論 文 の 内 容 の 要 旨

論文題目
Unsupervised Induction of Natural Language Discourse Structure Based on Rhetorical Structure Theory
（修辞構造理論に基づく談話構造の教師なし解析）

氏　　名　　西田　典起

Natural language text is generally coherent. Discourse coherence can be represented as discourse structures, which discourse parsing aims to analyze automatically for given text.

Despite the promising progress achieved in recent decades, discourse parsing still remains a significant challenge. The difficulty is due in part to the high cost and and low reliability of hand-annotated discourse structures.

This thesis tackles the problems by introducing unsupervised discourse parsing. Unsupervised discourse parsing is a novel technology that automatically induces discourse structures for input texts without relying on human-annotated discourse structures. Based on Rhetorical Structure Theory (RST), we assume that coherent text can be represented as a tree structure. The leaf nodes correspond to non-overlapping clause-level text spans (called elementary discourse units in RST), while the internal nodes consist of three orthogonal elements: (1) discourse constituents, (2) discourse nuclearities, and (3) discourse relations.

Based on this assumption, we first break down the unsupervised discourse parsing problem into smaller subtasks, each corresponding to one of the three orthogonal elements. Then, we propose unsupervised algorithms for the three subtasks. The unsupervised algorithms are developed based on our prior knowledge of the target discourse elements.

Experimental results demonstrate that our unsupervised algorithms outperform and improve unsupervised baselines. Moreover, our unsupervised algorithm induces more accurate discourse constituents than recent fully supervised parsers. We also analyze what can or cannot be captured by our unsupervised algorithms, and find that careful consideration of prior knowledge is crucial in unsupervised discourse parsing.