Abs-170

Amal Alsaif, Katja Markert

# Annotating Discourse Connectives in MSA: Disagreement Cases in the LADTB

Discourse relations such as CAUSAL or CONTRAST relations between textual units play an important role in producing a coherent discourse. They are widely studied in theoretical linguistics (Halliday and Hasan, 1976; Hobbs, 1985), where also different relation taxonomies have been derived (Hobbs, 1985; Knott and Sanders, 1998; Mann and Thompson, 1988; Marcu,2000). Discourse relations can be signalled by explicit lexical indicators, so-called discourse connectives (Marcu, 2000; Webber et al., 1999; Prasad et al..2008a). Our study is based on Leeds Arabic Discourse Treebank â€“ LADTB- a recent annotation effort of discourse connectives of MSA (Alsaif and Markert,2010). It provides a new annotation layer above the existing layers of annotation (syntax and morphology) in the Arabic news corpus Penn ATB, Part1 (Maamouri and Bies, 2004) by annotating all discourse connectives, the relations they signal and the two arguments they relate.

In the first such study for Arabic, 107 potential discourse connectives and 18 discourse relations were analyzed following similar annotation principles of Penn DTB project for English (Prasad et al..2008a); taking into account properties specific to Arabic. In particular, we deal with the fact that Arabic has a rich morphology: we therefore include clitics, prepositions and nouns as connectives as well as a wide range of nominalizations as potential arguments. A dedicated discourse annotation tool is developed for Arabic which is based on plain text; for unrestricted discourse annotation. Both the human identification of discourse connectives and the determination of the discourse relations they convey are reliable (Alsaif and Markert,2010). We measure also the inter-annotator agreement of identifying the text spans of the arguments individually in different ways (i) the exact match of textual units in the argument text (ii) the average of overlapping syntactic tree nodes of the two arguments and (iii) the match of syntactic heads of arguments. We show that although there is no high agreement on the exact textual units, annotators reliably agree on the syntactic heads a part from disagreements for some connectives at beginning of paragraphs. The syntactic head seem to be in the most cases the text expressing the core proposition in the discourse.

We report the disagreement and ambiguity cases in our human annotation in terms of identifying (i) discourse connectives,(ii) relations and (iii) related arguments. Our results show that Arabic has a higher ambiguity than in English; connectives in PDTB are almost unambiguous a part from few discourse connectives such as since, while. Moreover, Arabic discourse tends to use longer and more complex sentences with many complements than in English. Thus annotators have disagreed in the definite boundaries of the arguments. In addition, there is a common usage of a coordinating conjunction wa/and at beginning of each paragraph if not every sentences, particularly in the news writing, without any specific discourse function rather than conjunction. Defining variant disagreement cases would help in understanding the language features in a comparative study with other languages and improving further annotation studies.

## References

A. Al-Saif, K. Markert.2010. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In International Conference on Language Resources and Evaluation (LREC 2010), Malta.

M.A.K. Halliday and R. Hasan. 1976. Cohesion in English. Longman London.

J.R. Hobbs. 1985. On the coherence and structure of discourse. Center for the Study of Language and Information, Stanford, Calif.

E.H. Hovy. 1993. Automated discourse generation using discourse structure relations. Artificial intelligence, 63(1-2):341â€“385.

A. Knott and T. Sanders. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. Journal of Pragmatics, 30(2):135â€“175.

M. Maamouri and A. Bies. 2004. Developing an Arabic treebank: Methods, guidelines, procedures, and tools. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (COLING), Geneva.

W.C. Mann and S.A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. Text, 8(3):243â€“281

D. Marcu. 2000. The theory and practice of discourse parsing and summarization. MIT Press.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008a. The

Penn discourse treebank 2.0. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).