日英特許データベースからのシソーラスの自動構築

間弓沙織[1]，難波英嗣[2]，竹澤寿幸[2]
1. 広島市立大学　情報科学部
2. 広島市立大学大学院　情報科学研究科

1. はじめに

本研究では，日英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，文献の検索や専門文書の執筆の際の情報源として，また，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ている．また，専門用語の翻訳の際には，正確な対訳辞書が必要不可欠であるが，既存の辞書に登録されていない用語が増加し続けており，対訳辞書を人手で継続，管理するには非常にコストがかかる．そのため，専門用語を特許文書から抽出し，正しい訳語を自動推定して，翻訳辞書作成を支援するシステムが求められている．

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．また，専門用語の訳語推定法については，統計的機械翻訳モデルを用いて訳語推定を行う手法，及び，既存の対訳辞書を利用した要素合成法を併用して，専門用語の訳語を推定する手法が提案されている[森下 2010]．

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し，日英特許データベースからそれぞれ上位，下位概念を獲得する．次に，統計的機械翻訳モデルを用いた訳語推定法に着目し，引用分析手法[Kessler 1963, Small 1973]と合わせて，日本語と英語の用語間の対応付けを行うことにより，日英特許シソーラスを自動的に構築する．これにより得られたシソーラスを用いることで，文献の検索や専門文書の執筆，訳語推定など，幅広く活用することが可能になると考えられる．

本論文の構成は以下のとおりである．2 節では，特許データベースからの上位，下位概念の抽出法を述べ，3 節では，日英の用語間の対応付け方法について説明する．4 節では，本研究で行った実験について述べ，5 節で実験結果からの考察を述べる．最後に 6 節で本稿をまとめる．

2. 上位、下位概念の抽出

日本文では「A や B などの C」，「A や B 等の C」，英文では，“A, such as B and C”といった定型表現に着目する．例えば，「染料や顔料などの着色剤」という文では，「着色剤」という上位概念に対して，「染料」「顔料」が下位概念であることが分かる．また，“pets, such as cats and dogs”という文では，“pets”という位概念に対して，“cats”“dogs”が下位概念であることが分かる．本研究では，このような定型表現に着目し，図 1 のような日本文特許データベースと図 2 のような英文特許データベースから上位，下位概念を獲得する．

1993-000024:【構成】　天然繊維、紙、パルプなどの天然素材の繊維の集合体で加工された開口率５～６０％の網状で厚み５～４０mm の芝養生マット。

図 1：日本文特許データベースの例

The grinding wheel 2 comprises a generally hourglass shape along its width and is made of a suitable abrasive material such as aluminum oxide or cubic boron nitride (CBN).

図 2：英文特許データベースの例

3. 日英の用語間の対応付け

3.1. フレーズテーブルを用いた対応付け

日英の用語対候補の作成には，統計的機械翻訳技術を用いる．統計的機械翻訳では，対象とする言語に関する文法的知識を必要としないため，容易に翻訳システムを構築することができる．本研究では，統計的機械翻訳ツールである GIZA++を使用し，翻訳モデル用に日英特許から抽出された 3,185,254 文対を用い，言語モデル用に 3,186,284 文の日本語特許文を用いて，フレーズテーブルの作成を行った．

以下に，作成したフレーズテーブルを用いた用語の対応付け方法について説明する．図 3 はその流れを示したものである．

(1) **翻訳**
テキストデータベースから獲得された日本語の上位概念，下位概念を，作成したフレーズテーブルを用いてそれぞれ単独で翻訳する．

(2) **上位，下位の候補を作成**
得られた訳語から，全ての組み合わせで上位，下位の候補を作成する．

(3) **対応付け**
得られた候補の中から，テキストデータベースから獲得された英語の上位，下位概念に当てはまるものがあれば，日英の用語を対応付けする．

上記の(1)～(3)のように日英の用語間の対応付けを行った結果，2,635 対の日英用語対が得られた．

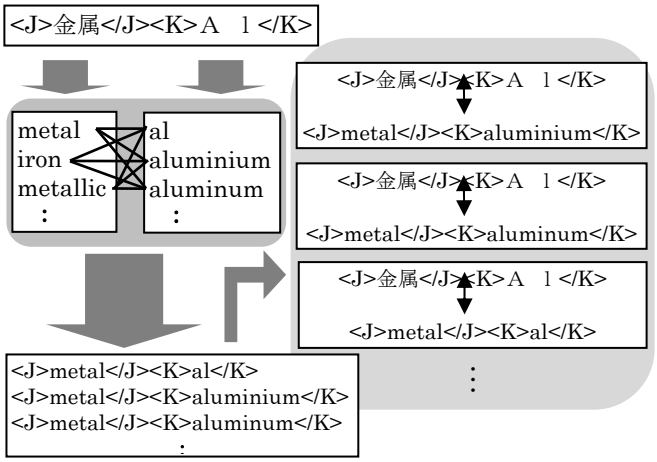

図 3：フレーズテーブルを用いた対応付け

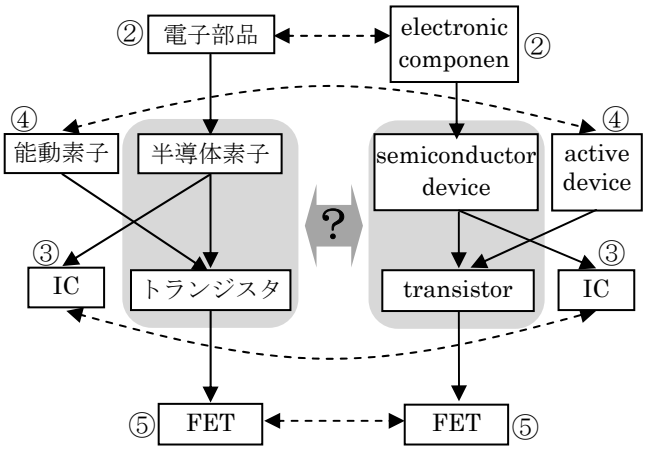

図 4：上位，下位関係を用いた対応関係の検出

3.2. 日英用語対の抽出

3.1 節のフレーズテーブルを用いた対応付けで得られた日英の用語対 2,635 対から，用語対候補を絞り込む．用いた素性は，以下の 5 種類である．

① 翻訳確率
② 日本語，英語の上位語の上位語の一致数
③ 日本語，英語の上位語の下位語の一致数
④ 日本語，英語の下位語の上位語の一致数
⑤ 日本語，英語の下位語の下位語の一致数

②～⑤については，それぞれの最大一致数で，個々の一致数を割った値を素性値とする．また，日英の用語が一致しているかどうかの判断は，4.1 節で作成したフレーズテーブルを用いて日本語の用語を翻訳し，英語の用語と比較することで行う．翻訳の際，フレーズテーブルに登録されている訳語の中で，最も翻訳確率（日→英の翻訳確率と英→日の翻訳確率の積）の高い訳語のみを使用する．

②～⑤を素性として用いたアイディアは，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．ここでは，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，日英対応関係を抽出する．図 4 は，「半導体素子 ＞ トランジスタ」という日本語の上位，下位概念と，“semiconductor device ＞ transistor”という英語の上位，下位概念を中心に，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 4 において，実線で結ばれたものは上位，下位関係を表し，点線で結ばれたものは日英対応関係を表す．このような関係が成り立っているとき，「半導体素子 ＞ トランジスタ」と“semiconductor device ＞ transistor”は，共通の上位語あるいは下位語を持ち，対応関係にあると考えられる．

本研究では，素性①と，②～⑤の素性のいずれかを用いた 2 種類の素性により，4 通りの組み合わせを使用する．以下，①，②の組み合わせを (a)，①，③の組み合わせを(b)，①，④の組み合わせを(c)，①，⑤の組み合わせを(d)とする．

用語対の抽出方法について，以下に(a)を例として説明する．

(1) **素性の和を計算**
素性①の値が a，素性②の値が b であるとき，$a^{\beta}+\alpha\times b$ を計算する．ここでβは，1/5，1/10，1/15，1/20 の 4 通りで計算する．また，αは 0.1，0.2，…，0.9 とする．

(2) **訓練**
(1)で求めた和が，ある閾値 x 以上のときに正解，x 未満のときに不正解とし，F 値を算出する．このとき，訓練用データを用いて x の値を 0～2 の間で 0.01 ずつ変化させ，F 値が最大となる x を求める．

(3) **評価**
(2)で得られた x を用い，テスト用データで評価を行う．

(b)，(c)，(d)についても同様の処理を行う．

4. 実験

3 節で述べた手法により，実験を行った．

4.1. 実験方法

◆ 実験に用いるデータ

フレーズテーブルを用いた対応付けにより得られた日英の用語対 2,635 対の正解判定を人手で行った結果を使用する．人手による判定結果を表 1 に示す．

表 1：人手による判定結果

正解	不正解	合計
982	1,653	2,635

◆ 比較実験

本研究では，用語間の上位，下位関係を用いた引用分析手法の有効性を確認するため，比較手法とし

て，引用分析手法を用いた素性を与えずに実験を行う．比較手法で用いる素性と，提案手法で用いる素性を以下の表 2 にまとめる．

表 2：実験に用いる素性

	素性①	素性②	素性③	素性④	素性⑤
提案手法(a)	○	○			
提案手法(b)	○		○		
提案手法(c)	○			○	
提案手法(d)	○				○
比較手法	○				

◆ 評価尺度

上記の実験に用いるデータを 4 分割し，そのうち 3 つを訓練用，1 つを評価用として，4 分割交差検定を行うことで評価を行う．

人手判定によって正解とした用語対数をP_m，システム判定によって正解とされた用語対数をP_sとし，さらに人手判定とシステム判定の結果が正解で一致する用語対数をP_{m-s}とする．評価には，表 3 に示す精度，再現率，F 値を用いた．

表 3：評価尺度

精度	再現率	F 値
$\frac{P_{m-s}}{P_s}$	$\frac{P_{m-s}}{P_m}$	$\frac{2 \times \text{再現率} \times \text{精度}}{\text{再現率} + \text{精度}}$

4.2. 実験結果

提案手法(a)～(d)と，比較手法によって得られた精度，再現率，F 値を以下の表 4 に示す．比較手法は(e)とする．表に示した値は，それぞれの実験において F 値が最大のときの結果である．

表 4：実験結果

	α	β	精度（%）	再現率（%）	F 値（%）
(a)	0.1	1/10	76.4	78.1	77.1
(b)	0.1	1/20	76.3	79.5	77.4
(c)	0.1	1/15	75.8	78.4	76.9
(d)	0.1	1/15	77.5	79.4	78.3
(e)	0	1/15	78.5	77.8	78.0

表 4 より，提案手法(d)において，比較手法より高い再現率，F 値が得られ，提案手法の有効性が確認された．

5. 考察

◆ 日英用語対の抽出

◇ システムが誤って正解と判定したもの

以下に，人手では不正解と判定したが，システムでは正解と判定された 226 件の検出誤りを種類ごとに分析し、主要な原因をいくつか示す．226 件の検出誤りは，大きく次の 5 種類に分類できる．

① 類似した用語（73.9%）

226 件のうち，167 件（73.9%）が「亜鉛」と“aluminum-zinc”のような類似した用語と対応付けされたものだった．類似の用語は上位，下位概念に同じ用語を持つ可能性が高いため，引用分析手法において一致数が多くなってしまったと考えられる．

② 抽出個所の不十分な用語（10.6%）

226 件のうち，24 件（10.6%）が「弾性体」と“elastic”のような抽出個所の不十分だと思われる用語と対応付けされたものだった．

③ 余分な単語が含まれている用語（4.4%）

226 件のうち，10 件（4.4%）が「金属」と“to metal”のような余分な単語が含まれている用語と対応付けされたものだった．

上記の①～③に共通した原因として，フレーズテーブルを用いた用語対の作成段階で，翻訳候補の全ての組み合わせで上位，下位概念の候補を作成したため，「類似した用語」や「抽出個所の不十分な用語」，「余分な単語が含まれている用語」と対応付けされたものが多かったと考えられる．この問題は，上位，下位概念の候補を作成する際に，全ての組み合わせを候補とするのではなく，翻訳確率を考慮して候補を作成することで改善できると思われる．

④ 上位語と下位語が同じ用語（4.4%）

226 件のうち，10 件（4.4%）が「車両 ＞ 自動車」に対して“vehicles ＞ vehicle”のような上位語と下位語が同じ用語と対応付けされたものだった．原因としては，フレーズテーブルを用いた用語対の作成段階で，全ての組み合わせで上位，下位の候補を作成したため，英語の上位，下位語が同じになってしまったと考えられる．そのため，類似した用語と対応付けされてしまい，それぞれの上位，下位概念に同じ用語を持つ可能性が高いため，引用分析手法において，一致数が多くなってしまったと考えられる．この問題は，上位，下位概念の候補を作成する際に，候補の中から上位語と下位語が同じものを削除することで改善できると思われる．

また，上記の①～④に共通している原因として，上位，下位概念の獲得の際の問題が考えられる．実際に本研究で抽出されたものには，上位，下位概念ではないものや，余分な語を含んでいるものがあった．

◇ システムが正解と判定できなかったもの

以下に，人手では正解と判定したが，システムでは不正解と判定された 201 件の再現できなかった用語対を原因の種類ごとに分析し、主要な原因をいくつか示す．201 件の再現できなかった用語対は，大きく次の 3 種類に分類できる（重複あり）．

① 複数形（33.3%）

201 件のうち，67 件（33.3%）が「金属」と“metals”のような複数形の用語と対応付けされたものだった．

② 元素記号（22.9 %）

201 件のうち，46 件（22.9%）が「銅」と“cu”のような元素記号と，元素の名称で書かれた用語が対応付けされたものだった．

③ 略語（21.4%）

201 件のうち，43 件（21.4%）が「ＣＤ」と“compact disk”のような略語である用語と対応付けされたものだった．

上記の①～③の用語は，訳語としてはあまり一般的ではない．①～③に当てはまらない用語対も，あまり一般的ではない訳語と対応付けされたものが多かった．よって，フレーズテーブルにおいて翻訳確率が低くなり，再現できなかったと考えられる．この問題は，フレーズテーブルを作成する際の学習データを増やすことで改善できると思われる．また，一般的でない表現は抽出された上位，下位概念も少なく，引用分析手法において一致数が少なくなったと考えられる．

◆ 比較実験

◇ 精度について

提案手法を用いた場合は 226 件，比較手法を用いた場合は，208 件の検出誤りがあった．検出誤りの中で，提案手法では不正解と判定したが，比較手法では正解と判定した用語対は 0 件であった．逆に，比較手法では不正解と判定したが，提案手法では正解と判定した用語対は 18 件であった．

提案手法において“vehicles ＞ vehicle”のように，上位語と下位語が同じ用語と対応付けされたものが誤って検出された．また「アルミニウム」に対して“aluminum film”のように，類似の用語が対応付けされたものも誤って検出された．これらの用語対は，上位，下位概念に同じ用語を持つ可能性が高いため，引用分析手法において一致数が多くなってしまい，比較手法において誤って検出してしまったと考えられる．これらの問題は，上位，下位概念の候補を作成する際に，候補の中から上位語と下位語同じものを削除したり，上位，下位概念の候補を作成する際に，翻訳確率を考慮して候補を作成したりすることで，ある程度改善できると思われる．よって，このような改善を行うことによって，提案手法が正しく正解を判定することにおいて有効となると考えられる．

◇ 再現率について

提案手法を用いた場合は 201 件，比較手法を用いた場合は，216 件の再現できなかった用語対があった．再現できなかった用語対の中で，比較手法では正解と判定したが，提案手法では不正解と判定した用語対は 0 件であった．逆に，提案手法では正解と判定したが，比較手法では不正解と判定した用語対は 15 件であった．

比較手法において，「記憶媒体」に対して“record medium”や，「車両」に対して複数形の“vehicles”のように，一般的でない訳語と対応付けされたものが再現できなかった．これらの用語対は，フレーズテーブルにおいて翻訳確率が低くなり，再現できなかったと考えられる．提案手法で再現できなかった用語対も，一般的でない訳語と対応付けされたものであったが，引用分析手法を用いることにより，ある程度問題が改善されることが確認できた．よって，引用分析手法を用いた日英用語対の抽出を行う本研究の提案手法は，より多くの正解を再現することにおいて有効であると考えられる．

6. おわりに

本研究では，日英特許データベースから上位，下位概念を獲得し，日英の用語間の対応付けを行うことにより，シソーラスの自動構築を行った．

上位，下位概念の抽出の際，定型表現に着目し，日英特許データベースから上位，下位の用語対を獲得した．また，日英の用語間の対応付けにはフレーズテーブルを用い，その後，引用分析手法を用いて日英用語対の絞り込みを行った．

実験の結果，提案手法において，精度 77.5%，再現率 79.4%，F 値 78.3%という結果が得られた．また，比較手法を用い，抽出した用語対の精度と再現率を比較することで提案手法が有効なものであることを示した．

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．

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.

[森下 2010] 森下洋平，梁冰，宇津呂武仁，山本幹雄，フレーズテーブル及び既存対訳辞書を用いた専門用語の訳語推定，電子情報通信学会論文誌 D，Vol.J93-D，No.11，pp.2525-2537，2010.