検索エンジンを利用した英文の名詞語彙誤り検出の検討

Vocabulary error detection of English noun phrases by a search engine

玉城悠仁
Haruhito Tamaki

岡山大学　太田研究室
Ohta Laboratory, Okayama University

概要 日本人が英文を執筆する際に英文中の誤りを発見するのは一般に困難である．これまでの研究では，検索エンジンを用いて語と語の共起の強さを得ることにより，名詞や冠詞などの誤りを検出していた．名詞の語彙選択誤りを検出する宮城らの研究では，文の先頭から順に名詞を 3 単語抽出し，それらの共起の強さを求めていた．本研究では名詞だけでなく形容詞との共起の強さも考慮することで名詞の語彙選択誤りの検出性能を改善した．その結果，誤り検出の F 値は 0.469 であった．

1　はじめに

現在, 日本人の多くが, 第二言語として英語を学んでいる．しかし，英語を母語としない日本人にとって，正しい英語の使い分けは難しく，日本人の英作文には様々な誤りがしばしば見られる．その中でも，名詞句に関する誤りは特に多い．

宮城ら[1][2]は，検索エンジンによって得られる検索結果数を用いて，英文中の名詞語彙誤りを検出する手法を提案した．宮城らは，検索結果数を用いて，名詞とその近くに存在する名詞との共起の強さに基づいて名詞句の語彙誤りを検出した．

本稿では，名詞句の語彙誤りの検出のために，宮城らの提案した手法を拡張して，より高い精度の語彙誤り検出方法を検討する．

2　宮城らの英文名詞語彙誤り検出手法

2.1 名詞の共起の強さ

宮城ら[1][2]は，英文中の名詞句の語彙誤りを検出するために, 名詞の共起の強さを利用した．宮城らの手法では，名詞の共起の強さを，検出対象の名詞とその他の名詞が一文中にともに出現する頻度を用いて定義する．例えば，“ My hobby is baseball. ”という英文では，“hobby” と “baseball” が一文中にともに出現する頻度を用いる．そして名詞の共起の強さは，検索エンジンの検索結果を用いて式(1)で算出する．

$$\text{共起の強さ} = \frac{\text{語 A と語 B を同一文中に含む検索結果の数}}{\text{取得した検索結果の数}} \quad (1)$$

ここで，取得した検索結果の数とは，語 A と語 B の検索を実行して, Bing Search API から取得した検索結果の数を示す．この検索結果は，最大 1050 件で，URL やスニペットなどの情報を持つ．そのスニペットを解析し，語 A と語 B が一文中にともに出現しているかどうか判定し，その頻度を求める．この共起の強さが閾値未満である場合，語彙誤りの可能性があると判断する．名詞と名詞の共起の強さで用いる閾値は KJ コーパスのうち実験データとして使用しない英文 34 文を用いて，0.098085 と定めた．

2.2 共起の強さに基づく名詞の語彙誤り検出

宮城らが提案した共起の強さに基づく名詞の語彙誤りの検出では，英文の先頭から順番に 3 つの名詞を選択して，その中の全ての組み合わせ（3 通り）の共起の強さを比較し，誤りを検出する．

例えば“In A, B have small C and D.”という文を考える．ここで A, B, C, D は名詞とする．宮城らのシステムではまず(A, B, C) の 3 語の組み合わせについて(A, B)，(B, C)，(A , C)の共起の強さを求める．次に(B, C, D) の 3 語の組み合わせについても同様にして共起の強さを求め，それらの求めた共起の強さに基づいて誤りを検出する．この際，3 つの名詞は，最初は 1, 2, 3 番目，次に 2, 3, 4 番目のよう先頭から順番に選出する．

3 つの名詞の共起関係を扱うため，名詞が 2 語以下しかない文については，誤り検出を行わない．

共起の強さを求めるために必要な検索結果を得るために，検索クエリを生成する．対象の英文を MontyTagger でタグ付けし，文中のすべての名詞を抽出する．3 つの名詞(A, B, C)で考えられる三つのペア(A, B)，(B, C)，(A , C)の検索クエリを生成する．生成したクエリを用いて得られた検索結果から各ペアの共起の強さを求める．そして共起の強さに基づいて，以下のように語彙誤りを判定する．ここで，「＝」は共起の強さが閾値以上であること，「－」は共起の強さが閾値未満であることを示す.

語彙誤りを検出するパターン

- A＝B, B－C, A－C のように一つの名詞のみが他の二つの名詞との共起の強さが小さい場合（この例では C がそれに該当するので誤りとして検出）

語彙誤りを検出しないパターン

- A＝B, B＝C, A＝C のように全て共起の強さが大きい場合
- A－B, B－C, A－C のように全て共起の強さが小さい場合
- A－B, B＝C, A＝C のように一つの名詞のみ他の二つの名詞との共起の強さが大きい場合

このように, 3 つの名詞間での共起の強さに基づいて誤りを検出する.

3　提案手法

宮城らの手法[1][2]では，名詞と名詞の共起の強さを調べることで語彙誤りを検出していた．しかし，語彙誤りの検出には名詞だけではなく他の品詞との共起の強さも有用な情報といえる．そこで本稿では，名詞と形容詞の共起の強さを扱うことで，宮城らの語彙誤り検出法の性能改善を試みる．

具体的には宮城らの手法を用い，英文の先頭から順番に名詞を選択する際に，形容詞も選択する．“In A, B have small C and D.”という例文では，最初に(A, B)，(B, small)，(A, small)の共起の強さを求め，その共起の強さに基づいて誤りを検出する．形容詞が含まれない部分については宮城らの手法と変わりはない．

また，名詞を代名詞とその他の名詞に分け，共起の強さの閾値を品詞の組み合わせごとに設定する．MontyTagger によって各単語がタグ付けされるが，名詞に付与されるタグは NN, NNS，NNP，NNPS，PRP のいずれかである．PRP は“I”, “they”, “it”などといった代名詞である．学習データで共起の強さを調べたところ，これらは他の名詞に比べ共起の強さの値が小さくなる傾向があったため，閾値も変える必要があると考えた．宮城らは名詞のみを扱い代名詞を区別しなかったが，本研究では名詞，代名詞，形容詞の組み合わせごとに学習データを用いて閾値を設定した．学習データには KJ コーパスのうちテストデータとして使用しない英文 46 文を用いた．設定した閾値を表 1 に示す．

表 1　提案手法の閾値

品詞の組み合わせ	閾値
名詞と代名詞	0.005346
名詞と名詞	0.101300
代名詞と形容詞	0.009931
名詞と形容詞	0.113912

4　評価実験

3 節で説明した提案手法について評価実験を行った．テストデータには，KJ コーパスの英文 30 文を用いる．この 30 文は全て名詞と形容詞を合わせて 3 語以上含んでおり，かつ，名詞の語彙誤りも含む．

KJ コーパスの名詞語彙誤りには欠落誤りや，スペルミスや品詞選択誤りにより，誤り箇所が名詞以外の品詞とみなされてしまうといったものが含まれている．例えば “So I do not like Susi from other.” は“places”が欠落しており，正しくは“So I do not like Susi from other places.” である．また “I and my friends went to the see and the pool, drove anyway, ate BBQ.” では“see”が正しくは“sea”である．しかしこの“see” では名詞ではなく動詞となってしまう．

実験に用いる 30 文にはこのような誤りを含まないものを選んだ．この 30 文は，30 件の名詞語彙誤りを含んでいる．宮城らの手法と提案手法による名詞語彙誤りの検出性能を評価する．

宮城らの手法と提案手法それぞれで 30 文の語彙誤りを検出した結果を表 2 と表 3 に示す．表 3 において誤検出とは，検出した誤りが誤りではないことを示す．

表 2　語彙誤り検出数

	検出	誤検出	非検出
宮城らの手法	7	7	20
提案手法	15	19	7

表 3　語彙誤り検出性能

	適合率	再現率	F値
宮城らの手法	0.500	0.233	0.318
提案手法	0.441	0.500	0.469

5　考察

表 3 より，形容詞と名詞の共起の強さを導入し，閾値の設定を細分化することで，再現率，F 値が向上したことがわかる．宮城らの手法では検出されなかった名詞の語彙誤りが検出できるようになり再現率が上がった．逆に誤検出も増えたため適合率は下がったが，F 値は向上した．

宮城らの手法では検出できなかったが，提案手法により名詞の語彙誤りを検出できるようになった文に“Because tomatoes in the shop have good taste and body.” という文がある．この文では“body” が誤りであり正しくは“appearance” である．提案手法がこの文で body の誤り検出の際に生成するペアは，(good, taste), (taste, body), (good, body) である．これらの中では(good, taste)の共起の強さの値は閾値より高く，(taste, body), (good, body)ではいずれも閾値より低かったため“body” が誤りと検出された．宮城らの手法では生成するペアが(shop, taste), (taste, body), (shop, body)であり，いずれも閾値未満であったため検出しなかった．よって，提案手法で形容詞との共起の強さを用いたため，検出が可能になった例といえる．他には“But the ear is very bad near Kawanishi station.” という文では“ear” が誤りであり正しくは“air” である．この例でも検出対象“ear”の近くにある“bad” が共起の強さに影響を与えており，形容詞との共起のスコアが名詞の語彙選択誤り検出に有効に働いていたことがわかった．

また，“I often think of stories in my brain.” という文では“brain” が誤りであり正しくは“head” である．この誤りは宮城らの手法と提案手法の両方で検出できていたが，宮城らの手法では “I” を誤りと誤検出していた．これは“I”が代名詞であるために他の名詞と比べ共起の強さの値が低く，誤りだと判定された．提案手法では代名詞のときの閾値を設定していたため誤検出しておらず，他の名詞と区別して代名詞の共起の強さを算出することが誤り検出の性能改善に有効であるとわかった．

提案手法では隣接する 3 語間の共起を計算しているため，離れた語の共起は考慮できていない．共起する名詞や形容詞は必ずしもこの 3 語以内近くに出現しないため，共起の強さを調べる際には，それほど近くにない名詞や形容詞との共起の強さを検討する必要がある．また，本実験ではテストデータから誤りが名詞としてタグ付けされていない文を除いており，実際そのような誤りは検出できない．また，提案手法は MontyTagger のタグ付けが正しい前提で判定しているため，正しくタグ付けできない文に対しては正しく検出されないおそれがある．検出精度を上げるとともにタガーの誤りも考慮した方法を導入することが今後の課題である．

6　まとめ

本稿では，検索エンジンの検索結果により英文中の語と語の共起の強さを算出し，英文名詞句の語彙誤りを検出する方法について述べた．本稿の提案は，名詞を代名詞とその他の，名詞に分け，さらに形容詞との共起の強さも考慮し，品詞ごとに共起の強さを比較して算出するようにしたことである．これにより F 値が宮城らの手法では 0.318 だったのに対し，提案手法では 0.469 に向上した．本稿では隣接する 3 単語間の共起の強さのみを用いたが，今後は離れた語との共起も考慮し，またタガーの誤りも考慮した方法をも導入することで名詞句の語彙誤り検出精度の向上を目指す．

参考文献

[1] 宮城雄太，新妻弘崇，太田学,“検索エンジンを用いた英文名詞句誤りの修正支援”, DEIM2015, D7－2, 2015.

[2] 宮城雄太,“検索エンジンを用いた英文名詞句誤りの修正支援に関する研究”，岡山大学工学部情報系学科特別研究報告書, 2015.