概念意味特徴の画像による表示とその解釈に関する分析

林 良彦
大阪大学 言語文化研究科
hayashi@lang.osaka-u.ac.jp

1　はじめに

言語の意味を画像などの言語以外のメディア情報と対応付けたり [4]，画像から抽出した高位の属性 (意味特徴) を意味表現に適用する [10] といったマルチモーダルな意味処理に関する研究が活発化しつつある．

ある条件 (例:人間) において特定の意味を適切に表現すると解釈される画像が，別の条件・状況においても同様の意味を想起させるものとして解釈されるとは限らない．このような内容 (content) と解釈 (interpretation) のギャップ [2] がどのような意味概念，あるいは，画像において起こりうるかを調べておくことは，言語と画像の対応付けの向上のために重要である．

そこで本研究では，[5] で収集した (a) 概念に対する意味特徴 (例:*beaver* builds dams; *beaver* chews on woords) に対して Web から得た画像 (以下，Web 画像)，および，(b) それらの概念意味特徴に対する適合度 (以下，Web 画像適合度) に関するデータを用い，Web 画像適合度の評定を行った評定者とは異なる評定者 2 名により言語注釈を付与し，画像検索におけるクエリとして用いた概念意味特徴と第三者の評定者による言語注釈との間の類似度と，Web 画像適合度の間の関連性の分析を試みた．その結果，両者の間には中程度の相関関係が認められること，適合度の予測においては，とくに動詞の類似度が重要な役割を果たすことが分かった．

2　言語注釈の付与

2.1　概念意味特徴の Web 画像適合度

[5] では，541 種の概念 (例:beaver)，これらに対する 7,526 件の概念意味特徴 (例:beaver builds dams; beaver chews on woods) を収集した McRae のデータベース [8] から，(1) 生物・無生物の何らかの動作・振る舞いを表す 535 件の概念意味特徴を 218 種の概念に対して抽出し，(2) これらの概念意味特徴から 4 通りの手段により検索クエリを生成し，(3) Web 画像検索エンジン[1]を利用して画像を収集し，(4) 得られた各画像の概念意味特徴に対する適合度を 0〜4 の 5 段階で評定した．本報告の分析では，このデータを用いる．

2.2　言語注釈付与の対象データ

上記のデータの中から，今回は工数上の制約から，特に以下の条件に適合する 239 件の概念意味特徴に対する 3,653 点の Web 画像を言語注釈を付与する対象データとして抽出した．また，[5] の検討において，比較的良好な画像が収集可能であった動詞を ”-ing 形” とする検索クエリ (例: beaver builds dams に対して”beaver building dams”) を用いて収集した画像を対象とした．

- NDCG (Normalized Discounted Cumulative Gain) [7] を，あるクエリ (すなわち，概念意味特徴) の検索結果集合の良否を表す指標 (すなわち，概念意味特徴の Web 画像適合度) を表す指標と考え，この値がある程度高い (≥ 0.4) 概念意味特徴を選択する．
- ただし，画像適合度が 0 の画像を含む概念意味特徴は除外する．

結果として抽出された 3,653 点の Web 画像において，画像適合度が比較的高い (≥ 3) の画像の件数は 2,886 件 (74.8%) であったのに対し，比較的低い (< 3) ものが 767 件 (25.2%) であった．

2.3　言語注釈の付与作業

2 名の評定者により言語注釈の付与を行った．これらの評定者は十分に英語に堪能であり，[5] の画像適合度の評定には関わっていない．画像検索におけるク

[1] Google Images (http://images.google.co.jp/) により，2012 年 11〜12 月に収集．

表 1: 概念意味特徴の表現パターンと画像件数.

表現パターン	例	件数
動作主+動作	tortoise swims	1,882
動作主+動作+対象	cat eats mice	926
動作主+動作 (+対象)	goat eats ϕ	681
動作主+動作+様態など	zebra travels in herds	164
計		3,653

画像適合度: 3

概念意味特徴: airplane crashes
言語注釈-1: A plane disintegrates.
言語注釈-2: The plane explodes.

図 1: Web 画像に対する画像適合度と言語注釈の例.

エリのもととなるそもそもの概念意味特徴との意味ある比較を行えるようにするため，以下のような方針で言語注釈の付与を行なってもらった.

- 概念意味特徴の言語表現の文型パターンを表 1 のように分類し，極力この文型に即した言語表現を付与する.
- ただし，当該の概念意味特徴を有するターゲット概念 (概念意味特徴において主語名詞となる) は，教示しない.

McRae のデータベースのデータベースから得られる概念意味特徴，これに対して得られた Web 画像及び画像適合度，2 名の評定者により付与された言語注釈の例を図 1 に示す.

3 言語表現の類似性と画像適合度の間の関連性の分析

3.1 関連性の分析の概要

まず，McRae のデータベースに与えられている概念意味特徴と評定者による言語注釈の間の (総合的な) 類似度を考え，これがいくつかの要素についての類似度から計算できると考える．次に，このようにして求めた総合的な類似度と当該の画像に対して与えられている適合度に相関関係があると仮定し，重回帰分析を行う．このような関連性の分析を行うことにより，上記の相関関係の仮定を検証し，さらに，画像適合度の予測に寄与しうる要素の議論が可能となる.

なお，重回帰分析の手法としては，L1 正則項を考慮に入れた線形重回帰分析 (Lasso 法)[6][2]，および，サポートベクター回帰 (SVR)[3]を適用した．本件では従属変数に相当する Web 画像適合度は離散的な段階であるが，5 段階という多段となっているため，これらをスコアとみなし，回帰分析を適用した.

3.2 言語表現間の類似度計算における要素

文の間の一般的な意味的な類似度に関する研究が活発化している [1] が，本検討では，表 1 に示したように，オリジナルの概念意味特徴の言語表現と今回付与した言語注釈の表現パターンが同様の構造に限定されていることから，両者の間の類似度を (1) 文字列の類似度，(2) 主語となる名詞の意味的類似度，(3) 動詞の意味的類似度，(4) 目的語の意味的類似度の 4 つの要素により計算できるものとして簡略化した．(2)〜(4) の文法的要素を抽出するために，Stanford CoreNLP[4]のパーサを利用し，SVO 以外の要素は無視した.

(1) の文字列の類似度については，Jaro-Winkler 尺度 [11] を用い，(2)〜(4) の文法的要素における単語の意味的類似については，いずれも WordNet に基づく類似度尺度 [3] である，Lin の類似度と Wu-Palmer の類似度の平均を用いた．これらの値は，0.0(類似度低)〜1.0(類似度高) の範囲をとる.

4 結果と考察

4.1 関連性の分析

表 2 に関連性の分析結果を示す．数値は全て，Pearson の相関係数 (いずれも $p < 0.001$) である．カラム rel は画像ごとの画像適合度と重回帰分析による予測値との相関であり，カラム NDCG は概念意味特徴ごとの NDCG 値と検索されている各画像の適合度に対する予測値との二乗平均誤差の間の相関である．なお SVR については 10 分割交差検定を行い，もっとも良い結果 (決定係数 R^2 が最大) が得られたモデル (このときの画像適合度と予測値との相関をカラム rel' に示

[2]AIC 基準による最適化を実施し，交差検定を行った場合と同様の結果を得ている.
[3]scikit-learn (`http://scikit-learn.org`)[9] を利用した.
[4]`http://nlp.stanford.edu/downloads/corenlp.shtml`

表 2: 関連性の分析結果 (Pearson 相関係数).

走行条件	Lasso 法		SVR		
	rel	NDCG	rel	NDCG	rel'
A: 評定者:a+b, 文字列類似度:あり	0.474	-0.549	0.574	-0.511	0.528
B: 評定者:a+b, 文字列類似度:なし	0.472	-0.549	0.556	-0.497	0.513
C: 評定者:a, 文字列類似度:なし	0.437	-0.556	0.481	-0.535	0.471
D: 評定者:b, 文字列類似度:なし	0.399	-0.561	0.425	-0.619	0.403
E: a, b 間類似度のみ	0.341	-0.552	0.371	-0.602	0.354

している) により，全データを再度予測させた結果を rel，NDCG カラムに示している．

この表からは，以下のことが言える．

- 全体として，画像適合度と適合度の予測値は中程度の正の相関を示す．つまり，画像適合度と，そもそもの概念意味特徴と評定者の言語注釈の類似度の間には一定の関連がある．Lasso 法と SVR による結果は，大きな傾向として類似しているが，SVR による場合の方が相関が高い傾向にある．
- 文字列類似度を用いる場合のほうが用いない場合よりも良い結果であるが，その差は大きくはない．これは，要素の意味的類似度のみで一定の予測ができることを示す．このため，以下では文字列類似度を用いない場合を検討する．
- それぞれの評定者の言語注釈を単独で用いる場合，評定者 a の言語注釈を用いる予測の方が画像適合度との相関が高いが，評定者 b の言語注釈を合わせて用いた場合，相関はさらに高くなる．すなわち，言語注釈の冗長性が有用である．
- 概念意味特徴 (すなわち画像検索クエリ) 単位でみた場合，その Web 画像適合度を近似すると考える NDCG 値と平均予測二乗誤差の間には中程度の負の相関が認められる．すなわち，概念意味特徴が”難しくなる”ほど，その画像に対する対する第三者による言語注釈は，オリジナルの概念意味特徴と類似していない傾向にある．
- NDCG 値と予測値の二乗平均誤差の間の負の相関は，画像ごとの予測との相関が低い評定者 b の言語注釈を用いる場合が最も大きい．この結果は一見予想と矛盾するが，総体的に評定者 b はランク上位に存在する画像に対して，そもそもの概念意味特徴と類似性が高くなる注釈を与えている傾向が予測される．
- 走行条件 A〜D は，そもそもの概念意味特徴と評定者による言語注釈との類似度に基づくが，走行条件 E は，両者による言語注釈の間の類似度だけを用い，そもそもの概念意味特徴を用いない場合の結果を調べてみたものである．これによっても，ある程度の予測はできている[5]が，その相関は決して高くはないことから，第三者による言語注釈間の類似度のみから画像適合度を予測することは困難であることが示唆される．

(a) "cheetah hunts" (b) "faucet leaks"

図 2: 概念意味特徴と Web 画像 (適合度:4) の例.

4.2 意味的類似度要素の寄与

表 3 に走行条件 A，B の場合の両手法による各意味的類似度要素の (偏相関) 係数を示す．この結果から示唆されることは，とくに概念意味特徴と言語注釈で用いられる動詞間の一致度，類似度が画像適合度の予測に大きく関わっているということである．

4.3 概念意味特徴に関する考察

概念意味特徴 (すなわち検索クエリ) に対して得られた画像集合から計算される NDCG 値と予測値の平均二乗誤差を乗じた値が大きいほど，検索された画像集合の概念意味特徴に対する適合度が高いにもかかわらず，第三者は別の見方をしうる可能性があることを示す．今回の結果に関して，このような概念意味特徴を調査してみたが，一定の明確な傾向が得られるまで

[5] ランダムな予測値系列を発生させた場合は，その相関係数はほぼ 0.0 であり，p 値も有意水準を全く満たさない

表 3: 各類似度要素に対する係数. (a,b は評定者; Lasso 法の切片はいずれの場合も 1.8〜2.0 程度)

走行条件	手法	a:文字列類似度	b:文字列類似度	a:主語	a:動詞	a:目的語	b:主語	b:動詞	b:目的語
A	Lasso 法	0.266	0.252	-0.329	0.966	0.051	0.164	0.653	0.0
	SVR	2.494	1.370	-0.305	6.629	-1.357	4.083	5.306	-0.917
B	Lasso 法	-	-	-0.273	1.011	0.060	0.193	0.685	0.0
	SVR	-	-	-3.446	0.258	-0.346	3.360	1.638	-3.387

には至っていない．そこで，以下ではいくつかの例を示し，そこから示唆される傾向について考察する．

図 2(a) は，”cheetah hunts” という概念意味特徴に関して，適合度:4 と判定された画像であるが，”cheetah chases prey”，および，”cheetah runs” という言語注釈が与えられている．hunt をするにはその対象 (獲物) が存在するので，前者の評定者はそのような状況をなるべく述べようとしたと推測できる．一方，後者の評定者は行為の対象よりもターゲットの概念 (cheetah) の動作 (run) に焦点を当てたと思われる．

図 2(b) は，”faucet leaks”という概念意味特徴に関して，適合度:4 と判定された画像であるが，”water drips”，および，”The faucet drips”という言語注釈が与えられている．まず，何をターゲット概念 (すなわち主語) と捉えるかが両者で異なっている．また，動詞はいずれも”drip”が用いられているが，これは,「液体が滴る」という一般的な物理現象を描写するのに対し，そもそもの概念意味特徴における”leak”は「蛇口・水道などから水漏れが生じている」という事態を描写しており，それらの意味レベルには差がある．

以上のように，評定者は画像に描写されている状況をなるべく忠実に描写しようとするが，注釈の表現がバリエーションを持つことは避けられないと言える．

5 おわりに

画像が単独で提示されたとき，その画像にどのような解釈を与えるかは個人による差が大きく，対応する言語注釈も豊富なバリエーションを有する．[2] は，本研究で扱うような概念意味特徴に対応する画像の意味は，図像的 (iconographic) であるとし，その注釈は複雑な言語構造を持ちうると指摘している．

今回の結果からは，このようなバリエーションには，ターゲットとなる概念の特性，画像に描写される行い・振る舞いの典型性，あるいは，顕在性，ターゲットと同時に描写されている事物の特性などのほか，画像の画像的な特徴 (構図や動きの表現など) も大きく関わっていることが示唆されるが，一方で，複数の評定者がほぼ同じ表現による解釈を与える画像が存在するのも事実である．今後は，言語表現の類似度の頑健性を高めるとともに，画像的特徴を要因として取り込んでいきたい．

謝辞

本研究は，JSPS 科研費#24650123 の助成を受けた．

参考文献

[1] Agirre, E., et al. 2012. SemEval-2012 Task 6: A Pilot on semantic textual similarity. *Proc. of *SEM 2012: The First Joint Conference on Lexical and Computational Semantics*.

[2] Alm, C.O. et al. 2006. Challenges for annotating images for sense disambiguation. *Proc. of the Workshop on Frontiers in Linguistically Annotated Corpora*, pp.1–4.

[3] Bundanitsky, A., and Hirst, G. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, Vol.32, No.1, pp.13–47.

[4] Fujita, S. and Nagata, M. 2010. Enriching dictionaries with images from the Internet. *Proc. of COLING 2010*, pp.331–339.

[5] 林 良彦. 2013. Web 画像による語義・概念の視覚的な提示に関する検討. 人工知能学会 インタラクティブ情報アクセスと可視化マイニング研究会, 第 5 回研究会研究発表予稿集, SIG-AM-05-04, pp.19–26.

[6] 川野秀一，他. 2010. 回帰モデリングと L1 型正則化法の最近の展開. 日本統計学会誌, Vol.39, No.2, pp.211–242.

[7] Manning, C.D. et al. 2008. *Introduction to Information Retrieval*, Cambridge University Press.

[8] McRae, K. et al. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments, and Computers*, 37(4):547–559.

[9] Pedregosa et al. 2011. Scikit-learn: machine learning in Python. *JMLR* 12, pp.2825–2830.

[10] Silberer, C., Ferrari, V., and Lapata, M. 2013. Models of semantic representation with visual attributes *Proc. of ACL 2013*, pp.572–582.

[11] Winkler, W. E. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proc. of the Section on Survey Research Methods (American Statistical Association)*, pp.354–359.