Modelling Parallel Texts for Boosting Compression*

Joaquín Adiego,† Miguel A. Martínez-Prieto,†
Javier E. Hoyos-Torío† and Felipe Sánchez-Martínez‡

†Dept. de Informática, Universidad de Valladolid, Spain.
{jadiego,migumar2}@infor.uva.es, javierht@gmail.com

‡Dept. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain.
fsanchez@dlsi.ua.es

Bilingual parallel corpora, also known as *bitexts*, convey the same information in two different languages. This implies that to model a bitext we can take advantage of the translation relationship that exists between the two texts; the text alignment task makes it possible to establish such a translation relationship. A *biword* is defined as a pair of words, each from a different text, that are mutual translations in the bitext; the use of biwords allows both texts in the bitext to be represented on a single model. Several biword-based schemes have been proposed leading to good compression ratios [1].

Bearing in mind Melamed's [2] affirmation which states that "the translation of a text into another language can be viewed as a detailed annotation of what that text means", we propose a new model for bitexts in agreement with this affirmation, dubbed MAR_1. The idea is to represent the words in the right text with respect to the preceding word in the left text; thus, a first-order model based on alignment relationships is proposed. While the gain due to this idea is clear when it is used as a preprocessing step to another compressor, the price is that we have to encode two symbols (left and right words) instead of just one (biword). As a previous step to another compressor, coding biwords with a single symbol may be a good idea; however, the greater size of the dictionary and the redundancy loss in the encoded stream may be a handicap.

Empirical results show that MAR_1 approximately encodes twice the symbols that a biword-based model; however, compression is better for small bitexts and for bitexts consisting of closely-related language texts. The results also show that when encoded models are used as compression boosters we achieve compression ratios improving state-of-the-art compressors up to 6.5 percentage points which are up to 40% faster.

References

[1] J. Adiego, N. R. Brisaboa, M. A. Martínez-Prieto, and F. Sánchez-Martínez. A two-level structure for compressing aligned bitexts. In *Proceedings of the 16th International Symposium on String Processing and Information Retrieval*, pages 114–121, Saariselkä, Finland, 2009.

[2] I. D. Melamed. *Emplirical methods for exploting parallel texts.* MIT Press, 2001.

*Work supported by Spanish projects TIN2009-14009-C02-01 and TIN2009-14009-C02-02. Miguel A. Martínez-Prieto is granted by JCyL and ESF. We thank Gonzalo Navarro for ideas and inspiration.