2T-06 コーパスを用いた意味カテゴリの代表的語彙の収集法*

秋葉 泰弘† 中岩 浩巳†
NTTコミュニケーション科学基礎研究所†

1 はじめに

機械翻訳など自然言語処理システムでは，様々な処理で多義の絞り込みが必要となるが，多義絞り込みの手がかりの1つとして，意味制約が使われる．例えば，NTTの日英機械翻訳システム ALT-J/Eでは，解析，変換，生成の様々なフェイズ(形態素解析，名詞句解析・変換，文や名詞句の構造変換など)で，意味制約が大きな役割を果たしている[3, 1].

意味制約として，シソーラス[4, 2](日本語語彙大系，分類語彙表，EDR概念辞書等)上の概念が利用される．ALT-J/Eの場合，意味制約として意味属性が使われている[3]．意味属性は約2700種類あり，これら意味属性は最大の深さ12段の木構造を成す[4].

意味制約として利用される概念は，一般的過ぎず，特殊過ぎず，シソーラス上適切な深さの概念を指定する必要がある．例えば，[1]では，意味制約として利用できる意味属性の深さを上位に制限した場合，言語知識の記述能力がどのように落ちるかを報告している．一方，シソーラスの下位の意味属性ばかり指定した場合，その意味制約による多義の絞り込み能力は高くなるが，正解を取りこぼす可能性が高くなる．

これまで多くの意味制約は，ルールや辞書の作成者により手作業で付与されており，正解の取りこぼしを避けるために，適切と判断される概念よりもやや上方の概念を割り振る傾向にある．意味制約を自動的に付与する研究も数々ある[5]が，それらの手法では獲得できない意味知識が数多くあり，その獲得のためには手作業により意味制約を付与せざるをえない．

意味制約の新規付与や，付与済意味制約の妥当性検証や適用分野向け調整を手作業で行なう際，シソーラス上の概念に対する代表語彙が役に立つ．以下本稿では，コーパスから各概念の代表語彙を収集する方法を提案し，その実験的評価を報告する．

2 提案手法

まず，本稿でいう概念の代表語彙についてALT-J/Eの意味属性を例に説明する．表1は，意味属性，〈解説〉[1]を意味属性に持つ単語のリストである．〈解説〉が意味制約として妥当か否かを検証するには，リスト中から処理対象分野で良く使われる語と使われる文脈を想定して，妥当性をチェックする必要がある．この時取り上げる語がここでいう代表語彙である．

バランスの取れたコーパスが存在して，単語に多義がないと仮定すれば，(1) コーパスを形態素解析し，(2) 名詞句の主名詞になり得る名詞を抽出し，(3) 抽出した名詞のコーパス中の頻度を数え，(4) 概念毎に高頻度語を代表語彙として抽出することにより，所望の代表語彙を自動収集できる．実際はバランスの取れた理想的なコーパスは存在しないが，(3) の頻度情報は，複数のコーパスを用いて単語の出現頻度の平均値を求め，その平均値で代用可能である．一方，表2に示すように，一般に単語は複数の概念(語義)に対応するため，文中の単語がどの概念として使われているか判らず，数え方に工夫が必要となる．そこで，どの分野の文書でも良く使われそうな語をALT-J/Eの意味属性各々から人手で選び出し，選ばれた語が何番目の語義のために選ばれたかを調べたところ，表3のようになった．例えば，多義3の欄は，多義が3つある語のうち第1語義で選ばれた場合が，38.1%あったことを示す．各欄の数の並びは，ほぼ等差数列で最大差が10%である．このような分布を利用すれば，よい頻度が算出可能である．

以上の分析に鑑み，本稿では以下の手法を提案する．

- **(S1)** 複数のコーパスを準備し，各コーパスを形態素解析.
- **(S2)** コーパス毎に，名詞句の主名詞になり得る名詞を抽出.
- **(S3)** どのコーパスからも抽出される名詞(共通名詞)だけを選択.
- **(S4)** コーパス毎の各共通名詞の頻度を基に，該共通名詞の期待頻度(頻度の平均値)を算出.
- **(S5)** 概念毎に，該概念に分類される各共通名詞の期待頻度を，該共通名詞の語義における該概念の順位に応じて，補正.
- **(S6)** 概念毎に，共通名詞の補正済み期待頻度の多い順に並べ替え.
- **(S7)** 概念毎に，上位語を代表語彙として抽出.

なお，(S5)における期待頻度の補正は，期待頻度にW_j/W_1 (j='共通名詞が持つ概念における対象概念の順位')を掛けて行なう．ここで，$\{W_i\}_{i=1,\cdots,a}$は，単調減少の有限等差数列で，公差$-r/(a-1)$ (aは対象名詞が持つ概念数で，rは終項W_aを初項W_1に比べてどれだけ小さくするかを規定するパラメータで，$0 \leq r \leq 1$なる実数)で，総和$\sum_{i=1}^{a} W_i = 1$である．但し，$a = 1$の時には$W_1 = 1$とする．

*Collection of Typical Words on each Semantic Attribute by Using Corpora.
Yasuhiro Akiba† Hiromi Nakaiwa†
† NTT Communication Science Laboratories, 2-4 Hikaridai Seika-cho Souraku-gun Kyoto 619-0237, Japan

[1] [4]では意味属性を1528解説(属性番号と属性名の組)と表記するが，本稿では属性番号を略し〈解説〉と表記する.

表 1: 一般名詞意味属性別単語表 [4] の例

見出し	見出しをその意味属性に持つ単語
〈解説〉	絵解 絵解き 解釈 解説 概説 開題 喝破 噛み砕き 噛砕き 勧説 関説 キャプション 砕き 砕き 訓釈 講説 再説 細説 釈義 釈明 シャレード 詳説 叙説 絮説 新説 図図解 図説 説明 総説 俗解 種明かし 種明し 通解 通釈 解き 説き 解き明かし 解き明し 解明かし 解明し 説き明かし 説き明し 説明かし 説明し 説き起こし 説き起し 説起こし 説起し 説き及び 説及び 読解 ナレーション 評釈 評説 敷延 敷衍 ブリーフィング プレゼンテーション 補説 明解 名訳 約説 訳解 読み解き 読解き ライナーノーツ 略説 略解 縷説 例解 例説

表 2: 単語体系 [4] の例：意味属性の並び順は，他の人間用辞書同様，使用頻度の高い語義の順である．

見出し語	見出し語が持つ意味属性
絵解き, 絵解	〈解説〉
解釈	〈理解〉 〈解説〉
新説	〈説〉 〈意見〉 〈理解〉 〈解説〉
説き	〈解説〉 〈説得〉
プレゼンテーション	〈示し〉 〈提案〉 〈解説〉

表 3: 人手作成した代表語彙の多義数毎の語義分布 (%)

	多義無	多義 2	多義 3	多義 4	多義 5
第 1 語義	100	53.7	38.1	31.1	25.3
第 2 語義		46.3	33.7	26.4	22.9
第 3 語義			28.1	22.7	19.4
第 4 語義				19.8	17.2
第 5 語義					15.2

3 実験評価

「CD– 毎日新聞 91,92,94 版」の新聞一年分を 1 つのコーパスと見做し，コーパスの組合せとr値の設定を色々変えて，一般名詞意味属性 [4] に対する代表語彙を提案手法を用いて収集し，収集された代表語彙を意味属性毎に比較した．収集に際し，単語と概念の対応として単語体系 [4] を用い，形態素解析として ALT–J/Eの形態素解析部 [3] を用いた．各意味属性に対する代表語彙としては，補正済み期待頻度の上位 10 個を抽出した．例えば，〈解説〉に対しては，表 4のような代表語彙が抽出された．第一欄はコーパスの組とr値を示し，第二欄は第一欄の設定で得られた代表語彙である．

表 5は，条件 A と条件 B 各々で抽出された代表語彙を，意味属性毎に比べ，どちらの代表語彙がより該意味属性の代表語彙として適切かを 1 名の評価者が判断した際，条件 A の方が優位であると判断された意味属性の数と，条件 B の方が優位であると判断された意味属性の数を示す．91&92(0.1)vs91&94(0.1) の行を見ると，91&94(0.1) の方が優位である意味属性が多く，連続しない 2 年の新聞を組み合わせた方が連続する 2 年分の新聞記事の場合より適切な代表語彙が得られると言える．これは，連続する 2 年の新聞記事をコーパスの組として利用した場合，似た記事が掲載されている可能があり，連続しない 2 年の新聞記事に比べ頻度に偏りが出るためと思われる．91&94(0.01)vs91&94(0) の行を見ると，91&94(0.01) の方が優位である意味属性が多い．また，91&94(0.1)vs91&94(0.01) の行を見ると，僅かではあるが 91&94(0.01) の方が優位である意味属性が多い．従って，r値 =0.1 である時，結果が一番良い．

表 4: 抽出された代表語彙の例

年版 (r値)	〈解説〉に対して抽出された語彙とその修正済期待頻度 () の中は抽出元の共通名詞数.
91&92 (0.1)	説明;2188.2, 解説;778, 解釈;473.3, 釈明; 205.5, 図;100.8, 説明し;97, ナレーション;38.5, 解き;30.15, 説き; 16.5, 解明し;15.5 … (24)
91&94 (0.1)	説明;2229.1, 解説;737.5, 解釈;523.2, 釈明;144, 図;120.4, 説明し;101, 解き; 34.2, ナレーション;30, 説き;21, プレゼンテーション;15.5 … (24)
91&94 (0.01)	説明;2670.5, 解説;737.5, 解釈;626.8, 図;165.7, 釈明;144, 説明し;101, 解き; 37.5, ナレーション;30, 新説;21.1, 説き;21 … (24)
91&94 (0)	説明;2724.5, 解説;737.5, 解釈;639.5, 図;172, 釈明;144, 説明し;101, 解き;38, ナレーション;30, 新説;22, プレゼンテーション;21 … (24)

表 5: 各適用条件で抽出された代表語彙を意味属性毎に比べ，優位な差があった意味属性数の比較．

条件 A	vs	条件 B	A 優位	B 優位
91&94 (0.1)	vs	91&92 (0.1)	244	143
91&94 (0.01)	vs	91&94 (0)	23	6
91&94 (0.1)	vs	91&94 (0.01)	116	31

また，意味属性が成す木構造上の位置により，抽出される代表語彙にどの様な違いが出るかを 91&94(0.1) の条件下で調査したところ，リーフ又はリーフに近い意味属性に対して抽出された代表語彙は，意味属性の属性名と代表語彙には意味的関連性が顕著に見られた．一方，ルート近くに位置する抽象度の高い意味属性に対しては，意味属性の属性名との意味的関連性が顕著ではない単語が代表語彙として抽出されていた．抽象度の高い意味属性に対する代表語彙の収集は，今後の課題である．

4 おわりに

自然言語処理システムで利用する意味制約について，その妥当性の検証や適用分野向け調整を厳密に行なう為の情報源として，コーパスから意味属性の代表語彙を収集する方法を提案した．新聞記事を用いて提案手法を評価したところ，隔年の新聞記事をコーパスとして利用し，提案手法のパラメータを旨く設定すれば，日本語語彙大系の意味属性が成す木構造のリーフ又はリーフ近い意味属性に対して，適切な語が代表語彙として収集することができた．

参考文献

[1] 池原悟，宮崎正弘，横尾昭男: “日英機械翻訳における意味解析用の言語知識とその分解能,” 情報処理学会論文誌, Vol.34, No.8, pp.1692–1704 (1993).

[2] 大井耕三，隅田英一郎，飯田 仁: “意味的類似性と多義解消を用いた文書検索手法,” 自然言語処理, Vol.4, No.3, pp.51–70 (1997).

[3] 八巻俊文他: “特集：日英機械翻訳技術,” NTT R&D, Vol.46, pp.1391–1432 (1997).

[4] NTTCS 研監修，池原悟他編集: 日本語語彙大系, 岩波書店, 東京 (1997).

[5] Ng, H.T. and Zelle, J.: “Corpus based approaches to semantic interpretation in natural language processing,” AI Magazine, Vol.18, No.4, pp.45-64 (1997).