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　５ 年　 ６ 月　 ９ 日現在

機関番号：１２６０３
研究種目：基盤研究(B)（一般）
研究期間：2019～2022
課題番号：１９Ｈ０４２２４

研究課題名（和文）大規模字幕コーパスからの単語・フレーズ・会話のボトムアップ言語教材自動抽出

研究課題名（英文）Study on Automatic extraction of language teaching materials from a large closed caption corpus by bottom-up assembly of linguistic units such as words, phrases, and conversations.

研究代表者

望月　源（Mochizuki, Hajime）

東京外国語大学・大学院総合国際学研究院・准教授

研究者番号：７０３１３７０７

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）：我々は，これまでなかったサイズの異なるnグラムの頻度を一括して比較可能な統合文脈nグラムを開発し，複数の語，単語を組み合わせた定型表現（フォーミュライックシーケンス，FS）を抽出した。FSの分散表現でのクラスタリングにより表層表現は異なるが機能的に類似した「機能別フレーズ集合」としてFSのクラスタを獲得できることが確認できた。コーパス内の会話部分とCan-doを対応づけした教師データを作成し深層学習モデルによる教材会話の自動抽出も試み一定の成果を得た。字幕コーパスは拡張を続け，22億語規模に拡大した。研究成果は，EDMEDIA,E-Learnなどの国際学会を中心に査読付き論文発表を行なった。

研究成果の学術的意義や社会的意義
大規模日本語会話コーパスの構築を続け，10年以上にわたる日本のテレビ番組の字幕データを整備し，59万8千番組，2億8百文，22億5千万語超に達した。また，コーパスの語彙調査を行い，テレビ字幕データが言語教材として十分に有益であることを確認した。これまで実現していなかったサイズの異なるnグラムの頻度を一括して比較可能な統合文脈nグラムを開発したコーパス内のすべての文から定型表現としてフォーミュライックシーケンス，FSの抽出を行い，FSが日本語教科書の重要フレーズを含むことを確認した。日本語Can-doに対応した会話データを教師データとして整備し，機械学習モデルでの会話セグメント自動抽出を行った。

研究成果の概要（英文）：We have developed an integrated context n-gram that enables batch comparison of frequencies of n-grams of different sizes, and extracted formulaic sequences (FS) that combine multiple words. By clustering FSs in distributed expressions, we confirmed that we could obtain clusters of FSs as "functionally similar phrase sets" with different surface expressions. We also tried to automatically extract the conversation parts of the corpus using a deep learning model, and obtained certain results. The subtitle corpus has been continuously expanded to 2.2 billion words. The research results were presented in peer-reviewed papers mainly at international conferences such as EDMEDIA and E-Learn.

研究分野：情報科学

キーワード：学習コンテンツ開発支援　eラーニング　日本語教育　自然言語処理　Formulaic Sequences

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

単語を基本単位とした分散表現が定着するとともに，単語の意味の数値化は，機械翻訳や対話などのさまざまな自然言語処理の応用に恩恵をもたらしている。同時に，新たな観点目的による様々な分散表現が研究課題として盛んに研究されている。言語会話教材の抽出も言語処理のひとつであり分散表現の利用による恩恵が充分に期待できる。複数の単語からなるかたまりである Formulaic Sequences, FS についても，テキストを FS 単位に分割することで分散表現が計算できる。FS は意味と発話意図，待遇表現などのフレーズを含むため，FS の分散表現はフレーズ間の関係を計算可能にする。一方，FS には，謝罪する表現，不満を表明する表現，感謝をする表現，など言語教育のキーフレーズとして有益なものも含まれるが，FS の分析はまだ充分でない。また，言語教育の重要な役割を果たす Can-do はキーフレーズと強く結びつき，Can-do と関連する会話には多くのフレーズが含まれる関係にあると思われるが，充分に分析がなされていない。例えば，テレビ字幕からの大規模会話コーパス内には，「レストランで料理を注文する」場面など Can-do の場面，機能に使用可能な会話が多く含まれており，内在するフレーズから Can-do と会話の関連性を図ることが可能であると考えられるが，各 Can-do と会話を直接結びつけるための手法がまだ完成していない。単語，フレーズ，会話をボトムアップに結びつけて言語教材を大規模コーパスから抽出する技術の確立が必要である。

２．研究の目的

我々はこれまで存在しなかった大規模会話コーパスをテレビ字幕データから構築しており，現在約３１万 9 千番組分，約１億１千 2 百万文分，１２億１千万語の規模になっている。この研究では，このコーパスをさらに拡張するとともに，コーパスから，語，フレーズ，会話という言語単位のボトムアップな組み上げによって，言語教育 Can-do に対応する言語教材の自動抽出を試みるとともに，これまで進んでいなかった Can-do とキーフレーズの対応付けも行う。具体的には，(1)現在の語の分散表現では１語に複雑な語義が含まれることになり，各語がどのような役割であるかに応じた違いが区別できない。そこで語の意味用法ごとに分割する計算方法を探る。(2)複数の語で構成され，ひとかたまりの意味を持つフレーズである Formulaic Sequences(FS)の全コーパスからの網羅的な抽出と語用論的意味分析による知識化を行い，語を FS に拡張した分散表現とその FS の用途の違いによる分割計算を可能にする。さらに，(3)深層学習による Can-do 会話教材の自動抽出をし，会話から Can-do に対応するキーフレーズを FS 分散表現を用いて抽出する手法の開発を行う。

３．研究の方法

研究組織は，代表者の望月が全体を統括し，分散表現分割，FS 分析と Can-do 教材抽出の 3 つのタスクに分ける。各タスクの役割は次の通りである。

分散表現分割タスクでは，語の多義性を反映した分散表現計算を扱う。word2vec に代表される分散表現では語の持つ複数の語義が単一のベクトルで表現されるという問題がある。このタスクでは，周辺文脈内の語に重要度の重み付けをする手法の検討，語義ごとの周辺文脈の情報を利用して語の分散表現を分割するための意味計算を検討する。

FS 分析タスクでは，開発済みの FS 抽出アルゴリズムの全コーパスへの適用，FS2vec に基づく FS クラスタリングの実装，FS クラスタの場面，機能に基づく分析， FS 分散表現を FS の用法ごとに分割する拡張と，各アルゴリズムの洗練を行う。また，FS を抽出する単位としてこれまでの形態素・単語単位に加えて，文字単位の n グラムを用いる手法も開発するとともに n のサイズを制限しない全 n グラムでの n グラムを扱えるようにする。

Can-do 教材抽出タスクでは，字幕コーパスから会話部分を取り出した上で教師データを作成し，深層学習による Can-do 会話教材の自動抽出手法を開発する。教材会話からのキーフレーズの抽出により Can-do ごとのキーフレーズリストを獲得する。また，FS を核としてキーフレーズを抽出することとキーフレーズ間の関係をとらえるための類似度計算手法を開発し，FS とキーフレーズを結びつける手法を検討する。

3 つのタスクのほか，研究の前提として，テレビ字幕データからの大規模会話コーパスの継続的な構築を行い研究に利用できる形で整備をする。

４．研究成果

開発済みであった MapReduce 型アルゴリズムを改良し，これまで単語単位での n グラムの計算に限っていたものを，文字単位での n グラムでも計算するように拡張した。また，n のサイズが 2 グラムから 9 グラムに限られていたが，文字単位，単語単位のどちらにおいてもサイズ 2 から，各文の最大サイズまでの制限なしの組み合わせで n グラムパターンを作成し頻度計算を行

えるようにアルゴリズムとプログラムの見直しを行い，改めて実装した。さらに，計算した n グラムパターンの中から，重要なパターンである Formulaic Sequences(FS)を抽出するアルゴリズムを再検討し，これまで収集した全コーパスへの適応を行った。この過程で，これまで異なる n グラムどうしの頻度による重要度を直接比較できなかった点を解決する，統合文脈 n グラム分析を開発した。この統合文脈 n グラムでは，まず「1 つの文に対してすべての n グラムを生成し，異なり n グラムごとに出現する行の ID を出現文リストとして登録する」。この処理をコーパス内の全文に適用した後，「出現文リストと n グラムパターンでソートし，同じ出現文リストを持つ n グラムを比較し，短い n グラムパターンが長い n グラムパターンに含まれるなら，短い n グラムパターンを削除する」。これにより，同一の出現文リストを持つ複数の n グラムパターンの中で最長のパターンが有効な n グラムとして抽出できる。この手法で 1 文からすべての n グラムを生成すると，総数は等差級数の和の公式に従う。仮に 1 文が 10 単語からなる場合，単語単位での全 n グラムは 55 個となる。2022 年の時点でこの問題に着手し，当時の我々の字幕コーパス全体では約 1.9 億文であったため，104 億以上の n グラムの全出現位置をリスト化しソート，マージする処理を行うプログラムを実装した。ここから統合文脈 n グラムにより，異なり n グラム数で 117 万超, 延べ n グラム数で 15 億 5 千万超の n グラムパターンを生成した。この中で頻度が 500 以上の n グラムパターンのみを高頻度 FS とした場合に，FS 内の単語をまとめて 1 単語と見立ててコーパス内の総単語数を再計算すると，元が 20.9 億語であったのに対し，16.1 億語に換算され，コーパス内に存在する FS が平均で 44.8%を占めるという結果を得た。この結果は，FS が占める割合が談話全体の 3 分の 1 から 2 分の 1 であるとする応用言語学の報告と一致している。

抽出した頻度上位 10,000 の FS と本学留学生日本語教育センターの教授陣によって作成された教科書「初級日本語」掲載の 334 キーフレーズの一致度を調査した結果，キーフレーズのほとんど（83.2%以上）に自動抽出した FS が含まれており統合文脈 n グラム分析で抽出した FS の質的な有効性を確認した。さらに文末に高頻度で出現する FS に注目すると「そうですね」「ありがとうございます」「しています」「お願いします」「があります」「てください」「でしょうか」のようになり，日本語に特徴的な文末表現を多く含むという新たな知見が得られた。FS のリストを元に，改めてコーパス全文を FS による分割を行い，単語単位に変わる，FS 単位の分散表現である FS2vec の計算を行った。この FS2vec を用いて，k-means 法により抽出された FS のクラスタリングを行うことで FS の分散表現が類似したクラスタ作成を行った。単語単位の分散表現では単語の意味のベクトル化が行われているため，FS 単位でも同様になると予想したが、クラスタリング実験の結果により，意味というよりも用法や機能面が類似した FS が同一クラスタを形成する確かな傾向が確認された。FS のベクトル化による類似度計算では，「表層表現は異なるが，そのフレーズが表す，機能が類似する」FS 間の類似度が高くなるという性質があることが確認できた。

語の分散表現の語義による分割については，語の初期の分散表現が，文脈を取り入れることで変化する BERT などの Transformer モデルが広く普及したことにより，研究開始当初に想定していた開発の必要がなくなり，BERT を用いた深層学習を採用することで，直接取り入れることが可能になった。深層学習による Can-do 会話教材の自動抽出のために，教師データの整備を行った。大規模字幕データ内に含まれる会話部分として，近年日本語を学ぶきっかけとして大きな役割を果たしているアニメーションの字幕に注目し，日本国内および国外で人気のあるアニメーションの中でランダムに 12 作品を選出し合計 100 話分の中から会話部分を人手によりを抽出した。会話の中で本学留学生日本語教育センターによる AJ Can-do リストの初級 1，初級 2，中級 1，中級 2，中上級，上級 1，上級 2 に含まれる Can-do 項目の内容として利用可能な会話を Can-do 会話として対応付けし抽出した。この作業により，19,808 文からなる 1,600 セグメントの会話セグメントと Can-do リストの対応付けを行い「Can-do タグ付与済みコーパス」を作成した。作成したコーパスを談話境界推定タスクの教師データとして BERT をファインチューニングし，自動的に Can-do 会話に対応する会話セグメントを取り出すモデルの構築を行った。談話境界推定タスクを，コーパス内の各文が Can-do 会話セグメントの開始文となるかいなかを判定するタスクと，連続する 2 文を与えその 2 文間に談話境界が含まれるか否かを分類するタスクの 2 種類のタスクとして規定し，それぞれの BERT モデルでの実験を実施した。クロスバリデーションによって実験をした結果，2 つ目のタスクである 2 文の文間関係を「連続」「非連続」に分類するモデルによる精度が最も高く，境界推定の値としてテストセットでの評価値が Precision 70.1%，Recall 63.9%，および F1-Score 66.9%の値を得た。今後更なる性能の向上が望まれるものの，日本語学習のための Can-do 会話を字幕コーパスから自動抽出する手法として一定の成果が得られた。

また，研究開始当初に現在約３１万 9 千番組分，約１億１千 2 百万文分，１２億１千万語の規模だったテレビ字幕コーパスの拡張を行い，約５９万８千番組，約２億８百万文，２２億５千万語超に達する大規模会話コーパスの構築を行った。字幕データ取得のための自動化プログラムを洗練し，継続的な字幕データ収集をより安定した状態で行えるようにシステムの更新を行った。期間中に字幕収集期間が 10 年を超えたことから，2013 年 1 月から 2022 年 12 月までの 10 年間のテレビ字幕全体を通した語彙調査を行い，日本語教材としての字幕データの有用性について確認を行った。語の出現頻度統計をとり 1 年ごとの各上位 100 位の語を比較したところ，10 年間でほとんど順位の変動がなく，頻度上位語による安定的な分布を示すことが確認された。

また，コーパス全体での異なり語数が 672,377 語で，延べ語数が 2,251,883,117 語であったのに対して，出現頻度上位 4,387 語の出現語数がコーパス全体の 90%に達し，上位 10,000 語では，累積頻度が 94.73%を占めるという結果が見られた。およそ 1.49%の語によってコーパス全体の 94.73%がカバーされるという結果が 10 年分の実データによって算定されたことになる。n グラム単位での比較では頻度の高い n グラムでは，ほとんどが文末表現に関係するパターンであるという傾向が見られた。また，字幕コーパス全体での 1 文あたりの平均語数は 10.8 語であったが，アニメーションや映画では平均 6.9 語と短く，ニュース・報道では 19.9 語と長いという放送番組のジャンルによる長さのばらつきが見られた。ジャンルの違いが反映された結果となった。我々の整備した字幕データには，さまざまなジャンルの番組がバランスよく含まれており，話し言葉の均衡コーパスとしての利用価値が高いことが確認できた。

各段階の研究成果については，EDMEDIA，E-Learn などの国際学会 を中心に査読付き論文発表を行った。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3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 望月 源	４．巻 106
２．論文標題 日本語テレビ字幕放送データからの言語データ抽出と特徴の分析	５．発行年 2023年
３．雑誌名 東京外国語大学論集	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 オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Hajime Mochizuki, Kohji Shibano	４．巻 Nov. 01
２．論文標題 Mining Formulaic Sequences from a Spoken Japanese Based on Consolidated Contextualized N-gram Analyses and Its Verification with Key Phrases in Japanese Language Textbooks	５．発行年 2022年
３．雑誌名 World Conference On Educational Media and Technology + INNOVATE LEARNING 2022	６．最初と最後の頁 909-916
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 オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Hajime Mochizuki and Kohji Shibano	４．巻 Nov. 4
２．論文標題 Incorporating a State-of-the-Art Speech Recognition to a Japanese Language e-Learning System	５．発行年 2019年
３．雑誌名 E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2019	６．最初と最後の頁 1157-1162
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　6件）

１．発表者名 Hajime Mochizuki, Kohji Shibano
２．発表標題 Investigation of Formulaic Sequences at The End of Sentence in Japanese Closed Caption TV Corpus
３．学会等名 2023 STEM/STEAM and Education Conference（国際学会）
４．発表年 2022年

１．発表者名
Hajime Mochizuki

２．発表標題
Real Word Statistics and End of Sentence Expressions in Japanese Closed Caption TV Corpus

３．学会等名
9th International Conference on Language, Literature and Linguistics (LLL2022)（国際学会）

４．発表年
2022年

１．発表者名
Hajime Mochizuki, Kohji Shibano

２．発表標題
Extracting Japanese Sentence-Ending Expressions using Formulaic Sequences with Consolidated Contextualized N-gram Analysis

３．学会等名
The 21st Annual Conference of Hawaii International Conference on Education,（国際学会）

４．発表年
2023年

１．発表者名
大河原龍太朗，望月源

２．発表標題
Can-do型日本語学習用資源としてのアニメーション字幕の分析

３．学会等名
言語処理学会第28回年次大会

４．発表年
2022年

１．発表者名
イーフエイチー，望月源

２．発表標題
テレビ字幕データを用いた感情分析による「ある日の日本の気分」推定に関する研究

３．学会等名
言語処理学会第28回年次大会

４．発表年
2022年

１．発表者名 Hajime Mochizuki
２．発表標題 Investigation of Words in a Japanese Closed Caption TV Corpus
３．学会等名 Hawaii University Conferences, STAM/STEAM Education Conference, 2019（国際学会）
４．発表年 2019年

１．発表者名 王 楽淑 and 望月 源 and 鈴木 美加
２．発表標題 中国語母語話者の日本語学習におけるL1L2字幕利用の考察
３．学会等名 言語処理学会第26回年次大会
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
研究分担者	芝野　耕司 (Shibano Kohji) (50216024)	東京外国語大学・その他部局等・名誉教授 (12603)	

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