

審査の結果の要旨

氏 名 西田 典起

本論文は, 「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)」と題し, 英文6章からなる. 談話構造解析はさまざまな自然言語処理タスクの根幹をなす重要な技術であるが, 既存の談話構造解析手法の多くは人手によってアノテーションされた談話構造データを用いた教師付き学習に基づいている. しかしながら, そのようなデータは現実的には大量に収集することが困難であり, また談話解釈の曖昧性ゆえにしばしば教師情報としての信頼性が低いという問題がある. この問題に対し, 本論文では, 人手による談話構造データに依存せず談話構造解析を行う Unsupervised Discourse Parsing (教師なし談話構造解析) という新たな枠組みを提案し, これを実現する方法論と手法について述べている.

第1章「Introduction」では, 本論文で取り組む談話構造解析という研究分野の背景を概観している. まず, 文章の一貫性の表現として談話構造という考え方があることと, そこではある種の木構造が仮定されていることを修辞構造理論 (RST) 等の理論を挙げながら説明している. また, 談話構造解析の既存技術および応用タスクについて, 具体的な研究例を紹介しながら説明している.

第2章「Unsupervised Discourse Parsing: An Overview」では, 既存の談話構造解析技術の問題点・限界点を踏まえ, 本論文では, 教師なし談話構造解析という新たな枠組みを提案し, これに取り組むことを述べている. そして, 本論文で仮定するRSTに基づく談話構造 (木構造) を相補的な要素に分解し, それぞれの要素に対して教師なしアルゴリズムを実現していくことで, 全体として教師なし談話構造解析を実現するというロードマップを提示している. また, 教師なし談話構造解析によって期待される技術的利点について述べている.

第3章「Unsupervised Learning for Discourse Constituency Parsing」では, 入力文章のラベルなし構成素構造 (木構造) を解析するための教師なしアルゴリズムを提案している. 提案手法では, 教師なし構文解析で提案されているEMアルゴリズムをベースに, 談話解析のための拡張を施すことで, 教師なし談話構成素構造解析手法を実現している. 具体的には, 談話構成素構造解析器を, Viterbi EMと呼ばれる手法を利用して訓練しており, Viterbi EMとヒンジ損失関数を統合することによって, 談話構成素の明示的な数え上げを回避することに成功している. また, 談話の構成素性のViterbi trainingに有効な初期化手法を, 文章構造に関する事前知識に基づいて4つ提案し, それら

の重要性を示している．実験によって，提案手法がベースラインおよび教師あり手法と同等，またはより優れた解析性能に到達することを定量的に示している．

第4章「Unsupervised Induction for Discourse Nuclearity Classification」では，談話の核性（従属関係）を同定するための教師なしアルゴリズムを提案している．RST-DTコーパスを構築したCarlsonら(2001)が核性をチェックするために提案した削除テスト(deletion test)を，自動的に行うことで教師なし核性分類をするというのが基本的なアイデアである．談話の削除不可能性は，抽出型文書要約技術における文の重要度尺度と再帰アルゴリズムを統合することによって計算する．核性は結合される談話構成素それぞれの削除不可能性を比較することによって行う．実験では，提案手法がベースライン手法の核性分類精度を上回り，また提案手法のもとで複数の相補的な文の重要度尺度を組み合わせることによってさらに正解率を向上できることを定量的に示している．

第5章「Unsupervised Pre-training for Discourse Relation Classification」では，テキスト領域間の談話関係を分類するための教師なし事前学習アルゴリズムおよびその半教師あり学習への応用を提案している．特に，従来から難しいとされている，明示的な談話標識が現れないが談話関係が存在している暗黙的談話関係分類に取り組んでいる．文章の一貫性を認識するためには，局所的な文間に対して，明確な談話関係でなくとも，ソフトな談話関係の存在を認識できることが必要であると考えられる．そこで，本論文では，教師なし一貫性モデリング(coherence modeling)の学習を通してソフトな談話関係知識を獲得し，それに基づいて暗黙的談話関係分類の性能を向上させることを目指している．また，局所的な一貫性として，トピックに関する一貫性と，順序に関する一貫性が考えられるため，それらのどちらに関する一貫性がより有効なのかについて調べている．実験の結果，一貫性モデリングによって暗黙的談話関係分類の正解率は向上し従来手法を上回ることで，またトピックに関する一貫性が順序に関する一貫性よりも有効であることを定量的に示している．

第6章「Conclusions」では，本論文の成果と，現状の限界，および今後の展望についてまとめている．

以上を要するに，本論文は，高コストでしばしば信頼性が低い教師付き談話構造データに依存せずに談話構造解析を行う，教師なし談話構造解析という新たなアプローチを開拓したものである．より具体的には，談話構造解析で仮定される木構造を構成素性，核性，談話関係という3つの相補的な要素に分解し，それぞれの要素に対応するサブタスクに対して教師なしアルゴリズムを提案することで，教師なし談話構造解析器というシステムの実現を目指している．これらの成果は高い研究的価値を有するのみならず，自然言語処理の幅広い応用アプリケーションへとつながり得ることから，情報理工学における創造的実践の観点からの価値が認められる．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．