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，最小限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，ブートストラップ法と機械学習を組み合わせた手法を用いることで，特許との親和性を保持したオントロジーの構築を目指す．

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]．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[4]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[5]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，(X, Y)で表されるインスタンスやその関係の判別に有用なパターンを自動的に抽出する半教師あり学習アルゴリズムであるブートストラップ法[6]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. ブートストラップ法

近年では，ブートストラップ法に基づいた学習アルゴリズムとして, Espresso アルゴリズムが提案されている[7]．このアルゴリズムでは，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．本研究では, Espresso アルゴリズムによるブートストラップ法を用いることで，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．

2.3. 機械学習による関係判別

Girju ら[8]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，ブートストラップ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)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

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	属性: 利用技術 定義域: トランスファー方式 値域: 意味解析

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

本研究の目標は，3.1 節で構築した 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から自動的に収集することである．図 3 にシステムの構成を示す．シードインスタンスには，Step 1 で得られた知識を用いる．パターン抽出では，用語間に存在する表現をパターンとして抽出し，インスタンス抽出では, パターンの前後に存在する名詞(句)を抽出する．

3.2.1 節では，パターンおよびインスタンスの抽出について，3.2.2 節では，獲得したインスタンスに対する，機械学習によるクリーニングについて述べる．

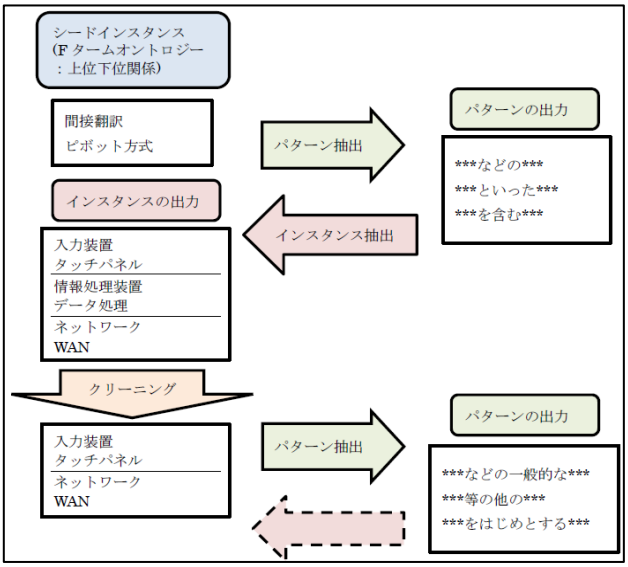

図 3: F タームからの知識をシードとして利用した用語間関係の獲得の概要

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

まず, Espresso アルゴリズムにおけるシードインスタンスとして，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．次に，そのインスタンス間に出現しているパターンを獲得し，Espresso アルゴリズムにより，パターンの信頼度を計算する．その後，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，信頼度を計算する．このように, F タームの体系を流用したパターンおよびインスタンス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，高精度なインスタンスを獲得することができる．

ここで本研究では，パターンの獲得について，図 4 に示すように,「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パターンを抽出する．例えば，(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，そのパターンの前後に出現する用語対を，インスタンスとして新たに抽出する．そのため，合計 6 種類のインスタンスが獲得される．最終的なインスタンス獲得では，カテゴリにおいて，「X<パターン>Y」と「Y<パターン>X」で獲得したインスタンス集合を，Espresso アルゴリズムにより算出された信頼度により足し合わせ，値が高い順に並び替える．

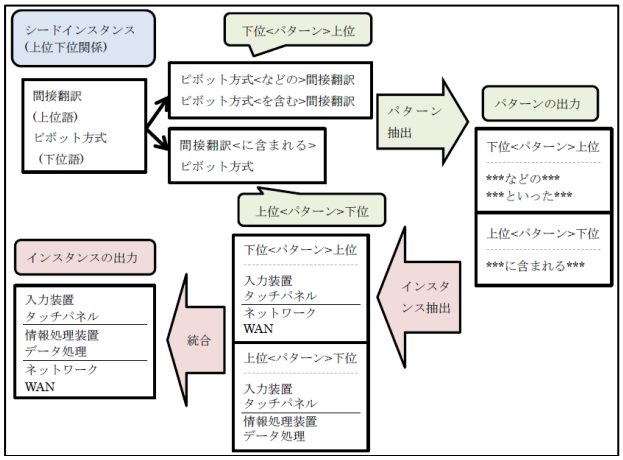

図 4: パターン・インスタンス抽出の詳細な概要

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 5 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」(または「X<パターン>Y」)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．また，機械学習に用いるパターンとして，直前のパターン抽出フェーズで獲得した 6 種類の集合を組み合わせる．

本研究のタスクでは，1 つのインスタンスに対して複数の用語間関係は存在しないと定めている．そのため本研究では，各カテゴリを判別する分類器を作成し，複数の分類器から正例と判断されたインスタンスを対象に，それぞれの超平面の距離を比較する．そして，最も計算結果が高かった分類器の表す用語間関係を，そのインスタンスの関係として判別する．

4. 実験

4.1. 実験方法

4.1.1. 実験条件

本研究の具体的な実験設定は以下のとおりである．なお，(3)において，本研究では，抽象的な用語や不要語を含むインスタンスは人手で除去した[a]．

[a] (データ，プログラム)や(情報，データ)といった抽象的なインスタンスに対して，その関係を人手で判別することは難しく，また，実際の特許検索でもほとんど使用されないと考えられる．

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 などの X	543	10	1
	Y といった X	134	6	0
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	9	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
上位<パターン>下位	X に含まれる Y	207	3	2
	X（例えば、Y	58	1	0
全体<パターン>部分	X が有する Y	4	23	3
	X を構成する Y	3	18	0
定義域<パターン>属性	X の具体的な Y	0	0	86
	X を実現する Y	0	0	28

図 5: 機械学習による用語間の関係判別

(1) シードインスタンスとして，F タームコードリストから，各カテゴリにおける 20 個の用語対を人手で選択する．
(2) パターン抽出で獲得した上位 50 個のパターンを用いてインスタンスを抽出する．
(3) インスタンス抽出で獲得した 3 種類のインスタンス集合から，それぞれ上位 100 個を選択する．
(4) 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を機械学習により判定する．

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)
- 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国際特許分類(IPC)コード G06F，G06K，G06T, G11C が付与されているデータを対象にした．これに関連し，上記の IPC コードを含む F タームテーマの範囲のものを F タームから抽出を行っている．

機械学習に用いる訓練用データとして, F タームから獲得した用語間関係リストにおいて，それぞれが名詞(句)のみで構成されているインスタンスのみを用いた．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．

評価用データには，4.1.1 節で述べた手順 3 で獲得した，各カテゴリにおける 100 個のインスタンス集合を用いた．また，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した．

4.1.3. 評価方法

以下に示す 2 種類の提案手法とベースラインにより，実験を行った．評価尺度には，精度と再現率を用いた．また，各手法により抽出したインスタンス数および実際に人手で正しいと判断した数を調べた．

提案手法

- **ESP+ML(*n*)**: Espresso アルゴリズムにより獲得したインスタンス集合を対象に，機械学習により，そのカテゴリが表す関係として本当に正し

いか判定する．n は，6 種類のパターン集合からそれぞれ上位 n 件を使用することを示す．

ベースライン

- **ESP**: 各カテゴリに対して，Espresso アルゴリズムにより獲得したインスタンス集合は，全て正しい関係によるものとみなす．

ESP+ML(n)手法において，本研究では，6 種類のパターン集合からそれぞれ上位 10 個から 100 個まで 10 刻みで使用し，どのように性能が変化するのか調べた．機械学習には TinySVM を用い，線形カーネルを使用した．また，機械学習による学習時間を考慮し，出現頻度に対する対数を求め，素性値とした．さらに，評価用データにおける素性を作成する際，2GB の特許全文データを用いた．

4.2. 実験結果と考察

各手法により獲得したインスタンスに対する実験結果を図 6 から図 8 に示す．それぞれの図による結果を比較すると，全てのカテゴリにおいて，機械学習を用いた場合，ベースラインより高い精度を示すことが分かった．特に，上位下位関係において，再現率を 85%以上保持しながら精度を改善しており，獲得できるインスタンスの数も 80 個以上と高い値を示している．定義域属性関係においても，高い再現率を保ちつつ，精度を 50%から最大 1.5 倍まで改善している．一方で，部分全体関係では，精度が向上したが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．しかし本研究では，より正確なインスタンスを獲得することを目的としている．そのため，1 回目の反復で獲得できなかったインスタンスは，機械学習により高い精度で判別したインスタンスを用いて，新たな反復で獲得すれば良いと考えられる．最後に，精度と正解数を考慮し，最も性能の高かった ESP+ML(90)手法を適用して獲得したインスタンスを表 1 に示す．

5. おわりに

本研究では，特許データベースを対象に，ブートストラップ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．

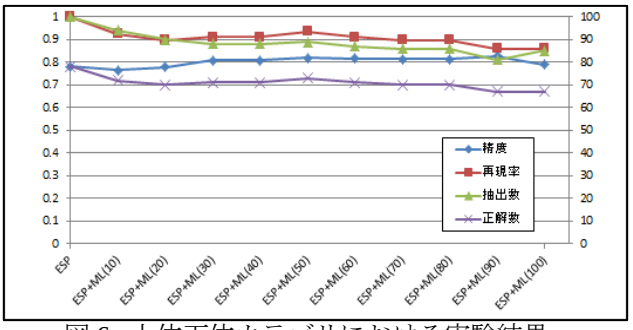

図 6: 上位下位カテゴリにおける実験結果

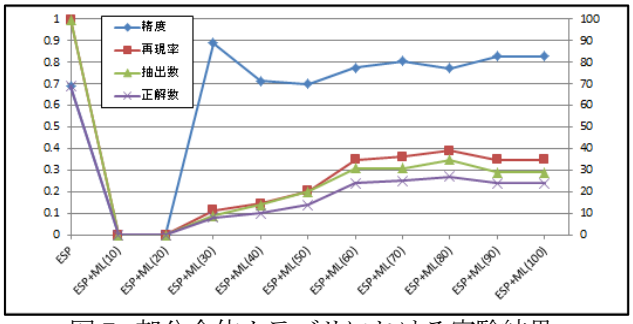

図 7: 部分全体カテゴリにおける実験結果

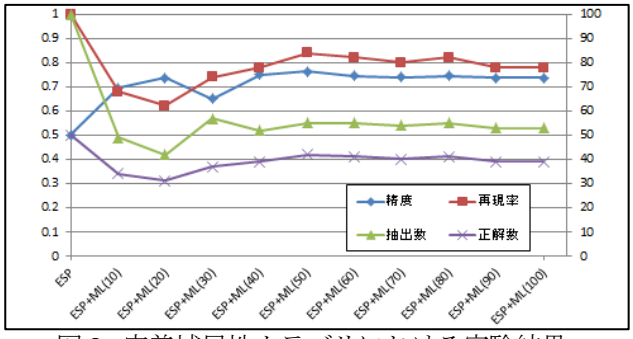

図 8: 定義域属性カテゴリにおける実験結果

表 1: 本手法により獲得されたインスタンス

カテゴリ	インスタンス
上位下位	(ポインティングデバイス，マウス) (入力装置，マウス) (入力装置，キーボード) (識別情報，ユーザ ID) (識別情報，パスワード) (記録媒体，IC カード) (表示装置，LCD)
部分全体	(プロセッサ，キャッシュメモリ) (記憶装置，記憶領域) (情報処理装置，入力装置) (IC カード，メモリ) (IC カード，IC チップ) (ページ，リンク) (メモリセル，トランジスタ)
定義域属性	(書き込み，手段) (最適化，方法) (サービス，手段) (サービス，形態) (情報処理装置，発明) (情報処理装置，形態) (タッチパネル，入力手段) (プリンタ，設定)

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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.