

言語処理分野の最前線 3
特許オントロジーの構築と
その特許検索への応用

難波 英嗣

広島市立大学大学院情報科学研究科 准教授

特許オントロジーの構築と その特許検索への応用

2015年2月24日

広島市立大学大学院情報科学研究科
難波英嗣

特許検索履歴の例

H01L23/00@A*[製品管理情報/TX+??製品管理
情報+特性/TX+??特性+性能/TX+??性能]*[バー
コード/TX+??バーコード]

上位—下位

対象—属性

研究の目的

オントロジーとは

文献検索や高度な言語処理を行なうための有用な情報源

オントロジー構築には非常にコストがかかる

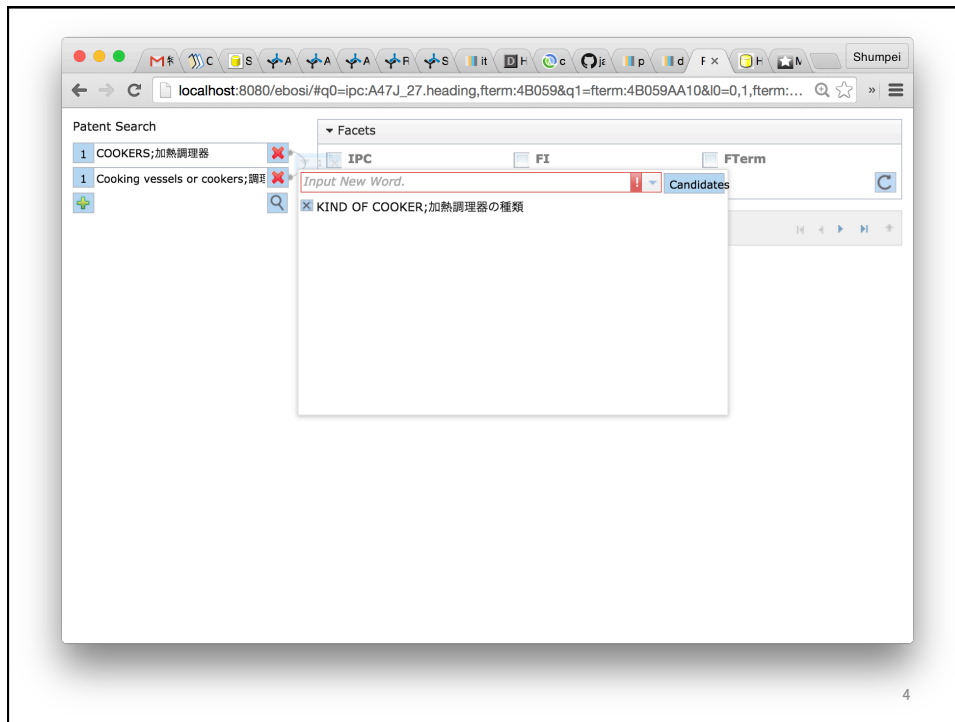
- 効率的なオントロジー構築手法の提案
- 特許検索への応用

2

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)

3



基本方針

オントロジー構築のためのFタームの利用

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

Fタームの構造そのものがオントロジーに近い体系→Fタームの体系をオントロジーの構築に流用

これに、テキストデータベースから自動構築された各種辞書をマッピングすることで、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

オントロジー構築のイメージ

[入力]

間接翻訳 — ピボット方式 上位下位
 間接翻訳 — トランスファー方式 上位下位
 ...

Fタームオントロジーを
入力とし

ブートストラップ+
機械学習により

[出力]

情報機器 — イメージスキャナ 上位下位
 コンピュータ機器 — パソコン 上位下位
 ...

Fタームオントロジーにはない
新たな知識を公開公報から抽出

テーマ“5B091(機械翻訳)”の Fタームコードの例

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

ドット表記には、上位・下位や
属性などの関係が含まれている。
その関係を明らかにし、オントロジー
構築のための知識抽出を目指す

Fタームから得られる知識

BA11 ・ 翻訳方式
BA12 ・ ・ 直接翻訳
BA13 ・ ・ 間接翻訳

属性: 方式
定義域: 機械翻訳
値域: 直接翻訳、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

テーマ“5B091(機械翻訳)”のFタームコードの例₈

Fタームから得られる知識

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

上位: 間接翻訳
下位: トランスファー方式、
ピボット方式

テーマ“5B091(機械翻訳)”のFタームコードの例₉

Fタームから得られる知識

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	属性: 利用技術
BA17	・ ・ ・ ・ ・ ピボット方式	定義域: トランスファー方式 値域: 意味解析

テーマ“5B091(機械翻訳)”のFタームコードの例₁₀

Fタームから得られる知識

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	属性: 利用技術 定義域: 意味解析 値域: 文脈解析
BA17	・ ・ ・ ・ ・ ピボット方式	

テーマ“5B091(機械翻訳)”のFタームコードの例₁₁

オントロジー構築のイメージ

[入力]

間接翻訳 — ピボット方式 上位下位
 間接翻訳 — トランスファー方式 上位下位
 ...

Fタームオントロジーを
入力とし

ブートストラップ+
機械学習により

[出力]

情報機器 — イメージスキャナ 上位下位
 コンピュータ機器 — パソコン 上位下位
 ...

Fタームオントロジーにはない
新たな知識を公開公報から抽出

(ステップ1) ブートストラップ

パターン抽出とインスタンス獲得を繰り返して、少量のシードインスタンスを反復的に増やす

初期シード

半導体メモリ
ROM

ROMなどの半導体メモリ

パターン抽出

などの

入力装置
キーボード

キーボードなどの入力装置
パソコンなどのキーボード

キーボード
パソコン

インスタンス
獲得

キーボードをはじめとする入
力装置

***をはじめと
する***

なぜ機械学習を用いるのか

パターン:「XなどのY」

上位下位関係の判別
に代表されるパターン

キーボードなどの入力装置

パソコンなどのキーボード

公開公報
データベース

入力装置 — キーボード 上位下位

キーボード — パソコン 上位下位



機械学習による統計的アプローチを
用いることで解決する

14

(ステップ2) 機械学習による クリーニング

Espressoで収集されたパターン

パターン ID	パターン (下位・上位)
1	XなどのY
2	Xに代表されるY
3	XといったY
4	X等のY
5	XのようなY

今, $(X, Y) = (\text{キーボード}, \text{入力装置})$ および $(\text{パソコン}, \text{キーボード})$ を入力とした時, XとYの間に上位, 下位関係が成立するか, Espressoで収集されたパターン集合を用いて調べたい

15

(ステップ2) 機械学習による クリーニング

Espressoで収集されたパターン

パターン ID	パターン (上位・下位)	X=キーボード Y=入力装置	X=パソコン Y=キーボード
1	XなどのY		
2	Xに代表されるY		
3	XといったY		
4	X等のY		
5	XのようなY		

「キーボードなどの入力装置」や
「キーボードに代表される入力装置」
という表現が公開公報データベース
の中に何回出現するか調べる

16

(ステップ2) 機械学習による クリーニング

これらの値の組み合わせでXとYの間に
上位・下位関係が成立するか判断する

Espressoで収集されたパターン

パターン ID	パターン (上位・下位)	X=キーボード Y=入力装置	X=パソコン Y=キーボード
1	XなどのY	1050	12
2	Xに代表されるY	320	0
3	XといったY	12	0
4	X等のY	1300	100
5	XのようなY	85	1

17

(ステップ2) 機械学習による クリーニング

これらの値の組み合わせでXとYの間に
上位・下位関係が成立するか判断する

Espressoで収集されたパターン

パターン ID	パターン (上位・下位)	X=キーボード Y=入力装置	X=パソコン Y=入力装置
1	XなどのY		
2	Xに代表されるY		
3	XといったY		
4	X等のY	1300	100
5	XのようなY	85	1

どのように組み合わせるかは機械学習
(SVM)で決める

18

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

19

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)

20

結論

特許の分類体系のひとつであるFタームを用い、
これに様々な知識体系を獲得することでオント
ロジーを構築する手法を提案

さらに、特許検索に応用

21