コーパスに基づくWordNetの多言語化

綱川　隆司　　梶　博行
静岡大学情報学部情報科学科

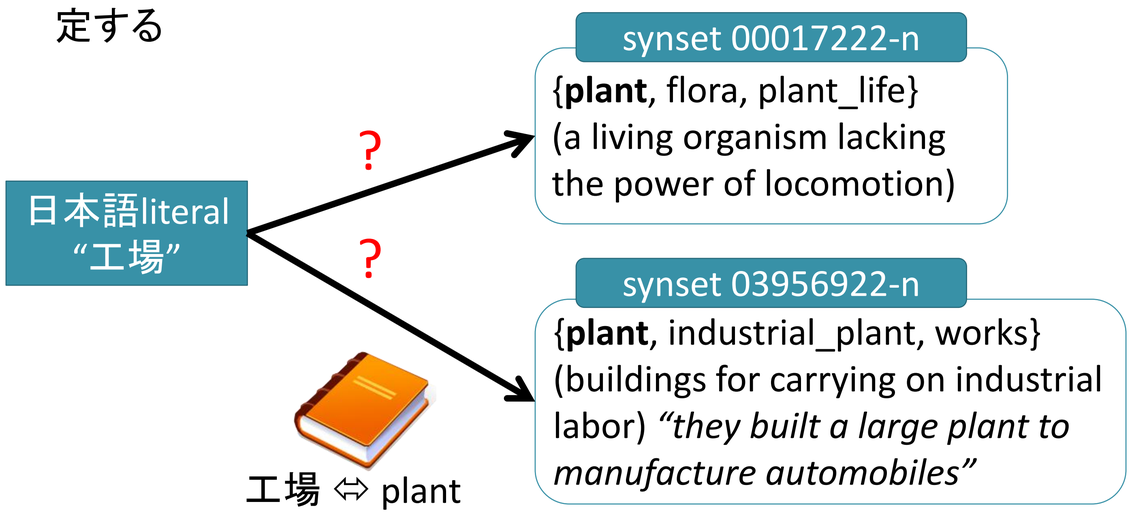
はじめに

背景

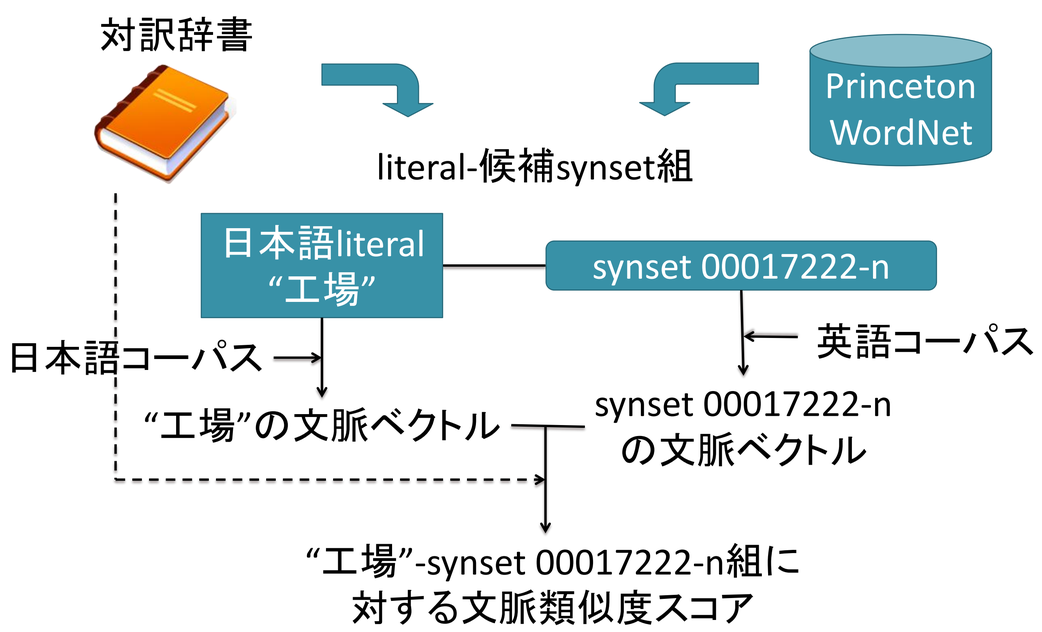
- WordNet等の意味論的情報をもつ語彙資源を多言語化することにより、言語横断的NLPの重要な基盤になることが期待される
- 英語によるPrinceton WordNet (Fellbaum, 1998) に対し、形式の互換性や言語間リンクをもつ他の言語版のwordnetが公開されている

「拡張的アプローチ」と曖昧性の問題

- 英語のliteralが曖昧性を持つ語の場合、もとのsynsetに対応しない他言語の語がliteralとして追加されてしまう
- コーパスから得られるsynsetの文脈と、追加する言語のliteralの文脈の類似度を測ることにより、追加すべきliteral-synset対を判定する



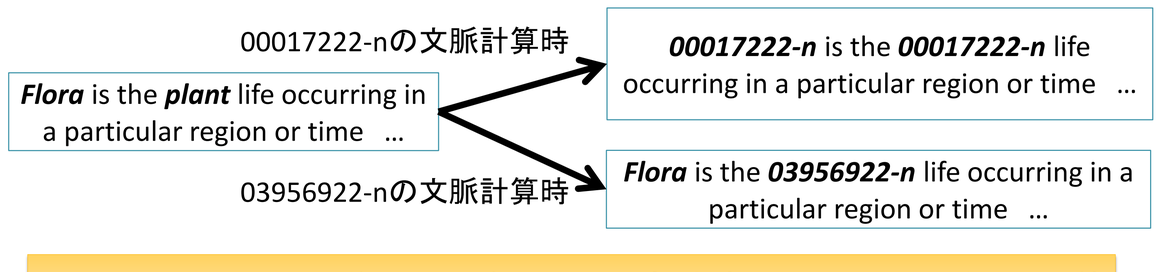
手法



Synsetに対する文脈の獲得

- Synsetの周辺文脈をコーパスから直接得るには、意味タグ付きコーパスを利用する必要があるが、利用可能なコーパスは限られ、また出現頻度の小さいsynsetは十分な文脈が得られない

→各synsetについて、それが含むliteralのコーパス中の出現をすべてsynsetに置き換えて文脈を求める



- Synsetのgloss（説明文・例文）に含まれる語や、下位・兄弟synsetのliteralの文脈を用いてバイアスをかける

$$\mathrm{Sim}_{\mathrm{bias}}(s, w_i) = \min(\mathrm{Sim}(s, w_i), \max(\max_{s' \in \mathrm{R}(s)} \mathrm{Sim}(s', w_i), \max_{w' \in \mathrm{G}(s)} \mathrm{Sim}(w', w_i)))$$

sim(w_1, w_2): 単語w_1とw_2の共起頻度に基づく関連度
sim(s, w): synset sと単語wのsynset置換による関連度
R(s): synset sの下位（＋兄弟）synsetのliteralの集合
G(s): synset sのglossに含まれる名詞の集合

日本語WordNetの「拡張的アプローチ」

- 日本語WordNet (Isahara et al., 2008) では、Princeton WordNetのsynset（同義語集合）に対して日本語の語や説明等を付加することで構築を行っている
- このような「拡張的アプローチ」により、各synsetに含まれる語(literal) について対訳辞書を用いて対応する他言語の語を付加することにより、低コストで形式の互換性および言語間リンクを保った他言語版wordnetが構築できる

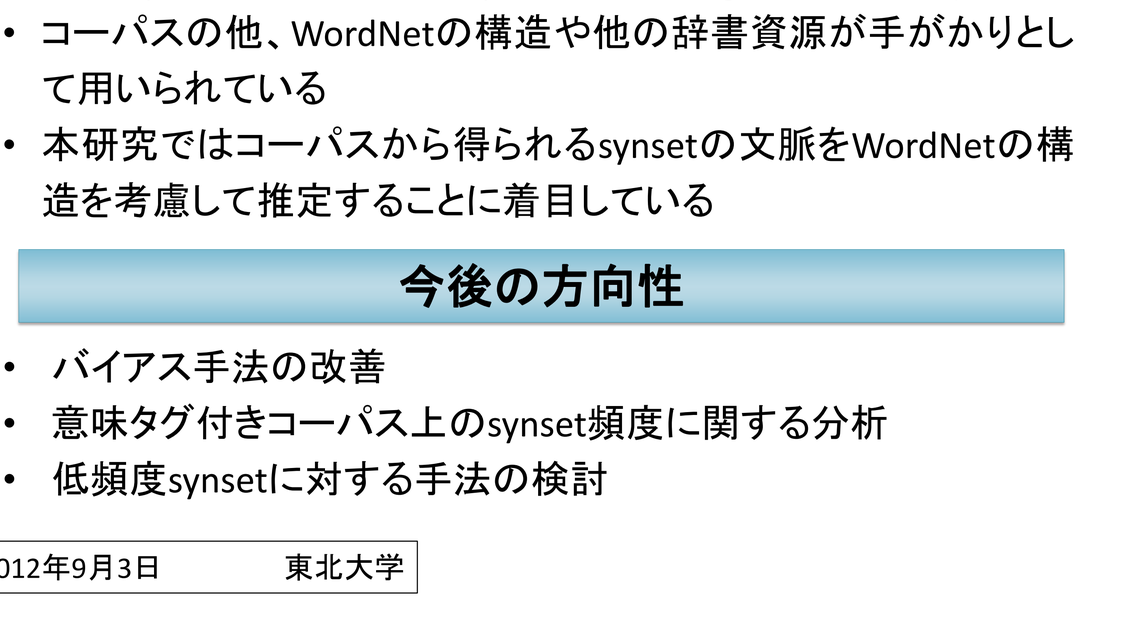
実験

実験設定

- Princeton WordNet 3.0のsynsetに対し、日本語を付与して日本語WordNet 1.1 を用いた評価を行う
- Synset 00002684-n {object, physical_object} の下位にあるsynsetについて、日本語literal-synset対とその類似度スコアを求める
  - 日本語literalはコーパス出現頻度上位で、以下の条件を満たす1871語
    - 対訳辞書に英訳が存在し、その英訳を含むsynset（候補synset）が2つ以上50以下ある
    - 候補synsetの中に日本語literalが含まれる「正解synset」が少なくとも1つある
  - 候補synset数の平均は12.83個、正解synset数の平均は2.66個
- 英語コーパス：Gigaword 4th Edition New York Times (2000-2008)
- 日本語コーパス：毎日新聞コーパス(2000-2008)
- 対訳辞書：EDR電子化辞書・日英対訳辞書
- 文脈ベクトルは、ウィンドウ共起頻度に基づく単語間関連度 discounted log-odds ratio (Evert, 2005) から求める
- 文脈ベクトル間の類似度: Dice coefficientの変形

実験結果

- Glossのみ、または下位・兄弟synsetのみによるバイアスをかけた場合に比べ、全てを考慮した場合に正解率を1～3ポイント程度改善する結果が得られた



関連研究

- 「拡張的アプローチ」によるWordNet多言語化、およびそれに伴う曖昧性解消に対処した多数の既存研究がある (Lee et al., 2000; Sathapornrungkij and Pluempitiwiriyawej, 2005; Kaji and Watanabe, 2006; Charoenporn et al., 2008; Montazery and Faili, 2010; Sagot and Fišer, 2011)
- コーパスの他、WordNetの構造や他の辞書資源が手がかりとして用いられている
- 本研究ではコーパスから得られるsynsetの文脈をWordNetの構造を考慮して推定することに着目している

今後の方向性

- バイアス手法の改善
- 意味タグ付きコーパス上のsynset頻度に関する分析
- 低頻度synsetに対する手法の検討

NLP若手の会第7回シンポジウム　　2012年9月3日　　東北大学