An Analysis of Expression Patterns for Establishing Research Significance

Kiyoko Uchiyama
Dept. Applied Computer Science
Shonan Institute of Technology
Fujisawa, Japan
e-mail: uchiyama@sc.shonan-it.ac.jp

***Abstract*—It is crucial to investigate essential information for comprehending contents of technical documents and academic papers in order to write a paper as novices. The previous works revealed the importance of grasping the logical structure and the knowledge of technical terms in the domain-specific field. As it takes a lot of time to acquire the knowledge of technical terms, a method which can be assumed the meaning of technical terms requires for effective reading. In this paper, we attempt to extract and analyze expression patterns of establishing discourse structure and reflecting author's intention in the Section Introduction of academic papers. The analysis carried out using by original categorization based on the existing model and reported the results.**

Keywords-Expression Patterns; Comprehensive Reading; Creating A Research Space (CARS) model.

I. INTRODUCTION

In academic education, some important assignments ask for reading academic papers and writing a report in specific field. Students in the engineering department have a lot of assignments that require reading of academic papers and technical reports related to the state of the art of technology.

Such documents, however, include various technical terms, which are unknown words for undergraduate students. A lack of knowledge of technical terms makes it difficult for novices to read the technical documents in the specific field. On the other hand, education for reading technical documents is not enough at the early stage of research.

We have proposed a method for reading technical documents, understanding technical terms and function words in logical structure. Firstly, regarding the technical terms, novices need to know the technical terms, which are basic and essential to a target field, in advance. However, the importance or essentiality of the terms in a target field remains unclear. We defined such technical terms as introductory terms. If novices do not have any knowledge of the introductory terms, they cannot comprehend the outline nor understand more difficult terms in a target field. We proposed various criteria for identifying the terms in a specific domain [1]. We proposed original criteria for the introductory terms: priority and compositionality and calculated the score based on C-Value [2]. C-Value is one of the term scoring methods and uses the type and token frequency for each constituent from the compound nouns in a corpus of the target field.

At first, we defined priority as a sort of ordering for learning in textbooks and attractive keywords in research papers. Secondly, concerning the compositionality, introductory terms tend to generate various new compound nouns by concatenating single words or word strings in prefix/suffix form. The introductory term candidates were calculated based on the type and token frequency occurred in academic papers and textbooks. As the result, we found that the frequency from the table of contents in textbooks was useful for extracting the introductory terms.

The subsequent analysis of the distribution of the terms has processed in a logical structure, such as "Abstract", "Introduction", and "Conclusion" [3]. The introductory terms tend to be included in the logical structure of "Abstract" and "Introduction", rather than that of "Experiment", "Discussion" and "Conclusion". It is assumed that novices can understand the outline of technical documents by effective reading the section of "Abstract" and "Introduction".

Based on those previous analysis, comprehensive reading in "Abstract" and "Introduction" section is necessary for novices in order to grasp the outline of the target paper effectively. As the "Abstract" section is too short to analyze the structure, "Introduction" is a target section.

As it takes a lot of time to acquire the knowledge of the technical terms in a specific field, a method requires other clues for comprehensive reading than the method by using the knowledge of technical term. That is to say, it is crucial that the meaning of technical terms can be detected by using functional words, phrases which establish the author's intention in the context.

In this paper, the expression patterns which reflect the discourse of the paper and the author's intention are analyzed in the context of the Section Introduction. The three steps are introduced for the analysis. Firstly, the role assigned to each sentence, in other words, discourse segment which dominates the context in the paper is processed. Secondly, based on the CARS model (detailed in Section 2), the following three types of expressions, which are related to construction and context of the paper are categorized for this analysis. Three types are (1) mutual expressions frequently used in academic field, (2) characteristic expressions in domain-specific field, and (3) reflecting expressions for establishing the author's intention. Finally, we analyze the relationship between the role of sentence and each type of expressions and organize the results by the previous two steps.

This paper is structured as follows. In Section 2, related works are summarized and our motivation to conduct our study. The analysis and results are described in Section 3 and Section 4 concludes our possible future work.

II. Related Works

There are several researches of rhetorical structure and writing strategy in academic papers. The existing researches are focusing on the role of each sentence and categorization of discourse segmentation related to our research.

A. Creating a Research Space (CARS) model

Based on existing analysis of the Section Introduction, we assumed that the CARS model proposed by Swales [4] can be applied to analyze the structure of target documents. CARS model consists of three moves that describe how research paper introductions are structured.

The three rhetorical moves are: (1) establishing a territory, (2) establishing a niche, and (3) occupying the niche. The model breaks down each of those moves into more detailed descriptions. The move1 establishing a territory includes three steps, claiming centrality, making topic generalizations, and reviewing items of previous research. After describing move1, authors try to write their refutation to earlier research, indicate a gap, raise a question and continue a tradition. Finally, authors reveal their findings or solution to help fill the gap in move2, by outlining purposes, announcing present research and main findings, indicating structure of the paper and evaluation of findings.

In establishing a niche of CARS model, authors claim their research advantages by showing that the previous research are not enough. Authors criticize the existing research by using words expressing a contrast evaluation, such as “less”, “little”, “fail”, “ignore” and “inefficient”. This sort of expressions might become clue words for novices to understand the author’s intention and find the originalities of the target documents.

B. The Role of Sentence in Discourse Segment

Swales’ CARS model has been used extensively by discourse analysis and annotation scheme for information retrieval of scientific papers. A Core Scientific Concepts (CoreSC) is one of the annotation schemes [5][6]. This annotation scheme adopts the view that a scientific paper is the human-readable representation of a scientific investigation. The CoreSC introduced 11 categories. Similarly, de Waad and Pander Maat categorized seven discourse segments: Fact, Problem, Goal, Method, Result, Implication and hypothesis [7][8]. The seven categories at the sentence level can be used for classifying the sentence in the Section Introduction.

We correspond Swales’ CARS model to seven discourse segments in each sentence for analyzing expression patterns in the Section Introduction.

III. Analysis and Result

We collected and analyzed academic papers in order to investigate the expressions which are structured. The Section Introduction were selected from the full text of the academic papers. The key expressions were extracted referring Swales’ CARS model and classified by the role in structure of Introduction.

A. Target Documents

One hundred academic papers written in Japanese which include a keyword “Natural Language Processing” in Information Processing Journal of Japan from 1998 to 2011 were collected. The 2000 sentences in the Section Introduction are target for this paper. The seven categories of annotation scheme for discourse segments is assigned to each sentence.

B. Analysis and Result

We analyzed three types of expressions. The first type is mutual expressions frequently used in academic field. This type can also define the role of sentence. For example, the expressions like “the purpose of this paper”, “we propose a method…” can be assigned the role of “Goal” to the sentences. The second type is characteristic expressions in domain-specific field. There are several kinds of words: clue words in wide range of field, such as “precision” “method” “automation” in information processing or engineering field., domain-specific technical terms which can be defined as introductory terms in our research, such as “morphological analysis” “parse” “corpus” in natural language processing field.

The third type is reflecting expressions for establishing the author’s intention which corresponds to Swale’s move 2: establishing a niche. The expressions include various part of speech, conjunction, adverbs, verbs and adjectives. The authors tend to use positive/negative words in each part of speech for describing their intention or emphasis point of their research. The words, such as “versatile”, “enormous”, “redundant”, “robust” and “exclusive” are observed characteristically in information processing field. Those expressions are commonly used for evaluating proposed method or research in contrast to the expressions “contrast or negative evaluation” widely used in academic field.

IV. Conclusion

In this study, we defined the expressions which constitute of context and the Section Introduction in academic paper as “establishing expressions”. The three steps were proceeded for analysis of each sentence applying the framework of Swale’s CARS model and discourse segments. The results of the analysis were that establishing expressions have common ones in academic field and specific ones in domain-specific field. We plan to further investigate the establishing expressions in some field, and confirm whether those expressions can be useful for effective reading.

REFERENCES

[1] K. Uchiyama, “A Study for Identifying Domain-Specific Introductory Terms in Research Papers,” Proc. 9th Terminology and Artificial Intelligence, pp.147-150, 2011.

[2] Katerina T. Frantzi and Sophia Ananiadou (1996). Extracting Nested Collocations, In proceedings of the 16th International Conference on Computational Linguistics (COLING 96):41-46.

[3] K. Uchiyama, “An Analysis of Domain-Specific Introductory Terms in Logical Structure of Scholarly Papers,” Proc. Workshop on Corpus Japanese Linguistics, pp. 195-198, 2012.

[4] J. Swales, Genre Analysis English in Academic and Research settings, Cambridge University press, 1990.

[5] M. Liakata, S. Teufel, A. Siddharthan, and C. Batchelor, “Corpora for Conceptualization and Zoning of Scientific Papers,” Proc. Language Resources and Evaluation (LREC2010), pp.2054-2061, 2010.

[6] M. Liakata, S. Saha, S. Dobnik, C. Batchelor, and D. R. Schuhmann, “Automatic Recognition of Conceptualization Zones in Scientific Articles and Two Life Science Applications,” Bioinformatics, 28(7), pp.991-1000, 2012.

[7] A. de Waard and H. P. Maat., “Categorizing Epistemic Segment Types in Biology Research Articles,” Proc. Workshop on Linguistic and Psycholinguistic Approaches to Text Structuring (LPTS 2009), 2009.

[8] A. de Waard and H. P. Maat, “Epistemic modality and knowledge attribution in scientific discourse: a taxonomy of types and overview of features,” Proc. Workshop on Detecting Structure in Scholarly Discourse, pp.47-55, 2012.