

ラベルなしデータからの 意味カテゴリタガールの学習

村本 英明 鍛冶 伸裕

吉永 直樹 喜連川 優

東京大学

研究の目的

- ウェブテキスト中の名詞(句)に対して適切な意味カテゴリをタグ付け
 - 例
 - イギリス<国>人科学者のニュートン<人>
 - 今月発売されたニュートン<出版物>の特集記事
 - 語義の曖昧性解消、情報抽出、機械翻訳等の基礎技術
 - 従来は新聞記事を対象に固有表現認識の枠組み研究されてきた

ウェブテキストを扱うための問題点

- 新聞記事とは異なる意味カテゴリを扱う必要性



人物, 場所,
組織, ...

自民党の渡辺美智雄元副総理・
外相は三日、栃木県鹿沼市で講
演し、「憲法...」と述べ、時代に応
じた改憲が必要との考えを示した。

新聞記事



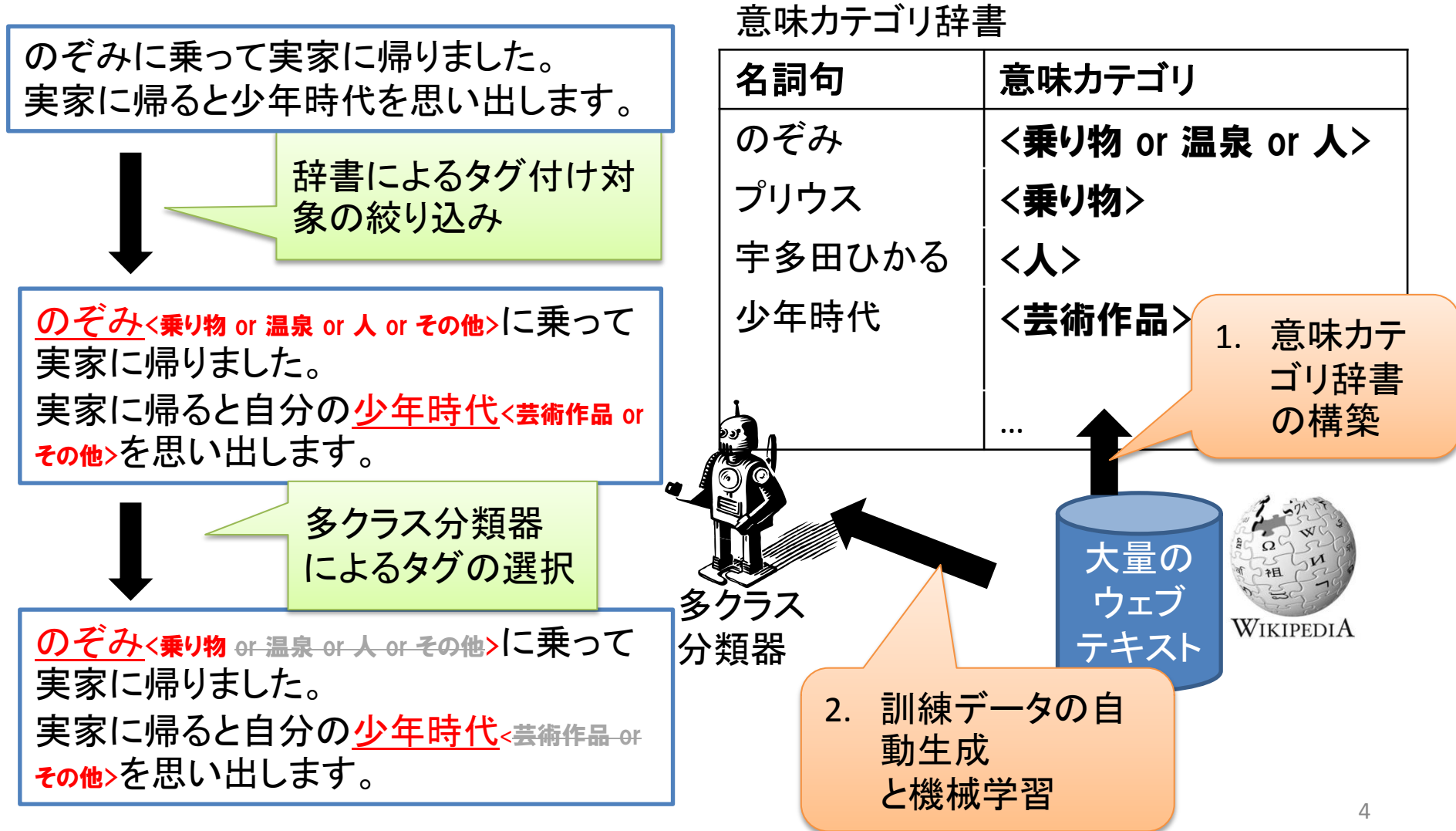
芸術作, 施設,
会社, 路線
乗り物, 食べ物...

日曜日は池袋のシネマサンシャ
インで踊る大捜査線を観てきま
した。織田雄二さんいいですね。
そのあと西口のマックで休憩。
ビックマック安いですね～

ウェブテキスト

- 教師あり学習は適用困難
 - － 新たに訓練事例を作成するコスト大

自動構築した辞書とウェブテキストからのタガー構築



1. 意味カテゴリ辞書の構築法(1)

ウィキペディアの記事カテゴリからの抽出

- 意味カテゴリとウィキペディアの記事カテゴリとの対応関係を人手で記述し、見出し語を抽出

＜芸術作品＞

意味カテゴリ

対応関係を人手で記述

記事カテゴリ

記事カテゴリ

取得できる名詞句

ウィキペディア フリー百科事典 メインページ コミュニティ・ポータル 最近の出来事 新しいページ 最近の更新 おまかせ表示 練習用ページ アップロード (ウィキメディア・コモンズ) ヘルプ 井戸端 お知らせ バグの報告 寄付 ウィキペディアに関するお問い合わせ ツールボックス 他の言語 Cesky Dansk English Español Euskara فارسی Suomi Français Հայերեն 한국어 Norsk (bokmål) Português Română Русский Slovenščina	Category:2000年の楽曲 2000年にリリースされた楽曲に関するカテゴリ。 サブカテゴリ このカテゴリへは次の1サブカテゴリしか属していません。 し • [×] 2000年のシングル (0) カテゴリ「2000年の楽曲」にあるページ 以下にこのカテゴリへ属しているページ 17 件中 17 件を表示しています。 * 2000年の音楽 あ • 新しい歌 (合唱) • ア・ミリオン・ライト・イヤーズ・アウェイ い • It's a Mystery • 生命の話をしよう う • うめばしのうた	さ • 桜 (コブクロの曲) じ • Just Be せ • Saywhatchugatta で • Deep Within と • 東京の女 (曲)	は • バンティング ひ • Been so long ふ • Fly away (J.S) • Flo jack へ • HEMAの鐘	ウィキペディア フリー百科事典 メインページ コミュニティ・ポータル 最近の出来事 新しいページ 最近の更新 おまかせ表示 練習用ページ アップロード (ウィキメディア・コモンズ) ヘルプ 井戸端 お知らせ バグの報告 寄付 ウィキペディアに関するお問い合わせ ツールボックス 他の言語 Cesky Dansk English Español Euskara فارسی Suomi Français Հայերեն 한국어 Norsk (bokmål)	Category:2001年の楽曲 2001年にリリースされた楽曲に関するカテゴリ。 サブカテゴリ このカテゴリへは次の1サブカテゴリしか属していません。 し • [×] 2001年のシングル (0) カテゴリ「2001年の楽曲」にあるページ 以下にこのカテゴリへ属しているページ 10 件中 10 件を表示しています。 * 2001年の音楽 う • Will be there ~恋心 え • Endless Moment き • 君がいないと	きの続き • キャラメルダンス こ • 言葉にできない は • バタフライズ む • Moon Clock
---	---	--	--	--	--	---

1. 意味カテゴリ辞書の自動構築法(2)

ウェブからの語彙統語パターンによる抽出

ウェブテキスト

テレビに歌手「宇多田ひかる」が登場した。
少年時代という曲が頭の中に浮かんできた。
プリウスなどの車はエコポイント対象と違います

語彙統語パターン

<人>抽出パターン

- ・Aという{歌手, 芸人, 俳優, ...}
- ・Aなどの{歌手, 芸人, 俳優, ...}
- ・{歌手, 芸人, 俳優, ...}「A」

<芸術作品>抽出パターン

- ・Aという{映画, 曲, テレビ番組, ...}
- ・Aなどの{映画, 曲, テレビ番組, ...}
- ・{映画, 曲, テレビ番組, ...}「A」

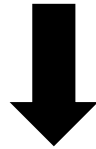
抽出される意味カテゴリ辞書

名詞句	意味カテゴリ
のぞみ	<乗り物 or 温泉 or 人>
プリウス	<乗り物>
宇多田ひかる	<人>
少年時代	<芸術作品>
...	...

2. 訓練データの自動生成 ナイーブな手法とその問題点

- カテゴリの曖昧性のない名詞句から訓練データを生成

のぞみに乗って実家に帰りました。
実家に帰ると少年時代を思い出します。
家族はプリウスに乗って遊びに出かけてました。



のぞみに乗って実家に帰りました。
実家に帰ると少年時代<芸術作品>を思い出します。
家族はプリウス<乗り物>に乗って遊び<芸術作品>に出かけてました。

複数の意味カテゴリを持つためラベル付けしない

一般語としても用いられるため、誤ってラベル付けされている例

意味カテゴリ辞書

名詞句	意味カテゴリ
のぞみ	<乗り物 or 温泉 or 人>
プリウス	<乗り物>
宇多田ひかる	<人>
少年時代	<芸術作品>
遊び	<芸術作品>
...	...

2. 訓練データの自動生成

訓練データの精度向上に関する検討

- 語彙統語パターンによる精度の高いラベル付け

のぞみに乗って実家に帰りました。
少年時代<芸術作品>という曲は良い曲ですね。
家族はプリウスに乗って遊びに出かけてました。

語彙統語パターンに
マッチしているのでラ
ベル付けする

語彙統語パターンに
マッチしていないので
ラベル付けしない

- 検討中の手法
 - 名詞句の係り受け先のエントロピーの利用
 - ウィキペディア記事の本文との類似度の利用

予備実験の設定

- コーパス
 - 2006年～2009年の4年分、20億文のブログ記事
- ルール
 - ウィキペディア:181個のルール
 - 語彙統語パターン:585個のパターン
- 意味カテゴリ辞書と訓練データの精度を検証
(分類器の学習は今後の課題)

ウィキペディアから得られた辞書と 語彙統語パターンで得られた辞書の比較

- 意味カテゴリ辞書の比較

	ウィキペディア	語彙統語パターン
精度	高	低
特徴	人名, 地名がメイン	芸術作品名(映画, TV 番組, 本)がメイン

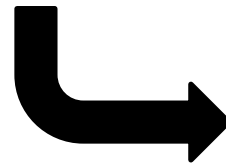
- 語彙統語パターンで新規に取得できたエントリ例

名詞句	意味カテゴリ
八海山	<食べ物>
グーグル革命の衝撃	<芸術作品>
ホームランバー	<食べ物>
〇〇野球同好会	<競技組織>

語彙統語パターンによって 誤って獲得された辞書エントリ

抽出元のテキスト

ニュースでアナウンサー「あけましておめでとう」を囃んで...
とてもこの監督「らしい」作品です
打撃が持ち味という選手...
「人材不足がボトルネック」という企業は
ロックの神様「エルビス・プレスリー」...
個性派女優という路線ですか？



誤って獲得された辞書エントリ

名詞句	意味カテゴリ
あけましておめでとう	<人>
らしい	<人>
人材不足がボトルネック	<法人>
エルビス・プレスリー	<神>
個性派女優	<路線>

訓練データの自動生成の誤り例

- 意味カテゴリ辞書が不十分である場合の誤り
 - でも、レッズも、ヤマ<神>からポンテにパスが通って、同点！
 - サチはコックス<法人>という、漕ぎ手に指示をする係り。
- 一般語との混在による誤り
 - 硫黄島で日本軍の指揮を執った栗林忠道中將が少年期<芸術作品>に書いた日記が...生家から見つかった
 - この街<芸術作品>にはいろんな思い出がある。
 - 虎連勝、M40<天体>が点灯

まとめと今後の課題

- 意味カテゴリタガーをラベルなしデータから学習する手法の提案を行った
- 意味カテゴリ辞書と訓練データの自動生成手法についての予備実験を行った
- タグ付け手法の検討, より高精度で大規模な訓練データを自動生成する手法の検討, タガーの評価手法の検討が今後の課題である

意味 カテゴリ	意味カテゴリ辞書				訓練データの自動生成				記述したルールの例	
	エントリ数		適合率		ラベル数		適合率		ウィキペディア記事カテゴリとの 対応関係	カテゴリ語
	wiki	blog	wiki	blog	語彙統語 パターン	フィルタ なし	語彙統語 パターン	フィルタ なし		
人	196,812	23,918	1	0.75	31,199	6881081	1	0.85	*年生	選手, 芸人, 歌手, 俳優, 女優, ...
神	3,031	4,301	1	0.5	1,537	197849	1	0.5	神, *の神	神, 神様
国際組織	148	159	1	0.95	21	2524	1	0.9	国際機関, 国際専門機関	国際組織, 国際機関
公演組織	726	0	1	NA	0	20610	1	1	*のオーケストラ	公演組織
家系	33	169	0.8	0.1	0	7425	1	0.05	*の士族, *の家族	家系
民族	0	776	1	0.1	0	0	1	NA	NA	民族, 州続, 国籍
競技組織	4,875	9,474	1	0.9	3,400	235137	1	1	*のサッカークラブ, *野球チーム...	競技組織, チーム, サッカークラブ, リーグ
法人	37,192	32,767	1	0.8	43,426	4336706	1	0.9	*法人, *の企業, *の企業グループ	法人, 社団法人, 企業, 会社, 株式会社...
政治的組織	3,021	1,668	1	0.75	1,198	317539	1	0.5	*の官公庁, *の政党, *の政治団体...	派閥, 政府, 政党, 内閣, 軍隊
温泉	1,310	3,454	1	0.85	1,090	132784	1	0.9	*の温泉	温泉
GPE	5,153	26,798	1	0.7	2,169	163323	1	1	*の町・字, *の群, *の都道府県, *の州...	市, 区, 町, 村, 都市, 都, 道, 府, 県, 州...
地域	158	7,506	1	0.85	16	66639	1	1	大陸, *の地方	地域, 地区, 大陸, 地域
地形	9,700	10,009	1	0.95	7,583	655816	1	1	*の山地, *の山, *の河川...	山, 島, 川, 河, 湖, 海, 湾
天体	355	793	1	0.8	1,023	11740	1	0.2	恒星, *の惑星, 88星座	天体, 恒星, 惑星, 星座
遺跡	1,071	817	1	0.85	119	10442	1	1	*の考古遺跡, *の古墳	遺跡, 古墳
GOE	38,732	26,064	1	0.85	11,505	2364355	1	1	*の大学, *の高等学校, *の小学校...	施設, 公共機関, 学校, 研究機関, 研究所...
路線	10,245	4,366	1	0.55	2,028	559937	1	1	*の鉄道路線, *の道路, 運河...	路線, 道路, 運河, 航路, トンネル
材料	741	3,392	1	0.9	278	551726	1	0.75	材料, *材料	燃料, 物質, 染料, 原材料
衣類	267	190	1	0.95	26	233711	1	0.95	装飾具, スポーツウェア, 履物...	衣類
貨幣	287	116	1	0.7	37	102646	1	0.85	硬貨, 金貨, 銀貨, *の硬貨	貨幣
医薬品	360	3,628	1	0.95	930	86608	1	1	一般用医薬品, 漢方薬, ...	薬, 医薬品
武器	3,610	3,392	1	0.4	433	404806	1	0.65	*兵器, *ミサイル	武器

意味 カテゴリ	意味カテゴリ辞書				訓練付きデータの自動生成				記述したルールの例	
	エントリ数		適合率		ラベル数		適合率		ウィキペディア記事カテゴリとの 対応関係	カテゴリ語
	wiki	blog	wiki	blog	語彙統語 パターン	フィルタ なし	語彙統語 パターン	フィルタ なし		
賞	1,272	1,380	1	0.85	733	63804	1	1	*の賞	賞
勲章	216	339	1	0.45	37	3877	1	1	*の勲章	勲章
罪	369	1,279	1	0.8	882	271730	1	1	犯罪	罪
キャラクター	1,421	11,611	1	0.85	2,334	228206	1	0.85	*キャラクター	キャラクター
乗り物	10,314	9,596	1	0.8	13,258	702851	1	0.65	*の車種, *の列車,	乗り物, 車, 自動車, 列車, 電車, ...
食べ物	3,190	21,112	1	0.95	4,527	3235723	1	0.85	ソフトドリンク, カクテル, 果実酒, 蒸留酒...	ソフトドリンク, カクテル, 果実酒, 蒸留酒...
芸術作品	36,918	265,111	1	0.95	739,210	13699253	1	0.05	*の美術館, 絵画作品, *テレビ番組...	芸術作品, 作品, 絵画, 絵, 番組, 映画...
出版物	2,751	23,692	1	1	51,229	337015	1	0.85	*の新聞, *年創刊の雑誌	出版物, 新聞, 雑誌
主義方式	14,326	13,332	1	0.6	10,160	5186531	1	0.75	*の文化, *の宗教, *科学...	主義, 文化, 宗教, 学問, 競技...
規則	2,415	3,702	1	0.6	3,898	246091	1	1	条約, *の法律	規則, 条約, 法令, 法律
称号	440	8,630	1	0.9	2,841	336741	1	0.75	職業	称号, 地位, 職業
言語	1,746	1,121	1	0.8	1,005	151711	1	0.85	*の方言, 語族, *の言語	言語, 言語
単位	1,410	2,984	1	0.3	2,182	937879	1	0.6	*単位	単位, 通貨
催し物	5,084	7,645	1	0.9	1,418	138928	1	0.95	*の祭, *大会, 議会, *の議会	催し物, 祭り, 祭, 競技会, 大会, 会議
事故事件	915	4,743	1	0.5	81	86517	1	0.7	*の戦い, 戦争	事件, 事故, 戦争
自然災害	879	813	1	0.8	150	45553	1	1	*の災害, *の地震	災害, 地震
元素	510	119	1	0.85	181	11064	1	0.95	元素	元素
化合物	2,340	209	1	1	115	88583	1	1	*の化合物, 化学物質	化合物
鉱物	505	309	1	0.9	251	17446	1	1	鉱物	鉱物
生物	3,897	10,147	1	0.9	3,333	887168	1	0.95	菌類, 節足動物, イカ, タコ...	生物, 菌類, 軟体動物...
生物部位	78	0	1	NA	0	89132	1	0.9	器官	NA
動物病気	332	4,863	1	0.95	1,434	233052	1	1	病気	病気
自然色	375	5,365	1	0.8	5,127	240952	1	0.95	色名	色
合計	561,859	409,530			952,404	44,583,213				