テキストからの類義語抽出手法とその評価

○ 福島 俊一
下村 秀樹

An Automatic Synonym Extraction System and Its Evaluation

Toshikazu Fukushima
Hideki Shimomura

This paper describes a system and methods to extract synonyms by analyzing Japanese text. One of the extraction methods is based on word co-occurrence similarity of synonyms. Another method detects explanatory expression patterns of synonyms. This system, which uses these two methods, has the advantage that it can apply to large-scale text corpora, because of its robustness. This paper also reports synonym dictionary construction process using this system. The former method tends to extract synonyms of general terms, and the latter method tends to extract abbreviations of proper nouns.

1. はじめに

情報化社会が進展し、コンピュータの自然言語処理能力に対する期待が高まってきている。コンピュータに曖昧性をもつ自然言語を扱う能力を与えているのは、言葉の意味や関係を記述した辞書・知識ベースである。類義語辞書はその１つであり、近年増大の一途をたどるフルテキストデータの検索[1]などで重要な役割を果たす。例えば、「コンピュータ」というキーワードでフルテキストデータベースを検索しようとしたとき、検索対象テキストのなかでは「コンピューター」や「計算機」と表記されているかもしれないし、「パソコン」や「ワープロ」などの言葉も関連性が強い。類義語（注：本論文では同義語も類義語に含めて扱う）の辞書を用いてキーワードを展開することで適切な検索が可能となる。また、類義語辞書の別な応用として、文書作成時に用語・表記を統一したり、逆に発想・表現をふくらませたりするのに利用することも考えられる。

従来、このような類義語辞書は人手によって言葉を収集・編纂するものであった。これに対して、大量に蓄積されたフルテキストデータから半自動的に類義語データを抽出するシステムが構築できるならば、人手による辞書作成の多大な労力を軽減できる。既存の類義語辞書の拡充や分野別拡張なども系統的に進められる。

このような方向について、単語の共起の類似性に着目したクラスタリングによりシソーラスの構築を目指す研究[2][3]が参考になる。ただし、質のよい小規模なデータでの実験ではなく、大規模なフルテキストデータに適用する場合は、ノイズや計算時間の問題も考えて方式の頑健性を高める必要がある。また、テキストから「国民総生産」と「GNP」のような言い換え表現の類を抽出するパタンマッチング的な手法[4]も参考になる。

本論文では、フルテキストデータを大量処理可能なことを必要条件とし、完全自動化までは望まなくとも、最終的に人手による選別・確認が入る半自動の類義語抽出システム[5]として実用性のあるものを目指した。

2. 類義語抽出システム

図１に筆者らの類義語抽出システムにおける処理の流れを示す。類義語抽出の方式には、共起類似性に基づく抽出方式Ａとパタンマッチングによる方式Ｂの２通りを採用した。いずれの方式でもランク付きの類義語候補リストを出力し、最終的には人手による選別・確認作業を行なう。

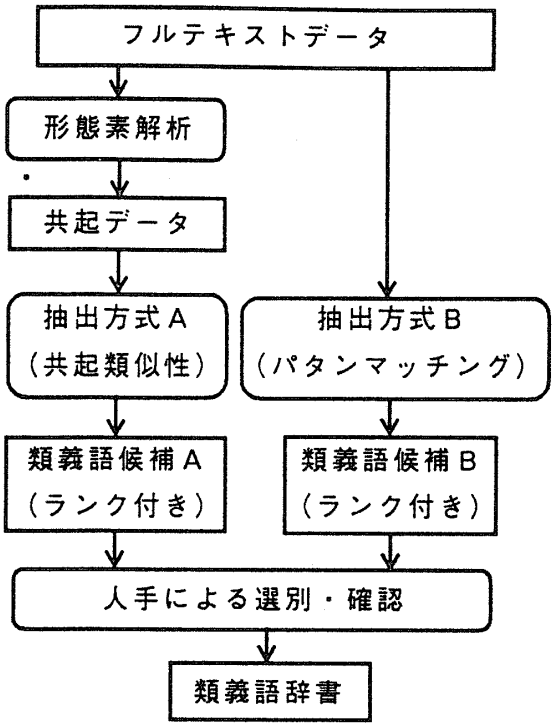

図１　類義語抽出システム

2.1 共起類似性に基づく類義語抽出

共起が類似しているとは、例えば、「製品」および「商品」という語に対して、各々次のような共起データが存在するとき、

製品－開発　製品－価格　製品－販売

商品－開発　商品－価格　商品－販売

「製品」に対する共起語の集合｛開発，価格，販売｝と「商品」に対する共起語の集合｛開発，価格，販売｝とに共通要素が多いことを意味する。このような考えに基いて、単語ｘと単語ｙの間の共起類似度S(x,y)を次のように定義できる。抽出方式Ａは、このS(x,y)によってランク付けした類義語ペアの候補ｘ・ｙを出力する。

$$S(x,y) = \sum_k W_k \cdot \frac{N(R(x,k) \cap R(y,k))}{N(R(x,k) \cup R(y,k))}$$

この式において、W_kは共起の種類ｋごとの重み係数（共起の種類の例は第３章参照）、N(R)は集合Ｒの要素数、R(x,k)は共起の種類ｋで単語ｘと共起する単語の集合を表わす。S(x,y)の定義は自然なものであるが、生のテキストデータを解析して取り出した共起データを用いて計算する際には、いくつかの問題が発生する。第一に、共起の可能性を有/無の２値で記述しているので、ノイズが大きな影響を及ぼす。かといって、テキスト中での共起頻度（多値）を共起類似度の定義に取り入れると（例えばベクトル角で定義）、頻度の極端に高い共起語だけに目を向けることになってうまくいかない。第二に、かなり大量のテキストを用いても共起の網羅はできないから、共起が成立し得ないのか、たまたま共起が出現しなかったのかを区別できない。その結果として、共起語が十分に収集されていない語どうしの共起語集合が偶然に一致して、共起類似度の高いペアになり得る。

そこで、次のような２つの対策によってノイズ除去を行なうことにした。

対策１：共起頻度が１の共起データは類義語抽出に用いない。

対策２：$N(R(x,k) \cup R(y,k)) \leqq 9$の単語ペアｘ・ｙは類義語ペア候補から除去する。

2.2 パタンマッチングによる類義語抽出

抽出方式Ｂでは、「GNP（国民総生産）」のような括弧を用いた説明的表現に着目し、テキストから文字列のパタンマッチングによって類義語ペア候補を抽出する。湯村ら[4]と同様、以下の手順をとった（各手順の詳細は必ずしも湯村ら[4]の通りではない）。

手順１：テキストから括弧内の文字列と直前の文字列のペアを抽出する。
手順２：類義語ペアになりにくい表現パタンを候補から削除する。
手順３：類義語ペア候補をランク付けする。

手順1において、括弧の直前の文字列は、字種（漢字、片仮名、平仮名、英字、数字）の変化点で切り出す。括弧の直前が漢字列・片仮名列・数字列の場合、さらにその直前が漢字列・片仮名列・数字列・英字列ならば、それも含めて切り出す。

手順2では、次のいずれかの条件に該当した類義語ペア候補を削除する。

削除条件１：ペアのどちらかの文字列長が１以下か21以上である。
削除条件２：ペアのどちらかが数字のみか平仮名のみで構成される。
削除条件３：ペアのどちらかが類義語ペアらしくない特徴表現を含む。

削除条件3の特徴表現とは、例えば、氏名を特徴付ける「選手」「容疑者」「敬称略」、時間を表わす「当時」「現地時間」などである。

手順3におけるランク付けは、類義語ペア候補の出現頻度を基本得点とし、英字を含むペア（「GNP（国民総生産）」など）にはさらに加点した。

3. 抽出精度と人手による選別・確認の負荷

前章で述べた類義語抽出方式Ａ・Ｂの出力する類義語ペア候補の品質（精度）を評価し、最終的に人手で行なう選別・確認作業にかかる負荷を測定する。その際、どれくらいまでを類義ととらえるかの判断基準を定めておく必要がある。ここでは、類義語辞書の用途を、第１章で述べたような検索条件キーワードの展開と考えて判断した。

まず、抽出方式Ａについては、田中が公開している共起データ[6]を用いて評価した。共起の種類は、四字漢字の複合語を構成する２語の共起や助詞「の」「が」「に」を介した共起の各々について、前の語からみた場合と後の語からみた場合の２通りずつ、計８通りを考えた（類似度計算式における重みは均等）。約67万件の共起データに現われた約19万種類の体言を対象に実行し、共起類似度が0.06以上となった約５千件のペアについて分析した（共起類似度が０でないペア数は約130万件）。その分析結果を図２に示す。

次に、抽出方式Ｂについては、約１年分の新聞記事を用いて評価した。そのうちの約１カ月分は、削除条件３の特徴表現セットを決める際の訓練データに使った（特徴表現数は160）。評価結果を図３に示す。

図４と図５に抽出方式Ａ・Ｂで得られた類義語ペア（人手による選別・確認済み）の例を示す。抽出方式Ａでは一般語の類義語、抽出方式Ｂでは固有名詞や専門用語の別称・略称が得られやすい傾向がある。

4. おわりに

テキストを解析して類義語候補を自動抽出し、最終的には人手による選別・確認を経て類義語辞書を作成するシステムを構築した。類義語抽出の方式は、従来アプローチ（２通り）を参考に、大量のフルテキストデータを処理できるように頑健性・実用性を重視して設計した。共起類似性に基づく類義語抽出方式Ａの精度は上位候補約1200ペアに対して17％程度、パタンマッチングによる類義語抽出方式Ｂの精度は上位候補約3000ペアに対して54％程度となった。１分につき14ペアをチェックするペースで作業したとき、方式

共起類似度	ペア候補数	類義語ペア確認数	割合	作業時間
0.10 以上	362 件	100 件	27.6 %	約 0.5 時間
0.08 以上（累積）	1245 件	216 件	17.3 %	約 1.5 時間
0.06 以上（累積）	5208 件	503 件	9.7 %	約 6.0 時間

図 2　抽出方式Aの評価結果

テキスト	ポイント 2 以上	類義語ペア確認数	割合	作業時間
1 カ月分（訓練）	854 件	603 件	70.6 %	約 1.0 時間
11 カ月分	3018 件	1620 件	53.7 %	約 3.5 時間

図 3　抽出方式Bの評価結果

ゆとり	余裕	影響	効果	ＡＩ	人工知能
応用	利用	価格	コスト	ＡＴ＆Ｔ	米国電信電話会社
会議	会談	価格	相場	ＩＣ	インタチェンジ
会議	協議	価格	物価	ＩＣ	集積回路
会議	討議	家族	家庭	くつろぎ	リラクセーション
会社	企業	恐れ	危機感	ＪＡＬ	日本航空
商品	製品	恐れ	警戒感	アパルトヘイト	人種隔離政策

図 4　抽出方式Aで得た類義語ペアの例　　図 5　抽出方式Bで得た類義語ペアの例

Aの候補リストからは 1 時間に約 140 件、方式Bの候補リストからは 1 時間に約 460 件の類義語ペアが得られたことになる。抽出精度の値としては高いものではないが、類義語辞書の作成効率を高めるのには役立つと考える。筆者らは、本システムを用いて作成した類義語辞書をフルテキスト検索システム[1]に搭載して利用している。

自動抽出の誤り原因としては、方式Aの場合、前述したようにテキストからは共起語を網羅的に収集できないことが大きい。また、多義語を意味ごとに分離して扱えない点にも問題がある。方式Bの場合は、括弧の直前の文字列の切り出し方法が粗いことや、類義語候補を絞り込むためのヒューリスティックスの改良・整備が課題である。

参考文献

[1] 福島ほか、テキストデータベース検索、 NEC 技報 :47(8)、1994 年。
[2] 白井ほか、実データからの言語知識の自動抽出と活用、情処研報 :NL−51−3、1985 年。
[3] P.Srinivasan, Thesaurus Construction, Information Retrieval (Prentice−Hall), 1992.
[4] 湯村ほか、テキストデータベースからの同意表現の抽出、情処 47 全大 :2M−1、1993 年。
[5] 下村ほか、共起類似性に基づく同義語の抽出、情処 47 全大 :1M−10、1993 年。
[6] 田中ほか、自然言語の解析による知識獲得と拡張、情処研報 :NL−67−4、1988 年。

所属： NEC 情報メディア研究所 (Information Technology Research Laboratories, NEC Corporation)