UNIVERSITY OF GENEVA

Automatic Structuring of Dialogs: the Problem of Topic Segmentation

PhD THESIS

Maria Georgescul

Review committee members:

Supervisor: Susan Armstrong, ISSCO/TIM, ETI, University of Geneva

Advisor: Alexander Clark, Department of Computer Science, Royal Holloway University of London.

Dan Cristea, Faculty of Computer Science, University of Iasi

Diane Litman, Department of Computer Science, University of Pittsburgh

Paola Merlo, Department of Linguistics, University of Geneva

President: Barbara Moser-Mercer, Interpretation Department, ETI, University of Geneva

Andrei Popescu-Belis, ISSCO/TIM, ETI, University of Geneva

GENEVA, SWITZERLAND
2006

Abstract

Thematic segmentation is an important initial step in applications such as information retrieval and document summarization. With the rapid growth of textual/audio/visual materials available in electronic form, the need for the automation of these tasks is evident. In this thesis we address the task of automatic text structuring into linear and non-overlapping thematic episodes. In particular, we investigate the appropriateness of using new discriminative and generative machine learning techniques, trying to exploit their complementary advantages. We attempt to enhance the topic segmentation performance on multi-party meeting recording transcripts, which pose specific challenges for topic segmentation models. Experiments are conducted primarily by exploiting features based on lexical reiteration, but also by using syntactic, prosodic and pragmatic structural information specific to multi-party dialogues. We also report on the performance of news story segmentation and on segmentation of datasets containing artificial thematic episodes generated by putting together blocks of text that have been randomly extracted from different documents. By evaluating using various corpora to assess the performance of various systems, we bring evidence that many previous research efforts related to topic segmentation have involved evaluation on artificial data which provide little conclusive evidence of system effectiveness. We also analyze whether existing evaluation measures are adequate for the task of topic segmentation and we propose and implement a new evaluation metric.

Résumé

La tâche principale abordée dans cette étude concerne le découpage d'un flux continu de textes ou transcriptions de dialogues parlés en segments relatifs au même sujet. Plus précisément, nous traitons de la segmentation automatique des textes en épisodes thématiques non superposés et ayant une structure linéaire. Nous proposons à cet effet des approches fondées sur l'apprentissage automatique en utilisant des modèles discriminatives et génératives. Nous nous intéressons en particulier au problème de la segmentation thématique dans le contexte des dialogues multi-locuteurs qui ont été préalablement enregistrés et transcrits. Nous procédons également à l'évaluation des algorithmes sur des ensembles de données contenant des transcriptions d'émissions radio/TV, ainsi que sur des données contenant des épisodes thématiques artificiels obtenus par la concaténation de fragments de textes qui ont été aléatoirement extraits à partir de textes narratifs. Par rapport aux autres types de textes, la segmentation thématique des transcriptions de dialogues multi-locuteurs pose des défis spécifiques que les expériences que nous avons entreprises mettent en évidence. Ces expériences montrent également que l'efficacité de certains systèmes sur des données artificielles ne s'etend pas nécessairement aux données réelles. Nous analysons également la fiabilité des mesures d'évaluation qui existent et nous proposons une nouvelle mesure d'évaluation plus fiable. Dans l'optique de trouver des méthodes paramétrables applicables à différents types de textes, nous avons exploité principalement les traits lexicaux. Cependant, pour les données contenant des transcriptions de dialogues multi-locuteurs, nous explorons également des traits lexicaux, acoustiques et syntaxiques.