

Integrating Logical Document Structure as a Cue for Antecedent Detection

Daniela Goecke
Universität Bielefeld
daniela.goecke@uni-bielefeld.de

15. November 2007

Abstract for *Processing Text-Technological Resources*

The paper proposes an approach for the integration of logical document structure information into an anaphora resolution system. The approach is based on previous work from the project A2 “Sekimo” on the integration of heterogeneous linguistic resources and attempts to close the gap between application domains of anaphora resolution [Machine Translation, Summarization, Text Mining; e.g. Steinberger et al., 2007, Boguraev and Kennedy, 1999] and the text size that most anaphora resolution systems focus on. These are mostly short dialogues or texts whereas the application domains focus not only on short texts but - especially in case of summarization - on long texts, too. For long texts, however, special attention has to be paid to the detection of antecedent candidates. In general, an anaphora resolution system can be subdivided into different steps: (1) detection of anaphoric discourse entities, (2) creation of an antecedent candidate list for each anaphoric discourse entity, (3) filtering the candidate list according to agreement and selectional restrictions, and (4) detection of the most likely candidate. The present paper focuses on the second step, i.e. on the creation of suitable candidate lists. Two different approaches exist for the creation of a candidate list: linear models and hierarchical models.

On the one hand, linear models define the search windows for antecedent candidates in terms of linear distance, i.e. in terms of tokens, markables, sentences, paragraphs and the like [e.g. Soon et al., 2001, Yang et al., 2004]. On the other hand, hierarchical models describe referential accessibility and referential distance in terms of hierarchical discourse structure [Polanyi, 1988, Cristea et al., 1998, 2000, Chiarcos and Krasavina, 2005]. Whereas hierarchical approaches that use rhetorical discourse structure information [e.g. *RST*, Mann and Thompson, 1988] lead to good results regarding the prediction of referential accessibility, a caveat to the use of discourse structure is the difficulty to create annotated data. Although discourse parsers exist, the

creation of annotated data is a time-consuming task [Carlson et al., 2003]. Therefore, we argue for an alternative to discourse structure: logical document structure. Logical document structure describes the structure of texts in terms of chapters, sections, paragraphs, lists, and the like and it is available either in terms of document markup (e.g. DocBook, LaTeX, HTML) or can be extracted from formatted texts (e.g. pdf files). Power et al. [2003] describe (logical) document structure to realise rhetorical structure¹. Thus, we assume logical document structure to give valuable cues for the detection of antecedent candidates in two ways: Logical document structure might have either (a) a direct influence on the discourse entities (or antecedent life span) or (b) an influence on the search window, e.g. in terms of different window sizes according to the NP type of the anaphor [Goecke and Witt, 2006]. Whereas for short texts linear models lead to good results, hierarchical models are needed in order to find antecedent candidates in long texts. Corpus investigation on a set of scientific and newspaper articles shows that only about 50% of the anaphoric elements find their antecedent within a fixed search window of 15 markables. However, it is not possible to create candidate lists by simply enlarging the search window, as the size of the candidate lists influences subsequent steps of the resolution process: The bigger the candidate list the higher the probability to choose a false candidate instead of the correct one. Therefore, we have to answer two restrictions when creating the candidate list. First, the candidate list should include the correct antecedent and, second, the candidate list should be as small as possible. The full paper describes the interrelationship of logical document structure and discourse structure in detail and presents results of a corpus study that investigates the impact of logical document structure for referential accessibility.

Literatur

- Boguraev, B. and Kennedy, C. (1999). Saliency-based content characterisation of text documents. In Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pages 99–110. MIT Press, Cambridge, Mass.
- Carlson, L., Marcu, D., and Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In van Kuppevelt, J. and Smith, R., editors, *Current Directions in Discourse and Dialogue*. Kluwer Academic Publishers, Dordrecht.
- Chiarcos, C. and Krasavina, O. (2005). Rhetorical distance revisited: a parametrized approach. In *Proceedings of Workshop on Constraints in Discourse*, pages 63–70, Dortmund, Germany.

¹Power et al. [2003] use the term *abstract document structure*

- Cristea, D., Ide, N., Marcu, D., and Tablan, V. (2000). Discourse structure and co-reference: An empirical study. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000)*, Luxembourg. An earlier version of this paper was presented at the ACL'99 Workshop on the Relation Between Discourse Structure and Reference, Maryland, June 1999.
- Cristea, D., Ide, N., and Romary, L. (1998). Veins theory: A model of global discourse cohesion and coherence. In *Proceedings of ACL/COLING'98*, pages 281–285, Montreal.
- Goecke, D. and Witt, A. (2006). Exploiting logical document structure for anaphora resolution. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.
- Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organisation. *Text*, 8(3):243–281.
- Marcu, D. (2005). Automatic discourse parsing. In *Encyclopedia of Language and Linguistics 2nd Edition*, volume 3, pages 649–654. Elsevier.
- Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638.
- Power, R., Scott, D., and Bouayad-Agha, N. (2003). Document structure. *Computational Linguistics*, 29(4):211–260.
- Soon, W. M., Lim, D. C. Y., and Ng, H. T. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.
- Steinberger, J., Poesio, M., Kabadjov, M., and Jezek, K. (2007). Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43(6):1663–1680. Special Issue on Summarization.
- Yang, X., Su, J., Zhou, G., and Tan, C. L. (2004). Improving pronoun resolution by incorporating coreferential information of candidates. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL04)*, Barcelona, Spain.