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　２ 年　 ６ 月 １６ 日現在

機関番号：３２６４２
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ００４８９

研究課題名（和文）文書類似度を利用した英語学習用例の自動生成

研究課題名（英文）Automatic collocation generation for English learners as a foreign language using document similarity analysis

研究代表者

来住　伸子（Kishi, Nobuko）

津田塾大学・学芸学部・教授

研究者番号：５０２４５９９０

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）：この研究では、潜在意味解析、語頻度、語頻度・逆文書頻度の３種類の文書類似度評価方法を利用して、英語学習者の興味や習熟度に適した用例を自動生成し、実際に学習者が用例を評価することを目指した。使用ハードウェア、ソフトウェアの改善により、先行研究より文書類似度計算を高速化できた。用例の難易度の推定にも、文書類似度を利用することにし、難易度の異なる複数の文書集合との距離から難易度を推定することにした。その結果、文書集合の種類や大きさが先行研究より増大した。そこで、潜在意味解析だけでなく、他の類似度計算方法、word2vecなどの浅い機械学習による類似度評価方法も利用することにした。

研究成果の学術的意義や社会的意義
この研究は、第2言語として英語を学ぶ学習者に、学習者の興味や習熟度にあった教材を自動生成する研究の一環として行っている。社会人や大学生の英語学習者の場合、本人の仕事や専門分野で実際に使われる表現の習得を効率的に行うことが望ましいが、適した教材（教科書、書籍、動画など）は非常に少ない。一方、Wikipediaや各種オープンコンテンツの普及により、英語テキストは入手しやすくなっている。そこで、情報検索分野で使われている、潜在意味解析、頻度分析などの手法を利用して、大規模テキストデータから、教材の素材となる用例（英語の分離）の自動抽出を行った。

研究成果の概要（英文）：This study uses three types of document similarity evaluation methods: latent semantic analysis, bag of words, term-frequency and inverse document frequency, to generate English collocations for the learners of English as a foreign language. In the previous study, we find the latent semantics analysis is more suitable for generating collocations for English for specific purposes. However, the generated collocations were not usable as real learning materials because the difficulty level of collocations are not considered, and the subject area is limited. In this study, we used more computational resources to increase the speed of calculation and the quantity of documents. Furthermore, we used two sets of documents: an easy set and a difficult set, to estimate the difficulty level of collocations based on the different similarities to the two sets. We also added other algorithms to calculate the similarity from shallow machine learning algorithms such as word2vec.

研究分野：情報工学

キーワード：英語学習　文書類似度　文書分類　潜在意味解析　教材自動生成　機械学習　語彙学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

大学生や社会人による外国語としての英語学習では、学習者の興味や専門分野、学習者の習熟度に合わせた適切な教材が必要とされている。しかし、外国語としての英語学習教材作成は、少数の出版社で限られた人材で行われており、種類、量ともに限られている。最近の話題、専門性の高い話題を扱った英語教材を見つけることは年々難しくなっている。

先行研究では、入手しやすくなった、Wikipedia などの英語によるオープンデータ、コーパスを利用し、情報検索やテキストマイニングのための各種アルゴリズムを利用して、与えられた語の並び（トピックに相当）に関連した英文用例（文やフレーズ）を生成した。

しかし、計算資源に限りがあったため、限られたトピックに関する用例しか生成できなかった。また、学習者の習熟度にあわせた難易度を考慮した用例の生成には至らなかった。

２．研究の目的

本研究では、先行研究より計算機資源を増やすことにより、大規模なテキスト処理を行ない、より広い範囲の内容の用例生成を可能にするとともに、難易度を考慮した用例生成を行うことを目指した。また、生成した用例を、学習者に評価してもらい、本研究で推定した難易度と、学習者が感じる難易度の違いなども評価することを目指した。

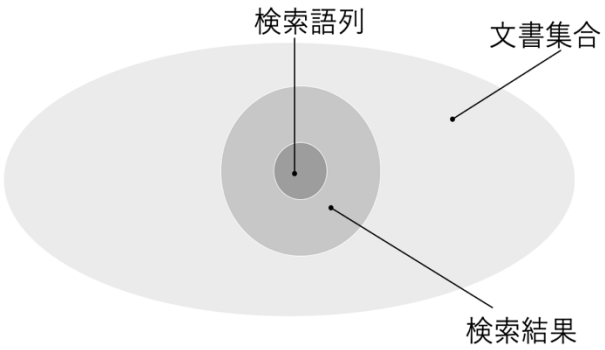

図１：通常の文書検索

図１は、現在使われることの多い、情報検索のイメージ図である。検索語列と、大規模な文書集合の各文書の距離を計算し、距離の短い文書を検索結果とする。距離の計算には、各種の方法があり、語彙頻度、語彙頻度―逆文書頻度、潜在意味解析、機械学習に基づく方法などが使用される。

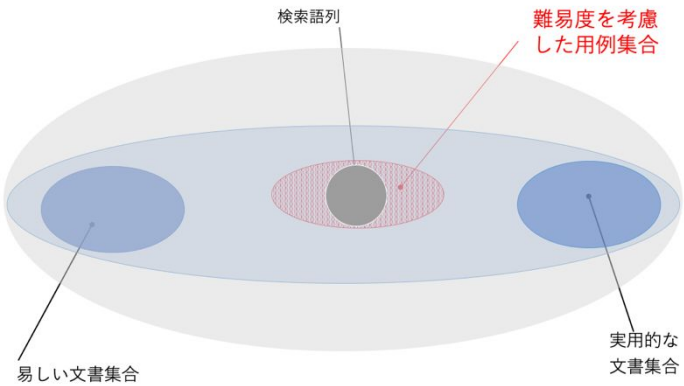

図２：英語学習者のための用例検索

図２は、本研究で行った用例生成の方法のイメージ図である。情報検索分野と同様の文書の距離計算を使用し、さらに、語のフィルタリングなどの手法が情報検索と異なるので、「文書類似度」と呼ぶことにした。また、易しい文書集合、実用的な文書集合（青の部分）、特定分野の文書集合（薄い青の部分）との距離計算を行うことにより、難易度を考慮した用例集合（赤の網掛け部分）を生成することを目指した。

３．研究の方法

より高性能な CPU を持ち、GPU 付きの計算機を購入し、計算環境を整備した。整備した計算機上で、オープンソフトウェアを利用して、語彙頻度(bag of words)，語彙頻度-逆文書頻度（tf-idf），潜在意味解析（latent semantic analysis）などのアルゴリズムを使って大規模テキストデータの文書類似度を計算した。

難易度は、易しい文書データ（例: Simple Wikipedia, Graded Reader） と母語話者向けの文書データ（例：通常の Wikipedia, 英語で書かれた専門書）の２種類の文書データそれぞれと、用例の文書類似度を計算し、二つの文書類似度から、文書の難易度を推定することにした。

４．研究成果

より大規模な文書類似度計算を行う環境は整備したが、難易度の推定や文書データの大規模化のためには、不十分であった。計算量が少なくてすむアルゴリズム、word2vec, phrase2vec などを使い、限られた範囲のテキストデータで、難易度推定を行なった。テキストデータの整備、アルゴリズムの選定評価、学習者が用例検索するためのソフトウェア開発、などに予想以上に時間がかかり、推定した難易度を実際の学習者に評価してもらうことは、研究期間内にできなかった。研究期間終了後も、このテーマに関する研究を継続し、学習者による難易度評価と、推定した難易度の関連を調査する予定である。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔図書〕　計1件

１．著者名 Nobuko Kishi, Mari Yoshida, Minori Yoshizawa, Aoi Yoshida	４．発行年 2019年
２．出版社 Vilnius University Faculty of Philosophy and Institute of Data Science and Digital Technologies	５．総ページ数 9(961)
３．書名 Visualization Tool for Finding Characteristics of Teaching and Learning Process of Scratch Programmers, Constructionism 2018 conference proceedings	

〔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	岸　康人 (KISHI YASUHITO) (50552999)	神奈川大学・付置研究所・研究員 (32702)	
研究分担者	田近　裕子 (TAJIKA HIROKO) (80188268)	津田塾大学・総合政策学部・教授 (32642)	
研究分担者	久島　智津子 (KUSHIMA CHIZUKO) (80623876)	津田塾大学・言語文化研究所・研究員 (32642)	