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
National Institute of Information and Communications Technology (NICT)

1 はじめに

[4] によって日本語版 Wikipedia から約 244 万個の上位語/下位語対が自動獲得され，鳥式改 [6] に利用されている．本発表では NICT が行なった上位語クリーニングの方法と結果を主に報告するが，これと並行して上位語が後述の G 評価をもつような 100 万対の上位/下位関係の適切性のチェックも行なっている．クリーニングの済んだデータは「高度言語情報融合フォーラム」を通じて配信する予定である．

1.1 作業の目的

[4] によって自動獲得されたデータを D とする (D はパターンマッチで自動獲得された事例を SVM で分類し，90% の精度で分離されたものである)．

D を構成する上位語の異なり数は 9.4 万，下位語の異なり数は 110 万で[1]，(1) に示すような事例からなる (これらは D から無作為に 30 個の対を選んだ結果である):

(1) 1. 現役選手:マット・モリス; 2. 大阪府出身の人物:金森又一郎; 3. 過去に在籍した選手/歴代監督: 船越優蔵; 4. 秋田県出身の人物: 高田斉; 5. ヒノキ科:ミヤマビャクシン; 6. キャスト:立花大介; 7. 船:将; 8. 日本の法学者:小菅成一; 9. アニメ作品:魔法遊戯; 10. 作品:マンガ; 11. 日本のインターチェンジ:利府塩釜インターチェンジ; 12. これまでの代理司会者:Mr. マリック 13. 作品:あくまこあくま; 14. 架空の惑星:バース星; 15. 中堅メーカー:宮島醤油; 16. 都市及び町:ジョージアナ; 17. 小惑星:菅野洋子; 18. 出演作品:華麗な休暇; 19. ラジオ番組のパーソナリティ・DJ:中村基樹; 20. フランスの彫刻家:フレデリク・バルトルディ; 21. 附属機関:神道博物館; 22. 他著:改訂電子回路; 23. 友好都市:島根県松江市; 24. パラグアイのサッカー選手:リカルド・バエス; 25. 邸宅:ベッキンガム宮殿; 26. 日本の鉄道駅:落合川駅; 27. キャラクター:イスピン・シャルル; 28. 歴代監督:ローラン・フルニエ; 29. 株式会社:アスピア 30. 元スピードスケート長距離選手:牛山貴広

1.1.1 D の性質

D は概念辞書の一種と見なすことができるが，採録している用語の範囲が広く，階層が浅いという点で従来シソーラスとは異なる．(1) を見ればすぐにわかるように D には従来のシソーラス [5, 10, 1] が記述して来なかった固有名や複合名詞が非常に多く含まれる[2]．従って，D をシソーラス (e.g., WordNet-Ja [1]) に接続することで，固有名辞書と上位オントロジーとの接続を実現できる可能性がある．

1.1.2 改良の必要性

獲得精度 90% とは言え，D はきれいなデータではない[3]．(1)-7，(1)-17 のような意味不明な対があるほかにも，上位語集合，下位語集合のおのおのに幾つかの難点がある．上位語集合の問題点は次の通り:

(2) 粒度が揃っていない
(3) 上位語集合に意味的に未飽和 (unsaturated) な (関係) 名詞 [11] がかなりの割合で含まれる．上位語の一部は，(4) に示す形で適当な項を補完するか，余分な要素を除去しないと有用性は低い:[4]
(4)
  a. (1)-1 の上位語「現役選手」⇒「〈チーム〉の現役選手」
  b. (1)-3 の上位語「過去に在籍した選手/歴代監督」⇒「〈チーム〉に過去に在籍した選手/歴代監督」
  c. (1)-6 の上位語「キャスト」⇒「〈作品か番組〉のキャスト」
  d. (1)-15 の上位語「中堅メーカー」⇒「〈業種〉の中堅メーカー」
  e. (1)-18 の上位語「出演作品」⇒「〈出演者〉の出演作品」
  f. (1)-21 の上位語「付属機関」⇒「〈帰属先〉の付属機関」
  g. (1)-22 の上位語「他著」⇒「〈著者〉の〈排除項〉の他の著 (作)」
  h. (1)-23 の上位語「友好都市」⇒「〈都市〉の友好都市」
  i. (1)-27 の上位語「キャラクター」⇒「〈作品〉のキャラクター」
(5) instance-of (事例化) の関係と member-of の関係の混在のような，意味関係の混同がある (例えば (1)-29 の「歴代監督」はグループを指示し，クラスは指示しないので，[歴代監督: 〈個人名〉] は正確には上位語と下位語の関係ではない[5])．

ただ instance-of 関係と is-a 関係 (本来の hypernym/hyponymy 関係) を区別することは難しく，かつデータの目的から考えると本質的でない可能性が高い．

下位語集合の問題点は以下の通り:

---

[1] 最大級の電子化辞書である EDR[10] の日本語単語辞書でも収録語数は 27 万である．

[2] 特に下位語集合では，固有名が占める割合が非常に高い．7 割程度は固有名である．上位語集合にも固有名がそれなりの割合で混入しているが，その場合には下位語が特殊のものが多い．

[3] 原因の一つは SVM に与える「正例」の中に未飽和な (関係) 名詞が少なくないことにある．だが，それらを除外することは，今のところ正例の不足を招く可能性が高く，被覆率の低下を招く可能性が高いという理由で現実的ではないと思われる．

[4] この知見は元々 [8] に由来する．

[5] 「選手」は個体の属性指定とグループの指定の両方が可能であるが，後者は派生的であると考えられる．

(6) 多くが固有名で，しかも未知語であるため，真偽性の判断以前に成語性の判断が難しい[6].

以上の問題点のうち，(2) と (3) を解決するため，§2 で説明する方法で上位語のクリーニング作業を行なった．

(5) は厳密には上位語固有の問題ではなく，特定の対で顕在化するものである．これは上位語の整備と同時並行して行なっている 100 万対の下位語のクリーニングの一部となる．

1.2 扱うべき問題

(2) と (3) に関する改良作業を報告すると同時に，NLP における名詞の処理に関するもう少し一般的な問題も提起しておきたい．

(3) に関しては，多くの場合で未飽和性は語彙的に成立していると言うより，修飾語句が付いて複合語になった段階で成立している点には注意が必要である．関係名詞の項の省略の可能性を考えると，かなりの割合の複合名詞が表層形のままでは未飽和な (関係) 名詞のままである可能性がある．これは，未飽和 (関係) 名詞の項構造解析が (動詞などの述語の項構造解析に劣らず) 意味処理上で重要である可能性を示唆する．

名詞 N の未飽和性は，N の特質構造 [3] と関連が深いがそれ以上の複雑性をもっている可能性がある．従って，名詞の意味論に関する一般的な枠組みを構築する必要がある (少なくとも特質構造理論の延長上に何かをしようと思っても，特質構造を自動獲得し，人手チェックしてデータベース化するという作業が必要だろう)．

事態 (喚起) 性名詞の分析に関する限り，NomBank [2] は先駆的な仕事であり，それから派生した SynCha [12] が日本語に関して先駆的な仕事であるが，関係名詞 (句) 一般の (共) 項構造に記述のレベルには達していない．名詞の共項構造の基礎理論が不在であることが原因になっていることは否定できない．

日本語に限らず，どの言語の言語理論，言語処理でも名詞の解析理論が全体的に未発達である理由は明白である: それは従来の言語理論があまりに述語偏重だったことの理論的な帰結である．「文は (主要部である) 述語の投射である」という (直観的には信憑性があるように思える) 想定が，述語以外の要素に (冗長性を許して，同時並行的に) 別個の (共) 項構造をもたせるという，並列/分散的な表示モデルを抑制してきたのは確実である[7].

だが，ここで次のように考えてみることは意味のないことではないだろう: 語彙素/形態素を品詞クラスに分けることは統語解析においては必須な処理だが，それは意味解析において本質的な処理だろうか? 少なくともそれが意味処理では非本質的である可能性は十分にある．その意味では，従来の言語処理と言語学が前提にしてきたカスケード/パイプライン型の処理モデルの限界がこの点で露呈しているとも言える．

2 上位語クリーニング作業

2.1 手順

2.1.1 上位語パスの追加

D の上位語 H の要素 $h \in H$ のそれぞれを形態素解析し，得られた品詞情報に基づいて h を階層化する (結果を h の上位語パスと呼ぶ)．例えば (1)-31 に上位語パスを追加したものは (7) である:

(7) 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 (=h). 元スピードスケート長距離選手:
i. 牛山貴広

上位語パスの最左に現われる上位語候補を「(上位語パスの) 最上位語」と呼び，もっとも右にある，元の上位語を「(上位語パスの) 最下位の上位語」と呼ぶ．

このようにして生成された上位語パスの構成要素の異なり数は 143,791 だった．

2.1.2 上位語パスの要素の人手評価基準と結果

自動的に生成された上位語パスのノード (h1, h2, . . .) に現われる上位語候補 h_i は，i) 成語性のない文字列であったり，仮に成語性があっても ii) 下位語 i の上位語として不適切だったり，iii) 未飽和な名詞だったりする．そのため，上位語パスに現われる全候補を (8) の Good (=G), Less Good (=LG), Dubious (=D), Bad (=B) の 4 段階で人手で評価した:

(8) B: 成語性がないもの (e.g,.「的人物」「展記念 CD」「ートの魔法アイテム」)
D: 成語性があるか判断が難しいもの (e.g., 「かけ丼」「史研究者」)
G: 成語性があり，かつ未飽和性を感じないもの (e.g., 「マラソン選手」)
LG: 成語性があり，かつ未飽和性を感じるもの (e.g., 「選手」「アメリカ合衆国の選手」)[8]

以上の手順によって，143,791 個の上位語候補集合から (異表記を含めて) 約 7.5 万の G 評価をもつ上位語集合を得た[9].

予備調査の段階で上位語パスの要素の評価の際に起こってくるとわかっていた様々な問題に対処するため，§2.2 で説明する前処理を行なった．

2.2 作業用データの準備

2.2.1 処理 1: サンプリング

鳥式改の元データ D の上位語 H の異なりは 94,744 個である．この上位語集合が重複なしに現われるように事例をランダムに取り出した．このデータを sampled-n=1 データとする．

[6] 知識のある人間が判断しても，外部資料を参照しないで自信をもって判定できるものは，1/3 程度しかない．

[7] もちろん，この想定には見かけの妥当性がある他にも，処理モデルの単純化を可能にするという利点があるが，IT の発達で処理が現実的になるにつれて，扱い切れない例外は増える一方である．

[8] (1)-12 のように特定の時点に言及する場合は未飽和表現には含めずに D 扱いしている．

[9] チェック自体が終わっていないため，これは正確な数字ではない．現在，これらの集合から良俗に反する用語 (e.g., 他者に対して明白な攻撃性をもつ用語や猥褻性のある用語) を分離する作業を進めている (執筆時点で未完了).

2.2.2 処理 2: 上位語パスの追加

sampled-n=1 のデータを基に ChaSen/IPADic を使って上位語パスを延ばした[10]．この結果として得られた最上位語の異なりは 11,949 個である[11]．

この作業では，候補の生成は上位語候補の品詞が正規表現で記述された主要部と一致する場合に，それまでに新しい主要部が現われた段階で結合させるという手順で生成した．主要部を認定する生起表現としては次の Strict, Tolerant, Loose, Very Loose の四つを試し，何度かの最適化のための試行の上で Tolerant を選んだ (Strict は精度が高くゴミが混ざらない代わり，主要部の候補の被覆率が十分ではなかった):[12]

(9) a. (Strict:) 未知語.*| 接頭.*名詞.*| 名詞.*(一般 | サ変—固有—語幹).*
b. (Tolerant:) 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).*
c. (Loose:) 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.*
d. (Very loose:) 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.*

上位語パスの追加により上位語の般化が行われた (参考値として，上位語が当時開発中だった WordNet (WN-Ja) v0.6-all [1] の語彙集合に (語義の曖昧性解消を考慮を入れないで) 含まれる割合は，パスの追加により約 50% から約 80% 強に向上した)．これにより (2) の粒度の分散の問題も部分的に解消されていると期待できる．

2.2.3 処理 3: 冗長な行の除外

パスの途中が異なっていても．パスの最上位語と下位語の対が一致するものを冗長な対と見なし，パスの短い方を取り除いた．

2.2.4 処理 4: 不適切な最上位語の除外

最上位語の異なり数は 11,949 であるが，その一部には上位語として適切でないものが含まれる．事前調査から不適と思われる (10) と (11) を合わせた 16 個の条件に該当する行を取り除いたデータを作業対象から分離し，独立に後処理することした:

(10) (1) .*など; (2) .*ほか; (3) .*他; (4) 類い; (5) 等 (6) もの; (7) モノ; (8) 物; (9) こと; (10) 事; (11) コト; (12) 名; (13) 呼称; (14) 総称; (15) 通称;
(11) (最下位の) 上位語に「・」を含む行

以上の前処理で sample-n=1 のうち，人手作業の対象とすべきものは 84,632 対に減った．

「・」を含む行を作業から外したのと同じ理由で，最下位の上位語が「A と B」「A や B」「A 及び B」「A 並びに B」のような形で選言を表わしている場合，§2.2.2 の上位語パス追加の自動処理で得られるのは，指示の般化ではなく特殊化である．しかし，「・」の場合と違って，これらの場合は自動処理で分離することは難しい．このため，作業の途中で選言の事例に当たったら，適宜，除外することにした．

2.2.5 処理 5: パス長による分類

効率化のため，上位語パスの長さ l に応じて類別した．l ごとのデータの分布は (12) にある通り:

(12) $l = 1$: 8,582; $l = 2$: 32,619; $l = 3$: 24,778; $l = 4$: 11,949; $l = 5$: 4,305; $l = 6$: 1,547; $l = 7$: 518; $l = 8$: 201; $l = 9$: 77; $l = 10$: 25; $l = 11$: 18; $l = 12$: 7; $l = 13$: 4; $l = 14$: 2 (合計: 84,632)

2.2.6 前処理 6: 「主な」などの限定要素の削除

上位語に「主な」「おもな」「主要な」「主要」を含む行を編集し，これらの部分文字列を削除した．この処理の理由は，i) これらの要素が上位語認定で意味が空虚な限定であり，ii) 作業者の評定の撹乱要因になっていることが後から判明したからである．

2.3 上位語性評定作業の中身

第一，第三著者が事前調査に基づいて用意した作業マニュアルに従って，四人の作業者に依頼した．作業は Microsoft 社の Excel を使った．これは，i) Visual Basis for Application (通称 VBA) を使ってワークシート上の定型作業の効率化が可能である; ii) 作業者に事前知識を期待できる (=評価用のツールを開発し，使用方法を指導する必要がない) という 2 点を重視したからである．

(8) に従った評定はかなり複雑な評価である．特に G と LG の区別に必要な未飽和性の概念は作業者にとって新奇なものであるため，作業者の判断が安定し，かつ作業者間で十分な一致率を見るまではそれなりの訓練と慣れを必要とした．予想されていたことだが，LG での不一致が目立った[13]．

予備調査から名詞の未飽和性は正確には程度の違いをもつ未飽和度であることが判っていた (例えば「社長」「部長」「局長」はいずれも未飽和であるが，未飽和性を感じる強さが異なる (社長 < 部長 < 局長 の順に強く，なぜか「社長」だけが例外的に未飽和性を感じさせない)．

その後の調査から，問題の未飽和性を感じる程度は統計指標 (e.g., 名詞の逆行エントロピー=過去に現われる語句の確立分布) で推測できる可能性も出てきた (ただし執筆時点では肯定的な予備調査が得られているだけ)．

2.3.1 作業結果の見本

(13) にクリーニングの結果の見本を示す:

(13) a. 革命 $_{G}$; 市民革命 $_{G}$: フランス革命
b. サイト $_{LG}$; インターネットサイト $_{G}$: ジョブシャワー
c. 企業 $_{G}$; 上場企業 $_{G}$; 東証一部上場企業 $_{G}$: 大阪瓦斯
d. アニメ $_{G}$; 版アニメ $_{B}$; 劇場版アニメ $_{G}$: 機動戦

[10] i) ChaSen/IPADic の他，ii) ChaSen/UniDic [7] v1.3.8 や iii) Juman/Jumandic も試した．iii) はスクリプトに使用した Juman には Python 用の binding がないため，実行速度の問題で主要部の正規表現の最適化では使わなかった．

[11] この数は解析で利用される辞書と未知語処理アルゴリズムに大きく依存する．

[12] なお，(9) の正規表現は IPADic, UniDic, Jumandic で共用できるようなものである．

[13] 約 2000 行に対する 4 人の作業者による評定で，G, LG, D, B のすべて評定に関して一致率を計算したところ，κ =0.492 だった．ただし，以下のように一部を欠損値とした場合，一致度は上がる: i) LG と D を欠損値とした場合，$\kappa = 0.916$; ii) LG を欠損値とした場合，κ =0.759; iii) D を欠損値とした場合，κ =0.562.

士ガンダム

e. 家 $_{LG}$; 演奏家 $_{G}$; 音楽の演奏家 $_{D}$; クラシック音楽の演奏家 $_{G}$: マグダ・タリアフェロ

f. アナウンサー $_{G}$; フリーアナウンサー $_{G}$: 朝山くみ

2.4 評価

今のところクリーニングの結果の直接の評価は行なっていないが，クリーニングで得られた上位語集合は，山田ら [9] の上位語推定研究で使われた．

3 今後の課題

3.1 下位語のクリーニング

最初に述べたように，上位語だけでなく下位語のクリーニングも必要であり，上位語に G 評価をもつ 100 万について，その作業を行なっているところである (執筆時点では未完了)．

3.2 正例の純化: 獲得精度の向上のために

(5) は，自動獲得精度向上に深く関係する本質的な問題である．実際，SVM に正例として与えられた「正解」の中に厳密には事例化の関係になっていない事例が含まれていた．これは被覆率の向上をもたらしているが，その一方で精度を下げているのは考えられることである[14]．だが，これは (6) から要求される下位語のクリーニング作業と独立に行なうことはできない．

3.3 WN-Ja との接続

整理された上位語と下位語の対の大規模データ D は固有名と普通名との繋がりをうまく表現している．従って，WN-Ja [1] のノードと D の上位ノードとが対応づけられることで，双方のデータの有用性が増すのは明らかである．語義の曖昧性解消など技術的に難しい点もあるが，それが実現できるよう作業を続ける予定である．

3.4 Wikipedia 用の上位オントロジーの構築

だが，WN-Ja との対応づけとは独立に Wikipedia の知識体系をうまく表現するような独自の上位語オントロジーを構築する必要もあるように思える．というのは，Wikipedia では架空の存在への言及が非常に多いからである．だが，これはこれで最上位の上位語の再解析という，新たな問題を提起する．

D の上位語の解析によって得られた最上位語の異なりは，どの形態素解析辞書を使うかで異なるため，「標準化」が必要である．ここで問題になるのは，形態素解析プログラムで語 (彙素)/形態素の認定基準が不統一 (かつ不明瞭) であることである．例えば Jumandic と UniDic 1.3.8 では「料理人」と「有名人」が 2 語 (彙素)/形態素だが，IPADic では「料理人」は 2 語 (彙素)/形態素，「有名人」は 1 語 (彙素)/形態素である．特に IPADic は分割不充分の印象を強くもつ[15]．これは UniDic と較べると明白である: IPADic 2.6.3 と UniDic 1.3.8 で得られた最上位語の異なりの数は，頻度 1 以上の場合で 12,846 vs. 10,589 (差は約 2 千)，頻度 2 以上の場合で 4,683 vs. 4,463 (差は約 200) である)．両者で約 1 割の違いがあり，頻度 1 の場合の差の 2 千は決して小さくない．

4 まとめ

本稿は [4] によって日本語 Wikipedia から自動獲得され，鳥式改 [6] に使われている約 244 万個の上位語と下位語の対の上位語部分のクリーニング作業の手順と結果を報告した．作業の結果，約 8 万の G 評価をもつ上位語が得られており，これらの上位語をもつ対の数は約 170 万 (元データの約 70%) である．本研究で報告した作業の結果に基づいて，現在，上位語が G 評価をもつような 100 万対の上位/下位関係の適切性のチェックを行っており，その産物は NICT が主催する「高度言語情報融合フォーラム」を通じて配信する予定である．

参考文献

[1] F. Bond, H. Isahara, K. Uchimoto, T. Kuribayashi, and K. Kanzaki. Extending the Japanese WordNet. 言語処理学会 15 回大会発表論文集. 2009. C1-4.

[2] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. Annotating noun argument structure for NomBank. In *Proceedings of LREC-2004*, Lisbon, Portugal, 2004.

[3] J. Pustejovsky. *The Generative Lexicon*. MIT Press, 1995.

[4] A. Sumida, N. Yoshinaga, and K. Torisawa. Boosting precision and recall of hyponymy relation acquisition from hierachical layouts in Wikipedia. In *Proceedings of LREC-2008*. 2008.

[5] NTT コミュニケーション科学研究所 (監修). 日本語語彙大系. 東京: 岩波書店, 1997.

[6] 鳥澤健太郎, 隅田飛鳥, 野口大輔, 柿澤康範, 風間淳一, De Saeger, Stijn, 村田真樹, 山田一郎, 塚脇幸代, 太田公子. ウェブ検索ディレクトリの自動構築とその改良 –鳥式改–. 言語処理学会 15 回大会発表論文集. 2009. P2-1.

[7] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源: 形態素解析用電子化辞書の開発とその応用, 10 2007.

[8] 黒田航, 井佐原均. 意味フレームを用いた知識構造の言語への効果的な結びつけ. 信学技報, Vol. 104 (416), pp. 65–70, 2004.

[9] 山田一郎, 鳥澤健太郎, 風間淳一, 黒田航, 村田真樹, Bond, F., 隅田飛鳥. 統語構造の特徴を利用した上位下位関係辞書の拡張. 言語処理学会 15 回大会発表論文集. 2009. B3-5.

[10] 情報通信研究機構. EDR 電子化辞書, 2003. [http://www2.nict.go.jp/r/r312/EDR/J_index.html].

[11] 西山佑司. 日本語名詞句の意味論と語用論: 指示的名詞句の非指示的名詞句. ひつじ書房, 2003.

[12] 飯田龍, 小町守, 乾健太郎, 松本裕治. 日本語書き言葉を対象とした述語項構造と共参照関係のアノテーション: NAIST テキストコーパス開発の経験から. 言語処理学会第 13 回年次大会発表論文集, 2007.

[14] もう一つの問題は，正解には未飽和な上位語を含んでいる率が高いという点である．だが，これは被覆率の低下という副作用なしに解決することは無理だと思われる．

[15] 例えば「露天風呂」「食虫植物」「節足動物」「社会学部」「桂冠詩人」「国務大臣」「政務次官」「放送時間」「総合大学」「森林公園」などが 1 語彙素である理由は不可解である．これは固有名認識を辞書の中でやろうとした結果なのかも知れないが，「二兎を追う者，一兎も得ず」になっているように思える．