D-003

潜在意味解析などを利用した英文用例の自動生成
Using Latent Semantic Analysis for Generating Multiword Expressions for English Language Learners

来住 伸子† 岸 康人‡ 久島 智津子† 田近 裕子†
Nobuko Kishi Yasuhito Kishi Chizuko Kushima Hiroko Tajika

1. はじめに

英語を母語としない英語学習者、とくに社会人の英語学習者が英語を効率的に学ぶには、実際に利用する可能性の高い語彙や用例を使った教材が必要である。そのような英語教材を作成する第一歩としてコーパスの利用が始まっており、出現頻度による語彙リスト作成や、人による語彙リスト作成が実際に行われている。しかし、中級レベル以上の学習者や、特定分野の英語を学びたい英語学習者に向いた語彙リストや、その語彙を文脈とともに学べるような用例データは、十分には作成されていないのが現状である。

そこで、入手しやすくなった英語テキストデータと、低価格な計算資源を活用して、個々の学習者に適した語彙リストとその用例データを自動生成するサービス COOLL4C を開発することにした。

この報告では、そのサービスの設計方針と、準備調査として行った、英語 Wikipedia テキストデータで使われる語彙頻度、語彙カバー率、潜在意味解析などの各種アルゴリズムを利用した類義語、同形異義語の占める割合について報告する。

2. 背景

2.1 コーパス

コーパスの普及により、書籍、新聞などで実際に使用された英語を低価格で入手できるようになった。Brown Corpus や British National Corpus がその先駆者である。現在入手しやすいコーパスは、Corpus of Contemporary American English (COCA) で、約 4.5 億語分のデータが提供されている[1].

2.2 学習語彙リストと語彙カバー率

外国語を学ぶ際には、語彙学習が欠かせない。そこで、コーパスを利用して、学習者が使う可能性が高い語彙リストを生成し、それを学ばせることが試みられている。Nation をはじめとする研究グループ[2,3]は、まず、コーパスを利用して、語彙を頻度順で順序をつけた。教養のある英語母語話者は、約 20,000 語の語彙を知っているが、そのうち、頻度の上位 2000 語と、Academic Vocabulary 636 語を知っていれば、大学の経済学の教科書に出てくる語彙の 91.2%を知っていることになる（語彙カバー率 91.2%）という報告がされている。英語母語話者が知っている語彙数はかなり多いが、大学レベルの教科書の読解には、かなり少ない語彙数で十分であることになる。

†津田塾大学 Tsuda College
‡神奈川大学 Kanagawa University

2.3 語彙と読解力

しかし、Nation らの提案したコーパスに基づく語彙学習は必ずしも普及していない。まず、受容語彙や発表語彙の総量が多くても、必ずしも読解力やコミュニケーション能力の向上に結び付かないことが指摘された [4]。また、コーパスと、実際に学習者が触れるテキストデータが一致していないと、語彙カバー率が低くなることがある。Nation の学習語彙リストは必ずしも日本人学習者に適していないということが動機となって、大学英語教育学会は、大学英語教育学会基本語リスト (JACET 8000)を発表した[5]。

現状では、特定の目的（大学入試、TOEIC 対策など）用に英語教員がコーパスを利用しつつ、語彙リストを編集する、という状況になっている。また、そうやって選んだ語彙を、単語カードで暗記するのではなく、語彙の意味に焦点をあてた理解可能なインプット(テキストや音声)を使って学ぶことが勧められている。

2.4 語彙リストと用例の自動生成

コーパスや、ユーザが用意したテキストデータを解析して出現頻度を表示するサービスはすでに提供されている。しかし、個々の学習者に適したテキストデータを、教員や学習者以外の人間が準備することは手間がかかる、ユーザインタフェースが英語圏の英語教員対象になっている、などの理由から、日本の英語教育では必ずしも利用されていない。

3. COOLL4C プロジェクトについて

3.1 プロジェクトの目的

大規模テキストデータから語彙リストと用例リストを自動生成するサービスを提供することを目指して、COOLL4C と名付けたプロジェクトを開始した。このサービスのユーザとして、どの英語テキストを読めるようになりたいか明確に指定できる中級の英語学習者と、学生に適した教材、とくに語彙と用例リストを作成したい英語教員の 2 グループのユーザを想定している。これらのユーザが興味を持っている大量のテキストから、学習すべき語彙や用例を簡単に入手できることを目指している。

3.2 使用データと主な利用技術

利用するテキストデータとして、第一段階は、英語 Wikipedia のアーカイブデータを主に利用することにした[6]。 前述の COCA は、現在は、英語の Wikipedia を含み、Project Gutenberg, Web から集めたデータなど、英語 Wikipedia より大きなテキストデータを提供している。しかし、本プロジェクトの第一段階では、日本の英語学習者

が読む可能性が高いデータに限定して、データ量を抑えることにした。また、著作権の面からも、Wikipedia から直接ダウンロードしたデータを利用するほうが公開しやすく、評価しやすい。将来は、著作権管理を行い、COCA 全体のデータや電子書籍データに対応することを考えている。

Wikipedia からダウンロードしたデータは,Wikipedia Extractor というライブラリ[7]を使い、記事の本文に該当するテキストデータを抽出した。編集データは含めていない。

次に、記事の本文データを、カンマや空文を元に文単位に分割した。一文を一文書として、語文書行列を生成した。語文書行列の生成、TF-IDF 値行列の生成、潜在意味解析には、gensim ライブラリを使用している。

Gensim ライブラリ[8] は、Python 上のベクトル空間処理ライブラリである。潜在意味解析のための行列変換や各種の行列空間での距離計算、類似度計算を行うことができる。genism が提供する各種アルゴリズムのうち、頻度数, TF-IDF, 潜在意味解析（LSA）の３種を COOLL4C 空間生成に使用している。３種の距離空間それぞれにおいて、入力した単語列に近い文を距離順に表示することにより、用例リストを表示する。そのため、用例と呼んでいるが、実際には、ユーザが入力した単語列と次のような関係のある、Multi-word expression や類義語を表示する可能性がある。

- 用例　collocations
- 2 語熟語　binomials
- 複合動詞 multi-word verbs
- 慣用句　idioms
- 定型表現 lexical bundles
- 類義語 synonymous words

4. 準備調査

4.1 調査の方法

このサービスの理想的な評価方法は、学習者にもたらす学習効果を測定することだと考える。サービスの完成前に、そのような測定は無理なので、語彙と用例リストについての準備調査を行うことにした。英語教員が学習者に学習を勧めたい表現が、自動生成した用例リスト（Multi-Word Expression）に含まれるかどうかを推定することを試みた。

まず、学部２年生に英文読解を Computer の歴史に関する本を使って教えている英語教員に面接し、学生に学習させたい点について意見を尋ねた。語彙学習については、次のような意見が得られた。

- 頻度が上位の語彙は必ず学んでほしい。
- 頻度が上位の語彙は、高校で学んだ意味とは異なる意味で使われる語彙（同形異義語）も多い。そのような語彙について用例も学んでほしい。
- 頻度が下位の語彙、初めて見る語彙は自分で調べる習慣をつけてほしい。
- 頻度リストに掲載されない、固有名詞や専門用語も自分で調べてほしい。が、専門教員の解説も必要である。

そこで、対象としている 英語 Wikipedia のデータの語彙頻度を調査し、語彙カバー率を調査した。次に、上位語（頻度の高い語）の一部に対して用例リストを３種のアルゴリズムで生成し、その中に同義語や、同形異義語を含まれる割合を調べた。

4.2 語彙カバー率

英語 Wikipedia の場合、上位 2000 語でカバーできる語彙は、前述の語彙学習に関する研究[4]でとりあげられたテキストと、ほぼ同じ割合になっている。しかし、Academic Word List や上位 8000 語を加えても、カバーできない語がより多く残る。これは Wikipedia で、固有名詞や専門用語が使われる割合が、通常のテキストより高いためと考えられる。

4.3 類義語や同形異義語の生成の割合

潜在意味解析(LSA)の大きな特長の一つは、同じ文脈で使われる異なる表現を抽出できる、意味的によく似た語彙を抽出できる点である[10]。つまり、類義語の生成には適したアルゴリズムであるが、同形異義語を排除することも多い. 一方、語彙頻度や TF-IDF を使ったアルゴリズムでは、類義語はあまり抽出しないが、同形語は確実に抽出する。

そこで、語彙頻度や TF-IDF では用例として抽出され、LSA では用例として抽出されなかったような文を、同形異義語を含む文として利用することの検討を始めることにした。

5. 今後の課題

準備調査の結果、使用するコーパスによって、語彙頻度順や語彙カバー率が、下位語において大きく異なることが分かった。英語 Wikipedia 以外のコーパス、COCA のコーパスなどに対して、同様の準備調査を早急に行う必要がある。

また、LSA は、類義語を含む用例抽出に適しているが、同形異義語の生成にはあまり適していないことが観察できた。頻度数や TF-IDF を利用したアルゴリズムとの併用方法の検討が必要である。

今後、COOLL4C の設計と実装をすすめ、実際の学習者や英語教員による、自動生成した語彙リストと用例の評価を実施したい。

謝辞

本研究は JSPS 科研費 253300417 の助成を受けたものです。

参考文献

[1] http://corpus.byu.edu/coca/
[2] I.S.P. Nation, Learning Vocabulary in Another Language, Cambridge University Press (2001)
[3] I.S.P ネーション, “英語教師のためのボキャブラリーラーニング”, 松柏社, 2005.
[4] Batia Laufer, Geke C. Ravenhorst-Kalovski “Lexical threshold revisited: Lexical text coverage, learners’ vocabulary size and reading comprehension”, Reading in a Foreign Language April 2010, Volume 22, No. 1
[5] 大学英語教育学会基本語改訂委員会 ,”“大学英語教育学会基本語リスト JACET List of 8000 Basic Words”, 大学英語教育学会 (2003).
[6] http://dumps.wikimedia.org/enwiki/20150403/
[7] http://corpus.byu.edu/
[8] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor
[9] https://radimrehurek.com/gensim/index.html
[10] Thomas K. Landauer et. al “Handbook of Latent Semantic Analysis” Psychology Press (2007)