Exploring automatic theme identification: a rule-based approach

Lara Schwarz, Sabine Bartsch, Richard Eckart and Elke Teich

Abstract. Knowledge about Theme-Rheme serves the interpretation of a text in terms of its thematic progression and provides a window into the topicality of a text as well as text type (genre). This is potentially relevant for NLP tasks such as information extraction and text classification. To explore this potential, large corpora annotated for Theme-Rheme organization are needed. We report on a rule-based system for the automatic identification of Theme to be employed for corpus annotation. The rules are manually derived from a set of sentences parsed syntactically with the Stanford parser and analyzed in terms of Theme on the basis of Systemic Functional Grammar (SFG). We describe the development of the rule set and the automatic procedure of Theme identification and assess the validity of the approach by application to some authentic text data.

1 Introduction

Text data applications of NLP, such as information extraction (IE) or document classification (DC), require a new look at issues of discourse parsing. While the focus in discourse parsing has been on qualitative analyses of *single* texts – for instance identifying the meaningful, coherent parts of a text (generic structure, rhetorical structure, logical structure; see e.g., Marcu (2000); Poesio et al. (2004)), interpreting reference relations (co-reference resolution) or analyzing information structure (e.g. Postolache et al. 2005) – the attention of NLP in IE/DC is on *sets* of texts and quantitative features. So far, the potential contribution of discourse knowledge for IE/DC applications has hardly been explored, since the predominant methods are string or word-based and even supervised data mining rarely employs information at higher levels of linguistic abstraction. Here, the bottleneck is often the lack of (large enough) corpora annotated in terms of discourse features.

The work reported on in this paper is placed in the context of enhancing corpora with linguistic features of discourse organization, an increasingly active research area (see e.g. Lobin et al. 2007; Lüngen et al. 2006; Stede and Heintze 2004). We report on the derivation of rules for automatic Theme identification from a set of sample sentences instantiating the principal Theme options of English. Our approach combines automatic syntactic parsing with annotation of Theme (cf. the work by Honnibal and Curran (2007) on enhancing the Penn Treebank in terms of features from Systemic Functional Grammar Halliday (2004), or Buráňová et al. (2000) on