平成25年度 情報工学コース卒業研究報告要旨

石川研究室 研究室	氏　　名	楊　遠　東
卒業研究題目	テキスト構造を特徴づける 手がかり表現の自動抽出	

文章というものは、単純に文を並べたものではなく、書き手の意図を読み手に伝えるための構造を持っている。文章を作成する時には、適切なテキスト構造を意識しながら書かなければ、読みやすい文章を作成できない。文と文の間にある関係を明確に述べるには、その関係を特徴づける接続詞や表現を使う必要がある。

本論文では、テキスト構造を特徴づける手がかり表現を自動抽出する手法を提案する。本手法では、テキスト構造を修辞構造理論（Rhetorical Sructure Theory, RST）により表現する。修辞構造理論はテキストの部分間の修辞関係を記述するフレームワークを提供する。修辞構造はテキスト内の文や節などを基本単位とし、テキストを木構造として表現する。修辞関係は2つのテキストスパンの間に成り立つ。修辞関係で結ばれたテキストスパンの重要な方を核という。テキストスパンの重要でない方を衛星という。修辞構造理論は文章の自動生成、分析、自動要約などによく利用されている。

本論文で提案する手法では、修辞構造が与えられたコーパスから統計情報を使って、修辞関係を特徴づける手がかり表現を自動的に抽出する。次の2ステップで実現する。

1. テキストに出現する単語列が言語的なまとまりを持つフレーズかどうかを統計情報を使って判定する。統計情報として、自己相互情報量（Pointwise Mutual Information, PMI）を用いて、コーパスからN-gramの単語列を取り出し、各N-gramについて、それを2つ分割してPMIにより、それらの結びつきの強さを計算する。また、出現頻度が低い単語列については、信頼できるPMIの値が得られないので、単語列が出現する頻度を使って、頻度が低い単語列を削除する。

2. 単語列がこの修辞関係を特徴付けているかを判定する。統計情報として、修辞関係と単語列の間のPMIを使う。

RST Discourse Treebankコーパスを用いて、評価実験を行った。コーパスから修辞関係とテキストスパンのペア18422個のデータを取り出し、提案手法により修辞関係を特徴づける手がかり表現を抽出した。実験の結果、本手法により、表1のような結果が得られ、本手法の有効性を確認できた。

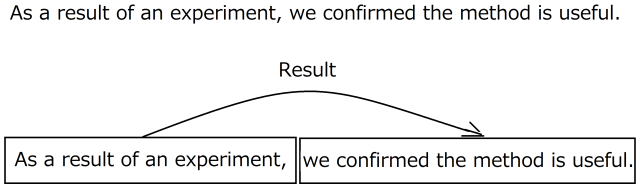

図1 修辞構造と手がかり表現の例

表1 本手法により抽出された表現の例

修辞関係	手がかり表現
Result	As a result of

+