英語論文作成のための用例文検索システム

江川　誠二 †　　加藤　芳秀 ‡　　松原　茂樹 §

†名古屋大学大学院情報科学研究科
‡名古屋大学大学院国際開発研究科
§名古屋大学情報連携基盤センター

Example Sentence Retrieval System for Writing Research Paper

Seiji Egawa †　　Yoshihide Kato ‡　　Shigeki Matsubara §

† Graduate School of Information Science, Nagoya University
‡ Graduate School of International Development, Nagoya University
§ Information Technology Center, Nagoya University

1　はじめに

本論文では，英語論文作成支援環境の提供を目的として，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．研究成果の発信には，英語による学術論文の作成が不可欠であり，英語ネイティブでない研究者のニーズに合致しているといえる．

用例文検索環境としては，従来より，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 [4]．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，ユーザ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った．

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．現在，コンピュータサイエンス分野の英語論文に含まれる約 18 万文を対象とした英文用例検索が実現され，実運用されている．

2　ESCORT: 英文用例検索システム

2.1　構文的情報を用いた文検索

これまでに，キーワードに基づく文検索システムが多く提案され，使用されてきた．これらのシステムのほとんどは，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．しかしながら，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，必ずしもユーザの意図に合致した検索結果を提示できるとは限らない．

例として，ユーザが develop, system をキーワードとして入力した場合を考える．以下の文はそのキーワードをすべて含む英文の一例である．

(1) The techniques developed here improved the system.

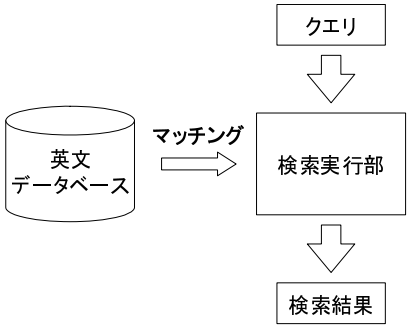

図 1: システムの構成

しかし，これがユーザの意図する英文である可能性は低いと思われる．なぜなら，この文は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，ユーザが複数のキーワードを入力した場合，それらの間に何らかの関係性を暗に想定しているものと予想される．

提案システムでは，そのような関係性は語彙的な依存関係によって捉えられると仮定し，キーワード間に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．これにより，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は，検索結果として提示される．

2.2　システムの構成

英文検索システム ESCORT の構成を図 1 に示す．システムは，英文データベース，ならびに検索実行部から構成される．

英文データベースには，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．すべての英文に対して，文中の各単語がそれぞれどのような関係にあるのかを表す依存構造を付与

してある．図 2 に依存構造の例を示す．たとえば，この例では，requirements は increase の主語であり，number は with の目的語である．

検索実行部は，まず検索キーワードをすべて含む文をデータベースから抽出する．次に，抽出した各文に対し，次節で説明する依存構造パターン同定アルゴリズムを適用し，入力されたキーワードが文中で形成する依存関係を同定する．最後に，同定された依存構造パターンに従って文を分類し，ユーザに提示する．

本システムの特徴は以下の通りである．

- キーワード間に文法的関係が存在する英文のみを検索できる．
- ユーザは英文法に精通している必要はなく，単にキーワードを入力するだけでよい．
- 検索結果は，構文的関係の種類に応じて分類される．

2.3 アルゴリズム

本節では，依存構造パターン同定アルゴリズムについて説明する．本アルゴリズムは，文献 [1] のアルゴリズムを拡張したものである．入力として，

クエリ $q_1 \cdots q_m$ ($q_i (1 \leq i \leq m)$ はキーワード)

文 $s = w_1 \cdots w_n$ ($w_j (1 \leq j \leq n)$ は単語と品詞の対)

文の依存構造 D

を受け取り，依存構造パターンの集合を出力する．ここで D は，文 s の単語間の依存関係の集合である．w_j が w_k に依存し，その関係の種類が r であるとき，(j, k, r) は D の要素である．

依存構造パターンは 5 項組 $d = (h, L, R, D_L, D_R)$ で，h は単語位置であり，これを d の主辞と呼ぶ．L，及び R は依存構造パターンのリストである．L 中の依存構造パターンの主辞が左から h に依存することを意味し，R の場合は右からの依存を意味する．

D_L，D_R は依存関係の種類のリストである．D_L の i 番目の要素は，L の i 番目の要素の主辞と h との依存関係の種類を表す．D_R についても同様である．

依存構造パターンは，クエリ $q_1 \cdots q_m$ に対して以下の操作をボトムアップに適用することにより生成する．

初期化

各 $q_i (1 \leq i \leq m)$，$w_j (1 \leq j \leq n)$ に対して，q_i が w_j の単語あるいは品詞とマッチするならば，q_i に対する依存構造パターンとして $(j, \varepsilon, \varepsilon, \varepsilon, \varepsilon)$ を生成する．

結合操作

$d = (h, L, R, D_L, D_R)$，及び $d' = (h', L', R', D'_L, D'_R)$ をそれぞれ，$q_i \cdots q_j$，及び $q_{j+1} \cdots q_k$ に対する依存構造パターンとし，d 中の最も右に出現する単語が，d' 中の最も左に出現する単語より左にあるとする．このとき，ある r が存在し，$(h, h', r) \in D$ かつ $R' = \varepsilon$ ならば，$q_i \cdots q_j q_{j+1} \cdots q_k$ に対する依存構造パターン $(h', dL', \varepsilon, rD'_L, \varepsilon)$ を生成する (図 3(a) 参照)．$(h', h, r) \in D$ ならばパターン $(h, L, Rd', D_L, D_R r)$ を生成する (図 3(b) 参照)．

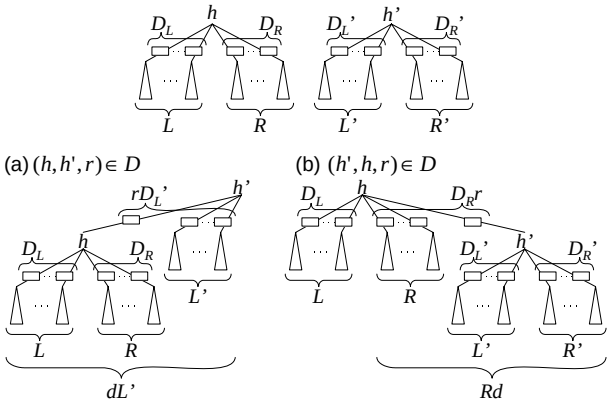

図 3: 結合操作

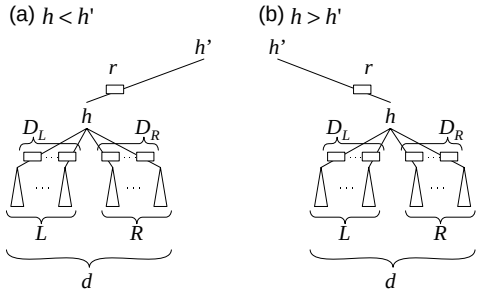

図 4: 補完操作

上記の操作だけでは，クエリ中のキーワード同士が直接依存関係を持たないような文を検出できない．そこで，文献 [2] では，補完操作を導入している．これを上記の操作と同様に拡張する．

補完操作

$d = (h, L, R, D_L, D_R)$ を $q_i \cdots q_j$ に対する依存構造パターンとする．$(h, h', r) \in D$ であるような h' と r が存在するとき，$h < h'$ ならば，$q_i \cdots q_j$ に対する依存構造パターン $(h'^*, d, \varepsilon, r, \varepsilon)$ を生成する (図 4(a) 参照)．$h > h'$ ならばパターン $(h'^*, \varepsilon, d, \varepsilon, r)$ を生成する (図 4(b) 参照)．

主辞 h' に付与された記号 * は，h' が補完操作によって導入されたことを意味する．

この操作により，クエリ中のキーワード間に直接の依存関係がないような依存構造パターンを生成できる．ただし，補完操作を繰り返すと，クエリ中のすべてのキーワードを含むあらゆる文について依存構造パターンが生成されてしまう．これを防ぐために，補完操作の回数をコストと考え，コストに対する閾値を設定し，コストが閾値以下の依存構造パターンのみを検索結果とする．

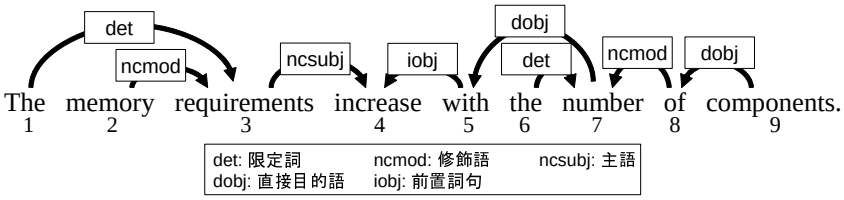

図 2: 依存構造の例

図 5: クエリ入力画面

3 実装

前節で提案した手法を用いて英文用例検索システム ESCORT を実装した．実装には Perl を用いた．システムは Web 上で利用できる．

検索対象として，コンピュータサイエンス分野の英語論文を用いた．PDF ファイルから英文を取り出すにあたり，まず PDF ファイルを pdftohtml ツール [6] で XML ファイルに変換し，その後，文献 [5] の手法を参考にして，text タグで囲まれた各文字列の始端と終端の座標をもとに段落 (本文) のみを抽出した．抽出した英文を，依存構造解析器 RASP[3] で解析した．文数は 185,488 文である．

検索システムのクエリ入力画面を図 5 に示す．画面中央の入力ボックスに検索クエリを入力し，コストの閾値を選択して，検索を行う．システムの検索結果は図 6 のように，同定された依存構造パターンごとに分類されて表示される．表示された文は代表例であり，その下のリンクから該当するすべての文を確認するための画面を表示できる．

4 動作例

本節では，システムの具体的な動作例を示す．

「〜の数とともに増加する」という用例を探すために，“increase 前置詞 number” というクエリで検索した場合を考える．“increase”， “前置詞”， 及び “number” をこの順で含む文は

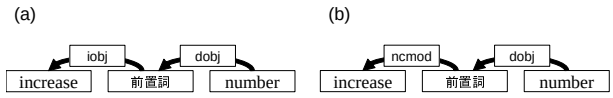

図 7: “ increase 前置詞 number ”に対する依存構造パターン

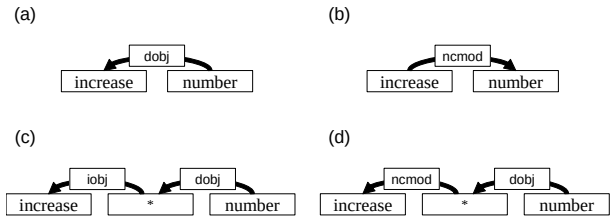

図 8: “ increase number ”に対する依存構造パターン

79 文存在した．このうち 27 文が，図 7 の 2 種類の依存構造パターンのいずれかに分類された (いずれもコスト 0)．

パターン (a) は “number” が前置詞を介して動詞の “increase” に依存するパターン，パターン (b) は “number” が前置詞を介して名詞の “increase” に依存するパターンである．分類された文数はそれぞれ 9 文，18 文であった．(a) に分類された文はすべて検索目的に合致しており，一方，(b) はいずれも目的に合わない文であった．

上述の例では，increase と number の間に前置詞が入ることをユーザが知っていることを仮定していた．しかし，ユーザがそれを知らないということも考えられる．そこで，そのような状況を想定し，同じく「〜の数とともに増加する」という用例を探すために，“increase number” というクエリが入力された場合を考える．コストの閾値が 0 では目的の用例は得られない一方，閾値が 1 のときの動作は以下の通りである．

“increase”，及び “number” をこの順で含む文は 212 文存在した．このうち 143 文が，11 種類の依存構造パターンに分類された．11 種類のうち主な 4 種類の依存構造パターンを図 8 に示す．それぞれのパターンは以下のような依存関係を示している．

(a) “number” が，動詞 “increase” の直接目的語である．

(b) 動詞 “increase” が，分詞として “number” を修飾する．

(c) “number” が，ある単語を介して動詞の “increase” に依存する．

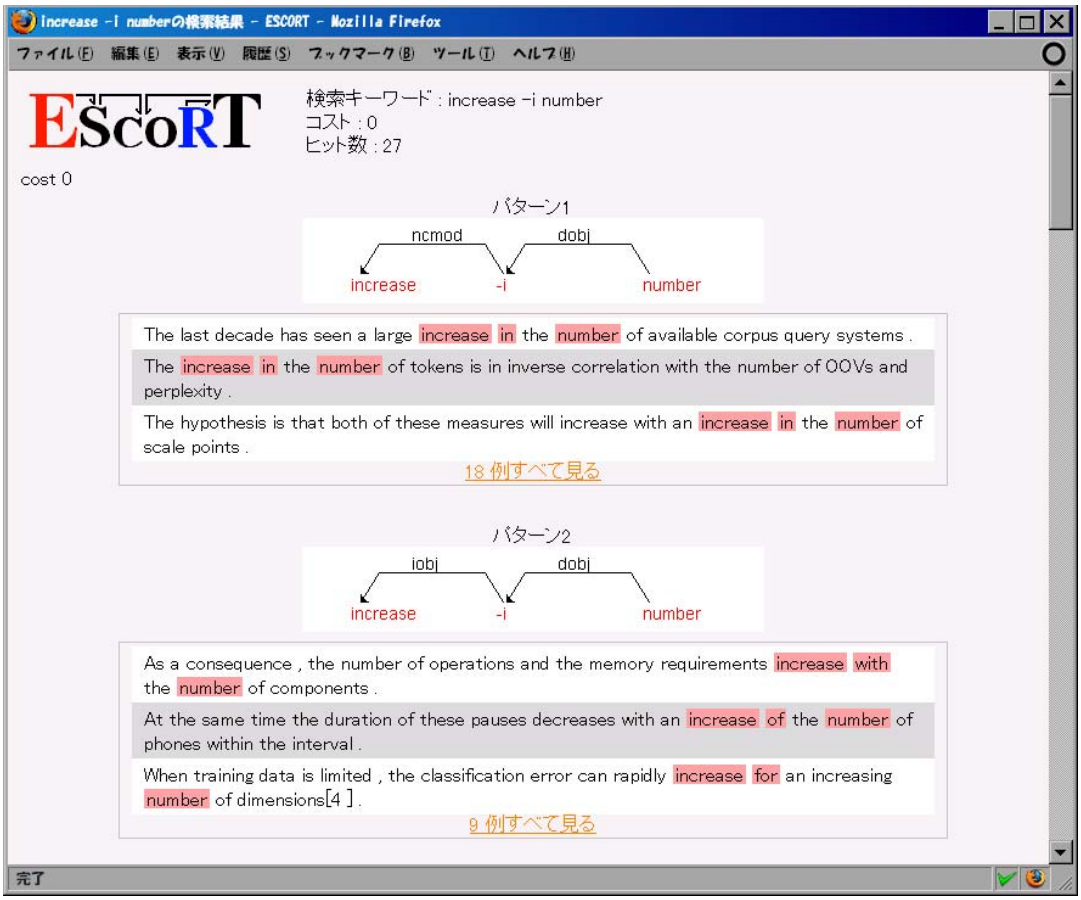

図 6: 検索結果画面

(d) “number” が，ある単語を介して名詞の “increase” に依存する．

(a)，(b) はコスト 0 の依存構造パターン，(c)，(d) はコスト 1 の依存構造パターンである．なお，クエリ中のキーワード以外の単語を*で表している．(c) に分類された文はすべて検索目的に合致しており，(c) 以外の 10 種類はいずれも目的に合わない文であった．

以上のように，ESCORT は，依存関係を用いることにより，ユーザが求めるよう例文を，そうでない文と区別して提示することができる．

5 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 [7] 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 [8]．今後の課題として，データベースの大規模化，検索の高速化があげられる．また，被験者実験による用例文分類の評価を予定している．

参考文献

[1] 加藤芳秀，松原茂樹，稲垣康善, “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.

[2] Y. Kato, S. Matsubara, Y. Inagaki, “A Corpus Search System Utilizing Lexical Dependency Structure,” In *Proceedings of LREC-2002*, pp. 2269-2272, 2006.

[3] T. Briscoe, J. Carroll, R. Watson, “The Second Release of the RASP System,” In *Proceedings of COLING/ACL-2006*, pp. 77-80, 2006.

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

[5] Y, Ishitani, “Logical structure analysis of document images based on emergent computation,” In *Proceedings of IEICE TRANS*, pp. 1831-1842, 2005.

[6] http://pdftohtml.sourceforge.net/

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

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