英文用例検索システム ESCORT:
論文データベースを用いた図書館サービス

○松原茂樹[1)], 江川誠二[2)], 加藤芳秀[3)]
名古屋大学情報連携基盤センター[1)],
名古屋大学大学院情報科学研究科[2)],
名古屋大学大学院国際開発研究科[3)]
〒464-8601 名古屋市千種区不老町
Tel: 052-789-4387 FAX: 052-789-4384
E-mail: matubara@nagoya-u.jp

ESCORT: English Sentence Retrieval System: Library Service using Article Database

MATSUBARA Shigeki[1)], EGAWA Seiji[2)], KATO Yoshihide[3)]
Information Technology Center, Nagoya University[1)],
Graduate School of Information Science, Nagoya University[2)],
Graduate School of International Development, Nagoya University[3)]
Furo-cho, Chikusa-ku, Nagoya 464-8601 Japan
Phone: +81-52-789-4387 Fax: +81-52-789-4384
E-mail: matubara@nagoya-u.jp

【発表概要】

本論文では、図書館サービスとして論文作成支援環境を提供することを目的に、英文検索システム ESCORT を提案する。ESCORT は、研究者が英語論文を作成する場面で参照するに相応しい英文用例を提示することにより、適切な英文作成へと研究者を導くことを目指している。英文検索環境としては、キーワードにより文を検索するシステムが提案されてきた。しかし、単に入力されたキーワードを含む文を検索するだけでは、利用者の要求に合致しない英文が多数提示されるという問題がある。それに対して ESCORT では、英文データベースに格納された英文の構文構造を参照することにより、ユーザによって入力されたキーワード間に構文的関係を見出すことができる英文のみを検索結果として提示する。コンピュータサイエンス分野の英語論文に掲載された約 50 万文を対象とした英語用例検索が実現され、実運用されている。

【キーワード】

情報検索, 自然言語処理, 構文解析, 学術論文, テキストコーパス, 機関リポジトリ

1. はじめに

近年、機関リポジトリ(institutional repository)に代表されるように、大学図書館において学術論文データを蓄積する動きが進みつつある。今後は、単にデジタルコンテンツを蓄積し配信するだけでなく、それらを効果的に活用した新しい図書館サービスの展開が望まれる。

本論文では、図書館サービスとして論文作成支援環境を提供することを目的として、英文検索システム ESCORT を提案する。ESCORT は、研究者が英語論文を作成する場面で参照するに相応しい英文用例を提示することにより、適切な英文作成へと研究者を導くことを目指している。研究者が研究成果を発信する上で、英語による学術論文の作成が不可欠であり、日本をはじめ、英語ネイティブでない研究者のニーズに合致したサービスの提供が可能となる。

英文検索環境としては、これまでにも、キーワードにより文を検索するシステムが提案されてきた(例えば、[1])。しかし、単に入力されたキーワードを含む文を検索するだけでは、利用者の要求に合致しない英文が多数提示されるという問題がある。

それに対して ESCORT では、英文データベースに格納された英文の構文構造を参照し、ユーザによって入力されたキーワード間に構文的関係を見出すことができる英文のみを検索結果として提示する。現在、コンピュータサイエンス分野の英語論文に掲載された約 50 万文を対象とした英語用例検索を実現しており、実運用されるに至っている。

2. ESCORT

2.1 構文情報を用いた文検索

これまで、KWIC などキーワードに基づく文検索システムが提案され、使用されてきた。基本的にこれらのシステムでは、ユーザが入力したすべてのキーワードを含む文を検索し提示する。しかし、キーワードを包含していることのみを考慮した検索では、必ずしも利用者の意図に合致した検索を実行することはできない。

例として、利用者が“develop system”をキーワードとして入力した場合を考える。以下の文

(1) The techniques developed here improved the system.

は検索クエリに合致した英文であるが、この文が利用者の意図する英文である可能性は低い。なぜなら、すべてのキーワードを含んでいるものの、それらの間に直接的な関係がないためである。このように、複数のキーワードを入力した場合、ユーザは暗に、それらの間に何らかの関係性を想定しているものと予想される。

本研究では、そのような関係性を語彙的な依存関係として捉え、依存関係が存在する英文のみを検索結果として提示する。この場合、上記(1)の文は検索結果には含まれないが、

(2)We have developed a document retrieval system.

のように、develop と system の間に依存関係が見出せる文については、検索結果として提示する。

2.2 システムの概要

英文検索システム ESCORT は、英文データベース、ならびに検索モジュールから構成される。

英文データベースは、英語で書かれた学術論文から抽出した英文が格納されている。近年の学術論文の多くは、PDF 形式で保存されており、PDF ファイルをテキストファイルに変換した上で英文を抽

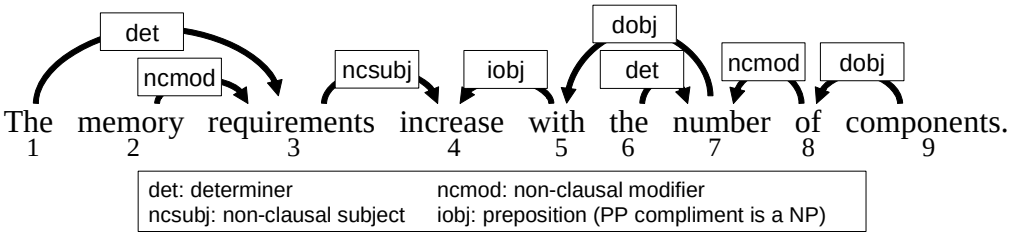

図1. 依存構造の例

出する。抽出したすべての英文に対して、依存構造解析を与えている。

一方、検索モジュールでは、まず、入力されたすべてのキーワードを含む英文をデータベースから抽出する。次に、その依存構造パターンを同定することにより、キーワード間になりたつ依存関係をもとめ、該当する英文を依存構造パターンごとに分類して提示する。

ESCORT の特徴は以下の通りである。

◎ キーワード間に構文的関係が存在する英文のみを検索できる。
◎ ユーザは単にキーワードを入力するだけでよく、英文法に精通している必要はない。
◎ 検索結果は、構文的関係の種類に応じて分類し、提示される。

2.3 アルゴリズム

依存構造パターンを同定するアルゴリズムは、文献[2]の拡張により実現した。入力は

■クエリ: $q_1 \cdots q_m (q_i (1 \le i \le m)$ は、キーワード)

■英文: $s = w_1 \cdots w_n (w_j (1 \le j \le n)$: は単語と品詞の対)

■依存構造: D

である。ここで、依存構造 D は、文 s に含まれる単語間依存関係の集合である。単語 w_i が w_j に依存し、かつ、その間に関係 r が成り立つなら、3項組 (i, j, r) は D の要素である。図1に依存構造の例を示す。

クエリを構成するすべてのキーワードが英文に存在し、かつ、それらの間に依存関係がなりたつかどうかを検査する。そのために、上述した入力のもとで、依存構造パターンの生成を実行する。

依存構造パターンとは、5項組 $d = (h, L, R, D_L, D_R)$ である。ここで h は、d の主辞(head)であり、文中における単語の位置を示す。L, R は依存構造のリストであり、D_L, D_R は、依存関係のリストである。L の i 番目の要素の依存構造パターンが h に左から依存し、その依存関係の種類が D_L の i 番目の要素にある。R, D_R については右からの依存である。

依存構造パターンの生成は、以下の手続きをクエリ $q_1 \cdots q_m$ にボトムアップに適用することによって実行する。

初期化(initialization): $q_i (1 \le i \le m)$, $w_j (1 \le j \le n)$ に対して、q_i が w_j の単語または品詞にマッチするとき、q_i に対する依存構造パターン $(j, \varepsilon, \varepsilon, \varepsilon, \varepsilon)$ を生成する。

結合(combining): $d = (h, L, R, D_L, D_R)$, ならびに $d' = (h', L', R', D_L', D_R')$ を、それぞれ $q_i \cdots q_j$ ならびに $q_{j+1} \cdots q_k$ の依存構造パターンであるとする。このとき、ある r に対して $(h, h', r) \in D$ で、かつ、$R' = \varepsilon$ なら、$q_i \cdots q_j q_{j+1} \cdots q_k$ に対するパターン $(h', dL', \varepsilon, rD_L', \varepsilon)$ を生成する(図2a参照)。$(h', h, r) \in D$ で

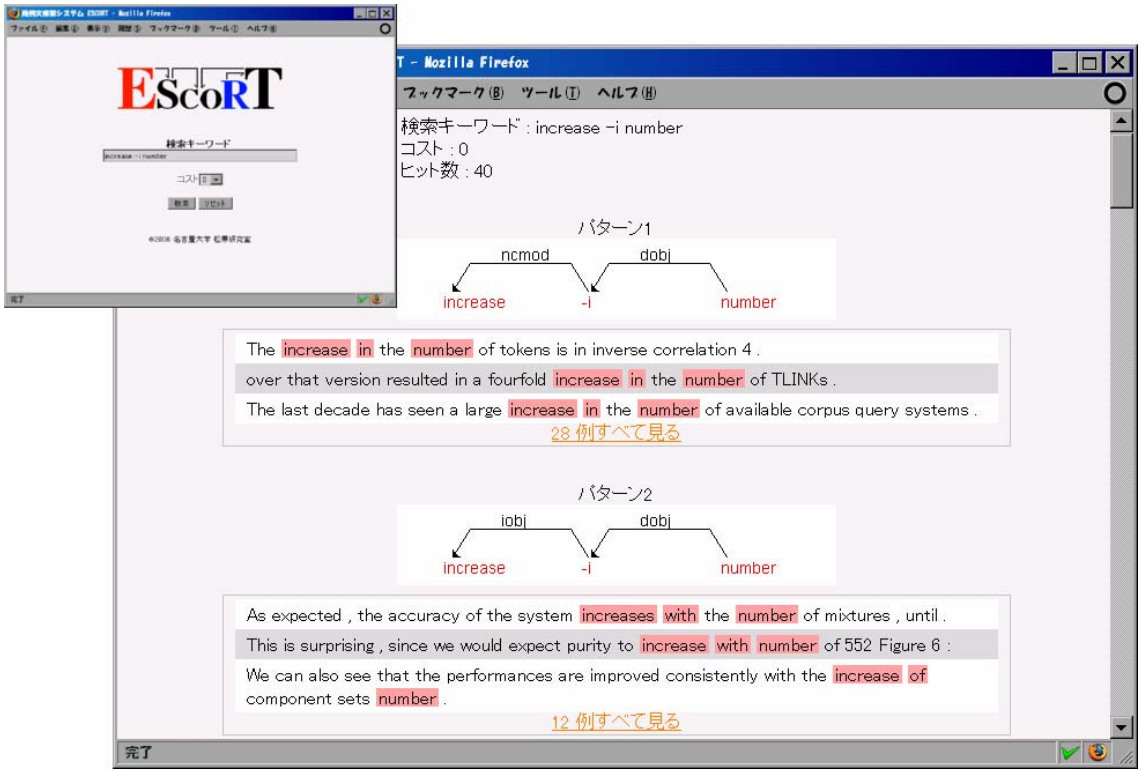

図3. ESCORT の検索画面と検索結果

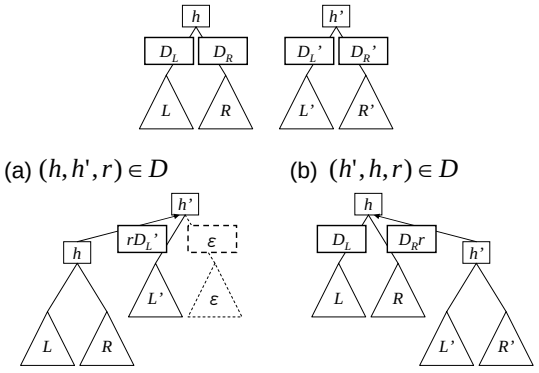

図２．結合操作

あれば、$(h, L, Rd', D_L, D_R r)$を生成する(図2b参照)。

3. 実装と動作例

3.1 システムの実装

前節のアルゴリズムに基づき英文検索システムを Web アプリケーションとして実装した。論文データベースとして、コンピュータサイエンス分野の国際会議録を使用した。PDF ファイルを pdftotext[3]を用いてテキストファイルに変換し、英文を取り出した。依存構造解析器 RASP[4]を用いて、すべての英文に対して依存構造を計算した。現在、502,235 文が対象となっている。

図３に ESCORT の検索画面を示す。検索結果は依存構造パターンに基づいて分割される。依存構造パターンごとにその代表的な文例が提示され、ユーザが望めば、該当するすべての英文を参照することができる。

3.2 動作例

例として、図１に示したような文を検索することを想定し、ユーザがクエリ“increase 前置詞 number”を入力したときの ESCORT の実際の動作について説明する。“increase”, “前置詞”,

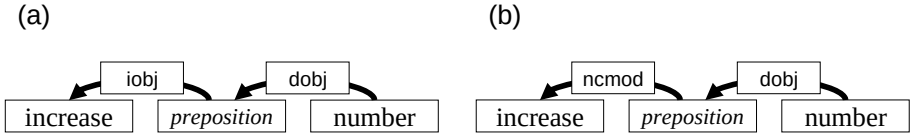

図4. “increase 前置詞 number” の依存構造パターン

“number” がこの順で出現する英文はデータベース内に 102 文あり、システムはそこから図4に示された2つの依存構造パターンを生成する。

パターン(a)は “number” が前置詞を介して動詞 “increase” に依存することを意味し、パターン(b)は、 “number” が前置詞を介して名詞 “increase”に依存することを意味している。(a)に 12 文が、(b)に 28 文が分類された。図1のような英文はパターン(a)に属しており、この例からもユーザが探している文を容易に見つけられることがわかる。

別の例として、「システムは～のニーズに応える」を英語で表現したときに使用する動詞を調べるために、“system –v needs” をクエリとして入力したときの動作を以下に示す。“system” が主語、“needs”が目的語として動詞に依存する用例として 6 文提示され、

“… advanced information systems to address complex information needs …”,

“… the systems did indeed satisfy the users’ specific needs.”

“… the system could potentially meet the users’ needs.”

などを参照することができる。その結果、上記の意図に合致した動詞として、“address”, “satisfy”, “meet” などが使用されていることがわかる。

4. おわりに

本論文では、英語ネイティブでない研究者による学術論文作成を支援するための環境として英文検索システムを提案した。キーワード間の依存関係を考慮することにより、ユーザの意図に合致した英文のみを検索結果として提示できる。

ESCORT は、名古屋大学情報連携基盤センターが運営する IT ラボ[5]のコンテンツの一つとして学内に公開され利用されている[6]。今後は、データベースの大規模化、ならびに、検索の高速化がはかられる予定である。

大学図書館における学術論文コンテンツの蓄積が進みつつある。本論文で提案した文検索機能を新しい図書館サービスとして本格的な運用を推進するとともに、英語学習者のための自習用コンテンツとして提供することを検討している。

5. 参考文献

[1] K. Tanaka and H. Nakagawa: A multilingual usage consultation tool based on internet searching: more than a search engine, less than QA, *Proceedings of WWW-2005*, pp. 363 -371 (2005).

[2] 加藤、松原、稲垣: 依存構造に基づくコーパス検索, 電子情報通信学会論文誌, Vol.J89-D, No.12, pp.2766-2770 (2006).

[3] http://www.foolabs.com/xpdf/

[4] T. Briscoe, J. Carroll, and R. Watson: The second release of the RASP system, *Proceedings of COLING/ ACL-2006*, pp. 77-80 (2006).

[5] http://lab.itc.nagoya-u.ac.jp/

[6] http://escort.itc.nagoya-u.ac.jp/