On the contribution of discourse structure to topic segmentation

Paula C. F. Cardoso[1],

techniques for topic segmentation, including those based on text structure, cue words and high-frequency indicative phrases for topic identification in a summarization system. Although the authors do not mention an evaluation of these features, they suggested that discourse structure might help topic identification. For this, they suggested using RST.

RST represents relations among propositions in a text and discriminates nuclear and satellite information. In order to present the differences among relations, they are organized in two groups: subject matter and presentational relations. In the former, the text producer intends that the reader recognizes the relation itself and the information conveyed, while in the latter the intended effect is to increase some inclination on the part of the reader (Taboada and Mann, 2006). The relationships are traditionally structured in a tree-like form (where larger units – composed of more than one proposition – are also related in the higher levels of the tree).

To the best of our knowledge, we have not found any proposal that has directly employed RST for topic segmentation purposes. Following the suggestion of the above authors, we investigated how discourse structure mirrors topic shifts in texts. Next section describes our approach to the problem.

3 Strategies for topic segmentation

For identifying and partitioning the subtopics of a text, we developed four baseline algorithms and six other algorithms that are based on discourse features.

The four baseline algorithms segment at paragraphs, sentences, random boundaries (randomly selecting any number of boundaries and where they are in a text) or are based on word reiteration. The word reiteration strategy is an adaptation of TextTiling[1] (Hearst, 1997)

The next algorithms are based on the idea that some relations are more likely to represent topic shifts. For estimating this, we have used the CSTNews (described in next section), which is manually annotated with subtopics and RST.

In this corpus, there are 29 different types of RST relations that may connect textual spans. In an attempt to characterize topic segmentation based on rhetorical relations, we recorded the frequency of those relations in topic boundaries. We realized that some relations were more frequent on topic boundaries, whereas others never occurred at the boundariet 609.31 Tm[occ)-2(ur)7ETon

his cor9.2

er they are closer or farther from one another), they score zero. However, it is also important to know how close the identified boundaries are to the expected ones, since this may help to determine how serious the errors made by the alg

References

Paula C.F. Cardoso, Erick G. Maziero, Maria L.R. Castro Jorge, Eloize M.R. Seno, Ariani Di Fellipo, Lúcia H.M. Rino, Maria G.V. Nunes, Thiago A.S. Pardo. 2011. CSTNews –