審 査 の 結 果 の 要 旨

氏名　Hugo Hernault（エルノー　ユーゴ）

本論文は「A Study on Semi-Supervised Approaches for Discourse Analysis（談話解析への半教師付き学習アプローチ)」と題し，英文で記されており，６章から成る．

第１章「Introduction（序論)」では，自然言語テキストの談話解析の役割とこれまでに提唱されてきた理論とモデルについて記し，既存研究をまとめている．そして本論文の貢献は，(1)テキストの効率的な談話ユニットへの分割法，(2)木構造の談話関係構造を生成する教師付き学習による談話解析器の開発，(3)十分な談話関係ラベル付き訓練用データが存在しない状況下で，談話解析器を構成するために重要となる大量の談話関係ラベル付与なしデータも活用する半教師付き学習法(semi-supervised learning)，であるとしている．

第２章「A Sequential Model for Discourse Segmentation（談話分割のための逐次処理モデル)」では，条件付き確率場(CRF: Conditional Random Field)に基づく基本談話ユニットへの逐次分割法を記している．このテキストの基本談話ユニットへの分割は談話分析の最初のステップとして重要である．本研究では談話分割器を RST(Rhetorical Structure Theory)談話 Treebank のデータを用いて CRF による訓練によって得ている．この提案手法による談話分割器は，実験により SVM(Support Vector Machine)等を用いる統計的手法やルールベース手法と比較して，優れていることを示している．提案手法による実験結果では F 値が 0.94 程が達成されるが，これは人間による判定の F 値の 0.98 に近くなっている．

第３章は「An Implemented Discourse Parser Based on Support Vector Machine（SVM により実装された談話構造解析器)」であり，RST に基づく木構造となる談話構造解析器の開発について記している．以前の教師付き学習による談話構造解析が単文レベルの解析か部分的に実装されたシステムであったのに対し，本論文の談話構造解析は複数文を含む文全体を解析する完全に実装されたシステムとなっており，性能も既存のものを上回ることを実験的に示している．

第４章「Co-occurrence-based Discourse Relation Classification（共起に基づく談話関係識別)」では，談話構造解析器の中核となる談話関係識別を行うための半教師付き学習法を記している．既存の多くの談話関係識別器の設計は，談話関係レベル付きコーパスを用いる教師付き学習に基づいてきたが，このような多量のコーパスは必ずしも存在するとは限らず，特に新たに設定された談話関係については利用できない．そこで，ここでは少数の談話関係ラベル付きテキストデータと多数の談話関係ラベル付与のないテキストデータを活用する半教師付き学習法を提案している．本章で考案している半教師付き学習法は，ラベルなしデータ中の特徴の共起情報を活用するものであり，この特徴の共起情報を利用して関係識別で用いる特徴ベクトルを拡大する．これにより，ラベル付きデータだけでは得られない特徴量も利用できることになり，性能の向上に繋がる．RST 談話 Treebank コ

ーパスと Penn 談話 Treebank　コーパスを用いた実験により，提案の半教師付き学習による談話関係識別器は識別性能向上をもたらすことを示している．例えば，約 1000 のラベル付きデータを用いる場合，提案の半教師付き学習による談話関係識別器はベースライン識別器に対して，マクロ平均 F 値で 50%程の識別性能向上をもたらすことを示している．これは出現が低頻度であるために，少数のラベル付きデータしか得られていない談話関係に対する場合などで特に有効となる．

第 5 章「Structural Learning-based Discourse Relation Classification（構造学習ベースの談話関係識別）」では，中心認識課題の部分問題である多数の補助問題に対する学習結果を利用する半教師付き学習手法（構造学習と呼ばれる）を，談話関係識別に適用するアプローチを記している．ここでは初めに，ラベルなしデータに対して多数の補助的識別問題を解き，次いでこれによって学習した多数の補助的識別器を特徴ベクトルの拡大に用い，談話関係識別器の訓練に使用する．前章と同じ特徴集合とラベルなしデータを用いた場合においては，補助的識別器の数は 13,000～15,500 程になっている．前章の共起に基づく手法では約 15,000 もの特徴が追加されるのと比較すると，構造学習ベースと呼ぶ本章の手法は性能向上のためには補助的識別器学習のためにより多くのラベルなしデータを必要とするが, 数学的処理により最終的に追加となる特徴量の数は 50 程の少数になる. 実験により，二つの半教師付き学習による談話関係識別の性能はほぼ同等になることを示しており，両者とも少数のラベル付きデータしか存在しない場合に，ラベルなしデータの利用により大きい性能向上が得られることを示している．

第 6 章「Conclusion and Future Work（結論と今後の研究）」では，本論文の研究成果をまとめ，今後の研究の方向と課題に言及している．

以上を要するに，本論文は自然言語テキストの談話関係解析において，CRF(Conditional Random Field)に基づく効率的な談話ユニットへの分割法，RST(Rhetorical Structure Theory)に基づく木構造の談話関係構造を生成する教師付き学習による談話解析法，十分な談話関係ラベル付きデータが存在しない状況下で談話関係識別器を構成するための大量の談話関係ラベルなしデータも活用する 2 種の半教師付き学習法 ― 特徴の共起を利用し特徴ベクトルを拡張する独自手法と多数の補助問題に対する学習結果を利用する構造学習法をベースとする手法 ― を考案，開発し，性能向上効果を実験的に示している．これは自然言語テキスト談話解析において，特に半教師付き学習を適用した先駆的な貢献と認められ，情報理工学における創造的実践の観点で価値が認められる．

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる．