様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　５月　２４日現在

機関番号：１７１０２
研究種目：基盤研究（B）
研究期間：2008～2011
課題番号：２０３２００８２

研究課題名（和文）　Web上からの母語話者／非母語話者英語論文コーパスの作成・公開とその利用

研究課題名（英文）　Building a Native/Non-native English Language Technical Paper Corpus from Web and its Release and Application

研究代表者
冨浦　洋一（TOMIURA YOICHI）
九州大学・システム情報科学研究院・教授
研究者番号：１０２１７５２３

研究成果の概要（和文）：
個人Webページで公開されている英語科学技術論文を，Web検索エンジンを用いて収集する手法，及び，品詞列の特徴に基づいて文書の英語の質を推定する統計的手法を開発した．さらに，これらを組み合わせ，Web上から各論文の英語の質情報を付与した大規模な英語科学技術論文コーパスを構築するシステムを開発した．また，Web上の文書を利用したコーパスの構築・公開に関して，著作権法上の問題や注意事項を検討した．

研究成果の概要（英文）：
We developed a method for collecting English language technical papers on the private web pages using web search engine and a statistical method for estimating the English quality of a document based on the characteristics about the sequences of part of speeches in the document. Furthermore, using these methods, we developed a system to build a large-scale English language technical paper corpus from Web, which includes the information about English quality for each paper. We also investigated copyright problems and what we should consider on building a corpus form Web and releasing it.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	3,300,000	990,000	4,290,000
２００９年度	2,700,000	810,000	3,510,000
２０１０年度	1,600,000	480,000	2,080,000
２０１１年度	1,900,000	570,000	2,470,000
年度			
総　計	9,500,000	2,850,000	12,350,000

研究分野：自然言語処理
科研費の分科・細目：言語学・外国語教育
キーワード：コーパス，Ｗｅｂ，英文の質判定，仮説検定，英作文支援，英語教育，著作権

１．研究開始当初の背景

日本人のような英語非母語話者に対する，英語学習の補助的な教材の開発や英文書作成支援システムの開発を行う際，言語資源（データ）としてWeb上の英文書を利用することは，その量および内容の豊富さから見て非常に有望である．言語資源としては，母語話者が書いた程度に良質な英文書（母語話者

文書）と非母語話者が書いた誤りや不自然さを含む英文書（非母語話者文書）双方が大量に必要となる．

研究代表者は既に，品詞列の情報を基に英文書の英文の質を推定する手法をいくつか開発しており，これらを基にした，さらに高精度な推定手法の開発と評価に取り組んでいた．

２．研究の目的

Web 上から科学技術論文（PDF ファイル）を収集し，各論文にその英語の質に関する情報を付与した大規模なコーパスを構築する手法を開発するとともに，構築したコーパスの著作権などを侵害しない公開方法について検討する．また，得られたコーパスの利用例として，非母語話者が犯しがちな不自然な表現の収集，非母語話者文書に固有の文法的・語彙的特徴の抽出等を行ない，これらの公開法を検討し，公開する．

３．研究の方法

Web 上から英語科学技術論文を収集するシステムを作成し，実際に us ドメインと jp ドメインを対象としてそれぞれ 500 論文程度を収集する．収集した論文の英語の質を英語母語話者に判定してもらい（英語論文校正会社に依頼），英語の質の情報が付与された小規模な英語科学技術論文コーパスを作成する．これを用いて，英文書の英語の質の推定システムを構築するとともに，その性能を評価する．

Web 上からの英語科学技術論文収集システムと英語の質の推定システムを組み合わせ，英語の質情報を付与した大規模な英語科学技術論文コーパスを構築するシステムを開発し，実際にコーパスの構築を行なう．

並行して，知的財産権等が専門の研究分担者を中心として，コーパス構築方法が著作権を侵害しないか，また，公開は可能か，可能とすればどのような公開法が良いか等を検討する．

さらに，構築されるコーパスを想定して，英文書作成支援システムへの利用や英語非母語話者に固有の文法的・語彙的特徴の抽出等を検討しておき，ある程度大規模なコーパスが構築された後，それらを実施する．

４．研究成果

(1) Web 上からの論文収集システム

個人 Web ページで公開されている英語科学技術論文を，Web 検索エンジンを用いて収集する手法を考案した．２ヶ月間の実働で，us ドメインから約６万，jp ドメインから約２万の論文を収集できた．

(2) 英語の質推定システム

Web 上から収集した 993 編の論文の英文の質を英語論文校正会社に依頼して判定してもらい，このデータを基に，英語論文を，英語母語話者レベルの英文書（母語話者文書）と英語非母語話者レベルの英文書（非母語話者文書），およびグレーゾーンの文書に分けた．母語話者文書と非母語話者文書を用いて，英語の質推定システムを構築するとともに，性能を評価した．精度を重視するようにメタパラメタを設定した場合で，母語話者文書と推定する場合の精度，再現率はそれぞれ 94%，25%であり，非母語話者文書と推定する場合の精度，再現率はそれぞれ 92%，22%であった．

また，性能向上のための学習データの増強を目的として，英語論文校正会社に依頼して，新たに 550 編の論文の英語の質判定を行なった．今後，学習データを増強したものに対して，性能評価を行ない，ツールキットを公開する予定である．

さらに，品詞列とは異なる特徴を利用した母語話者文書／非母語話者文書の判別器と組み合わせることにより，判別性能の向上が期待できる．その試みとして，研究協力者（小林雄一郎（大阪大学・大学院言語文化研究科博士後期課程）により，談話標識の分布の相違に基づく判別実験を行った．約８割の精度を得ており，２つの判別器を統合することにより，判別性能の向上が期待できる．

(3) コーパスの構築

上記(1)(2)のシステムを組み合わせ，英文の質情報付き科学技術論文コーパスを構築するツールキットを作成し，実際にコーパスの構築を行なった．構築したコーパスは，現時点で，約８万編の英語科学技術論文から成る．この内，前述の(2)と同様の精度を重視するようにメタパラメタを設定した場合に，母語話者文書と推定されるものは，約 14,000 編，非母語話者文書と推定されるものは約 3,100 編である．

(4) 著作権上の検討

コーパス構築のための Web 上の論文の複製は，「私的使用のための複製」には該当しないと考えられるため，複製には著作権者の許諾を必要とする．一方，著作権法の改正（平成 22 年 1 月に施行）があり，「情報解析のための複製」などの権利制限が盛り込まれた．検討の結果，情報解析を行おうとする者が論文を収集することは，「情報解析のための複製」に該当すると考えられるが，収集して作成したコーパスを公開（あるいは譲渡）

するには，やはり著作権者に許諾を取る必要があると考えられる．

また，コーパスを利用して作成した英語学習の補助的な教材などは，その内容次第で新たな著作物と考えられるか否かが異なるため，一概に公開の可否を決めることはできない．内容を検討して判断するとともに，たとえ，著作権を侵害しない場合でも，不自然な表現集のような場合は，著者の個人的感情にも配慮する必要がある．

(5) コーパス構築ツールキットの公開

上記で述べたように，情報解析のために構築したコーパスであっても，それを著作権者の許諾なしに公開することはできない．そこで，構築したコーパスを公開する代わりに，上記(3)のコーパス構築ツールキットを公開することとした．すでに，マニュアルや契約書を整備しており，増強した学習データを使用した場合の精度等を調査した後，2012 年度を目処に，研究代表者の研究室 Web ホームページにツールキット入手方法等の情報を掲載する予定である．

著作権法 47 条の 7 で，情報解析のための複製は許されることとなったが，構築したコーパスの公開等を考慮すると，まだ権利制限が不十分と考えられる．誰のどのような権利を守るかは非常に難しい問題であるが，今後，著作権法の更なる改正について検討が進むことを期待する．

(6) コーパスを利用した研究

現在のコーパスの規模では，不自然な表現か否かを判定することは困難である．このため，その代わりに，構築したコーパスを用いた，形容詞と名詞からなる共起表現の形容詞の適切な代替候補を提示する作文支援システムを作成し，不自然な共起表現とその修正のデータ（英語論文校正会社に依頼して作成）を利用して性能評価を行った．代替候補の形容詞を 100 提示した場合に，その中に英語論文校正会社の proof reader が示した修正後の形容詞が含まれる割合は 62%であった．候補の中には非母語話者でも意図するものと明らかに意味が異なることが分かる簡単な形容詞が多数出現しているため，100 候補を確認することは容易である．この意味で実用に耐えるとも考えられるが，意図するものと明らかに意味が異なる形容詞を候補から削除する方法などの検討も必要と考えられる．

また，(2)で述べた 993 編の論文を分類した母語話者論文と非母語話者論文に基づき，英語母語話者と英語非母語話者とにおける『名詞用法の差異』および『学校文法項目の使用多寡の傾向』の調査を行い，英語教育上有用な知見を得た．

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９件）

① 田中省作，安東奈穂子，冨浦洋一，コーパス構築と著作権：Web を源とした質情報付き英語科学論文コーパス，英語コーパス研究，第 19 巻，pp. 31-41（印刷中）（査読有）

② M. Shibata, T. Funatsu, Y. Tomiura, Extraction of Alternative Candidates for Unnatural Adjective-Noun Co-occurrence Construction of English, *Procedia - Social and Behavioral Science*, Vol. 27, pp. 32-41 (2011)（査読有）
DOI: 10.1016/j.sbspro.2011.10.580

③ 田中省作，柴田雅博，冨浦洋一，Web を源とした質情報付き英語科学論文コーパスの構築法，英語コーパス研究，第 18 巻，pp. 61-71（2011）（査読有）

④ 田中省作，Web コーパスの言語情報処理基盤，英語コーパス研究，第 18 巻，pp. 97-111（2011）（査読有）

⑤ 安東奈穂子，著作権法のもとでの情報解析，人工知能学会誌，第 25 巻，pp. 634-652（2010）（査読無）

⑥ M. Shibata, Y. Tomiura, T. Mizuta, Identification among Similar Languages Using Statistical Hypothesis Testing, *Proc. of Pacific Association for Computational Linguistics*, pp. 47-52 (2009)（査読有）

⑦ 冨浦 洋一，青木 さやか，柴田 雅博，行野 顕正，仮説検定に基づく英文書の母語話者性の判別，自然言語処理，Vol. 16, pp. 23-46（2009）（査読有）

〔学会発表〕（計１３件）

① M. Shibata, T. Funatsu, Y. Tomiura, Extraction of Alternative Candidates for Unnatural Adjective-Noun Co-occurrence Construction of English, Pacific Association for Computational Linguistics (PACLING'11), 2011.7.19, Malaysia

② 小林雄一郎，田中省作，冨浦洋一，ランダ

ムフォレストを用いた英語科学論文の分類と評価,情報処理学会人文科学とコンピュータ研究会第 90 回研究発表会, 2011 年 5 月 21 日, 同志社大学

③ 田中省作, Web コーパスの言語情報処理基盤, 英語コーパス学会第 35 回大会シンポジウム, 2010 年 4 月 24 日, 兵庫県立大学（兵庫県）

④ 田中省作, Web を源とした英語科学論文コーパスの構築 ―技術的方法論と法的観点からの検討―, 英語コーパス学会第 34 回大会, 2009 年 10 月 3 日, 青山学院大学(東京都)

⑤ M. Shibata, Y. Tomiura, T. Mizuta, Identification among Similar Languages Using Statistical Hypothesis Testing, Pacific Association for Computational Linguistics (PACLING'09), 2009.9.1, Hokkaido University

〔その他〕
ホームページ等
作成中（http://nlp.inf.kyushu-u.ac.jp）

６．研究組織
(1)研究代表者
冨浦　洋一（TOMIURA YOICHI）
九州大学・　　システム情報科学研究院・教授
研究者番号：１０２１７５２３

(2)研究分担者
田中　省作（TANAKA SHOSAKU）
立命館大学・文学部・准教授
研究者番号：００３２５５４９

後藤　一章（GOTO KAZUAKI）
摂南大学・外国語学部・講師
研究者番号：９０３９７６６２

羽山　恵（HAYAMA MEGUMI）
獨協大学・外国語学部・准教授
研究者番号：６０４０９５５５

安東　奈穂子（ANDO NAHOKO）
九州大学・大学院法学研究院・専門研究員
研究者番号：５０３８０６５５

柴田　雅博（SHIBATA MASAHIRO）
九州大学・情報基盤研究開発センター・学術研究員
研究者番号：００４５２８１３