CD-003

ソーシャルメディアの画像タグ共起情報からの単語分散表現の生成
Generating word embeddings from social-media originated image tag co-occurrnces

長谷川 美夏 † 小林 哲則 † 林 良彦 †
Mika Hasegawa Tetsunori Kobayashi Yoshihiko Hayashi

1 はじめに

単語に対して，その意味を適切に表す表現を与えることは，自然言語処理における最も基本的な問題の一つである．現在では，大規模なテキストコーパスを用いた統計的処理または機械学習により，単語に対する実数値ベクトル (分散表現) を与える手法が広く用いられている [1]．一方で，単語の意味表現には人間の持つ知覚や経験に基づく情報を取り入れるべきという立場 [2] もあり，画像に代表される非テキストモーダルな情報を統合する研究も盛んになっている [3, 4]．

本稿では，これらとは異なる非テキストモーダル情報として，ソーシャルメディア上で画像に付与されたタグから単語の意味表現を生成する手法を提案する．このタグ情報には，投稿画像に対する投稿者の投稿意図が反映されていると考えられるため，テキストや画像の特徴抽出から得られるものとは異なる属性が得られることが期待できる．Flickr の画像タグを収集した YFCC100M [5] という大規模なデータセットを用いた意味的類似度・関連度タスクの実験結果からは，テキストコーパスから抽出した分散表現に匹敵する性能が得られた．また，テキストから得た分散表現では困難とされてきた，対義語の同義語・関連語からの弁別に適した意味表現である可能性が示された．

2 関連研究

単語の意味表現は大規模なテキストコーパスを用いて構成するのが一般的であるが，人間が単語の語義を決定する際に知覚情報 (特に視覚情報) が大きな影響を与えることから [6]，テキストと画像をモダリティとして構成したマルチモーダル意味表現が注目を集めている．

一方で，意味表現を構成するための情報源としては，上記のような知覚情報の他に，ソーシャルメディアから獲得できる情報を用いることが考えられる．例えば，Flickr や Instagram といった画像投稿型のソーシャルメディアにおいては，画像の他に，タグ，アノテーション，また，位置情報や日付などの要素も含まれる．特に，タグやアノテーションには，投稿した画像が広く検索され閲覧されるようにという投稿者の意図が言語表現として反映されるため，従来の言語・画像特徴が持つ特性とは異なる意味表現を構成できる可能性がある．

2.1 マルチモーダル意味表現

言語情報と視覚情報によるマルチモーダル意味表現の構成手法には，それぞれの特徴をよりお互いの情報が補完しあう関係として働くよう重み付けや特異値分解をして連結するものや [7, 8]，ニューラルネットワークにより 2 つのモダリティを同時に学習するもの [3, 4] がある．構成した表現は，単語のペアの類似度・関連度を人手でスコアリングしたデータセット (詳細は 4.2 で述べる) との意味的関連度の推定タスクで評価される．

視覚情報の情報源として用いられる画像データセットにおいては，収録される各画像に対して，それを表す言語表現が与えられている必要がある．従来研究においてよく用いられてきた画像データセットには，ImageNet [9]，ESP-Game dataset [10]，Flickr [11] などがある．ImageNet は，WordNet における synset が指示する対象オブジェクトが中心に位置した高品質な画像を収集している．一方，ESP-Game dataset や Flickr における画像は，特定の対象オブジェクトを対象としたものではなく，1 枚の画像につき複数タグが付与されている．これらには，背景情報に関するタグや状況を表す形容詞のタグも含まれる [12]．このように画像の性質は大きく異なるが，Kiela ら [7] の実験結果では同程度の推定精度が得られており，マルチモーダル意味表現を構成する際の画像情報源としての優劣はつけがたい．

2.2 ソーシャルメディア情報の利用

ソーシャルメディアの情報は人間の生活に密接していること，情報がリアルタイムに増えていくことから，さまざまなタスクの学習データ源として近年注目を集めている．ソーシャルメディア情報を用いて何らかの意味を抽出しようとする研究には，Instagram のタグ情報のみを用いて，その共起性から単語間の情報量の大小といった語間関係抽出を行うものや [13]，YFCC100M の位置情報と画像情報を用いて地域と時間による単語の概念の分析を行うもの [14] がある．

†早稲田大学理工学術院 Faculty of Science and Engineering, Waseda University

3　提案手法

本研究では，画像投稿型のソーシャルメディアを情報源とし，そこから得られる情報から単語の意味表現を構成することを目的としている．ただし，収録されている画像から得られる視覚情報は一切用いず，画像に付与されたタグの共起情報のみから単語の意味表現を構成し，テキストコーパスから抽出した意味表現との違いを調査する．このタグの単語共起情報には，視覚的に共起しやすいオブジェクトの情報，投稿画像を広く見て欲しいという投稿者の意図を反映した情報が含まれると考えられる．つまり，画像に付与されたタグ情報は，画像本体に対する代替物の役割を果たしているとみなす．

3.1　タグの単語共起行列

1 枚の画像につき複数タグが付与されている画像データセットのタグ情報のみを利用した単語共起行列を構成する．付与されているタグのうち N 単語について，$N \times N$ の単語共起行列 M を作成し，同じ画像に付与されたタグを数え上げる．しかし，この単語共起行列は N の値が大きいと疎な行列となる上，どの単語とも比較的共起しやすい単語の情報，即ちその単語の意味を決定するにおいて大きな影響を与えないような情報が含まれている．そこで，それらの影響を軽減するため，正の自己相互情報量をとり，特異値分解によって次元を圧縮する．

3.2　正の自己相互情報量の計算

タグの単語共起行列 M について正の自己相互情報量 (Positive PMI: PPMI) を取る．単語 w_i，単語 w_j の自己相互情報量は，独立事象の場合は 0，独立事象よりも共起が少ない場合負の値となるため，単語の共起性から意味表現を得るならば PPMI の情報で十分である．単語共起行列 M について，単語 w_i と単語 w_j の PPMI を以下の式で定める．

$$M_{w_i,w_j} = \max\left(0, \log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}\right) \quad (1)$$

$$= \max(0, \log f(w_i,w_j) + \log f(*,*) - \log f(*,w_j) - \log f(w_i,*)) \quad (2)$$

$f(w_i,w_j)$ は単語 w_i と単語 w_j の共起回数，$f(*,*)$ は全単語出現回数，$f(w_i,*)$, $f(*,w_j)$ はそれぞれ単語 w_i，単語 w_j の出現回数である．これを全単語について導出する．PMI により，単語の出現頻度に左右されない単語間の共起の強さを計量することができる．

3.3　特異値分解

正の自己相互情報量を計算した $N \times N$ の単語共起行列 M_p について，特異値分解をおこない低次元 $d(d < N)$ に圧縮する．特異値分解の効果として，特異値が上位の情報のみを残すことで，重要度の低い情報を削減できる．つまり，単語同士が類似しているものはよりコサイン類似度が高く，類似していないものはよりコサイン類似度が低く計算されることが期待される．

$$M = U \cdot \Sigma \cdot V^T \quad (3)$$

$$\approx U^{(d)} \cdot \Sigma^{(d)} \cdot V^{T(d)} \quad (4)$$

ここで単語 w_t のベクトル表現 v_{w_t} を $v_{w_t} = U^{(d)}_{w_t} \cdot \sqrt{\Sigma^{(d)}}$ とする．これにより，N 単語に関して d 次元の意味表現ベクトルを得る．

4　評価実験

本研究では，画像に付与されたタグの共起情報から単語の意味表現を構成し，従来研究と同様に，意味的関連度のタスクにより評価を行う．

4.1　タグ情報源

本研究では，Yahoo Flickr Creative Commons 100M (YFCC100M)[5] というデータセットを用いる．このデータセットには，Flickr に投稿された約 9,930 万枚の画像と約 70 万の動画について，画像タイトル，ユーザータグ，マシンタグ，URL などのメタデータが含まれているが，本研究ではユーザータグ (投稿者が投稿時に付与したタグ) の情報のみを用いて単語の意味表現を構成する．ユーザータグが付与されているデータは全部で 68,971,123 件 (画像 68,552,616 件，動画 418,507 件) あり，1 データにつき平均 7 個のタグが付与されていた．この約 6,900 万件のタグ共起データから，1,500 件以上の画像に付与されていて，アルファベットのみからなる単独の単語として現れる 20,943 単語について単語共起行列を構成した．この際，全ての単語は小文字表記に統一している．

図 1: YFCC100M の例

4.2　評価データセット

さまざまな目的で開発された，以下の単語ペアの類似度・関連度のデータセットを評価に用いる．単語の実数値ベクトルのコサイン類似度で単語ペアの類似度・関連度の推定値を表し，スピアマンの順位相関係数によってゴールドデータとの一致度を評価する．

- **YP130** [15]: WordNet の動詞表現類似度の能力を調査する目的で，動詞表現について関連性をネイティブスピーカーに評価させたデータを収録している．
- **WordSim353** [16]: 単語ペアの関連度を評定したデータを収録している．単語のペアは同義語，反義語，上位・下位概念などさまざまな 9 つの関係に分類できる．
- **SimLex999 [17] / USF Assoc [18]**: 単語ペアに対して類似度 (similarity) と関連度 (relatedness) や連想強度 (association) を明確に区別して評価している．MEN や WS353 と違い，比較する単語の品詞は必ず同一である．
- **MEN** [8]: マルチモーダルな意味表現を評価する目的で作られている．単語ペアの品詞は必ずしも同じではなく，収録単語は ESP-Game や Flickr のタグに由来しているため具体概念に偏っている．
- **SemSim / VisSim** [3]: 意味空間モデルの開発と評価を目的としている．名詞単語のペアについて意味的に類似しているか，視覚的に類似しているかを別々に評価している．

5 実験結果

5.1 意味的関連度の推定精度とデータ数

表 1 は，各評価データセットに対する相関係数を意味表現ごとに比較している．ここで，YFCC 由来の意味表現は 1,000 万件分のタグデータを用いて構成したものであり，表の 2 列目は評価データセットに収録される単語ペアの数と，評価実験に用いた 20,943 種類の単語で構成可能なペアの網羅度を示す． 比較対象の単語意味表現として，Wikipedia 2009 [1]と GoogleNews [2]による 300 次元のコーパス由来の Skip-Gram 分散表現を用いた．

本報告の主眼である YFCC による MEN における相関係数は 0.81 であり，コーパス由来の 2 つの意味表現に比べ特に高い結果を得た．具体概念中心のデータセットである MEN の評価で相関係数が高い表現が得られたことより，ソーシャルメディアのタグ共起情報は具体概念の意味表現において有益であると考えられる．また，SimLex999 における相関係数は，GoogleNew の結果にはやや劣るものの Wikipedia の結果よりは高く，意味的類似度の観点でも既存の表現と同程度に性能が高い表現が得られているといえる．YP130 は単語の網羅性が 16%低いことから，画像投稿型のソーシャルメディアにおいてタグ情報に動詞表現を付与する可能性は低いと言える．

次に，単語共起行列の導出時に用いるデータ数をそれぞれ，1 万，10 万，100 万，1,000 万，6,900 万 (全部) 件としたときの，単語の意味表現の相関係数の推移を図 2 (縦軸:相関係数の値，横軸:データ数を対数表記) に示す．この結果から，全てのデータセットについて，1,000 万件まではデータ数が増えるほど相関係数が向上するが，それ以上は変わらない傾向にあることがわかる．すなわち，全データについて調査しなくても 1,000 万件程度で全体を調査するのと同等の意味表現を得ることができることが分かる．ただし，飽和に達するデータ量は対象の単語数にも依存すると想定される．

[1] http://mattmahoney.net/dc/textdata
[2] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

表 1: 意味的類似度・関連度の推定精度 (相関係数)

dataset	pairs	YFCC	Wiki	GNews
YP130	130 (16%)	0.47	0.35	0.24
WS353	353 (66%)	0.65	0.74	0.70
SimLex999	999 (54%)	**0.45**	0.39	0.49
USF Assoc	999 (54%)	0.34	0.38	0.44
MEN	3000 (96%)	**0.81**	0.74	0.77
SemSim	7576 (62%)	0.62	0.63	0.72
VisSim	7576 (62%)	0.49	0.50	0.55

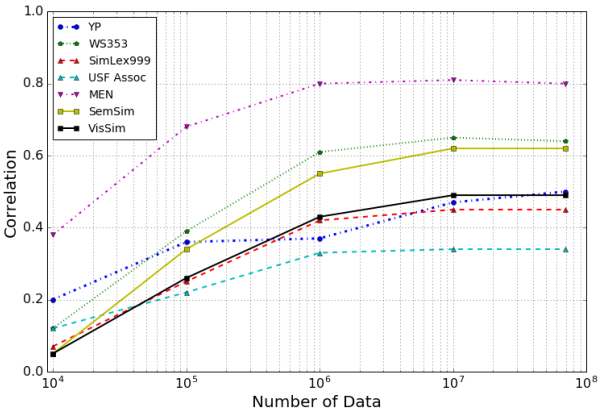

図 2: データ数と意味的関連度の推定精度の関係推移

5.2 画像タグ情報を用いた意味表現の特徴

画像タグ情報を用いた意味表現の特徴をさらに調べるため，ベクトル空間上で類似した単語がどのような意味関係にあるかを調査した．具体的には，比較する 3 つの意味表現 (YFCC, Wiki, GNews) 全てに存在する単語 11,928 語について，コサイン類似度が近い単語を求め，WordNet において以下の意味関係にある単語の平均逆順位を求めた．この結果を表 2 に示す．なお，各単語に対する意味関係の該当単語数の平均は，反義語 (antonym): 1.65 件，類義語 (synonym): 3.27 件，下位概念語 (hyponym): 5.34 件，上位概念語 (hypernym): 3.07 件であった．

この結果から考察されることは以下の 3 点である．

(1) ソーシャルメディアのタグ情報に由来する YFCC による意味表現を用いる場合，他の意味表現を用いる場合に比べて

反義語のランキングが低くなる傾向が確認できる．従来より，テキストコーパスから抽出した意味表現においては，同義・類義語を対義語やその他の意味関係による語から弁別することが課題となっており，その意味で反義語の類似度を低く押さえられるという傾向は望ましい傾向であるといえる．

(2) 類義語 (synonym) に関する結果は，YFCC と GNews において同程度である．これより，同じオブジェクトを表す可能性が高い類義語は共起しやすいこと，あるいは，より多くの人に検索させるために類義語のタグを網羅的につけていることが示唆される．

(3) 一方で，上位概念語 (hypernym) や下位概念語 (hyponym) のランクは，YFCC による意味表現を用いる場合の方が高い傾向にある．ソーシャルメディアにユーザーがタグを付ける際に，検索されやすさを考慮して，適当な上位語やより具体的な概念である下位概念語をタグ情報に含めていることが推測される．

表 2: 単語属性ごとの平均逆順位

	pairs	YFCC	Wiki	GNews
antonym	798	0.05	**0.18**	0.13
synonym	3593	0.15	0.09	**0.16**
hyponym	1900	**0.11**	0.04	0.07
hypernym	4163	**0.06**	0.02	0.04

6 おわりに

本稿では，ソーシャルメディアにおける画像タグ共起情報のみから構成した意味表現が，単語の意味的類似度・関連度の推定において，コーパス由来の分散表現と同等の性能を発揮することを示した．また，構成した意味表現はソーシャルメディアのタグ由来である特徴を反映し，類義語と反義語の弁別に適している可能性があること，一方で上位・下位概念語の関連度を高く評価する傾向にあることが確認できた．今後の方向性としては，従来から研究されてきた，テキストコーパスから得られる言語特徴，画像から抽出する画像特徴に加え，画像タグから抽出される情報を適切に組み合わせることにより，新たな意味表現の構成について研究することが考えられる．なお，本稿では英語を対象としアルファベットのタグのみを用いて実験を行ったが，YFCC のデータには多言語のタグが付与されているため，画像を仲介とする多言語の意味表現への展開も考えられる．また，ソーシャルメディアの特徴として日々情報が増えることや時代と共に情報の質が変化していくことが挙げられるため，意味表現の時間変化についても調査する価値があると考えられる．

参考文献

[1] M Baroni, G Dinu, and G Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Acl*, pages 238–247, 2014.

[2] Lawrence W Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59(August):617–645, 2008.

[3] Carina Silberer, Vittorio Ferrari, and Mirella Lapata. Visually Grounded Meaning Representations. *IEEE Trans. Pattern Anal. Mach. Intell.*, X(X):1–1, 2016.

[4] 長谷川 美夏, 小林 哲則, and 林 良彦. 画像によって単語意味表現をエンハンスするニューラルネットワークモデル (ViEW model). *NLP2017*, 4(C):1–4, 2017.

[5] Bart Thomee, David A Shamma, Gerald Friedland, Douglas Poland, and Damian Borth. YFCC100M : The New Data in Multimedia Research. *Commun. ACM*, 2016.

[6] Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behav. Res. Methods*, 37(4):547–559, 2005.

[7] Douwe Kiela and Leon Bottou. Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. *Emnlp-2014*, pages 36–45, 2014.

[8] Elia Bruni, Daniel Gatica-perez, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Intell. Res.*, 49(December):1–47, 2014.

[9] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 248–255, 2009.

[10] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. *Proc. 2004 Conf. Hum. factors Comput. Syst. - CHI '04*, pages 319–326, 2004.

[11] Shane Bergsma and Randy Goebel. Using Visual Information to Predict Lexical Preference. *Ranlp*, 6(2008):399–405, 2011.

[12] Douwe Kiela, Stephen Clark, Anita L Ver, Stephen Clark, Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics. *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 34(2):447–456, 2016.

[13] 前西 鷹 and 田島 敬史. SNS 上のタグ付き写真データセットからの語間関係抽出. *DEIM Forum 2017*, 8:1–8, 2017.

[14] ボルドビレグサイハン, 及川 雄介, 伊藤 祥文, and 柳井 啓司. 位置情報付き画像を用いた単語概念の時間変化の分析. *DEIM Forum 2016*, 8(2):103–110, 2016.

[15] Dongqiang Yang and David Mw Powers. Verb Similarity on the Taxonomy of WordNet. *Proc. GWC-06*, pages 121—-128, 2006.

[16] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131, 2002.

[17] Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Comput. Linguist.*, 41(4):665–695, 2015.

[18] D. L. Nelson, C. L. McEvoy, and T. a. Schreiber. The University of South Florida free association, rhyme, and word fragment norms. *Behav. Res. Methods, Instruments, Comput.*, 36(3):402–407, 2004.