様式 C-7-2

自己評価報告書

平成２１年　４月　１０日現在

研究種目：特定領域研究
研究期間：2006～2010
課題番号：18061005
研究課題名（和文）　書き言葉コーパスの自動アノテーションの研究

研究課題名（英文）　Research on Automatic Annotation of Written Language Corpora

研究代表者
松本　裕治（MATSUMOTO YUJI）
奈良先端科学技術大学院大学・情報科学研究科・教授
研究者番号：10211575

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：自然言語処理，言語学，機械学習，コーパス，アノテーション

１．研究計画の概要

言語学から言語処理研究にいたる様々な基礎・応用分野に役立つコーパスへのタグ付けを行うための支援環境を構築する．形態素，構文，意味，文脈情報等の様々なレベルのタグ付けを対象とし，(1)タグ付けの自動化と，(2)コーパスへタグ付けを行う際の効率や精度を管理・維持するための支援環境の構築という２つの次元から問題を整理する．前者については，単語分かち書き，品詞付与，文節や固有表現の解析，係り受け解析等の構文解析，用言および体言に対する項構造解析，照応解析等の指示対象の解析，文書構造や談話構造解析など，さまざまな言語情報についてのタグ設計およびタグ付け基準の設定を行う．異なるレベルの整合性を保ちつつ記述するための統合的なタグ付け方式の設計を行う．後者については，設計されたタグ方式に従ってコーパスを作成しつつ，タグ付きコーパスからの機械学習に基づいてタグ付けの自動化を行う言語解析システムの構築と、タグ付きコーパスを管理し，利用するための支援ツールの設計と開発を行う．

２．研究の進捗状況

コーパスに対する自動タグ付けツールの開発：日本語係り受け解析の性能向上を目指して，新たな学習方式を提案した．また，従来から問題である並列構造の範囲同定の問題に対して類似系列のアラインメントの自動化に基づく新たな手法を提案した．文節情報のタグ付けについて検討し，文節まとめ上げのための自動ツールの設計を行った．照応解析および述語項構造解析についても新たな手法を提案し，さらに，事態名詞の項構造解析への拡張を行った．事象間の時間関係解析のため，局所的情報と大域的情報の効果的な融合法を提案した．
コーパス管理ツールの開発：形態素・係り受け解析済みコーパスを管理し，検索等の利用機能を提供するコーパス管理ツールの設計と実装を行った．また，ネットワーク経由での利用が可能になるように拡張した．当初２年間で開発したツールの機能の再検討を行い，基本機能をドットネットフレームワーク上で再実装した．の実装を行った．さらに，検索結果表示の高速化や検索履歴を表示する機能を達成した．これとは別に，セグメントとリンクに基づく汎用のタグ付けツールを設計開発し，本研究で目標としている様々なタグのアノテーション作業に供するようにした．データ構造の設計の詳細化や多重に埋め込まれたタグの記述にも対応可能になるよう拡張した．
タグ付きコーパスの構築：述語項構造および照応情報を付与したコーパスを構築した．また，大規模な固有表現抽出を行ない，コアデータの一部に対して，固有表現タグ付け作業を行った．
談話構造アノテーションツールの開発：文関係，共参照，項構造を記述するためセマンティックエディタの設計と実装を行った．一般化された木構造表示ユーザインタフェースを実装し，談話構造や意味構造を編集する機能を実装した．

３．現在までの達成度

②おおむね順調に進展している

（理由）

日本語の基本的な言語解析である単語分

かち書き，品詞付与，文節解析，係り受け解析に対して，機械学習に基づき，実用レベルの精度での自動解析を行う手法の提案と自動解析ツールの開発を行うことができた．照応解析と項構造解析については実用レベルには達していないものの，現在のところ最高の性能を示す手法を提案している．並列構造解析など当初明確に計画していなかった問題についても，これまでの性能を上回る手法の提案を行うことができた．

タグ付きコーパス管理ツールについては，品詞，文節，係り受け解析済みのコーパスに特化したコーパス管理ツール「茶器」を実装し，第一バージョンを一般公開した．利用者からのフィードバックを受けて，新たなシステムとして再設計を開始している．汎用アノテーションツール SLAT の実装を行い，当初計画していた機能をほぼ実現し，機能の見直しと改良を行う段階に進んでいる．両ツールとも領域内外の研究者や学生向けの講習会を開催し，ツールの普及と利用者からの要望の収集に努めている．談話構造アノテーションツールについても，当初計画通り，文間，共参照，項構造を記述するためセマンティックエディタの実装を行うことができた．

４．今後の研究の推進方策

コーパスへのタグ付けツールおよびタグ付きコーパスの管理ツールの構築が進んできたため，今後は，領域内外に公開されているコアデータに対する種々のタグ付け作業を開始し，コアデータの一部をタグ付きコーパスとして公開するとともに，この作業を通じてツールの問題点の洗い出しとその解決，および，改良点の発見とその実装を行う予定である．また，タグ付け作業の経験に基づき，効率的なタグ付け作業方法と支援法を明確にしていく．特に，短単位への分割や文節解析については，コーパス班および電子化辞書班との連携を強化していく．

これまで個別に開発してきたコーパス管理ツールの統合をシステムレベルあるいはデータレベルで行い，種々のレベルのアノテーションが共存できる環境を構築していく．

５． 代表的な研究成果

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 9 件）

① 岩立将和，浅原正幸，松本裕治，トーナメントモデルを用いた日本語係り受け解析，自然言語処理，Vol.15, No.5, pp.169-185, 2008. 査読あり

② Takenobu Tokunaga, Chu-Ren Huang, Yat Mei Lee. Asian language resources: the state-of-the-art. Language Resources and Evaluation. Vol.42. No.2. pp.109-116. 2008, 査読あり

③ 橋本泰一，吉田恭祐，野口正樹，徳永健伸，田中穂積．関係データベースを用いた構文木付きコーパス検索手法．自然言語処理．Vol.14. No.4. pp.3-22. 2007. 査読あり

〔学会発表〕（計 29 件）

① Masakazu Iwatate, Masayuki Asahara, Yuji Matsumoto, Japanese dependency parsing using a tournament model, Proceedings of the 22nd International Conference on Computational Linguistics, pp.361-368. 2008.8.21, Manchester. 査読あり

② Tokunaga Takenobu, Dain Kaplan, Chu-Ren Huang, Shu-Kai Hsieh, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Kiyoaki Shirai, et al., Adapting International Standard for Asian Language Technologies. Proceedings of the Sixth International Language Resources and Evaluation, 2008.5.29, Marrakech, Morocco, 査読あり.

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕

本研究課題に関連した研究に対して，以下の賞を受賞した．

① 岩立将和，浅原正幸，松本裕治，言語処理学会第 14 回年次大会優秀発表賞，「トーナメントモデルを用いた日本語係り受け解析」2008 年 3 月

② 松本裕治, 2007 年度 日本 OSS 貢献者賞，「日本語形態素解析システム「茶筌(ChaSen)」の開発をはじめとした OSS への貢献」2007 年 10 月

③ 飯田龍，情報処理学会平成１９年度山下記念研究賞，「NAIST テキストコーパス：述語項構造と共参照関係のアノテーション」2007 年 8 月

④ 飯田龍，小町守，乾健太郎，松本裕治，言語処理学会第１３回年次大会優秀発表賞，「日本語書き言葉を対象とした述語項構造と共参照関係のアノテーション：NAISTテキストコーパス開発の経験から」2007 年 3 月