7H－1

Pictorial Dictionary に基づく場面情報の結束性によるそれ自身の語義の自動付与*

角田 達彦，羽柴 正輝，長尾 眞†
京都大学 工学部

1 はじめに

既出版の文書データをコーパスとして自然言語処理に利用する際に，利用形態に応じて形態素・構文・意味解析のための情報を付与する必要があるが，多くの場合は人間が判断して割り当てるものであり，大変労力を必要とする．特に語義などの意味を付与することは，定義文を解釈する必要もあり，時間のかかる作業であるため，自動付与が求められる．

本論文ではコーパスの一つとして，場面知識の情報源に用いる OXFORD-DUDEN Pictorial English Dictionary (OPED)[1] を対象にする．このような場面知識は自然言語処理の意味の曖昧性を解消するために有効な文脈情報の一つである [2]．だがこの辞書には図版とそれに対応する語が列挙されているのみで，語義はあらためて割り当てる必要がある．そこでここでは，場面内の語彙的結束性を利用して WordNet[3] に基づく語義を自動的に推定する方法を提案する．語義推定のために異なる二種類の計算方法を用いて実験を行ない，人間が判断した正解に対する再現率と適合率によって評価した．以下それぞれの手法，実験結果，評価，問題点を考察する．

2 語彙的結束性による場面情報の語義の自己推定

OPED は図 1 に示されるように，日常生活を図版にし，その図版中に出てくる物体に番号をふり，その物体の名前を示す単語が対応づけてある辞書である．単語は対象物体を直接指し示す名詞と，その名詞を修飾する名詞・形容詞・動詞などの自立語が主であり，他に前置詞や冠詞などの付属語も含まれる．しかしそのほとんどは名詞である．そこで修飾／被修飾によらず名詞を語義推定の対象とする．

語義推定の説明のため 'Table' という語を例にとると，その意味は「机」や「表」など複数ある．また 'chair' という語にも「椅子」や「議長」など複数の意味がある．だがこれらの単語が同時に現れた場合には，共通する「家具」の概念が重なり，それに対し「表」の数学的な概念や「議長」の社会的関係を示す概念は重なりにくい．このように単語集合内の語彙的結束性を利用し，正しい意味を推定できる可能性が高い．

語義は WordNet の出力によって定義する．WordNet は語の概念をまず表 1のような 25 個の概念に分類し，その概念から対象単語までのパスと，最上位概念までのパスを出力する．尤度の計算では WordNet のこれらの特徴を利用する．

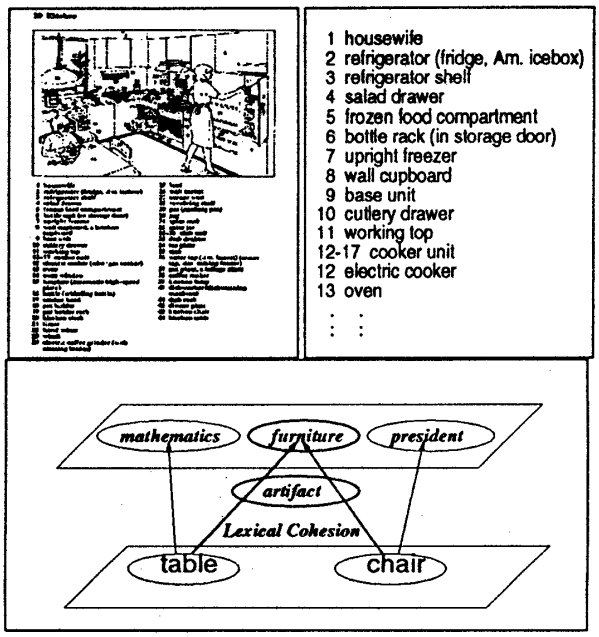

図 1: (左上) OPED での台所の図版, (右上) テキスト部分抜粋, (下) 語彙的結束性を利用した語義推定の概念図

表 1: WordNet での名詞の概念の基本分類（25 種類）

action	cognition	group	person	relation
animal	communica.	location	plant	shape
artifact	event	motive	possession	state
attribute	feeling	natur. obj.	process	substance
body	food	natur. phe.	quantity	time

```
Sense 4 table
  = > furniture, furnishing, piece of furniture, ...
    = > instrumentality
      = > artifact, article, ...
        = > object, inanimate object, ...
          = > entity
```

上の例は 'table' の名詞の 4 番目の意味の出力結果である．下側に向かう程上位の概念を示し，ここでは 'entity'（実体）が最上位概念である．基本 25 分類の一つを太字で示した．

語義の選択は尤度（$L(\theta)$, θ は語義）の比較に基づく．対象とする単語の各語義の尤度を，場面全体の準意味分布から求める．場面全体の準意味分布の求め方は，前述のように，場面に現れる全単語の全語義（名詞のみ，正解／不正解混合）を WordNet によって自動的に出力し，それらの語義の各階層の各ラベルの場面全体での頻度を数える：

*Method of Assigning Senses to Words in Scene Information Based on Pictorial Dictionary by its Lexical Cohesion

†Tatsuhiko Tsunoda, Masateru Hashiba, Makoto Nagao
Faculty of Engineering, Kyoto University
{tsunoda,hashiba,nagao}@kuee.kyoto-u.ac.jp

表 2: 基本統計頻度情報

図版	単語数 対象	単語数 解析可	語義数 平均	推定なしの 適合率
Restaurant	167	160	3.0	58.0%
Living-room	62	60	3.2	61.3%
Kitchen	91	89	2.9	56.7%
Hall	49	47	3.8	51.9%
Dining-room	61	60	4.2	46.0%
Bank	71	62	5.4	46.0%
Bed-Room	53	52	3.0	60.5%
全体	554	530	3.9	55.0%

$$n(S,C) = \sum_{w_h \in C} \sum_i \sum_j \delta(M(w_h,i,j),S)$$

ただし，この式の中の w_h は場面に出てくる各単語，i は各単語中の語義の順番，j は各語義の各ラベルの階層，$M(w_h,i,j)$ は各ラベルの名前，$n(S,C)$ は任意の概念ラベル S の場面 C 全体における頻度である．この準意味分布を用いて各場面での語義の推定を行なう．各単語の各語義の尤度は，これらの頻度に重みをかけ，足し合わせることによって求める．

場面 C で各ラベル M につける重み $W(C,M)$ は次の二つの方法によって与える（後で別々に評価する）．

手法 1. WordNet 25 分類での頻度統計を利用
$\forall C, \forall S \in S_{25},\ W(C,M) = \delta(M,S)$
手法 2. 25 分類（含）の下位概念全体の頻度統計を利用
$\forall C, \forall S\ subcat.of\ S_{25},\ W(C,M) = \delta(M,S)$

上記の頻度にこの重みをかけ足し合わせ，尤度を求める：

$$L(C,w,i) = \sum_j W(C,M(W,i,j))n(M(w,i,j),C)$$
$$L(C,w,i^*) = \max_i L(C,w,i)$$

これに従い各語の語義（i^* によって示される）が出力される．

3 実験結果と考察

実験の対象とする場面は，様々な概念が混合し語義の推定が難しいと思われる 7 場面を OPED から選択した．各場面の準意味分布の作成の際は，'(', ')' などの記号などの無意味記号を除き，すべての語を WordNet に入力し，名詞と仮定した場合の全語義を出力したデータに基づいた．これは人手による部分を極力なくすためである．そして評価対象は，人間が場面中に現れると判断した名詞のみとした．その基本情報および全部の語義を出力したときの適合率を表 2に示す．

表 3および図 2に今回の 2 つの手法による推定の，実験結果を示す．各手法の出力と人間が正解とみなすものは，いずれも複数回答が有り得るので，再現率と適合率で評価した．

両手法とも一部の推定誤りのため再現率は低下したものの，適合率は推定なしの場合に比べて向上している．手法 1 は WordNet の 25 の基本分類という荒いカテゴリの頻度の高さによって尤度を求めるため，手法 2 に比べて回答数が多くなり，再現率が高く適合率が低いという結果である．手法 2 は手法 1 の特定をさらに詳細化するために再現率が低くなり，適合率は若干高い．しかし絞り込みすぎが生じているため推定精度のばらつきが高くなっている．両手法での推定の失敗は個々の概念の結束性からの逸脱によるものが多い．例えば Restaurant の 'mat' は全体の結束性のため「人工物」である「床のマット」,「スポーツ用マット」などが回答されている．しかし場面ではグラスの下に敷くマットを示す「用品」であったため，推定に失敗している．また再現率／適合率の悪い Bank（銀行）では，サービス業務などの行為が多いため，本来受け付け窓口のカウンターを示す 'counter' が「なぐる」カウンターに推定されるなど，物体と抽象的概念の干渉が顕著であった．手法 2 は手法 1 の分類の詳細化であるため，このような影響を特に受けやすいことが実験結果からわかる．

表 3: 実験結果

図版	手法 1 再現率	手法 1 適合率	手法 2 再現率	手法 2 適合率
Restaurant	89.1%	69.6%	80.2%	81.5%
Living-room	92.5%	77.6%	83.3%	86.7%
Kitchen	94.9%	74.8%	66.7%	70.2%
Hall	77.9%	65.3%	59.6%	61.7%
Dining-room	90.0%	68.0%	65.8%	68.3%
Bank	56.7%	48.3%	39.0%	47.5%
Bed-Room	94.2%	77.7%	83.7%	84.0%
全体	86.3%	69.1%	70.4%	73.2%

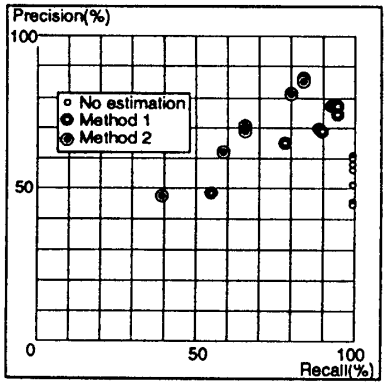

図 2: 実験結果．再現率と適合率による各手法の比較

4 おわりに

本論文では場面知識の情報源に用いる OXFORD-DUDEN Pictorial English Dictionary の語義の自動的自己推定を行なう手法の提案をし，実験・評価を行なった．実験の結果，推定精度の向上が見られた．手法により再現性と適合性のトレードオフがあるため，目的に応じて選択する必要がある．今後，重みの調整による語義推定の精度向上の上，全場面について同様の調査を行なう予定である．

謝辞：データの評価を助けて下さった京都大学 工学研究科 電子通信工学専攻の増田 智君にお礼を申し上げます．また評価対象の辞書データは東京大学 工学部の田中英彦研究室にて入力したものを使わせて頂きました．大変感謝致します．

参考文献

[1] Oxford University Press. The OXFORD-DUDEN Pictorial English Dictionary. 日本出版貿易株式会社, 1981.

[2] 角田達彦, 田中英彦．場面情報に基づく英語名詞の語義の優先づけ方法と評価．情報処理学会第 5 0 回全国大会, Vol. 3, pp. 83–84, 3 1995.

[3] Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. and Tengi, R. Large Five Papers on Wordnet. CSL report 43, Cognitive Science Laboratory, Princeton University, 1993.