様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ８ 日現在

機関番号：１４３０１
研究種目：基盤研究(B)
研究期間：2012～2014
課題番号：２４３０００５３

研究課題名（和文）多様なテキストへの高次アノテーションに基づく文脈理解モデルの明確化

研究課題名（英文）Modeling Context Analysis based on Semantic Annotation on Diverse Documents

研究代表者
黒橋　禎夫（Kurohashi, Sadao）

京都大学・情報学研究科・教授

研究者番号：５０２６３１０８

交付決定額（研究期間全体）：（直接経費）　　11,800,000円

研究成果の概要（和文）：自然言語の意味解析の研究には意味関係を付与したコーパスが必要であるが，従来の意味関係のタグ付きコーパスは新聞記事を中心に整備されてきた．しかし，文書には多様なジャンル，文体が存在し，その中には新聞記事では出現しないような言語現象も出現する．本研究では，従来のタグ付け基準では扱われてこなかった現象に対して新たなタグ付け基準を設定し，ウェブを利用することで多様な文書の書き始めからなる意味関係タグ付きコーパスを構築した．さらに，2段階のクラウドソーシングにより，談話関係タグ付きコーパスを構築する手法を考案した．

研究成果の概要（英文）：To study semantic analysis of natural language texts, a corpus annotated with semantic relations is required. Although existing corpora annotated with semantic relations have been restricted to newspaper articles, there are texts of various genres and styles containing linguistic expressions that are missing in newspaper articles. In this research, we defined annotation criteria for linguistic phenomena which have not been treated using existing criteria. We built a diverse document leads corpus annotated with semantic relations. We also proposed a novel approach for rapidly developing a corpus with discourse annotations using crowdsourcing.

研究分野：自然言語処理

キーワード：自然言語処理　文脈理解　談話解析　照応解析　コーパス

１．研究開始当初の背景

（1）情報爆発，情報洪水が叫ばれる今日，我々は膨大なコストをかけて情報の入手・取捨選択を行っており，計算機に情報の組織化や推論の能力を持たせることへの期待が高まっている．また，情報機器の氾濫とともに高齢者などのデジタル・ディバイドの問題も深刻化しており，情報機器が自然言語インタフェースを備えることへの要望も高い．

（2）これらを実現する基盤技術は自然言語処理，すなわち計算機によって日本語や英語などの自然言語の構造と意味を解析・理解する技術である．情報を伝える単位は，単語や文ではなく文章であるにも関わらず，文章の解析は高い精度では実現されていなかった．

２．研究の目的

計算機による自然言語の形態素・構文解析は，コーパスに言語解釈をタグとして付与し，学習・評価を行うという方法論により，90 年代以降に急速に進展した．しかし，言語が情報を伝える単位である文章の解析，すなわち文脈処理については，複雑さやコストの問題からタグ付与コーパスを構築することが進展せず，研究が停滞している状況にある．本課題では，ウェブから多様な文章を収集し，書き始めの 3 文程度に対して照応関係と談話関係のアノテーション（タグ付与）を行い，これによって文脈理解モデルの明確化を行うことを目的とする．3 年間で 5,000 記事（15,000 文）のコーパスを構築し，研究コミュニティに広く公開する．

３．研究の方法

言語情報を付与する文章は，ウェブから大量のテキストを収集することにより多様性を確保する．まず，共参照，述語の項の省略，名詞橋渡し照応などの照応関係の基準を明確化した上で，5,000 記事，15,000 文のアノテーションを行う．さらに，主題連鎖，理由，例示などの談話関係について，照応関係との関係性を考慮しつつアノテーション基準を設定し，同一の 5,000 記事についてアノテーションを行う．コーパス構築においてはアノテーションの一貫性を高めることがもっとも重要であるため，作業ツールの構築・修正，実験的アノテーションの実施，作業者との十分な打ち合わせ，自動解析システムとの差分のチェックなどを行いながらコーパス構築を進める．

４．研究成果

（1）多様なウェブページの先頭 3 文からなるコーパス，5,000 文書に対して，形態素・構文情報と，著者・読者等の外界ゼロ照応を含む省略・照応情報に関するアノテーションを専門家によって行い，省略・照応関係コーパスを完成させた．また，重要な論理関係として根拠・条件と転換の 2 種類の談話関係に注目し，上述の 5,000 文書を含む 10,000 文書(30,000 文)に対して，談話関係の有無の判定とタイプの判定を 2 段階で行うクラウドソーシングにより談話関係アノテーションを行った．このコーパスを京都大学ウェブ文書リードコーパス(Kyoto University Web Document Leads Corpus)と名付け，研究代表者の研究室ウェブページで公開した．

（2）これまでの日本語の意味関係解析の研究で主に用いられてきたのは新聞記事コーパスであった．しかし，テキストには新聞記事以外にも百科事典や日記，小説など多様なジャンルがある．これらの多様なテキストの中には依頼表現，敬語表現など新聞記事ではあまり出現しない言語現象が出現し，意味関係と密接に関係している．例えば例（1）の「買いたかった」の動作主が著者となることは意志表現に，「教えてください」の動作主が読者，受け手が著者になることは依頼表現に密接に関係している．

> 例（1）今日はソフマップ京都に行きました。時計を買いたかったのですが、この店舗は扱っていませんでした。時計を売っているお店をコメントで教えてください。

このような言語現象と意味関係の関係を明らかにするためには，多様なテキストからなるタグ付きコーパスの構築とその分析が必要となる．そこで本研究ではニュース記事，百科事典記事，blog，商用ページなどを含むウェブページをタグ付け対象として利用することで，多様なジャンル，文体の文書からなる意味関係タグ付きコーパスの作成を行った．

ウェブに存在する文書には，コーパスとして利用するには不適切な文書も多数存在している．そこで，テキストのみでは内容の理解が困難な文書，過度にくだけた文体でタグ付けが困難な文書を，まず簡単なルールで自動フィルタリングし，その後残った文書を人手で確認しコーパスとして適切な文書についてのみタグ付けの作業を行うこととした．

多様な文書を含むタグ付きコーパスの構築を行うためには，多数の文書に対してタグ付け作業を行う必要がある．この際，1 文書あたりの作業量が問題となる．形態素，構文関係のタグ付けは文単位で独立であり，文書が長くなっても作業量は文数に対して線形にしか増加しない．一方，意味関係のタグ付けでは文をまたぐ関係を扱うため，文書が長くなると作業者が考慮すべき要素が組み合わせ的に増加する．このため 1 文書あたりの作業時間が長くなり，文書全体にタグ付けを行うと，タグ付けできる文書数が限られてしまう．そこで，先頭の 3 文に限定してタグ付けを行うことで 1 文書あたりの作業量を抑える．意味関係解析では既に解析した前方の文の解析結果を利用する場合があり，先頭の解

析誤りが後続文の解析に悪影響を与える．先頭数文に限定したコーパスを作ることで，文書の先頭の解析精度を上げることが期待でき，全体での精度向上にも寄与できると考えられる．

（3）本研究のタグ付け対象には新聞記事ではあまり出現しない言語現象が含まれる．その中でも特に重要なものとして文章の著者・読者の存在が挙げられる．著者や読者は，省略されやすく，モダリティや敬語などと密接に関係するなど，他の談話要素とは異なった振る舞いをする．新聞記事では，客観的事実を報じる内容がほとんどのため，社説を除くと記事の著者や読者が談話中に出現することはほとんどない．そのため，従来のタグ付け基準では 著者や読者などを外界の照応先として定義していたが，具体的なタグ付け基準についてはあまり議論されてこなかった．一方，本研究で扱うウェブではブログ記事や通販ページ，マニュアルなど著者や読者が談話中に出現する文書が多く含まれ，その中には従来のタグ付け基準では想定していなかった言語現象および意味関係が出現する．そのため，著者・読者が出現する文書でのタグ付け上の問題点を分析し，タグ付け基準を設けることが重要となる．

著者・読者が出現する文書へのタグ付けにおける１つ目の問題は，文章中で著者・読者に対応する表現である．例（2）では，「僕」は著者に対応し，「皆さん」は読者に対応した表現となっている．

例（2）僕は京都に行きたいのですが，皆さんのお勧めの場所があったら教えてください。

本研究ではこのような著者や読者に対応する表現を著者表現，読者表現と呼ぶこととした．著者表現，読者表現は外界ゼロ照応における著者や読者と同様に談話中で特別な振る舞いをする．例えば例（2）の「教えてください」のように，依頼表現の動作主は読者表現に，依頼表現の受け手は著者表現になりやすい．本研究で扱う文書は多様な著者，読者からなり，著者読者，読者表現も人称代名詞だけでなく，固有表現や役割表現など様々な表現で言及され，語の表層的な情報だけからは簡単に判別できない．そこで本研究では著者表現，読者表現をタグ付けし，著者・読者の談話中での振る舞いを明らかにした．

2 つ目の問題は項を明示していない表現に対する述語項構造のタグ付けである．日本語では一般的な事柄に対して述べる場合には，動作主や受け手などを明示しない表現が用いられることが多い．従来の新聞記事を対象としたタグ付けでは，[不特定-人]を動作主などとすることでタグ付けを行ってきた．一方，著者・読者が談話中に出現する場合には，一般的な事項について述べる場合でも動作主などを著者や読者と解釈できる場合が存在する．例（3）の「公開する」の動作主であるガ格は，不特定の人が行える一般論であるが，著者自身の経験とも読者が将来する行為とも解釈することができ，作業者の解釈によりタグ付けに一貫性を欠くこととなる．

例（3）ブログに記事を書き込んで、インターネット上で公開するのはとても簡単です。

本研究ではこのような曖昧性が生じる表現を分類し，タグ付けの基準を設定した．

（4）文章中の談話関係について、アノテーション手続きを簡単化することにより，クラウドソーシングによる談話関係タグ付きコーパス構築を行った．関係をもつスパン（談話関係の基本単位）を同定するのは高コストであるので，自動分割した節を単位とし，この区切りは高精度であるので修正しないこととした．タグ付け対象の各文書は 3 文からなるもので，そこに含まれる節は 5 節までとして簡単化した．談話関係タイプの付与については，一回に任意の節間に成り立つ関係を判定するのはクラウドソーシング上では困難であるので，談話関係有無の判定と談話関係タイプの判定の 2 段階クラウドソーシングで行った．談話関係のタグセットとしては，Penn Discourse Treebank をベースに簡単化し，2 階層からなるタグセットを設計した．上位タイプは，“根拠・条件”，“転換”，“その他”の 3 タイプからなる．このように簡単化した問題設定ではあるが，ここで得られた知見やモデルは一般的な談話関係解析に発展させることが十分に可能である．

このような方法により，10,000 文書(30,000 文)からなる談話関係タグ付きコーパスを 2 段階のクラウドソーシングによって作成した．クラウドソーシングに要した時間は合計 8 時間弱であり，従来の大規模な談話関係アノテーションと比べて非常に高速に構築することができた．

また，得られた談話関係アノテーションを用いて，談話関係解析システムを構築した．談話関係解析の機械学習モデルを作成し，交差検定によってその精度を評価した．その結果，“根拠・条件”タイプについて 37.9%の F 値で解析できることがわかった．これは，英語の談話関係解析器の精度と近い値であり，得られたコーパスの有用性が示されたと考えられる．

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 2 件）
①萩行正嗣，河原大輔，黒橋禎夫，外界照応および著者・読者表現を考慮した日本語ゼ

ロ照応解析，自然言語処理，Vol.21， No.3，pp.563-600，2014，査読有.
http://doi.org/10.5715/jnlp.21.563

②萩行正嗣，河原大輔，黒橋禎夫， 多様な文書の書き始めに対する意味関係タグ付きコーパスの構築とその分析,自然言語処理，Vol.21，No.2，pp.213-248，2014，査読有.
http://doi.org/10.5715/jnlp.21.213

〔学会発表〕（計 7 件）
①澤田晋之介，小浜翔太郎，河原大輔，黒橋禎夫，クラウドソーシングによる確率的アノテーションを利用した談話関係解析，情報処理学会 第 77 回全国大会，京都大学(京都府・京都市)，2015.3.19.

② Daisuke Kawahara，Yuichiro Machida，Tomohide Shibata，Sadao Kurohashi，Hayato Kobayashi and Manabu Sassano，Rapid Development of a Corpus with Discourse Annotations using Two-stage Crowdsourcing，In Proceedings of the 25th International Conference on Computational Linguistics (COLING2014)， pp.269-278， Dublin (Ireland)，2014.8.25.

③河原大輔，町田雄一郎，柴田知秀，黒橋禎夫，小林隼人，颯々野学，2 段階のクラウドソーシングによる談話関係タグ付きコーパスの構築，情報処理学会 第 217 回自然言語処理研究会，オホーツク・文化交流センター（北海道・網走市)，2014.7.4.

④萩行正嗣，河原大輔，黒橋禎夫，著者・読者表現および外界ゼロ照応を考慮したゼロ照応解析，言語処理学会 第 20 回年次大会，pp.721-724，北海道大学（北海道・札幌市)，2014.3.20.

⑤Masatsugu Hangyo，Daisuke Kawahara and Sadao Kurohashi，Japanese Zero Reference Resolution Considering Exophora and Author/Reader Mentions，In Proceedings of EMNLP 2013: Conference on Empirical Methods in Natural Language Processing，pp.924-934，Seattle (USA)，2013.10.19.

⑥ Masatsugu Hangyo，Daisuke Kawahara，Sadao Kurohashi，Building a Diverse Document Leads Corpus Annotated with Semantic Relations，In Proceedings of 26th Pacific Asia Conference on Language Information and Computing，pp. 535-544，Bali (Indonesia)，2012.11.8

⑦萩行正嗣，河原大輔，黒橋禎夫，多様な文書の書き始めに対する意味関係タグ付きコーパスの構築，情報処理学会 第 206 回自然言語処理研究会，東京工業大学（東京都・目黒区)，2012.5.10.

〔図書〕（計 1 件）
黒橋禎夫，自然言語処理，放送大学教育振興会，195 ページ，2015.

〔その他〕
ホームページ:
http://nlp.ist.i.kyoto-u.ac.jp/index.php?KWDLC

６．研究組織
(1)研究代表者
黒橋 禎夫（KUROHASHI，Sadao）
京都大学・情報学研究科・教授
研究者番号：50263108

(2)研究分担者
河原 大輔（KAWAHARA，Daisuke）
京都大学・情報学研究科・准教授
研究者番号：10450694

柴田 知秀（SHIBATA，Tomohide）
京都大学・情報学研究科・助教
研究者番号：70452315
(2014 年 3 月まで)