自然言語処理 113－12
（1996. 5. 29）

語彙的結束性による図解辞書中の名詞の語義の推定

角田 達彦　　増田 智
京都大学 工学研究科 電子通信工学
〒606-01 京都市左京区吉田本町
e-mail: tsunoda@kuee.kyoto-u.ac.jp

要旨
既出版のコーパスや辞書から知識を抽出するとき，多義性解消は大変労力を必要とするため，自動化が求められる．本稿では，場面を知識化するために有効な図解辞書 OXFORD-DUDEN Pictorial English Dictionary の図版中の名詞の多義性解消を自動的に行なう方法を提案する．各図版中の全名詞の全語義を WordNet シソーラスを調べて列挙し，そのシソーラス中の各カテゴリの頻度を数える．そしてそれを用いて元の各名詞の意味を推測する．WordNet の構造により，2 種類の手法を提案し比較する．提案手法を図解辞書に適用した結果，再現率は約 9 割，適合率は約 7 割という結果が得られた．本稿ではさらに，提案手法の問題点と限界点を考察する．
キーワード　場面，知識，辞書，多義性解消，語彙的結束性，シソーラス

Sense Estimation of Nouns in Pictorial Dictionary by Lexical Cohesion

Tatsuhiko TSUNODA　　Satoshi MASUDA
Department of Electronics and Communication, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto 606-01, Japan
e-mail: tsunoda@kuee.kyoto-u.ac.jp

Abstract
Word sense disambiguation in knowledge source requires too much labor. We propose a method of automatical disambiguation of nouns in OXFORD-DUDEN Pictorial English Dictionary, which is useful for constructing scene knowledge. The method uses lexical cohesion of nouns in each picture. To detect the lexical cohesion, it uses hierarchical structure of WordNet. Combination of word senses in each picture supports the most salient meanings of each noun. According to the structure in WordNet, we propose two types of calculation. We applied our method to seven pictures in the dictionary; the recall ratio was about 90 % and the precision was about 70 %. We also discuss the problems and limits of our method.
keywords　scene, knowledge, dictionary, disambiguation, lexical cohesion, thesaurus

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ため，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．本稿では特に，場面知識 ([2, 3]) の情報源に用いる辞書 OXFORD-DUDEN Pictorial English Dictionary (OPED)[1] を対象にし，語義を付与することを目的にする．

辞書中の単語の多義性を解消する試みに関しては，Bruce らの研究 [4] がある．そこでは，単語の上位概念を，その単語の各語義のそれぞれに対して取り出すことを試みている．まず，Longman Dictionary Of Contemporary English (LDOCE) を調べ，単語の定義文の中に含まれている，その単語の上位概念を表わす語を特定する．上位概念を表わす語は，やはり単語であるため，多義であるが，元の単語の語義につけられている意味コードと分野コードを使って，多義性を解消する．

だが，本稿で扱う図解辞書 OPED のように，図版とそれに対応する単語が列挙され知識のフレームになっているだけであり，他の辞書の語義を割り当てる必要がある場合には，Guthrie らの方法は使うことができない．そこで本稿では，図版内の語彙的結束性を利用して WordNet[5] 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，2章では本稿で扱う図解辞書の特徴を示し，3章で語彙的結束性による各図版中の名詞の語義の推定方法を述べる．4章で実験とその結果を示し，5章で問題点を考察する．

2 図解辞書 OPED の特徴

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．

1 housewife
2 refrigerator (fridge, Am. icebox)
3 refrigerator shelf
4 salad drawer
5 frozen food compartment
6 bottle rack (in storage door)
7 upright freezer
8 wall cupboard
9 base unit
10 cutlery drawer
11 working top
12-17 cooker unit
12 electric cooker
13 oven
⋮ ⋮
44 kitchen table

図 1: 図解辞書 OPED の 'kitchen' の図版のテキスト部分抜粋

それは Man and his Social Environment などの 11 クラスに大きく分けられており，さらに Kitchen などの 384 図版に小さく分かれている．

もともと物の名前を出てくる場面と対象物の形から引くのが目的のものであるので，通常，言語では明示的には扱われない常識の一部を含んでいると考えることができる．角田らの研究 [2, 3] では，この図版が空間的情報としての場面の近似であるとして扱われている．具体的には，図版ごとに現れる単語に対し，語義を割り当てておく．またその語義を集めて意味的なまとまりの現れやすさも計算しておく．そして入力された文章中の場所格などの情報から場面が決定されたと仮定すれば，それに対応する図版を調べ，入力文章中の単語の語義を，直接推定するか，あらかじめ計算してあった意味的なまとまりの現れやすさをみて，語義の優先づけをする．このように，この辞書を一つの知識であるとして，言語処理に適用することが行なわれている．

だが，このような処理を行なうには，図版に現れる単語の語義を特定しておく必要がある．例えば Kitchen の図版には，図 1の 44. kitchen table のように，さまざまな物体や行為が（絵と）言葉で示されている．しかし，この 'table' という単語が「机」をさすか，「表」をさすか，あるいは他の概念をさすかはわからない．そこで本稿は，シソーラスや一般辞書で，どの概念をさしているかを特定することを目的

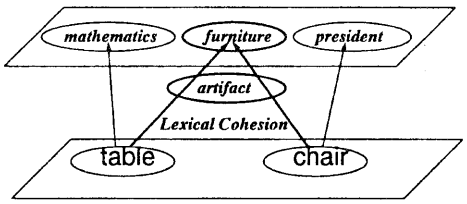

図 2: 語彙的結束性を利用した語義推定の概念図

とする.

多義性解消を行なう対象とする単語に関しては，この辞書の図版でほとんどを占め，知識として有用であることが確かめられている名詞に絞ることにした．本来，図版中の各物体や概念の主体となる語は被修飾語の名詞であるが，修飾語の名詞も知識として使えることが多いため，その図版で表わされる場面に密接に関連するものに限り，推定の対象として含めた．注釈語(図1の 2 の中の Am.= 米語) などは全て省いた．

3 語彙的結束性による図版中の語義の自己推定

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[1]，その概念から対象名詞までのパスと，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．

例えば，図 3は 'table' の名詞の 2 番目と 4 番目の意味の出力結果を示したものである．下側の 4 番目の

[1] ただし，実際に調査した結果，food は substance の下にあり，また communication は relation の下にあることがわかったので，含まれている側の food と communication の分類は使わず，substance と relation でそれぞれまとめて扱った．

表 1: WordNet での名詞の概念の基本 25 分類

action	food	process
animal	group	quantity
artifact	location	relation
attribute	motive	shape
body	natural object	state
cognition	natural phenomena	substance
communication	person	time
event	plant	
feeling	posession	

図 3: WordNet で調べた table の 2 番目と 4 番目の意味の表現

Sense 2
table, tabular array
= > array, arrangement
= > **group, grouping**
:
Sense 4
table
= > furniture, furnishing, ...
= > instrumentality
= > **artifact, article, ...**
= > object, inanimate object, ...
= > entity

意味の方を例にとって説明する．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．また上側は下位の概念を示し，table 自体は 'furniture' (家具) の中に含まれていることがわかる．基本 25 分類の一つである 'artifact' (人工物) の部分は階層の中ほどにあり，この図では太字で示してある．

語義の選択は尤度 ($L(\theta)$, θ は語義) の比較に基づく．対象とする名詞の各語義の尤度を，図版全体の近似的な意味分布から求める．その求め方は，前述のように，図版に現れる全名詞の全語義 (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図版全体での

頻度を数える（式(1)）．

$$n(S,C) = \sum_{w_h \in C} \sum_{i} \sum_{j} \delta(M(w_h,i,j),S) \quad (1)$$

ただし，この式の中の w_h は図版に出てくる各名詞，i は各名詞を調べたときの語義の順番，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 の図版 C 全体における頻度である．この近似的な意味分布を用いて各図版での語義の推定を行なう．各名詞の各語義の尤度は，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．

図版 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

$$\forall C, \forall S \in \mathbf{S}_{25}, \quad W(C,M) = \delta(M,S) \quad (2)$$

手法 2. 25 分類（含）の下位概念全体の頻度統計

$$\forall C, \forall S \; subcat.of \; \mathbf{S}_{25}, \quad W(C,M) = \delta(M,S) \quad (3)$$

上記の頻度にこの重みをかけ足し合わせ，式(4)に従って尤度を求める．

$$L(C,w,i) = \sum_{j} W(C,M(W,i,j))n(M(w,i,j),C) \quad (4)$$

そして式(5)のように最大の尤度を持つ語義を求める．

$$L(C,w,i^*) = \max_{i} L(C,w,i) \quad (5)$$

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．

この様子を具体的な例で説明する．手法 1 では，図版中の名詞を全て列挙し，さらにそれぞれの全ての語義を列挙する．そして基本 25 分類（group や artifact など）の部分を調べ，それぞれの分類の頻度を数えあげておく．そして例えば，図 3であげた 'table' の 2 番目と 4 番目の意味のうち，どちらを解として出力するかを決める場合を考える．この場合は，group と artifact のそれぞれのカテゴリの頻度を比較し，頻度が多い方を選択する．例えば Kitchen という図版では artifact が多いので「机」を示す 4 番目の意味になるし，Mathematics という図版では group が多いので「表」を示す 2 番目の意味になる．

手法 2 では，基本 25 分類を含め，それ以下の概念の分類についても，それぞれ頻度を数えあげておく．

表 2: 基本統計頻度情報

図版	語数 対象	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	86	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	527	3.9	55.0%

次に，'table' の 2 番目の意味のパスのうち，group 以下のパス（'table, tabular array' まで）の頻度を全て足す．同様に，'table' の 4 番目の意味に対しても，artifact 以下のパス（'table' まで）の頻度を全て足し，値の大きい方の語義をとる．

4 語義推定の実験とその結果

実験の対象とする図版は，様々な概念が混合し語義の推定が難しいと思われる 7 図版を OPED から選択した．

各図版の近似的な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．

評価対象は，人間がその図版で表わされる場面に現れると判断した名詞のみとした．OPED の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．そこでここでは，その図版に直接関係すると判断した名詞を語義推定の対象とする．その基本情報を表 2に示す．

多義性解消の場合，人間が正解と判断する語義は，各名詞に対して複数ありうる．また，多義性解消処理からも，複数の解が出力されうる．このため，各名詞に対して次のような再現率と適合率を定義する．

$$\text{再現率} = \frac{\text{処理の出力のうち正解の数}}{\text{人間が正解と判断する語義の数}} \quad (6)$$

$$\text{適合率} = \frac{\text{処理の出力のうち正解の数}}{\text{処理が出力する語義の数}} \quad (7)$$

さらにここでは，複数の名詞をまとめて評価するため，それぞれの平均を求める．以下，再現率と適合率は，それぞれ平均をさすものとする．

表2に，推定なしで全部の語義を出力したときの適合率を示す．推定なしの場合には，適合率は平均して55 % ほどしかない．これに対し，今回提案した2つの手法を適用したところ，手法1では再現率9割以上，適合率7割以上という推定率が，そして手法2では再現率，適合率ともに7割以上という結果が得られた．表3に，今回の2つの手法による推定の詳しい実験結果を示す．

全体の語義の分布であるが，Restaurant, Living-room, Kitchen, Hall, Dining-room, Bed-room では，全語義を列挙した場合も，人間が判断して正解を列挙した場合も，artifact（人工物）が最も多い．また，Bank では，どちらの場合も relation（関係）が最も多い．それぞれの図版で2番目に多いものは，単純に全語義を列挙した場合と，人間が判断して正解を列挙した場合とでは，異なる場合も多くなっている．したがって分布が一致しているとは必ずしもいえないという結果が得られた．

上の実験の結果のうち，推定の誤り例を次章で考察する．

5 考察

両手法とも適合率は推定なしの場合に比べて向上し，7割以上になった．手法2では再現率が大きく低下し，7割になったが，手法1では9割以上を保っている．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，適合率の平均値は若干高い．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ため，有意に手法2の方が適合率が良いとは言えない．

表3: 実験結果

	手法1		手法2	
図版	再現率	適合率	再現率	適合率
Restaurant	90.0%	70.2%	80.2%	81.5%
Living-room	100 %	84.7%	83.3%	86.7%
Kitchen	98.3%	78.6%	66.7%	70.2%
Hall	95.7%	65.3%	59.6%	61.7%
Dining-room	91.5%	69.2%	65.8%	68.3%
Bank	72.5%	58.2%	39.0%	47.5%
Bed-Room	94.2%	77.7%	83.7%	84.0%
全体	91.5%	72.0%	70.4%	73.2%

表4: 概念の結束性からの逸脱の例

語	正解	出力（誤り）
foliage	植物	デコレーション
mat	グラスマット	床用マット
napkin	テーブルナプキン	おむつ
roll	パン	フィルム
water	水	水道管
set	組	機器，ステージ

(1) 概念の結束性からの逸脱

推定の失敗のうち，両手法で共通したものは，個々の概念の結束性からの逸脱によるものが多い．Restaurant, Living-room, Kitchen, Hall, Dining-room, Bed-room では，artifact（人工物）というカテゴリが最も多いため，artifact が語義にあれば，それを必ず選ぶようになっている．このため，表4に示すような誤りが起きた．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 Restaurant の図版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，'water' という語は，Kitchen では水道管で正しいが，Restaurant では水を答えなければならず，誤っている．このように，図版ごとの全体の意味の傾向とずれる語義を持つ語は，この方法では推定が困難である．

表 5: 抽象的関係の推定の誤りの例

語	正解	出力（誤り）
counter	受け付けカウンター	なぐる
glass	ガラスの	鏡，グラス，顕微鏡
service	サービス（業）	救済
accounts	登録	報告
company	会社	友好関係
place	代替，社会的地位	場所
draft	推奨，保証	請求書

(2) WordNet の問題

WordNet の分類が必ずしも正しくない場合がある．基本 25 分類に分けられておらず，推定できなかったものが 3 つあった．他に，Restaurant での 'card' のように，artifact（人工物）に分類されてよいものが，relation に分類されているように，分類が図版に合っていないと判断されるものもある．汎用のシソーラスを作った場合に避けられない問題であり，状況に応じた精緻なシソーラスが望まれる．

(3) 抽象的関係の推定の困難さ

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，推定が難しい．例えば，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．また，relation（社会的関係）を表わすカテゴリと，action（行為）を表わすカテゴリは，動詞が名詞化したものが多く，意味が派生的で，人間でも正解の特定や分類が困難なものが多い．誤りの例の一部を表5に示す．

(4) 手法 2 の再現率の低下の原因

手法 2 の再現率を下げている原因は，a) 複数の正解があったときに，絞り込み過ぎて解を少なくしてしまった場合と，b) 間違った解に絞り込んで正解を選択しなかった場合とに大きく分かれる．a) の，複数の正解があるにも関わらず絞り込み過ぎた場合には，同時に誤った解を適切に除くことができた場合には，適合率が上昇する．しかし，手法 1 のおおまかな 25 分類で既に正解に絞り込んでいたような場合には，適合率は上昇しない．また，b) の，間違った解に絞り込んでしまう場合には，適合率は下降する．表3を見ると，図版によってそれぞれどの場合に当てはまるかが異なっているのがわかる．平均値で見る限りは，手法 1 と手法 2 とでは適合率は有意な差がない．しかし，再現率では手法 1 の方が 2 割程度高いため，手法 1 を用いて正解を多く残し，それを人間が判断する方が良いと思われる．

6 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を提案した．

実験の結果，推定なしの適合率 55% に対し，提案手法では再現率 9 割以上，適合率 7 割以上となり，推定精度の向上が見られた．

今後，重みの調整による語義推定の精度向上の上，全図版について同様の調査を行なう予定である．

謝辞：評価対象の辞書のデータは東京大学 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, 英語名詞の多義性解消における文脈としての場面情報の評価, 自然言語処理, Vol. 3, No. 1, (1996), pp. 3–27.

(3) 角田達彦, 羽柴正輝, 図解辞書と LDOCE の分野コードに基づく場面知識による英語名詞の多義性解消, 電子情報通信学会研究報告 NLC96-85, (1996).

(4) Bruce, R. and Guthrie, L., Genus Disambiguation: A Study in Weighted Preference, *Proceedings of COLING-92*, (1992).

(5) 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).