Webから辞書への語義別画像付与の試み
- 基本語意味データベース Lexeed および Wikipedia を対象に -

藤田 早苗
sanae@cslab.kecl.ntt.co.jp,
永田 昌明
nagata.masaaki@lab.ntt.co.jp
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

1 はじめに

Web上には膨大な画像が流通しており、これらを既存の辞書と関連付けること ができれば、文字列情報からだけでは得られない、視覚的な情報を利用できる ようになる。例えば、語を調べるとき、適切な画像も同時に表示されれば、子供や、母国語 話者以外でも、語義が理解しやすいと考えられる。 特に、多義性を持つ語の場合、語義毎に適切な画像を提示できれば、 より直観的に異なりを理解しやすいと考えられる。 そのため、本稿では、辞書の語義にできる限りまんべんなく画像を付与することを 考える。

辞書(シソーラス)に画像を付与する研究には、 Bond et al. (2009)や、 藤井ら(2005)、**ImageNet**(Deng et al., 2009) などがある。 Bond et al. (2009)は、日本 語 **WordNet**[1]に、Open Clip Art Library (**OCAL**)から獲得した画像を付与している。ここで、**OCAL**と**WordNet**の階層構造を比較して 画像候補を得た後、各synset(意味クラス)の画像として適切かどうかを人手判 断している。 **OCAL**は著作権フリーで再配布可能という利点があるが、画像の種類が限られるため、 辞書の語義を広くカバーするのは難しい。

藤井ら(2005)は、Webから獲得した画像を、辞典検索システム**CYCLONE**[2]の語義と対応 付けている。 画像が対応すべき語義は、画像のリンク元テキストの多義性解 消により推定している。 しかし、複数語義の出現頻度に大きな差がある場合、 見出し語のみで検索しても、ほとんど、最もメジャーな語義の画像しか得られ ない[3] ため、マイナーな語義にも適切な画像を獲得することは難しい。

一方、近年は、大勢の利用者によってデータ共有やタグ付与を行う仕組み (フォルクソノミー)が発達し、 タグを付与された大量の画像や映像が蓄積されてきている (井手ら, 2009)。例えば、**ImageNet**(Deng et al., 2009)は、 **WordNet**(Fellbaum, 1998)の体系から選択した、一部の synset [4] に大量の画像を付与している。 画像はWebから検索で収集し、 対象synsetの画像として適切かどうかをAMT [5] を利用して人手判断している。 この手法は、大量のデータを精度良く集めることができるため、 非常に有望である。 しかし、現在は対象synsetが限定されているため、 語義の網羅性には疑問が残る。 Deng et al. (2009)との比較は、3章でも述べる。

本稿では、幅広い語義に画像を付与するため、Webからの画像獲得を行う 。 対象辞書は、基本語意味データベース**Lexeed** (笠原ら, 2004)および、Web上のユーザー参加型辞書 **Wikipedia**[6] である (§ 2)。 また、語義毎に適切な画像を付与するため、 語義文から抽出した情報を用いて、 あらかじめ、語義毎に検索語の拡張を行い、 語義を示す適切な画像をWebから獲得する実 験を行う(§ 3)。

2 言語資源

2.1 基本語意味データベース:Lexeed

基本語意味データベース**Lexeed** (笠原ら, 2004)は、日本人の95%以上が知っ ている語を心理実験により選定、収録している。各エントリは、 見出し語、品詞、各語に対するなじみ深さの度合を表す親密度の情報を持ち、 語義毎に定義文と例文を持つ。 また、「檜」プロジェクトにより、定義文から、上位語や同義語も抽出してお り(「檜」オントロジー(Bond et al., 2004))、日本語のシソーラスである語彙大系(池原ら, 2997)とのリンク等も存在する。 図 1に、簡略化したLexeedのエントリの例を示す。

2.2 Web辞書:Wikipedia

本稿では、 **Wikipedia** の曖昧さ回避のページを利用した[7]。 曖昧さ回避のページは、語の複数の意味をリストアップ したページであり、多くは、各語義の詳細なページへのリンクが 存在する。 曖昧さ回避のページの簡単化した例を、図 2に示す。 図 2のように、 記述方法は多様であり、リンクも存在する場合としない場合がある。

```
1  '''EU'''
2  * [[欧州連合]]
3  * [[Europa Universalis]]シ リ ー ズ - [[パ ラドックスインタラクティブ]]の[[歴史シミュレーションゲーム]]
4  * [[愛 媛 大 学]](Ehime University) - [[愛 媛県]][[松山市]]にある日本の[[国立大学]]
5  '''Eu'''
6  * [[ユウロピウム]]の元素記号
7  * [[ユーフォニアム]] - 金管楽器
```

但し、行頭の番号は参照しやすくするため、便宜的に付与した
また、[[ ]]は、**Wikipedia** のサイト内リンクを示す

図 2: **Wikipedia** の例:「EU（曖昧さ回避）」の一部

2.3 Lexeed と Wikipedia の比較

表 1は、**Lexeed** と **Wikipedia** の曖昧さ回避の ページ、両方に共通する見出し語数や語義数を調べたものである。

[1] http://nlpwww.nict.go.jp/wn-ja/
[2] http://cyclone.cl.cs.titech.ac.jp/
[3] 例えば、見出し語「アーチ」で検索した場合、少なくとも上位500画像中に‘‘本累打’’を示す画像はなかった。
[4] Deng et al. (2009)の段階では、 *mammal, bird, fish* など、12 の部分木に含まれる、5,247synsetが対象
[5] Amazon Mechanical Turk, http://www.mturk.com/
[6] http://ja.wikipedia.org/
[7] 利用したバージョンは20091011

見出し語　アーチ　(品詞: 名詞)
親密度　5.531 [1–7] (≥ 5)
語源　*arch*

語義 1
- 定義文　上部$_1$ を 弓$_1$ の 形$_1$ に して 支え$_2$ やすく した 建物$_1$ 。 また$_9$ 、 その$_3$ 建築$_1$ の 様式$_2$ 。
- 例文　あの$_2$ 橋$_1$ は 2 つ の アーチ$_1$ で 出来$_4$ て いる 。
- 上位語　様式$_2$, 建物$_1$
- 意味属性　⟨865:**家屋（本体）**⟩ (⊂ ⟨2:**具体**⟩), ⟨2435:**類型**⟩ (⊂ ⟨1000:**抽象**⟩)

語義 2
- 定義文　骨組み$_2$ の 上$_4$ を 、 装飾$_1$ として 、 杉$_1$ など の 青葉$_2$ で 覆っ$_2$ た 門$_2$ 。
- 例文　運動会$_1$ で は 、 入場$_1$ 用 と 退場$_1$ 用 に 2 種類$_1$ の アーチ$_2$ を 作っ$_2$ た 。
- 上位語　門$_2$
- 意味属性　⟨891:**門**⟩ (⊂ ⟨2:**具体**⟩)

語義 3
- 定義文　野球$_1$ で 、 本塁打$_1$ 。 ホームラン$_1$ 。
- 例文　4 番 バッター$_1$ が ライト$_4$ スタンド$_2$ に 逆転$_3$ の アーチ$_3$ を 放っ$_4$ た 。
- 上位語　本塁打$_1$,　同義語　ホームラン$_1$,　分野　野球$_1$
- 意味属性　⟨1680:**スポーツ**⟩ (⊂ ⟨1000:**抽象**⟩)

図 1: **Lexeed** & 檜 の例:エントリ「アーチ」の一部

表 1: **Lexeed** と**Wikipedia** (曖昧さ回避)

	Lexeed	**Wikipedia** (曖昧さ回避)	見出し語 共通
見出し語数	29,272	33,299	2,228
語義数	48,009	197,912 [a]	19,703
平均語義数	1.6	5.9	8.8
最大語義数	57	320	148
最大語義数の見出し語	取る	本町	フジ
単義語数	19,080	74[b]	2
定義文の平均語数[c]	14.4	10.7	11.0

[a] リストが215,883 あり、そのうち語義らしきものをヒューリスティックに抽出。例えば、図2の場合、行1,5,8 を除いたもの
[b] 語義が一つしかリストされていないものを単義とした
[c] 茶筌によって形態素解析した

共通する見出し語は非常に少なく、1割に満たない [8]。 曖昧さ回避のページに限らず、全見出し語と比較した場合でも、 **Lexeed** の 29,272 エントリ中、16,685 エントリ(57 %) には、**Wikipedia** に共通の見出し語がない [9] 。 また、表 1から、**Wikipedia** は語義数が非常に多い事がわかるが、 大半の語義は固有名詞である。 例えば、「ヒマワリ」の場合、**Lexeed** では単義語だが、**Wikipedia** では67の語義がリストさ れている。このうち、‘‘植物’’と‘‘ひまわり油’’以外は、全て、 楽曲名や組織名などの固有名詞である。 逆に、「アーチ」は、**Wikipedia** では、 ‘‘建築用語’’の語義(**Lexeed** のアーチ$_1$に対応)しかなく、 曖昧さ回避のページもない。

このように、**Lexeed** と**Wikipedia** は非常に収録語、収録語義の傾向が異なる。本稿で は、こうした傾向の異なる辞書の語義に対し、それぞれ適切な画像を付与でき るかどうか、また、どのようにすればより良い画像を得られるかを調べるため、 実験を行う。

[8] 共通の見出し語は、ボロリ、そば、サイクル、フリーキック、フクロウ、ビルマ、など。**Wikipedia** のみの見出し語は、イソップ、ビオ、竜門の滝、AT、オンエアバトル、ラドン、など
[9] 後払い、超過、ユーモラス、とんぼ返り、総決算、抜擢、など

3 画像付与実験

本稿では、辞書の各語義毎に、適切な異なる画像を付与する方法を提案する。見出し語のみで画像検索した場合、ほとんど、最もメジャーな語義の画像しか得られないため、あらかじめ、検索語を語義毎に拡張しておく。本稿では、定義文から得られる情報を用いて、大きく分けて2通りの拡張を行った。主に同義語(SYN)と、主に上位語を含む関連語(LNK)による拡張である。

検索語の拡張方法には様々な手法が提案されているが(Voorhees, 1994),(海野ら, 2008)、同義語による拡張は、テキスト検索の評価で一定の効果を得ている。また、**ImageNet**では、上位語[10] による拡張を行っている。

3.1 実験・評価方法

画像検索[11] で得た画像のうち、リンク切れやアクセス制限のある画像を除いて、語義毎に先頭から5つの画像を獲得し、対象語義を示すのに適切かどうかを評価する。

評価は人手で行い、評価対象の画像が対象語義を示すのに適切(T)か、許容範囲(M)か、不適切(F)か評価した。関連する画像でも、対象語義を表すのに不適切だと思われれば、F (不適切)と評価している。例えば、たまねぎの画像として、たまねぎを使った料理の画像はFと評価している。また、不適切な場合には、その理由も付与した。

3.2 実験：Lexeed

本稿では、**Lexeed** の定義文から上位語や同義語を抽出した檜オントロジー(Bond et al., 2004) の情報を用いて、検索語の拡張を行う[12]。表 4に、檜オントロジーの内訳を示す。SYN は、表 4の同義語、略称、別称など、および、辞書記載の表記ゆれを用いて拡張する。LNK は、上位語や分野など、それ以外の関係を用いて拡張する。但し、LNK, SYN のどちらの方法でも、検索語が拡張できなかった語義は、4,205 語義存在する。更に、対象語が単義語の場合

[10] **WordNet** の定義文に当たるgloss の中で、上位synset名と一致する語
[11] Google AJAX images API, http://code.google.com/intl/ja/apis/ajaxsearch/ を利用
[12] 画像検索は、2009年9月に実施した。

表 2: 語義を表す画像の割合(適合率): **Lexeed**

対象語分類		拡張方法	F (不適切) No.	F (不適切) %	T (適切) No.	T (適切) %	M (許容範囲) No.	M (許容範囲) %	MT (M + T) No.	MT (M + T) %	Total
具体物	単義 (MC)	SYN	18	24.0	36	48.0	21	28.0	57	76.0	75
		LNK	82	33.5	112	45.7	51	20.8	163	66.5	245
		MONO	42	16.8	181	72.4	27	10.8	208	83.2	250
		BL	46	18.4	171	68.4	33	13.2	204	81.6	250
	多義 (TC)	SYN	94	38.7	88	36.2	61	25.1	149	61.3	243
		LNK	111	44.4	92	36.8	47	18.8	139	55.6	250
		BL	180	72.0	53	21.2	17	6.8	70	28.0	250
区別なし	単義 (MA)	SYN	32	42.7	21	28.0	22	29.3	43	57.3	75
		LNK	138	57.5	54	22.5	48	20.0	102	42.5	240
		MONO	98	40.0	98	40.0	49	20.0	147	60.0	245
		BL	112	44.8	86	34.4	52	20.8	138	55.2	250
	多義 (TA)	SYN	122	49.0	64	25.7	63	25.3	127	51.0	249
		LNK	150	60.2	52	20.9	47	18.9	99	39.8	249
		BL	201	80.7	36	14.5	12	4.8	48	19.3	249

表 3: 語義を表す画像の割合(適合率): **Wikipedia**

対象語分類	拡張方法	F (不適切) No.	F (不適切) %	T (適切) No.	T (適切) %	M (許容範囲) No.	M (許容範囲) %	MT (M + T) No.	MT (M + T) %	Total
Lexeedと共通	SYN	98	40.8	119	49.6	23	9.6	142	59.2	240
	LNK	92	41.8	107	48.6	21	9.5	128	58.2	220
Lexeedと非共通	SYN	100	41.2	103	42.4	40	16.5	143	58.8	243
	LNK	96	41.0	93	39.7	45	19.2	138	59.0	234

表 4: 檜オントロジーの内訳

タイプ	No.	%	例
上位語	47,054	69.1	アーチ$_1$, 様式
同義語	14,068	20.6	アーチ$_3$, ホームラン
分野	1,868	2.7	アーチ$_3$, 野球
下位語	757	1.1	売り買い, 売る
部分全体	686	1.0	赤身, 魚肉
略称	383	0.6	亜, アジア
別称など	216	0.3	差し込み, コンセント
その他	3102	4.6	包み焼き, 魚
Total	68,134	100	

は、標準表記による検索も行う(MONO)。但し、標準表記がひらがなの場合のみ、表記ゆれを用いて拡張する[13]。

具体物か抽象物か、単義語か多義語か、によって傾向が変わると考えられることから、**Lexeed** の見出し語を、具体物、かつ、単義語(MC)、あるいは、多義語(TC)、具体物かの区別なし、かつ、単義語(MA)、あるいは、多義語(TA) の4通りに分け、その中から、ランダムに50語義ずつ対象語義を選択した。ここで、**Lexeed** の語義にリンクした意味属性が〈2:**具体**〉配下のみなら具体物だと仮定している[14]。

3.3 評価結果と議論 : Lexeed

表 2は、獲得した画像のうち、対象語義を表す画像の割合(適合率)を示している。ベースライン(BL)は、見出し語(標準表記)のみを検索語として画像検索した場合に、対象語義に分類された数である。表 2によると、LNK よりSYN の方がいずれも適合率が高い。つまり、より、語義を絞る効果があったと考えられる。また、多義語(TC, TA)の場合、拡張の効果が高く、語義毎の検索語拡張によって、各語義に適切な画像を獲得しやすいといえる。但し、具体物の単義語(MC)の場合、LNK,SYN共に適合率が下がっている。これは、MCでは、対象語義自体がメジャーな語義になるためだと考えられる。但し、MONO だけは良くなっており、ひらがなの標準表記以外への拡張は語義を絞る効果があったといえる。

TC のLNKを対象に、Fになった原因を分析する(表 5)。表 5で、「画像で語義表示は困難」は、24.3%を占める[15]。表 5の原因2-5(33.3 %)は、検索条件の改善によって対処できると考えられる。特に、「上位語等により対象がぼやけた」が10%以上あった。上位語は、効果的な場合も多いが、その上位語に含まれる他の語が出てくる事も多い。例えば、「煮干し」を、上位語「食品」で拡張すると、「煮干しを使った食品」が多く出現した。こうした原因で、LNK よりSYN の方が結果が良かったと考えられる。また、拡張に使った語の語義がマイナーだった場合や、検索対象の語義が圧倒的にメジャーな語義の場合、拡張によってむしろ適合率を下げる結果となった。そのため、メジャーな語義かどうか判断する、マイナーな語義の拡張語は使わないといった工夫が必要である。

3.4 実験: Wikipedia

図 2のように、**Wikipedia** では、サイト内リンクは[[ ]]で囲まれている。LNK では、この、[[ ]]で囲まれた部分を拡張に利用する。全語義中、95.5 % には[[ ]]で囲まれた部分がある[16]。また、[[ ]]の出現数は、1語義平均0.95個だが、最大48 個出現した。但し、[[2010年]]や[[1990年代]]のような、時間表現へのリンクは用いない。

[13] 例えば、標準表記「とんぼ」の表記ゆれは「蜻蛉」など

[14] 例えば、図1のアーチ$_2$ (〈891:**門**〉 ⊂ 〈2:**具体**〉)など

[15] 「画像で語義表示は困難」と判定された語義は、各50語義のうち、MC が3、TC が9、MA が10、TA が16だった。

[16] 実際にリンク先ページが存在するのは、85.4%

表 5: F(不適切)の原因: TC, LNK の場合

番号	原因	No.	%
1	画像で語義表示は困難	27	24.3
2	上位語等により対象がぼやけた	12	10.8
3	検索語がマイナー語義だった	11	9.9
4	展開しないほうがよいだろう	8	7.2
5	元の語義があまりにマイナー	6	5.4
6	その他	47	42.3
	Total	111	100

表 6: F(不適切)の原因: **Lexeed** と共通見出し語, LNK の場合

番号	原因	No.	%
7	検索語が少ない（定義文から利用可能)	14	15.2
8	検索語が不適切（定義文から利用可能）	10	10.9
2	上位語等により対象がぼやけた	5	5.4
9	T(適切)かもしれない	5	5.4
6	その他	58	63
	Total	92	100

また、同義語らしきものをヒューリスティックに獲得し、SYN に利用した。例えば、- や– で区切られた先頭部分[17]や、矢印(→)などで明示された参照先名[18]、見出し語が英数字の場合に各文字を含む部分[19] などである。これにより、197,912語義のうち、98.0% に同義語らしき語を獲得できた。実験対象の語義は、**Lexeed** と共通する／しない見出し語から、それぞれランダムに50語義選択した[20]。

3.5 評価結果と議論: Wikipedia

表 3 は、獲得した画像のうち、対象語義を表す画像の割合(適合率)を示している。BL はないが、評価時に、拡張しなかった場合の検索結果も表示し、比較しており、拡張しない方が良いものは全くなかった[21]。これは、語義数が多い分、圧倒的にメジャーな語義である確率が低かったためだと考えられる。

表 3 によると、SYNとLNK に大差はなかった。これは、SYN で利用した検索語の多くが、内部リンクにもなっていたため、差が出にくかったのだと考えられる。しかし、T (適切)のみに着目すると、若干SYNの方が適合率が高く、同義語を抽出する過程で、語義と関連の薄いリンクを排除する一定の効果はあったものと考えられる。

Lexeed でも、上位語を用いた拡張(LNK)より、同義語を用いた拡張(SYN)の方が適合率が高かった。本評価方法では再現率は計算できないが、上位語を用いた場合、再現率は上がっているかもしれない。**ImageNet**は上位語を用いて拡張しているが、上位語を用いた拡張は、より多くの画像を獲得するためには、適切かもしれない。しかし、本稿のように、辞書に画像を付与する場合、語義を適切に示す画像が数個程度獲得できれば良いため、再現率より、語義を絞ることによって適合率が上がる方が好ましい。

表 6は、**Lexeed** と共通の見出し語のLNK を対象に、F(不適切)となった原因を分析したものである。**Lexeed** とは対称的に、「画像で語義表示は困難」な語義はなかった。しかし、「語義を示す画像」とは何か、という判断が難しい語義も多かった。例えば、表 6で、「T(適切)かもしれない」と判断されているのは、人名に対し、代表作の画像が出てきたものである。本稿では、人名には、その人自身の画像のみT(適切)としたが、作家やアーティストなどに、代表作の画像が付与されるのは、むしろ良いとも考えられる。

また、本稿では、画像は5つ獲得したが、画像はひとつで十分な語義もあった。例えば、アルバムの名の場合に、アルバムのジャケットの画像などである。反対に、複数のタイプの異なる画像を付与した方がいいと場合もあった。例えば、市の場合に、地図、風景、市役所、市章の画像などである。そのため、アルバム名やアーティスト名、市町村名など、よく出現するクラスに語義を大別し、必要な画像タイプを決めることも考えられる。

[17]図 2の3, 4, 7, 9行目のような形式。

[18]例えば、「イヌ」‘‘[[十二支]]の一つ。→[[戌]]。’’など

[19]例えば、「CS」‘‘computer science’’など

[20]画像検索は、2009年12月に実施

[21]対象語義は、結果的にほとんどが具体物であり、全て多義語であることから、**Lexeed** でのTC と同じような条件だと考えられる。

4 まとめ

本稿では、辞書の語義毎に、語義を示す適切な画像をWebから獲得する実験を行った。対象辞書は、**Lexeed** と**Wikipedia** という、非常に傾向の異なる辞書を用いた。語義毎に画像を得るため、定義文を用いて見出し語を拡張し、画像検索を行った。拡張には、同義語(SYN)と、上位語を多く含む関連語(LNK)を用いた。但し、**Wikipedia**では、LNKとして、内部リンク名を用いている。画像検索の結果得られた画像に対し、該当語義を示すのに適切かどうかを人手評価した。評価の結果、いずれも同義語を用いた方が適合率が高く、特に、**Lexeed**では、大きな差がついた。また、多義語に対して拡張の効果が高く、辞書の各語義に適した画像を獲得するという目的に効果をあげた。

今後は、画像で表示が困難な語義の特徴分析を進める。また、検索語の拡張方法を更に改善したい。特に**Wikipedia**に対しては、あらかじめ、各語義を、アルバム名やアーティスト名、市町村名など、よく出現するクラスに大別し、検索語や必要必要な画像タイプを決める等の工夫が必要である。

参照文献

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the Japanese WordNet. In *ACL-IJCNLP-2009 ALR*, pp. 1–8.

Francis Bond, Eric Nichols, Sanae Fujita, and Takaaki Tanaka. 2004. Acquiring an Ontology for a Fundamental Vocabulary. In *COLING-2004*, pp. 1319–1325.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE CVPR*. URL `http://www.image-net.org/`.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Ellen M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *ACM SIGIR-1994*, pp. 61–69.

池原悟, 宮崎雅弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林 良彦. 1997. 日本語語彙大系. 岩波書店.

海野裕也, 宮尾祐介, 辻井潤一. 2008. 自動獲得された言い換え表現を使った情報検索. 言語処理学会第*14*回年次大会, pp. 123–126.

笠原要, 佐藤浩史, Francis Bond, 田中貴秋, 藤田早苗, 金杉友子, 天野昭成. 2004. 「基本語意味データベース:lexeed」の構築. In *2004-NLC-159*, pp. 75–82.

藤井敦, 石川徹也. 2005. テキスト処理による画像の多義性解消と事典検索サイトへの応用. 言語処理学会第*11*回年次大会*(NLP-2005)*, pp. 1002–1005.

井手一郎, 柳井啓司. 2009. セマンティックギャップを越えて: 画像・映像の内容理解に向けて. 人工知能学会誌, pp. 691–699.