大規模コーパスを用いた言語処理に基づく複合名詞辞書の構築

Development of Noun Compounds Dictionary by Corpus Processing

岩田 陽也[*1], 梅村 祥之[*2]
Yoya IWATA[*1], Yoshiyuki UMEMURA[*2]
[*1] 大学院工学系研究科
[*1]Graduate School of Science and Technology
[*2] 広島工業大学
[*2]Hiroshima Institute of Technology
Email: m161501@cc.it-hiroshima.ac.jp

あらまし：本稿では大規模コーパスを用いた英語複合名詞辞書の構築を行う. 日本人英語学習者にとって複合名詞の使い方は難しいのに, 複合名詞一覧などの学習教材がない. そこで, 本研究ではコーパスから複合名詞を自動獲得する. 手法には主に出現頻度を用いた. その結果, コーパスを用いての複合名詞の獲得が可能であることを確認し, 既存の辞書に記載されていない複合名詞を獲得出来た.
キーワード：言語処理，大規模コーパス，複合名詞， 辞書構築， 英語学習

1. はじめに

多くの日本人にとって英語で複合名詞を書くことは難しい. 例えば, 「学生用のアパート」と英語で書く場合, “an apartment for students”, または複合名詞を用いて, ”a student apartment”, と書くことも出来る. 複合名詞は物事を簡潔に表す事が出来るという特徴があり, 言語学的な研究[1]がなされている． しかし, 日本人の英語学習者が英語で複合名詞を書く場合は単純に単語を直訳してつなげれば良いわけではない. 英語学習者は自ら複合名詞を生産する能力を獲得する必要がある. しかし, 既存の辞書では複合名詞は数例しか掲載されておらず, 複合名詞を生産する能力の獲得が困難である.

そこで, 本研究では, 初級単語 2 語を様々に組み合わせることにより, 数多くの複合名詞を生成し, コーパスでの出現頻度を基に複合名詞を自動獲得する試みを行った[2-3]. その結果, コーパスからの自動獲得する方法が有効であることがわかった. それを受けて, 本稿ではコーパスから大規模な複合名詞辞書を自動獲得する.

2. 複合名詞の自動獲得の方法及び結果

コーパスには英文の大規模コーパスである英語版 Wikipedia[4]を利用する. Wikipedia は文章の質が個人向けブログ等よりも高く, 2015 年 12 月 1 日版で全体で約 30 億ワードを含む. 大規模なデータが提供されているので使用する. 事前にコーパス中に含まれている xml タグや記号等の不要部分を除去する. Wikipedia を英文形態素解析ツールである Stanford POS tagger[5]を用いて, 文中の品詞の同定を行い, 3 語以上の名詞からなる複合名詞を除外し, 2 語の一般名詞からなる複合名詞を大規模に獲得する. 対象を一般名詞 2 語に限定した理由は, 固有名詞を扱うと, “Bill Crinton”, のような人名や企業名が大量に獲得され, 学習用の辞書を換算する目的にそぐわないからである. また, 3 語以上の複合名詞を扱うと, 獲得数が莫大に増えるので, 今回は基礎検討として 2 語の複合名詞を扱うことにする. この時点で, 複合名詞の出現頻度の文単位での頻度の計測を行う. 4,231,591 語の複合名詞を獲得した. この時点では特殊記号を含む語が混入しているので正規表現を用いてアルファベット以外の文字を含む語を除外する. その結果, 3,798,716 語となった. このようにして獲得された複合名詞の妥当性を調査する.

3. 複合名詞の調査の方法及び結果

前章で機械的に獲得した複合名詞を対象に, 次の 5 つに分類して調査する.

分類Ⅰ）誤抽出を調べる
分類Ⅱ）Ⅰ)以外で, 掲載不要なものを調べる
分類Ⅲ）Ⅰ), Ⅱ)以外で,
　辞書にあるものを調べる
分類Ⅳ）Ⅰ)~Ⅲ)以外で,
　Google 翻訳で訳が分かるものを調べる
分類Ⅴ）Ⅰ)~Ⅳ)以外で,
　例文から分かるものを調べる
分類Ⅵ）Ⅰ)~Ⅴ)以外

分類は次のように行った. 誤抽出か否かは, 作業者がコーパス中の例文を見て, 文脈から判断する. 専門的すぎる, 日常ではなく特殊な場面でのみ使われる複合名詞は掲載不要と見なす. 掲載が必要と見なされた複合名詞を weblio 英和・和英辞典[6]で掲載されているか否かを調べる． 次に, 辞書に掲載されていないものの内, Google 翻訳で和訳を行い, 意味が分かるかを調べる. Google 翻訳で訳が分からなければコーパス中の文を調べ, 文脈から正しい意味が分かるかを調べる.

今回は基礎検討として, 調査対象を 3,798,716 語

からランダムにサンプリングした 1,000 語の複合名詞とする．作業者は1名で，英語力はTOEICスコアが 600 点以上である．

調査対象の 1,000 語に対し，上記のやり方に沿った構成率の調査を行った．その結果，掲載不要のものや誤抽出はわずかで，全体の約 8 割が既存の辞書に掲載されていないが，日常で使われているものだった． 以下に割合及び結果の例を示す．

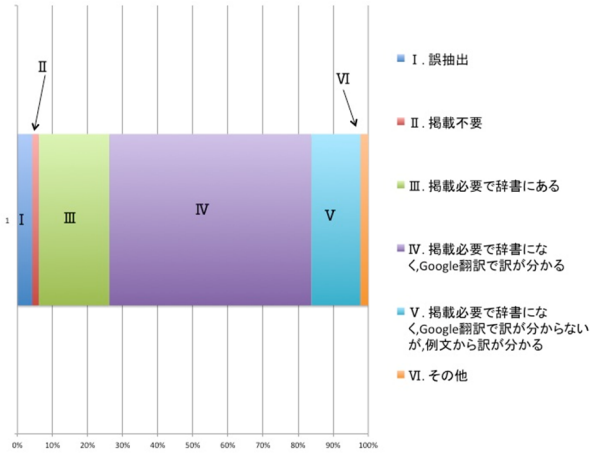

図 1 抽出した複合名詞の割合

・分類Ⅳ： Ⅰ~Ⅲ以外で， Google 翻訳で訳が分かる複合名詞

見出し語：treaty talks
Google 翻訳での和訳: 条約交渉
名詞 1 の頻度: 37,240
名詞 2 の頻度: 43,056
共起頻度: 11
例文:
Finally, in May 1846 Buffalo Hump became convinced that even he could not continue to defy the massed might of the United States, and the State of Texas, so he led the Comanche delegation to the treaty talks at Council Springs that signed a treaty with the United States.
最後に、1846 年 5 月、バッファロー・ハンプ氏は、米国とテキサス州の大衆を無視し続けることはできないと確信し、カウンシル・スプリングスの条約協議にコマンチ代表団を導いた。 アメリカ。

・ 分類Ⅴ： Ⅰ~Ⅳ以外で，例文から訳が分かる複合名詞

見出し語 5: road hat
Goolge 翻訳での和訳: 道路帽子
名詞 1 の頻度: 278,145
名詞 2 の頻度: 19,053
共起頻度: 1
例文:
The road hat is black with a matching brim and button, the O bolt logo on the front, and a gold New Era flag logo on the left side.
道路帽子は黒色であり、縁と縁が一致しており、「O ボルト」は、 正面にはロゴ、左側には金ニューエラフロゴがあります。

4. 考察

Wikipedia から獲得した複合名詞に関して大まかに印象を評価したところ，質が高いことが分かった．また，8 割程度正確な辞書を機械処理で作れることが明らかになった．Google 翻訳の性能向上が見込まれるため，この割合が今後上昇し，現在 10 数%存在する分類Ⅴが減少することが見込まれる．また，共起頻度が 1 のものでも事前の推測に反し，掲載不要とならない語が多数見られたのは着目に値する．

以上から，本手法を用いて，大規模な複合名詞辞書を機械的に構築することが可能で，かつ，有用な辞書が構築される見通しを得た．通常の中英和辞典が 15 万語程度を収録することから，今回獲得した複合名詞 380 万語という規模は格段に大きいとは言い切れない．しかし，今回の実験を通じて，機械処理のみで，8 割程度正確な辞書を構築できる手法が明らかになったため，今後，複合名詞辞書の規模を拡大するには，一般名詞が 2 語連続するという条件に替え，形容詞や現在分詞形なども取るように設定すると，獲得数の数 10 倍，数 100 倍の獲得が可能であると考える．

5. まとめ

日本人英語学習者にとって複合名詞の扱いは難しいにも関わらず，複合名詞を一覧できる教材や辞書がない．そこで，コーパスから複合名詞を自動獲得し， 辞書を自動作成した.その結果，コーパスから数 100 万規模の複合名詞を獲得することができた．自動獲得した複合名詞の妥当性をサンプリング調査によって調べたところ，8 割程度の妥当性を持つことが判明した．

機械処理だけで，これだけの品質を持つ複合名詞辞書を構築できる目処を立てることができたため，今後は電子辞書や Web サービスのような形態で複合名詞辞書を提供することを検討してゆきたい．

参考文献

(1) Réka Benczes: Metaphor-and metonymy-based compounds in English: a cognitive linguistic approach, Acta Linguistica Hungarica, Vol.52, No.2-3, pp.173-198 (2005)
(2) 岩田陽也，梅村祥之：”大規模コーパスを用いた言語処理に基づく英語複合名詞の学習支援”,JSiSE Research Report,vol.31,no.5, pp.53-58 (2017)
(3) 岩田陽也，梅村祥之：”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”,教育システム情報学会 2015 年度学生研究発表会 (2016)
(4) https://en.wikipedia.org/wiki/Main_Page
(5) http://nlp.stanford.edu/software/tagger.shtml
(6) http://ejje.weblio.jp/