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　元 年　 ６ 月 １９ 日現在

機関番号：１２６０３
研究種目：基盤研究(B)（一般）
研究期間：2015～2018
課題番号：１５Ｈ０２７９４

研究課題名（和文）大規模会話コーパスのＦＳ２ｖｅｃ処理によるＣＥＦＲ　Ｃａｎ-ｄｏ言語教材の開発

研究課題名（英文）Development of CEFR Can-do Language Learning Materials by FS2vec Processing of Large-scale Spoken Language Corpus

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）　　11,800,000円

研究成果の概要（和文）：我々は，日本語テレビ字幕(CCTV)コーパスからフォーミュライックシーケンス(FS)を抽出する方法を開発した。本研究では，FSの候補として，重要なn-gramをCCTVコーパスから抽出する。各n-gramの出現頻度を計算するため，我々は大量のn-gramをソートし，マージするプログラムをMapReduceアルゴリズムに基づき開発した。会話セグメント内の話題や場面による分類を行い，対応可能なCan-doの存在を確認した。字幕コーパスは拡張を続け，13億語規模に拡大した。研究成果は，AAAL，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
これまで存在していなかった大規模な日本語会話コーパスの構築を続け，６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た。規模は３５万番組，１億２千４百万文，１３億３千６百万語超に達した。この大規模なコーパスから，日本語学習教材にも応用できる特別な意味を持つ複数単語のまとまりであるFormulaic Sequence（定型表現）を大量に抽出した。定型表現を核にして，コーパス内の会話セグメントを取り出し，セグメント内の定型表現が表す機能と，各セグメントの話題，場面をCan-doと対応づけることで有益な教材が作成できることを確認した。

研究成果の概要（英文）：We developed a method for extracting formulaic sequences from Japanese closed caption TV Corpus. In this research we extract signifiant n-grams as candidates for formulaic sequences of continuous words from a CCTV corpus. To calculate n-gram frequencies we developed programs to sort, marge, and count based on the MapReduce algorithm. We examined clustering of discourse segments by topics and scenes and confirmed the existence of suitable can-do statements for them. We have been continuing to build the CCTV corpus.
The total number of words in our corpus has reached over 1,300 million morphemes. Regarding the research results, we presented peer-reviewed papers mainly on international academic societies such as AAAL, 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

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

Big Data がビジネス分野でも学問分野でも大きな注目を集めている。言語学や計算言語学では，Google が公開した Web 1T は従来の British National Corpus の 1 万倍の規模であり，この Big Data を用いての統計的言語処理や用例ベースドアプローチでの研究が進んでいる。言語教育でも，2001 年に最新版が発表されたヨーロッパ発の言語教育の世界標準である CEFR("Common European Framework of Reference for Languages: Learning, Teaching, assessment")では「できる(Can-do)」言語教育に変わってきている。言語学，計算言語学分野で成功を収めている大規模コーパスの利用が期待できるが，十分な規模と種類をもつ話し言葉コーパスの不在から，CEFR 以降の言語教育の革新は進んでいない。言語教育の革新には，(1)大規模話し言葉コーパスの構築，(2)語単位ではなく，会話パターンの抽出を可能にする Big Data 処理アルゴリズム，及び(3)これらを活用できる会話分類のための word2vec アルゴリズムの改良が必要である。

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 5 万 3 千時間分，約 3 千 2 百万文分，3 億 3 千万語(異なり語数 35 万種)の規模になっている。本研究では，このコーパスを更に拡張する。また構築した大規模コーパスから，複数単語を組み合わせた数百万通りの文字列パターンの頻度統計をとり，意味ある文字列パターンである Formulaic Sequence(FS)を抽出する。本研究では，この Big Data に対し MapReduce アルゴリズムを開発して処理する。抽出した FS を含むコーパス内の会話セグメントと，既存日本語教科書の Can-do 会話例を，FS の類似性に基づいて対応付ける。この意味的類似性の判定には,語単位の word2vec を FS 単位に拡張した FS2vec を新たに開発して用いる。これらにより，大規模会話コーパスから FS を核として，Can-do に対応した日本語教材を作成する。

３．研究の方法

研究組織は，代表者の望月が全体を統括し，談話処理班，談話対応班，日本語教材班の三班で構成する。各班の主な役割は次の通りである。

談話処理班では，談話抽出と談話解析の二つの副課題に分かれる。談話抽出は，字幕コーパスから会話のまとまり(会話セグメント）と FS を取り出す。より困難な課題として，会話セグメント内の 1〜5 単語 N グラムの網羅的な組み合わせでの頻度集計を行なう。我々の会話コーパスでは，約 3 千 2 百万文内の数百万パターンを扱う必要のある Big Data であるため，MapReduce 型のアルゴリズムを開発し，各パターンの出現位置や頻度から決まり文句や定型表現と看做せる Formulaic Sequence(FS)の抽出を行なう。談話解析では，会話セグメントの記述内容や含まれる FS から「話題」「場面」「目的」を推測する手法を確立する。

談話対応班は Can-do 会話例と会話セグメントの対応付けを行なう。Can-do では，(a)「飲食店などで店員に、料理や飲み物などを短い簡単な言葉で注文することができる。」(b)「デパートなどの店員に、買いたい物の売り場がどこにあるかなどについて質問し、いくつかの簡単な答えを理解することができる。」などのように「〜できる」目標が文によって定義されている。この「できる」(Can-do)の定義文と会話例を，会話コーパスから抽出した会話セグメントや FS パターンと比較し，対応付けを行なう。日本語教材班は Can-do に対応付けられた会話セグメントの有効性を評価し，日本語教材化する。

また，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

研究開始当初に 5 万 3 千時間分,3 億 3 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 35 万番組，約 1 億 2 千 4 百万文分，13 億 3 千 6 百万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実装し，継続的な字幕データ収集を可能とした。収集した元の字幕データはテレビ画面の表示サイズの制限があるために完全な文ではなく断片化していることが多い。そこで，字幕データ内で断片化している複数の文字列をつないで文を復元するプログラムを実装した。文取得プログラムでは字幕の表示時間の連続性および，句読点表示の有無などに基づいて処理を行なった。ただし，句読点を用いていない番組については，個別の番組データの表記を参照し，規則性が見られた番組についてはプログラムに反映させることで，文復元の精度を向上させた。

また、MapReduce 型アルゴリズムのプログラムを開発し，単語の N グラムの組み合わせパターンを作成し，頻度統計を計算した。N グラムは当初約 3 千 2 百万文を対象に N=2 から 5 の範囲で開始したが，最終的には約 1 億 2 千 4 百万文を対象に N=2 から 9 までの範囲に対応した。コーパス内の文に出現するこの N グラムパターンの中から、出現文が完全一致するパターン群で最長のパターンを残すアルゴリズムにより，Formulaic Sequences(FS)の重要候補を取り出す

ことができた。この FS を形態素解析辞書として整備し，取り込むことで，形態素の代わりに FS を用いて文を分割できるようにした。全ての文を FS で分割したコーパスによる FS の頻度データを用いることで，word2vec を FS ベースに拡張した FS2vec の実現を行った。word2vec の distance に相当する計算を FS2vec を用いて行い，FS のグループ化を行ったところ，word2vec の word とは異なり，word 単位よりも長い表現となる FS では距離の近いものどうしが意味の近いグループになるとは限らず，意図や目的を同じくする表現のグループ化が見られるという FS に特有の特徴がみられた。また，比較的長い文字列で構成され，出現頻度も多い FS に絞り込むことと，各 FS の字幕データでの出現について番組ジャンルによる偏りを考慮することで，より実用的な FS のリストを取り出すアルゴリズムの実装も行なった。この FS 抽出アルゴリズムは抽出の際に FS の出現位置を記録する記号化の書式を研究開始当初の方式から改善し，今後増え続けるデータの世代管理を可能にする改良を行なった。この改良により，これまで，データ量の関係で形態素単位でのNグラムに限定していたFSの抽出を文字単位でのNグラムで行うことが可能になった。

さらに，FS をキーとした会話セグメントの検索，FS2vec 型の Doc2vec によるセグメント間類似度計算，SVD での次元縮退，K-means 法によるクラスタリングの一連のプログラムを完成させた。これにより特定の FS から開始して，同一 FS 表現を含む会話セグメント集合を取り出し，さらに取り出された会話セグメントを全 FS の類似度によってクラスタリングすることで，話題，場面の異なるクラスタを生成できるようになった。具体的な対象として，ドラマ，バラエティ，情報番組の中で「いらっしゃいませ」「ください」を含む約 1800 会話セグメントを 15 クラスタに分割し，会話セグメント内の話題や場面による分類と対応可能な Can-do を調査した。大量の事例の中に，特定の Can-do との対応付けが可能な事例が存在することを確認した。

また，各クラスタ内の文について，存在する FS を取り出したところ，1 つのクラスタに複数回出現する FS が一定数存在すること，また FS の多くは文末の位置に出現すること，話題や場面に関する語までを含む FS は多くないことがわかった。この結果から，会話セグメントと Can-do の対応として，会話セグメント内の FS が Can-do の機能面に関係し，FS 以外の名詞周辺の語は Can-do の話題や場面に関係すると想定して対応づけを行うことで，教材作成が行えることが確認できた。各段階の研究成果については AAAL，AiLA，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　14　件）

① Hajime Mochizuki，Investigation of Words in Japanese Closed Caption TV Corpus，2019 Hawaii University Conferences，STEAM Education Conference，査読有，2019，12 pages.
② Hajime Mochizuki and Kohji Shibano，Analyzing Usefulness of Dialogues from Closed Caption TV Corpus as an Example of Can-do Statements for Language Learning，2018 Hawaii University Conference，Arts，Humanities，Social Sciences & Education (AHSE)，査読有，2018，11 pages.
③ Hajime Mochizuki and Kohji Shibano，Searching Discourse Segments for Formulaic Sequences in a Closed Caption TV Corpus for Language Learning，World Conference on E-Learning in Corporate，Government，Healthcare，and Higher Education 2017，査読有，2017，pp.19-27.
④ Hajime Mochizuki，Augmented Reality Applications for Multilingual Learning with Intuitive Understanding，Proceedings of World Conference on Educational Media and Technology (EDMEDIA) 2017，査読有，2017，pp.1205-1213.
⑤ Hajime Mochizuki and Kohji Shibano，The Acquisition of a Japanese Practical Formulaic Sequences List from a Closed Caption TV Corpus，2017 Hawaii University Conferences，STAM/STEAM Education Conference，査読有，2017，6 pages.
⑥ Hajime Mochizuki and Kohji Shibano，Extracting Formulaic Sequences Containing Useful Expressions for Language Learning from Closed Caption TV Corpus，World Conference on E-Learning in Corporate，Government，Healthcare，and Higher Education，E-Learn 2016，査読有，2016，pp.29-37.
⑦ Hajime Mochizuki，Development of AR Materials for Understanding Roles of Japanese Particles，2016 Hawaii University Conferences，STEAM Education Conference，Hawaii，USA，査読有，2016，6 pages.
⑧ 芝野耕司，データベース・データ工学研究者・技術者にとっての SQL 規格の意味，日本データベース学会　Newsletter，無，Vol.9，No.1，2016.
⑨ Hajime Mochizuki and Kohji Shibano，Analyzing Attractiveness of Specific Location Names of Tourist Destination from a Closed Caption TV Corpus，2016 Hawaii University Conferences，Arts，Humanities，Social Sciences & Education (AHSE)，査読有，2016，12 pages.

⑩ Hajime Mochizuki and Kohji Shibano, Detecting Topics Popular in the Recent Past from a Closed Caption TV Corpus as a Categorized Chronicle data, the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KMIS), 査読有, 2015, pp.342-349.
⑪ Hajime Mochizuki, Development of Sample Code Stocks for AR Applications in Language Learning, the 6th International Conference on Education and Management Technology-ICEMT2015, 査読有, 2015, 6 pages.
⑫ Hajime Mochizuki and Kohji Shibano, Development of a Closed Caption TV Corpus Retrieval System to Seek Video Scenes Containing Useful Expressions for Language Learning, World Conference on Educational Media and Technology (EDMEDIA), 査読有, 2015, pp.1760-1768.
⑬ Keiko MOCHIZUKI, Hiroshi SANO, Ya-Ming SHEN, Chia-Hou WU, ″Cross-Linguistic Error Types of Misused Chinese Based on Learners' Corpora″, 査読有, Computational Linguistics and Chinese Language Processing, Vol. 20, No. 1, June 2015, pp.97～113.
⑭ Hajime Mochizuki and Kohji Shibano, Re-Mining Topics Popular in the Recent Past from a Large-Scale Closed Caption TV Corpus, International Journal of Future Computer and Communication, 査読有, vol. 4, no. 2, 2015, pp. 98-103.

〔学会発表〕(計 13 件)

① Hajime Mochizuki, Building a Very Large Spoken Language Corpus from Closed Caption TV and Extracting Practical Formulaic Sequences for Language Learning, The 10th International Conference on Advanced Computer Theory and Engineering, 招待講演, Korea, August 2017.
② Hajime Mochizuki and Kohji Shibano, Discourse Segment Clustering with Word Embedding based on Formulaic Sequences for Language Education, 2017 International Conference on Education and Multimedia Technology (ICEMT 2017), 査読有, Singapore, July 2017.
③ Kohji Shibano, Analyzing formulaic sequences in spoken Japanese from a large Japanese TV closed caption corpus, The 18th World Congress of Applied Linguistics (AILA 2017), 査読有, 23-28 July 2017, Rio de Janeiro, Brazil.
④ XIAO Tingting, Kohji Shibano, Developing Intimacy by Style-shifting in Japanese: A TV Subtitle Corpus-based Study, The 2017 conference of the American Association for Applied Linguistics (AAAL 2017), 査読有, 18-21 March, 2017, Portland, USA.
⑤ Hajime Mochizuki, Simplified AR Language Learning Environment using 3D Letter String Objects, International Conference on Virtual and Augmented Reality Simulations (ICVARS 2017), 査読有, Australia, February 2017.
⑥ Hajime Mochizuki and Kohji Shibano, Modification of word2vec by Formulaic Sequences and Extraction of Useful Expressions for Language Learning from Closed Caption TV Corpus, The IAFOR International Conference on Language Learning, 査読有, USA, January 2017.
⑦ Hajime Mochizuki, Development of a Closed Caption TV Corpus Retrieval System for Language Learning, 8th International Conference on Education Technology and Computers (ICETC 2016), 査読有, September 2016.
⑧ Hajime Mochizuki, Straightforward Expansion of word2vec by Formulaic Sequences in CCTV corpus, Ninth International Conference on Advanced Computer Theory and Engineering, 査読有, Hong Kong, August 2016.
⑨ Hajime Mochizuki, Japanese Language Learning System for Understanding a Sentence that has Correct Syntax but has Semantic Errors, 2016 the 2nd International Conference on Information Technology (ICIT 2016), 査読有, Melbourne, Australia, 3-4 March 2016.
⑩ 芝野耕司,SQL 言語の開発と日本の講演, DEIM2016 日本データベース学会功労賞記念講演, 招待講演, March 1, 2016.
⑪ Hajime Mochizuki and Kanetaka Abe, Building an event log management system to acquire users' operation behavior on a large number of client PCs, the 2nd International conference on Networks and Information Security (ICNIS), 査読有, Singapore, October 2015.
⑫ 芝野耕司, 日本語話し言葉コーパスの 構築と会話用例検索システム, 6th CASTEL/J Hawaii, 査読有, 2015 年 08 月 07 日~2015 年 08 月 08 日, Honolulu, USA.
⑬ Kohji Shibano, A Quantitative Formulaic Analysis of Large TV Closed Caption Corpus - Pragmatic Use of Utterance End in Japanese Animation Languages, 14th International Pragmatics Conference, 査読有, 26-31 July 2015, Antwerp Belgium.

〔図書〕(計 2 件)

① 芝野耕司共訳，日本規格協会，JIS X 3005-2: 2015 データベース言語 SQL 第 2 部:基本機能 (SQL/Foundation)，2015,1112
② 芝野耕司共訳，日本規格協会，JIS X 3005-14: 2015 データベース言語 SQL 第 14 部:XML 関連仕様 (SQL/XML)，2015，334

６．研究組織

(1)研究分担者

研究分担者氏名：芝野 耕司
ローマ字氏名：Kohji Shibano
所属研究機関名：東京外国語大学
部局名：その他部局等
職名：名誉教授
研究者番号（8 桁）：50216024

研究分担者氏名：佐野 洋
ローマ字氏名：Hiroshi Sano
所属研究機関名：東京外国語大学
部局名：大学院総合国際学研究院
職名：教授
研究者番号（8 桁）：30282776

研究分担者氏名：藤村 知子
ローマ字氏名：Tomoko Fujimura
所属研究機関名：東京外国語大学
部局名：大学院国際日本学研究院
職名：教授
研究者番号（8 桁）：20229040