様式 C-7-2

自己評価報告書

平成２３年　５月　９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 Technical Paper Corpus from Web, Opening it to Researchers, and Application of it

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・准教授
研究者番号：１０２１７５２３

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．研究計画の概要

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．

本研究では，英文の質情報が付与された科学技術論文コーパスの構築と公開，およびその利用に関して，以下を行う．

(1) Web 上から英語で書かれた大量の科学技術論文を収集するシステムを作成する．
(2) 品詞列の情報を基に英文書の英文の質を高精度で推定するシステム（母語話者性判別システム）を開発する．
(3) 上記(1)のシステムを用いて，Web 上から大量の英語科学技術論文を収集し，上記(2)のシステムを用いて，英文の質を推定し，英文の質情報付き英語科学技術論文コーパスを構築する．
(4) 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．
(5) 得られたコーパスから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，これらを公開する．

２．研究の進捗状況

(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２万の論文を収集した（まだ収集できる見込みである）．

(2) 母語話者性判別システム
Web 上から収集した 993 編の論文の英文の質を英文校正会社に依頼して判定してもらい，開発した判別器の性能をこのデータを基に評価した．精度を重視するようにメタパラメタを設定したときの母語話者判定に対する精度，再現率はそれぞれ 94%，25%であり，非母語話者判定に対する精度，再現率はそれぞれ 92%，22%であった．

(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た．これを用いて，これまでに，約 14,000 編の母語話者論文と約 3,100 編の非母語話者論文を取り出している．

(4) コーパスの構築・公開に関する法的検討
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，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．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．

(5) 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ため，代わりに，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構築し，英語校正会社に依頼して作成した評価データを利用して性能評価を行い，実用化に向けた検討を行った．また，非母語話者文書に固有の文法的・語彙的特徴の抽出も試行段階ではあるが進めている．

３．現在までの達成度

(1) Web 上からの論文収集システム：②（おおむね順調） ∵ 進捗状況に示した通り．
(2) 母語話者性判別システム：②（おおむね順調） ∵ 進捗状況に示した通り．
(3) コーパスの構築：②（おおむね順調） ∵ 進捗状況に示した通り．
(4) コーパスの構築・公開に関する法的検討：②（おおむね順調） ∵ 進捗状況に示した通り．
(5) コーパスを利用した研究：③（やや遅れている） ∵ 予定していた非母語話者固有の文法的・語彙的特徴の抽出がまだ試行段階であるため．

４．今後の研究の推進方策

まず第１に，以下の項目と並行して，英語科学技術論文の収集を引き続き行う．

母語話者性判別システムについては，精度は高いものの再現率が非常に低い．これは，英語論文は数十万規模で収集できると考え，再現率を犠牲にして精度を高くしたためである．より多くの論文を母語話者論文あるいは非母語話者論文としてコーパスに含めることができるように，最終年度は，品詞列の性質だけでなく，談話標識の分布などの特徴も取り入れて，精度を落とさず再現率を向上させる判別システムの改良を試みる．

法的検討の結果，コーパスを構築するのではなく，コーパス構築のためのツールキットを公開するというように方針を転換したため，ツールキット公開に向けてマニュアル等のドキュメントを整備する．

構築するコーパスを利用した研究はやや遅れている．最大の原因は十分な規模のコーパスを研究分担者に提供することができなかったためであるが，昨年度末までにある程度の規模のコーパスが構築できたため，最終年度は本格的な実験や検討に着手できるものと思われる．

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 5 件）

① 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，2011 年（印刷中），査読有

② 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，2011 年（印刷中），査読有

③ 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 年，査読無

④ 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 年，査読有

⑤ 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 年，査読有

〔学会発表〕（計 4 件）

① 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）

② 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）

③ 水田貴章，母語話者／非母語話者コーパスを用いた不自然な英語表現の抽出，電気関係学会九州支部連合大会（第 62 回），2009 年 9 月 28 日，九州工業大学（福岡県）

④ 水田 貴章，仮説検定に基づいた言語識別，情報処理学会自然言語処理研究会，2008 年 11 月 27 日，九州大学（福岡県）