SHUYUAN CAO, IRIA DA CUNHA, MIKEL IRUSKIETA
Pompeu Fabra University, National Distance Education University, University of the Basque Country, Spain

A Spanish-Chinese parallel corpus for natural language processing purposes

"Corpus-based discourse analysis for Natural Language Processing (NLP) is becoming more and more popular. In several NLP domains discourse information is being used, such as automatic summarization, machine translation (MT) and machine translation evaluation, information extraction, etc. Therefore, the aim of this work is to introduce a new Spanish-Chinese parallel corpus annotated with discourse information, which could help for the development of NLP applications concerning this language pair, such as MT or bilingual summarization.

The theoretical framework of this work is the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). RST addresses coherence by means of relations between text spans in a tree-like structure. Texts are segmented into discourse units and coherence relations are established between these discourse units. As these relations are recursive, they form text structures linking units or groups of units (spans). The units can be Nuclei or Satellites; satellites offer additional thematic information about nuclei. Relations can be nucleus-satellite (e.g. ANTITHESIS, CAUSE and EVIDENCE) and multinuclear (e.g. CONJUNCTION, JOINT, and SEQUENCE). This parallel corpus has been annotated manually with RSTTool (O'Donnell, 2000) and follows the annotation method by Iruskieta, da Cunha and Taboada (2015), in order to annotate and create a gold standard corpus under RST. Three kinds of information for Spanish and Chinese in the corpus will be provided to NLP researchers: (a) discourse segmentation, (b) central unit (CU) and (c) discourse relations.

Firstly, segmentation is the first step of discourse analysis and is useful for different NLP tasks, such as, evaluation of automatic segmentation systems, development of discourse parsers and text summarization. In our corpus, the users can compare the discourse similarities and differences between the parallel Spanish-Chinese texts by counting the quantity of EDUs, the number of discourse markers (DM), the position of DMs, etc.

Secondly, the CU is the key information or the main topic of a text, which can be applied to automatic summarization, development of intelligent systems and sentiment analysis. Genre, domain and discourse structure determine the position of the CU in a text; thus, by consulting the CU of the texts in the corpus, users can know how to organize the information of texts in different genres and domains. A good translation of the main topic or CU is also fundamental for a MT system.

Thirdly, discourse relations show the coherence of a language and are useful for several NLP tasks, such as, discourse parsing, information extraction, automatic summarization and evaluation of MT. Discourse structure can be used to detect discourse similarities and differences between Spanish-Chinese regarding discourse relations, type of relations and used translation strategies, among other issues."