Sentence Alignment for Spanish-Basque Bitexts: Word Correspondences vs. Markup similarity

Arantza Casillas[1], Idoia Fernández[1], and Raquel Martínez[2]

[1] Dpt. de Electricidad y Electrónica, Facultad de C. y Tecnología
Universidad del País Vasco
arantza@we.lc.ehu.es webfeani@lc.ehu.es
[2] Escuela Superior de CC. Experimentales y Tecnología
Universidad Rey Juan Carlos
r.martinez@escet.urjc.es

Abstract. In this paper, we present an evaluation of two different sentence alignment techniques. One is the well-known SIMR algorithm based on word correspondences on both sides of a bitext. The other one is the ALINOR algorithm, which is based on the similarity of the markup on both sides of a bitext. Both algorithms are accurate in 1-1 alignment, but ALINOR works slightly better in the case of N-M alignment.

1 Introduction

Corpora containing bilingual versions of the same text entity (bitext) are a very useful source of data. The bitext increases its value by obtaining aligned pairs of source and target language sentences. These aligned sentences are immensely valuable for different Natural Language Processing applications, such as example and memory based machine translation, multilingual information retrieval, and bilingual terminology extraction.

On the one hand, manual alignment is a very expensive process. On the other hand, automatic alignment can be obtained given limited time and resources. Several automatic techniques have been presented in the relevant literature: length-based statistical approach, pattern recognition, lexical approach, and combinations of the foregoing are the main ones. The ideal sentence alignment algorithm should be accurate and independent of the language pair.

In this work, we compare our sentence alignment algorithm ALINOR (presented in [Martinez 1998a] and [Martinez 1998b]) with the well-known Smooth Injective Map Recognizer (SIMR) algorithm of I. Dan Melamed ([Melamed 1996], [Melamed 1997]). ALINOR is based on evaluating the similarity of the linguistic and extra-linguistic markup on both sides of a bitext. It is quite language independent and obtains good accuracy rates without extra bilingual knowledge. SIMR generates word correspondences relying on cognates and, in addition, it can use a translation lexicon. If sentence boundary information is provided to the algorithm, the output corresponds to sentence alignment.

In this paper, we present the results of the evaluation of the two algorithms running on a Spanish-Basque parallel corpus. Section 2 briefly describes the