用語の属性を考慮した上位，下位概念辞書の構築

福田 悟志[1]　難波 英嗣[1]　竹澤 寿幸[1]　橋田 浩一[2]

[1]広島市立大学大学院 情報科学研究科　[2]東京大学大学院 情報理工学系研究科

1.はじめに

本研究では，特許公開公報から構築されたシソーラスにおける上位，下位概念に対して，上位語と下位語に共通する属性(ファセット)を発見する手法を提案する．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として，「A や B などの C」といった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ものが挙げられる．例えば，「テンキーやフラットキーボードなどのキーボード」という文の場合，「キーボード」が上位概念，「テンキー」，「フラットキーボード」が下位概念となる．しかし，どのような側面でこれらの概念が上位，下位関係であるのかを裏付けることは難しい．そこで本研究では，定型表現を用いて抽出した概念が要素技術である場合，それに付随する属性に着目する．ここで，本研究における「要素技術」とは，研究で使用されたアルゴリズムやツール，技術的手法を指し，要素技術から得られる特徴・性質を「属性」とする．例えば，「キーボード」を要素技術とした時，「操作性」，「使い勝手」，「誤操作」など，主にユーザのキー操作に関するものが属性となる．また，「テンキー」や「フラットキーボード」の場合，「操作性」，「使い勝手」，「操作感」など，「キーボード」と同様，ユーザ操作に関するものが属性となる．このことから，「キーボード」と「フラットキーボード」，「テンキー」は，「操作性」という共通の属性をもつ上位，下位関係であると示すことができる．そのため本研究では，特許公開公報から，要素技術とその属性を自動的に抽出し，上位，下位関係にある概念間に共通する属性を発見する手法を提案し，属性を含めた上位，下位概念辞書の構築を行う．

しかし，「AやBなどのC」の定型表現を用いた際，「コンピュータやワープロなどのキーボード」という文の場合，「キーボード」が上位概念，「コンピュータ」，「ワープロ」が下位概念と誤って抽出されてしまう．ここで，本研究の手法を用いた時，「コンピュータ」に付随する属性は，「操作性」の他に，「セキュリティ」，「処理時間」など，ユーザ操作以外に関する様々な種類の属性が存在すると示すことができる．また，一般的な上位，下位概念の定義から，下位概念に位置する要素技術は，上位概念のそれを発展，改良したものと考えることできる．そのため，下位概念における属性の種類は，上位概念より限定されると考えることができるため，定型表現における「キーボード」と「コンピュータ」は，誤った上位，下位関係であるとみなすことができる．同様に，共通する属性が存在しない場合，概念間に関係性がないとみなすことができるため，誤った上位，下位関係であると判断できる．このように，本研究の手法を用いることで，上位，下位概念から，誤った上位，下位関係を効率的に除去することができることを示す．

2.関連研究

本節では，「上位，下位概念の抽出」，および「属性を考慮した様々な知識源からの知識獲得」に関する研究について述べる．

2.1. 上位，下位概念の抽出

大量のテキストデータから，上位，下位関係を獲得する手法はこれまでに数多く提案されている．近年では，HTML 文書の構造を利用する手法[Shinzato 2004]や，Wikipedia の構造を利用する手法[Oh 2009, 山田 2012]が提案されているが，論文や特許などのテキストベースから獲得する場合，「A や B などの(等の)C」といった定型表現を用いる手法[相澤 2006, Hearst 1992, Nanba 2007]が一般的である．

特に，Nanba は，上記の定型表現を用い，1993 年から 2002 年までの 10 年分の特許公開公報から，異なり語数 1,825,518 語，上位，下位概念 7,031,159 で構成されるシソーラスを構築している．本研究では，Nanba の手法を用いて，1993 年から 2012 年までの 20 年分の特許公開公報を用いてシソーラスを構築し，このシソーラスから，誤った上位，下位概念の自動抽出を行う．

2.2. 属性を考慮した様々な知識源からの知識獲得

様々な資源から知識獲得を行う研究が数多く行われている．Web の検索履歴から知識獲得を行う場合[小町 2008, 関口 2010]，以下のような手順が用いられることが多い．

(1)「東京　大阪　名古屋」など，性質の似た用語集合をシードとしてシステムに与える．
(2) 各シードの用語と共起する語句を，2 語から構成される検索質問の履歴データから抽出する．
(3) 手順(2)で抽出された語集合と共起頻度の高い語句を，手順(1)で入力したシードの関連語として出力する．

ここで，手順(2)で抽出される共起語の多くは，手順(1)で入力されるシードの属性(例えば，この例の場合「航空券」や「名所など」)であり，用語の属性を考慮している．

一方，Wikipedia から知識獲得を行う場合，山田ら[山田 2012]は，Wikipedia 記事のタイトルと，Wikipedia 記事から獲得した上位，下位関係の上位概念，下位概念は，「対象，属性，属性値」の 3 つ組で解釈できると述べている．例えば，「黒澤明」が記事タイトル，「作品→七

人の侍」がその記事から獲得された上位下位関係である時，「作品」と「七人の侍」を「黒澤明」という対象の属性，属性値と解釈している．

この他にも，Web 上における商品ページから，エントロピーを用いたパターンのスコア付けにより属性を抽出する研究[坂地 2010]や，ネットオークションの出品情報に多数存在する属性情報を，機械学習により自動抽出する研究[塚原 2009]，Web における表形式や箇条書きなどのレイアウトから，機械学習やパターンを用いて属性関係を抽出する研究[Chen 2000, Yoshida 2004, 吉永 2007]が存在するが，本研究では，特許公開公報を対象としている．そこで本研究では，福田ら[福田 2013]が提案した手法を用いる．

福田らは，特定分野の特許公開公報から，機械学習を用いて，要素技術とその効果を示す表現を自動的に抽出する手法を提案している．福田らは，論文と特許において，"を用いた"や"に基づいた"などの直前には要素技術を表す用語が出現し，"が可能になる"や"ができる"の直前には効果を表す表現が出現する傾向があることに着目し，要素技術を示す手がかり語の有無，および効果を示す手がかり語の有無を素性として機械学習に用いている．本研究では，福田らの手法を用いて，特許公開公報から要素技術と属性を自動的に抽出する．さらに，要素技術間における属性の共通性を考慮することで，対象の上位，下位関係に共通している属性を発見し，上位，下位関係における新たな知識獲得を行う．

3. 特許公開公報からの要素技術と属性の抽出方法，および誤った上位，下位概念の検出手段

3.1. 特許公開公報の構造解析におけるタグの定義

本研究では，福田ら[福田 2013]の手法を用いて，機械学習による特許公開公報の構造解析を行う．以下に，本研究で使用する構造タグを示す．

- **TECHNOLOGY**：要素技術を示す．
- **EFFECT**：効果(新しい機能の追加，新しく得られた物質，問題点の抑制や解決したこと)を示す．
- **ATTRIBUTE，VALUE**：例えば，「処理速度(ATTRIBUTE)が向上(VALUE)」のように，「属性」と「属性値」の対で表現する．

上記のタグを付与した例を図 1 に示す．

PM 磁束制御用コイルを設けて<TECHNOLOGY>閉ループフィードバック制御</TECHNOLOGY>を適用するため，<EFFECT><ATTRIBUTE>電気損失</ATTRIBUTE>を<VALUE>最小化</VALUE></EFFECT>できる．

図 1 特許への要素技術と効果に関するタグ付与の例

3.2. 要素技術－属性リストの構築

本節では，特許公開公報からの要素技術と属性を抽出する手順について述べる．

まず，特許公開公報に含まれる3つの項目【発明が解決しようとする課題】【課題を解決するための手段】【発明の効果】に対して，3.1.節で定義したタグを付与する．次に，【発明が解決しようとする課題】，および【課題を解決するための手段】の項目から，TECHNOLOGY タグが付与された箇所を抽出し，【発明の効果】の項目から，EFFECT タグ内における，ATTRIBUTE タグが付与された箇所を抽出する．その後，抽出された要素技術と属性をそれぞれ対応付ける．

この時，本研究では，次の点について考慮を行う．一般に，特許文の特徴として，文意を一義的に解釈させるために，1 文を長く記述し，複雑な複合名詞や長い名詞句を多用する傾向がある．ゆえに，TECHNOLOGY タグが付与される箇所も長くなる．例えば，「上記課題を解決する為に，本発明は手書きによる画像情報を入力する画像入力手段と，文字コードを入力するキーボードを有し，」という文の場合，「手書きによる画像情報を入力する画像入力手段」と「文字コードを入力するキーボード」の箇所に対して TECHNOLOGY タグが付与される．しかし，特許から抽出される要素技術表現の多くは，上記のように，「X を Y する Z」や「X を Y させる Z」，「X を Y するための Z」といった修飾表現で構成されている．そのため本研究では，TECHNOLOGY タグが付与された箇所における末尾の名詞句を要素技術とみなし，属性と対応付ける．図 2 は，その一例であり，各行，「要素技術 属性 頻度」を示している．

キーボード	操作性	65
キーボード	使い勝手	11
キーボード	誤操作	4
コンピュータ	消費電力	24
コンピュータ	セキュリティ	13
コンピュータ	誤操作	1
フラットキーボード	操作感	1
フラットキーボード	操作性	1

図 2 要素技術－属性リストの例

3.3. 属性を用いた誤った上位，下位概念の検出手段

上記の手順から作成した「要素技術－属性リスト」を用いて，以下の 2 種類による判断基準から，誤った上位，下位概念を検出する．

(1) 上位，下位概念における要素技術間の類似度を測定する．具体的には，要素技術における属性の頻度から，各要素技術における属性の TF-IDF 値を算出し，コサイン類似度を用いて類似度を計算する．
(2) 特定の要素技術における属性の種類を調べる．そして，上位，下位概念の要素技術間における属性の異なり数を比較する．

4. 実験

3 節で述べた手法の有効性を調べるための実験を行った．

4.1. 実験条件

実験データ

2 節で述べた Nanba の手法[Nanba 2007]を用いて，1993 年から 2012 年までの特許公開公報のデータから

上位，下位概念のシソーラスを構築した．以下に，本研究のシソーラスにおける上位，下位関係の異なり数，および全用語数を示す．

- 上位，下位関係(異なり数)：15,682,721 関係
- 全用語数(異なり数)：4,398,498 語

評価データは以下の手順で作成した．

(1) 特許公開公報から作成した上位，下位関係から，「キーボード」という単語が上位概念にある下位概念の用語を抽出する．
(2) その中から要素技術を表す用語を抽出する．この時，一文字の用語は除去する．

以上述べた手順により，手順(1)で得られた 490 語から，250 語の下位概念を抽出した．その一部を図 3 に示す．各行，「特許公開公報における上位，下位関係の頻度 下位概念」を示している．出現頻度が最も高い下位概念は「テンキー」であり，上位，下位関係として 76 回出現していることが分かった．また，上位，下位概念の要素技術に付随している属性は，178,120 語(異なり)であった．

76	テンキー
58	パーソナルコンピュータ
30	コンピュータ
27	タイプ
23	携帯電話
15	ワープロ
14	コピーキー
7	テンキーボード
1	小型キーボード
1	フラットキーボード

図 3　上位概念「キーボード」における下位概念の例

評価方法

3.3 節で述べた 2 種類の判断基準を適用し，検出した下位概念の評価，考察を行う．その後，最終的に獲得した上位，下位関係から，概念間に共通する属性を用いた上位，下位概念辞書の構築について述べる．

4.2. 実験結果および考察

誤った上位，下位概念の検出

まず，コサイン類似度の値が 0，すなわち，上位概念「キーボード」の属性と全く共通性を持たなかった下位概念について調べる．類似度が 0 であった下位概念は 34 語であった．これらを人手で判断し，全ての下位概念が「キーボード」と上位，下位関係でないことを確認した．図 4 に，特許公開公報における上位，下位関係の頻度が 2 以上であった下位概念を列挙する．

図 4 の例において，「ワープロ」，「電卓」，「読影レポート」など，特許公開公報において，高い頻度で出現するが，実際には正しい上位，下位関係でない下位概念が出力されていることが分かる．特に，「ワープロ」と「電卓」は，4.1.節の手順(1)で獲得した 490 語のうち，それぞれ 7 番目，14 番目に頻度が高い用語である．この結果から，要素技術間における属性の共通性が無いものを誤った上位，下位概念であると判断することは有用であるといえる．また，図 4 において，「JIS」という用語が出力されている事が分かる．これは，上位概念が「キーボード」である場合，「JIS 配列キーボード」のことを指すと解釈することもできるが，「JIS 規格」や「シフト JIS」などと解釈することもでき，この場合は「キーボード」の下位関係には当たらない．このように，複数の意味や用途を示唆する用語は，除去することが望ましいといえる．

15	ワープロ	2	クリック音
9	電卓	2	JIS
3	読影レポート	2	携帯型電話器
3	ノートパソコン	2	情報端末装置
2	アイコンデータ	2	FDD コントローラ

図 4　上位概念「キーボード」の属性と全く共通性を持たなかった下位概念と上位，下位関係の頻度

次に，類似度の値が高かった結果，すなわち，システムが上位概念「キーボード」と属性の観点から高い共通性を持っていると判断した下位概念について調べた．図 5 に，「キーボード」と類似度が高かった上位 10 件の結果を示す．各行，「特許公開公報における上位，下位関係の頻度　類似度　下位概念」を示している．図 5 において，「入力」や「選択」など，「キーボード」と全く関係のない下位概念が上位に出力されていることが分かる．

7	0.4634	タッチパネル
2	0.3736	操作
1	0.3558	入力
1	0.3004	操作キー
2	0.3001	操作ボタン
1	0.2990	選択
4	0.2984	ボタン
2	0.2853	表示画面
1	0.2830	設定
2	0.2736	ポインティングデバイス

図 5　上位概念「キーボード」との類似度が高かった下位概念とその値，および上位，下位関係の頻度

次に，「キーボード」における属性の異なり数，および下位概念の要素技術における属性の異なり数について調査した．下位概念における属性の異なり数が多い上位 9 件の結果，および上位概念「キーボード」における属性の異なり数を図 6 に示す．図 6 から，「入力」や「選択」に付随する属性の数は，「キーボード」の数の 40 倍以上あることが分かった．また，「キーボード」の属性の数より多い種類の属性を持つ用語(53 語)について調べた結果，「キーボード」の下位概念に当てはまる用語は存在しなかった．

図 5 に示す実験結果から，付随する属性が 662 種類以上存在する下位概念を除去した．その結果を図 7 に示す．各行，「特許公開公報における上位，下位関係の頻度　類似度　下位概念の用語」を示している．

27997	入力	11798	電源
22219	選択	9634	端末
18777	スイッチ	7744	装置
16747	設定	4390	コンピュータ
14565	操作	661	キーボード

図 6　下位概念に付随する属性の種類

2	0.2736	ポインティングデバイス
1	0.2461	フラットキーボード
3	0.2194	マウス
1	0.2187	入力キー
2	0.2136	プッシュボタン
1	0.2013	キー操作部
1	0.1833	操作卓
1	0.1801	ジョイスティック
1	0.1789	選択キー
6	0.1615	ワンタッチキー

図 7　属性の種類が「キーボード」より多い下位概念を除去した後の類似度結果

図 5 では出現しなかった「フラットキーボード」が，図 7 では上位に出現していることが分かる．これは，「キーボード」の下位概念に当たる用語であると判断できる．このため，対象となる上位概念の用語に付随する属性より多い種類の属性を持つ下位概念の用語を類似度結果から除去することで，人手での上位，下位関係の負担を軽減することが出来ると考えられる．

属性付き上位，下位概念辞書の構築

ここで，上位概念「キーボード」と下位概念「フラットキーボード」に付随する共通の属性について調べた．その結果，「操作性」と「操作感」の 2 種類の属性が共通していることが分かった．この結果を用いることで，構築した上位，下位概念シソーラスにおける「キーボード」と「フラットキーボード」に対する新たな知識獲得を行うことができると考えられる．すなわち，「キーボード」と「フラットキーボード」は，「操作性」と「操作感」という 2 つの属性の側面を持つ上位，下位関係であるとみなすことができる．そこで本研究では，本手法を適用して獲得した 163 件の上位，下位関係を対象に，人手で「キーボード」の下位概念として正しいと判断した用語から，それぞれに共通する属性を抽出した．表 1 にその一部を示す．このような属性付き上位，下位関係を収集することで，どのような側面で上位，下位関係にあるのかを明確にする上位，下位概念辞書を構築することができる．

5. おわりに

本研究では，上位，下位概念の用語を要素技術とみなし，それに付随する属性の共通性，および属性の種類を比較することで，誤った上位，下位概念の検出を行った．上位概念を「キーボード」とした時，合計 87 件の誤った上位，下位関係を検出することができた．

また，上位，下位概念に共通する属性を発見することで，特許シソーラスにおける上位，下位概念に対する新たな知識獲得を行うことができることを示した．

表 1　人手により作成した属性付き上位，下位概念

上位概念	下位概念	属性
キーボード	フラットキーボード	操作性，操作感
キーボード	タッチキーボード	操作性，キー操作性
キーボード	テンキー	操作性，操作ミス，コスト，使い勝手
キーボード	小型キーボード	誤入力
キーボード	テンキーボード	操作性，精度

参考文献

[相澤 2006] 相澤彰子 (2006) "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.

[Chen 2000] Chen, H.-H., Tsai, S.-C. and Tsai, J.-H. (2000) "Mining tables from large scale html texts". *Proc. COLING*.

[福田 2013] 福田悟志, 難波英嗣, 竹澤寿幸 (2013) "論文と特許からの技術動向情報の抽出と可視化". 情報処理学会論文誌データベース, Vol.6, No.2, pp.16-29.

[Hearst 1992] Hearst, M.A. (1992) "Automatic Acquisition of Hyponyms from Large Text Corpora". *Proc. 14th International Conference on Computational Linguistics*, pp.539-545.

[小町 2008] 小町守, 鈴木久美 (2008) "検索ログからの半教師あり意味知識獲得の改善". 人工知能学会論文誌, Vol.23, No.3.

[Nanba 2007] Nanba, H. (2007) "Query Expansion using an Automatically Constructed Thesaurus". *Proc. 6th NTCIR Workshop Meeting*, pp.414-419.

[Oh 2009] Oh, J.-H., Uchimoto, K. and Torisawa, K. (2009) "Bilingual Co-Training for Monolingual Hyponymy-Relation Acquisition". *Proc. ACL 2009: IJCNLP*, pp.432-440.

[坂地 2010] 坂地泰紀, 小林暁雄, 関根聡, 竹中孝真 (2010) "商品ページからの属性・属性値抽出と同一商品クラスタリング手法". 言語処理学会第 16 年次大会, pp.371-374.

[関口 2010] 関口裕一郎, 田中智博, 内山匡, 藤村滋, 望月崇由, 鈴木智也 (2010) "検索クエリログのセッション情報を利用した属性語句抽出". DEIM Forum.

[Shinzato 2004] Shinzato, K. and Torisawa, K. (2004) "Acquiring Hyponymy Relations from Web Documents". *Proc. 20th International Conference on Computational Linguistics*, pp.73-80.

[塚原 2009] 塚原裕常, 宮崎林太郎, 西村純, 前田直人, 森辰則, 小林寛之, 石川雄介, 田中裕也, 翁松齢 (2009) "ネットオークションの出品情報文書からの 2 段階属性抽出". 言語処理学会第 15 回年次大会, pp.400-403.

[山田 2012] 隅田明日香, 吉永直樹, 鳥澤健太郎(2012) "Wikipedia の記事構造からの上位下位関係抽出". 自然言語処理, Vol.16(3), pp.3-24.

[Yoshida 2004] Yoshida, M., Torisawa, K. and Tsujii, J. (2004) "Integrating tables on the World Wide Web". *Translations of the Japanese Society for Artificial Intelligence*, pp.548-560.

[吉永 2007]吉永直樹, 鳥澤健太郎 (2007) "Web からの具体物の属性・属性値情報の自動獲得". 言語処理学会第 13 回年次大会, pp.887-809.