泉　朋子（黒橋教授）

「Normalization and Similarity Recognition of Complex Predicate Phrases Based on Linguistically-Motivated Evidence」

（言語学的特徴を用いた述部の正規化と同義性判定）

平成 26 年 1 月 23 日授与

ブログ、Twitter、対話ログなど大量のテキストデータから有益な情報を抽出するためには、計算機による深い意味理解が必要である。特に、「つながらない」「わからない」など「どうした」を表す述部は、文の核情報を表しており、これら述部の意味を正しく計算機が扱えることで、意見マイニングや、QA/対話システム、検索エンジンなど後段の言語処理技術の精度を大幅に向上させることが可能である。しかし、これらの述部は、その表現の多様性ゆえに、計算機で意味を扱う事（すなわち同義性を認識する事）が困難である。例えば、「つながらない」という表現も、「接続を行うことができない」「つながらないんですが」など様々な言い方で表現される。

本論文では、文の「どうした」を表す述部に焦点を当て、これらの多様な述部表現を、言語学的な分析をもとに、同じ出来事を表す単純な表現に正規化し、さらに文脈をも考慮し同義となる述部を自動で認識する高精度な同義判定技術を提案した。

本論文で対象にする述部の多様性には、「形態素レベルでの多様性」「構造レベルでの多様性」「意味レベルでの多様性」という 3 つの要因が関連している。第 2 章では、「形態素レベルでの多様性」に焦点をあて、日本語の文末表現の正規化を提案している。出来事の意味に影響を与える文末表現を、形式意味論の「時制」「モダリティ」「否定」という 3 つの軸をもとに定義し、これらに属す表現のみを最低限残す正規化ルールを確立し、述部を「同じ出来事を表す最も単純な表現」に正規化した。第 3 章では、「構造レベルでの多様性」に焦点をあて、「解約を行う」「納得が行かない」などの機能動詞構造の正規化を提案した。動詞の文法機能のみを保持しつつ、単純な述部に言い換えるための正規化パターンを構築し、さらに、機能動詞構造と本動詞構造の曖昧性解消に関して、大規模な新聞・ブログコーパスを用いて「曖昧性解消辞書」を構築した。第 4 章では、「意味レベルでの多様性」に焦点を当て、「メモリを消費」と「メモリを食う」のような文脈によって同義になり得る述部をも対象とした、述部の同義性判定手法を提案した。辞書定義文、用言属性、分布類似度、機能表現といった異なる言語情報から、同義述部の特徴を抽出し、同義判定の素性として用いた。さらに、今まで言語処理では困難であった、反義関係を表す述部と同義関係を表す述部の識別を、反義関係に特化した言語的特徴を用いることで、正しく判別することを可能にし、既存手法に比べて高精度に述部の同義性を判定することができた。

今後は、これら述部の同義性判定技術を用いて、QA システムや検索エンジンなど、上位アプリケーションでの効果を検証するとともに、同義だけではなく反義・含意・推意といったより深い述部の意味関係の認識・獲得技術の研究を行う。

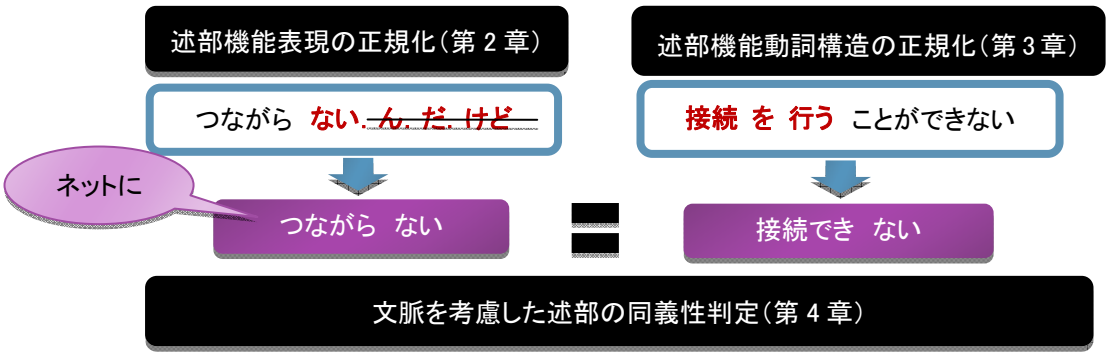

図 1：述部の正規化と同義性判定