共起類似性に基づく同義語の抽出

1M－10

下村 秀樹　福島 俊一
(NEC 情報メディア研究所)

1. はじめに

同義語辞書は，データベース検索等でその性能（ユーザインタフェース）を向上させるのに有効である．従来から，係り受けや共起関係の類似性に着目したクラスタリングに基づき同義語が抽出できるとの報告[1,2]がある．これらは，人手で細かく分析された質の良い（したがって小規模にならざるを得ない）データでの試みが中心であるが，規模の大きな辞書を作る際には，容易にかつ大量に集められるデータを前提とした方が望ましい．一方，筆者らもクラスタリングに基づいた同義語抽出の予備実験を行ったが，同義語はクラスタリングの初期段階（1クラスタあたり数語）にも多く見られ，概念体系の構築までは望まず同義語の抽出だけを目標とするならば，計算量の大きいクラスタリングは必須ではないとの知見を得た．

本研究では，大規模な同義語辞書の作成を目標とし，容易にかつ大量に得られるデータから比較的高速な処理で同義語候補を得る方式を検討する．具体的には，単語共起データを元に，クラスタリングのような階層化を行わずに，単語間の共起類似性に基づいた同義語対の抽出を試みた．ただし，現実の問題として同義語を完全に自動抽出することは難しいので，人間による選別作業も併せて検討した．以下，同義語候補抽出方式の概要とともに同義語の抽出実験，人間による選別の結果を報告する．

2. 同義語候補自動抽出

2.1 共起類似性と同義語

同義語候補の自動抽出は，共起類似性に基づいて行う．すなわち，文章中に現れる各語について，特定の文脈関係（例えば「ある助詞を挟んで後に現れる」など）を満たす共起語を集め，それが似ている語の対を同義語候補として抽出する（図1参照）．図1の例ではある文脈の前側の語についての共起類似性を考えたが，同様に後ろ側の語についての共起類似性も考えることができる．

A Synonym Extraction Method Based on Co-Occurrence Similarity
Hideki Shimomura and Toshikazu Fukushima
Information Technology Research Laboratories,
NEC Corporation

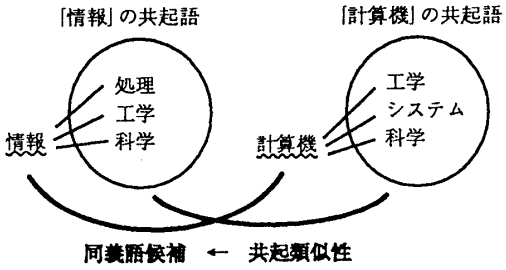

図1　共起類似性と類似度

2.2 類似度

本研究での共起類似性は次式に示す類似度 sim で定義する．

単語 a の共起語の集合： A
単語 b の共起語の集合： B
$sim(a,b) = n(A \cap B) / n(A \cup B)$
（ただし，n(X) は集合 X の要素数）

これは，共起語集合の重なりの度合いを意味しており，大きいものほど共起類似性が高い．図1の例だと，類似度は $2/4 = 0.5$ となる．

3. 同義語候補の自動抽出実験

3.1 実験データ

同義語候補の自動抽出実験のデータには，田中が公開している「語と語の関係」データ[3]を用いた．このデータは格助詞に関する共起の対や4字漢字列（前後2文字ずつで意味を持つ対）のデータを新聞等から大量に収集している．各単語の共起語の集合は，このデータから機械的にかつ容易に得ることができる．

3.2 予備実験

同義語候補の抽出能力を確認するために，4字漢字列のデータの一部について，実際に自動抽出処理を行った．しかし，(1)共起頻度の低い統計的信頼性のないデータがノイズになること，(2)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一致し高い類似度を生むこと，によって，類似度と同義語の相関がほとんど見られなかった．そこで，(1)共起頻度が1の信頼性の低

い共起対は元の共起データから除去する，(2)類似度計算時に共起語集合の和集合の要素数（類似度定義式の分母）が 9 以下のデータを除去する，というノイズ除去の対策を講じた．この結果，類似度上位の 273 対中に意味的に関連が認められる対（同義語以外も含む）が 69 (25.3%) となった．

3.3 類似度計算と複数データからの類似度の併合加算

予備実験の結果をふまえ，4 字漢字列のデータと格助詞（“の” “が” “に”）の共起データから体言の同義語候補を抽出した．まず，元の共起データ（共起対合計約 67 万，体言約 18 万種）に対して，各共起データごとに予備実験と同様のノイズ除去を行って類似度を計算した．次に，さまざまな文脈から見て類似度が高いものは同義語である可能性が高いと考え，各共起データから同じ同義語候補が得られていれば，その類似度を同じ重みで加算し，一つにまとめた．処理したデータの量を図 2 に示す．ノイズ除去によって約 2 万種の体言間の類似度を求めたことになるが，この結果約 130 万対の類似度 0 でない同義語候補が得られ，その類似度上位 345 対のうち 126 対 (36.5%) に意味的関連が見られるようになった．

4. 人手による同義語選別

次に，人間がキーボードを使って同義語を選別するツールを作成し，類似度の高い同義語候補から類義語を選別した．選別の結果得られたものの内訳を図 3 に，同義語の例を図 4 に示す．全体では，5208 対から 178 対の同義語（対義語，同語別表記等の関連の深い語も含めると 505 対）が得られた．作業時間は全体で 6 時間強（平均 15 対／分）であったが，類似度の高い最初の約 1200 対（作業時間 1.5 時間弱）から同義語の半数以上（90 対）が得られた．

5. おわりに

同義語辞書の作成を目的とし，共起語の類似性から同義語候補を得る方式を検討した．実験では，約 18 万（ノイズ除去後は約 2 万）種類の体言間の共起類似性を計算し，その結果を人間が選別して 200 弱の同義語（意味的関連の深い語も合わせると 500 強）の対を 6 時間程度の単純作業で得ることができた．本方式では，語の共起関係が網羅的に現れた大量のデータを用いることにより，さらに多くの同義語候補が得られると思われる．今後はそのデータの収集とともに，ノイズ除去方法や類似度定義の改良を検討することが課題となる．

参考文献

[1]白井他：実データからの言語知識の自動抽出と活用，情報処理学会 NL 研究会，51-3 (1985)
[2]藤原：共起パタン分布に基づく単語間類似度法を用いた動詞・名詞のクラスタリング法，人工知能学会第 2 回全国大会，8-3 (1988)
[3]田中他：自然言語の解析による知識獲得と拡張 －四字漢字列をもちいて－，情報処理学会 NL 研究会，67-4 (1988)

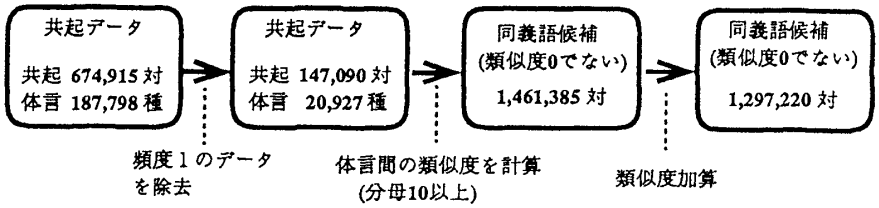

図 2 同義語候補自動抽出での処理データ数

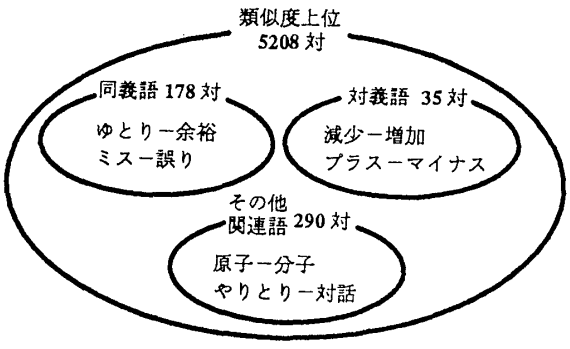

図 3 人手による同義語選別作業の結果

ゆとり	余裕	苦情	要請
意見	主張	苦情	要望
意向	意思	経費	費用
運動	活動	研削	切削
応用	利用	見解	考え
家	自宅	原因	理由
会議	会談	現況	現状
会議	協議	現状	実情
会議	討議	使用	利用
会社	企業	試験	実験
解決	決着	質疑	質問
議論	論議	状況	状態
競争	争い	進展	前進
協議	論議	人々	人たち
協議	話し合い	生産	製造
金	資金	節減	節約

図 4 得られた同義語の例