鳥式改の上位語データの人手クリーニング

黒田 航　李 在鎬　野澤 元　村田真樹　鳥澤 健太郎

NICT

2009/03/02

言語処理学会第15回年次大会, 鳥取大学

作業の目的と内容

- 目的
  - Sumida & Torisawa (2008) が日本語 Wikipedia から自動獲得した上位語/下位語対 (約240万個) の上位語集合の整備 (と体系化の下準備)
- 内容
  - 成語性の低い上位語の除去と非飽和名詞句の区別
  - 大規模な固有名辞書をシソーラス (e.g., Bondら(2008) *WordNet-Ja*) =上位オントロジーと接続するための下準備

元データ (断片)

- 現役選手: マット・モリス
- 大阪府出身の人物: 金森又一郎
- 過去に在籍した選手/監督: 船越優蔵
- ヒノキ科: ミヤマビャクシン
- キャスト: 立花大介
- 船: 将
- 日本の法学者: 小菅成一
- アニメ作品: 魔法遊戯
- 日本のインターチェンジ: 利府塩釜インターチェンジ
- これまでの代理司会者: Mr. マリック
- 作品: あくまこあくま
- 架空の惑星: バース星
- 中堅メーカー: 宮島醤油
- 都市及び町: ジョージアナ
- 小惑星: 菅野洋子
- 他著: 改訂電子回路
- 出演作品: 華麗な休暇
- 友好都市: 島根県松江市

要件の定義 1/2

- <競技のチーム>の現役選手: マット・モリス
- 過去に<競技のチーム>の在籍した選手 OR 過去に<競技のチーム>の在籍した監督: 船越優蔵
- ヒノキ科の植物: ミヤマビャクシン
- <作品 OR 番組>のキャスト: 立花大介
- 船: 将 [意味不明]
- これまでの<番組>の代理司会者: Mr. マリック
- <作者>の作品: あくまこあくま
- <業種>の中堅メーカー: 宮島醤油
- 都市 OR 町: ジョージアナ
- 小惑星: 菅野洋子 [意味不明]
- <著者>の<著作>の他の著: 改訂電子回路
- <出演者>の出演作品: 華麗な休暇
- <都市>の友好都市: 島根県松江市

要件の定義 2/2

- 上位語の非飽和性 (西山 2003) [重度の問題]
  - <競技>の<チーム>の現役選手, <作者>の<分野>での作品,
  - ヒノキ科の植物
  - 特殊な場合として未解消な相対指示性をもつ上位語
    - これまでの司会者, 放送予定の番組, 放送中の番組
- 対応の不適格性 [軽度の問題]
  - 船: 将, 小惑星: 宮島洋子

作業の設計 1/3

- 問題1と問題2は別にする
- 本発表では問題1の解決のための約94,000個の上位語のクリーニング作業の手順と結果を報告
- 問題2も別系統で作業中
  - 90万の上位語・下位語対の対応評価が進行中

作業の設計 2/3

- 元データ
  - h: 元スピードスケート長距離選手, i: 牛山貴広
- から次を生成
  - h1: 選手; h2: 長距離選手; h3: スケート長距離選手; h4: スピードスケート長距離選手; h5: 元スピードスケート長距離選手,
  - i: 牛山貴広

作業の設計 3/3

- h5からh1; h2; ... ; h5のような上位語パスを自動生成し，パスの要素からなるべく多くの用語を取り出す
- 上位語の主要部を取るだけでは
  - 未飽和名詞だけが獲れても嬉しくない
  - 上位オントロジーと固有名を接続している中間オントロジーの情報を損失

上位語パスの例

sample-pla

	A	B	C	D	E	F	G	H	
1	ID	h1	h2	h3	h4	h5	h6	i	r
2	2314041	路線	計画路線	道路計画路線	規格道路計画路線	高規格道路計画路線	地域高規格道路計画路線	根室中標津道路	
3	2315403	路線	指摘された路線	点が指摘された路線	問題点が指摘された路線	関わり問題点が指摘された路線	転換に関わり問題点が指摘された路線	伊勢線	
4	2316572	路線	中の路線	計画中の路線	中および計画中の路線	建設中および計画中の路線	現在建設中および計画中の路線	仙台市営地下鉄・東西線	
5	2321611	狼	人狼	上の人狼	インターネット上の人狼	国内でのインターネット上の人狼	日本国内でのインターネット上の人狼	チャット形式	
6	2323345	論	方法論	開発方法論	ソフトウェア開発方法論	指向ソフトウェア開発方法論	オブジェクト指向ソフトウェア開発方法論	オブジェクトモデル化技法	
7	2323500	論理	古典論理	非古典論理	拡張としての非古典論理	論理の拡張としての非古典論理	古典論理の拡張としての非古典論理	時相論理	
8	2323502	論理	古典論理	非古典論理	代替としての非古典論理	論理の代替としての非古典論理	古典論理の代替としての非古典論理	線形時相論理	
9									

評定作業の実際

- 上位語パスの要素を人手で4つのタイプに分類
  - *Good* terms [薄い緑色]: 独立した概念を表わす語
  - *Less Good* terms [濃い緑色]: 非飽和な概念を表わす語
  - *Dubious* terms [空色]: 成語性が怪しいもの，取り決めで格下げした語など
  - *Bad* terms [無色]: 成句性のない文字列，最下位の上位語に対して上位語にならないもの (否定がからむと起こる現象)

G, LG, D, B の例

- *Good*: 秋田県出身の人物, 日本の鉄道駅, 駅, 醤油の中堅メーカー
- *Less Good*: 出身の人物, 登場人物, 中堅メーカー, 誌
- *Dubious*: かけ丼
- *Bad*: 的人物, 非古典的論理の上位語としての古典的論理
  - 表層形で獲得できる名詞句が Less Good である割合はかなり高い

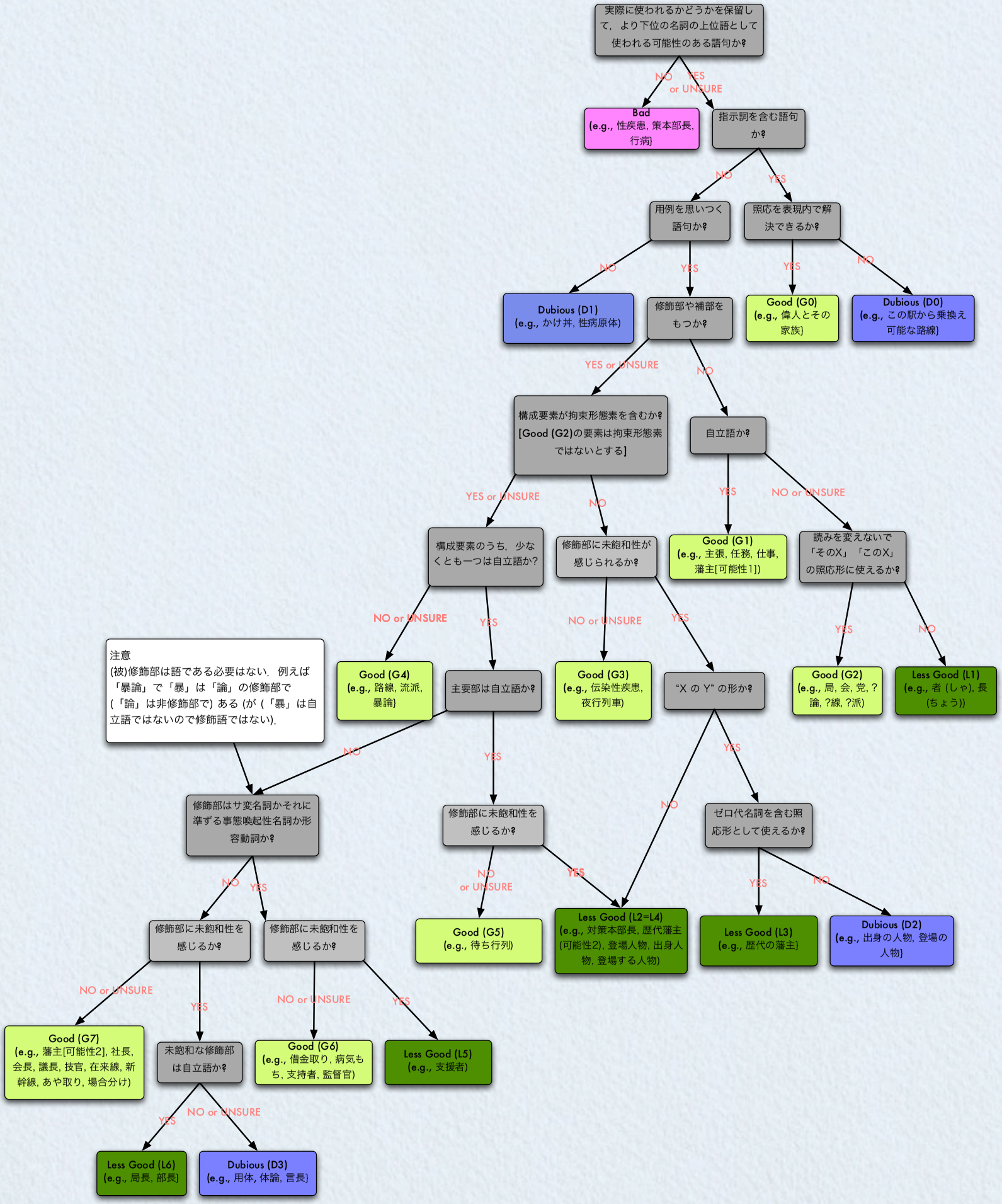
cleaned-4b-check.xls

A	B	D	F	O	P	Q	R	S	T
ID	Length	Annotator	Checker	パス編集	h1	h2	h3	h4	i
376	4	jhl	sh	1	アルバム	美奈子のアルバム	本田美奈子のアルバム	AVE_MARIA	
0351	4	jhl	nw		紙	専門紙	競馬専門紙	中央競馬専門紙	競馬ニホン
0375	4	jhl	nw		紙	専門紙	食品専門紙	水産食品専門紙	みなと新聞
0424	4	jhl	nw		紙	専門紙	流通専門紙	青果物流通専門紙	農経新聞
9742	4	jhl	nw		紙	季刊紙	情報季刊紙	生活情報季刊紙	グリエツィ
0477	4	jhl	nw		紙	総合紙	産業総合紙	医薬品産業総合紙	薬事日報
1091	4	jhl	nw		紙	夕刊紙	タブロイド判夕刊	駅売りタブロイド判夕刊紙	夕刊フジ
1506	4	jhl	nw		詩	引用詩	中引用詩	劇中引用詩	ボブ・ディラン
1868	4	jhl	nw		詩	交響詩	作交響詩	代表作交響詩	ロシア
7319	4	jhl	nw		試	模試	実施されていた模試	かつて実施されていた模試	横国大入試プレ
7314	4	jhl	nw		試	模試	試験の模試	資格試験の模試	全国大検模試
7315	4	jhl	nw		試	模試	試験の模試	入学試験の模試	愛知全県模試
4841	4	jhl	nw		試験	士試験	鑑定士試験	不動産鑑定士試験	短答式試験
4844	4	jhl	nw		試験	資格試験	語に関する資格試験	朝鮮語に関する資格試験	韓国語能力試験
4889	4	jhl	nw		試験	性能試験	安全性能試験	衝突安全性能試験	オフセット前面衝突
4958	4	jhl	nw		試験	認定試験	資格認定試験	教員資格認定試験	高等学校教員資格試験

サンプル

細則

- 作業マニュアルを準備し，アノテーターを手取り足取り指導
- 相手にするデータは複雑であり，細則がいろいろある


実際に使われるかどうかを保留して，より下位の名詞の上位語として使われる可能性のある語句か？
NO
YES or UNSURE
Bad (e.g., 性疾患, 策本部長, 行病)
指示詞を含む語句か？
NO
YES
用例を思いつく語句か？
照応を表現内で解決できるか？
NO
YES
YES
NO
Dubious (D1) (e.g., かけ丼, 性病原体)
修飾部や補部をもつか？
Good (G0) (e.g., 偉人とその家族)
Dubious (D0) (e.g., この駅から乗換え可能な路線)
YES or UNSURE
NO
構成要素が拘束形態素を含むか？ [Good (G2)の要素は拘束形態素ではないとする]
自立語か？
YES or UNSURE
NO
YES
NO or UNSURE
構成要素のうち，少なくとも一つは自立語か？
修飾部に未飽和性が感じられるか？
Good (G1) (e.g., 主張, 任務, 仕事, 藩主[可能性1])
読みを変えないで「そのX」「このX」の照応形に使えるか？
NO or UNSURE
YES
NO or UNSURE
YES
YES
NO
注意 (被)修飾部は語である必要はない．例えば「暴論」で「暴」は「論」の修飾部で (「論」は非修飾部で) ある (が (「暴」は自立語ではないので修飾語ではない).
Good (G4) (e.g., 路線, 流派, 暴論)
主要部は自立語か？
Good (G3) (e.g., 伝染性疾患, 夜行列車)
“X の Y” の形か？
Good (G2) (e.g., 局, 会, 党, ?論, ?線, ?派)
Less Good (L1) (e.g., 者 (しゃ), 長 (ちょう))
NO
YES
YES
NO
修飾部はサ変名詞かそれに準ずる事態喚起性名詞か形容動詞か？
修飾部に未飽和性を感じるか？
ゼロ代名詞を含む照応形として使えるか？
NO
YES
NO or UNSURE
YES
YES
NO
修飾部に未飽和性を感じるか？
修飾部に未飽和性を感じるか？
Good (G5) (e.g., 待ち行列)
Less Good (L2=L4) (e.g., 対策本部長, 歴代藩主 (可能性2), 登場人物, 出身人物, 登場する人物)
Less Good (L3) (e.g., 歴代の藩主)
Dubious (D2) (e.g., 出身の人物, 登場の人物)
NO or UNSURE
YES
NO or UNSURE
YES
Good (G7) (e.g., 藩主[可能性2], 社長, 会長, 議長, 技官, 在来線, 新幹線, あや取り, 場合分け)
未飽和な修飾部は自立語か？
Good (G6) (e.g., 借金取り, 病気もち, 支持者, 監督官)
Less Good (L5) (e.g., 支援者)
YES
NO or UNSURE
Less Good (L6) (e.g., 局長, 部長)
Dubious (D3) (e.g., 用体, 体論, 言長)

作業日程

- 前半 (5月-8月)
  - 黒田 航, 李 在鎬 (週3日), 野澤 元 (週1日)
- 後半 (8月-10月)
  - 8月中旬から派遣作業者 (4人) を導入
  - 10月に一通り作業完了
  - その後は新規に獲得された追加データ () で同様の作業
    - 新規な上位語 55,194, 共有 38,253,

結果 1/3

- 前処理で84,642になった行 (元データの段階で94,744) のうち74,564行を処理
  - パス長=1で下位語が上位語と同じ語で終わらない行は未処理
- 作業結果を整理したものを NICT 主催の「高度言語情報融合フォーラム」で配信する予定

結果 2/3

Length	Count
1a	2,495
2	30,968
3	23,614
4	11,112
5	4,230
6-12	2,145
Total	74,564

- パス長さごとの処理行数
- Length=1a (下位語と上位語の終わりが共通) は上位語と下位語をh2と見なして評価

結果 3/3

Class	Count
Good	76,373
Less Good	26,828
Dubious	8,614
Bad	19,529
Total	131,344

- パス要素の評定値の分布
- G, L, D, B の数えは重複
  - 次の理由で同一のパス要素に異なる評価があてられている場合がある
  - 意味の曖昧性がある場合
  - 評定が不統一な場合

評定後の上位語パスの例

- 空手家:女性空手家:コンタクト系女性空手家
- 学校:特別支援学校:京都府の特別支援学校
- 学校:特別支援学校:石川県の特別支援学校
- 学校:聾学校:大学に附属する聾学校:国立大学に附属する聾学校
- 学校:高等学校:佐倉高等学校:千葉県立佐倉高等学校
- 学校記念館:高等学校記念館:旧制高等学校記念館
- 家:皇女を通じて近親に当たる家:天皇の皇女を通じて近親に当たる家:昭和天皇の皇女を通じて近親に当たる家
- 密度:質量密度:単位体積あたりの質量密度

下位語は非表示

前処理

前処理の重要性

- 上位語パスの要素数は約240,000個で，作業量は膨大
- 作業内容を工夫しないと (指定された期限の半年では) 終わらない!!
- 前処理で作業量を減らすことが肝腎

行なった前処理

- 前処理 1: 下位語のサンプリング
- 前処理 2: 上位語パスの追加とパス長での分類
- 前処理 3: 冗長な行の除外
- 前処理 4: 有用性の低い上位語をもつ行の除外
- 前処理 5: WordNet-Ja (Bond et al. 2007)との対応の表示
- 前処理 6: 「主な」のような限定詞の削除
- 前処理 7: 「メダ」「ギタ」のような不完全要素の補完

前処理 1/4

- 元データの上位語の異なり数は 94,744
  - 下位語の異なり数は110万程度
- 上位語の異なりを確保して対をサンプリング (n=1)
  - 下位語の頻度は考慮に入れず

前処理 2/5

- 上位語パスの構築
  - 元データの上位語を形態素解析し，品詞情報に基づいて段階的に複合表現を抽出
- パスの長さでデータを分類
  - もっとも長いパスは14
- この処理で上位語の異なり数は11,949に縮約
  - この数は IPA Dic でのもの

主要部認定のための正規表現

- Strict: r"未知語.*|接頭.*名詞.*|名詞.*(一般|サ変|固有|語幹).*"
- Tolerant: r"未知語.*|接頭.*名詞.*|名詞.*(一般|サ変|固有|語幹|非自立|接尾|副詞可能).*"
- Loose: r"未知語.*|接頭.*名詞.*|名詞.*(一般|サ変|固有|語幹|非自立|接尾|副詞可能).*|.*副助詞.*"
- Very loose: r"記号.*|未知語.*|接頭.*名詞.*|名詞.*(一般|サ変|固有|語幹|非自立|接尾|副詞可能).*|.*副助詞.*"

最上位語集合

hypernym-roots-via-chasen-v1.xls

	A	B	C	D	E	F	G	H	I	J
1	hypernym root	len	freq	freq>2	freq>4	freq>12	validity	included	Transform	Suggestion
2	者	1	2286	1	1	1	1	1		人物: 者
3	作品	2	1997	1	1	0	1	1		
4	番組	2	1944	1	1	0	1	1		
5	家	1	1600	1	1	1	1	1		人物: 家
6	選手	2	1502	1	1	1	1	1		人物: 選手
7	人物	2	1491	1	1	0	1	1		人物
8	施設	2	1165	1	1	0	1	1		
9	こと	2	1110	1	1	0	1	0		
10	会社	2	871	1	1	0	1	1		
11	もの	2	846	1	1	0	1	0		
12	学校	2	807	1	1	0	1	1		
13	種類	2	743	1	1	0	1	0		
14	企業	2	723	1	1	0	1	1		人物?: 団体: 企業
15	人	1	690	1	1	0	1	1		人物<=人
16	メーカー	4	624	1	1	0	1	1		人物?: 団体: メーカー
17	有名人	3	611	1	1	1	1	1	人: 有名人	人物: 有名人
18	ソフト	3	574	1	1	0	1	1		
19	ゲーム	3	523	1	1	0	1	1		
20	シリーズ	4	514	1	1	0	1	1		
11937	シミュレーションロールプレイングゲーム	19	1	0	0	0	1	1		
11938	ソニー・ミュージックエンタテインメント	19	1	0	0	0	1	1		
11939	コロムビアミュージックエンタテインメント	20	1	0	0	0	1	1		
11940	リアルタイムストラテジーコンピューターゲーム	22	1	0	0	0	1	1		
11941	タクティカルエスピオナージアクションゲームシリーズ	25	1	0	0	0	1	1		
11943	Total		94649	3107	1896	670		17282		
11944								0.18259		
11945										

最上位語集合

- A. 上位語パス (=最下位の上位語) の異なり数: 94,649
  - そのうち18%が[人物]に関するもの
- B. 頻度が2より大きな最上位語の異なり数: 3,107 (3.28%)
- C. 頻度が4より大きな最上位語の異なり数: 1,896 (2.00%)
  - B を基に人手でオントロジーを構築することが可能

パス長分類

length-class-ranks.xls

	A	B	C	D	E	F
1	length	lines	cells	rank	ratio	note
2	3	24778	74334	1	31.86%	
3	2	32619	65238	2	27.96%	
4	4	11949	47796	3	20.49%	
5	5	4305	21525	4	9.23%	
6	6	1547	9282	5	3.98%	
7	1	8582	8582	6	3.68%	上位語と下位語が同一形態素で終わるクラスと終わらないクラスに分割
8	7	518	3626	7	1.55%	
9	8	201	1608	8	0.69%	
10	9	77	693	9	0.30%	
11	10	25	250	10	0.11%	
12	11	18	198	11	0.08%	
13	12	7	84	12	0.04%	
14	13	4	52	13	0.02%	
15	14	2	28	14	0.01%	
✳						
17	Total	84632	233296		100.00%	

前処理 3/5

- 情報不足の行の削除
  - 上位語と下位語が同一な行
  - (元データで上位語と下位語が非同一だが) 上位語パスの最上位語と下位語の対が同一な行

前処理 4/5

- 不適切な(最)上位語をもつ行の除外 (別に処理する)
  - (1) 等; (2) など; (3) ほか; (4) 他; (5) 類い; (6) もの; (7) モノ; (8) 物; (9) こと; (10) コト; (11) 事; (12) 名; (13) 呼称; (14) 総称; (15) 通称
  - (16) 上位語に “・” が含まれる行
- 「主な」を含むパス要素の削除
- 「メダ」や「ギタ」で終わる行を編集

前処理の効果

- 以上の前処理により，処理すべき行は94,744行から84,642に減少
- 更に連言的や選言的な用語は遭遇する度に隔離した
  - X及びY, X並びにY, XとそのY, etc
  - 分離して，後処理に回す

上位語パス追加の効果

- 上位語パスを追加する前の WordNet-Ja (v0.6-all) の被覆率は50%程度だった
- 上位語パスの追加で，最上位語にある上位語の80%強が WordNet-Ja に対応語をもつようになった
  - ただし語義の区別は考えないでの話

今後の課題 1/2

- 上位語のオントロジー構築
  - 曖昧性を解消しWordNet-Ja と対応づける
  - 上位語パスの最上位に現われる語彙素 (e.g, 症, 家) の体系化
  - Wikipedia 特有の概念 (e.g., 作品の登場人物, 歴史上の存在, 架空の存在) に適応する必要あり
- 多言語化
  - 英語版 Wikipedia から獲得したデータとの対応づけ

今後の課題 2/2

- 未飽和名詞句の自動獲得
  - 名詞Nが非飽和名詞であるならばNに先行する文脈でノ以外の助詞が生起する割合が低い

THANKS FOR YOUR ATTENTION

付録 1

付随する問題

- 上位語オントロジーを整備するには，最上位語集合を標準化する必要がある
  - 形態素解析のレベルで誤解析がデータの「汚れ」につながっている例は稀ではない

最上位語の標準化

- 分類ランク名のクラス名への変換
  - Xの種類⇒X, X属⇒X
- OR
  - X属 ⇒ X属のY
- 一語の語彙素/形態素の曖昧性の解消
  - 族⇒部族, 族⇒種族
  - 法⇒法規, 法⇒方法・技法,

同義性判定

- 略語の補足
  - ソフト⇒ソフトウェア
- メトニミー的同義性の認定
  - サイト⇒サービス, コンテンツ⇒サービス [文脈自由]
  - システム⇒サービス, 技術⇒サービス [文脈依存]
- WordNet-Ja を使えば(半)自動化できる??

浮上中の意外に厄介な問題

- 形態素解析プログラムで単語性/形態素性の認定基準が不統一で不明瞭
  - IPA Dic では「料理人」は2語, 「有名人」は1語
  - Juman と UniDic では「料理人」と「有名人」が2語
- 上位オントロジー構築のためには (多少の曖昧性があっても良いから) 語より細かい意味認定単位) (e.g.., 人, 者, 物, 所) が欲しい

IPA DIC の複合単語LEN=4

- おとぎ話, 露天風呂, 三和酒類, 情報処理, 魚形水雷, 合い言葉, 大和言葉, 西太平洋, インド洋, 日本石油, 底引き網, 精神療法, 中国地方, 断崖絶壁, 産経新聞, 脊椎動物, 節足動物, 軟体動物, 観葉植物, 顕花植物, 被子植物, 裸子植物, 食虫植物, 多肉植物, 炭水化物, 水酸化物, 幕僚監部, 音楽学部, 社会学部, 練り製品, 財務諸表, 軽便鉄道, 奄美諸島, 大東諸島, 南西諸島, テレビ塔, 宇治山田, 岩波書店, 京成電鉄, 阪急電鉄, 宮崎交通, 三重交通, 鶴見緑地, 流通団地, 工業団地, パン生地, 名古屋帯, 君主政体, 毎日放送, 長距離走, 南北戦争, 戊辰戦争, 水中翼船, 内分泌腺, 名所旧跡, 中性子星, 吟遊詩人, 桂冠詩人, 太政大臣, 国務大臣, 一休宗純, 慶應義塾, 二十八宿, 浄土真宗, 百人一首, リンパ腫, 天台座主, 軽自動車, 第一人者, 変わり者, ならず者, 秋葉神社, 氷川神社, 株式会社, 廃止当時, 固有名詞, 学園都市, 浮世草子, アミノ酸, 一夫多妻, 心筋梗塞, 信用組合, 近世以降, えびす講, 地方銀行, 太上天皇, 朝鮮学校, 小中学校, 二十四孝, 太皇太后, 出入り口, 作り物語, 軍記物語, ラテン語, ドイツ語, 信用金庫, 都道府県, 治外法権, 起承転結, 三十六計, 掛け時計, 正多角形, 日本航空, 筆記用具, 飛び道具, 七つ道具, ミニ四駆, テレビ局, 森永乳業, 人身御供, 伊勢神宮, 潮見が丘, 劇団四季, セスナ機, 五星紅旗, 金管楽器, 休憩時間, 経過時間, 計算時間, 作業時間, 放送時間, 警視総監, 政務次官, 事務次官, 雌阿寒岳, 量子力学, 近畿大学, 国際大学, 東洋大学, 単科大学, 総合大学, 短期大学, 形而上学, 帝王切開, 七つの海, 統一教会, 創価学会, かがり火, 判定結果, ズボン下, 気管支炎, 海の公園, 森林公園, 緑地公園, 運動公園, 雙葉学園, 原生花園, 最寄り駅, 美福門院, 市立病院, 福井病院, 義務教育

その他

- Len = 4 の場合ほど顕著ではないが，Len = 5, 6, 7, ..., 16 にも解析されない複合語がある
- Len =8
  - 田園調布雙葉学園, 日本テレビ放送網, 日本民間放送連盟, 薄膜トランジスタ
- Len = 16
  - 徳間ジャパンコミュニケーションズ

付録 2

下位語性評価の問題

- 試行から次の問題が浮上
  - 下位語候補が本当に下位語になっている率 (下位語獲得の精度) は思ったほど高くない
    - Length=4 の場合の試行で 60% 程度
  - i が何を表わしているか不明な場合が圧倒的に多い
    - 評定支援ツールが不可欠

L=4の場合

	A	B	C	D	E	F
1	評価3	説明	例	集計	Rate	Note
2	x	i は未知語	h5:ナチス体制下で製作された映画/i: オリンピア	10069	90.77%	全体に対する割合
3	1	i は hx の下位語	h2: スポーツセダン/i: いすゞジェミニ	318	57.30%	{0,0.1,0.5,1}の合計に対する割合
4	0.5	i は hx の関連語だが,下位語ではない特定の関係がある	h5: 東京オリンピック選手団/i: 岡野功	118	21.26%	{0,0.1,0.5,1}の合計に対する割合
5	0.1	i と hx がどんな関係か不明確か,iは hx の下位語でない	h1: テーマ/ h2: 最終テーマ/ i: 炎; 機動警察派とレイバー/i: 機動警察派とレイバー the Movie	46	8.29%	{0,0.1,0.5,1}の合計に対する割合
6	0	i は hx の下位語ではない	h5: コピーコントロールヲ実施している会社/i: 完全撤退	73	13.15%	{0,0.1,0.5,1}の合計に対する割合
7	{0,0.1,0.5,1}の小計			555	5.00%	
8	対象外	最下位の上位語が Dubious か Bad		469	4.23%	
9	{0,0.1,0.5,1,対象外}の小計			1024	9.23%	
10	総計			11093		length=4の場合のみ

再獲得のため提案

- 規模の拡大のために Wikipedia データの獲得をやり直すなら
- 上位語候補 h と下位語候補 i の対を獲得するのではなく，階層パスを tuple で獲得すべき
  - 獲得時には後処理で有効な上位語と下位語の対を同定することを前提にする

Haskell (例)

- 1 概要
- 2 構文
  - 2.1 代数的データ型
  - 2.2 カリー化と関数の部分適用
  - 2.3 型クラス
  - 2.4 リストとリスト内包表記
- 3 実例
  - 3.1 より複合的な例
- 4 批判
- 5 実装
  - Glasgow Haskell Compiler
  - Gofer
  - HBC
  - Helium
  - Hugs
  - Jhc
  - nhc98
  - yhc
- 6 関連
- 7 参照
- 8 外部リンク

獲得の対象

- (Haskell, 実装, Glasgow Haskell Compiler) のような tuples
- 獲得された tuples を加工して (Haskell の実装: Glasgow Haskell Compiler) のような対を生成
  - 「実装」が未飽和なサ変名詞であることを利用するなどして自動化も可能?