日本語キーワードをクエリとする言語横断型英文用例検索システム

葛原 和也*，江川 誠二，加藤 芳秀，松原 茂樹 (名古屋大学)

Cross-Lingual English Sentences Retrieval System with Japanese Queries

Kazuya Kuzuhara, Seiji Egawa,Yoshihide Kato, Shigeki Matsubara (Nagoya University)

1 はじめに

英語で論文を書く際，実際の英語論文に現れる用例を検索し，自分が書いた文の用法を確かめることは有用である．英文用例検索システム ESCORT[1] では，入力された英語キーワード間の依存関係の種類ごとに分類して用例を表示できる [2]．

しかし，自分の書きたい英文の内容が日本語では分かっているものの，ある日本語単語に対応する英単語が複数思いつく場合，または，思いつかない場合，英単語を適切に入力することは難しい．そのため，日本語をクエリとして認めることが考えられるが，日本語のキーワードを入力する上で，訳語選択と語順の違いが問題となる．

そこで本稿では，日本語キーワードをクエリとする言語横断型英文用例検索を行うために，上述の問題を解決し，ESCORT への日本語キーワード入力を効率化する手法を提案する．この手法を用いて日本語キーワードによる検索を行った結果，複数語のクエリによる検索において検索時間が大幅に短縮された．

2 ESCORT

2.1 ESCORT の概要

ESCORT[1] では，英語キーワード系列をクエリとして受け取り，英語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関係のパターンを同定し，英文をパターン別に分類して出力する [2]．

2.2 ESCORT における言語横断検索

本手法では，ESCORT への日本語キーワードの入力を可能にするために，入力された日本語単語を，それぞれ和英辞書を用いて英単語に翻訳し検索を実行する．そのとき，訳語選択および語順が問題となる．

訳語選択の問題 ある日本語単語に対応する英単語は辞書データ中に複数存在することが多い．しかし，どの英単語を用いるのが適切であるかをシステムが判断することは難しい．

語順の問題 一般に，日本語と英語とは語順が異なる．そのため，入力した日本語キーワードの順序と検索対象の英文上の英語キーワードの順序が一致しているとは限らない．

これらの問題を解決する方法の一つとして，考えられるすべての組み合わせを ESCORT に入力する方法が考えられるが，処理時間等の理由により困難である．このため，組み合わせを絞る必要がある．

3 言語横断検索の効率化

本節では，日本語での検索を効率よく行えるための，ESCORT の拡張について述べる．

前節で述べた問題点を解決するために，和英辞書「英辞郎」[3] と ESCORT の英単語に対するインデックスから，日本語単語に対するインデックスを作成した．インデックスは文データベース中の文番号及び文中の単語位置からなる．例えば，ある日本語単語に対し，2 つの訳語が存在した場合，その日本語単語に対するインデックスは 2 つの英単語のいずれかが出現する文番号と単語位置である．

作成されたインデックスを用いた ESCORT の検索は次の 2 つの手順で行う．まず，各日本語単語のインデックスを比較し，各単語に共通して出現する文番号と，それぞれの単語の単語位置を取り出す．次に，得られたすべての文番号に対して，単語位置の小さい順に日本語単語を並び替える．

これらの操作により，入力した日本語単語に対応する英単語がすべて出現している文を選択し，それらの英単語の出現順に絞って検索を行える．そのため，不要な組み合わせの検索を避けることができ，効率よく検索を行うことが可能である．

4 実装と動作例

前節で提案した手法を用いてシステムを実装した．出力では，語順によって分けることにより，分かりやすく分類を提示できるように工夫した．

動作例として,『効率的な手法を提案する。』という表現を含む英文用例検索を行うために，“効率的な 手法 提案する” というクエリで検索を行ったときのシステムの出力画面を Fig.1 に示す．

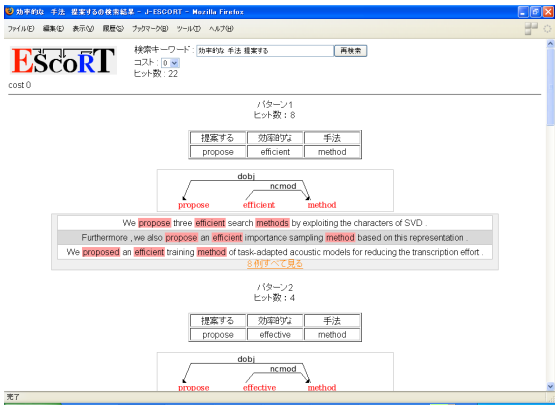

Fig. 1: Result for a query “効率的な 手法 提案する”

出力結果より，“効率的な” という日本語に対して，“efficient”,“effective” の複数の訳語での検索結果が示されている．また、入力された日本語キーワードの語順と，日本語単語に対応する英単語の文中での出現順が異なっていることが分かる．

クエリ “効率的な 手法 提案する” に対して，ESCORT にすべての組み合わせを入力した場合，結果を出力するまでに約 9 分かかったのに対し，本手法を用いることにより約 3 秒で結果を出力することができた．以上より，日本語単語での用例検索が効率よく行えることを確認した．

5 おわりに

本稿では，ESCORT への日本語キーワード入力を可能にする手法を提案した．今後は，入力された日本語単語に対応する英熟語を検索できるよう拡張する予定である．

文献

(1) http://escort.itc.nagoya-u.ac.jp/
(2) 江川他: 言語処理学会第 13 回年次大会，pp.294-297，2007
(3) 英辞郎，アルク，2005