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 specific techniques based on the use of grammatical categories, lexical categorisation and text processing, for the enhancement of the nal translation [1]. 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. Table 1 shows the improvement achieved in terms of BLEU after using the mentioned techniques in 2000 sentences from El Periódico.

<table>
<thead>
<tr>
<th></th>
<th>es2ca</th>
<th>ca2es</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-II</td>
<td>83.91</td>
<td>83.23</td>
</tr>
</tbody>
</table>

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 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”.

4. References