様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １３ 日現在

機関番号：２７１０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ１１４４６

研究課題名（和文）Natural language processing for academic writing in English

研究課題名（英文）Natural language processing for academic writing in English

研究代表者
Ｇｏｈ　ＣｈｏｏｉＬｉｎｇ（Goh, Chooi Ling）

北九州市立大学・国際環境工学部・特任准教授

研究者番号：９０５３１６１６

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：英語を母国語としない研究者は、自分の研究内容や結果を英語で表現することに困難を感じている。そのため、国際論文を提出する際に不採択されてしまう可能性がある。本研究の目的は、自然言語処理技術を用いて英論文作成を支援するシステムを設計し実装することである。学術論文には通常、共通の語彙と特有の書き方がある。研究者は簡単な英語で自分の考えを大まかに伝えることができるが、同じ考えをより高いレベルの流暢かつ適切な表現が存在する可能性がある。本研究では、学術論文に特化した語彙の候補や、スタイルの変換を提案する支援システムを提供する。

研究成果の学術的意義や社会的意義
この研究成果は、不確かで単純な文章で書かれた研究アイデアを専門的な文章に変換することができ、研究者はより短期間かつ低コストで多くの論文を発表することが可能となる。これにより、研究者は自分の研究成果をより普及し、世界的なランキングへの参加につなげることができるだろう。特に、若手研究者にとっては、専門的な論文の書き方を学ぶことができる。従って、科学論文の特有な表現や語彙を知ることになる。

研究成果の概要（英文）：Researchers who are non-native speakers of English experience difficulties to describe their work and results in English. Due to the low level of proficiency, they are more probably rejected for international publication. The purpose of this research is to design and implement a computer aided system, using natural language processing techniques, to help researchers to write their articles in English. Scientific articles usually have common vocabularies and specific writing styles. Researchers may be able to roughly convey their ideas in simple English, but more fluent and adequate ways of describing the same idea may exist at a higher proficiency level. In this research, we provide a system to help to change the writing styles by proposing some possible vocabulary candidates which are specific to scientific writing.

研究分野：自然言語処理

キーワード：単語の埋め込み　マスク言語モデル　語彙バンドル　自動文章生成　英論文執筆支援

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

研究結果を発表するのが研究者にとってとても重要である。個人の研究業績では研究論文の数に依存するが、大学や研究機構の評価では全体の研究成果に依存する。論文の数は多ければ多いほど評価が高くなり、国際会議や論文誌で発表するのはもっと自分の研究が知られていてもっと高く評価される。しかし、多くの国際会議や論文誌などは英語で書いた論文しか認めない。十分な英語力を有していない研究者には不利になる。英語の水準は母国語話者より低いため、流暢でない文章になり、研究のアイデアや方法、結果などを正確に伝えられず、この原因で論文が不採択になる場合がある。研究結果が良くても論文にならなければ、評価されない。英語が母国語でない話者にとって英語で論文を書くことは容易ではないし、時間とコストがかかる。

時間：英語を母国語としない研究者にとって、英語での論文執筆は、英語を母国語とする研究者と比べて余分な時間を必要とする。この余分な時間は、研究時間を犠牲にして失われる。

コスト：英語を母国語としない研究者は、通常、プロの校正者に論文の校正を依頼する場合がある。これにはコストがかかる。この場合も、直接的な研究コストを犠牲にして失われる。

以上の理由を踏まえて、英語論文執筆の際に、文章を自動生成するツール、あるいは語彙や、類似文の検索などの支援ツールが必要になってくる。

２．研究の目的

国際誌に掲載されるための条件としては、研究成果の内容と文章の質の２つである。もちろん研究成果についてはここでは考慮できないが、英語での文章の質を向上する重点を置くことにした。本研究課題の目的は、英語を母国語としない研究者のために、自然言語処理の技術を用いて、剽窃を避けながら、英語で学術論文を書くための支援システムを設計・実装することである。

現存の機械翻訳システムの精度はかなり向上しているが、正しく学術論文の書式やスタイルになっていない場合がよくある。この研究では、自然言語処理の技術を通じて正しい学術論文のスタイルや語彙などの提案をする。方法としては、大量の既存の高品質の、すでに発表された論文の中から似ている文章や句を学習して、適切な語彙やスタイルを提案する。こうやって、質の高い論文を書くことができ、より高い評価が得られ、論文が採択されるまで至る可能性が高くなると考える。英語が母国語でない研究者にとって役に立つだろう。

３．研究の方法

日本語の話者なら、多くの場合は機械翻訳システムを用いて英語に翻訳する。その不完全な翻訳文を支援ツールで語彙を編集したり、スタイルを変更したり、よりよい文章を作成する。それぞれの分野では独自のスタイルや、語彙が決められている。まず自然言語処理の流れに従い、自然言語処理研究分野への応用、すなわち NLP4NLP を行う。この研究では ACL Anthology Reference Corpus で提供される NLP 分野の研究論文を用いる。ACL は自然言語処理分野のプレミアム・カンファレンスであり、論文の英語レベルには定評があるから。

多くの論文誌は、標準的なステレオタイプの文体が決められている。各論文は、いくつかのセクションから構成されている。おおよそ、序論、先行研究、方法、結果、考察、今後の課題、結論である。さらに、各セクションの中で使われる表現も極めて標準的である。よりよい文体を提案するために、コーパスを上記のセクションに分類し、セクションごとに学習を特化させて、その結果、目的のセクションの文体に応じて、より適切な文章が生成できる。

剽窃を避けるため、提案する文章は慎重に構成する必要がある。語彙の束は剽窃とみなされないので、剽窃のない束のデータベースを構築する。これらの語彙のバンドル（lexical bundles）は、既存の論文によく見られるものであっても、ユーザに提案するのに安全に使用することができる。このデータベースは、研究論文コーパスから出現頻度の高い長い N-gram を収集し構築される。

単語の分散表現、あるいは単語埋め込みモデル（Word2Vec）では単語と単語の間に関係性が見えてきる。類義語だけではなく、文脈によって、単語間の類似度も測れる。それを利用して、シソーラスや辞書引きに代えて、代用語を検索することにする。単語より長い単位のフレーズや文章なども、単語埋め込みの拡張モデルとして文埋め込みモデルを利用すると考える。

最近の研究では深層学習が主に主流になって、この研究も同じ方向で行うことにした。特に大量のテキストから学習された事前に学習済み言語モデルを用いてたくさんの言語タスクの精度が向上することを示された。2018 年に公開した、文脈を考慮した分散表現である BERT はそれのひとつである。さらに、科学分野に特化したモデルとして、SciBERT というモデルを利用することも可能である。

４．研究成果

（１）学習コーパスの整備

ACL Anthology Reference Corpus（ACL-ARC）で提供された研究論文をダウンロードし、それぞれの特徴を学習させるためセクション（Abstract, Introduction, Body, Conclusion）ごとに分割した。以下、このコーパスを利用して、各データの抽出や、学習に用いた。

（２）語彙のバンドル（lexical bundles）

（a）用語集：ACL-ARC から、自然言語処理分野の語彙の束を収集した。出現頻度が高くて長い N-gram を抽出し、人手でチェックしたバンドル集合（Salazar により）から学習した機械学習モデルを用いて、それらを真のバンドルと誤ったバンドルに分類した。検証実験の結果、最適なモデルは 76%の精度を達成した。このモデルを用いて、ACL-ARC から 18,000 以上の語彙バンドルを抽出し、研究室のホームページにて一般に公開した。（国際会議 ICACSIS 2019 で発表）

（b）語彙バンドルの典型性：学術論文の異なるセクションにより使用する語彙バンドルに差異がある。本研究では自然言語処理系論文のセクションにおける語彙バンドルの典型性を定量化する方法を提案した。この指標は個々の KL-Divergence スコアと、あるバンドルがある種類のセクションに出現する確率の積である。得られた結果は、人間の目から見て納得するものであった。頻度による単純なランキングよりも、正しいセクションで使われる典型的な語彙バンドルをより高い位置にランクを付けることができた。提案した典型度指標は、学術論文執筆支援システムにおいて、執筆者が作成するセクションに典型的な語彙バンドルを提案し、より適切な語彙バンドルを検索するために有用であると考えられる。（国際会議 ICNLP 2020 で発表）

（３）セクションの文章生成

自然言語処理分野の学術論文の要約と結論の文の間で、学術論文のスタイル変換に焦点を当て、双方向のスタイル変換が可能かどうかを検討した。この課題を解決するための学習ペアが存在しないため、まず、文のベクトル表現間の余弦類似度によって測定される文の類似性に依存し、科学論文のコーパスから対になっていない類似文のデータセットを構築した。次に、このデータセットでディープラーニングの手法のひとつ、CycleGAN アーキテクチャを用いたモデルを設計し、生成器が Transformer として実装された。本手法により、変換された文の主な相違点は各セクションに特徴的な時制と語法であることがわかった。（国際会議 ICACSIS 2020 で発表）

（４）語彙検索及び類似文検索

（a）英語を母国語としない研究者は、学術論文を英語で作成する際に常に何らかの問題に直面する。ほとんどの場合、それは語彙の不足、あるいは代替表現方法の知識の不足が原因である。特定の分野の学術論文に使われる代用語を探すために、シソーラスや辞書引きに代えて、単語埋め込みを利用することを提案した。単語埋め込みには、意味的に類似した単語だけでなく、似たような単語ベクトルを持つ他の単語も含まれている可能性があり、より良い表現になる可能性がある。また、特定の分野の学術論文のコーパスに対して学習させた単語埋め込みモデルは、その文体に適合し、その分野に適した類似の表現を提案する可能性がある。実験の結果では、自然言語処理領域で学習した単語埋め込みモデルは、特定の文脈でターゲット単語を置き換えるために使用可能性のある代用語を提案することが有効であると結論付けることができた。（国際会議 PACLING 2019 で発表）

（b）英語を母国語としない者が英語で研究成果を論文執筆する場合、不適切な表現を使いがちである。論文執筆を支援するために、マスクされた言語モデルを使用することを検討した。マスク言語モデルは、文の前後の文脈を与えることで、空白を埋めるために有用な単語を予測することができる。また、事前に学習済みマスク言語モデル SciBERT や ACL-ARC で学習した単語埋め込みモデル Word2vec のように、特定の領域で学習したモデルを用いることで、学術論文でよく用いられる語彙の選択を制御し、その領域における学術論文の書き方の習熟度を向上させることが可能である。実験では、自然言語処理分野の国際誌と国内誌の 2 誌でテストを行い、英語を母国語とする人とそうでない人の両方が執筆した論文から抜粋した抄録を対象に実験を行った。言語モデルの予測の妥当性は、原文と全く同じ動詞がいくつ予測できたかを数えることと、また、置換後の出力の流暢さを測ることによって評価した。その結果、単語埋め込みとマスク言語モデルの両方の特徴を含む学習・文章作成システムを設計する動機付けとなる有望な結果を得た。（国際会議 PACLIC 2021 で発表）

（c）学術論文の書き方を学ぶ方法の 1 つとしては、過去に出版された論文を参照することである。同じ分野の論文では、たいていの場合、似たような表現、語彙、書き方が使われる傾向がある。本研究では、文埋め込みを利用し、過去の論文から類似の文章を検索し、学術論文の執筆を支援するのに有効であることを示した。二人の英語を母国語としない大学院生に、システムが提案する類似文だけを参考にして、元の文を修正してもらった。10 文のうち、9 文が修正可能だった。この結果から、文埋め込みによって提案された類似文は、論文執筆の支援に有用であると考えられる。しかし、検索結果の精度がまだ低いため、元の文と関係のない文も表示されてしまうという問題がある。また、執筆者にとって、たくさんの類似文に目を通すのは時間の無駄であり、フラストレーションの原因になる。したがって、もっと精度を上げ、表示する例文を少なくすることが課題に残る。（国内会議 NLP 2020 で発表）

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Chooi-Ling Goh and Yves Lepage	４．巻 1215
２．論文標題 An Assessment of Substitute Words in the Context of Academic Writing Proposed by Pre-trained and Specific Word Embedding Models	５．発行年 2020年
３．雑誌名 Communications in Computer and Information Science	６．最初と最後の頁 414～427
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-981-15-6168-9_34	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　5件）

１．発表者名 Chooi-Ling Goh
２．発表標題 Applying Masked Language Models to Search for Suitable Verbs Used in Academic Writing
３．学会等名 The 35th Pacific Asia Conference on Language, Information and Computation (PACLIC 2021)（国際学会）
４．発表年 2021年

１．発表者名 Haotong Wang, Yves Lepage and Chooi-Ling Goh
２．発表標題 Typicality of lexical bundles in different sections of scientific articles
３．学会等名 The 2nd International Conference on Natural Language Processing (ICNLP 2020)（国際学会）
４．発表年 2020年

１．発表者名 Haotong Wang, Yves Lepage and Chooi-Ling Goh
２．発表標題 Unpaired abstract-to-conclusion text style transfer using cycleGANs
３．学会等名 The International Conference on Advanced Computer Science and Information Systems (ICACSIS 2020)（国際学会）
４．発表年 2020年

１．発表者名
Chooi-Ling Goh

２．発表標題
Cloze Test for Verbs in Academic Writing by Masked Language Models

３．学会等名
言語処理学会第27回年次大会

４．発表年
2021年

１．発表者名
Chooi-Ling Goh and Yves Lepage

２．発表標題
Extraction of lexical bundles used in natural language processing articles

３．学会等名
The International Conference on Advanced Computer Science and Information Systems (ICACSIS 2019) (国際学会)

４．発表年
2019年

１．発表者名
Chooi-Ling Goh and Yves Lepage

２．発表標題
An assessment of substitute words in the context of academic writing proposed by pre-trained and specific word embedding models

３．学会等名
The 16th International Conference of the Pacific Association for Computational Linguistics (PACLING 2019) (国際学会)

４．発表年
2019年

１．発表者名
Chooi-Ling Goh and Yves Lepage

２．発表標題
Finding similar examples for aiding academic writing using sentence embeddings

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
Chooi-Ling Goh and Yves Lepage

２．発表標題
Word Embeddings in place of Dictionary Lookup in the context of Academic Writing

３．学会等名
言語処理学会第25回年次大会

４．発表年
2019年

１．発表者名
Tianjiao Li and Yves Lepage

２．発表標題
Informative Sections and Relevant Words for the Generation of NLP Article Abstracts

３．学会等名
言語処理学会第25回年次大会

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Kakenhi Kiban C 18K11446 - Experimental Data
http://lepage-lab.ips.waseda.ac.jp/en/projects/kakenhi-kiban-c-18k11446/
AwTool Website
https://cl-lab.net/app/awtool/
CL-Laboratory
https://cl-lab.net/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	ＬＥＰＡＧＥ　ＹＶＥＳ (Lepage Yves) (70573608)	早稲田大学・理工学術院（情報生産システム研究科・センター）・教授 (32689)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関