Bruno Cartoni (University of Geneva, Switzerland), Thomas Meyer (Idiap Research Institute, Martigny, Switzerland)

Building "directional corpora" for unbiased contrastive analysis

Large multilingual parallel corpora are easily available and vastly used in Statistical Machine Translation (SMT) and can also constitute an interesting field of investigation for empirical contrastive studies (i.e. the systematic analysis of linguistic phenomena in two (or more) languages in order to highlight differences and similarities see (Granger 2003) for an overview on corpus and contrastive analysis).

However, in such corpora, language information (i.e. setting the original language in which the text has actually been written) is scarcely provided. A few researches do take into account the translation direction of a parallel corpus and analyses the influences it can have. E.g. (Ozdowska 2009) discusses the implications of translation directions used in training language and translation models for SMT. In the field of contrastive analysis, recent studies on large corpora tend to incorporate the directionality of a corpus (like in Johansson 2006), revealing sometimes important discrepancies between analyses performed on translated or original text (and their counterpart) like in (Degand 2005). Making use of multilingual parallel corpora in linguistic investigation would consequently require methodological precaution and some preprocessing.

In this work, we introduce the notion of directional corpora, as parallel corpora where the source language (i.e. the language in which the text and/or speech has been produced) is clearly identified. We present an experiment that has been performed to extract directional corpora out of an existing parallel corpus (namely Europarl (Koehn 2005)). This specific multidirectional parallel corpus contains scarce information about the original language in which each statement was made, and simple extraction of existing language tags would gather only a small amount of directional data. The scarcity of the language information is uneven within the language pairs, so we automatically gathered all the tag information in all the file sets, mutually 'correcting' all the tags and discarding diverging information.

Doing so, we significantly increased the unidirectional extraction in terms of number of words (e.g. from an English to French directional corpus of 5,609,994 English token, we result with a new directional corpus of 6'358'597 English token).

Extraction and correction techniques will be presented, together with experiments on specific linguistic phenomena (namely discourse markers) that have been performed on the extended directional corpora extracted in this study, which shows interesting discrepancies in the results in translated languages and in original languages. Further methodological issues are also addressed, such as the “translational origin” of the translated data: In multilingual corpora such as Europarl, while source language is clearly identified as the “original” and target language as the “translated”, there is no evidence that the target language has been directly translated from the source language, or through a pivot language. This aspect would require other methodological precautions.

References:

Degand, Liesbeth. (2005), De l'analyse contrastive Ã la traduction: le cas de la paire puisqueaangezien, in Geoffrey Williams, ed., La linguistique de corpus, Presses universitaires de Rennes.

Granger Sylviane. (2003) ‘The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies?’ In S. Granger, J. Lerot & S. Petch-Tyson (eds) Corpus-based Approaches to Contrastive Linguistics and Translation Studies. Rodopi: Amsterdam & New York, 17-29.

Johansson, Stig. (2006), ‘How well can well be translated? On the English discourse particle well and its correspondances in Norwegian and German’, in Karin Aijmer & Anne-Marie Simon-Vandenbergen, ed., Pragmatic Markers in Contrast, Elsevier

Koehn Philip. Europarl: A Parallel Corpus for Statistical Machine Translation, MT Summit 2005.

Ozdowska, Silvia. (2009). Donnes bilingues pour la TAS francais-anglais : impact de la langue source et direction de traduction originales sur la qualite de la traduction. Actes de la conference Traitement Automatique des Langues Naturelles, TALN'09, Senlis, France, juin 2009.