論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Citation Block Determination in Academic Texts
著者(和文)	Kaplan Dainan
Author(English)	Dainan Kaplan
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第10254号, 授与年月日:2016年3月26日, 学位の種別:課程博士, 審査員:徳永 健伸,徳田 雄洋,宮崎 純,村田 剛志,藤井 敦
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第10254号, Conferred date:2016/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	要約
Type(English)	Outline

Citation Block Determination in Academic Texts

Dainan Kaplan

This work presents a novel approach to citation block determination (CBD) using textual coherence. CBD is the task of identifying a citation block, the span in sentences referring to a work as cited by a citation anchor (e.g. "(Ingsoc, 1984)") within the running text of research papers. Identifying citation blocks is crucial for research utilising information present in citation blocks, such as research paper summarisation, idea attribution, sentiment analysis, and other citation-based analysis research. As recent studies suggest that the majority of salient information related to a citation block is beyond the initial sentence containing the citation anchor, detection of the citation anchor alone is insufficient, and makes the task of CBD all the more pertinent. We propose the use of textual coherence for CBD, which concerns itself with how text is composed to form a meaningful whole.

Hitherto CBD has used only ad-hoc features targeting aspects of a citation block itself, such as binary features indicating the presence of another citation anchor or author name. Results from previous work also indicate ample room for improvement and CBD can still be seen as in its infancy. This work proposes a novel approach that utilises textual coherence to capture salient aspects of the citation block for determining its boundaries. To our knowledge this is the first work that attempts to do so.

Textual coherence (TC) is a field of research that describes in general how a given text is structured, both in terms of meaning and discourse, and how this structure builds up to create a unified, meaningful (i.e. "cohesive") whole. For example, TC describes how discourse connectives like "however" and "in addition" produce a flow of statements within a work describing how each is related to one another (e.g. "contrast" or "conjunction"), and how lexical-chains, such as the repetition of words, will be present in order for a text to be comprehensible. We hypothesise that TC will be effective for CBD because it is the formalisation of how a text is built up from smaller constituents into a cohesive whole. By definition a text must be cohesive as a whole in order to be comprehensible to the reader, or it would appear as a series of disconnected, unrelated fragments. As citations are objective-driven, i.e., they are introduced into the discourse for the purpose of saying something about them, it follows that they should also be cohesive as a whole. We hypothesise therefore that TC should be effective for CBD if we can properly exploit certain TC aspects of a text. One main challenge becomes distinguishing between generally coherent text, and the coherence specific to a citation.

We begin first with adding to existing ad-hoc features that target specific aspects of the citation block itself; the new additions can be categorised into three groups; the first lexicalises previously existing binary features, bringing a simple and controlled form of textual cohesion into the features. Other features target pronominal expressions that are a type of reference coherence. The third group targets specific aspects of anchor sentences so the models can hopefully learn to distinguish between different types of citations. Our best proposed model here achieves upwards of 20% lift over the baseline.

After this experiment, we move on to experiment with a more extensive type of TC feature utilising coreference-type features. Coreference describes how entities are referenced from one sentence to the next, and if this can be utilised effectively, should yield promising results. Unfortunately, due to limited coverage of anaphoric resolution in the academic domain, many coreferences are missed and despite high precision, this method suffers from comparatively low recall. From here we then present a battery of feature sets utilising various aspects of textual coherence. These feature sets include relational coherence features, specifically location and discourse structure features, and lexically motivated features using entity-grids, point-wise mutual information, and topic models. The location features allow us to segment citations based on where they occur in a paper, which has been shown to affect the citation style; the discourse features try to capture salient aspects of the flow of discourse. The lexically motivated features try to capture other, latent aspects of the text near citations by looking at how words/phrases appear and how information-rich they are with respect to the rest of the papers. Here the best performing model among proposed achieves lift of around 10% above baseline.

We then combine the extended citation specific features with the best performing textual coherence features to see how their synthesis compares to the results from previous experiments. Combination of methods yields higher recall (around 2%-4%) over the best proposed from previous chapters without compromising much on recall (.1%-.6%).

We conclude with a summary of this work's findings, illuminating our systematic approach to classification of feature sets for CBD, and our proposal of textual coherence for CBD, and finally suggest possible directions for future work.