2B2-01

国語辞典の語釈文の解析と語義のシソーラスへのマッピング

Linking word senses in a dictionary to semantic classes in a thesaurus

正津康弘 Syotu Yasuhiro　白井清昭 Shirai Kiyoaki　徳永健伸 Tokunaga Takenobu　田中穂積 Tanaka Hozumi

東京工業大学 大学院情報理工学研究科
Department of Computer Science, Tokyo Institute of Technology

Linguistic knowledge plays crucial role in natural language processing. Constructing large linguistic knowledge requires a lot of human effort and much cost. There has been many attempts to construct linguistic knowledge automatically, which can be categorized into two classes, namely, attempts to extract knowledge from annotated corpora, and attempts to enlarge existing knowledge by using annotated corpora. This paper describes algorithms to enlarge existing linguistic knowledge by combining with another linguistic resources. More concretely, this algorithm links a word sense defined in a monolingual dictionary to semantic classes in a thesaurus. Experiments showed that the precision of linking was 85.5% and its coverage was 61.4%.

1. はじめに

情報検索や翻訳などの自然言語処理アプリケーションにおいて，言語知識は非常に重要な役割をはたす．しかしながら，言語知識を人手によって構築するには多大な労力を要する．この問題を解決するためにコーパスや機械可読辞書などの言語データからさまざまな言語知識を自動的に獲得する研究が数多くおこなわれてきた [8]．言語知識を自動構築する手法は大きくわけて，コーパスなどの比較的「生」の状態に近い言語データから知識を獲得するアプローチ [1, 5, 7] と既存のシソーラスや辞書などの知識を基礎として，それらを拡張するアプローチ [2, 6, 4] に分類できる．前者のアプローチでは知識の骨格から獲得することを狙っているので十分な精度が得られない場合が多い．本論文では，基本的に後者のアプローチをとるが，既存の知識を拡張する際の知識源として，コーパスなどの「生」に近い言語データを用いるのではなく，別の既存の知識と組み合わせることによって知識を拡張する．

本論文では，日本語の国語辞典と日本語の語を分類したシソーラスを対象とし，国語辞典で定義された名詞の語義とシソーラスの名詞の意味クラスの対応関係を自動的につけることにより言語知識を拡張する手法を提案する．国語辞典は語の意味 (語義) を自然言語表現によって定義しており，その語自身に関する記述が中心である．また，ひとつの語 (見出し語) について一般には複数の語義を定義している．語を定義する語釈文には異なる語義の間の関係についての記述も含まれることがある．これに対してシソーラスは語をあらかじめ定められた意味クラスにそって分類したものであり，語と他の語の関係を中心に整理されているという特徴がある．このように国語辞典とシソーラスはそれぞれ語の別の観点から記述しているといえる．このように性質の異なる言語知識の間の対応関係を同定することによって，各語に関する情報を豊かにすることができる．

2. 言語データ

本論文では，RWCP によって形態素タグが付与された岩波国語辞典第 5 版のデータ [3] と NTT によって作成された日本語語彙体系 [?] を使用した．岩波国語辞典は 51,438 語の名詞の見出し語を持っている．各見出し語にはひとつ以上の語義が定義されている．一方，日本語語彙大系は，2,710 の名詞意味クラスを定義しており，264,312 語の名詞をこれらの意味クラスに分類している．各語はひとつ以上の意味クラスに含まれることもあり，語が含まれる意味クラスはその語の上位概念に相当する．また，日本語語彙体系は意味クラスを節点とする木構造を成しており，親の意味クラスは子の意味クラスよりも上位の概念を表している．日本語語彙体系の一部を図 1 に示す．

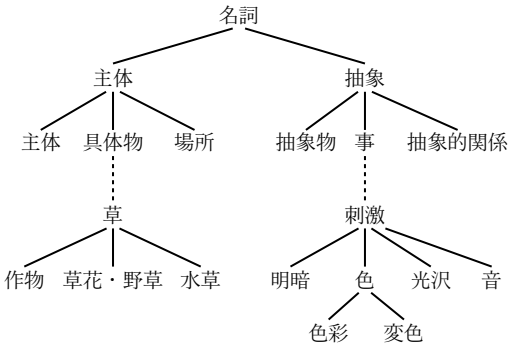

図 1: 日本語語彙体系の一部 (茜の意味クラス)

3. 上位語を利用したリンク

本節では，国語辞典の語釈文から見出し語の上位語を抽出し，これを用いて辞典の各語義にシソーラスの意味クラスを対応付ける方法について述べる．

3.1 上位語の抽出

各語義の語釈文の最初の文の末尾が次のような形をしているとき，⟨名詞⟩ を上位語として抽出する [?, ?, ?].

- ⟨名詞⟩
- ⟨名詞⟩ すること
- ⟨名詞⟩ をすること
- ⟨名詞⟩ の一つ

連絡先: shotsu@cl.cs.titech.ac.jp

- ⟨名詞⟩ の一種
- ⟨名詞⟩ の略
- ⟨名詞⟩ の〜称

(例) アーク灯： 放電を利用した 電灯
悪循環　： 互いに影響し合って，とめどなく 悪化 すること
青写真　： 図面などの複写に使う 写真 の一種
赤切符　： もとの汽車の 三等 乗車券 の通称

ただし，⟨名詞⟩ が非自立語や「こと，事，もの，物，略，一つ，一種，語，方，〜称」などの語の場合は抽出しない．

3.2 シソーラスのパス長に基づく対応付け

各語義にシソーラスの意味クラスを対応付ける方法として，見出し語の意味クラスと抽出した上位語の意味クラスをつなぐパスの長さが最小になるような意味クラスの組を見つけ，その見出し語の意味クラスを語義に対応付ける，という方法が考えられる．

しかし，日本語語彙大系では，葉の意味クラスの深さが均一ではないため，階層一段階分のクラスの意味の変化の度合いが均一ではない．たとえば，“悪人” と “魔物・化け物” はどちらも葉の意味クラスだが，これらのふたつの語から根節点 “名詞” までパスは以下のようになっている[*1]．

- 悪人→悪人等→善人・悪人等→人間＜性向＞→人間＜能力・性向＞→人間→人→主体→具体→名詞
- 「魔物・化け物→準人間→人→主体→具体→名詞

これを見ると，同じ一段階分でも “悪人” → “悪人等” の間の方が “魔物・化け物” → “準人間” の間に比べて意味の変化が少ないことがわかる．このような格差があるため，純粋にパスの長さだけで判定しようとすると間違いが生じる可能性が高い．

また，パスの長さが対応させると必ず最小のものを，語義に対応するシソーラスの意味クラスが存在しない場合も，誤った意味クラスを対応付けてしまうという問題が起きる．たとえば，国語辞典の「藍」の語義は「秋，穂状の赤い小花をつける，たで科の一年生植物.」と定義されているが，シソーラスの「藍」の意味クラスは “染料” と “色彩” しかなく，植物を表す意味クラスは定義されていない．しかし「藍」と「植物」はどちらもシソーラスに登録されている名詞であるため，ふたつの単語の意味クラスをつなぐパスは存在する．そのため，国語辞典で定義されている語義は植物を表すにも関わらず，この語義をシソーラスの “染料” か “色彩” のどちらかの意味クラスに対応付けてしまう．

以上の問題に対処するために本論文では，以下のような方法で対応付けをおこなう．

(1) シソーラスのすべての意味クラス c に以下のような重み $W(c)$ を付ける．

$$W(c) = \begin{cases} 100 & (d < 3) \\ 1/d & (d \geq 4, C = \phi) \\ \dfrac{|C|}{\displaystyle\sum_{c_i \in C} \dfrac{1}{W(c_i)}} & (d \geq 4, C \neq \phi) \end{cases}$$

d は c の深さ，C は c の子クラスの集合，$|C|$ は c の子クラスの数である．意味クラスの深さは根節点を 0 として計算する．

(2) 見出し語の意味クラスと上位語の意味クラスをつなぐパスの上の節点の重みの総和を，ふたつのクラスの間の距離とする．

(3) 距離が最小になる組の見出し語側の意味クラスを語義に対応付ける．ただし，最小の距離が 100 以上の場合は対応付けをおこなわない．

根から葉に至るまでのパスが長い (葉の位置が深い) ほど階層一段階分の意味変化の度合は小さい傾向がある．そのため，このような重み付けをすることによって変化の格差の影響を少なくすることができる．また，深さ 3 以下の意味クラスを経由しなければならないような意味クラスの組は近い意味を持つとはいいがたい．そこで，このような場合に対応付けをおこわないようにすれば，シソーラスに存在しない語義に誤って対応付けをする場合を減らすことができる．

3.3 表記のゆれと複合語の扱い

岩波国語辞典と日本語語彙大系の間には表記方法などの違いがあるため，対応付けの際に次のような調整をおこなった．

- 国語辞典の見出し語に異なる表記方法がある場合は，すべての表記方法で距離を測定する．(例：明 (か) り→明かり，明り)
- 本来は漢字で表記すべき上位語がひらがなで表記されている場合は岩波国語辞典によって漢字に変換する．同音異義語がある場合は，すべての語で距離を測定する．
- シソーラスに存在しない複合名詞が上位語になる場合は，複合名詞を構成する名詞の一番左の名詞から削除し，シソーラスに存在するかどうかをその都度調べる．(例：一年 | 生 | 植物→生 | 植物→植物)

3.4 実験結果

実験の結果，国語辞典の 27,853 個の語義に対応付けをすることができた．国語辞典にもシソーラスにも存在する名詞の語義の総数は 55,875 個なので被覆率は 49.8%となる．対応付けをおこなった語義をランダムに 100 個選出し，それらの正解不正解を主観的に判断し，精度を求めた．その結果，上位語による対応付けの精度は 84.5%となった．

4. 語釈文の動詞を利用したリンク

本節では，語釈文の最初の文の末尾が「⟨動詞⟩ こと」というパタンを利用して語義にシソーラスの意味クラスを対応付ける方法について述べる．

4.1 動詞に基づく対応付け

語釈文が「⟨動詞⟩ こと」というパタンの語義を持ち，同じ動詞を共有する見出し語のグループは，共通の意味クラスを持っていることが予想できる．さらに，この共通の意味クラスは「⟨動詞⟩ こと」という概念を表わす意味クラスであることが予想できる．たとえば，語釈文の最後が「飲む こと」となっている語義を持つ見出し語のほとんどは，シソーラスの “飲み” という意味クラスを持っている．そこで,「⟨動詞⟩ こと.」の ⟨動詞⟩ を共有する見出し語の意味クラスの分布を利用して対応付けをおこなう．

(1) 語義の語釈文の最初の文の末尾が「⟨動詞 v⟩ こと」となっている見出し語 w が，意味クラス c を持つ頻度 $fr(v, c)$ を求める．

*1 シソーラスの意味クラスは “...” でくくって表わす

(2) $fr(v,c) \geq 2$ となる (v,c) の組全てに対して以下の式により $P(c|v)$ を求める．

$$P(c|v) = \frac{fr(v,c)}{\sum_{c_i \in C_v} fr(v,c_i)}$$

(3) 見出し語 w の意味クラスが $c_{w1} \ldots c_{wn}$ で，w の語義「〈動詞〉こと」の v に対応する意味クラスが $c_{v1} \ldots c_{vm}$ のとき，各意味クラス間の距離 $d(c_{wi}, c_{vj})$ を 3.2 の方法で求める．

(4) $d'(c_w, c_v) = d(c_w, c_v)/P(c_v|v)$ を求め，最小になる $d'(c_w|c_v)$ の c_w を語義に対応付ける．

$d(c_w, c_v)$ を $P(c_v|v)$ で割ることによって，c_v が v の見出し語の意味クラスとしてあまり頻繁に現れない場合に c_w と c_v の間の距離を大きくし，対応付けの優先度を低くしている．$fr(c,v)$ が 1 の場合に無視するのは，自分しか持っていない意味クラスを対応付けてしまうのを防ぐためである．この対応付けの際にも，3.3 と同様に，見出し語の異表記に対する調整をおこなった．

4.2 結果

実験の結果，6,486 個の語義に意味クラスを対応付けることができ，精度は 90.0%となった．3. 節の結果とあわせると，合計 34,339 個の語義にリンクできたことになる．被覆率は 61.4%，精度は 85.5%となる．

5. Linksense によるリンク

Jen Nan Chen と Jason S. Chang は辞典とシソーラスの間のリンクをおこなうアルゴリズム Linksense を提案している [?]. 3. 節や 4. 節の方法では，語釈文の最初の文の末尾付近の 1 語から語義と意味クラスの距離を判断しているのに対し，Linksense は語釈文中のすべての名詞から語義と意味クラスの近さを判断するアルゴリズムである．Chen らは *Longman Dictionary of Contemporary English* (LDOCE) と *Longman Lexicon of Contemporary English* (LLOCE) を用いた実験をおこなっている．Linksense アルゴリズムを岩波国語辞典と日本語語彙体系に適用し，我々の手法と比較をおこなった．

5.1 Linksense による対応付け

語義 d と意味クラス c の近さ $Sim(D,c)$ を次の式で求める．

$$Sim(d,c) = \frac{\sum_{d_i \in KEY_d} 2 \cdot w_{d_i} \cdot In(d_i, c)}{|KEY_d| + 1}$$

ここで KEY_d は語義 d の語釈文に含まれる名詞の集合，$|KEY_d|$ は名詞の数，w_k は k の持つ意味クラスの数の逆数である．また，B 自体か B の先祖か子孫の意味クラスを語義 a が持っているならば，$In(a,B) = 1$，そうでないならば，$In(a,B) = 0$ とする．そして，$Sim(d,c)$ が最大になる意味クラス c を語義 d に対応付ける．$Sim(d,c)$ がすべてある閾値 (ここでは 0 とする) 以下ならば対応付けをおこわない．

5.2 表記のゆれの扱い

Linksense についても対象が日本語なので，対応付けの際に 3.3 と同様に以下のような調整をおこなった．

- 見出し語に異なる表記がある場合はすべての表記で距離を測定する．
- 本来は漢字で表記すべき上位語がひらがなで表記されている場合は岩波国語辞典によって漢字に変換する．同音異義語がある場合は変換をおこなわない．
- 語釈文中に見出し語と同じ名詞が現れた場合は無視する．
- かぎ括弧 “「」” の間に見出し語と同じ名詞が単独で現れた場合，かぎ括弧内の文は用例の文であることが多いのでかぎ括弧内の名詞はすべて無視する．

5.3 結果

実験の結果，38,045 個の語義に意味クラスをリンクさせることができた．被覆率は 68.1%，精度は 79.5%となった．

6. アルゴリズムの比較

表 1 は，我々の手法と Linksense を被覆率と精度について比較したものである．我々の手法と Linksnse の両方で対応付けをおこなうことができた語義や片方でしか対応付けできなかった語義について調べたところ，表 2 のようになった．

アルゴリズム	語義数	被覆率	精度
提案手法	34,339	61.4%	85.5%
(上位語による)	27,853	49.8%	84.5%
(動詞による)	6,486	11.6%	90.0%
Linksense	38,045	68.1%	79.5%

表 1: 提案手法と Linksense の被覆率と精度

対応付けできたアルゴリズム	語義数	比率	精度	
			提案手法	Linksense
両アルゴリズム	27,199	48.7%	86.1%	86.4%
(重複あり)	24,457	43.8%	87.0%	87.0%
(重複なし)	2,607	4.7%	82.0%	85.0%
提案手法のみ	7139	12.8%	83.0%	—
Linksense のみ	10,846	19.4%	—	57.0%

表 2: 両アルゴリズムの結果の重複

これらの結果から，それぞれのアルゴリズムに次のような特徴があることがわかる．

- 提案手法は語義との関連が最も強い意味クラスひとつ対応付けることが多いが，Linksense は語義と関連のある意味クラスなら関連の強さに関わらず複数と対応付ける傾向にある．
  例) 愛嬢：親がかわいがっている娘．まなむすめ．
  提案手法：“娘” という意味クラスを対応付ける
  Linksense：“女” と “娘” のふたつつの意味クラスを対応付ける
- Linksense は語義と関連はあるが上位概念としてはふさわしくない意味クラスを対応付けてしまうことが多い．
  例) 愛書：本が好きなこと．
  Linksense：“出版物” と “本（内容）”

- 提案手法は必ず語義の語釈文の最初の文に記述されている概念を表す意味クラスに対応付けようとするが，Linksenseは第二文以降に記述されている概念に対する対応付けを行うことがある．
  例) 赤樫：ぶな科の常緑高木．高さ一〇メートル以上になり，材は赤く，質が堅いので用途が広い．
  提案手法：“樹木（その他）” Linksense：“材木”

- どちらのアルゴリズムも，見出し語が持つ意味クラスの中に正解となる意味クラスがあるのに別の意味クラスを対応付けてしまうという間違いよりも，見出し語が正解となる意味クラスを持っていないのにどれかの意味クラスを対応付けてしまったという間違いの方が多い．

7. おわりに

参考文献

[1] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of English words. In *Proceedings of 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, 6 1993.

[2] Philip Resnik. A class-based approach to lexical discovery. In *Proceedings of 30th Annual Meeting of the Association for Computational Linguistics*, pages 327–329, 7 1992.

[3] RWCP テキスト・サブ・ワーキンググループ. RWC テキストデータベース，rwc-db-text-96-2, 1996.

[4] T. Tokunaga, A. Fujii, M. Iwayama, N. Sakurai, and H. Tanaka. Extending a thesaurus by classifying words. In *Proceedings of the ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources*, pages 16–21, 1997.

[5] Takenobu Tokunaga, Makoto Iwayama, and Hozumi Tanaka. Automatic thesaurus construction based on grammatical relations. In *Proceedings of IJCAI '95*, pages 1308–1313, 1995.

[6] Naohiko Uramoto. Positioning unknown words in a thesaurus by using information extracted from a corpus. In *Proceedings of COLING '96*, pages 956–961, 1996.

[7] Akira Ushioda. Hierarchical clustering of words. In *Proceedings of COLING '96*, pages 1159–1162, 1996.

[8] 松本 裕治 and 徳永 健伸. コーパスに基づく自然言語処理の限界と展望. 情報処理, 41(7):793–796, 2000.