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[*] 栗林 孝行[**] Francis BOND[**,#] 神崎 享子[*****] 井佐原 均[**,##]

*京都工芸繊維大学/早稲田大学総合研究機構, **NICT MASTAR Project,

#Nanyang Technological University, ****国立国語研究所, ##豊橋科学技術大学

言語処理学会第17回年次大会 (2011/03/10 (Sat), 豊橋技術科学大学)

発表の概要

- **日本語ワードネット (JWN)** (Bond et al. 2008, *et seq*; 栗林ら 2010) **の最新更新の報告**
  - 並行コーパスへのタグづけ
  - 異表記対応 (ただし現時点では限定的)
  - 形容詞の定義の見直しと見出し語形の修正
  - サ変名詞の見出し語形の修正
- **それぞれについて説明**
- **補足**
  - JWNは Princeton WordNet (Fellbaum, ed. 1998)の日本語訳

並行コーパスへの語義タグづけ

なぜ語義タグづけか?

- **語義タグつきコーパスへの需要は大きい**
  - 語義頻度を知りたいとかジャンルごとの語義分布を知りたいという需要は以前から
  - このデータがあれば語義の曖昧性解消タスクの精度向上が可能
- **解決策1**
  - SemCor (Miller et al. 1993, http://www.cse.unt.edu/~rada/downloads.html#semcor) の日本語化 (Tim Baldwin が担当し作業が進行中)
    - **SemCor は Brown Corpus の一部 (360,000語) にWNの語義タグを付与したdata**
- **解決策2**
  - 並行コーパスへの*X*WN語義タグづけ (*X*= English, Japanese, Chinese)

RWCコーパスとの比較

- **RWCコーパスは**
  - 岩波国語辞典の語義 ID が付与されている3000個の新聞記事データ
- **RWCの嬉しくない点**
  - 使われている語義 (岩波国語辞典) が多言語対応ではないので，並行コーパスへの語義タグづけには不適
  - RWCコーパスはフリーではない
    - **値段はともかく，商用利用不可というのは痛い...**

英日中の並行テキスト

- **『シャーロック・ホームズ』**
  - 「まだらの紐」と「踊る人形」(合わせて1400文)
- **『伽藍とバザール』(769文)**
- **『京都大学テキストコーパス』 (最初の1000文)**

語義タグづけの手順 (日本語の場合)

1. **茶筌 (Chasen) を使って文単位で単語分割と品詞付与**
   - 内容語 (N, V, A, R)のみをタグづけの対象とする
2. **a. JWNに適切な語義があれば選択して保存する**
   - アノテーションツール上で “s” (for “save”) を選択
2. **b. JWNに適切な語義がなければ拡張候補として保存**

3. **作業者が3の結果を分類**
   - synset を一つに絞るか synset が未定義だと指定するか
4. **2回以上出現する未知語と見出し語が不足している語についてWNのエントリーを作成**
5. **2に戻って拡張したWNで再度語義タグ付与**

タグづけツール in Action

知見と今後の予定

- **知見**
  - 語の多義性の度合いは英語が最高
    - それに中国語と日本語が続く
  - 英語より中国語と日本語の曖昧性が低いのは漢字の効果?
    - これはタグ付与作業の精度に依存する
  - 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
- **今後の予定**
  - 語義タグつきコーパスとタグづけツールを 2011 年中に公開予定

異表記対応

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

- **異表記**
  - かわいい, 可愛い
  - きれい, 奇麗, 綺麗
  - すてき, 素的, 素敵
  - スイート, スウィート
- **異表記の取りこぼし**
  - 愛苦しい
  - ステキ
  - キレイ
- **見出し語が直観に合わない**
  - 大切な, 貴重な, 綺麗な, 可憐な, 幼気な, 素的な

異表記対策

- **v1.1 版まで**
  - 概念と見出し語を直接対応づけ
- **v1.2 以後**
  - 概念と見出し語の間に “標準表記” を入れる
- **定義**
  - 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
  - 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

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

- v1.2以降のJWNの構造は
  01808671-a の synset は表 2 や表 3 のようになる

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

複合表現 (MWEs) の扱い

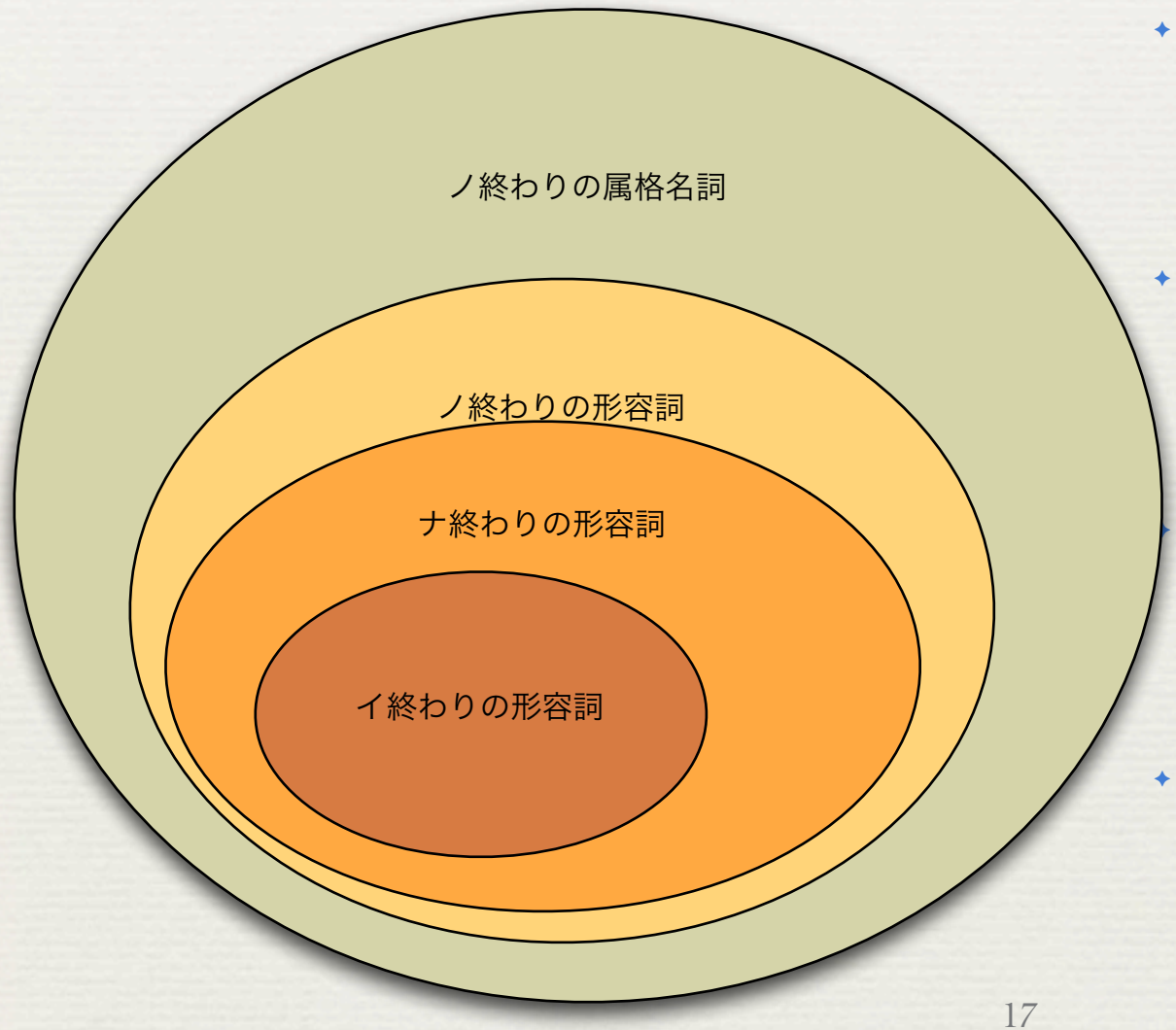
- **方針**
  - 見出し語 w が茶筌で二形態素以上 $u+v+\cdots$ に分割され，それが正しい分割なら，w と $u+v+\cdots$ の両方を異表記集合に追加
- **例えば**
  - 〈機械 翻訳 0〉には〈機械翻訳 0 キカイ ホンヤク きかいほんやく 機械 翻訳 〉を入れる
- **効果**
  - 茶筌の辞書に解析対象の MWE がなくてもマッチ可能に

形容詞の定義の見直しと
見出し語形の修正
+サ変名詞の扱いの変更

品詞情報のズレ

- **JWN には名詞 (-n), 動詞 (-v), 形容詞 (-a), 副詞(-r) の四品詞しかない**
  - JWNは Princeton WN 3.0 の品詞情報を継承
- **問題**
  - 日本語の標準的な品詞体系とPWNの品詞とのズレをどうする?
- **具体的には**
  - 橋本文法の形容詞動詞=UniDicの形状詞の扱いをどうする?
  - サ変名詞の扱いをどうする?

形容詞の階層



- イ形容詞 (=学校文法で言う形容詞)
  - 丸い, 悪い, 軽い, 苦しい
- ナ形容詞 (=学校文法で言う形容動詞)
  - 大きな, 真な
- ノ形容詞
  - 真の, 本当の, ウソの, 突発性の
- 注意
  - ノで終わるものすべてが名詞というわけではない

決め手になる例

- ナ形とノ形の意味の差別化を伴う共存
- 例: 真な ≠ 真の
  - *真な勇者 << 真の勇者
  - 真な命題 > ?真の命題

分類の詳細

1. イ形容詞: “+(し) い” で終わる用言
2. ナ形容詞: “+な”で終わる用言 (i.e., 形容動詞)
3. ノ形容詞: “+の” で終わる用言
4. ナノ形容詞: “+な” と “+の” のいずれでも終わるが, “な” 終わりの方が自然な場合
5. ノナ形容詞: “な” と “の” のいずれでも終わるが, “の”終わりの方が自然な場合
6. その他: “たる” や “なる” で終わる形容詞

- 益岡・田窪 (1992) とJUMANの辞書構築方針を参考にした

ナ形とノ形の優先度つきの共存

- ナ形 > ノ形
  - 様々な >> ?様々の
  - 甘々な >> ?甘々の
  - 色々な >> ??色々の
- ナ形 < ノ形
  - ?別々な << 別々の
  - フサフサな << フサフサの
  - ??生煮えな << 生煮えの

境界例

- いや待て．ノ形とナ形は語義が違う???
  - バラバラ
    - バラバラの死体 > バラバラな死体
    - バラバラの意見 < バラバラな意見
  - モジャモジャ
    - モジャモジャのヒゲ >> モジャモジャなヒゲ
    - (毛が)モジャモジャの犬 < (毛が) モジャモジャな犬

境界例

- イ形とナ形と共存
  - 大きい, 大きな
  - 身近な, 身近い*

サ変名詞の見出し語形修正

手順

- サ変活用動詞なら手をつけない
  - 例: 発する
- 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
- 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
- 見出し語の重複を解消
  - 例: {要望, 要望+する} ⇒ 要望+する

修正後の表記

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

修正後の表記

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

まとめ

現状

- 英日中の並行コーパスへの語義タグづけを開始
  - タグつきコーパスとタガーは H23年度内に公開予定
- 限定的だが異表記対応を行なった
- 形容詞類の語尾 (e.g., “な”, “の”) の追加した
- サ変名詞にダミー動詞 “する” を追加した

参照文献

- Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto (2008). "Boot-strapping a WordNet using multiple existing WordNets." In *Proc. of the 6th In- tern. Conf. on Language Resources and Evaluation (LREC-2008)*, 2008.
- Fracis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki (2009). "Enhancing the Japanese WordNet." In *Proc. of The 7th Workshop on Asian Language Resources*, pp. 1–8, Singapore, 2009.
- Francis Bond, Hitoshi Isahara, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki (2009). "Extending the Japanese WordNet." In 言語処理学会 15 回大会発表論文集, pp. 80–83.
- Christiane Fellbaum, *ed*. (1998) *WordNet: An Electronic Lexical Database*. MIT Press.
- 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 (2010). "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.
- 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.
- 益岡 隆志 and 田窪 行則. 基礎日本語文法 (改訂版). くろしお出版, 1992.
- George Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. (1993). "A semantic concordance." In *Proc. of the 3 DARPA Workshop on Human Language Technology*.

Thank You
for
Your Attention