL)*, pages 311–318, Philadelphia, July.

L. J. Rodríguez and I. Torres. 2003. Comparative study of the baum-welch and viterbi training algorithms applied to read and spontaneous speech recognition. In *1st Iberian Conference on Pattern Recognition and Image Analysis*, Puerto Andratx (Mallorca), Spain.

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

E. Vidal. 1997. Finite-state speech-to-speech translation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 111–114, Munich, Germany, April.