The TALP on-line Spanish-Catalan machine-translation system

Marc Poch, Mireia Farrús, Marta R. Costa-jussà, José B. Mariño, Adolfo Hernández, Carlos Henríquez, José A. R. Fonollosa

Center for Language and Speech Technologies and Applications (TALP), Technical University of Catalonia (UPC), Barcelona, Spain

{mpoch, mfarrus, mruiz, canton, adolfohh, carloshq, adrian}@gps.tsc.upc.edu

Abstract

In this paper the statistical machine translator (SMT) between Catalan and Spanish developed at the TALP research center (UPC) and its web demonstration are described.

1. Introduction

Statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The statistical approach contrasts with the rule-based approaches to machine translation as well as with example-based machine translation. During the last three years we have developed new SMT architectures of general interest as well as specific adaptations and modules for the Spanish-Catalan pair.

2. System description

The TALP translation system is based on an N-gram translation model integrated in an optimized log-linear combination of additional features improved by specic techniques based on the use of grammatical categories, lexical categorisation and text processing, for the enhancement of the nal translation [1].
The translator is an N-gram-based SMT system. Such an approach is faced using a general maximum entropy approach in which a log-linear combination of multiple feature functions is implemented. This approach leads to maximising a linear combination of feature functions:

$$\tilde{t} = \underset{t}{argmax} \left\{ \sum_{m=1}^{M} \lambda_m h_m(t,s) \right\} \quad (1)$$

where the argmax operation denotes the search problem, i.e. the generation of the output sentence in the target language, hm (t, s) are the feature functions and m are their corresponding weights. The main feature function is the Ngram-based translation model which is trained on bilingual n-grams. This model constitutes a language model of a particular bi-language composed of bilingual units (translation units) which are referred to as tuples. In this way, the translation model probabilities at the sentence level are approximated by using n-grams of tuples such as described by the following equation:

$$\tilde{t}_1^I = \underset{t_1^I}{argmax}\{p(s_1^J, t_1^J)\} = K = \quad (2)$$

$$= \underset{t_1^I}{argmax} \prod_{n=1}^{N} p((s,t)_n | (s,t)_{n-x+1}, K, (s,t)_{n-1})$$

where the *n-th* tuple of a sentence pair is referred to as *(s,t)n* . The system is trained with the aligned Spanish-Catalan parallel corpus taken from El Periódico newspaper, which contains 1.7 million sentences. To improve it several techniques based on the use of grammatical categories, lexical categorisation and text processing are used. Most of these techniques are based on preprocessing the text that is used as input data for the baseline system and postprocessing the translation. Table1 shows the improvement achieved in terms of BLEU after using the mentioned techniques in 2000 sentences from El Periódico.

	es2ca	ca2es
N-II	83.91	83.23

Table 1: BLEU results in both directions of translation.

3. Demonstration

The demonstration of the system consists of a website that allows the user to execute translations between Catalan and Spanish pair of languages. As shown in Figure Fig. 1 the user can type text directly or send a text file to be translated in both directions. The web has an online spell checker for both Spanish and Catalan to avoid typical mistakes that would end in a bad translation.
There is an option called Log that allows the user to see data being preprocessed and postprocessed before and after being translated by the machine translator. This is very useful to see the changes that improvement techniques have on the data. The demonstration can be found online at “http://www.n-ii.org”.

Figure 1: *Capture image of the demonstration.*

4. References

[1] Mireia Farrús, Marta R. Costa-jussà, Marc Poch, Adolfo Hernández, and José B. Mariño, “Improving a Catalan-Spanish Statistical Translation System using Morphosyntactic Knowledge”, EAMT Barcelona 2009.