

様 式 C - 7 - 1

平成 26 年度科学研究費助成事業（科学研究費補助金）実績報告書（研究実績報告書）

1. 機関番号

1	4	6	0	3
---	---	---	---	---

 2. 研究機関名 奈良先端科学技術大学院大学
3. 研究種目名 基盤研究(A) 4. 研究期間 平成 26 年度～平成 28 年度
5. 課題番号

2	6	2	4	0	0	3	5
---	---	---	---	---	---	---	---
6. 研究課題名 構文パターン獲得と並列構造解析による統語的依存構造解析の高精度化

7. 研究代表者

研究者番号	研究代表者名	所属部局名	職名
1 0 2 1 1 5 7 5	マツモト ユウジ 松本 裕治	情報科学研究科	教授

8. 研究分担者

研究者番号	研究分担者名	所属研究機関名・部局名	職名
9 0 3 1 1 5 8 9	シンボ マサシ 新保 仁	情報科学研究科	准教授
8 0 6 3 7 3 2 2	ドゥ ケビン Duh Kevin	情報科学研究科	助教
2 0 7 3 4 7 8 4	シンドウ ヒロユキ 進藤 裕之	情報科学研究科	助教

9. 研究実績の概要

<p>長い文の統語解析の高精度化へ向けて、英語の複単語表現および構文パターン抽出のための資源構築に取り組んだ。複単語表現に関しては、英語の句動詞の網羅的な収集とそれぞれの例文の収集、および、例文中の句動詞候補が句動詞として使われている正用例かそれとも字義通りの意味として使われている負用例かアノテーション作業を行った。これを学習データとして、句動詞の正負用例を分類する実験を行い、高い精度で分類が可能であることを確認した。構文パターンの収集については、英語学習のための構文例文集から複文の例文を収集し、自動解析を行った結果の誤り修正と構文パターンのアノテーション作業を約 300 文に対して行った。また、アノテーション作業を行うためのツールの開発を行った。</p> <p>また、大規模の言語データを対象にして、高頻度語を階層的にまとめ上げることにより、頻出パターンを抽出する言語モデルを提案し、言語モデルとして有用であることを確認した。</p> <p>並列構造解析に関して、依存構造解析アルゴリズムを拡張することにより、英文中の並列構造を明示的に識別することを可能にするアルゴリズムを提案し、英語の標準コーパスの一つである Penn Treebank 中の並列構造をもつ文の大半をカバーできることを確認した。</p> <p>また、英語および日本語の依存構造解析の性能向上のため、形態素解析レベルの曖昧性を保持しつつ依存構造解析を実行するためのいくつかの手法の提案と実験を行った。</p>

10. キーワード

(1) 自然言語処理	(2) 並列構造解析	(3) 依存構造解析	(4) 多言語処理
(5) 機械学習	(6) 構文解析	(7)	(8)

11. 現在までの達成度

(区分)(3) やや遅れている。

(理由)

複雑な構文パターンや並列構造を解析するための基礎データの蓄積に、想定以上の作業量と時間を要したため、本年度の研究時間の多くをデータ収集とデータへのアノテーション作業に費やすることになった。そのため、英語の構文パターンのカバレッジに関する調査を行うことができなかった。
並列構造解析に関しては、アルゴリズムの設計は完了し、動作確認および解析時間の評価も行った。当初の予定通り推移しているが、実験の効率化のための実装の問題が新たに浮上したため、今後効率化への考察が必要であることが明らかになった。

12. 今後の研究の推進方策

(今後の推進方策)

英語の構文パターンの収集を引き続き行う。構文パターンの例文を収集したコーパスをより網羅的に作成し、構文パターンのアノテーション作業および辞書管理システムにおける構文パターンの表現方法の設計を行い、単語、複単語表現、構文パターンをすべて辞書項目として統一した表現方法で辞書管理システムに格納できるよう、辞書管理システムの拡張を行う。英語と並行して、日本語の構文パターンの収集と例文へのアノテーション作業の仕様を決め、日本語複文コーパスの作成を開始する。
複単語表現や構文パターンを利用して依存構造解析の性能を向上させるアルゴリズムの設計を実装を行う。平行して、複単語表現や構文パターンの半自動抽出に関する研究を進め、辞書項目の拡張と依存構造解析のカバレッジと性能向上を並行的に達成する方法について検討する。
並列構造を考慮したグラフに基づく依存構造解析アルゴリズムの効率的な実装を行い、実験により性能の評価を行う。並列構造解析の性能を評価するためのデータの収集を行い、正解がアノテーションされた評価コーパスを構築する。
並列構文解析の基本技術となる単語列や表現列のアラインメントに利用するため、単語や表現間の意味的な類似性を計算する方法を、分散表現を基盤とする手法に基づいて設計し、実験により評価を行う。

13.研究発表(平成26年度の研究成果)

(雑誌論文) 計(0)件 うち査読付論文 計(0)件

著者名		論文標題			
雑誌名	査読の有無	巻	発行年	最初と最後の頁	
掲載論文のDOI(デジタルオブジェクト識別子)					

(学会発表) 計(4)件 うち招待講演 計(0)件

発表者名		発表標題	
大内啓樹, Kevin Duh, 進藤裕之, 松本裕治		Supertagの曖昧性を考慮した依存構造解析	
学会等名	発表年月日	発表場所	
情報処理学会研究報告 第218回自然言語処理研究会	2014年09月01日～2014年09月02日	首都大学東京(東京都八王子市)	

発表者名		発表標題	
Xiaoyi Wu and Yuji Matsumoto		A Hierarchical Word Sequence Language Model	
学会等名	発表年月日	発表場所	
28th Pacific Asia Conference on Language, Information and Computation	2014年12月12日～2014年12月14日	Phuket (Thailand)	

発表者名		発表標題	
俵雄貴, 東藍, 松本裕治		係り受け情報を利用した日本語形態素解析	
学会等名	発表年月日	発表場所	
情報処理学会 第220回自然言語処理研究会	2015年01月19日～2015年01月20日	九州大学医学部百年講堂(福岡県福岡市)	

発表者名	発表標題	
駒井雅之, 進藤裕之, 松本裕治	英語の句動詞表現の同定とコーパス構築	
学会等名	発表年月日	発表場所
言語処理学会第21回年次大会	2015年03月17日～2015年03月19日	京都大学(京都府京都市)

〔図書〕計(0)件

著者名	出版社		
書名	発行年	総ページ数	

14. 研究成果による産業財産権の出願・取得状況

〔出願〕計(0)件

産業財産権の名称	発明者	権利者	産業財産権の種類、番号	出願年月日	国内・外国の別

〔取得〕計(0)件

産業財産権の名称	発明者	権利者	産業財産権の種類、番号	取得年月日	国内・外国の別
				出願年月日	

15.備考

--