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　３ 年　 ６ 月 １５ 日現在

機関番号：６２６１５
研究種目：基盤研究(B)（一般）
研究期間：2018～2020
課題番号：１８Ｈ０３２９７

研究課題名（和文）定型表現集の活用を支援する言語処理基盤技術の研究

研究課題名（英文）Natural Language Processing Technologies for Formulaic Expressions

研究代表者
相澤　彰子（Aizawa, Akiko）

国立情報学研究所・コンテンツ科学研究系・教授

研究者番号：９０２２２４４７

交付決定額（研究期間全体）：（直接経費）　　13,200,000円

研究成果の概要（和文）：本研究では、英作文における定型表現集の活用を支援するための自然言語処理技術の研究開発に取り組み、特に学術論文に焦点をあてて、論文の文中の定型表現および伝達機能の抽出および検索手法を提案した。本研究を通して、手法の構築や評価に広く利用可能なコーパスを構築・公開するとともに、伝達機能の抽出における深層学習モデルの有効性や定型表現抽出における文法知識の活用法を示すことができた。

研究成果の学術的意義や社会的意義
英語による文書作成を支援するための定型表現集が数多く出版されているが、その大半は電子化されておらず、電子化されていたとしても執筆途中に必ずしも気軽に利用できるものではない。本研究では、体系化された定型表現を大規模なドメイン・コーパスに対応付けることによって、ドメインに特化した言い回しや伝達機能などを含む豊富な文脈情報が獲得できることを示した。本研究で得られた知見は、これまで困難であった定型表現の予測や検索の実現に結び付くものである。

研究成果の概要（英文）：This research aims to develop natural language processing technologies to construct a formulaic expression database for English writing assistance. Focusing on academic paper writing, we proposed a method for extracting and retrieving formulaic expressions with their communicative functions. We also constructed an annotated corpus of sentences with communicative functions that can be used to training and evaluating. Our research demonstrated the effectiveness of deep learning models in extracting communicative functions and grammatical knowledge in extracting formulaic expressions.

研究分野：情報通信／知能情報学、ヒューマンインターフェースインタラクション、データベース

キーワード：定型表現　執筆支援　ドメイン・コーパス　辞書自動構築　意味表現

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

非母語話者が英語論文の執筆に要する時間は平均でも母語話者の２倍ともいわれ、言語構造が大きく異なる日本語話者の場合は数倍程度の時間を費やしていることが予想される。また、執筆した英語の質の問題から、論文の添削や英文校正に要するコストも大きい。

論文を対象とした自動翻訳や英文自動誤り訂正に関する研究は従来から行われているが、これらの研究では、完成された「文」を入力として、意味的・構文的な対応が保証された「文」を出力することを目標としている。しかし実際の執筆作業では、必ずしも「文」を先に作成するわけではない。まず、言いたいことの骨格を念頭に目的にあった定型表現を含む例文を見つけ、次に、それを編集して使うことがしばしば行われる。このような方法は、誤りを回避して効率的に執筆を行う上で有効である。

ところが、自然言語処理の研究分野では、例文の書き換えによる英文執筆支援の技術はほとんど検討されてこなかった。その理由として、(1)「定型表現」の言語処理に使えるリソースが整備されていないこと、(2)定型表現は文の役割を示す機能的なものであり、内容語に注目する従来手法では類似検索が困難であること、をあげることができる。

２．研究の目的

以上の背景に基づき、本研究では「英作文における定型表現集の活用を支援するために必要な自然言語処理技術の研究開発」という新たな課題に挑戦する。

英語による文書作成を支援するための定型表現集が数多く出版されているが、その大半は電子化されておらず、電子化されていたとしても執筆途中に必ずしも気軽に利用できるものではない。本課題では、この問題を解決するため、ユーザの入力に基づき適切な定型表現を提示する手法の開発に取り組む。具体的には、大量のドメイン・コーパスを計算機で解析して、定型表現やその機能を自動的にラベル付けするための言語処理手法を研究する。体系化された定型表現を大規模なドメイン・コーパスに対応付けることによって、ドメインに特化した言い回しを含む豊富な文脈情報を獲得することができ、これまで困難であった定型表現の予測や検索の実現に結び付く。

３．研究の方法

本研究では「定型表現アノテーション付き文コーパスの構築」、「定型表現アノテーション付き文コーパスの構築」、「定型表現抽出手法の開発」、「伝達機能ラベル付き定型表現データベースの構築」４つの研究課題を設定した。

(1) 定型表現アノテーション付き文コーパスの構築

定型表現の抽出・検索手法の分析や比較のためには、人手でアノテーションしたコーパスが必要であるが、そのような目的で使えるコーパスは存在しなかった。そこで、論文から抽出した文に対して、その中に含まれる定型表現の機能ラベルを人手で付与したコーパスの構築を目指す。

(2) 定型表現を含む文の伝達機能の予測

与えられた入力文に対して、定型表現の伝達機能ラベルを予測する手法を研究し、(1)で得られた正解付きコーパスを用いて有効性を評価する。

(3) 定型表現抽出手法の開発

論文からの定型句の抽出に関する従来研究では、高頻度 n グラムに注目した手法が多く用いられるが、これらは「よく使われる言い回し」全般を対象とするものであることから、本研究では伝達機能を持つ定型表現を対象とした抽出手法を新たに開発する。

(4) 伝達機能ラベル付き定型表現データベースの構築

定型表現の機能的な意味を、指定されたドメインのコーパス中での用例と対応づけてデータベース化し、その有用性を評価する。

４．研究成果

本研究で設定した４つの研究課題に対する成果を以下にまとめる。

(1) 定型表現アノテーション付き文コーパスの構築

既存の定型表現集である Academic Phrasebank を参照して、代表的な 39 個の機能表現ラベルを選び、自然言語処理分野の論文アーカイブである ACL Anthology から計 397 個の例文を抽出したコーパスを構築・公開した。また、これらの例文を用いて、文の機能表現に関する２択問題を設計し、代表的な埋め込み表現を用いた場合のベースライン性能、および人間のアノテーターによる正解率をあわせて公開した（発表文献：LREC-2020）。

(2) 定型表現を含む文の伝達機能の予測

伝達機能に基づく文分類を教師あり学習を用いて行う手法を提案した。(1)の知見に基づき、計算言語学、分子化学、腫瘍学、心理学の４つの学術分野について、正解機能ラベルを付与したコーパスを作成して性能評価を行った結果、SciBERT を追加学習することで高い分類性能が得られることを示した。また、４つの学術ドメインに関する分析によって、伝達機能の予測はドメインの違いに比較的頑強で、異なるドメインで学習した場合でも高い分類性能が得られるとの知見を得た（発表文献：EACL-2021）。

(3) 定型表現抽出手法の開発

予備的な試みとして、コーパス中の文の各単語に対して定型表現の一部であるかそうでないかのラベルを付与する逐次ラベリング問題の枠組みを用いて、機械学習を適用する手法を提案した（COLING-2018）。しかしながら、定型表現の正解データの作成は困難を伴うことが判明したことから、次に固有表現抽出および依存構造解析を用いた定型表現抽出法を新たに提案した。人手評価により、既存の定型句評価手法と比較して、提案手法が伝達機能付き定型表現に適した抽出法となっていることを示した（発表文献：SDU-2021）。

(4) 伝達機能ラベル付き定型表現データベースの構築

これまでの研究成果に基づき、機能的な意味を付与した大規模な定型表現データベースを実際に構築し、機能的な意味に基づく定型表現検索により、従来のキーワード検索では得られない多様な表現が得られることを検証した（発表文献：EACL-2021）。

これらの研究を通して、論文の文中の定型表現および伝達機能の抽出および活用に関して、今後の研究にも活用できるコーパスを構築・公開するとともに、伝達機能の抽出における深層学習モデルの有効性や定型表現抽出における文法知識の活用法を示すことができた。本研究で対象とした伝達機能の個数はまだ限られたものであるため、その拡大が今後の課題としてあげられる。

５．主な発表論文等

〔雑誌論文〕　計10件（うち査読付論文　9件／うち国際共著　2件／うちオープンアクセス　10件）

１．著者名 Kenichi Iwatsuki and Akiko Aizawa	４．巻 -
２．論文標題 Using Formulaic Expressions in Writing Assistance Systems.	５．発行年 2018年
３．雑誌名 The 27th International Conference on Computational Linguistics (COLING 2018)	６．最初と最後の頁 2678-2689
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yang Zhao, Zhiyuan Luo, and Akiko Aizawa	４．巻 -
２．論文標題 A Simple Language Model based Evaluator for Sentence Compression.	５．発行年 2018年
３．雑誌名 The 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)	６．最初と最後の頁 170-175
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/P18-2028	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Takuma Udagawa abd Akiko Aizawa	４．巻 33
２．論文標題 A Natural Language Corpus of Common Grounding under Continuous and Partially-Observable Context.	５．発行年 2019年
３．雑誌名 The 33rd AAAI Conference on Artificial Intelligence (AAAI-19)	６．最初と最後の頁 7120-7127
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yang Zhao, Xiayu Shen, Wei Bi, Akiko Aizawa	４．巻 -
２．論文標題 Unsupervised Rewriter for Multi-Sentence Compression.	５．発行年 2019年
３．雑誌名 The 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)	６．最初と最後の頁 2235-2240
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/P19-1216	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Saku Sugawara, Pontus Stenetorp, Kentaro Inui, Akiko Aizawa	4．巻 34
2．論文標題 Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets.	5．発行年 2020年
3．雑誌名 The 34th AAAI Conference on Artificial Intelligence (AAAI-20)	6．最初と最後の頁 8918-8927
掲載論文のDOI（デジタルオブジェクト識別子） 10.1609/aaai.v34i05.6422	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Takuma Udagawa, Akiko Aizawa	4．巻 34
2．論文標題 An Annotated Corpus of Reference Resolution for Interpreting Common Grounding.	5．発行年 2020年
3．雑誌名 The 34th AAAI Conference on Artificial Intelligence (AAAI-20)	6．最初と最後の頁 9081-9089
掲載論文のDOI（デジタルオブジェクト識別子） 10.1609/aaai.v34i05.6442	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Kenichi Iwatsuki, Florian Boudin, and Akiko Aizawa	4．巻 -
2．論文標題 An Evaluation Dataset for Identifying Communicative Functions of Sentences in English Scholarly Papers.	5．発行年 2020年
3．雑誌名 The 12th International Conference on Language Resources and Evaluation (LREC 2020)	6．最初と最後の頁 1712-1720
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Takuma Udagawa, Takato Yamazaki, and Akiko Aizawa	4．巻 -
2．論文標題 A Linguistic Analysis of Visually Grounded Dialogues Based on Spatial Expressions.	5．発行年 2020年
3．雑誌名 EMNLP 2020 (2020 Conference on Empirical Methods in Natural Language Processing)	6．最初と最後の頁 750-765
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/2020.findings-emnlp.67	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Kenichi Iwatsuki and Akiko Aizawa	4．巻 -
2．論文標題 Extraction of Formulaic Expressions from Scientific Papers.	5．発行年 2021年
3．雑誌名 The AAAI-21 Workshop on Scientific Document Understanding (SDU-2021))	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Kenichi Iwatsuki and Akiko Aizawa	4．巻 -
2．論文標題 Communicative-Function-Based Sentence Classification for Construction of an Academic Formulaic Expression Database.	5．発行年 2021年
3．雑誌名 The 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021): Main Volume	6．最初と最後の頁 3476-3497
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 篠田一聡，相澤彰子
2．発表標題 深くて早い言語理解の実現に向けて．
3．学会等名 NLP若手の会（YANS）第14回シンポジウム
4．発表年 2019年

1．発表者名 山崎 天，相澤 彰子
2．発表標題 タスク指向対話におけるニューラル句抽出を用いた End-to-End 発話生成
3．学会等名 第11回対話システムシンポジウム
4．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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

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