On the Use of Words Alignments to Enhance Bitext Compression  
Miguel A. Martínez-Prieto, Joaquín Adiego, Felipe Sánchez-Martínez, Pablo de la Fuente and Rafael C. Carrasco  
19th Data Compression Conference. Snowbird, Utah, USA. March 16-18, 2009

Multilingual Parallel Corpora  
Represents the same information in different languages, i.e.:  
- Official texts of the European Union.  
- Daily newspapers in Spain.  
- Information on Internet.

Bitexts: Bilingual Parallel Corpora  
Two texts that are mutual translations.

Bitexts Compression  
Bitexts show two different representations of the same information. A shared representation of both texts should improve compression.

Pipeline Processing for Bitext Compression  
Alignment Stage  
Obtains a biword representation of the bitexts.  
A biword is a pair made of two words, each one from a different text or empty, that are mutual translations in the bitext.  
We use GIZA++ to compute the word alignments from which biwords are generated.

Compression Stage  
Mapping biwords on a limited-size dictionary: mppm variation. Features:  
- Each biword is encoded using a 2-bytes code.  
- When dictionary is full, a LRU policy is applied in order to replace biwords.  
- Each 2-bytes code is compressed using PPMDi.  
- Boosting on PPM.

Comparison and compression ratios achieved with different bitext collections

Dictionary & output sizes evolution

Comparison of bitexts compression ratios achieved with different bitext collections

Dictionary & output sizes evolution

Dictionary & output sizes evolution