英語動詞の適切な使い分けを支援するシステムの開発

豊辻 宏旨 † 松崎 拓也 ‡ 佐藤 理史 ‡
† 名古屋大学 工学部 電気電子・情報工学科 ‡ 名古屋大学大学院 工学研究科

1 はじめに

英語が母語でない者にとって、書こうとしている英語表現が正しいのか判断するのは難しいことがある。その際に、用例を見ることは非常に有用である。しかし、既存の用例検索システムでは、意図する意味に対して不適切な表現を入力しても、適切な表現を得ることはできないという問題がある。そこで、本研究では、動詞と目的語を入力として、目的語と適切なコロケーションとなる類似する動詞を提示するシステムを提案する。

用例検索システムの例として、松原らによる ESCORT [1]、岡崎らによる PoEc [2] などがある。その他にも様々な用例検索システムがある。[1] [2] しかし、現状のシステムでは、単語での AND 検索や OR 検索、もしくはワイルドカードと組み合わせたクエリでの検索しかできない。そのために、誤った表現から適切な表現を得る、あるいは意味的に類似する、よく使われる表現を得るといったことはできない。例として、「ラッシュを避ける」という意味を表す英語の用例を調べたい場合に、誤った表現である “escape the rush” を入力したとする。このとき、たとえば Weblio での検索結果は 0 件であり、この表現が誤りらしいと分かる。しかし、適切な表現である “avoid the rush” を得ることはできない。また、「習慣をやめる」という意味を表現したい場合に、“stop the habit” を入力すると、それが含まれる例文を検索することができ、適切な表現であることが分かるが、“break the habit” などの類似する表現を得る事はできない。

英語を書く際に、動詞の選択は特に難しいと思われる。意味的には適切であっても、使い方や目的語との関係によって適切な表現にならないことがあるためである。本研究では、動詞と目的語の関係に着目し、目的語に対して適切な動詞を選択する。動詞の修正や候補の提示を行う研究としては、Liu らの英語学習者作文の動詞選択誤りの自動訂正 [3] がある。周辺の単語は考慮しているが、対象動詞が 50 個に限定されているということが本研究と異なる。

本研究では、まず英語学習者が動詞と目的語との関係において、誤った動詞を含んでいる例を調査した (§2)。次に、動詞と目的語を入力とし、入力の動詞に対し、類似する動詞候補の中から、目的語との組み合わせが適切となるものを提示するシステムを開発した (§3)。そして、学習者の誤りに対して、システムが適切な動詞をどの程度提示することがきるか評価を行った (§4、§5)。

2 学習者コーパスでの誤りの調査

実際に、非英語話者の動詞選択の誤りにどのようなものがあるのかを学習者コーパスを用いて調査した。

2.1 調査対象

コーパスとして、Cambridge Learner Corpus (CLC) [4] Konan-JIEM (KJ) [3] 、NICT-JLE (JLE) [5] を用いた。CLC と KJ は英語学習者による自由英作文をまとめたものである。NICT-JLE は英語学習者に対してインタビューテストを行い、発言を書き起こしたものである。それぞれ、誤りの部分にはタグ付けがされており、誤りの種類と修正の結果が書かれている。まず、これらのコーパスから、動詞の選択が誤っているものを抜き出した。CLC では、タグが RV (動詞の置換が必要) となっているもの、KJ、JLE では、タグが v⊥xc (動詞-語彙選択誤り) となっているものを抜き出した。表 1 に、誤り全体に対する動詞選択誤りの割合と、全ての誤りタイプの内、動詞選択誤りが何番目に多い誤りなのかを示す。これらの動詞選択誤りの中から、修正結果が普通名詞、固有名詞、代名詞が目的語になっているものを人手で集めた。なお、修正結果が自動詞になっているものは除いた。

[1] Weblio 英語用例検索: http://ejje.weblio.jp/sentence/
[2] 英辞郎 on the WEB: http://www.alc.co.jp/
[3] GSK2015-A Konan-JIEM Learner Corpus Fourth Edition: http://www.gsk.or.jp/catalog/gsk2015-a/

表 1: 学習者コーパスにおける動詞選択誤りの割合

コーパス	誤り総数	誤り割合 (%)	誤り数の順位	抽出数
CLC	3580	6.4	2/75	322
KJ	291	7.0	6/20	47
JLE	925	6.5	4/47	10

表 2: テストセットの誤りの分類

誤りの分類	誤りの例	修正例	総数
選好的	stop habit	break habit	74
目的語との相性	take progress	make progress	147
語法的	occur trouble	cause trouble	17
文脈的	make picture	take picture	124
その他	change visit	reschedule visit	17

2.2 動詞・目的語の関係における誤り分類

2.1 節で述べた、英語学習者コーパスから抽出した修正例 (計 379 個) を選好的誤り、目的語との相性誤り、語法的誤り、文脈的誤り、その他の 5 つに分類した。表 2 に分類例と各カテゴリに分類された修正例の数を示す。

選好的誤りは、元の動詞と目的語の組み合わせも誤りとは言えないが、よりふさわしい組み合わせとなる動詞があるものである。例えば、JLE において、「習慣をやめる」という意味で “stop habit” が break habit に修正されているが、“stop habit” が使われることもある。しかし、break のほうが一般的である。目的語との相性誤りに分類されるのは、CLC でみられた誤りである、“take progress” など、動詞と目的語の組み合わせとして使われることのないものである。この例は「進歩する」という意味を表現しようとしたものである。しかし、その意味では、“make progress” が正しい。また語法的誤りの例としては、KJ でみられた “occur trouble” などがある。「問題を引き起こす」という意味を表わそうとしたものであるが、occur は自動詞であり、他動詞 cause を使わなければならない。別の誤り例としては、CLC での “put effort” がある。「努力する」という意味であるが、“put effort into ...”(「... に努力する」) という形で使わなければならない。その対象を明示せず、単に努力するという意味では”make effort” が適切である。文脈的誤りは、CLC において “make picure”(「絵を描く」) が “take picture”(「写真を撮る」) と修正されているように、修正前・後のいずれも英語フレーズとしては意味をなすが、修正前の動詞・目的語ペアが文脈で意図した意味を表していないものである。

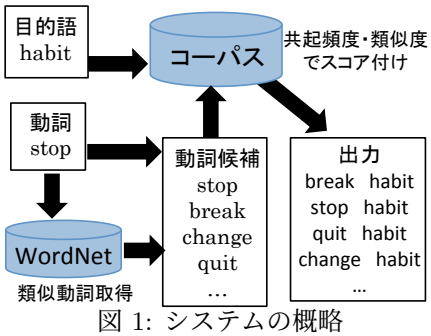

図 1: システムの概略

3 目的語に対してより適切な動詞を提示するシステム

3.1 提案手法の概要

2 節で述べた分類のうち、選好的誤りと目的語との相性誤りに分類されるものを想定される入力として、適切な動詞の候補の提示システムを提案する。システムの入力は、動詞と目的語である。目的語は名詞のみであり、動詞と目的語は両方とも標準形とする。まず、WordNet から入力された動詞と意味的に類似する動詞を取得し、提示する動詞の候補とする。次にコーパスでの共起頻度や word2vec [6] を用いて算出した動詞の類似度をもとに、候補となる動詞から、入力された目的語との組み合わせが適切と考えられるものを提示する。

3.2 コーパスの整形

英語話者の英語における動詞と目的語のペアの頻度を得るために、British National Corpus（BNC）[4] を構文解析し、動詞と目的語の関係になっているものを取り出した。構文解析器として、Enju [7] を用いた。取り出した動詞と目的語 (名詞) はそれぞれ標準形に直し、その共起頻度とコーパス中における出現位置をまとめた。この動詞と目的語の組み合わせは延べ 5,648,516 個あり、異なり数は 1,453,222 個であった。

3.3 動詞候補の取得

入力の動詞に対して、意味的に類似した動詞候補を得るために WordNet を利用した。入力された動詞に加えて、入力の動詞の同義語 (synset)、上位語 (hypernym)、下位語 (hyponym)、同族語 (coordinate terms) を動詞候補とした。

[4]The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

3.4 候補動詞に対するスコア付け

候補動詞から、目的語と組み合わせが適切な動詞を選択するためにいくつかの異なるスコアを定義し、比較した。以下、それぞれについて述べる。

共起頻度 (CO) 各候補に対して、入力された目的語がコーパス中において、動詞・目的語関係になっている頻度をスコアとした:

相互情報量 (MI) 各候補と入力された目的語の相互情報量をスコアとした:

$$I(v,n) = \log \frac{p(v,n)}{p(v)p(n)}$$

$p(v,n)$ は、動詞と目的語が、動詞・目的語の関係としてコーパスに出現する確率である。$p(v)$、$p(n)$ はそれぞれ、動詞と名詞のコーパスで出現する確率である。

相互情報量と共起頻度の積 (MI plus) 相互情報量に共起頻度をかけた値をスコアとした:

$$c(v,n)I(v,n) = c(v,n) \log \frac{p(v,n)}{p(v)p(n)}$$

$c(v,n)$ は動詞と目的語が、動詞・目的語の関係としてコーパスに出現する頻度である。

χ 二乗値 (χ^2) χ 二乗値をスコアとした:

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})}$$

N は全ペア数であり、O_{11}、O_{12}、O_{21}、O_{22} はそれぞれ、候補のペアが動詞と目的語になる確率、目的語が候補動詞以外の動詞と目的語の関係となる確率、候補動詞が入力目的語と異なる目的語をとる確率、候補動詞と入力目的語が含まれないペアとなる確率である。

3.5 フィルタリング

スコアでランク付けした候補動詞から不適切なものを除くことを試みた。具体的には、word2vec を用いて、入力された動詞と候補動詞の類似度を測り、値が小さいものを候補から除外するようにした。word2vec の学習には、BNC を用いて、cbow モデル (200 次元) で学習を行った。類似度は、コサイン類似度を使用し、値が負となるものを候補から除外した。

4 実験

3.4 で示した 4 つのスコアに基づくシステムの出力とフィルタリングの効果を評価した。テストセットとして、2 節で述べた、英語学習者コーパスから抽出した修正例を用いた。適切な動詞が、コーパスでの修正例だけとは限らないが、評価のためにそれを「正解」とみなす。

4.1 スコア付け方法の比較

テストセットの中で、選好的誤りと目的語との相性誤りに分類される 220 個について評価を行った。目的語と誤りの動詞を入力し、正解の動詞が提示されている順位とその平均逆順位を求め、評価した。その結果を表 3 に示す。表中の R@n は正解の動詞が n 位以内に現れる割合 (百分率) である。

結果より、選好的誤りと目的語の相性の誤りに対して、最も結果がよかったのは、CO(フィルタリングあり) であった。使用者が、5 位までの候補を見れば 51.4%ほど、コーパスでの修正例と同じものを得られることがわかった。また、フィルタリングを行うと、上位に入る割合を改善することができ、5 位までを考慮すれば、半数の誤りに対して正解を得ることができた。

4.2 誤りタイプごとの結果の比較

それぞれのスコア付け方法において、誤りの分類ごとにどのように結果が異なるのかを平均逆順位により評価を行った。表 4 にその結果を示す。

結果より、選好的誤りと目的語との相性誤りに対しては、CO が最も性能がよいことがわかる。また、どのスコア付け方法においても、目的語との相性誤りよりも語法的誤りに対する性能のほうがよいという結果が得られた。選好的誤りと目的語との相性誤りを想定する入力としていたが、語法的誤りもある程度まで修正できていることがわかる。

5 実験結果の検討

今回行った実験では、動詞・目的語の誤りのみを扱い、目的語は名詞となるものしか扱っていない。また、提示候補の 5 位以内に正解が 5 割程度しか現れていないので、実用には及ばない。

CO は動詞・目的語となる頻度のみしか考慮していないため、目的語に対して適切な動詞の出現頻度が低い場合は有用ではないと考えられる。一方、MI では、動詞の出現が低頻度であるが目的語との共起頻度が多いものを得られると思われたが、結果としては良くない。これは、動詞・目的語となる頻度に対して、動詞の出現頻度の方がスコアに大きく影響を与えているためだと考えられる。MI では、動詞の頻度が極端に少なく、共起頻度も少ないものが上位に入ってしまった。テストセットに含まれる誤りの例が比較的に出現頻度の高い動詞になっていることも MI の結果が悪かった原因の一つだと考えられる。

表 3: スコア付け方法の比較

スコア	フィルタリング	R@5	R@10	R@20	R@100	平均逆順位
CO	なし	51.4	60.0	68.2	71.8	0.379
CO	あり	51.8	60.5	68.2	28.2	0.382
MI	なし	27.3	40.9	50.5	70.5	0.175
MI	あり	29.5	41.8	51.8	70.5	0.187
MI plus	なし	48.6	57.7	63.6	70.5	0.353
MI plus	あり	50.0	57.7	64.1	70.5	0.355
χ^2	なし	41.8	50.9	60.0	70.9	0.319
χ^2	あり	42.3	51.8	60.9	70.9	0.325

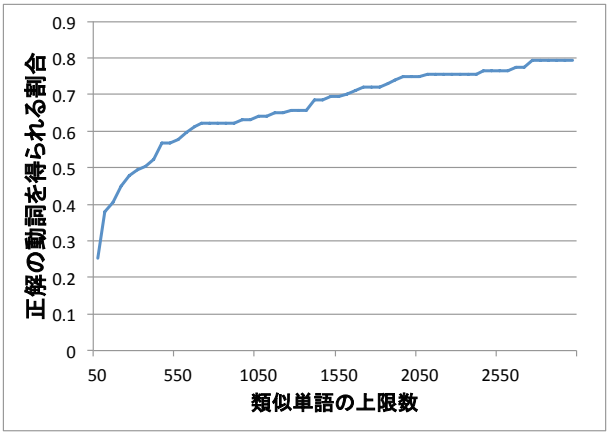

図 2: word2vec による類似単語数の上限と正解動詞の得られる割合

正解となる動詞が得られない原因のひとつは候補動詞中に正解がないことである。テストデータの内 29.3%のペアが、候補動詞の中に正解となる動詞が入っていなかった。これらの中には、look-watch のように意味的に類似するものがある。これらに対処するためには、英和辞書と和英辞書を用いて、入力された単語を和訳し、再び英訳することで、その候補を増やすことができると考えられる。また、word2vec を用いて候補動詞を取得することも考えられる。表 5 は WordNet で正解の動詞を得られない場合において、誤りの動詞に対して、word2vec を用いて、何個の類似単語まで見れば正解の動詞を取得できるのかを示している。類似単語を 500 個まで追加すれば、WordNet で得られない正解のうち半分近く得ることができることがわかる。

誤りの分類ごとの結果の比較では、選好的誤りに対する結果が良かった。選好的誤りにおいては、誤りの動詞と入力動詞が類似していることが多いためだと考えられる。表 5 に誤りの動詞を入力として、正解の動詞を WordNet で得られない割合を誤りタイプごとに示す。選好的誤りでは、他の誤りよりも正解の候補を得られていることが多い。

表 4: 誤りタイプごとの比較

スコア	フィルタリング	選好的	目的語との相性	語法的	文脈的	その他
CO	なし	**0.413**	0.359	0.398	0.318	0.266
CO	あり	**0.413**	**0.364**	0.400	**0.319**	0.273
MI	なし	0.183	0.170	0.146	0.081	0.182
MI	あり	0.189	0.186	0.187	0.091	0.199
MI plus	なし	0.380	0.337	0.423	0.254	0.278
MI plus	あり	0.382	0.340	**0.425**	0.255	0.286
χ^2	なし	0.349	0.302	0.357	0.184	0.302
χ^2	あり	0.356	0.308	0.361	0.192	**0.304**

表 5: WordNet で正解の動詞を得られない割合

誤りの分類	正解を得られない割合
選好的	19%
目的語との相性	28%
語法的	29%
文脈的	34%
その他	47%

6 おわりに

本研究では、学習者コーパスを調査し、誤りから目的語として名詞を取っているものを抽出し、分類した。誤りの分類のうち、選好的誤りと目的語の関係の誤りとなっているものを対象に、適切な動詞を提示するシステムを開発した。学習者コーパスから得られた誤り例によってシステムを評価したところ、動詞と目的語の共起頻度のみでスコア付けを行う方法が最も良い性能を示した。しかし、実用には十分ではない。候補集合を増やすと共に、スコア付けの精度を向上させていく必要がある。また、自動詞の誤りにも対応していくことも検討したい。

参考文献

[1] 加藤芳秀, 江川誠二, 松原茂樹, 稲垣康善. 依存構造に基づく用例文検索手法とその評価. 電子情報通信学会論文誌, Vol. 92, No. 3, pp. 417–427, 2009.

[2] 高松優, 岡崎涼太, 乾健太郎. 英作文支援のための用例検索システムの構築. pp. 361–364. 言語処理学会 第 18 回年次大会発表論文集, 2012.

[3] Xiaohua Liu, Bo Han, Kuan Li, Stephan Hyeonjun Stiller, and Ming Zhou. Srl-based verb selection for esl. In *Proc. EMNLP*, pp. 1068–1076, 2010.

[4] Diane Nicholls. The cambridge learner corpus: Error coding and analysis for lexicography and elt. *Proc. Corpus Linguistics 2003 conference*, pp. 572–581, 2003.

[5] Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. The nict jle corpus: Exploiting the language learner ’s speech database for research and education. *International Journal of the Computer, the Internet and Management*, Vol. 12, No. 2, pp. 119–125, 2004.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, Vol. abs/1301.3781, , 2013.

[7] Yusuke Miyao and Jun’ichi Tsujii. Feature forest models for probabilistic hpsg parsing. *Comput. Linguist.*, Vol. 34, No. 1, pp. 35–80, 2008.