大規模コーパスと国語辞典の統合的利用による シソーラスの自動構築

大福 泰樹　　河原 大輔　　黒橋 禎夫
東京大学工学部　東京大学大学院情報理工学系研究科
{hiroki, kawahara, kuro}@kc.t.u-tokyo.ac.jp

1 はじめに

語の同義・類義関係や上位下位関係をまとめたシソーラスは言語処理の様々な場面で用いられる重要なリソースである．

これまでに分類語彙表，EDR 概念辞書，NTT 日本語語彙大系，WordNet など，様々なシソーラスが人手によって構築されてきた．これらは膨大なコストと，人間が意味について徹底的に考えて大規模なものを記述する際に常に生じる微妙なバランスの上になりたっている．そのため，部分的に不満足な部分があっても修正することは容易でなく，言語の経年変化への対応や新たな専門分野シソーラスの構築などを柔軟に行うことも難しい．

これに対して，シソーラスを自動で構築する試みも様々に行われており，それらは辞書を用いるものと，大規模テキストを用いるものに大別できる．

国語辞典を用いれば各見出し語の定義文からその上位語を取り出すことができる [5, 4]．しかし，上位語が一気に「こと」「もの」のような非常に抽象的な語に飛んでしまう場合や，意味を捉える，すなわち上位語とその修飾という形で語の定義を与える際に様々な視点があるために，必ずしも所望のまとまりが得られないという問題がある．たとえば，単純には「エンジン」「資本主義」「宅配便」の上位語がすべて「しくみ」となってしまう．また，当然のことながら未知語には対応できない．

一方，大規模コーパスにおける共起情報をもとに類義語を取り出す研究も行われている [2]．この場合には，共起分布が非常に近いものは確かに類義語である場合が多いが，次のレベルには類義語的なものとそうでないものがまじってしまい，いわば，ぼんやりとした結果しかえることができない．また，“A such as B and C” のようなパターンでコーパスから上位下位関係を取り出す手法もあるが [1]，基本的な語彙についてはこのようなパターンでは収集が困難であるという問題がある．

シソーラスの自動構築はこのような状況であり，国語辞典を用いる方法も大規模コーパスを用いる方法も，どちらも実用的に使えるもの，現在の人手のシソーラスの代替となるものを作り出すには至っていない．

本研究では，国語辞典のシャープであるがぶれの大きい情報と，コーパスの漠然とした情報を統合することで，上で述べたような問題点を解消し，実用レベルのシソーラスを自動構築する方法を提案する．まず国語辞典から確実な上位下位関係のまとまり，すなわちシソーラスの核のようなものを抽出し，次にコーパスを用いてそれらの核に属する未知語を獲得したり，核を統合することを行う．全体の枠組みを図 1 に示す．本研究では対象を名詞に限定する．

なお，現段階では適当な粒度の類義語のクラスターを求めることを目標としており，人手シソーラスのような深い階層構造を求めることは考えていない．

2 国語辞典からの上位下位関係の抽出

国語辞典では，定義文末尾の語が見出し語の上位語である場合が多いが，他にも上位語を示すさまざまな表現があり，また，上位語ではなく，同義語，下位語が示される場合もある．そこで，次の例に示すようなパターンとのマッチングによって，各語義の第一文から見出し語との関連語を取り出す．

同義語

〜のこと／略。
〜のくだけた／乱暴な／丁寧な言い方。

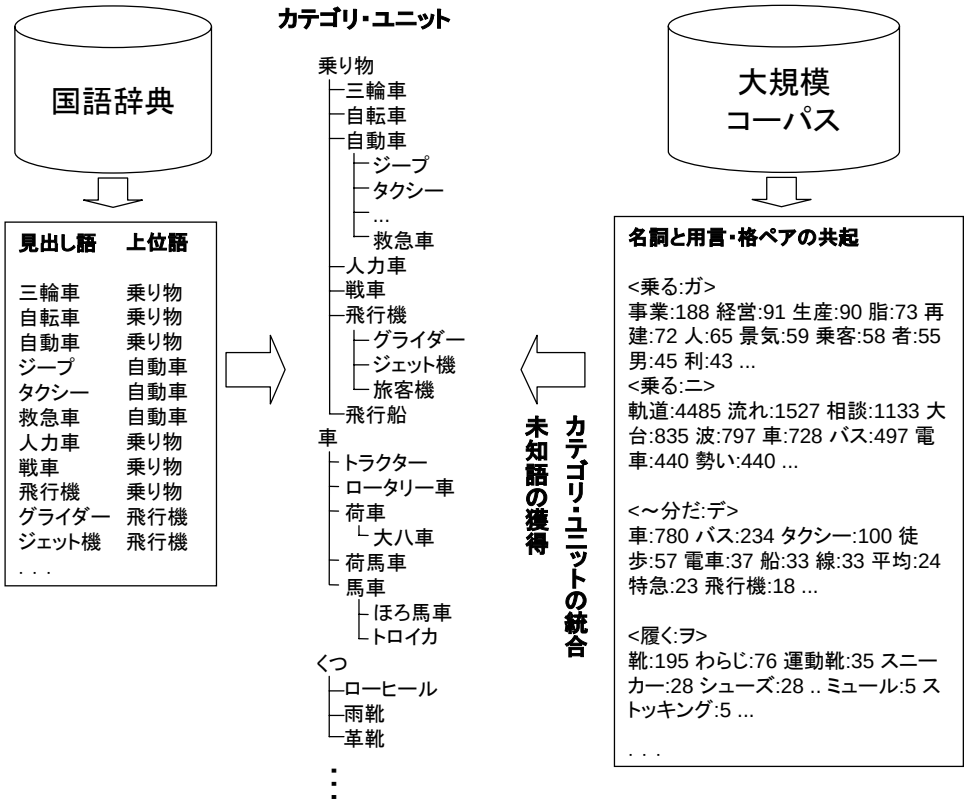

図 1: シソーラスの自動構築の概要

下位語

〜など。
〜などをまとめていうことば。

上位語

〜の一種。
〜の一つ。
〜。(他のどのパターンともマッチしない場合は末尾の語を上位語とする)

このようにして取り出した関連語のうち，見出し語と上位語との関係をまとめることによってシソーラスを作ることができるが*，そのような素朴な方法には次のような問題がある．

1. 「こと」「もの」など，非常に抽象的な語が上位語となり，その下に多様な語が集まってしまう．
2. 「しくみ」「集まり」など，ある視点からは確かに上位語であるが，シソーラスとしてはその視点ではまとめたくないというものができてしまう．
3. 多義性の問題．たとえば「車」は複数の語義があり,「輪」と「乗り物」という二つの上位語がえられる．一方,「車」は多くの語の上位語となっており，この関係を単純にまとめてしまうと「輪→車→トラクター ロータリー車 荷車...」と「乗り物→車→トラクター ロータリー車 荷車...」というまとまりができてしまう．

そこで，国語辞典からのシソーラスとしては，次のような上位下位関係を削除し，安定で確実な部分だけを利用することとした．

- 「こと」「もの」や,「しくみ」「集まり」など，特定の語 (約 40 語) との上位下位関係は利用しない．
- 定義が複数ある見出し語（多義語）は，それと上位語との関係は利用せず，下位語との関係のみ利用する．たとえば「車」は「輪」にも「乗り物」にもつなげず,「車→トラクター ロータリー車 荷車...」の関係だけを取り出す．

この結果，比較的小規模な上位下位関係の木が多数えられる (図 1 参照)．以下では，この各木をカテゴリ・

*現在のところ見出し語と上位語との関係だけを扱っており，同義語や下位語との関係，さらに各語義の第二文以降は今後扱っていく予定である．

表 1: 名詞と用言・格ペアの共起

	⟨ 乗る:ニ ⟩	⟨ 待つ:ヲ ⟩	⟨〜分だ:デ ⟩ ···	［計］
車	728, 4.8	11, 0.0	780, 8.4 ···	39395
バス	497, 6.1	111, 4.9	234, 8.6 ···	10401
···				
［計］	20724	11440	1847 ···	31817706

ユニットとよぶこととし，本文中では『乗り物 (三輪車 自転車 自動車 ...)』のように二重かぎ括弧でかこみ，最上位語を先頭，その直接・間接の下位語を丸括弧でかこんで表示する．

3 共起情報に基づく処理

3.1 コーパスからの共起情報の収集

名詞を特徴付ける共起関係として用言との係り受け関係を考える．大規模コーパスを構文解析し，その結果から係り受け関係が確実なものだけを収集する．その方法は我々が格フレームの自動構築で用いた方法と同じものである [3]†.

この結果，用言と格のペア (用言・格ペア) ごとに次のような共起データが収集される．

＜乗る:ガ＞ 事業:188 経営:91 生産:90 脂:73 再建:72 人:65 景気:59 乗客:58 者:55 男:45 利:43 ...

＜乗る:ニ＞ 軌道:4485 流れ:1527 相談:1133 大台:835 波:797 車:728 バス:497 電車:440 勢い:440 ...

＜〜分だ:デ＞ 車:780 バス:234 タクシー:100 徒歩:57 電車:37 船:33 線:33 平均:24 特急:23 飛行機:18 ...

3.2 共起分布に基づく類似度の計算方法

用言・格ペアに対する名詞 (または名詞集合) の共起分布の類似度を計算する方法として，Hindle の方法を用いる [2].

たとえば,「車」と「バス」の用言・格ペアに対する共起が表 1 のようにえられたとする．ただし，表 1 の最下行は用言・格ペアの頻度，最右列は名詞の頻度，最下行最右列は共起データの総数 (次式では N と表す)，他の各欄の値は，用言・格ペアに対する名詞の頻度と，次式で計算される名詞と用言・格ペアの相互情報量である．

$$\log_2 \frac{\text{用言・格ペアと名詞の共起頻度}/N}{(\text{用言・格ペアの頻度}/N)\cdot(\text{名詞の頻度}/N)}$$

このとき，用言・格ペアごとに 2 つの名詞の小さい方の相互情報量の和をとったものを 2 つの名詞の類似度とする．表 1 の例の場合には $4.8 + 0.0 + 8.4 + \cdots$ となる．

3.3 未知語の獲得

前節で定義した類似度に基づいて，各カテゴリ・ユニットに対して，それに属する未知語 (国語辞典の見出し語でない語) の獲得を行う．

この場合，分布を比較する一方は，名詞ではなく，カテゴリ・ユニットに含まれる名詞の集合とし，名詞集合の頻度を考える．たとえばカテゴリ・ユニット『乗り物 (三輪車 自転車 自動車 ...)』の場合は,「乗り物」「三輪車」「自転車」などの頻度の和を用いる．

カテゴリ・ユニットと分布が似ている可能性がある語は，カテゴリ・ユニット中の語と同じ用言・格ペアに出現する語に限られるので，それらの語をすべて取り出して比較を行う．すなわち,『乗り物』の未知語を調べる場合には，＜乗る:ニ＞，＜待つ:ヲ＞などに含まれる「ヘリ」「トレンド」「ドナー」などの語を調べる．そして，類似度が閾値以上のものがあれば，そのカテゴリ・ユニットに属する語と判断する．

なお，閾値を考える際に，カテゴリ・ユニット（の名詞集合）ごとの頻度の違いを正規化するために，類似度は，カテゴリ・ユニットと用言・格ペアの相互情報量の和で正規化することとする．

3.4 カテゴリ・ユニットの統合

用言・格ペアに対する共起分布の比較は名詞集合と名詞集合の間でも行うことができるので，まったく同じ計算方法によってカテゴリ・ユニット間の類似度を計算し，類似度が閾値以上のカテゴリ・ユニットを統合する

4 実験と考察

実験は，国語辞典として例解小学国語辞典 [6]，コーパスとして新聞記事 25 年分 (毎日新聞と日本経済新

†ただし，格フレーム構築では既存のシソーラスを用いてクラスタリングを行い，各用言の意味・格パターンの区別を行ったが，今回はシソーラスの学習を目的としているので，既存のシソーラスに依存するクラスタリングは行わず，用言ごとのまとまりとした．

聞) を用いた．

まず，国語辞典からの上位語・下位語の抽出と整理により，名詞約 8800 語に対して約 1600 のカテゴリ・ユニットがえられる．ただし，このうち約半数は 2 語，すなわち一つの上位下位関係のみからなるカテゴリ・ユニットであったので，以降の処理ではそれらは対象外とし，3 語以上からなる約 800 のカテゴリ・ユニットのみを処理対象とした．

次に，各カテゴリ・ユニットについて未知語の獲得を行った．この結果について，正規化された類似度が 0.30 以上の語をランダムに 300 個，人手で評価したところ，81.3%が妥当な語 (そのカテゴリ・ユニットに属すると考えてよい語) であった．具体例としては次のようなものが抽出された．

『伝記 (自叙伝 自伝 立志伝)』
エッセー:0.46 回顧録:0.44 評伝:0.43 回想録:0.38

『くつ (ローヒール 雨靴 革靴)』
ブーツ:0.43 ミュール:0.37 ゾウリ:0.37 ヒール:0.36 履物:0.35 ゲタ:0.32 ...

『もめごと (小競り合い 内輪もめ)』
内紛:0.76 係争:0.64 不祥事:0.61 ゴタゴタ:0.60 崩落:0.56 市街戦:0.56 縄張り争い:0.55 あつれき:0.54 ...

最後に，カテゴリ・ユニットの統合結果について，類似度が 500 以上の 246 ペアを調査したところ，91.5%について妥当であると判断された．具体的には，類似度の高いものから順に次のようなペアがえられた．

1994.11 『乗り物 (三輪車 自転車 自動車 ...)』↔『車 (トラクター 荷車 荷馬車 ...)』

1496.91 『文字 (くさび形文字 エジプト文字 ローマ字 ...)』↔『図形 (五角形 四辺形 台形 ...)』

1324.27 『気分 (住み心地 寝心地 雰囲気 夢心地) 』↔『感じ (ぬくもり スリル ニュアンス 印象 ...)』

1322.30 『乗り物 (三輪車 自転車 自動車 ...)』↔『バス (スクールバス マイクロバス ワンマンカー)』

1232.73 『町 (ニュータウン 下町 近郷近在 ...)』↔『市 (政令指定都市 朝市 年の市)

1189.99 『乗り物 (三輪車 自転車 自動車 ...)』↔『車両 (ディーゼルカー 貨車 列車 汽車 ...)』

『車』は 2 節で議論した例で,「輪」と「乗り物」の多義性があるため独立のカテゴリ・ユニットとなるが，共起情報によって『乗り物』と統合された．4 番目の『バス』(「自動車」と「風呂」と「声」) も同様の例である．

このような多義語の扱いは,「多義語であっても，それが最上位語になってカテゴリ・ユニットを形成する場合には一つの意味で用いられる (ことが多い)」という仮定に基づいている．これらの例のように多くの場合はそれが成り立つが，5 番目の『市』の例はそれが成り立たない場合で,「政令指定都市」と「朝市」では「市」の意味が異なっており，これらを一つのカテゴリ・ユニットにすること，さらにそれを『町』と統合することは適切ではない．しかし，そのような例は少数であった．

5 おわりに

まだ基本的な検討と実験，確認ができた段階であるが，このように国語辞典と大規模コーパスを統合的に利用することにより，人手のシソーラスに置き換え可能なクオリティのものが自動構築できる見通しをえた．

今後，国語辞典からの関連語の取り出しをより精密にした上で，これまでに人手のシソーラスを利用していた並列構造解析，格解析，省略解析などで自動構築シソーラスの有効性を検証していく予定である．

参考文献

[1] Marti Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th COLING*, pp. 539–545, 1992.

[2] Donald Hindle. Noun classification from predicate argument structures. In *Proceedings of the 28th ACL*, pp. 268–275, 1990.

[3] Daisuke Kawahara and Sadao Kurohashi. Fertilization of case frame dictionary for robust Japanese case analysis. In *Proceedings of the 19th COLING*, pp. 425–431, 2002.

[4] Jun-ichi Nakamura and Makoto Nagao. Extraction of semantic information from an ordinary English dictionary and its evaluation. In *Proceedings of the 12th COLING*, pp. 459–464, 1988.

[5] Hiroaki Tsurumaru, Toru Hitaka, and Sho Yoshida. An attempt to automatic thesaurus construction from an ordinary Japanese language dictionary. In *Proceedings of the 11th COLING*, pp. 445–447, 1986.

[6] 田近洵一（編）. 例解小学国語辞典. 三省堂, 1997.