検索エンジンを利用した英作文支援システムの構築

佐藤　学† 安藤　進†† 山名　早人†††

† 早稲田大学理工学研究科

†† 翻訳者

††† 早稲田大学理工学術院、国立情報学研究所

E-mail: †manavy@yama.info.waseda.ac.jp, ††sando@inter.net, †††yamana@waseda.jp

1. はじめに

近年、英語で文書を書く機会が増加しており、効率的に自然な英文を作成するためのシステムが求められている。英文を作成するためのシステムには、機械翻訳、翻訳支援と２つのアプローチがあるが、近年では特に翻訳支援でのアプローチが盛んになっている。

英作文支援には、大きく分けて英語の単言語コーパスを用いる方法と、文の対応付けがされた日英対訳コーパスを用いる方法があり、双方に異なる利点がある。単言語コーパスを用いる方法は、コーパスデータを大量に入手できるため、表現の一般性を調べるのに適している。一方、文の対応付けがされた対訳コーパスを用いる方法は、日本文と英文を比較しながら用法を参照できるため、ユーザが効率的に意味を把握できるという利点を持つ。単言語コーパスを用いた研究例としては、TransAid [1] ,WebLEAP [2],Kiwi [3] などがある。しかしこれらの手法は、何れもコーパス内での出現回数を元に機械的に選択した例文を表示したり、入力文に対しコーパス内での出現文書数が低い箇所を示す、検索結果から用例を抽出するといった機能に留っている。このため、ユーザーが自身で表現の一般性を比較し判断を行うための、判断基準を得ることができないという問題を持つ。

本論文では、こうした問題を解決するために、検索エンジン特有の機能を利用し表現の一般性を調べる機能や、対訳コーパスから効率的に訳語の用法を発見できる機能を実現するシステムを提案する。

2. 検索テクニックを利用した表現の一般性の確認

ユーザが自身で表現の一般性の判断を行うための判断基準を得る手法として、検索エンジン特有の機能を利用してフレーズ候補を抽出し、フレーズ検索での検索結果数の比較を行い、語彙の組み合わせを検討する方法がある。[4]

例えば、「汗で濡れる」という表現を英訳する場合、”wet * sweat”と in の部分をワイルドカードに置き換えて Google で検索すると、検索結果要約の中で”wet with sweat”,”wet from sweat”を発見できる。これらを英訳の候補とし、それぞれフレーズ検索を行と、結果は表 1 のようになる。

表 1　前置詞の検討

検索文字列	検索結果件数
”wet with sweat”	3,280
”wet from sweat”	854

表 1 から、「with」を使った場合の方がヒット件数が多いことがわかり、「with」がよく使われているということがわかる。前置詞は状況によって使い分けられるため、必ずしも多いものが正解というわけではないが、極端に少ないものは用法が間違っている可能性が高いという判断材料を得ることができる。しかし、こうしたフレーズを検索結果要約から探し、一つずつ検索していたのでは非常に効率が悪いという問題がある。

こうした問題に対し、以下では、検索エンジン特有の機能を利用し表現の一般性を調べる方法の自動化を行うシステムの提案、評価を行う。

3. ワイルドカード検索を用いた表現の一般性の確認

２節で紹介した検索テクニックを自動化する機能について述べ、本機能の処理の流れを図 1 に示す。

まず、ユーザはテキストボックスにワイルドカードを含む英語フレーズを入力する。たとえば、「汗で濡れる」というフレーズの英訳を考える際、適切な前置詞について検討したい場合、「wet * sweat」というフレーズを入力する。抽出するフレーズのワイルドカード部分を前置詞に限定する場合は、

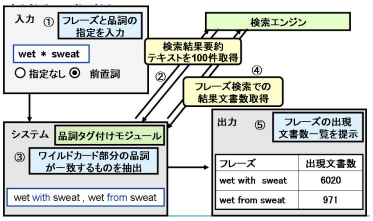

図 1　ワイルドカード検索を利用した語彙組み合わせの検討

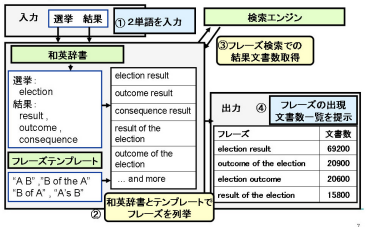

図 2　和英辞書とフレーズテンプレートを用いた複合語の検討

前置詞をチェックする。

フレーズの検索には GoogleAPI [5] を使用する。GoogleAPI による検索では一度に得られる検索結果数が 10 件に制限されているため、提案するシステムではフレーズを受け取ると、検索結果の 1-10 件目、11-20 件目といった具合に、GoogleAPI を用いて並列に Google へクエリを送信し、合計 100 件の検索結果の要約テキスト（スニペット）から、太字になっている該当フレーズを抽出する。

抽出したフレーズの一覧には、ワイルドカード部分の語が意図しない品詞のものも含まれるため、品詞タグ付けモジュールにより、品詞を取得し、指定したものと一致するフレーズのみ抽出する。品詞タグ付けモジュールには Brill's tagger [6] を用いた。

次に、抽出した全てのフレーズを新たなクエリとして検索を行い、検索結果文書数を取得する。取得したフレーズと検索結果文書数の一覧をユーザに提示することにより、ユーザはフレーズ選択の基準として検索結果数一覧を利用できる。ここで、自動的に出現文書数の上位のフレーズのみを選択せず、一覧表示を行うのは、実際には必ずしも出現文書数が多いものが最適であるとは限らないため、ユーザの判断に任せるためである。

4. 和英辞書とフレーズテンプレートを用いた複合語の検討

本節では、和英辞書とフレーズ検索を用いて、複合語の語彙組み合わせの検討を行う機能について説明する。

本機能の処理の流れを図 2 に示す。

日本語の名詞 2 語もしくは、日本語を含む英文フレーズを入力とし、和英辞書とフレーズテンプレートを用いてフレーズの候補を生成し、検索エンジン内の出現文書数一覧を提示する。

システムへの入力は日本語の 2 語もしくは、日本語を含む英文フレーズとする。

例えば、「医療施設」という複合語を英語で表現したい場合、「医療　施設」という 2 語を入力する。ただし、「医療」には「medical」だけ該当させたいという場合は、「medical 施設」と入力し、医療の訳語を固定させることもできるようにした。

システムにフレーズが入力されると、システムは日本語で入力された語を判別し、和英辞書から検索、候補の一覧を作成する。和英辞書には、EDR 日英対訳辞書 [7] を用いた。

そして、「A B」「A's B」「B of A」「B of the A」といったフレーズ構造のテンプレートを定義しておき、それぞれ和英辞書で一致した単語を挿入してフレーズの一覧を作成する。フレーズの候補一覧が生成されると、候補全てに対し、検索による文書数を調べ、結果数が多いものから順に出力する。ユーザは結果文書数を比較し、最適と思うフレーズを選択し、英文作成に用いることができる。

5. 可読性を考慮した検索結果からの例文抽出

次に、前節で述べた 2 手法によりフレーズの候補を参照する際、抽出されたフレーズの利用例（例文）が確認できるようにする。しかし、利用例を表示するにあたり、検索結果をそのまま表示したのでは読みづらい。また、単純に 1 文を抜き出しても、記号や略語、難しい語を多く含む文など、読みづらいものを表示してしまい、本来の目的である、フレーズの利用方法・前後関係などについての確認を行うのに効率が悪い。そこで、提案手法では、検索結果要約テキストから、1 文を抽出し、「読みやすい」順にランキングし表示する。

本機能の処理の流れを図 3 に示す。

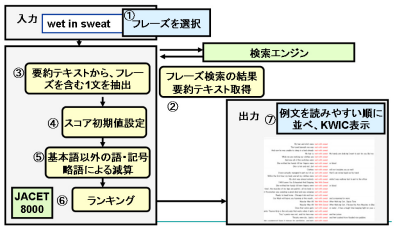

図 3　可読性を考慮した検索結果からの例文抽出

5.1 可読性の定義

読みやすい例文を上位表示するアルゴリズムを考えるにあたり、まず「読みにくい文」の指標となる項目を以下のように定義する。

- 語数が多すぎる文
- 記号や略語を多く用いている文

- 難しい語を多く用いている文
- 固有名詞や専門用語を多く用いている文

スコアの初期値を 100 と設定し、抽出した文それぞれについて、上記各項目に該当する場合、それぞれの重みに応じたスコアを減点するという方法によって読みやすい文を上位表示するようにした。1 文内の語数については、5 語以上、20 語以内で構成されている文について抽出を行った。次に、抽出した文について、記号や略語によるスコアの減点を行う。なお、記号や略語は、可読性に与える影響が大きいため、各記号・略語に対し 2 点の減点とした。記号や略語の定義は、英字、「,」 ，「’」以外の文字全てを対象とし、全て大文字から構成されている語を略語とて扱った。

さらに、難しい語、固有名詞や専門用語を多く用いている文については、JACET8000 [8] という英単語リストを元に、スコアを計算した。その詳細について次節で述べる。

5.2 JACET8000 を用いたスコア計算

語の難易度を定義するために、大学英語教育学会の定める基本英単語リスト JACET8000 を用いる。JACET8000 は 1000 語ずつ 8 段階に分かれており、level1 が中学校レベル、level2,3 が高校レベル、lavel4,5 が大学受験・大学一般レベル、level6,7,8 が大学一般教養・英語専攻レベルとなっている。

JACET8000 に出現しない単語は、固有名詞や難しい専門用語などに該当する可能性が高いため、JACET8000 の高 level の語や JACET8000 に出現しない語を減算の対象とした。

ここでは、基本〜大学一般レベルである level1 〜 5 までに含まれている語についてはスコア操作をせず、それ以外の語について単語が含まれる度に 1 点減点した。

以上の手法で計算したスコアが高い順に例文をランキングし、入力フレーズが中央に整列する KWIC 形式で表示を行う。

6. 検索語の訳語対応を自動取得する対訳文検索

5 節でフレーズの用法を効率的に調べるために、検索結果からの例文抽出の機能を構築した。しかし、単語の使用方法を調べる上では、対訳コーパスを利用した方が、使用方法と意味を同時に理解できるので効率がよい。

また、ユーザが対訳コーパスからの検索を行う場合、原文から検索し、検索語に対応する訳語の使用方法を把握することを目的とする場合が多い。このため、単純に原文と訳文を並べるのみでは、訳語の使用方法を把握するという観点で効率が悪い。そこで本研究では、自動的に検索語と対応する訳語を強調する仕組みを取り入れた、具体的には、左右 2 列の KWIC 表示によって結果を提示する機能を構築した。本機能の処理の流れを図 4 に示す。

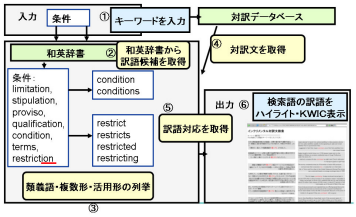

図 4　検索語の訳語対応を自動取得する対訳文検索

6.1 使用データ・モジュール

対訳文データには、NICT 内山氏が作成した、日英新聞記事の対応付けデータ [9] を用いた。また、訳語対応の一致をとるために、EDR 日英対訳辞書を使用した。さらに、EDR によって見出し語が発見できなかった際の補完用として、EDICT [10] のデータも使用した。検索については、Senna [11] を組み込んで日本語全文検索を可能にした MySQL を使用した。Senna は転置インディックス部分のみに特化した検索モジュールで、DBMS 等に組み込んで使用することができる。1 対複数文対応のデータを使用したため、日本文と英文の対応付けデータ 1 件を 1 レコードとした。

6.2 検索語の訳語パターンの展開

検索語に対応する訳語が動詞の場合、辞書には原形しかエントリが存在しない。しかし、実際の文中では活用形の変化をしている場合があるため、活用形のパターンを列挙して、それぞれ比較を行う必要がある。また、検索語に対応する訳語が名詞として抽出された場合でも、実際の文中では動詞として使用されている場合がある。こうした場合に対応するため、「ion」「ing」などの接尾辞がある場合について、上記のような名詞と動詞の相互変換を行った上で対応する訳語の検索を行う。

6.3 文全体を参照する 2 列 KWIC 表示

検索結果の文を参照する際、検索語を縦一列に並べた KWIC 表示にすることにより検索語の前後の文脈を一度に参照しやすくなり、検索語の用法を把握するのに効率がよい。しかし、一般的な KWIC 表示のように文の途中部分だけ抜き出す形にしてしまうと、文全体の構造を参照することができない。検索語を縦一列に並べつつ文全体を表示するために、検索語の横に並ぶ文字数の最大値を設定し、最大値を超えないように単語の配置を行った。最大値を超えるものについては、上下に並べて表示する。

図 5 のように、日本文の検索語が縦一列に並ぶため、ユーザは視線を縦に動かしていくだけで、自分の想定した用法での検索語を探すことができる。そして、自分の想定した用法での検索語を発見した場合、横に表示された英訳文を参照し検索語の周辺文脈から検索語の訳語の用法を参照する。

訳文の意訳の程度により、訳語対応を取得できない場合も

図 5　文全体を表示する 2 列 KWIC 表示

あるが、新聞記事のようにある程度原文に忠実な訳と、辞書に載っている検索語であれば、高い確率で訳語対応を取得できる。

7. 評　　価

評価実験として、18 人の理工系の大学生・大学院生に WEB 上で本システムを試用してもらい、アンケート方式のユーザ評価を行った。

各機能ごとに、「英文書作成時における有用性」を 4 段階（とても有用である、有用である、有用でない、全く有用でない）で評価してもらい、比較対象として、WEB 上で利用可能な機械翻訳サービスの Excite 翻訳 [12] と、人手による編集の対訳コーパスをベースにした多摩美術大学英語用例データベース [13] についても有用性の評価を行ってもらった。単一言語コーパスを利用して語彙組み合わせの検討を行うシステムでは、WEB 上で公開されているものが見つからなかったため、比較対象にはできなかった。

有用性の項目を 1 から 4 の数値として換算し、Excite 翻訳と提案システムの各機能に対して t 検定による比較を行った。帰無仮説を「2 群の平均は等しい」、有意確立を 5 %とし、両側検定での境界値と t 値を比較した。

表 2　Excite 翻訳との英文書作成時の有用性の比較

比較する機能	既存	提案	境界値	t 値
1. ワイルドカード検索を利用した検討	2.89	3.33	2.03	2.49
2. 和英辞書とフレーズテンプレートを利用した検討	2.89	3.44	2.03	3.04
3. 可読性を考慮した検索結果からの例文抽出	2.89	2.78	2.04	-0.44
4. 検索語の訳語対応を自動取得する対訳文検索	2.89	3.17	2.03	1.29

表 2 の項目は左から、提案システムの比較する機能、既存システム (Excite 翻訳) に対する評価の平均値、提案システムの評価の平均値、両側検定での境界値、t 値である。「ワイルドカード検索を利用した検討」と、「和英辞書とフレーズテンプレートを利用した検討」の 2 つの機能で、t の絶対値が境界値を上回っており、提案システムに優位性が見られた。

同様に、多摩美術大学英語用例データベースについても、英文書作成時における有用性の比較を行った。その結果、表 3 のように「可読性を考慮した検索結果からの例文抽出」以外の 3 つの機能について、提案システムに優位性が見られた。「可読性を考慮した検索結果からの例文抽出」の機能については、英文だけ読んでも意味や用法が直感的に把握できないという問題から、評価の平均値が高くなかったものと考えられる。

表 3　多摩美術大学英語用例データベースとの比較

比較する機能	既存	提案	境界値	t 値
1. ワイルドカード検索を利用した検討	2.72	3.33	2.03	3.45
2. 和英辞書とフレーズテンプレートを用いた検討	2.72	3.44	2.03	3.98
3. 可読性を考慮した検索結果からの例文抽出	2.72	2.78	2.05	0.22
4. 検索語の訳語対応を自動取得する対訳文検索	2.72	3.17	2.03	2.07

8. おわりに

本稿では、検索エンジンを利用して語彙の組み合わせの相性を検討する機能や、対訳コーパスから効率的に訳語の用法を学習できる機能により、英文書の作成を支援するシステムの構築を行い、有用性があることを明らかにした。今後は、受動態と能動態の選択のように、大幅な文型の変化にも対応する語彙組み合わせの検討など、有効に適用できる範囲を広げるとともに、英語の知識を問わず誰でも使いやすい英作文支援システムを構築する必要があると考えられる。

参考文献

文　　献

[1] Sawa Takakura, Takeshi Ito and Teiji Furugori:”TransAid: a writer’s aid system for translating Japanese into English”, WA2E3, The Proceeding of IEEE SMC 2002 Vol.6, Oct.(2002)

[2] Takashi Yamanoue, Toshiro Minami, Ian Ruxton and Wataru Sakurai: ”Learning Usage of English KWICly with WebLEAP/DSR”, Proceedings of the 2nd International Conference on Information Technology and Applications (ICITA-2004), 14-6, Harbin, China January. 8-11 (2004)

[3] Kumiko Tanaka-Ishii, Masato Yamamoto, Hiroshi Nakagawa:“ Kiwi: A Multilingual Usage Consultation Tool based on Internet Searching ”, Proceedings of the Interactive Posters/Demonstrations, ACL-03, Sapporo, 105-108, 2003.

[4] 安藤進著, ”翻訳に役立つ Google 活用テクニック”, 丸善, ISBN4-621-07294-3 (2003.10)

[5] GoogleAPI http://www.google.com/apis/

[6] Eric Brill:”A Simple Rule-Based Part of Speech Tagger”,Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing,(1992)

[7] 通信総合研究所,EDR 電子化辞書仕様説明書 (2003)

[8] 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”(2003)

[9] Masao Utiyama and Hitoshi Isahara. (2003) Reliable Measures for Aligning Japanese-English News Articles and Sentences. ACL-2003, pp. 72–79

[10] Jim Breen:”building an electronic japanese-english

[11] Senna 組み込み型全文検索エンジン http://qwik.jp/senna/

[12] EXCITE 翻訳 http://www.excite.co.jp/world/

[13] 多摩美術大学英和訳用例データベース http://studio1.idd.tamabi.ac.jp/corpus/yourei/index.htm