A sentence compression module for machine-assisted subtitling

Nadjet Bouayad-Agha, Angel Gil,
Oriol Valentin, and Victor Pascual

Universitat Pompeu Fabra,
Barcelona, Spain
nadjet.bouayad@upf.edu, firstname.lastname@upf.edu

Abstract. We present in this paper a sentence compression module used in a machine-assisted subtitling application developed in the European e-content project e-title. Our approach to compression and the architecture of the system are motivated by the commercial and multilingual nature of the project, that is, the need to output reasonable compressions and the ability to add new strategies, genres and languages easily. The compression module currently works for the Catalan and English languages and uses the Constraint Grammar engine for linguistic preprocessing and for the linguistically motivated compression rules, thus providing a homogenous format throughout the compression process. The compression rules were implemented based on a corpus of automatically aligned <script,subtitle> pairs of films for both languages. We performed for both languages an automatic quantitative evaluation of the compression using the aligned corpus and a qualitative manual evaluation of grammaticality and informativeness.

1 Motivation

We present in this paper a sentence compression module used in a machine-assisted subtitling application developed in the European e-content project e-title.[1] This application integrates speech-text synchronisation, machine translation and sentence compression to assist subtitlers in the different stages of the subtitling process. Our approach to compression and the architecture of the system are motivated by the commercial and multilingual nature of the project, that is, the need to output reasonable compressions and the ability to add new strategies, genres and languages easily. We surveyed various approaches to sentence compression developed in Natural Language Processing, for example (Jing, 2000; Zechner, 2001; Hori and Furui, 2002, Knight and Marcu, 2002, Vandeghinste and Pan 2004), and compiled the following list of desiderata for our system:

Grammaticality: the compression module should preserve grammaticality. We found that some approaches guarantee the grammaticality of the output but at the cost of heavier linguistic machinery such as full parsing and subcategorization information (Jing,2000).

[1] EDC22160. Jan.2004–Jan.2006.