日本語ワードネットの異表記対応と並行コーパスへの語義タグづけ

黒田 航
kow.kuroda@gmail.com
京都工芸繊維大学
早稲田大学

栗林 孝行
kuribayashi@nict.go.jp
jwordnet@gmail.com
NICT MASTAR Project

Francis BOND
bond@ieee.org
Nanyang Technological University
NICT MASTAR Project

神崎 享子
kanzaki@ninjal.ac.jp
国立国語研究所

井佐原 均
isahara@tut.co.jp
豊橋技術科学大学, NICT MASTAR Project

概要

日本語ワードネット (Japanese WordNet: JWN) は，バグフィックスの他に次の三点の改良を経て 1.2 に更新された．1) 異表記処理，2) 形容詞の定義の見直しと見出し語の修正，3) 英語/日本語/中国語の並行語義タグづけ (対象は「シャーロック・ホームズ」「伽藍とバザール」「京大コーパス」で合計 3,000 文強)．本発表ではそれぞれの改良の詳細について解説し，頻度の低い見出し語の表示抑制を含めた将来の改良の予定を紹介する．

1　はじめに

日本語 WordNet (以後，JWN と略記) は，2008 年に 1.0 が公開されたフリーのシソーラスである [2]，それ以来，何度か改訂を重ねている [1, 3, 12]．JWN は Princeton WordNet (以後，PWN と略記) [6] Version 3.0 の日本語訳である．過去 3 年の開発で，PWN の見出し語 (lemmas) の日本語化だけでなく，語釈文 (glosses) の日本語化も完了した [12]．更に，今回から日英中の並行コーパスへの語義タグづけ作業に着手した．今後は，SemCor (PWN の語義タグが付与された Brown Corpus の部分コーパス) の日本語化を予定している．

JWN の開発は確実に進んでいるが，次のような課題も抱えている．まず，PWN の日本語化として見る限り，現状の JWN には次の課題があると認識している:

(1) 半自動処理を元にしているため，
  a. 英語の見出し語の日本語訳が誤っている
  b. 語釈文に誤訳が残っている．

これらは一朝一夕には解消できない問題で，時間をかけて地道に修正して行く予定である．

現状の JWN には (1) に挙げた他にも，辞書としての使い勝手を大きく左右する問題がある．具体的には次である:

(2) a. 異表記が混在している．
  b. 形容詞と副詞の表示法と分類法に難がある．
  c. 実際の使用が稀で，事実上「無用な」見出し語 (e.g., サ変名詞としての「信憑」) がある．

今回は (2a) と (2b) の問題の解消に着手した．(2c) の問題は時間の都合で先送りしたが，重要な課題だと認識している．

以下では，2 で並行コーパスへの語義タグづけ作業を説明し，3 で異表記対応と形容詞の語尾の追加を説明する．

2　並行コーパスへの語義タグづけ

現在，JWN には語義の頻度の情報がない．その情報があれば synset を表示する時に，よく使う語をあまり使わない語 (頻度ゼロ) から区別をするのは簡単である (曖昧性解消のベースラインとして最高頻度を使うのは一般的である)．それとは別に，機械学習ベースの多義曖昧性解消システムを構築するには，語義タグづけされたコーパスが必要である．これらの理由から私たちは語義タグつきコーパスの開発に着手した．

今回のリリースで対象とするのは「シャーロック・ホームズ (Sherlock Holmes)」の「まだらの紐 (Speckled band)」と「踊る人形 (Dancing men)」「伽藍とバザール (The Cathedral and the Bazaar)」 [9] (769 文=全文)，「京都大学コーパス」 [7] (最初の 1000 文) で，合計 3,000 文強である．多くの人が利用できるようにするため，なるべく再配布可能な文書を選択した．

多言語で間の意味の扱いにも興味があるので，英語/日本語/中国語で並行して語義タグづけをすることにした．文数はまだ少ないが，来年以降も付与づけを継続する．タグづけされたコーパスとタグづけツールは 2011 年中に開示する予定である．

以下では日本語の語義タグづけ作業の詳細を紹介する．英語と中国語への言及は比較に留める．

2.1　タグ付与処理

並行コーパスへのタグ付与は (3) の手順で進められた:

(3) 1. 文単位で茶筌で解析し，分割と品詞を付与した．分かち書きの誤りを事後的に少し修正した．茶筌の品詞に基づいてタグづけ対象とする内容語と対象としない機能語とが区別された．
  2. 内容語が WN の見出し語に一致するなら，その synset の集合を保存した．ただし
    i. 複合表現 (multiword expression: MWE) は 3 語までのスキップを許した[1]．
    ii. 茶筌の品詞で WN の語義を絞らなかった[2]．
    WN にない内容語は拡張のための候補として保存した．
  3. 作業者は，synset の集合を一つに絞るか語義が WN で

[1] スキップを利用してうまくマッチした例は日本語にはなかったが，英語には幾つか存在した．
[2] 品詞が合わないタグが 2 割を越えた．

定義されていないと指定する[3]．整合性のために，作業者は文ごとではなく語ごとに作業をした．

4. 2 回以上出てくる WN にない語と語義が足りない語に対して，WN のエントリを作成する．
5. (3b) に戻って，拡張した WN でもう一度語義タグを付与する．再びタグづけする場合は synset の候補が増えた場合にのみ作業者に付与を依頼する．

WN を定期的にアップデートして行く予定であるが，それに合わせたコーパスの効率良いアップデートを目ざしている．これによって動的にコーパスを作成する [8]．タグ付与の早さは 100 語/時間で「檜」プロジェクトの最終スピードの半分ぐらいである [4]．スピードの差は (i) MWE も考量し (ii) 不足した語義を考量しているためと思われる．

多義性は英語が一番高く，それに中国語と日本語が続く．中国語と日本語は漢字があるため曖昧性は英語より少ないと思われる．中国語が日本語より曖昧である理由は名詞と動詞の曖昧性が原因である．

3 異表記対応と形容詞類の語尾の追加

表 1 “かわいい” の検索結果 (JWN 1.1)

Synset	Lemmas	Gloss
01808671-a:	かわいい, 甘美, スイート, 愛くるしげ, 芳しい, 愛おしい, 美味い, めんこい, スウィート, 可愛い, 愛くるしい, 香ばしい, かわいらしい, 愛らしい	感覚に気持ちよい
01462324-a:	かわいい, 可愛い, 愛しい, 大切	心から愛されている
00148642-a:	かわいい, 貴重	明らかにうまく魅了する
01459755-a:	かわいい, 可愛らしい, 愛々しい, 幼気, 愛くるしげ, 愛おしい, 愛愛しい, かわゆい, 可愛い, 愛くるしい, 愛しい, 愛らしい	特に無邪気でナイーブな態度で愛らしい
⋮	⋮	⋮
00219809-a:	かわいい, 素敵, 可愛らしい, 佳, 奇麗, 素適, 可憐, 愛々しい, 幼気, 美しい, すてき, ...	目と同様に心にうったえる

JWN の見出し語は今まではすべて同格に扱われてきた．これが原因で，幾つかの厄介が生じる．これを表 1 に示した「かわいい」の検索結果を用いながら説明する．具体的には (4)–(8) に示した 5 つの問題がある:

(4) 異表記の関係が明示されていない．表 1 について言えば，
   a.「かわいい」と「可愛い」は異表記
   b.「きれい」と「奇麗」と「綺麗」は異表記
   c.「すてき」と「素的」と「素敵」は異表記
   d.「スイート」と「スウィート」は異表記
   これらが纏められていないため，異なり語がどのぐらいあるかは数えられない．
(5) 異表記の取りこぼしがある．表 1 について言えば，
   a.「愛くるしい」があるのに「愛苦しい」がない．
   b.「すてき」「素敵」「素的」があるのに「ステキ」がない．
   c.「きれい」があるのに「キレイ」がない．
   こうなると WN を引くときやコーパスに付与しようとする時に見つかるべき語が見つからない．
(6) 誤表記か，標準的でない異表記がある．表 1 について言えば，稀な表記である「素適」が無条件に「素敵」と「素的」の異表記と認定するのは難点がある．
(7)「佳」のように，単独の語ではなく，接頭辞あるいは接尾辞としてのみ使う要素も見出し語となっている．
(8) 表記が直観的でない，あるいは日本語を母語としない話者にとってかなり不親切である．表 1 について言えば，見出し語が「大切な」「貴重な」「綺麗な」「可憐な」「幼気な」「素的な」となっていない．また，表 1 には実例がないが，「本当の」ように「の」で終わり「な」で終わらない形容語の語形が示されていない (cf. ??本当な).

(4) と (5) の問題を解決するために異表記処理を加えた．(7) と (8) の問題を解決するために形容詞の語尾を追加した．(6) については，先送りした使用頻度の低い見出し語の抑制で実現したいと考えている．

3.1 異表記対応

異表記を扱うために JWN の構造を一階層を多く設定した．1.1 版まで JWN は概念 (synset) の体系と概念と語のマッピングの二つからできていた．今後は概念と見出し語の間に標準表記を入れることにする．標準表記は標準語形と番号からなる．標準表記は基本的に同じ発音のものを一緒にする．今後，WN をオンラインで参照する時のデフォルト形は標準表記とする．例えば 01808671-a の synset は，表 2 や表 3 のようになる．

宮崎ら [15] が指摘したように，最頻の表記が常に標準形になるとは限らない．このため，どの表記が標準表記になるかは以下の条件で決める:

(9) a. jumandic に標準表記の定義があればそれに従う [13]
   b. jumandic に定義がない場合には
      i. 漢字を平仮名より優先する
      ii. 新しい字体を古い字体より優先する
      iii. カタカナなら長い方を選ぶ

語の異表記集合を得るために，JMdict [5] と jumandic を参照している．同一 synset で JMdict または jumandic で同じ語として扱われているなら，異表記集合に入れる．

01808671-a	綺麗＋な	0
01808671-a	甘美＋な	0
01808671-a	スウィート＋な	0
01808671-a	可愛い	0
...		

表 2 概念標準表記表

[3] あるいは，前処理に誤り (分かち書きや品詞) があると指定する．

綺麗	0	キレイ	綺れい 奇れい き麗 きれい
甘美	0	カンミ	かんみ
スウィート	0	スウィート	スイート
可愛い	0	カワイイ	かわいい
...			

表 3　異表記集合の表

複合語表現 (MWE) の扱いは次の通り: JWN 内の見出し語が茶筌で二形態素以上に分割され，かつそれが正しい分割なら，分割版も MWE として異表記集合に追加する．例えば，⟨機械翻訳 0⟩ には ⟨機械翻訳 0 キカイホンヤク きかいほんやく 機械_翻訳 ⟩ を入れる．これで茶筌の辞書に解析対象の MWE がなくてもマッチできるようになる．

ところで，このように WN のデータ構造を変えるとなると，それを参照しているツールも一緒に更新する必要があるため，リリースはいつもより時間がかかる．

3.2　形容詞類の語尾の追加

日本語の学校文法 (正確には「橋本文法」) では，形容詞は「(し) い」で終わる用言である[4]．これが意味するのは，

(10)　a.「きれいな」や「華麗な」のように「な」で終わる用言 (i.e., 形容動詞あるいは UniDic [14] で言う形状詞) は形容詞ではない．
　　b.「本当の」や「突発性の」のように「の」で終わる用言は形容詞ではない．

これは日本語の記述文法としては問題ないかも知れないが，JWN の見出し語形の整理に関しては厄介な問題を生じさせる．PWN の品詞は n (名詞), v (動詞), a (形容詞), r (副詞) の四つしかない (不変化詞は PWN に収録されていない)．このため，「な」で終わる修飾語と「の」で終わる修飾語を n か a のいずれかに分類する必要が生じる．これらすべてを n と分類するのは情報損失が大きい．純然たる名詞のほとんどは「な」で終わらないからである (cf. *学校な机 vs 学校の机)[5]．

すでに述べたように，「大切」「貴重」「綺麗」「可憐」「幼気」「素的」に語尾「な」がついておらず，「本当の」や「突発性の」に語尾「の」がついていない理由は，これらが日本語の学校文法では形容詞ではないからである．だが，これは多国語の概念を対応づけるという Global WordNet の使用目的に合った設定だとは言えない．

JUMAN の辞書 (jumandic) [11] は [16] に従い，次のような形容詞の品詞体系を採用している:

(11)　a.「(し) い」で終わる用言はイ形容詞
　　b.「な」で終わる用言 (i.e., 形容動詞) は「ナ形容詞」
　　c. のように「の」で終わる用言を「ナノ形容詞」[6]．

[4] 学校文法で設定されている品詞は，i) 名詞, ii) 動詞, iii) 形容詞, iv) 形容動詞, v) 副詞, vi) 連体詞, vii) 感動詞, vii) 接続詞, ix) 助詞, x) 助動詞の 10 種である．

[5] ただし「の」で終わることは名詞性の十分条件ではない．

[6] jumandic では「本当の」はナノ形容詞と分類される．「?本当な」という語形は容認度が低いので，これは本来なら「ノ形容詞」か「ノナ形容詞」と分類すべきところかと思われる．

私たちは jumandic の判断を参考にしつつ，形容詞の下位分類として次の 6 つを設けた:

(12)　a. イ形容詞: 「(し) い」で終わる用言
　　b. ナ形容詞: 「な」で終わる用言 (i.e., 形容動詞)
　　c. ノ形容詞: 「本当の」のように「の」で終わる用言
　　d. ナノ形容詞: 「な」と「の」のいずれの形でも終わることが可能であるが，「な」終わりの方が「の」終わりよりも自然な場合 (e.g., X な ≫ ??X の).
　　e. ノナ形容詞: 「な」と「の」のいずれの形でも終わることが可能であるが，「の」終わりの方が「な」終わりよりも自然な場合 (e.g., ??本当な ≪ 本当の)
　　f. その他: 「たる」や「なる」で終わる形容詞

(13)–(16) に，i) ノ形容詞，ii) ノナ形容詞，iii) ナノ形容詞，iv) ナ終わりとノ終わりが別な形容詞の幾つかの例を挙げる:

(13) ノ形容詞の例
　　a. 色取々の, ボウボウの, もじゃもじゃの
　　b. 粘着性の, 突発性の
　　c. 層状の, 液状の
(14) ナノ形容詞の例
　　a. 様々な ≫ ?様々の
　　b. 甘々な ≫ ?甘々の
　　c. 色々な ≫ ?色々の
(15) ノナ形容詞の例
　　a. 別々の ≫ ?別々な
　　b. フサフサの ≫ ?フサフサな
　　c. 生煮えの ≫ ?生煮えな
(16) ナ終わり形とノ終わり形で意味が異なる場合
　　a. 真の (勇者) ≠ 真な (命題)

この語は JWN 内では以下のように表記している: (a) イ形容詞とその他はそのまま．(b) 形容動詞はそれぞれ「+な」「+の」「+{ な, の }」「+{ の, な }」で表記する．この形は人間が見てもわかりやすいし，機械処理の際にも語尾が区別しやすい．形容動詞の切れ方は形態素解析システムによって語幹が違う．例えば「綺麗な」は茶筌では「綺麗」と「な」の 2 語に分かれ，「綺麗」が語幹だが，JUMAN では一語扱いで「綺麗だ」が語幹である．JWN はどちらにも対応できる．

3.3　サ変名詞の語尾の追加

表 4　“依頼” の検索結果 (JWN 1.1): *がついた語は現代語ではサ変名詞用法が稀有なので，この synset 中の表示を抑制する可能性がある．

07185325-n:	依頼, 申出, 申入れ, 要求, 申込, 申し出で, 求, 要望, ...	言葉による依頼
00688377-v:	信任＋する, 見込む, 頼む, 信憑＋する*, 依頼＋する, 見こむ, ...	信用，または信頼する
00753428-v:	要望＋する, 要請＋する, 頼む, 求める, ...	（人に）何かをするよう頼む

サ変名詞は同時に名詞かつ動詞の基体であるという二重の性質をもつ品詞である．これを 名詞 (n) か動詞 (v) の一方に排他的に分類するのは無理がある．JWN の検索では表 4 の “依頼”

の検索結果にある通り，サ変名詞は n と v の両方に一致する．

ただ，表示の統一感を出すため，§3 で説明した形容詞の語尾に「な」や「の」の追加するのと同様に，v の方では見出し語形に「＋する」を追加した．これにより，例えば 00753428-v の「要望」「要請」「依頼」は「要望＋する」「要請＋する」「依頼＋する」と表示される．この語尾も語を形態素の結果にマッチする場合は無視される．

サ変名詞に「する」を付与する作業は，(17) の手順で自動で処理した後に，結果を人間が見直して行なった:

(17) a. サ変活用動詞なら手をつけない[7]．
例: 発する
b. それ以外の「する」で終わるものは「＋する」に変換
例: 要望する ⇒ 要望＋する
c. ひらがなのイ段で終わらないものに「＋する」を追加
例: 要望 ⇒ 要望＋する
d. 見出し語の重複を解消
例: { 要望, 要望＋する } ⇒ 要望＋する

この処理には以下のような効果があった．(i) 表記の揺らぎ (e.g.,「する」の有無) による見かけの異なり語数を数千減らせた．(ii) サ変名詞から発生した動詞とサ変活用動詞は区別が明確になった．(iii)「要望する」のような場合はタグ処理ではマッチできるようになった．とは言え，今でも (英語版の PWN のように) すべての活用形を収録しているわけではない．

4 今後の予定

4.1 使用頻度の低い見出し語の削減

複数の日英辞書から自動的に日本語見出し語を取り出しているため，利用されない見出し語が少なからず存在する．「信憑」がサ変名詞として見出し語に挙がるが，これは現代語では通用しない用法である．

4.2 副詞の語尾の追加

今回の更新では，形容詞の概念を「な」や「の」で終わる修飾語を含むように拡張した．同じことが副詞についても必要である．だが，「しっかり (と)」のように，ゼロ語尾「しっかり」と語尾「と」のある「しっかりと」のいずれもが可能であることを明示することが必要な場合がある．

副詞の語尾の明示化は，形容詞類に較べてに対する需要大きくない点と合わせて，語尾のパターンの特定が形容詞類の場合より難しくなることが見込まれているため，先送りしている．

4.3 派生関係のリンクの追加

サ変名詞とは逆の派生関係についても使い勝手向上のための考慮が必要である．形態論が豊かな言語は，派生関係のリンクを辞書の外部にもっていた方が効率がよい．例えば表 4 で「申 (し) 出」と「申 (し) 入れ」はそれぞれ「申 (し) 出る」と「申 (し) 入れる」からの派生形であることがわかると有益である．

4.4 英語の異表記

日本語同様，英語にも異表記がある．例えば color と colour や to-night と tonight．これも日本語と同じように独立の語と異表記を区別する必要がある．

[7] サ変活用動詞は日本語文法 Jacy [10] の kurusuru-stem 配下になるかで判別した．チェックで多少のミスを見つけたので，文法の管理者に知らせて修正した

5 終わりに

本発表は，日本語ワードネット (JWN) の開発の最新情報を伝えた．日英中の並行コーパスへの語義タグづけ作業を開始した．異表記対応と形容詞の語尾の追加を行なった．

参考文献

[1] F. Bond, H. Isahara, S. Fujita, K. Uchimoto, T. Kuribayashi, and K. Kanzaki. Enhancing the Japanese WordNet. In *The 7th Workshop on Asian Language Resources*, pp. 1–8, Singapore, 2009. ACL-IJCNLP 2009.

[2] F. Bond, H. Isahara, K. Kanzaki, and K. Uchimoto. Boot-strapping a WordNet using multiple existing WordNets. In *Proc. of the 6th Intern. Conf. on Language Resources and Evaluation (LREC-2008)*, 2008.

[3] F. Bond, H. Isahara, K. Uchimoto, T. Kuribayashi, and K. Kanzaki. Extending the Japanese WordNet. In **言語処理学会** *15* **回大会発表論文集**, pp. 80–83, 2009.

[4] Francis Bond, Sanae Fujita, and Takaaki Tanaka. The Hinoki syntactic and semantic treebank of Japanese. *Language Resources and Evaluation*, 42(2):243–251, 2008.

[5] James Breen. JMdict: A Japanese-multilingual dictionary. In *Proc. of the Workshop on Multilingual Linguistic Resources*, MLR '04, pp. 71–79, Stroudsburg, PA, USA, 2004.

[6] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[7] Sadao Kurohashi and Makoto Nagao. Building a Japanese parsed corpus — while improving the parsing system. pp. 249–260. Kluwer Academic Press, 2003.

[8] Stephan Oepen, Dan Flickinger, and Francis Bond. Towards holistic grammar engineering and testing: Grafting TreeBank maintenance into the grammar revision cycle. In *Beyond Shallow Analyses — Formalisms and Statistical Modelling for Deep Analysis (Workshop at IJCNLP-2004)*, Hainan Island, 2004.

[9] Eric S. Raymond. *The Cathedral & the Bazaar*. O'Reilly, 1999.

[10] Melanie Siegel and Emily M. Bender. Efficient deep processing of Japanese. In *Proc. of the 3rd Workshop on Asian Language Resources and Intern. Standardization at the 19th Intern. Conf. on Computational Linguistics*, pp. 1–8, Taipei, 2002.

[11] 京都大学. 日本語形態素解析プログラム寿満 (juman). `http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html`.

[12] 栗林 孝行, Francis Bond, 黒田 航, 内元 清貴, 井佐原 均, 神崎享子, and 鳥澤健太郎. 日本語ワードネット 1.0. In **言語処理学会第** *16* **回年次大会発表論文集**, pp. 978–981, 2010.

[13] 岡部 浩司, 河原 大輔, and 黒橋禎夫. 代表表記による自然言語リソースの整備. In **言語処理学会第** *13* **回年次大会発表論文集**. 2007.

[14] 小椋 秀樹, 小磯 花絵, 冨士池 優美, and 原 裕. 『現代日本語書き言葉均衡コーパス』形態論情報規程集, 2008. UniDic の単位に関して.

[15] 宮崎 正弘, 池原 悟, and 横尾 昭男. 単語結合型辞書引きを用いた日英機械翻訳辞書の構成. **電子情報通信学会，** *NLC91-18*, pp. 15–22, 1991.

[16] 益岡 隆志 and 田窪 行則. **基礎日本語文法**. くろしお出版, 改訂版 edition, 1992.