

6P-9

意味分類コードを用いた 翻訳支援のための類似例文検索

青山典生 浅井泰博 兵藤安昭 池田尚志

岐阜大学工学部

1. はじめに

日本語文を外国語に翻訳する際、その日本語文に類似した対訳付き用例文(例えば、「意味分類コードを用いた例文検索」に対して、「～を使った…検索は…」の翻訳文)が提示されれば訳文を作るのに重要な参考になる。我々は、そのような翻訳支援のための類似例文検索システムを試作したので報告する。

本システムでは、形態素解析済みの例文データベースを用いる。この例文データベースは、類似文を検索するための方法として、単語を意味分類コードでコード化している。

検索は、類似単語が出現する文を例文データベースから検索してくる1次検索と、1次検索されたものの中から更に細かなパターンの比較を行なう2次検索の2段階で行なっている。

また、ユーザの意図にそぐわない検索結果が出てしまう場合もあるので、ユーザの検索意図を細かくシステムに伝えるためのユーザインタフェースの開発も行なっている。

2. 検索システム

本システムの全体像を図1に示す。

2.1 意味分類コードを用いた例文データベース

より正確な検索のためには構文・意味解析の施されているデータを用いるのが望ましいが、現状では大規模用例データに対して自動的に正確な構文、意味解析を行なうことは困難である。本システムでは、形態素解析を施した大規模用例データベースを構築することとして、実用的なシステムの開発を目的とした。

さらに類似文検索するためのてがかりとして、

全ての単語を意味分類コードで置き換えてある。

自立語の意味分類コードには分類語彙表[1]を用いた。文中の語が多義語である場合、形態素解析では語義が一意に決定できないため、すべての語義でコード化した。また、サ変動詞は全て名詞形でコード化してあり、検索時に動詞形の同類語を考慮することにした。

機能語は派生的なものをグループにまとめコード化してある[2]。例えば、「ている」「てはいる」「ではいる」「てもいる」「でもいる」「ている」「でいる」などは、同じグループになるようにコード化してある。

検索のためのインデックスファイルを、自立語については分類語彙表の項目レベルで作成した。

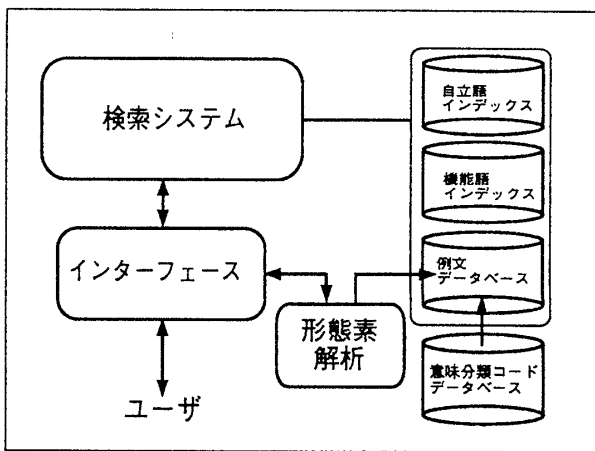


図1 システム全体像

- ・自立語の一般化(同じ品詞):
[輸出 ↔ 貿易, 輸入, 取り引き, ...]
- ・自立語の一般化(別の品詞):
[禁止 ↔ 止める, 禁じる, 取り締まる, ...]
- ・サ変動詞の扱い(名詞+する):
[輸出 ↔ 輸出する]
- ・複合語の扱い:
[例文検索 ↔ 例文 の 検索]
- ・語順の変化:
[学校へ自動車で行く ↔ 自動車で学校へ行く]

図2 1次検索での類似表現例

Example-Text Retrieval for Translation Aid by Using Semantic Classification Code.

Norio Aoyama, Yasuhiro Asai, Yasuaki Hyodo, Takasi Ikeda
Gifu University

1-1 Yanagido, Gifu, 501-11, Japan

機能語については“は”や“が”などの頻出する語を除いて例文検索に有用であると思われる語についてインデックスファイルを作っている。

2.2 1次検索

1次検索では、検索パターンに関係すると考えられる例文をできるだけ広く検索する。そのため、ユーザが入力する文を形態素解析して意味分類コードに置き換え、自立語の項目レベルのインデックスを用いて、検索パターン中に含まれる語と同類の語(分類語彙表中で同一の項目に属する語)が含まれる文を全て検索する。また、サ変動詞とその名詞形、複合語とその構成語の関係、あるいは分類語彙表では名詞 \leftrightarrow 動詞の対応がとれるようなコードづけがされていることなどを考慮して広く類似検索を行なっている(図2)。

ここで、ある一定量を越える大量の例文が検索されてしまう場合には、同類自立語を多く含む文を優先する。

機能語に対する1次検索も可能であるが、機能語の用例は非常に大量に検索されてしまうので、ユーザが特に指示しない限り機能語についての1次検索は行っていない。

2.3 2次検索

1次検索の結果に対して、検索パターンに、より類似した例文を優先順位をつけて出力するために2次検索を行なう。

2次検索は、1次検索の出力結果と検索パターンを比較して、自立語の一致度、語順、機能語の一致度、自立語と機能語の組合せの一致度について、それぞれに付与してある重みを用いて全体の類似性についての評価値を計算する(図3)。重みは既定値を与えてあるが、ユーザが変更することも可能である。

2.4 インターフェース

本システムのインターフェースには Emacsを用いた。検索結果がユーザの意図にそぐわない場合には、ユーザはワイルドカード化などを含む検索パターンの編集、2次検索の重みの変更などを行ない、ユーザの意図をシステムに明確に示した上

例: 検索文 "自動車で学校に行く"

2次検索の対象文:

A. "自転車で大学に出かける"

B. "自動車学校に行く"

であった場合の文の評価値は、

自立語の一致度の比較 : $A < B$ (Bがやや大)

自立語の語順の比較 : $A \approx B$ (同じ)

文節単位の一貫性の比較: $A >> B$ (Aが大)

文節単位の評価値を高く設定してあるため、総合で A.文の評価値は B.文の評価値より高くなる。

図3 2次検索の評価例

で、1次検索あるいは2次検索をやりなおすことができる。

2.5 検索実験

現在「講談社和英辞典」の中の7万用例文を対象に実験中であるが、たとえば検索文「この文の意味を教えてください」に対して、次のような文を検索することが出来る。

この語の意味を教えてください

Please tell me the meaning of this word.

この語句の出典を教えてください

Please give the source of this phrase.

その適切な英語の訳語を教えてください

Will you tell me the proper English for it?

3. おわりに

意味分類コードを用いた翻訳支援のための類似用例検索システムについて述べた。

現在、「講談社和英辞典」の7万用例文を対象にシステムの構築、実験を行なっている。

本システムでは自立語ベースの検索を行なったが、翻訳においては、文の骨格パターンに対する対訳用例が有用である場合が多いので、機能語をベースとした検索を充実することが必要である。

ユーザの検索意図をシステムに伝えるためのインターフェースはさらに検討が必要である。

「講談社和英辞典」磁気テープ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] 国立国語研究所:分類語彙表
- [2] 兵藤,池田:スロット表現による機能語辞書の整理,情報処理学会第45回全国大会,5F-10,1992