構文付きコーパスを対象とする用例検索システム

7J－2

兵藤安昭　河田実成　応江黔　池田尚志
岐阜大学工学部

1　はじめに

ある単語が現れるか否かだけでなく，単語間の関係すなわち係り受け関係をも指定して用例を検索することができれば，ある言い回しを含んだ文を検索することや，ある意味内容を含んだ文を検索することなど，検索対象をさらに絞り込んだ高度の検索をすることができる．そのためには，単語，品詞などの形態素情報をタグとして付加するだけでは不十分で，構文情報が付与されたコーパスを作成しなければならない．しかし，完全な構文解析には意味情報も必要であり，長文を含む大量のテキストに対して安定的に高い精度で構文解析を行なうことは現在のところ未だ困難である．

そこで我々は，必ずしも常に完全な構文木ではないが，場合によっては部分的に曖昧さを残したままの解析木を，表層的な情報のみを用いて安定的に求める骨格構造解析手法を開発し[1]，これを用いて構文付きコーパスを構築することを試みた．さらに，この構文付きコーパスを対象として，分類語彙表の意味分類を利用した意味コード化をも加え，類似用例検索システムの構築を行った．

2　構文付きコーパス

本実験では，講談社和英辞典とオーム社科学技術和英大辞典の中の英語対訳付き用例文約8万文に対して骨格構造解析を行ない図1のような構文付きコーパスを作成した．

骨格構造データは，原テキストに骨格構造解析を施したものである．データは次のような形式で登録されている．

(@文節 ((自立語 係り先文節) 機能語 文節カテゴリ)

また，類似用例の検索を実現するために，分類語彙表[2]を用いて，文節内の自立語に対して，意味分類を用いた番号付けによってコード化を施した(意味分類コード)．

```
0:原テキスト
   彼の考えに合うように計画を立てなさい
1:骨格構造データ
(@1 ((彼 @2) の 体体) @2 ((考え @3 @5) に 体用)
 @3 ((合う @5) ようだ 用用) @4 ((計画 @5) を 体用)
 @5 ((立てる) なさい 用終))
2:意味分類コード化データ
(@1 (((1100001 1200003) @2) の 体体)
 @2 (((1306103) @3 @5) に 体用)
 @3 (((2112003 2155001 2375001) @5) ようだ 用用)
 @4 (((1308401) @5) を 体用)
 @5 (((2122002 2151301 2154004 …)) なさい 用終))
3:対訳英文
Make your plan on his ideas.
```

図1: 構文付きコーパス (講談社和英辞典)

表1に，この8万文の中から抜き出したそれぞれ100文について，形態素解析，文節カテゴリ付け，骨格構造解析での正解文数を示す．講談社和英辞典については，骨格構造解析結果が正しい解析木を含んでいるものは100文中で97文，科学技術和英大辞典では90文であった．誤りのほとんどは，形態素解析あるいは文節カテゴリ付けの失敗によるものであり，これらが正しく行われば骨格構造解析そのものは，98%以上の精度で正しい解析木を含むものが得られる．

表1: 解析精度

	100文中の正解文数		
	形態素解析	カテゴリ付け	骨格構造解析
講談社	98/100	98/98	97/98
オーム社	95/100	92/95	90/92

3　類似用例検索システム

3.1　システムの概要

システム構成図を図2に示す．入力された検索対象文は，入力文解析部で解析され，その構造がインタフェースのウインドウ上に表示される．この構造表示の上で，さらに検索したい部分構造を特定することができる．検索は2段階に分けて行われる．まず始めに，検索パターン中の自立語ま

Similar Sentence Retrieval System over a Large Copus with Syntactic Structure
Yasuaki Hyodo, Mithunari Kawada, Jiangqian Ying, Takashi Ikeda
Faculty of Engineerring, Gifu University
Gifu-shi,501-11,Japan

たは機能語が出現する文を索引表を用いて検索する (一次検索). 次に, 検索された文を対象として, 検索パターンと構造的に一致するか否かの照合を行う (二次検索).

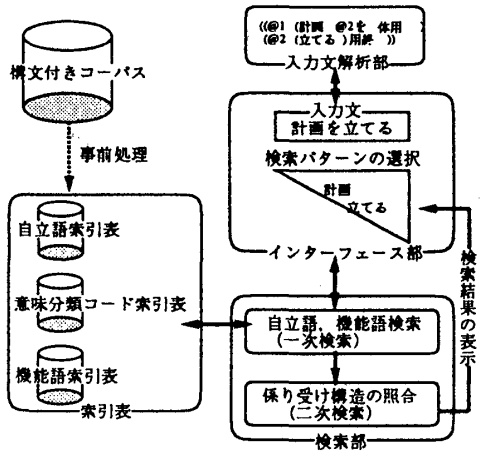


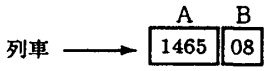
図 2: システム構成図

3.2 検索アルゴリズム

1. 自立語, 機能語による絞り込み (一次検索)

本システムは自立語に対して, (1) 単語による完全一致検索, (2) 意味分類コードによる類似用例検索を可能としている.

類似用例検索の場合の一次検索は, 各単語の意味分類コードを用いて検索を行う. その際, 図 3 のように, 曖昧レベル 1 検索では, 意味分類コード A, B が一致する単語を含む用例を, 曖昧レベル 2 検索では, 意味分類コード A のみが一致する単語を含む用例を検索する. これにより, 意味分類コードに対して単純な前方一致検索を行うだけで, その一致位置により 2 つのレベルの曖昧検索が可能となる.



曖昧レベル 1 検索	曖昧レベル 2 検索
1465 08 電車	1465 03 車
1465 08 機関車	1465 03 バイク
1465 08 トロッコ	1465 09 新幹線

図 3: 分類語彙表による類似検索

2. 構造検索 (二次検索)

二次検索では, 検索パターンと一次検索で抽出された文との間に構造的一致があるか否かを検査する. 例えば, 検索パターンが図 4 のように # に続く番号で表現されているとすると一次検索において, # 1 と @3, # 2 と @4, # 3 と @5 が一致していることがわかる. 従って, ここでは # 1 と # 3 の構造が @3 と @5 の構造と, # 2 と # 3 の構造が @4 と @5 の構造と同じであれば構造が一致したといえる.

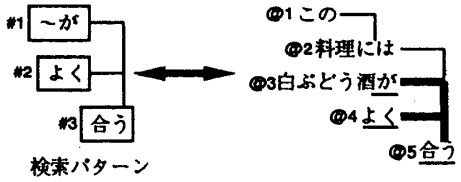


図 4: 係り受け構造の照合

3.3 検索例

以下に「(A) について」と「(B) を」が「行う」に係る用例の検索結果の一部を示す. この例では, 単語検索により, 検索パターンと構造的に一致しない用例が 11 文検索されたが, 構造を指定することにより, これらの用例は棄却される.

1. 単語検索：23 文 (構造が一致しない用例：11 文)
   - この図表にある燃費節減の数字は, 自動変速機と最小のエンジンを積んだ自動車について, 1981 年に EPA が行った燃費節減概算額に基づく
   - この状態での溶接強度を確かめるために, 一部分について簡単な引張り試験も行った

2. 構造検索：12 文
   - この件について十分協議を 行った
   - ACRS は, 原子力安全問題について自主的な再検討を行い, ···

4 おわりに

本論文では, 骨格構造手法による構文情報付きコーパスの構築と, これを対象とした構文指定による類似用例検索システムについて述べた. 今回の実験で用いたコーパスには, 英語対訳が含まれているので, 今後は, 類似用例検索による翻訳支援への応用システムについても検討していく予定である.

科学技術和英大辞典 (電子化版) の使用を認めていただいた (株) オーム社, ならびに, 講談社和英辞典 (電子化版) の使用を認めていただいた電総研自然言語研究室に感謝します.

参考文献

[1] 兵藤, 池田：表層的情報と N 近傍ブロック化手法による日本語長文の骨格構造解析, 情報処理学会論文誌, 36 (9)(掲載予定), 1995.

[2] 国立国語研究所：分類語彙表, 秀英出版, 1964.