ラベル付き依存関係に基づく英文用例検索システム

江川　誠二†　　加藤　芳秀‡　　松原　茂樹§

†名古屋大学工学部　‡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

egawa@el.itc.nagoya-u.ac.jp

1　はじめに

大規模なコーパスを効果的に活用するためには，コーパス検索環境の利用が不可欠である．これまでに著者らは，依存構造を用いたコーパス検索手法を提案している [1]．この手法では，キーワード系列をクエリとして受け取り，キーワードをこの順で含むコーパスの各文において，キーワードが形成する依存構造パターンを同定し，文をパターン別に分類する．ユーザは，キーワードを入力するだけで依存構造を考慮した検索を行うことができる．

しかしこの手法では，各キーワード間の依存関係の有無のみを考慮し，その種類は利用しておらず，関係の種類に応じた分類はできない．

そこで本稿では，依存関係の種類を活用してコーパス中の文を分類する方法を提案する．従来手法では，依存構造パターンを組み立てる際に依存関係の向きの情報のみを使用していた．これに対して，本手法では，依存関係の種類の情報まで考慮してパターンの生成を行う．これにより，より細かく分類された検索結果をユーザに提示できる可能性がある．英文用例検索システム ESCORT を実装した．システムの動作例により，本手法が従来手法に比べて，英文用例を有効に分類することを確かめた．

2　依存構造に基づくコーパス検索システム

本節では，著者らが提案した従来の検索システム [1] について説明する．システムは，ユーザから入力としてキーワード系列を受け取る．コーパスの各文にあらかじめ付与された依存構造を参照しながら，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．同定された依存構造パターンに従って文を分類してユーザに提示する．

依存構造パターンの同定は，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 に依存するとき，その単語位置の対 (j, k) は D の要素である．

依存構造パターンは 3 項組 $d = (h, L, R)$ 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．

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 $(j, \varepsilon, \varepsilon)$ を生成する．

結合操作　$d = (h, L, R)$，及び $d' = (h', L', R')$ 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，$(h, h') \in D$ 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', ε) を生成する．$(h', h) \in D$ ならばパターン (h, L, Rd') を生成する．

結合操作の様子を図 1 に示す．(a) に示すような 2 つの依存構造パターンが存在するとき，h が h' に依存しているならば (b) のように結合する．h' が h に依存するなら (c) のようになる．

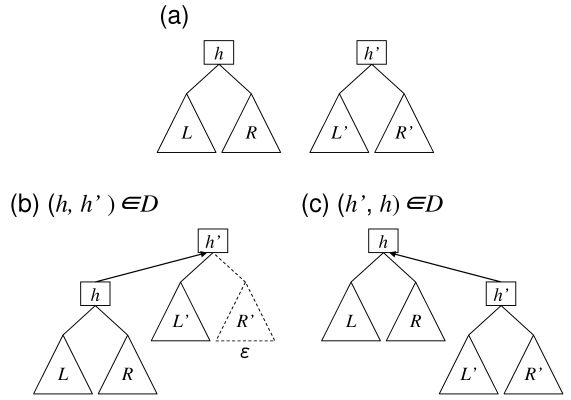

図 1: 結合操作

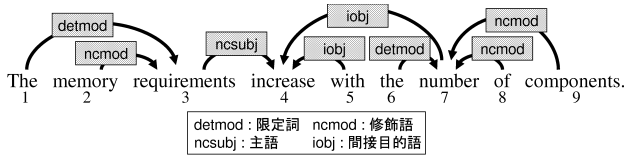

図 2: The memory requirements increase with the number of components. に対する依存関係

例として，図 2 の示す文とその依存構造，及びクエリ“increase 前置詞 number”が与えられたときについて考える．図 2 の依存構造において，“前置詞”が“increase”に右から依存し，“number”が“increase”に対してさらに右から依存するという関係が成り立つので，キーワード“increase”，“前置詞”，及び“number”を結合し，図 3 の依存構造パターンを生成する．

クエリ中のキーワードが直接，依存関係を持たないような文も同様に検出するには，次に定義する補完操作を用いる [2].

補完操作 $d = (h, L, R)$ を $q_i \cdots q_j$ に対する依存構造パターンとする．$(h, h') \in D$ かつ $h < h'$ ならば，$q_i \cdots q_j$ に対してパターン (h'^*, d, ε) を生成し（図 4(a) 参照），$h > h'$ ならば，パターン (h'^*, ε, d) を生成する（図 4(b) 参照）．

主辞 h' に付与された記号 $*$ は，h' が補完操作により導入されたことを表す．依存構造パターンの生

図 3: “increase 前置詞 number”に対する依存構造パターン

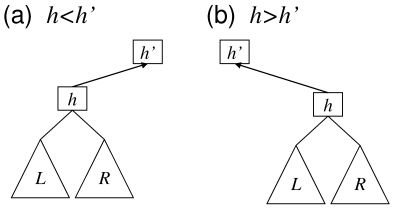

図 4: 補完操作

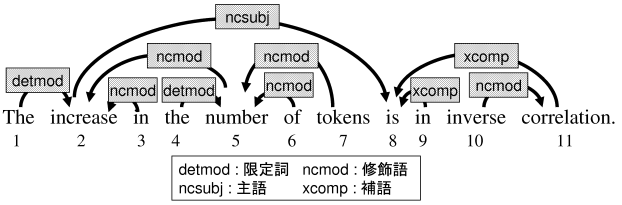

図 5: The increase in the number of tokens is in inverse correlation. に対する依存関係

成に用いられた補完操作の回数をコストとし，コストがある閾値を上回らないように制限することにより，キーワード同士の依存関係のパスが大きすぎる文の検出を防いでいる．

クエリ中のキーワードをすべて含む依存構造パターンが生成された文のうち，同じパターンを持つ文同士をまとめて提示することにより，ユーザは検出された文におけるキーワード間の関係を容易に知ることができる．

3 依存関係の種類を考慮した依存構造パターンの同定

文献 [1] の手法では，依存関係の種類を考慮した分類ができない．例えば，図 2，及び図 5 の文と依存構造に対して，クエリ“increase 前置詞 number”を用いてパターンを同定すると，“increase”と“前置詞”の間，及び“increase”と“number”の間の依存関係に異なるラベルが付与されているにも関わらず，ともに図 3 のパターンが生成される．クエリ中のキーワード間の関係を用いてさらに細かく分類することにより，より正確な分類ができると考えられる．

そこで本手法では，依存構造パターンを拡張し，依存関係の種類によって区別する．

以下では，単語 w_j が単語 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 の要素であるとする．

依存構造パターンを 5 項組 (h, L, R, D_L, D_R) に拡張する．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

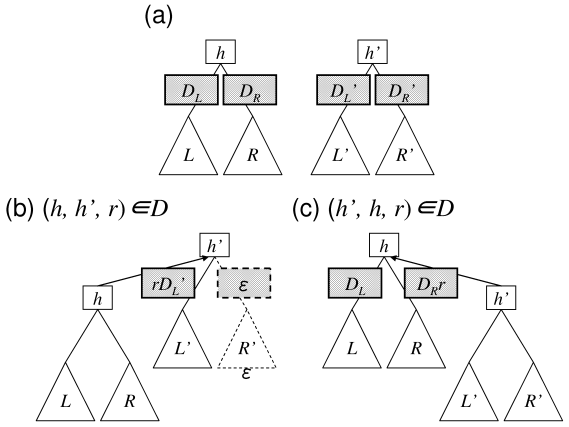

図 6: 拡張された結合操作

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．

結合操作，及び補完操作において，関係の種類を D_L，あるいは D_R に記録することにより，関係の種類を考慮したパターンを生成できる．すなわち，結合操作，及び補完操作を次のように拡張する．

結合操作 $q_i \cdots q_j$ に対する依存構造パターン $d = (h, L, R, D_L, D_R)$ と $q_{j+1} \cdots q_k$ に対するパターン $d' = (h', L', R', D'_L, D'_R)$ に対して，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 6(b) 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 6(c) 参照）．

補完操作 $q_i \cdots q_j$ に対する依存構造パターン $d = (h, L, R, D_L, D_R)$ があり，ある r が存在し，$(h, h', r) \in D$ かつ $h < h'$ ならば，$q_i \cdots q_j$ に対してパターン $(h'*, d, \varepsilon, r, \varepsilon)$ を生成し，$h > h'$ ならば，パターン $(h'*, \varepsilon, d, \varepsilon, r)$ を生成する．

例として，図 2 の依存構造と文，及びクエリ“ increase 前置詞 number ”に対する依存構造パターンの同定を考える．まず，クエリ中のキーワード“ increase ”，“ 前置詞 ”，及び“ number ”に対してそれぞれ，

$$(4, \varepsilon, \varepsilon, \varepsilon, \varepsilon) \quad (1)$$

$$(5, \varepsilon, \varepsilon, \varepsilon, \varepsilon) \quad (2)$$

$$(7, \varepsilon, \varepsilon, \varepsilon, \varepsilon) \quad (3)$$

を生成する．次に，依存関係 $(5, 4, iobj)$ が成り立つので，(1) と (2) を結合して“ increase 前置詞 ”に対するパターン，

$$(4, \varepsilon, (5, \varepsilon, \varepsilon, \varepsilon, \varepsilon), \varepsilon, iobj) \quad (4)$$

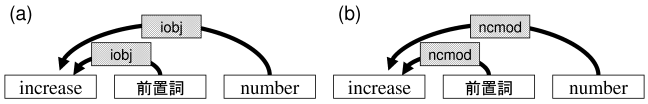

図 7: “ increase 前置詞 number ”に対するラベル付き依存構造パターン

を生成する．さらに，依存関係 $(7, 4, iobj)$ が成り立つので，(4) と (3) を結合して“ increase 前置詞 number ”に対するパターン

$$(4, \varepsilon, (5, \varepsilon, \varepsilon, \varepsilon, \varepsilon)(7, \varepsilon, \varepsilon, \varepsilon, \varepsilon), \varepsilon, iobj\ iobj) \quad (5)$$

を生成する（図 7(a) 参照）．

一方，図 5 の依存構造に対しては，

$$(2, \varepsilon, (3, \varepsilon, \varepsilon, \varepsilon, \varepsilon)(5, \varepsilon, \varepsilon, \varepsilon, \varepsilon), \varepsilon, ncmod\ ncmod) \quad (6)$$

を生成する（図 7(b) 参照）．

このように，依存関係の種類を考慮したパターンを生成できる．

4 実装と動作例

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．検索対象として，自然言語処理分野の論文 PDF ファイルを pdftotext ツール [3] でテキストファイルに変換し，依存構造解析器 RASP[4] で解析した結果を用いた．文数は 454,217 文である．

検索システムの入力画面を図 8 に示す．画面上段は検索クエリの入力ボックスである．画面中央部のチェックボックスにより，検索の対象とするコーパスを選択する．システムの検索結果は図 9 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．

以下に，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 100 文存在した．キーワード間の関係を RASP が誤って解析した 31 文を除く 69 文について，システムによる分類結果を考察する．

69 文のうち 27 文は，図 7 に示した 2 種類の依存構造パターンに分類された（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)

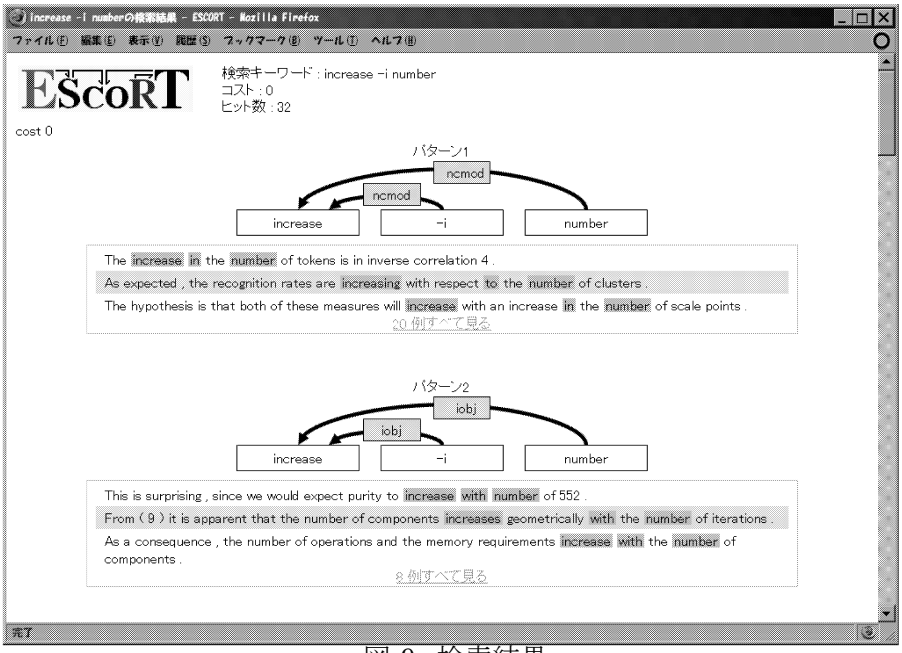

図 9: 検索結果

図 8: 英文用例検索システム ESCORT の入力画面

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 8 文，19 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．

これら 2 種類のパターンは依存関係の向きが同じであるため，従来のシステムでは同一のパターンとなり，文は分類されることなく提示される．依存関係の種類の導入によりそれぞれ分類し，目的の用例 8 文を，他の用例と分けて提示することができた．

5 おわりに

本稿では，依存構造パターンに基づく文の分類において，依存関係の種類を利用して，より細かく分類する手法を提案した．提案手法に基づいて英文用例検索システム ESCORT を実装し，検索実験によりその有効性を確認した．

今後の課題として，細かく分類されすぎている依存関係の種類をまとめることや，依存構造解析の誤った文を自動で検出し排除することなどが挙げられる．

謝辞 本研究の一部は，名古屋大学学術振興基金，並びに科学研究費基盤研究 (A)(2)(課題番号 : 1620001)，若手研究 (B)(課題番号 : 17700145) の助成を受けています．

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] 加藤芳秀，松原茂樹，稲垣康善，“依存構造に基づく英語用例検索システム，”言語処理学会第 12 回年次大会，pp. 668-671 (2006).

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

[4] Briscoe E. and Carroll J, “Robust Accurate Statistical Annotation of General Text,” Proceedings of LREC-2002, pp.1499-1504 (2002).