

K_047

WWW言語資源を利用した英文作成支援ツール

- ユーザの語彙レベルに基づく適応型用例提示 -

A Support Tool for English Composition using Language Resources on the WWW - Adaptive Example Sentence Display Based on User's Vocabulary Level -

安藤 一秋†

Kazuaki Ando

綱嶋 祐一†

Yuuich Tsunashima

溝渕 昭二‡

Shoji Mizobuchi

1. はじめに

近年、Web上の膨大なテキストデータを言語資源 (Webコーパス) として、英語学習[1]や英文作成支援[2-6]に利用する研究が注目されている。Webコーパスは内容が日々増え続け、自動更新される動的データといえる。一方で、質的な問題も避けては通れないが、膨大な情報量という魅力は大きい。

本研究では、Webコーパスとオンライン辞書や論文アーカイブなど電子化言語資源をフル活用して、英文作成のための有益な情報提示を目的とした英文作成支援ツール[5]の構築を目的としている。本稿では、システムの内、Google Web APIs (Google API)[7]を利用したWeb・用例検索モジュールの新機能である、ユーザの語彙レベルに適応した用例提示を中心に説明する。

2. 英文作成支援ツールの概要

現在開発中の英文作成支援ツール[5]は、英文作成 (EWS)、Web・用例検索 (WCS)、マイニング (DM) の3つのモジュールで構成される。EWSモジュールは、入力された和表現・和文を形態素解析し、各種の電子化言語資源を利用して、訳語候補の自動検索と選択支援を行う。WCSモジュールは、Webコーパスや静的コーパスを利用して、入力された英表現や英文の妥当性検証と用例検索・提示を行う。DMモジュールは、WCSモジュールで構築された検索ログをマイニングし、ユーザの苦手な単語や表現、構文パターンをフィードバックすることで英語の学習支援を実現する。

図1にWCSモジュールを示す。本モジュールでは、「翻訳に役立つGoogle活用テクニック[6]」から検索式が自動生成できるものをベースに実装した[5]。このテクニックの基本的な考え方は、ヒット数や snippet を利用して妥当性を検証するものである。具体的な機能[5]としては、ワイルドカード及びその候補指定などがある。

ユーザが検証したい表現や文を入力して実行すると、検索式が自動生成される。次にGoogle APIで検索し、検索結果から有用情報 (ヒット数や用例など) を抽出する。この情報はキャッシュとして利用する他、用例検索やマイニングデータとして活用するために検索ログとして保存する。この機能は従来システム[2-4]に見られない特徴的なものである。

本モジュールでは、複数の検索式で実行した場合でも検索結果を自動統合し、ユーザが閲覧しやすい形 (ヒット数ランキング、KWIC表示、ドメイン分類など) で提示

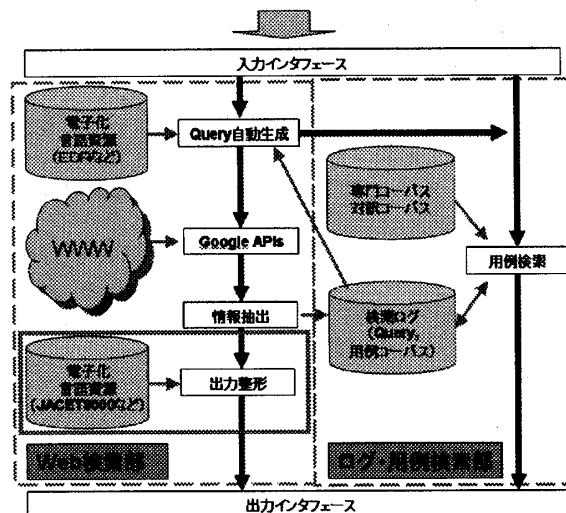


図1 Web・用例検索 (WCS) モジュールの概要

できる。また、入力を n-gram に分割し、それらのヒット数を利用して妥当性を検証する WebLEAP[2]型の提示と接続スコアに基づいて誤り箇所を特定する機能も実装した。

更に新機能として、個人に適応した用例提示機能を試作した。本稿ではこの機能を中心に説明する。

3. 語彙レベルに基づく個人適応型用例提示

3.1 英文の難易度

Web検索部の基本出力は、ヒット数でランキングした結果と Snippet から自動抽出した用例である。前回までのツール[5]の場合、用例は、Googleのランキングを保持した形でトップレベルドメイン毎に分類提示していた。しかし、英語を苦手とするユーザにとっては、検索の適合度に基づくランキングより、可読性が高く自身の英語力に対して難易度が高過ぎない、つまり、ユーザに適応した用例提示の方が望ましい。

英文テキストの難易度[8]は、語彙の難しさと構文構造の難しさの2種類の要因に基づいていると考えられる。英語が母国語である場合は、難易度の設定に単語の文字数、文の長さ、基本語の含まれる割合などが利用される。しかし、日本人の場合は、上記以外に、語彙や関係詞・分詞などの構文構造も英文を難しくする要因となる。

検索エンジンに基づくツールでは処理の大部分を検索時間が占有する。ユーザが快適にツールを利用するためには即応性が問われるため、可能な限り軽い処理で難易度・適応度を判定できることが望ましい。そこで、手始めに、構文構造は考慮せず、用例に含まれる語彙から難易度を推定する。また、難易度だけでなく、ユーザの英語レベルに適応した順で用例を提示することを目指す。

† 香川大学工学部, Faculty of Eng., Kagawa University

‡ 近畿大学理工学部, School of Sci. and Eng., Kinki University

3.2 適応度判定

語彙を利用して、ユーザに適応した用例提示を行うためには、ユーザにとって難しい語彙を設定する必要がある。

「ユーザが知らない語彙=難しい語彙」とする方法が考えられるが、知らない語彙群の構築が難しい。そこで、日本人英語学習者のために作成された JACET8000[9]を利用する。JACET8000は、British National Corpusとサブコーパスを分析し、日本人が国際コミュニケーションを行う上で真に必要な8000語を選定したものである。JACET8000は学習の利便性を高めるために、1000語ずつ8レベルに分類されている。レベル1は中学基本単語、レベル2は高校初級、レベル3はセンター試験、レベル4は、大学一般教養初級、レベル5は大学一般教養、レベル6は英検準1級、レベル7は英検1級、レベル8は一般英語学習の最終到達目標レベルに分類されている。

本研究では、ユーザの英語レベルと JACET8000 のレベルを対応させることで、用例の適応度を判定する。具体的には、JACET8000を基にユーザが設定した自身の英語レベルと用例中に含まれる語彙のレベル差、設定レベル以上の語彙が用例に含まれる割合を利用して、用例の適応度を算出する。

i 番目の用例に含まれる単語数を n_i 、用例 i の j 番目の単語を w_{ij} 、ユーザが設定したレベルを x とした場合、用例 i の適応度 AV_i は、以下で計算される。

$$AV_i = \frac{\left(\max_i + \sum_{j=0}^{n_i} \text{Score}(w_{ij}, x) \right)}{\max_i + n_i} * \frac{n_i - \text{over}_i}{n_i}$$

ここで、 $\max_i$ は、 $\alpha n_i(9-x)$ で得られる値で、 α は定数である。 over_i は、用例 i に含まれる単語の内、レベルが x を越えた単語数である。 $\text{Score}(w_{ij}, x)$ は以下の式で計算される。

$$\text{Score}(w_{ij}, x) = \begin{cases} 0 & (L(w_{ij}) < x) \\ 1 & (L(w_{ij}) = x) \\ -\alpha(L(w_{ij}) - x)^2 & (L(w_{ij}) > x) \end{cases}$$

ここで、 $L(w_{ij})$ は、 w_{ij} に対するJACET8000上の語彙レベルである。

以上の式に基づいて、検索結果に含まれる用例の適応度を計算し、適応度の高い順に提示する。その際、ユーザが設定したレベルより高いレベルに属する語彙が用例に含まれている場合は、その語彙のレベルを表示すると共に、オンライン辞書へのリンクを付与する。これにより、定義・意味などを調べる負荷が軽減できる。

3.3 簡易実験

実験には、5つのフレーズ(deal with など、平均7個の用例を保持)を利用する。英会話講師1名にGoogleがランキング(Google順位)したデータに対し、語彙に基づいて再ランキングしてもらい、その結果を正解データとする。Google順位と適応度で再ランキングした場合の精度を比較する。 α の値は予備実験により、1/4とした。

実験の結果、Google順位の精度は32.9%であった。しかし、再ランキングした場合の精度は66.9%で、よりユーザに適応した順で用例を提示できることが分かった。結果の例を図2に示す。実験結果から、いくつかの問題点も見えた。1つは、一般名詞と固有名詞の誤判定である。今回

key = "deal with" Level=3

search: 1/4 Page: Top

Title	excerpt
NBC strikes deal with YouTube CNET News.com	By landing a joint-marketing deal with NBC, YouTube is flexing some of its ... "That's what this deal with NBC really means. YouTube needs to partner with ..."
0.73673	By landing a joint-marketing deal with NBC, YouTube is flexing some of its ... "That's what this deal with NBC really means. YouTube needs to partner with ..."
0.95452 x 0.85714	By landing a joint-marketing deal with NBC, YouTube is flexing some of its ... "That's what this deal with NBC really means. YouTube needs to partner with ..."
ESPN.com - NCAA - Big Ten has new 10-year deal with ESPN, ABC	Big Ten has new 10-year deal with ESPN, ABC. ESPN.com news services. ESPN and ABC Sports will continue to broadcast Big Ten football, men's and women's
0.82439	Big Ten has new 10-year deal with ESPN, ABC. ESPN.com news services. ESPN and ABC Sports will continue to broadcast Big Ten football, men's and women's
0.88364 x 0.85455	Big Ten has new 10-year deal with ESPN, ABC. ESPN.com news services. ESPN and ABC Sports will continue to broadcast Big Ten football, men's and women's
0.73089	The activity around her blog led to some media coverage and she soon found herself signing a "really obscure book deal" with Little, Brown and Company.
0.83342 x 0.87500	The activity around her blog led to some media coverage and she soon found herself signing a "really obscure book deal" with Little, Brown and Company.

(a) 適用する前の結果

key = "deal with" Level=3

search: 1/4 Page: Top

Title	excerpt
Yahoo! News strikes deal with video-sharing site Tech News on ZDNet	The agreement is the latest sign that at least some in Hollywood may be ready to deal with Internet sites that some in the movie business consider a threat ...
0.85601	The agreement is the latest sign that at least some in Hollywood may be ready to deal with Internet sites that some in the movie business consider a threat ...
0.98398 x 0.85455	The agreement is the latest sign that at least some in Hollywood may be ready to deal with Internet sites that some in the movie business consider a threat ...
Yahoo! News strikes deal with video-sharing site Tech News on ZDNet	Yahoo! News strikes deal with video-sharing site Tech News on ZDNet
0.82439	Yahoo! News strikes deal with video-sharing site Tech News on ZDNet
0.88364 x 0.85455	Yahoo! News strikes deal with video-sharing site Tech News on ZDNet
ESPN.com - NCAA - Big Ten has new 10-year deal with ESPN, ABC	Big Ten has new 10-year deal with ESPN, ABC. ESPN.com news services. ESPN and ABC Sports will continue to broadcast Big Ten football, men's and women's
0.82439	Big Ten has new 10-year deal with ESPN, ABC. ESPN.com news services. ESPN and ABC Sports will continue to broadcast Big Ten football, men's and women's
0.83342 x 0.87500	Big Ten has new 10-year deal with ESPN, ABC. ESPN.com news services. ESPN and ABC Sports will continue to broadcast Big Ten football, men's and women's

(b) 適用後の結果

図2 適応度に基づくランキング結果の例

JACET8000に未登録の語彙はレベル9、固有名詞はレベルなしとして扱った。しかし、本来、レベル9として扱って欲しい語彙がレベルなしと判定され、ランキングが正しくできなかった事例も見られた。固有表現抽出技術を応用するなど、判定に何らかの工夫が必要である。また、未登録語ではあるが、語彙としては簡単なものが、レベル9と判定される問題もあった。これも改善が必要である。

4. おわりに

本稿では、WWW言語資源を利用した英文作成支援ツールの内、英文の語彙レベルを利用した適応型用例提示について報告した。英文の難しさは語彙だけで判定できないため、今後は、熟語の利用や軽い構文解析を利用する方法の考案と共に、使いやすい機能を充実させる。

謝辞

本研究は、文部科学省科学研究費補助金(若手研究(B)16700562)の補助を受けて実施した。

参考文献

- [1] 齋藤他, 改訂新版英語コーパス言語学, 研究社, 2005.
- [2] Yamanoue et al., "Learning Usage of English KWICly with WebLEAP/DSP", Proc. of the 2nd International Conference on Information Technology for Application, pp.380-385, 2004.
- [3] 相原 他, "英作文支援システムの構築", 情処第67回全国大会論文集, 5P-7, 2005年3月.
- [4] 佐藤他, "検索エンジンを利用した英作文支援システムの構築", 言処第12回年次大会, pp.664-667, 2006年3月.
- [5] 安藤他, "インターネット言語資源を利用した英文検証支援ツール", 情処第68回全国大会, p.445-446, 2006年3月.
- [6] 安藤進, "翻訳に役立つGoogle活用テクニック", 丸善, 2003.
- [7] Google Web APIs <http://www.google.com/apis/>
- [8] 永田他, "日本人英語学習者を対象にした英文難易度判定手法", 信学論D-II, No.6, pp.1329-1338, 2004年6月.
- [9] 相澤他, 『大学英語教育学会基本語リスト』に基づくJACET8000英単語, 桐原書店, 2005年11月.