特許データベースからのシソーラスの自動構築

難波英嗣[1]，奥村学[2]，新森昭宏[3]，谷川英和[4]，鈴木泰山[5]
1. 広島市立大学　情報科学部
2. 東京工業大学　精密工学研究所
3. インテック・ウェブ・アンド・ゲノム・インフォマティクス
4. IRD 国際特許事務所
5. ピコラボ

1. はじめに

本研究では，特許データベースからシソーラスを自動的に構築する手法を提案する．シソーラスは，文献を検索したり，特許や論文等の専門文書を執筆したりする上で有用な情報源として活用されている．例えば，JST が提供する文献検索サービス JDReamII[1]では，ユーザが検索を行う際の支援機能のひとつとして，シソーラスが利用できる．また，シソーラスは，情報検索や機械翻訳など計算機で言語処理を行う際の知識源としてもしばしば利用されている．しかし，シソーラスを人手で構築し，更新することは非常にコストがかかるため，テキストデータベースから，シソーラスを自動的に構築するという研究が近年活発に行われるようになってきている．

テキストデータベースからシソーラスを構築する代表的な手法は，「A や B などの C」や “A such as B, C” などの定型表現に着目して，用語の上位，下位概念を自動的に抽出するというものである [Hearst 1992, 安藤 2003, 相澤 2006]．また，この他にも HTML の構造を利用した抽出方法[新里 2005]や，用語の定義文を利用した方法[大石 2006]なども提案されている．本研究では，定型表現に基づいて上位，下位概念を獲得する手法に着目する．

定型表現に基づいた従来手法では，上位，下位関係ではない用語対を誤って抽出してしまうという問題があった．また，この他にもこの手法には様々な問題や改善すべき点が存在すると考えられる．そこで，本研究では，まず，定型表現を用いて上位，下位概念を獲得し，それを分析することで，定型表現を用いた手法にどのような問題があるのか整理する．次に，これらの問題のうちのいくつかについて，改善手法を提案する．

本論文の構成は以下のとおりである．次節では，従来の定型表現を用いた手法で上位，下位概念を獲得し，その問題点を整理する．3 節では，2 節で指摘した問題点のうちのいくつかを改善する手法を提案する．4 節では，提案手法の有効性を調べるために行った実験について述べる．5 節で本稿をまとめる．

[1] http://pr.jst.go.jp/jdream2/

2. 定型表現を利用した上位，下位概念の自動獲得

安藤ら[安藤 2003]は，上位，下位概念を獲得するために複数の定型表現を用いている．本研究でも「などの」「等の」「といった」「のような」の 4 種類の定型表現に着目し，特許公開公報(1993～2002 年)から，これらの表現を含む文を収集した．その結果，「などの」「等の」を含む文が 29,641,887 文，「といった」を含む文が 844,790 文，「のような」を含む文が 9,725,720 文収集された．実際に収集された文を見ると，「のような」と「といった」を含む文にはノイズが多く含まれていることがわかった．また，抽出された文数も「などの」と「等の」を含む文数と比べると件数は少なかったため，「などの」と「等の」の 2 つだけで十分な量の上位，下位概念が獲得できると判断した．表 1 に，実際に獲得した上位，下位概念の概要を示す．

表 1　獲得された上位，下位概念

上位，下位関係(異なり数)	7,031,159
全用語数(異なり数)	1,825,518
1 語以上下位語を持つ用語数	833,215
1 語以上上位語を持つ用語数	1,236,663

獲得された上位，下位概念を分析し，定型表現を利用した手法には，少なくとも以下に示す 5 つの問題点があることが分かった．

- [問題点 1] 上位，下位関係にない
  上位，下位関係にないものが誤って抽出される．例えば，「パソコンなどのキーボード」という表現から「パソコン」の上位概念として「キーボード」が誤って抽出されてしまう．
- [問題点 2] 抽出個所の誤り
  「英語や日本語などの複数の言語」から，「英語」と「日本語」の上位語として「複数」が抽出される．「複数」以外にも「様々」，「色々」などの表現が抽出されることもある．さらに「英語や日本語などの複数言語」のように，「複数」という語が上位語「言語」と結びついて複合名詞になる場合もある．
- [問題点 3] 直接上位，下位関係にない
  上位，下位関係として誤りと断定できないが，直接上位，下位関係にあるとは言えないもの．

例えば,「ワードプロセッサ」の直接の上位語は「文書編集装置」 や「文書作成支援装置」が妥当であると考えられるが,「情報処理装置」などの用語も抽出されるケースである.

- [問題点 4] 多義語
  多義語の区別ができない. 例えば,「カッター」には「切断手段」と「衣類」の 2 つの語義が存在する.
- [問題点 5] 同義語
  特許用語は，学術用語と異なり，ほぼ同じ意味で別の表現の用語が数多く存在する．例えば「フロッピーディスク」に対して「磁気記録装置」と「磁気記憶装置」など.

本稿では，これらの問題点のうち，問題点 1，3，5 に取り組む. 次節では，3 つの問題点の改善方法について述べる.

3. 定型表現を利用した概念獲得手法の改善

3.1. [問題点 1：上位，下位関係の判別]

「パソコンなどのキーボード」のように「などの」の前後の用語が上位，下位関係にないものの検出方法について述べる.「パソコンなどのキーボード」と同種の間違いは，相澤[相澤 2006]も指摘しており，新聞記事から抽出された誤り例として「台湾や日本などのアーチスト」や「フランスや英国などのクラブ」などを挙げている.

これらの表現が上位，下位関係であるかどうかは,「などの」の個所を「の」に置き換えた表現が日本語として自然であるかどうかで判断できる. 例えば,「パソコンなどのキーボード」の「などの」を「の」に置き換えると「パソコンのキーボード」となる. この表現が日本語として自然である場合,「パソコン」と「キーボード」は上位，下位関係にない．相澤が誤りとして挙げている上の 2 例についても「台湾のアーチスト」や「フランスのクラブ」という表現は日本語として自然であるため,「台湾」と「アーチスト」や「フランス」と「クラブ」は上位，下位関係にないと判断できる．一方「機械翻訳などの自然言語処理」の場合,「などの」を「の」に置き換えると「機械翻訳の自然言語処理」という不自然な日本語となるため,「機械翻訳」と「自然言語処理」は上位，下位関係にあると判断される.

「などの」を「の」に置き換えた表現が自然であるかどうかの判断は，実際にデータベース中にそのような表現が存在するかどうかで調べることができる．ただし，調べるときには次の点を注意する必要がある．例えば「アルミニウムなどの金属」の場合,「などの」を「の」に置き換えると「アルミニウムの金属」という文字列になる．この表現がデータベース中にあるかどうかを grep コマンドなどで調べると,「アルミニウムの金属片」という表現の一部とマッチしてしまい,「アルミニウム」と「金属」は上位，下位関係にないと判定されてしまう．そこで,「の」の前後の名詞句（「アルミニウム」と「金属片」）が完全に一致するかどうかで判断する必要がある.

3.2. [問題点 3：上位下位関係の優先度比較]

藤原ら[藤原 1996]は，二つの異なる概念体系を統合する際，一方の概念体系に

「通信」＞「多重通信」

という関係があり(“＞”は上位下位関係を示す),他方に

「通信」＞「通信技術」＞「通信伝送方式」＞「多重通信」

という関係がある場合,「通信」＞「多重通信」は冗長と考え，削除するという方法を提案している. この方法を用いると，ある特許から

「情報処理装置」＞「ワードプロセッサ」

という関係が抽出されても，別の特許から

「情報処理装置」＞「文書編集装置」＞「ワードプロセッサ」

という関係が抽出されていれば,「情報処理装置」＞「ワードプロセッサ」のように直接上位，下位関係にないものを削除することができる. しかし，実際にこの手法を用いたところ，直接上位，下位関係にあると考えて差し支えない関係まで削除されるケースが少なからずあった．そこで，本研究では，直接上位，下位関係にないものを削除するのではなく，表示する順序を変える(優先度を下げる)ことで問題点 3 の改良を試みる.

例えば，ある用語 X について,「X などの Y」という表現がデータベース中に m 回出現しているとする．Y の下位語が n 個存在する場合，X に対する上位語 Y のスコアを「$m/\sqrt{n}$」で計算する．スコアは，m の値が大きく，かつ用語 Y の下位語があまりない時に大きくなる．たとえ m の値が大きくても,「情報処理装置」のように下位語が多い用語は$\sqrt{n}$ が大きくなるため,「$m/\sqrt{n}$」の値は小さくなる.

3.3. [問題点 5：同義語の自動抽出]

本節では，例えば,「文書編集装置」と「文書作成装置」のような同義語対を自動的に検出する手法について述べる. これらの用語は，どちらも「ワードプロセッサ」の上位語であるが，同一文書内に共出現することはほとんどないため，共起を使った手法などで抽出するのは困難である. そこで，本研究は，用語の上位，下位関係を使って自動抽出する手法を提案する．図 1 は,「文書編集装置」と「文書作成装置」という 2 つの用語を中心に，これらと上位，下位関係にある用語の一部を示したものである. 図のように,「文書編集装置」と「文

書作成装置」が似たような意味を持っているのであれば，数多くの上位語あるいは下位語を共通に持つと考えられる．

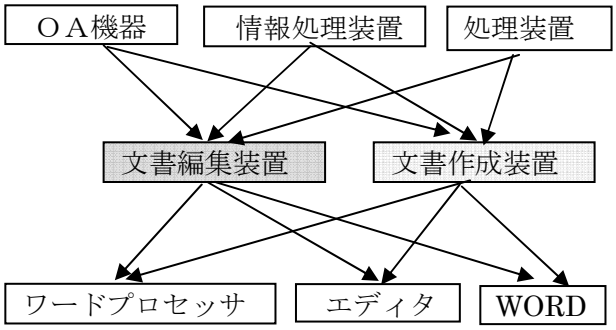

図 1 用語間の上位，下位関係を用いた同義語の検出

上で述べたアイディアは，引用分析研究における書誌結合[Kessler 1963]と共引用分析[Small 1973]に基づいたものである．引用分析とは，論文間の引用・被引用関係を用いて，論文間の関係を分析する方法である．書誌結合は，論文間の関連度を測る時に，2 論文間でどれだけ同じ論文を引用しているか，という基準に基づいている．一方，共引用分析は，2 論文がどれだけ他の論文で共に引用されているか，という基準に基づいた手法である．本提案手法は，用語間の上位，下位関係を論文間の引用関係と見なし，引用分析手法を用いて，同義語の抽出を行うものである．

4. 実験

3 節で述べた手法の有効性を調べるために実験を行った．

4.1. [問題点 1：上位，下位関係の判別]の改良結果

3.1 節で述べた手法を用い，表 1 に示す上位，下位概念の改善を試みた．その結果，上位，下位関係にある用語が誤って削除されるケースが非常に多いということがわかった．例えば，「アルミニウムなどの金属」から抽出された「金属＞アルミニウム」という関係が正しいかどうか判定するために，「アルミニウムの金属」という表現が特許データベース中に含まれるか調べたところ，次のような表現が見つかった．

「アルミニウムの金属の円柱」
「アルミニウムの金属の板」

「アルミニウムの金属」では不自然でも，後ろに「の円柱」や「の板」が続くことで不自然でなくなる場合がある．このようなケースは，事前に想定していなかったが，文節間の係り受け関係を考慮することでこの問題に対処することが可能になると思われる．

しかし，この他にも以下のような例が存在した．

「例えばアルミニウムの金属やプラスチック等の合成樹脂でもかまわない。」
「前記芯材 30 は、アルミニウムの金属からなり、」

これらの表現は，日本語として不自然と感じられるが，実際にデータベース中に存在する以上，アルミニウムと金属は上位，下位関係にないと判断されてしまう．ただ，このような例は，数としては非常に少ないため，頻度が非常に低い事例は存在しないと考えることで，不自然な日本語に対処できる可能性はある．

4.2. [問題点 3：上位下位関係の優先度比較]の改良結果

藤原らの手法で冗長な関係を削除した結果，上位，下位関係の異なり数が 5,162,513 となり，削除前と比べ約 30%の関係が削除された．表 2 に，削除前，削除後，提案手法の 3 通りの手法で「ワードプロセッサ」の上位語の抽出結果を示す．

表 2　3 通りの手法による「ワードプロセッサ」の上位語の抽出結果

順位	削除前	削除後	提案手法
1	情報処理装置	プリンタ機構	プリンタ機構
2	情報処理機器	文書データ処理装置	文書作成装置
3	ＯＡ機器	液晶ディスプレー	情報処理機器
4	電子機器	編集入力機	文書処理装置
5	文書処理装置	文書情報処理装置	情報処理装置

表 2 において，削除前の手法では，上位に「情報処理装置」や「情報処理機器」など，ワードプロセッサとは直接上位，下位関係にない用語が抽出されているのが分かる．一方，削除後の手法では，これらの語が削除され，ワードプロセッサの直接の上位語である「文書データ処理装置」が 2 位にランクされているが，削除前の 5 位の用語「文書処理装置」も誤って削除されている．提案手法は，「情報処理機器」や「情報処理装置」といった多くの下位語を持つ用語の順位が下がり，逆に「文書作成装置」や「文書処理装置」といった用語の順位が上がっているのが分かる．以上の結果から，少なくとも「ワードプロセッサ」に関しては，提案手法が有効であることが確認できる．

4.3. [問題点 5：同義語の自動抽出]の改良結果

3.3 節で述べた書誌結合および共引用分析を用いて同義語の抽出を試みたところ，どちらの手法でも同義語が抽出できていたものの，「アルミニウムと鉄」や「赤と緑」など，共通の上位概念または下位概念をもつ兄弟関係にある用語対が，同義語対と共に数多く抽出された．そこで，引用分析

の結果から兄弟関係にある用語対の除去を試みた．本研究では「A や B などの C」という定型表現に着目して上位，下位概念を抽出しているが，この表現において，A と B は兄弟関係にあると考えられる．そこで，定型表現から兄弟関係にある用語対を抽出しておき，引用分析の結果から，兄弟関係にある用語対を除去すれば，効率的に同義語の抽出ができると考えられる．

実際に，上記の定型表現から兄弟関係にある用語対を抽出したところ，5,046,426 個の用語対が得られた．この用語対を用い，引用分析結果の中から，兄弟関係にない用語のみを抽出した結果の一部(上位 15 件)を表 3 に示す．

表 3　同義語の抽出結果

順位	抽出された用語対	
1	ヒータ	ヒーター
2	医薬	医薬品
3	コントロールユニット	制御ユニット
4	伝導材料	伝導体
5	合穴	固定部
6	昇圧回路	高圧回路
7	不凍液	ブライン
8	デジタル表示	グラフ表示
9	電圧検出回路	波形整形回路
10	宝石類	宝飾品
11	樹脂液	感光液
12	偏光素子	複屈折板
13	ハロゲン化銀	銀塩
14	セラミックス	セラミック
15	活性溶媒中	溶媒中

この図からも分かるとおり，結果の上位には高い割合で同義語対が含まれている．この結果を用いてシソーラス中の同義語のノードを結合し，再度引用分析手法を適用することで，新しい同義語対を検出できる可能性がある．

5. おわりに

本研究では，文書集合から定型表現を用いて上位，下位概念を獲得する従来手法の問題点を 5 点指摘し，そのうちの 3 点について改善する方法を提案した．また，特許公開公報(1993～2002 年)から獲得した上位，下位概念について，提案手法がある程度従来手法の問題点を改善できることが分かった．

謝辞

本研究で用いた米国特許データは，国立情報学研究所の許可を得て，NTCIR テストコレクションを利用させていただいた．本研究は，NEDO 産業技術研究助成事業の支援を受けて行われた．

参考文献

[Hearst 1992] Hearst, M.A., “Automatic Acquisition of Hyponyms from Large Text Corpora,” Proceedings of the 14th International Conference on Computational Linguistics, pp.539-545, 1992.

[Kessler 1963] Kessler, M.M., “Bibliographic Coupling between Scientific Papers,” American Documentation, Vol.14, No.1, pp.10-25, 1963.

[Small 1973] Small, H., “Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents,” Journal of the American Society for Information Science, Vol.24, pp.265-269, 1973.

[相澤 2006] 相澤彰子，類語関係抽出タスクにおけるコーパス規模拡大の影響，情報処理学会研究報告 自然言語処理,NL-175,pp.91-98,2006.

[安藤 2003] 安藤まや，関根聡，石崎俊，定型表現を利用した新聞記事からの下位概念単語の自動抽出，情報処理学会研究報告 自然言語処理，NL-157，pp.77-82，2003.

[大石 2006] 大石康智，伊藤克亘，武田一哉，藤井敦，単語の共起関係と構文情報を利用した単語階層関係の統計的自動識別，情報処理学会研究報告，SLP-61，pp.25-30，2006.

[新里 2005] 新里圭司，鳥澤健太郎，HTML 文書からの単語間の上位下位関係の自動獲得，自然言語処理，Vol.12，No.1，pp.125-151，2005.

[藤原 1996] 藤原譲，劉野，頼静絹，意味関係記述のための概念構造モデル，情報処理学会研究報告 情報学基礎，FI-043，pp.109-114，1996.