様式 C-19

科学研究費補助金研究成果報告書

平成２３年５月１日現在

機関番号：14603
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 of 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

研究成果の概要（和文）：日本語コーパスに対する様々な言語情報のアノテーションを自動的に行う言語解析ツールの開発，および，アノテーションの誤り修正やアノテーションを施されたコーパスの柔軟な利用や管理を行うためのコーパスツールの開発を行った．具体的には，形態素解析，係り受け解析，並列構造解析，固有表現認識，述語項構造解析，照応・共参照解析，事象間時間関係解析などの自動解析，および，これらのアノテーションを施したコーパスを構築した．

研究成果の概要（英文）：We developed various automatic annotation systems for Japanese corpora, as well as corpus annotation assistance tools for error correction of annotation and for flexible use of annotated corpora. The automatic annotation systems we developed range over morphological analysis, syntactic dependency analysis, coordination structure analysis, Named Entity recognizer, predicate argument structure analysis, anaphora and co-reference analysis, temporal relation analysis of events, and so on. We also developed annotated corpora with those information.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	12,000,000	0	12,000,000
2007 年度	19,700,000	0	19,700,000
2008 年度	21,200,000	0	21,200,000
2009 年度	20,000,000	0	20,000,000
2010 年度	18,800,000	0	18,800,000
総　計	91,700,000	0	91,700,000

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：コーパス，形態素解析，統語解析，述語項構造解析，アノテーション，自然言語解析，照応解析，固有表現認識，機械学習

１．研究開始当初の背景

言語学研究および言語処理研究には，大規模なテキストデータ，すなわち，コーパスが重要な役割を演じる．研究開始当初は，大規模な新聞記事データや著作権の切れた小説などの大規模電子化データが研究利用可能な形で存在したが，著作権の不明確なインターネット上の言語データや著作権によって保護された書籍や雑誌のデータを自由に研究利用することはできなかった．また，日本語形態素解析や係り受け解析などの自動解析ツールがいくつか公開されていたが，いくつかの品詞セットや異なる解析規準に基づいており，統一した処理が行われているわけではなかった．また，これらのツールは，自

然言語処理研究者には使用上の困難はなかったものの，解析システムの出力は決して扱い易い形になっておらず，プログラミングに不慣れな言語学者にとっては使い勝手のよいものとは言えない状況にあった．

このような状況を踏まえ，次のような目標を発想するにいたった．

(1)斉一の単語分かち書き規準に従った，日本語解析ツールを，形態素解析，係り受け解析に留まらず，当時考えられていた様々なレベルの言語解析を対象にした自動解析ツールを構築し，広く公開すること．

(2)言語学研究の基本コーパスとして資するため，かつ，上記の言語解析システムを機械学習に基づいて構築するための学習データとなるアノテーション付きコーパスを公開すること．

(3)自動あるいは人手によってアノテーション付けされたコーパスを，プログラミング経験のない言語学研究者にも柔軟に検索・利用できるツールを構築することによって，容易に使いこなせることを目標としたコーパス利用ツールを構築すること．

２．研究の目的

本研究プロジェクトで開発する日本語書き言葉コーパスを有効に利用するために，様々な情報を付与（タグ付け）することが重要である．本研究の目的は，言語学から言語処理研究にいたる様々な基礎・応用分野に役立つアノテーションを行うためのコーパス解析・利用のための支援環境を構築することとした．本研究で行うアノテーションは，形態素，統語構造，意味，文脈情報等の様々なレベルのタグ付けを対象とした．その際，(1)アノテーションの種類と規準の設定，(2)実際に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つの次元から整理することとした．前者については，単語分かち書き，品詞付与，活用や派生等の語の内部構造の解析，係り受けや句構造，並列構造等の統語構造解析，用言だけでなく体言に対する項構造解析，照応解析等の指示対象の解析，事象間の時間関係解析など，現在考えられているほとんどすべての言語情報についてのアノテーション設計とアノテーション付与の規準の設定を行うことを目標とした．

このような異なるレベルの情報を整合性を保ちつつ記述するための統合的なアノテーション方式の設計を行うこととした．後者については，設計されたアノテーション方式に従ってコーパスを作成しつつ，アノテーション付きコーパスからの機械学習に基づいてコーパスへのアノテーションを行う言語解析システムの構築を行うことを目指した．

また，アノテーションが付与されたコーパスを管理し，タグ付け誤りの発見や修正機能を有し，アノテーション付きコーパスの整合性を維持しつつ管理する支援ツールの設計と開発を行うこととした．

３．研究の方法

本研究では，形態素，構文，意味，文脈情報等の様々なレベルのアノテーションを考えており，各レベルのアノテーション規準を設定することにした．研究計画で予定していた種々の自動解析システムおよびアノテーションが施されたコーパスの作成は，研究分担者で手分けして開発することとした．単語，形態素構造，句チャンキング，文節係り受け構造解析システムと同アノテーションコーパスは，松本・浅原が担当した．固有表現認識システムと同アノテーションコーパスは，橋本が担当した．述語項構造解析，照応解析，共参照関係解析は，乾・小町が担当した．述語間の時間関係解析は，浅原が担当した．

コーパスへのアノテーション作業とアノテーション付けされたコーパスを利用するためのツールとして，形態素解析，係り受け解析済みコーパスに特化したコーパス管理・利用ツールを松本・浅原が開発した．コーパスへのアノテーションをセグメントへのアノテーションとセグメント間の関係のアノテーションという形で一般化した汎用アノテーションツールを徳永が開発することとした．また，複数のコーパス解析ツール，コーパス管理ツールを相互運用するための汎用ツールの開発を橋田が行うこととした．

分担研究者間の意志統一およびツール間の整合性の維持を円滑に行うため，数ヶ月毎にツール班会議を開催することとした．また，他班の研究者へのツールの普及を促進するため，適宜ツール講習会を開催することとした．

４．研究成果

本特定領域研究で構築された日本語コーパスへの自動アノテーションツールとしては，予定していた言語解析ツールを一通り開発した．

日本語コーパスの一部であるコアデータに対して，種々のアノテーションを付与したタグ付きコーパスを構築した．形態素情報および長単位（文節情報）の付与については，コーパス班が担当することになっていたので，ツール班では，形態素・文節情報より上のアノテーションを担当し，そのための様々な言語解析ツール，アノテーション支援ツールの構築，および，コーパスへの具体的なアノテーション作業を実施した．

構築したツールの主なものは，自動言語解析ツールとしては，日本語係り受け解析，固有表現解析，述語項構造解析，照応・共参照解析，モダリティ解析ツールがあり，これらの解析ツールの構築を機械学習の利用によって行うため，および，性能評価のため，それぞれに対応するアノテーション付きコーパスを構築するとともに，自然言語解析ツールとして実装した．コアデータ（100 万語）のすべてに対しては，固有表現のアノテーションが完了したが，その他のアノテーションについては，コアデータの一部に留まった．当初最優先に考えていた係り受け解析は，文節定義や係り受け情報の規準設定の遅れなどによりコアデータの７割程度にしか人手によるチェックを行えなかったが，並列構造に関しては，全体の確認を修了した．係り受け解析については，プロジェクト終了後も継続する計画であり，2011 年度前半にはコアデータ全体のアノテーションを完了させることを考えている．述語項構造解析およびそれより上位の解析については，コアデータの感性を待つ前に，新聞記事データへのアノテーション作業を開始し，約 100 万語規模のコーパスへのアノテーション作業を行った．今後，コアデータへの自動アノテーションと修正作業を行うことを考えている．

コーパスアノテーションの支援ツールとしては，形態素，文節，係り受け解析に特化したコーパス管理ツール「茶器」を開発した．種々の検索機能，誤り修正機能，統計処理機能を有しており，自動アノテーションを施したコーパスの柔軟な利用，および，アノテーション情報の洗練を行うことが可能である．当初は予定していなかった並列構造のアノテーション機能や文末情報の挿入・削除機能などを実装し，形態素解析，文節まとめ上げ，並列構造，係り受け解析などに関するアノテーションのすべてを単独で取り扱うことができるシステムとして完成させた．

コーパスへのアノテーションは，コーパスの一部（セグメント）へのラベル付け，および，セグメント間の関係（リンク）へのラベル付けの２種類に大別できる．コーパスへのすべてのアノテーションをこれらの汎用的なアノテーションとして実現した汎用コーパスアノテーションツール「Slate」をWebブラウザ上で動作可能なシステムとして構築した．セグメント名やリンク名などコーパスに依存する情報は利用者が定義することができ，自然言語処理で考えられているほとんどすべてのアノテーション作業が可能なツールとなっている．Slateと茶器のデータベーススキーマは統一的な規準で定義されており，例えば，茶器でアノテーションを行ったコーパスをSlateで取り扱うことが可能となるように設計されている．

また，様々なタグ付きコーパスやコーパス構築支援ツールの相互運用を可能にするためのツールを設計し，構築した．

５．主な発表論文等

（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 13 件）

① 飯田龍，小町守，井之上直也，乾健太郎，松本裕治．「述語項構造と照応関係のアノテーション：NAISTテキストコーパス構築の経験から」 自然言語処理，Vol.17，No.2，25-50，2010，査読有

② Ai Azuma and Yuji Matsumoto，“A generalization of forward-backward algorithm，” Transactions of the Japanese Society for Artificial Intelligence，Vol.25，No.3，.494-503，2010，査読有

③ 渡邉陽太郎，浅原正幸，松本裕治．「述語語義と意味役割の結合学習のための構造予測モデル」．人工知能学会論文誌，Vol.25，No.2，.252-261，2010，査読有

④ 小町守，飯田龍，乾健太郎，松本裕治．「名詞句の語彙統語パターンを用いた事態性名詞の項構造解析」．自然言語処理，Vol.17，No.1，.141-159，2010，査読有

⑤ 吉川克正，Sebastian Riedel，浅原正幸，松本裕治．「Markov Logicを利用した時間的順序関係の同時推論」．人工知能学会論文誌，Vol.24，No.6，.521-530，2009，査読有

⑥ Vera Sheinman，Takenobu Tokunaga．“AdjScale：Visualizing differences between adjectives for language learners，” IEICE Transaction of Information and Systems，Vol. E92-D，No.8，1542-1550，2009，査読有

⑦ 松本裕治，大山浩美．「言語処理による作文支援・語彙学習への可能性について」．日本語教育「特集 作文教育のための語彙研究」，Vol.140，37-47，日本語教育学会，January 2009，査読有

⑧ 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.5，169-185，2008，査読有

⑨ Takenobu Tokunaga，Chu-Ren Huang，Yat Mei Lee．“Asian language resources：the state-of-the-art，” Language Resources and Evaluation，Vol.42，No.2，109-116，2008，査読有

⑩ 渡邉陽太郎，浅原正幸，松本裕治．“グラフ構造を持つ条件付確率場によるWikipedia文書中の固有表現分類，” 人工知能学会論文誌，Vol.23，No.4，245-254，2008，

査読有
⑪ 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, 3-22, 2007, 査読有
⑫ Ryu Iida, Kentaro Inui, Yuji Matsumoto. “Zero-anaphora resolution by learning rich syntactic pattern features,” ACM Transactions on Asian Language Information Processing (TALIP), Vol 6, Issue 4, Article 12, 2007, 査読有
⑬ Chu-Ren Huang, Takenobu Tokunaga, Sohpia Yat Mei Lee. “Asian language processing: current state-of-the-art,” Language Resources and Evaluation, Vol.40, No.3-4, 203-218, 2006, 査読有

〔学会発表〕(計 29 件)

① Tokunaga Takenobu, Yasuhara Masaaki, Terai Asuka, David Morris, Anja Belz. ″Construction of bilingual multimodal corpora of referring expressions in collaborative problem solving″, Proceedings of the Eighth Workshop on Asian Language Resources, 38-46, Beijing, China, 2010.08.21, 査読有
② Yotaro Watanabe, Masayuki Asahara, Yuji Matsumoto, “A structured model for joint learning of argument roles and predicate senses,” Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 98-102, 2010.07.12, 査読有
③ Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee and Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, David Traum, “Towards an ISO Standard for Dialogue Act Annotation,” Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta, 2010.05.21, 査読有
④ Dain Kaplan, Ryu Iida, Takenobu Tokunaga, “Annotation Process Management Revisited,” Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta, 3654-3661, 2010.05.20, 査読有
⑤ Ai Azuma, Yuji Matsumoto. “A Generalization of Forward-backward Algorithm,” In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Bled, Slovania, 99-114, 2009.09.08, 査読有
⑥ Tokunaga Takenobu, Dain Kaplan, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Virach Sornlertlamvanich, Thatsanee Charoenporn, Xia Yingju, Chu-Ren Huang, Shu-Kai Hsieh, Shirai Kiyoaki. “Query Expansion using LMF-Compliant Lexical Resources,” Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, Singapore, 145-152, 2009.08.07, 査読有
⑦ Ryu Iida, Kentaro Inui, Yuji Matsumoto. “Capturing Salience with a Trainable Cache Model for Zero-anaphora Resolution,” In Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009), Singapore, 647-655, 2009.08.04, 査読有
⑧ Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, Yuji Matsumoto. “Jointly Identifying Temporal Relations with Markov Logic,” In Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009), Singapore, 405-413, 2009.08.03, 査読有
⑨ Vera Sheinman, Tokunaga Takenobu. ″AdjScale: Differentiating between similar adjectives for language learners″, Proceedings of 1st International Conference on Computer Supported Education (CSEDU 2009), INSTICC Press, Lisbon, Portugal, 229-235, 2009.03.24, 査読有
⑩ Masakazu Iwatate, Masayuki Asahara and Yuji Matsumoto. “Japanese dependency parsing using a tournament model,” In Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008), Manchester, UK, 361-368, 2008.08.21, 査読有
⑪ Tokunaga Takenobu, Dain Kaplan, Chu-Ren Huang, Shu-Kai Hsieh, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Kiyoaki Shirai, others. “Adapting International Standard for Asian Language Technologies,” Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008.05.28, 査読有
⑫ Masaki Noguchi, Kenta Miyoshi,

Takenobu Tokunaga, Ryu Iida, Mamoru Komachi, Kentaro Inui. “Multiple Purpose Annotation using SLAT - Segment and Link-based Annotation Tool -,” Proceedings of 2nd Linguistic Annotation Workshop, Marrakech, Morocco, 61-64, 2008.05.27, 査読有

⑬ Kiyoaki Shirai, Takenobu Tokunaga, Chu-Ren Huang, Shu-Kai Hsieh, Tzu-Yi Kuo, Virach Sornlertlamvanich, Thatsanee Charoenporn. “Constructing Taxonomy of Numerative Classifiers for Asian Languages,” Proceedings of the Third International Joint Conference on Natural Language Processing, Hyderabad, India, 397-402, 2008.01.10, 査読有

⑭ Mamoru Komachi, Ryu Iida, Kentaro Inui and Yuji Matsumoto. “Learning Based Argument Structure Analysis of Event-nouns in Japanese,” Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING), Melbourne, Australia, 2007.09.19, 査読有

⑮ Ryu Iida, Mamoru Komachi, Kentaro Inui, Yuji Matsumoto. “Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations,” ACL Workshop `Linguistic Annotation Workshop', Prague, Czech Republic, 2007.06.29, 査読有

⑯ Yotaro Watanabe, Masayuki Asahara, Yuji Matsumoto. “A Graph-Based Approach to Named Entity Categorization in Wikipedia Using Conditional Random Fields”, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 2007.06.29, 査読有

⑰ Koiti Hasida, Noriaki Izumi, Akira Mori. “CBTO: Compositional Business-Task Organization,” W3C Workshop on Declarative Models of Distributed Web Applications, Dublin, Ireland, 2007.06.05, 査読有

〔図書〕（計１件）

松本裕治, 乾健太郎, 徳永健伸, 他６名「言語と情報科学」, 朝倉書店, 2011 (in press)

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

〔その他〕
ホームページ等（関連ツールの情報）
茶器：
http://sourceforge.jp/projects/chaki/
Slate：
http://www.cl.cs.titech.ac.jp/slate/
相互運用ツール：
http://u-compare.org/japanese.html
拡張固有表現コーパス：
http://riverstone.star.titech.ac.jp/taiichi/tokutei/ene/
照応・述語項構造コーパス：
http://cl.naist.jp/nldata/bccwj/

６．研究組織
(1)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

(2)研究分担者
徳永　健伸（TOKUNAGA TAKENOBU）
東京工業大学大学院・情報理工学研究科・教授
研究者番号：20197875

乾　健太郎（INUI KENTARO）
東北大学大学院・情報科学研究科・教授
研究者番号：60272689

橋田　浩一（HASIDA KOITI）
独立行政法人産業技術総合研究所・サービス工学研究センター・次長
研究者番号：00357766

浅原　正幸（ASAHARA MASAYUKI）
奈良先端科学技術大学院大学・情報科学研究科・助教
研究者番号：80379528

橋本　泰一（HASHIMOTO TAIICHI）
東京工業大学・総合プロジェクト支援センター・特任准教授
研究者番号：10345382

小町　守（KOMACHI MAMORU）
奈良先端科学技術大学院大学・情報科学研究科・助教
研究者番号：60581329