On the Use of Word Alignments to Enhance Bitext Compression*

Miguel A. Martínez-Prieto,† Joaquín Adiego,† Felipe Sánchez-Martínez,‡
Pablo de la Fuente† and Rafael C. Carrasco‡

†Depto. de Informática, Universidad de Valladolid, Spain.
{migumar2,jadiego,pfuente}@infor.uva.es
‡Dept. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain.
{fsanchez,carrasco}@dlsi.ua.es

The amount of information that is stored in digital form in more than one language is growing very fast as a consequence of the globalization. Furthermore, there are countries and supra-national entities whose legislation enforces the translation (and storage) of all the official texts into all their official languages.

Two texts that are mutual translations are usually referred to as a *bilingual parallel corpus* or, in short, as a *bitext*. Compressing independently the two texts of a bitext is far form efficient, since the information conveyed by both texts, the meaning, is similar. We take advantage of this fact to devise a bitext compression algorithm that compresses and stores the two texts that form a bitext simultaneously.

In our approach, a single model is used to represent both bitext components. For this purpose, we define a *biword* as a pair made of two words, each one from a different text, that are mutual translations in the bitext. This new concept allows one to represent with a single symbol two words with high mutual information.

The algorithm consists of a simple *processing pipeline* with two stages. The first one (*preprocessing*) performs a text (sentence and word) *alignment* in which no pre-existing resources are used; it takes as input the bitext and outputs its biword-based representation. The second stage (*compression*) implements a customization of the `mPPM` model in which biwords are used as symbols and a limited length dictionary is used to obtain two-byte codewords.

We carried out an exhaustive experimentation with seven language pairs (`es-ca`, `es-gl`, `es-en`, `es-fr`, `es-it`, `es-pt` and `fr-en`) and bitexts of different sizes. For comparison purposes we compressed the concatenation of the two texts of each bitext using general-purpose compressors. Large improvements are obtained for bitexts of closely-related languages (up to 82.73% for `es-gl`) due to the monotonicity of the word alignments obtained in the preprocessing stage. For less-related language pairs smaller improvements, between 11.12% (`es-pt`) and 2.48% (`fr-en`), are achieved because some non-monotonic word alignments are discarded to ensure that the original bitext can be fully recovered when decompressing. Multiword units may be used in the biwords in order to reduce the number of discarded word alignments.

*Funded by Spanish projects TIN2006-15071-C03-01, TIN2006-15071-C03-02 and VA012B08. Miguel A. Martínez-Prieto is granted by JCyL and ESF. We thank Mikel L. Forcada for inspiration.