テキストデータベースからの同意表現の抽出

2M－1

湯村　武、余田直之、野崎康夫、茂木　健、西田行輝
三洋電機(株)　情報通信システム研究所

１．はじめに

一般的な自然言語処理システムにおいて、利用者（管理者）は、使用環境に応じて辞書をカスタマイズする必要がある。ただ、これからの自然言語処理システムでは、このような辞書データのカスタマイズ、メンテナンス作業の負担をできるだるけ軽減できるようなメカニズムが望まれる。

例えば、知的文書検索や文書要約などの応用を考えた場合、処理テキスト中から同意表現を自動的に認識して、辞書に反映させることができれば、精度の高い処理が期待できる。本稿では、一般的なテキストデータベースから表層的な情報をもとにして、同意表現を抽出する試みについて述べる。

２．目的と概要

知的文書検索などの応用を考えた場合、テキスト中に同意表現が明示されていても、同意語辞書に記述されていないければ同意語による検索はできない。もし、テキスト中に明示されている同意表現を自動的に抽出（認識）できれば、同意表現による検索が可能になるはずである。ここでは、基本的に表層的な情報を利用した以下の２つの枠組を用意して同意表現の抽出を試みた。

(1). 括弧表現の前後からの抽出
(2). 同意表現に特徴的な言い回しの利用

３．括弧表現の前後からの抽出

実際のテキスト中で、括弧表現の前後、

ＡＡＡ（ＢＢＢ）　　・・・・［AAA,BBBは名詞］

で、"ＡＡＡ"と"ＢＢＢ"が同意表現（ここでは略称、別称、言い換えなど）である場合がある［例：ＡＩ（人工知能）］。ここで問題となるのは、"ＡＡＡ"と"ＢＢＢ"が同意関係であるのか、それ以外の関係であるのかという判定が難しい点である。

Automatic Extraction of Homonym Expression from Text Database
Takeshi YUMURA, Naoyuki YODEN, Yasuo NOZAKI, Takeshi MOGI, Yukiteru NISHIDA
Information & Communication System Research Center, SANYO Electric Co.,Ltd.

そこで、さまざまなテキストから、前記表現パターンを収集して、"ＡＡＡ"と"ＢＢＢ"の部分の分析を行ない、次のように整理した。

(a). 別称，略称など
　国民総生産（ＧＮＰ），東京証券取引所（東証）
(b). よみがな
　轍（てつ），彼岸（ひがん）
(c). 日時の言い換えなど
　一九七三（昭和四八）年，当時（一九五〇年）
(d). 補助説明など
　排出削減（フロンの削減等）
(e). その他
　東京（圏）

このような分類をベースに以下の２段階のステップで同意表現候補の絞り込みを行なう枠組を作成した。

(1). 同意表現になりにくい候補の削除
(2). 同意表現になりやすさの評価

３．１　同格表現になりにくい候補の削除

できるだけ表層情報だけで処理しようとしようという立場から、(1)文字種（ひらがなや数字の存在）、(2)文字列の長さ、(3)特徴表現（例：平成～、～年、～社長、当時、～等）など同意表現になりにくい条件をチェックして、同意表現の候補から削除する。

３．２　同格表現になりやすさの評価

３．１の絞り込みによって残されたものの中から以下のような条件を満たしているかどうかによって同格表現になりやすさ（確からしさ）を求める。

【条件１】ＡＡＡとＢＢＢが入れ替わって出現している表現の存在
　例：欧州共同体（ＥＣ）⇔ ＥＣ（欧州共同体）
【条件２】出現頻度（ＡＡＡとＢＢＢが同じ）
【条件３】文字列の構成要素の包含関係
　例：東京証券取引所（東証）
【条件４】文字種による判断
　例：ＭＢＸ（ビーメックス）

このような諸条件を満たしているかどうかでポイントを与え、同意表現になりやすさを数値で示すことにした。

3．3　実験および評価

以上の枠組を使って、新聞の社説[1]を対象に同意表現の抽出実験を行なった。

同意関係候補として得られたデータをポイントの高いものから並べると、表1のような結果（ポイント上位15まで）が得られた。さらに、一定のポイント以上の抽出データについて、同意関係にあるかを調べたところ、

・95%　（249／ポイント上位261）

・91%　（580／ポイント上位637）

の精度で同意表現を抽出することができていることが明らかになった。（ここでは、本来ピックアップすべきデータの漏れについては集計していないが、ポイントが高いほど確からしいということはいえる）

[見出し]	[同義候補]	[ポイント]
欧州共同体	ＥＣ	1752
国民総生産	ＧＮＰ	752
戦略防衛構想	ＳＤＩ	704
朝鮮民主主義人民共和国	北朝鮮	528
関税貿易一般協定	ガット	396
国連平和維持活動	ＰＫＯ	256
主要先進国首脳会議	サミット	184
アパルトヘイト	人種隔離	184
日本電信電話	ＮＴＴ	152
中距離核戦力	ＩＮＦ	150
政府開発援助	ＯＤＡ	144
北大西洋条約機構	ＮＡＴＯ	142
東南アジア諸国連合	ＡＳＥＡＮ	140
新興工業経済地域	ＮＩＥＳ	128
国際通貨基金	ＩＭＦ	108

表1　同意関係候補（上位15）

この実験で、同意関係ではないのにピックアップされた例としては、

・ソビエト（議会）　・技術（ハード）

などがあった。

また、これ以外の問題点として、名詞句のスコープの問題があった。［例：薬物の不正使用（ドーピング）］

4．同意表現に特徴的ないい回しの利用

テキストにおいて同意表現を明示している表現パターンを利用することによって、同意表現を抽出できると考えられる。例えば、次のテキストでは、同意関係（「ＥＣＵ」と「欧州通貨単位」）が明示されており、これを利用しようとする試みである。

『統一通貨として合意されているＥＣＵは「欧州通貨単位」の略称だが、・・・・』

このような表現から、以下のようなステップで同意関係を抽出する枠組を開発した。

(1). 形態素解析

(2). 簡単な構文解析（品詞の曖昧さの解消、名詞連続のまとめ上げなど）

(3). 典型的な言い回しパターンを利用した抽出規則

構文解析および同意表現抽出規則の記述には、独自の文法記述言語の枠組を用いた。図1に同意表現のうちの略称表現を抽出するための規則の一例を示す。

```
Ryakushou1.rr() {
 vi;  @x;
 cnd; [ROOT #1 NP1 #2  (RYAKU  NP2) #3];
      NP1.topic == 'T';
      RYAKU.jlex == "略称"|"略語"|"略";
      NP2.jbkk == 'NO';
 sub; @x = call(RyakushouPrt( [NP1 NP2] );
 act; [ ** ];
}
```

図1　略称の抽出規則例

この規則は、指定された構造とパターン照合を行い、照合に成功したら、その結果を出力するものである。このような抽出規則を用いて新聞の社説[1]を対象に抽出実験を行なった。

［結果］抽出規則数も少ないこともあって、数十程度の同意関係データを抽出したにすぎず、同意関係データの収集という立場からすると効率は悪かった。

ただ、文書処理の中で形態素解析の枠組が用意されていれば、付加的に同意表現を認識しておくことは無駄ではないと思われる。

5．おわりに

これまで述べてきたような表層的な情報を利用するだけでも、ある程度の精度で同意表現を抽出することができることが明らかになった。完全な自動抽出は難しくとも、抽出された候補データを人間がチェックすることで実用化することは可能であると考える。

今後は、さまざまな種類のテキストを用いて有効性を検証していきたいと考えている。

参考文献

[1] 朝日新聞社説（1985~1991年）