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 *

角田 達彦，羽柴 正輝，長尾 眞†
京都大学 工学部

1 はじめに

語義の曖昧性の解消は，解決困難な問題の一つである．文脈に依存する場合は，適切な知識を用いて文章を解析する必要があるため，特に難しい．このような知識の一つとして，典型的な場面に関する情報を知識として持つことが考えられる [1]．それを近似するため，図解辞書の図版に現れる名詞の語義や，それが使われる分野を図版ごとに集め，知識化する．図解辞書は，OXFORD-DUDEN Pictorial English Dictionary(OPED) [2] を用いる．語義は，英英辞典 Longman Dictionary Of Contemporary English (LDOCE) [3] によって定義する．分野の割り当てには，LDOCE の電子化版の，各語義につけられた分野コードを用いる．

本稿では，このような場面知識をあらかじめ構築しておき，その知識により，文章中に現れる名詞の語義を推定する手法を提案する．文中の場面が段落ごとに同定できたものと仮定し，その段落の中の全名詞に知識を適用する．場面に関連する名詞の多義性が解消でき，場面に関連しない名詞に対する誤りが少ないことが望ましい．提案手法を物語文に適用し，これらの評価を行なう．

2 場面知識の構築

場面に関する知識は，図解辞書 OPED に基づいて作る．この辞書は，物の名前を，それが現れる場面や，物の形などの種類から人間が引くことを目的に出版されたものである．日常生活に出てくる場面が網羅的に図版に描かれ，典型的な物体には，それを表わす語が割り当てられている．この図版全体を，人間が場面ととらえているものの近似であるとし，登場する物を，場面に現れやすい物であると仮定する．登場する物の名前は，名詞によって表現されているものがほとんどである．そこで，これらの名詞のそれぞれに語義を割り当てたものと，図版に現れる語の語義の意味的なまとまり（分野）をとらえたものとの，2 種類の知識を作る．意味的なまとまりを知識にする理由は，図版にはたまたま現れなかったが場面に関連すると思われる語（未登録語）に対しても，多義性解消の処理が行なえるようにするためである．そのまとまりは，LDOCE の電子化版で各語義につけられている「分野コード」によってとらえる．分野は FU（家具：Furniture）など，100 種類ほどある．ただし，語義によっては，分野コードがつけられていないこともある．

場面知識の構築の手順は次のようになる．

A.	図解辞書の各図版に現れる名詞を列挙する．
B.	上記 A の各名詞に語義を割り付ける． LDOCE 中の語義の中から，図版中で使われていると思われる語義を，人手で選ぶ．
C.	上記 B で得られた各語義についている分野コード（LDOCE 電子化版中）を調べる． 図版ごとに，現れる分野コードをまとめる．

例えば，図解辞書の「台所」の図版には，机が描かれ，それには 'table' という語がつけられている．この場合の 'table' の語義を人間が判断すると，LDOCE 中では，'1. a piece of furniture...' である．そこで，これらを対にして「台所」の場面の「語義フレーム」に蓄える（手順 B）．また，この語義 1 を LDOCE 電子化版で調べると，'1.[FU] a piece of furniture...' のように，分野コード FU が割り当てられているので，この分野コードを「台所」の場面の「分野フレーム」に加える（手順 C）．以上を図版中のすべての名詞に対して行ない，「語義フレーム」と「分野フレーム」のそれぞれをまとめる．その結果，次のような知識が得られた．

```
[台所:  語義:  [table, 1. a piece of furniture...],
               [tea, 3. a hot brown drink...],        (1)
               ...
        分野:  [HH, FO, FU, HR, EG,BO]]
```

ただし，HH: 家にある物, FO: 食べ物, FU: 家具, HR: 時間, EG: 工学,BO: 植物であり，分野は頻度順に並べてある．この頻度を用いて多義性解消の細かい制御を行なうことも考えられるが，今回は評価が繁雑になるため，行なわないことにした．寝室など，他の場面に対する知識も同様に，それぞれ別々に作る．

*Disambiguation of Noun Sense by Scene Knowledge Based on Pictorial Dictionary and LDOCE Subject-codes

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

3 場面情報による多義性解消

文章中の名詞の多義性を解消する過程はまず，文を形態素解析し，名詞を取り出すことから始まる．そしてその名詞の語義を，上で構築した場面知識によって推定する．ここでは，LDOCE 中の語義の中から選ぶことを，語義の決定と定義する．多義性解消は，次のような手順にしたがって行なわれる．

1. 形態素解析により名詞を取り出し，原形化．
2. 場面知識の語義フレーム（[名詞, 語義] の対）の中に，調査中の名詞が見つかれば，それに対応する語義を出力し，次の名詞を調べる．
   ↓ 見つからないとき
3. LDOCE で各語義の分野コードを調べ，
   (a) 場面知識の分野フレーム（[分野 1, ...]）に登録されている分野を持つ語義を，すべて出力する．
   (b) 分野フレームに登録されている分野を持つ語義が一つもなければ，全ての語義を出力する．

形態素解析は，E.Brill の開発した Part-Of-Speech Tagger と，Penn 大で開発された morph を使う．

場面知識の適用であるが，まず，場面知識の語義フレームの中で，処理対象の名詞と同じ名詞を探す．見つかれば，その語義を出力する．例えば，文章中の 'table' という名詞の多義性を解消したいとする．その段落は台所の場面であったとする．2 章の「場面知識の構築」で作った，台所の語義フレーム (1) を見ると，[table, 1. a piece of furniture...] というスロットがあるため，'1. a piece of furniture...' をそのまま出力する．もしも，この語義フレームに 'table' という名詞が登録されておらず，見つからなかった場合は，次に分野フレームを探す．分野フレームに FU という分野（家具：furniture）があり，かつ，'table' の語義の中に FU という分野コードが割り当てられているものがあれば，その語義を出力する．各単語で，場面知識のもつ分野と一致する分野がつけられた語義が複数ありうるが，それらの語義を全て出力する．分野フレームにも照合する分野を持つ語義が一つもなかった場合は，ここではその名詞は処理の対象としないと考え，その名詞の持つ語義を全て出力し，他の処理にゆだねる．

4 物語文を対象とした実験と評価

4.1 評価対象とした文章の中の名詞の性質

今回提案した多義性解消手法を評価するための文章は，モンゴメリー作「赤毛のアン」の英語原作より，台所の場面を含む 7 段落と，寝室の場面を含む 7 段落を人間が判断して切り出したものである．

表 1: 評価の対象とした名詞の，語義数の分布と平均：「場面関連」は，人間が判断して場面と関係があると思われる語について，また「非関連」は，場面と関係がないと思われる語について，それぞれ調べたものである．各名詞の語義数に対して，頻度をとったものが示されている．

評価の対象とした名詞の語義数の分布と平均				各名詞の語義数					語義数の平均
				1	2-5	6-10	11-20	≥21	
単語数	場面関連 65 個	台所	32	5	18	3	6	0	5.5
		寝室	33	8	9	5	10	0	6.9
	非関連 101 個	台所	66	9	33	11	7	7	7.4
		寝室	35	7	11	4	7	6	10.6
	全体		166	30	71	23	30	13	7.6

その文章を形態素解析した結果，台所の文章の中より 109 語，および寝室の文章の中より 84 語の，あわせて 193 語が名詞と判定された．そのうち，人間が判断して名詞であり，かつ LDOCE 中に正解と判断できる語義が存在したものは，台所の場面では 98 語，寝室の場面では 68 語の，あわせて 166 語であった．今回の評価では，この 166 語を用いた．

これら評価対象となる名詞の語義数を調べた結果，表 1のように，語義が一つのものは約 2 割あり，49 個の語義をもつものまで，幅広く分布している．全体の語義数の平均は 7.6 個であった．人間が判断し，それぞれの場面に関連する語と，関連しない語とに分けて評価を行なう．場面に関連する語の典型的な例は，「台所」では 'dish（皿）' であり，「寝室」では 'bed（ベッド）' である．一方，これらの場面に関連しない語の典型的な例は，'thing（物事）' や 'face（顔）' などである．今回対象とした名詞 166 個のうち，場面に関連する語は 65 個（39.2 %）で，場面に関連しない語は 101 個（60.8 %）であった．

表 1からわかるように，場面に関連しない語の方が，場面に関連する語よりも，語義数が多いという傾向がある．これは，場面に関連しない語の方が抽象的な名詞が多く，使われ方が多様であるために，語義数が多くなっているためである．しかし，場面に関連する語にも，語義数の多いものが多くある．このため，場面に関連する語の多義性を解消することは意味がある．

4.2 評価方法とその目的

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．

$$\text{再現率} = \frac{\text{処理の出力のうち正解の数}}{\text{人間が正解と判断する語義の数}} \quad (2)$$

$$\text{適合率} = \frac{\text{処理の出力のうち正解の数}}{\text{処理が出力する語義の数}} \quad (3)$$

表 2: 評価した名詞全体に対する多義性解消の結果

	全体 166 個	LDOCE の先頭語義	今回提案する手法	重要性
再現率	台所 98 個	66.3 %	89.8 %	
	寝室 68 個	84.6 %	93.4 %	◎
	平均	73.8 %	91.3 %	
適合率	台所 98 個	68.3 %	51.3 %	
	寝室 68 個	86.8 %	58.5 %	△
	平均	75.9 %	54.3 %	

表 3: 場面に関連する名詞に対する多義性解消の結果

	場面関連 65 個	LDOCE の先頭語義	今回提案する手法	重要性
再現率	台所 32 個	79.7 %	90.6 %	
	寝室 33 個	97.0 %	97.0 %	◎
	平均	88.5 %	93.8 %	
適合率	台所 32 個	81.3 %	79.5 %	
	寝室 33 個	97.0 %	80.6 %	◎
	平均	89.2 %	80.0 %	

さらにここでは，複数の名詞をまとめて評価するため，(2)，(3) の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．

本研究の目的は，場面に関連する名詞の多義性解消である．したがって，場面に関連する名詞に対する処理の結果は，再現率と適合率がともに高い方が良い．一方，場面に関連しない名詞に対しては，ここでの処理によって生じる誤りがなるべく小さい方が良いため，再現率が高い方が良い．だが適合率は，低くとも良い．場面に関連しない名詞に対しては，無理に語義を絞り込むことをせず，全語義を残しておき，他の処理で多義性を解消することを想定しているためである．

4.3 多義性解消の結果と考察

本稿で提案する手法と比較するため，各名詞の語義として LDOCE 中で先頭に登録されている語義を出力する方法も評価することにした．その理由は，LDOCE は，語義が日常使われやすい順に登録されているからである．したがって，他に知識がない場合には，先頭の語義を解とみなす方法は妥当である．以降，この手法と，今回提案する手法とを対比させながら評価をすることにする．

(A) 対象名詞全体に対する多義性解消の結果

まず，対象とした名詞全部に対して多義性解消を行なった結果を示す．評価した名詞は 166 個である．そのうち，台所の場面を含む段落に現れた名詞は 98 個，寝室の場面を含む段落に現れた名詞は 68 個であった．これらの名詞を評価した結果，表 2に示されるように，提案手法では，全体の再現率は 91.3 % であった．一方，先頭の語義をとる手法では，全体の再現率は 73.8 % であった．この結果から，提案手法では，少なくとも誤りを出力する割合が小さいことがわかる．適合率に関しては，提案手法では 54.3 %，先頭の語義をとる手法では 75.9 % という結果が得られている．

提案手法の適合率が低くなっているのは，以降の評価でわかるように，場面に関連しない名詞に対して，無理な絞り込みをせず，全ての語義を残しておくという，安全な結果になったからである．したがって，この数値が低いことは，提案手法が本質的に悪いことを示すものではない．また，先頭の語義をとる手法の再現率と適合率をともに悪くする原因となっているものは，場面に関連しない名詞に関しても語義を一つに絞り込み，その結果誤ったためである．以降，これらの名詞を人間が判断し，場面に関連する名詞と関連しない名詞とに分け，詳細な評価を行なう．

(B) 場面に関連する名詞に対する多義性解消の結果

上記の (A) の 166 個のうち，場面に関連すると人間が判断した名詞をとりだし，評価した結果を示す．場面に関連する名詞は 65 個である．そのうち，台所の場面を含む段落に現れた名詞は 32 個，寝室の場面を含む段落に現れた名詞は 33 個であった．これらの名詞を評価した結果，表 3に示されるように，提案手法の再現率の平均は 93.8 % であり，適合率の平均は 80.0 % であった．一方，先頭の語義をとる手法の再現率の平均は 88.5 % であり，適合率の平均は 89.2 % であった．これらは本研究で扱う場面の情報が有効であるかを直接評価するものであるため，大変重要な指標となる．

まず，寝室の場面では，先頭の語義をとる手法が大変高い再現率と適合率を示している．これは，寝室にある物は，あまり多義なものがなく，先頭の語義が示す典型的な意味で用いられていることが多いからである．それに対し，台所の場面では，「食べる」という行為が行なわれるために，文章中に出てくる語には，意味が本来の意味から派生したものがあった．今回誤ったものは，'dish'（先頭語義では「皿」であるが，文章中では「料理」の意味で使われた），'meal'（先頭語義：「食べ物」→ 文章中：「食事」），'chamber'（「閉空間」→「部屋」），'light'（「ランプ」→「ロウソク」）であった．また，'tea' は，日常使われる頻度が高いのは「茶葉」であるが，台所で使われるのは，「紅茶」であるために誤っている．

これに対し，提案手法では，先頭の語義をとる手法よりも再現率が高い．これは，上記の 'meal' や 'tea'

などの，場面に特有な語を，正しく推定できたためである．だが，適合率は，提案手法は先頭の語義をとる手法よりも値が低い．その原因となった語は，'bread', 'butter', 'wall', 'floor' などであった．これらの語に共通することは，LDOCE 中では，各語の語義のいずれも，分野コードがふられていないことである．すなわち，これらの語は，どの意味で使われる場合にも，特定の分野に特有なものではないことを示している．提案手法では，これらの語は，場面知識の語義フレームにも登録されていなかったため，語義を特定することはできなかった．この場合には，他の語義を支持することもないため，絞り込みはせず，全ての語義を出力するという安全な方向に向かう結果になっている．このため，少なくとも誤らないために，再現率は高いが，多義性解消できる語がその分少なくなるため，適合率が下がる．

また，提案手法で正解できたものは，65 個のうち 61 個であったが，その内訳は，語義フレームによる正解 21 個，分野フレームによるもの 31 個，そして全語義出力によるもの 9 個であった．この結果から，最初から場面に現れる語を登録しておく語義フレームだけでは不十分で，場面に含まれる分野を知識にした，分野フレームも必要であることがわかる．

(C) 場面に関連しない名詞に対する多義性解消の結果

上記の (A) の 166 個のうち，場面に関連しないと人間が判断した名詞をとりだし，評価した結果を示す．場面に関連しない名詞は 101 個である．そのうち，台所の場面を含む段落に現れた名詞は 66 個，寝室の場面を含む段落に現れた名詞は 35 個であった．これらの名詞を評価した結果，表 4に示されるように，提案手法では，再現率は 89.6 % であった．一方，先頭の語義をとる手法では，全体の再現率は 61.4 % であった．この結果から，提案手法では，少なくとも誤りを出力する割合が小さいことがわかる．適合率に関しては，提案手法では 37.7 %，先頭の語義をとる手法では 64.4 % という結果が得られているが，この値は全く重要でない．提案手法の適合率が低くなっているのは，無理な絞り込みをせず，全ての語義を残しておくという，安全な結果になったからである．したがって，この数値が低いことは，提案手法が本質的に悪いことを示すものではない．場面に関連しない名詞は，扱う対象とせず，無理な絞り込みを行なわずに，そのまま素通りさせたいという，本研究の目的に適っている．

提案手法で誤ったものは，'parlour'（場面知識:「飲物」→ 文章中:「部屋」），'collar'（「工業製品のリング」→「シャツのカラー」），'glass'（「グラスコップ」→「ガラスの」），などである．人間でも判断に迷うものが多い．それに対し，先頭の語義をとる手法では，これらの名詞に加えて，'appearnce', 'east', 'flood', 'one', 'world', 'note', 'indication', 'presence' など，もと

表 4: 場面に関連しない名詞に対する多義性解消の結果

場面非関連 101 個		LDOCE の先頭語義	今回提案する手法	重要性
再現率	台所 66 個	55.3 %	89.4 %	◎
	寝室 35 個	72.9 %	90.0 %	
	平均	61.4 %	89.6 %	
適合率	台所 66 個	57.6 %	37.7 %	×
	寝室 35 個	77.1 %	37.6 %	
	平均	64.4 %	37.7 %	

もと多義・抽象的・派生的であるものの語義を推定しようとしたため，誤ったものがほとんどである．本研究で提案する手法では，これらの語を推定することを自動的に避けることができたため，再現率が高いという望ましい結果が得られた．

5 おわりに

図解辞書と LDOCE 電子化版の分野コードを組合せることによって場面知識を構築し，場面に関連する英語名詞の多義性解消を行なう方法を提案した．場面知識は，名詞の語義を列挙したものと，それに基づいて場面に特有な分野を集計したものから成る．物語文の中の名詞に適用した結果，場面に関連する名詞に対し再現率 9 割および適合率 8 割という結果が，そして場面に関連しない名詞に対し再現率 9 割という結果が得られた．以上から，場面に関連する名詞の多義性解消を行ない，場面に関連しない名詞の語義の推定の誤りを少なくするという目的に，本研究で提案した手法が適うものであることが確かめられた．文中の場面の特定と，場面に関連しない名詞の多義性解消が今後の課題である．

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加藤 哲哉君にお礼を申し上げます．また図解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] 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, pp. 3–27, 1 1996.

[2] Oxford University Press. The oxford-duden pictorial english dictionary. 日本出版貿易株式会社, 1981.

[3] Longman dictionary of contemporary english. Longman Group Ltd., 1978.