F タームに基づいたオントロジーの構築

福田悟志[†1]　難波英嗣[†1]　竹澤寿幸[†1]　乾孝司[†2]
岩山真[†3]　橋田浩一[†4]　藤井敦[†5]

†1 広島市立大学大学院 情報科学研究科
†2 筑波大学大学院 システム情報工学研究科　†3 日立製作所 中央研究所
†4 東京大学大学院 情報理工学系研究科　†5 東京工業大学 情報理工学研究科

1. はじめに

本稿では，特許データベースから様々な文献に利用できるようなオントロジーを構築する手法について述べる．オントロジーとは，文献の検索や高度な言語処理に重要な情報源である．しかし，オントロジーを人手で構築し，更新することは非常にコストがかかる．一方で，テキストデータベースからシソーラスやオントロジーを自動構築する様々な手法が提案されているものの，人手による構築作業に取って代わるレベルまでには至っていない．そこで本稿では，最小減の労力で効率的にオントロジーを構築する枠組みについて述べる．

オントロジーを効率的に構築するため，我々は特許分類コード体系のひとつである F タームに着目する．F タームとは，特許を目的・利用分野・材料といった様々な観点から分類することを目的として日本国特許庁が構築した特許の分類体系のひとつである．F タームの詳細については 3 節で述べるが，実は，F タームの構造そのものがオントロジーに近い体系になっている．そこで本研究では，F タームの体系をオントロジーの構築に流用する．これをベースに，ブートストラッピング法と機械学習を組み合わせた手法を用いて，特許との親和性を保持しながら，学術論文など他のジャンルの文献にも利用可能なオントロジーの構築を目指す．

本論文の構成は以下のとおりである．次節では，関連研究について述べる．3 節では，F タームに基づくオントロジー構築手法を提案する．提案手法の有効性を確認するために行った実験について 4 節で報告し，5 節で本論文をまとめる．

2. 関連研究

2.1. 用語間関係の判別

大量のテキストデータから用語間の関係を判別する手法はこれまでに数多く提案されている．一般的に，論文や特許などのテキストベースを対象とする場合，X を上位語，Y を下位語とした時，「Y などの(等の)X」といった定型表現を用いる手法が一般的である[1, 2, 3, 4]．上記の定型表現を用いることで，X と Y は上位下位関係であることを判別することができる．この他にも，安藤ら[5]は，「Y という X」「Y のような X」「Y といった X」などのパターンも上位下位関係を判定するために有用であることを分析している．Kozareva ら[6]は，「X such as Y」「X are Y that」「X including Y」「X like Y」「such X as Y」という 5 種類の上位下位関係を表す表層パターンを用いることで，X と Y の位置関係を判別している．しかし，大量のテキストデータを対象に様々な用語対を判別する場合，上記で述べたパターン以外にも有用なものは数多く存在すると考えられる．また上位下位関係以外の様々な関係(例 : 部分全体関係)においても有用な判別パターンが存在すると考えられるが，これらを網羅的に人手で収集することは困難である．そのため本研究では，ブートストラッピング法[7, 8]により，複数の用語間関係を判別するために有用なパターンを網羅的に収集する．

2.2. ブートストラッピング法

本研究では，シードインスタンス集合を入力とした，パターン抽出とインスタンス抽出の 2 つのフェーズを繰り返すブートストラッピング法を考える．シードとして(X, Y)を与えた場合，パターン抽出フェーズにおいて，X と Y に挟まれているパターンを抽出する．そして，抽出したパターン集合から，「Y などの X」のようにインスタンスと共起しやすいパターンをいくつか選択する．インスタンス抽出フェーズでは，選択したパターンと共起するインスタンスを獲得する．そして，獲得したインスタンス集合からいくつか選択し，再びパターンを抽出する．このような処理を，停止条件が満たされるまで繰り返す．

本研究では，上位下位，部分全体，定義域属性関係という 3 種類の関係を対象に，そのカテゴリに属する新たな用語対を抽出することを目的としている(3.1 節で詳しく述べる)．このように複数のカテゴリを対象とした場合，特定のカテゴリに属するいくつかの用語対をシードとして用いる事が一般的である．

複数のカテゴリを対象にしたブートストラッピング法によるパターン・インスタンスの獲得を行う研究は数多く存在する．Krishnan ら[9]は，医学分野の特許文書集合を対象に，ブートストラッピング法を用いることで用語間の Treatment 関係(例:「A cure B」)と Causal 関係(例 :「A impact B」)を表す動詞(句)を抽出している．小町ら[10]および伊藤ら[11]は，Tchai アルゴリズムと呼ばれるブートストラッピング法を用いて，旅行，金融，番組名，芸能人などのカテゴリを対象に，Web 検索履歴から関連性の高いキーワード群を抽出している．Abe ら[12]は，動名詞を伴う複数の事態間関係(行為効果関係，部分全体関係)を収集するために，ブートストラッピング法を用いている．また，Kiso ら[13]は，ブートストラッピング法における重要な問題の一つである「各カテゴリにおける良質な(シード)インスタンスをどのように発見するのか[a]」を，HIS アルゴリズム[14]と組み合わせることで解決している．

上記で述べた研究におけるアプローチは全て Espresso アルゴリズム[15]に基づいている．これは，近年注目されている非常に精巧なブートストラッピング法のひとつである．次節では Espresso アルゴリズムの詳細を述べる．

2.3. Espresso アルゴリズム

公開公報から新たな用語間関係を収集するためのブートストラッピング法として, 本研究では Espresso アルゴリズムを適用する．Espresso アルゴリズムは，少量のシードインスタンスを用いて反復的に表層パターンの抽出を行い，多くの新たなインスタンスを収集する手法である．このアルゴリズムでは，従来のブートストラッピング法で問題となっていた意味ドリフトを考慮している．意味ドリフトとは，反復過程において複数のカテゴリで出現するようなパターン(ジェネリックパターン)やインスタンスを獲得してしまい，徐々にシードと関連性の低いものに移り変わっていく現象である．従来では，ジェネリックパターンを排除することで意味ドリフトを抑えることを行っていたが，獲得できるインスタンスの数が減少するという問題が新たに発生し，その結果，精度は高いが再現率が十分でないという欠点があった．このような意味ドリフトによる問題を軽減するために，Espresso アルゴリズムでは，スコアリング関数を用いて相互再帰的にインスタンスとパターンのスコアを定義している．これは，信頼度の高いパターンと頻繁に共起するインスタンスは信頼度が高く，信頼度の高いインスタンスと共起するパターンは非常に信頼性があるという考えに基づいている．パターン p およびインスタンス i ($i=\{x,y\}$ (*x,y*: インスタンスにおける用語))のスコアをそれぞれ$r_\pi(p)$，$r_\iota(i)$とした時，以下の式を用いて信頼度を計算する．

$$r_\pi(p)=\frac{1}{|I|}\sum_{i\in I}\frac{pmi(i,p)}{\max pmi}r_\iota(i) \quad (1)$$

$$r_\iota(i)=\frac{1}{|P|}\sum_{p\in P}\frac{pmi(i,p)}{\max pmi}r_\pi(p) \quad (2)$$

$$pmi(i,p)=\log\frac{|x,p,y|}{|x,*,y||*.p,*|} \quad (3)$$

ここで，*P* はパターン集合，*I* はインスタンス集合であり，|*P*|と|*I*|はパターンとインスタンスの数を表す．$|x,p,y|$はインスタンスを伴うパターン p の頻度を表している．また，$|x,*,y|$はインスタンスの頻度，$|*.p,*|$はパターンの頻度を表す．*pmi(i,p)*はインスタンスとパターン間の自己相互情報量(PMI: Pointwise Mutual Information)を表しており，max *pmi* は全てのインスタンスとパターンの組み合わせの間における *pmi* の最大値である．なお，$r_\pi(p)$，$r_\iota(i)$の初期値はそれぞれ 1 である．

Espresso アルゴリズムでは，反復過程において(1)式と(2)式を適用することで，精度を高く保ちながら再現率を大幅に向上させている．本研究では，Espresso アルゴリズムによるブートストラッピング法を用いることで，特許との親和性を保持した新たな用語間関係を獲得することを目指す．

2.4. 機械学習による関係判別

Espresso アルゴリズムを用いることで, シードインスタンスとの関連性の高いパターンを獲得することができるが，獲得したパターンを用いてインスタンスを抽出する場合，特定のカテゴリに特化したようなインスタンスが必ず獲得されるとは限らない．例えば，上位下位関係のカテゴリに属するシードインスタンスを用いて「Y といった X」というパターンが獲得されたとする．このパターンを用いて新たなインスタンスを獲得する場合,「キーボードといった入力装置」という文から(入力装置，キーボード)というインスタンスが獲得されるが,「キーボードといった入力部」という文が存在する場合，(入力部，キー

[a] 一般的には，特定のカテゴリ内で頻出するものを選択する方法，人手により選別する方法，ランダムに選択する方法が挙げられる．

ボード)というインスタンスが抽出される．このインスタンスは，部分全体関係にあるものと考えられるため，上位下位関係のカテゴリから除去する必要があるが，収集した全てのインスタンスを人手で判定することは困難である．

Girju ら[16]は，C4.5 と呼ばれる分類器を用いてインスタンスを分類する手法を提案している．この手法では，Iterative Semantic Specialization (ISS)手法により，パターンによる分類ルールを学習しており，高い精度と再現率を示している．しかし，訓練用データの作成に多大なコストを要しており，人手によるタグ付けを行う必要がある．また，対象としている用語間関係が部分全体関係のみである．本研究では，F タームにおける複数の用語間関係を対象としており，ブートストラッピング法により獲得した複数のカテゴリにおけるパターン集合を組み合わせて機械学習に適用することでインスタンスの判別を行う点で異なる．また，本研究で用いる訓練用データは，F タームに基づいたオントロジーを用いるため，非常に信頼性が高いという点が挙げられる．

3. F タームに基づくオントロジーの構築

本節では, F タームに基づくオントロジーの構築手法について述べる．本研究では，以下のステップによりオントロジーの構築を行う．

- Step 1: F タームからの知識抽出
- Step 2: Step 1 で得られた知識(用語間関係)をシードとして新たな用語間関係を獲得

各ステップにおける詳細を次節で述べる．

3.1. F タームからの知識抽出

3.1.1. F タームとは

F タームは，特許を目的・効果・構成などの様々な観点から分類することを目的とした分類体系であり，技術分野を示すテーマコードと観点の集合から構成される．ここでは，機械翻訳分野の F タームを例に説明する．機械翻訳には 5B091 という 1 つのテーマコードが, また「言語」(AA00),「処理対象要素」(AB00),「翻訳方式」(BA00)などの 9 個の観点が設けられている．ある機械翻訳システムについて考えた場合，そのシステムの対象言語は何か，どんな仕組みで翻訳するのか，などの属性が存在するが，これがそれぞれ「言語」(AA00)や「翻訳方式」(BA00)などの観点にあたると考えて良い．

F タームでは，観点が階層化されており，例えば，「言語」(AA00)という観点には，この観点を具体的に示す「・多言語間」(AA01)や「・2 言語間」(AA03)といった F タームコードが存在する．F タームコード間で一般/具体関係がある時には，ドットレベル記法で表すことになっている．図 1 の例では,「翻訳方式」(BA11)の下位分類として「直接翻訳」(BA12)と「間接翻訳」(BA13)があり，さらに「間接翻訳」の下位分類には「トランスファー方式」(BA14)と「ピボット方式」(BA17)がある．

BA11	・翻訳方式
BA12	・・直接翻訳
BA13	・・間接翻訳
BA14	・・・トランスファー方式
BA15	・・・・意味解析
BA16	・・・・・文脈解析
BA17	・・・ピボット方式

図 1 テーマ”5B091(機械翻訳)”の F タームコードの例

3.1.2. F タームからの知識抽出

本研究で構築するオントロジーでは, 3 種類の用語間の関係「上位・下位」「属性・定義域・値域」「全体・部分」を扱う．図 1 のドットレベル記法では明示されていないこれらの関係を人手で判断し，図 2 のような知識を獲得する．

関係 1	属性: 方式 定義域: 機械翻訳 値域: 直接翻訳，間接翻訳
関係 2	上位: 間接翻訳 下位: トランスファー方式
関係 3	上位: 間接翻訳 下位: ピボット方式
関係 4	属性: 利用技術 定義域: トランスファー方式 値域: 意味解析
関係 5	属性: 利用技術 定義域: 意味解析 値域: 文脈解析

図 2 図 1 から得られる知識

3.2. F タームからの知識をシードとして利用した用語間関係の獲得

Step 1 で得られた知識(用語間関係)をシードとし，F タームオントロジーには存在しない新たな知識を公開公報データベースから自動的に収集する. F タームからの用語間関係をシードとして利用した用語間関係の獲得は，以下のステップから構成される．

- Step 2-1: Espresso アルゴリズムによるパターン・インスタンスの獲得
- Step 2-2: 機械学習による Step 2-1 で得られたインスタンスのクリーニング
- Step 2-3: 新たなインスタンスが獲得されなくなるまで Step 2-1，Step 2-2 を繰り返す

図 3 に上記のステップの流れを概略図として示す．本手法におけるパターン，インスタンス獲得はパターンマッチを用いて行う．また，インスタンス抽出フェーズでは，パターンの前後に存在する名詞(句)を抽出する．Step 2-1 および Step 2-2 における詳細を次節で述べる．

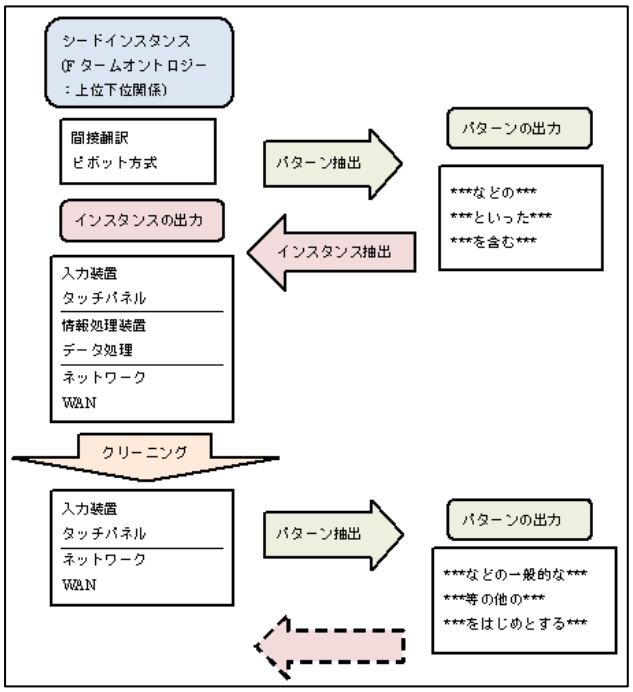

図3　F タームからの知識をシードとして利用した用語間関係の獲得の概要

3.2.1. Espresso アルゴリズムによるパターン・インスタンスの獲得

Espresso アルゴリズムにおけるシードインスタンスとして，本研究では，3.1 節で構築した F タームオントロジーから特定の関係(上位下位，部分全体，定義域属性)に属する少数の用語対を用いる．そして，そのインスタンス間に出現しているパターンを獲得し，(1)式を用いてパターンの信頼度を計算する．次に，パターンの前後に出現する用語対をインスタンスとして獲得し，パターンの信頼度と(2)式を用いてインスタンスの信頼度を計算する．その後，獲得されたインスタンスを用いてパターンを収集する．このように, F タームの体系を流用したインスタンスおよびパターンの抽出と信頼度の計算を繰り返すことで，特許との親和性を保持しながら，特定の関係に属するインスタンスを獲得することができる．

ここで, F タームからのシードインスタンスを用いて獲得するパターンについて，「X<パターン>Y」と「Y<パターン>X」という 2 種類の組み合わせから，それぞれのパターンを抽出する．例えば，F タームにおいて上位下位関係を持つ(間接翻訳，ピボット方式)というシードインスタンスから，「ピボット方式<パターン>間接翻訳」および「間接翻訳<パターン>ピボット方式」という組み合わせにより，パターンをそれぞれ抽出する．この処理を他の 2 種類の関係においても適用するため，合計 6 種類のパターン集合が獲得される．その後，そのパターンを抽出した組み合わせからインスタンスを新たに抽出する．そのため，6 種類のインスタンスが獲得される．ここで，より精巧なパターンおよびインスタンスを獲得するために，獲得されたパターン・インスタンスの選定を行う．具体的には，獲得したパターンまたはインスタンス集合において, 2 種類以上の関係に属するパターン(インスタンス)を除去する．これは，複数の関係に出現するパターンは，特定の関係のみに出現するインスタンスを発見する場合において有益でないと考えられるためである(インスタンスの選定においても同様)[17].

3.2.2. 機械学習による獲得したインスタンスのクリーニング

機械学習によるインスタンスの判別に対する概要を図 4 に示す．本研究では，各カテゴリにおけるシードインスタンスを用いて獲得したそれぞれのパターン集合を組み合わせて用語間の関係を判別する．各セル内の値は，「Y<パターン>X」という表現が公開公報データベース中で何回出現しているかを示している．これらの値の組み合わせにより，各カテゴリにおいて収集したインスタンスの用語間が統計的に成立しているかどうかを機械学習により判断する．

機械学習による用語間関係の判別を行うときに重要となることは，カテゴリ間で収集したパターンをどのように組み合わせるかということである．本研究では，各カテゴリにおいて獲得したパターンが他の同様の組み合わせによるカテゴリ内にどのくらい存在しているかを統計的に判定することに焦点を当てる．そのため，「下位－上位」「部分－全体」「属性－定義域」のカテゴリで収集したインスタンスの判定を行う場合には，これらのカテゴリ内で獲得したパターンを組み合わせる．同様に，「上位－下位」「全体－部分」「定義域－属性」のカテゴリにおけるインスタンスの判別では，これら 3 種類のカテゴリにおいて獲得されたパターンを組み合わせる．

カテゴリ	パターン	X=入力装置 Y=キーボード	X=入力部 Y=キーボード	X=入力装置 Y=発明
下位<パターン>上位	YといったX	134	6	0
	Yなどの他のX	24	1	0
	YのようにX	1	0	3
部分<パターン>全体	Y等からなるX	30	34	0
	Yで構成されたX	5	1	0
	Yを含めたX	0	1	0
属性<パターン>定義域	Yにおけるx	1	0	156
	Yにおいて、X	0	0	46
	Yにより、X	0	0	2

図 4 機械学習によるインスタンス内の用語間の関係判別

3.3. 複数言語による用語間関係の判別

本アプローチの特徴は，言語に依存していないという点である．ブートストラッピングアプローチにおけるパターン抽出では，パターンマッチにより 2 つの用語間に存在するものを抽出しており，インスタンス抽出では，パターンの前後にある名詞(句)を抽出している．また，機械学習アプローチでは，パターンとインスタンスの組み合わせの出現回数から統計的に用語間の関係を判別している．そのため，シードインスタンスとして用いる F タームを対象の言語に翻訳することで，様々な言語で記述された特許文書に対しても同様のアプローチを適用できると考えられる．

さらに，複数の言語を対象に，それぞれの抽出したパターンやインスタンスを比較することで，より正確な用語間の関係を判別することができると考えられる．例えば，「スマートフォン等の携帯端末」と「スマートフォン等のバッテリー」という文について考える．上記の 2 文は日本語公開公報に存在するため，(携帯端末，スマートフォン)だけでなく，(バッテリー，スマートフォン)に対しても上位下位関係が成立してしまう．しかし英語特許中に，「mobile computers, such as smartphones」という文が存在し，「batteries, such as smartphones」が存在しない場合，(携帯端末，スマートフォン)は上位下位関係として成立するが，(バッテリー，スマートフォン)は上位下位関係でない確率が高くなると考えられる．

ここで，どのように複数の言語間を対応付けるのかについて考える．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専門用語の訳語を推定する手法およ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[18]．本研究では，抽出したパターンやインスタンスを専門用語とみなしたとき，統計的機械翻訳モデルを用いる手法について着目する．これは，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ためである．また，統計的機械翻訳ツールには GIZA++[b]を使用し，入力言語を日本語，出力言語を英語としたとき，約 90%の精度で翻訳することができると報告されている．

本研究でシードインスタンスとして用いた F タームには日本語版を翻訳した英語版がある．これをブートストラッピングアプローチのシードインスタンスとして用いる．そして 2 種類の言語(日本語，英語)によるパターンとインスタンスをそれぞれ抽出する．そして，統計的機械翻訳モデルにより，言語間のパターンとインスタンスに対して対応付けを行う．その後，機械翻訳により，抽出したインスタンスの関係を判別する．このように，複数の言語による，より多くの根拠を利用することで，正確な用語間の自動判別を実現できると考えられる．

4. 実験

3 節で提案した手法のうち，各カテゴリにおけるシードインスタンスを用いて獲得したパターンと，機械学習によるインスタンスの判別の有効性を調べるために実験を行った．なお，本実験では，3.1 節と 3.2 節で述べた手法に対する実験のみを行った．

4.1. 実験方法

4.1.1. 実験条件

本実験のブートストラッピングアプローチにおける実験設定は以下のとおりである．

- シードインスタンスとして，F タームコードリストから，各カテゴリに属する 20 個の(名詞(句)で構成されている)用語対をそれぞれ人手で選択し，パターンを抽出する．
- インスタンス抽出フェーズにおいて，ランク付けされた上位 50 パターンを用いる．

本実験では，1 回目の反復により獲得したパターンとインスタンスを用いて提案手法の有効性を確かめる．また，機械学習に用いるパターンとして，Espresso アルゴリズムによる信頼度によってランク付けされた各パターン集合から，それぞれ上位 100 パターンを使用した(表 1)．

4.1.2. 実験データ

本実験では，以下の 2 種類のデータを用いた．

- 日本国特許全文データ：公開公報 1993-2012 年(396,532 文書，約 14GB)

[b] http://www.fjoch.com/GIZA++.html

表 1 機械学習に用いるパターンの例

下位<P>上位	部分<P>全体	属性<P>定義域
を行う	を格納する	を備える
等の他の	で構成された	を提供する
といった	等からなる	を有することを特徴とする
を保持する	として多用されている	を備えたことを特徴とする
上位<P>下位	全体<P>部分	定義域<P>属性
、すなわち	に格納されている	を実行する
、例えば、	が有する	を実現する
を用いているが、	には、複数の	の基本的な
システムにおける	が記憶する	を行うように制御する

- F タームから獲得した用語間関係リスト : 11,842 個 (102 テーマ)

特許全文データに関して，本研究では，情報分野に関連する IPC コード G06F，G06K，G06T，G11C が付与されているデータを対象に実験を行った．これに関連し，上記の IPC コードを F タームのテーマの範囲に含むものを F タームリストから抽出を行った．その結果，F ターム全 2,790 テーマ中，102 テーマが選出され，これらのテーマに関連する F タームコード 11,842 個が抽出された．

機械学習に用いる訓練用データとして，シードインスタンスを除いた F タームコードを使用した．機械学習に用いるインスタンスにおいて，それぞれが名詞(句)のみで構成されているもののみを対象とした．その結果，上位下位，部分全体，定義域属性関係において，2,719 個，664 個，415 個の F タームコードを機械学習の素性に用いた．

4.1.3. 評価方法

評価用データは，以下の手順により作成した．

1. 各カテゴリに対する反復過程においてランク付けされた 6 種類のインスタンス集合からそれぞれ上位 100 個および上位 500 件から 600 件までのインスタンスを選択する[c]．
2. 各インスタンスに対して，そのカテゴリが表す関係として本当に正しいかどうか人手で判定する．

[c] 上位に出現したインスタンスは一般的な表現のものが多く，下位にランク付けされた結果ほど特徴的なインスタンスが出現している傾向があった．そのため，各ランクの位置におけるインスタンスに対して，本手法がどのくらい性能を示すのかについて調査した．

評価尺度として，精度と再現率を用いた．また，機械学習を用いない場合をベースラインとする．

4.2. 実験結果

実験結果を表 2 に示す．ベースラインと比較すると，「下位－上位」「属性－定義域」「定義域－属性」に関して，再現率を 70-80%程度に保ちながら精度を向上させていることがわかる．この結果から，機械学習による用語間関係の判別は有効であることがわかる．しかし，部分全体関係に関して，精度が向上しないまま再現率が大幅に低下している．これは，部分全体関係を判別するような特徴的なパターンが上位に出現していないからだと考えられる．また，本実験では，機械学習に用いるパターンの数や組み合わせ方法，ブートストラッピングアプローチにおけるパラメータ(シードインスタンス数，インスタンス抽出に用いるパターンの数)を固定していたため，これらの最適な値の設定を調査する必要がある．最後に，本手法により獲得したパターンおよびインスタンスを表 3 に示す．

表 2 実験結果

上位 1-100 件	提案手法		ベースライン	
	精度	再現率	精度	再現率
下位<P>上位	0.576	0.792	0.480	1.000
上位<P>下位	0.526	0.788	0.520	1.000
部分<P>全体	0.500	0.314	0.510	1.000
全体<P>部分	0.698	0.448	0.670	1.000
属性<P>定義域	0.973	0.706	0.510	1.000
定義域<P>属性	0.893	0.848	0.790	1.000
上位 500-600 件	提案手法		ベースライン	
	精度	再現率	精度	再現率
下位<P>上位	0.513	0.848	0.460	1.000
上位<P>下位	0.400	0.829	0.410	1.000
部分<P>全体	0.417	0.227	0.440	1.000
全体<P>部分	0.529	0.383	0.470	1.000
属性<P>定義域	0.857	0.732	0.410	1.000
定義域<P>属性	0.700	0.672	0.470	1.000

4.3. 考察

前節でも述べたように，機械学習によるインスタンスの判別性能をさらに向上させるためには，そのカテゴリに対するより特徴的なパターンを用いる必要があると考えられる．例えば，「を添付した」というパターンは「部分<パターン>全体」の組み合わせからのみ獲得できる特徴的なパターンであるが，Espresso アルゴリズムによる信頼度の値は低かった．このような特徴的なパターンの信頼度を向上させるための方法として，情報利得を用いたリランキング方法が考えられる．

表 3 本手法により獲得されたインスタンス

	下位<P>上位		部分<P>全体		属性<P>定義域	
上位 1-100 件	上位：処理 下位：送信	上位：情報 下位：アドレス	全体：データ 部分：タグ	全体：テーブル 部分：データ	定義域：プログラム 属性：機能	定義域：処理 属性：手段
上位 500-600 件	上位：処理 下位：書き込み	上位：操作 下位：削除	全体：レジスタ 部分：アドレス	全体：テーブル 部分：キー	定義域：ＩＣカード 属性：手段	定義域：システム 属性：構成
上位 1-100 件	上位：入力装置 下位：キーボード	上位：処理 下位：表示処理	全体：テーブル 部分：ブロック	全体：キャッシュ 部分：アドレス	定義域：受信 属性：受信手段	定義域：アクセス 属性：手段
上位 500-600 件	上位：アドレス 下位：物理アドレス	上位：構成 下位：組合せ	全体：画面 部分：検索結果	全体：記憶部 部分：テーブル	定義域：情報処理装置 属性：形態	定義域：記憶 属性：記憶手段
	上位<P>下位		全体<P>部分		定義域<P>属性	
上位 1-100 件	上位：情報 下位：電子メール	上位：処理 下位：取得	全体：キーボード 部分：キー	全体：レジスタ 部分：値	定義域：アクセス 属性：手段	定義域：アクセス 属性：方法
上位 500-600 件	上位：情報 下位：識別情報	上位：情報 下位：識別情報	全体：メモリセルアレイ 部分：メモリセル	全体：ＲＯＭ 部分：制御プログラム	定義域：アプリケーション 属性：機能	定義域：認証 属性：手段
上位 1-100 件	上位：不揮発性メモリ 下位：フラッシュメモリ	上位：コンピュータ 下位：サーバ装置	全体：チャネルＭＯＳトランジスタ 部分：ゲート	全体：プリンタ 部分：印刷データ	定義域：移行 属性：手段	定義域：バックアップ 属性：機能
上位 500-600 件	上位：情報 下位：個人情報	上位：情報 下位：購入者	全体：フラッシュメモリ 部分：ブロック	全体：制御部 部分：制御プログラム	定義域：リフレッシュ 属性：手段	定義域：ジョブ 属性：要求

情報利得を用いることで，特定のパターンに対するカテゴリへの曖昧さ(エントロピー)を把握し，曖昧さが少ないパターンを選定することができると考えられる．表 4 に，3 種類のカテゴリ(「下位－上位，部分－全体，属性－定義域」または「上位－下位，全体－部分，定義域－属性」)を対象とした各カテゴリにおけるパターンのリランキング結果の例を示す．なお，リランキング方法として，Espresso アルゴリズムによる信頼度と情報利得値を掛け合わせ，値の高い順に並べ替えている．また，これらのパターンをインスタンス抽出フェーズに用いることで，Espresso アルゴリズムによるパターン集合だけでは獲得できない特徴的なインスタンスを抽出することができると考えられる．

表 4 情報利得による各カテゴリ内のパターン集合のリランキング結果

下位<P>上位	部分<P>全体	属性<P>定義域
システムの	として添付された	により生成された
手段の	を添付した	を通じて
に代わる	を除いた当該	にて
装置及びその	を除いた前記	により、前記
上位<P>下位	全体<P>部分	定義域<P>属性
を行い、	に添付されている	を受け付ける
部における	に添付された	の構成を示す
は従来の	から分離された	の一実施例を示す
手段による	に記憶された	の一実施例の

5. おわりに

本研究では，特許データベースを対象に，ブートストラッピング法と機械学習を組み合わせた手法を提案した．本手法では，F タームに基づいたオントロジーをシードインスタンスとして使用しており，特許との親和性を保持した新たな用語間関係を獲得できることを示した．また，本手法は言語に非依存である．今後は米国特許を対象とした実験を行い，統計的機械翻訳を用いて 2 種類の言語間を対応付けることで正確な用語間の自動判別を行うことを目指す．

参考文献

1. Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora, *Proceedings of the 14th International Conference on Computational Linguistics*, pp. 539-545, 1992.
2. 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.
3. Nanba, H.: Query Expansion using an Automatically Constructed Thesaurus, *Proceedings of the 6th NTCIR Workshop Meeting*, pp. 414-419, 2007.
4. Kozareva, Z. and Hovy, E.: Learning Arguments Supertypes of Semantic Relations using Recursive Patterns, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1482-1491, 2010.
5. 安藤まや, 関根聡: 上位語・下位語を含む連体修飾表現の言語的分析, 言語処理学会第 10 回年次大会, 2004.
6. Kozareva, Z. and Hovy, E.: A Semi-Supervised Method to Learn and Construct Taxonomy using the Web, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1110–1118, 2010.
7. Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics (SCL'95)*, pp. 189-196, 1995.
8. Abney, S.: Understanding the Yarowsky algorithm, *Journal of Computer Linguistics*, Vol. 30, No. 3, pp. 365-395, 2004.
9. Krishnan, A., Cardenas, A.F. and Springer, D.: Search for Patents using Treatment and Causal Relationships, *Proceedings of the 3rd International Workshop on Patent Information Retrieval*, 2010.
10. 小町守, 鈴木久美: 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, 2008.
11. 伊藤淳, 戸田浩之, 廣嶋伸章, 望月崇由, 鈴木智也, 筧捷彦: クエリログをコーパスとした意味知識獲得法の改善, 第2回データ工学と情報マネジメントに関するフォーラム(DEIM2010), 2010.
12. Abe, S., Inui, K. and Matsumoto, Y.: Acquiring Event Relation Knowledge by Learning Co-occurrence Patterns and Fertilizing Co-occurrence Samples with Verbal Nouns, *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pp. 497-504, 2008.
13. Kiso, T., Shimbo, M., Komachi, M. and Matsumoto, Y.: HITS-based Seed Selection and Stop List Construction for Bootstrapping, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2, pp. 30-36, 2011.
14. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Enviroment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604-632, 1999.
15. Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, pp. 113-120, 2006.
16. Girju, R., Badulescu, T. and Moldovan, D.: Automatic Discovery of Part–Whole Relations, *Journal of the Computational Linguistics*, Vol. 32, No. 1, pp. 83-135, 2006.
17. Curran, J.R., Murphy, T. and Scholz, B. Minimising.: Semantic Drift with Mutual Exclusion Bootstrapping, *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, pp. 172-180, 2007.
18. 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.