Wikipedia 英語版コーパスを利用した冠詞と複合名詞に関する英語学習用問題自動生成システム

The System of Automatically Article and Noun Compounds Problem Generation with Wikipedia

岩田　陽也[*1], 梅村　祥之[*1]
Yoya Iwata[*1], Yoshiyuki Umemura[*1]
[*1] 広島工業大学　大学院工学系研究科
[*1]Hiroshima Institute of Technology Graduate School of Science and Technology
Email: m161501@cc.it-hiroshima.ac.jp

あらまし：本稿では英語学習用問題の自動生成システムを扱う．日本人英語学習者にとって可算,不可算,複合名詞の判別は難しい.そこで本稿では単語辞書から名詞を取り出し,表現の出現頻度を調べ,複合名詞問題を生成した．生成表現が学習に効果的か否か例を挙げて提示した.また冠詞の問題を生成し優位差検定を行った.その結果，複合名詞問題では適切な問題を生成できた.
キーワード：学習支援システム，問題生成，英語表現

1. はじめに

多くの日本人の英語学習者にとって名詞が可算名詞なのか,不可算なのかの判断は難しく,冠詞の誤用につながる．また，判定,複合名詞の書き方の判定は難しい.

自然言語処理により冠詞を機械判定する研究が行われているが[5],問題生成を扱った研究は比較的少ない．そこで，本研究の 1 つは，冠詞を判定する際に必要となる可算名詞と不可算名詞の問題の自動生成を扱う.

複数の名詞からなる複合名詞はことばを数限りなく生成できる便利な表現である．自然言語処理の分野で複合名詞の微細構造を解析する研究が行われている.文献[6]は日本語複合名詞の解析の研究である.複合名詞には，例えば evening news のような「名詞 1 名詞 2」の場合と news of evening のような「名詞 2 of 名詞 1」の 2 種類がある．学習者はどちらが適切であるのか分らない場合がある．本研究では，この 2 者を判別する問題を自動生成する.

2. 複合名詞問題生成方法

2.1 問題生成法

複合名詞問題を生成するために使用する単語を設定する.「英辞郎第三版[1]」中の初級に該当する標準語彙水準SVL12000 が 1 の単語 1,000 語を抽出する.ここから品詞が名詞のみである 186 個を取り出す.186 個の名詞単語から 2 語を取り出して複合名詞表現を作成する．作成できる表現の数は順列の計算により 186×185=34,410 通りである．以下，この段階で生成された表現を「単純生成表現」と呼ぶ．このようにして作成した複合名詞表現にはとてもあり得ないような不適切な表現が数多く含まれ問題文として適切でない.

英語版 Wikipedia[4]には約 30 億ワードという大量のテキストが含まれている.そこで,これを使い,表現の出現頻度を利用して適切な表現を抽出する.まず，第 1 段階として，Wikipedia の一部である約 1 億 4 千万ワードのテキストから名詞 2 語からなる複合表現「名詞 1 名詞 2(例: potato salad)」または「名詞 2 of 名詞 1(定冠詞，不定冠詞，無冠詞のいずれかを伴う)（例: salad of potato)」の出現頻度を計測する．その結果，複合表現が 1 回でも出現した複合表現数は単純生成表現 34,410 表現のうち,1,707 表現となった．なお，Wikipedia の一部を使ったのは，計算時間短縮のためである.

次に，詳細な計算として Wikipedia 全文を使い,先の 1,707 表現について，出現頻度を計測する．さらに，問題としての適切さを考えると，「名詞 1 名詞 2」と「名詞 2 of 名詞 1」の出現頻度の比率が 1:1 に近い場合，どちらの表現も使えるため，問題として不適切である．そこで,「名詞 1 名詞 2」の出現頻度と「名詞 2 of 名詞 1」の出現頻度の常用対数値が 1.5 より大きい，あるいは-1.5 より小さいという条件を設ける．この条件と,「名詞 1 名詞 2」あるいは「名詞 2 of 名詞 1」の出現頻度の合計が 50 回以上という 2 条件で抽出する．その結果，先の 1,707 表現から 351 表現に絞られ,最終結果となった.以下,これを「抽出表現」と称する.

2.2 抽出結果

単純生成表現から 5 例を示す.

〈名詞 1 名詞 2〉	〈名詞 2 of 名詞 1〉
car spoon	spoon of car
library shirt	shirt of library
hair zoo	zoo of hair
notebook baker	baker of notebook
lion sugar	sugar of lion

あり得ないような表現が多く生成されてしまう.

次に，得られた抽出表現から 5 例を示す.

〈名詞 1 名詞 2〉	〈名詞 2 of 名詞 1〉
evening news	news of evening
potato salad	salad of potato
television movie	movie of television
farm animal	animal of farm
student apartment	apartment of student

妥当な表現が数多く得られる．また，これらは英和辞典のエントリーにも掲載されておらず，コーパスを利用して初めて得られるものである.

4. 可算／不可算問題生成方法

4.1 問題生成法

可算名詞か不可算名詞かを判定する問題を自動生成するために,まず始めに,対象とする単語を設定する．設定にあたって,「英辞郎第三版[1]」に収録された単語の中で初級に該当する標準語彙水準 SVL12000 が 2 の単語 1,000 語を抽出する．その中から，品詞に名詞を含む語で，固有名詞以外の語を抽出する．ここで，固有名詞の認定は簡易的に大文字で始まる語とする．次に，辞書を用い，それらの名詞が可算名詞か不可算名詞か一意に定まるのか，語義によって可算，不可算が異なるといった理由等により一意に定まらないかを判定し，一意に定まる語を抽出する．可算，不可算の記載は辞書によって異なる．今回，WordReference[2]を用いる．この段階で抽出された語彙セットを以下,「全体選定語彙」と称する．この段階が，本研究における語彙選択の妥当性を評価する際の比較対象となる語彙セットである.

以下が，可算，不可算の判定問題に相応しい学習者にとって判定が難しいと思われる語彙を選定するステップである．具体的な物を表す名詞は容易に可算名詞と推測できると考え，抽象名詞を選定する．具体的には，シソーラス辞書 WordNet[3]を用いてシソーラス階層の中に具体物であることを示す「physical entity」が含まれれば選定しないようにする．複数の語義がある場合，いずれかの語義に「physical entity」があれば選定しない.

最後に，Wikipedia 英語版コーパスから，対象とする単語を含む文を検索し，その単語の品詞が名詞である頻度を調べる．Wikipedia コーパスは複合名詞の問題作成処理と同じであるが，その一部である約 1 億 word を使う.出現頻度が少ない語を削除する．削除の閾値として 100 語以下を設定する．以上で問題用の語彙が抽出される．これを「提案選定語彙」と称する.

4.2 評価実験の方法

本語彙選定方法の妥当性を評価するために，次のように問題を生成し，実験参加者による評価実験を行う.「全体選定語彙」は難易度が低く,「提案選定語彙」は難易度が高くなり，学習用の問題として妥当となることが事前予測である.

「提案選定語彙」に含まれる可算名詞，不可算名詞それぞれと同数,「全体選定語彙」中からランダムに，可算名詞，不可算名詞を選択し,「提案選定語彙」全体とランダムに混ぜ合わせて可算，不可算名詞判定問題を作成する．実験参加者は，各語に対し「可算名詞」,「不可算名詞」,「可算，不可算を判定できない」,「この単語を知らない」の 4 択で解答する.

5. 可算／不可算問題の生成・評価結果・考察

英辞郎のレベル 2 は 1,000 語を含む．品詞に名詞を含む語を抽出した結果，725 語抽出された．そのうち可算，不可算が一意に決まる語として，418 語に絞られた．次に，抽象名詞に絞り，247 語となった．最後，コーパス中の出現頻度が閾値以上のものとして 84 語に絞られた.

この中からアルファベット順に並べたときの先頭と最後から計 10 語を示すと次の語となる.

accent，accident，advice，alphabet，anger，
vote，warning，wedding，win，wish

評価実験用にこの 84 語を「提案選定語彙」に採用し，それと同数の可算，不可算名詞を含むように「全体選定語彙」をランダムに選定し，140 語の問題となった．84 語の 2 倍にならないのは，重複が 28 語含まれたためである.

先と同じ実験参加者 7 名に解答してもらい,「全体選定語彙」と「提案選定語彙」で正答率と誤答率を調べた．その結果,「提案選定語彙」の誤答率が 37%で,「全体選定語彙」の誤答率が 35%となった．誤答率が高いということは問題の難易度が高いことを意味する．しかし，比率に関する有意差検定を行った結果，$p=0.54>0.05$ となり有意差は認められなかった.

6. まとめ

複合名詞問題を自動生成したところ，生成された問題は妥当な問題が大半であることがわかった．可算，不可算名詞の判定問題を自動生成した結果，やや難易度の高い問題が得られた．しかし，統計検定による有意差は得られなかった．さらに詳細な検討が必要である.

参考文献

[1] Electronic Dictionary Project 監修: 英辞郎第三版，アルク(2007).
[2] http://www.wordreference.com
[3] https://wordnet.princeton.edu
[4] https://en.wikipedia.org/wiki/Main_Page
2015 年 12 月 1 日版
[5] 竹内他:前方文脈を考慮した冠詞の推定, 言語処理学会第 19 回年次大会(2013).
[6] 小林他:名詞間の意味的共起情報を用いた複合名詞の解析，自然言語処理，Vol.3,No.1(1996).