君の名は －画像認識対象の名称獲得－

那須川 哲哉　村岡 雅康
日本アイ・ビー・エム株式会社 東京基礎研究所

1. はじめに

近年、深層学習などの技術の進展によって画像認識の精度が大幅に向上し、画像上の多様な物体を特定できるようになってきている。視覚的に特定できる対象の情報を自然言語でやり取りするためには、その対象の言語表現、すなわち名称が必要となる。

本稿では、画像認識対象の名称を大量のデータから獲得するタスクを提案し、このタスクの意義や性質を検討した上で、百万件規模の実データを用いて、このタスクを試行した結果を示す。

2. 画像認識対象の名称獲得タスクと意義

ある画像に写っている物体を認識したり、特定の物体が写っている画像を選択したりすることを実現する画像認識器には、多くの場合、画像ラベルとして、認識対象を示す言語表現が紐付けられている。例えば、ラーメン画像認識器に「ラーメン」という表現を紐付けることで、ある画像にラーメンが映っていると人間が解釈できるようになっている。しかし、その認識対象の言語表現が「ラーメン」のみであるとは限らない。その画像を見る人により「中華そば」や「味噌ラーメン」と表現されたり、あるいは「らーめん」と表記されたりする可能性がある。

同じ物を見ても、それを伝えるのに、人により、状況により、異なる表現が使われるのが自然言語の特徴である。例えば、図1は、見る人や伝える相手表現する際の状況などによって、「犬」「トイプードル」「プーちゃん」「動物」「dog」「Hund」といった多様な表現と対応付けられることになる。但し、ある物体に対し、（固有名詞などの特殊な場合を除き）任意の表現が紐付けられるわけではない。言語や文化などに応じたコミュニティでの共通表現を用いることによって、その物体に関するコミュニケーションが可能になる。従って、ある対象を何と表現するかは、コミュニケーションのために重要な知識であり、その知識が不足しているとコミュニケーションが困難になる。例えば、ある物品を購入しようと店に行って、それが見つからない場合、その名称を知らなければ、売り場や在庫の有無を店員に訊く際に苦労することになる。

画像認識技術が発達し、任意の画像に対応する認識器を比較的容易に構築できるようになり、言わば機械に視覚能力を持たせることができるようになった。そこで、視覚能力と言語能力をつなげるための

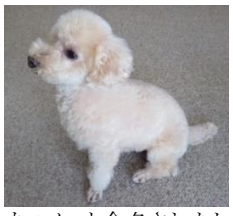

図1：「プーちゃん」と命名されたトイプードル

知識として、その認識対象の言語表現、すなわち名称を、人間が教えるのではなく、データから獲得しようというのが本稿で提案するタスクである。この知識は、例えば、「ハワイ語の雨や風を表す言葉は100を越える」といった報告[1]があるように、言語、地域、文化などに依存し、このタスクを通じて獲得されるデータ自体に学術的な価値が認められる可能性がある。また、画像及びその認識は言語に依存しないため、多様な言語において同じ対象の名称を獲得すれば、翻訳知識として、多言語翻訳や多言語検索に応用できそうである。

産業応用としては、商品パッケージや社名ロゴなどの画像に対する名称から、商品名や社名の通称を調査し、SNS上での評判分析などへ活用することが考えられる。企業活動のグローバル化が進展する中、自社や競合他社の製品が海外でどう呼ばれているかを把握して市場調査に活かしたり、海外の製造現場における機器や環境の呼称を把握して技術支援やマニュアル整備などの製造効率向上に活かしたりといった応用において、言語非依存の画像を活用できるメリットが今後高まっていくと考えられる。

3. 名称獲得タスクの基本的性質と特徴

本タスクを実現するためには、基本的に画像とテキストが紐付いているデータが必要となる。次節における試行例では画像と共につぶやかれたツイートのデータを用いるが、ツイートに限らない。画像付きのブログ、画像とそのキャプションなど、画像とテキストの内容に関連性があるデータであれば、活用できる可能性がある。ネット上で流通するデータに占める動画の割合が高まる中、将来的には、動画における、キャプションや音声をベースにしたテキストと、そのテキストが生じた期間の一連の画像を活用するアプローチが有望そうである。

画像認識器が何らかの対象を認識した画像に紐付いたテキストから、その対象を示す表現（名称）を獲得するためには、その表現が、対象画像の内容を示していることを把握する必要がある。しかし、任

意のテキストから、特定の表現が対象画像を示すことを認識するのは容易でない。例えば、図1においては、キャプション中の表現から、この画像が「トイプードル」であると判断できそうであるが、図や画像の名称がキャプションに含まれるとは限らない。「2019年1月1日撮影」といった、画像内容とは直接関係無いキャプションが付く可能性もある。

次節では、統計的に、特定の画像認識内容と関連性の高い表現を名称候補とするアプローチを示す。例えば、リンゴ画像に紐付いているテキストには、「りんご」や「林檎」「ふじ」といった、リンゴの名称を示す表現が出現する確率が高いという仮説に基づき、大量のデータから、画像認識内容の名称を特定しようというアプローチである。

ここで注意すべき点として、ある表現が名称として通用するための共通認識が挙げられる。例えば、リンゴは、日本語を使うコミュニティにおいては「リンゴ」で通用するが、英語を用いるコミュニティでは、「apple」になる。従って、名称にはコミュニティにおける共通性が必要であり、コミュニケーションに利用可能な名称を特定する上では、複数の話者や筆者のデータを用い、世の中で通用する名称であることを確認する必要がある。この性質を本稿では「通用性」と呼び、利用目的に応じて通用性が高い表現を名称として獲得することを目指す。

通用性と並ぶ、もう一つの留意点として、話者や筆者の知識により、同じ対象に対し異なる名称が用いられる点が挙げられる。例えば、図1を見て、単に「犬」と表現する人もいれば、「プードル」、あるいは「動物」と表現する人もいると考えられる。こういった表現の使い分けは、見る人や、伝える相手、表現する際の状況などに依存するため、名称・呼称としては全て正しい表現と考えられる。しかし、その表現が示す概念の粒度や限定性が異なっている点を考慮する必要がある。同じ対象に対して多様な名称が存在し、その粒度が異なる場合、その名称が示す概念の間には、基本的には包含関係が成立すると考えられる。図1の例の表現であれば、各表現の概念の包含関係は下記のようになる。

動物 ＞ 犬 ＞ プードル ＞ トイプードル

この性質を本稿では「限定性」と呼ぶことにする。画像認識対象の名称獲得においては、この「通用性」と「限定性」を考慮する必要がある。

4. ツイートからの画像認識対象の名称獲得

2018年6月にTwitter APIを用いて、日常生活に関する「うどん」「カバン」「ポスト」などの約百語のキーワードのいずれかを含む約百万件（1057955件）のツイートを取得した[1]ところ、その12%弱（125590件）のデータに画像（のURL）が紐付いていた。この画像に対して、IBM Watson Visual Recognition[2]（以降、WVRと略記）による画像認識を適用したところ、各画像に関し、最も確信度の高い認識結果として、2528種類のラベルが付与された。高頻度上位10件のラベルを表1に示す。

表1：ツイート画像に付与された画像認識ラベル

画像認識ラベル	ツイート件数
food	25381
spectrum of colors	13310
nutrition	9463
person	6825
dish	4731
garment	3610
building	2988

この約百万件のツイートをIBM Watson Explorer Advanced Edition Analytical Components V12.1 [2]に投入し、各画像ラベルと各ツイート中の名詞表現との相関分析を行った。例えば、画像認識ラベルがfeline（ネコ科動物）のツイートが169件存在した。この169件中、3件以上のツイートに出現する名詞71語の中で、felineラベルとの自己相互情報量(PMI)が高い（正の値を取る）表現として、「猫」「ネコ」「黒猫」「毛」「ガムテープ」「朝」などが抽出された。すなわち、felineの画像と紐付いたツイートにおいては、「猫」「ネコ」「黒猫」といったfelineの名称を示す表現が他のツイートよりも高い確率で出現しているという結果が得られた。しかし、単に関連性が高いというだけでは、「毛」「ガムテープ」「朝」など、felineの名称とは言えない表現も抽出された。そこで、PMIの値を用いて抽出された、各画像認識ラベルと関連性の高い名詞表現を画像検索エンジン[3]にかけ、各名詞と紐付く100画像を取得し、取得した各画像を画像認識エンジンにかけ、元のラベルと同じ認識結果になるかどうかを検査する処理系を構築した。つまり、「猫」「ネコ」「黒猫」で検索された画像の大半は、画像認識エンジンにかけると、認識結果としてfelineというラベルが付与されるので、felineの名称であると判断し、「毛」「ガムテープ」「朝」で検索された画像の大半は、画像認識エンジンにかけると、feline以外の認識ラベルが付与されるので、felineの名称ではないと判断するようにした。

上記の手法を用いて、前述の約百万件のツイートデータから画像ラベルと名称候補となる名詞表現のペアを抽出した。画像認識器WVRは、認識結果として画像ラベルと共に0から1までの間の値の確信度を

[1] データの重複を避けるため、リツイートなどは取得対象外にした。

[2] https://www.ibm.com/watson/services/visual-recognition/

[3] ここでは、https://www.google.co.jp/imghp を利用した。

出力するため、名称候補で検索した100件の画像に対する画像ラベルの確信度の総和を求め、その値が20未満のペアは対象外とした。また、特定の画像ラベルを与えられた画像に紐付くツイートの3件以上に同じ名称候補が含まれていないケースも対象外とした。結果的に87種類の画像ラベルに対し、合計1362件の名称候補が抽出された。その一部を表2に示す。

表2：約百万件のツイートから抽出された名称候補

画像ラベル	ラベル付けされた画像に対する名称候補
bicycle	電動自転車, 自転車, チャリ, ロード, サイクリング, アシスト, 電動
hat	ハット, 帽子, キャップ
Sunset	sunset, 夕日, 夕陽, 夕焼け, 日の出, 日没, 夕方, 雲海
vegetable	ピーマン, ネギ, キャベツ, ほうれん草, もやし, 野菜, ポテト, サラダ

5. 獲得された名称の評価と考察

前節の抽出結果に含まれる87種類の画像ラベルのうち、personというラベルに関しては、名称候補に姓名や役職などが含まれ、他のラベルと性質が異なるため、対象外にすることにした。残る86種類に対する1301件の名称候補ペアからランダムサンプリングで10%の131件を選択し、画像ラベルの示す概念に対する名称としての妥当性を人手で評価した。このサンプル評価の目的は、今後大規模なデータの評価をクラウドソーシングで行なうための準備と評価者間の判断の一致度(inter-agreement)の確認である。評価者に対し、画像認識ラベルと抽出名称候補のペア131件を下記形式で提示するようにした。

「チンチラ」は「animal」(の一種)。
「サンゴ礁」は「nature」(の一種)。
「トンネル」は「road」(の一種)。

各ペアに対し、著者2名を含む5名の評価者が下記の選択肢から一つの値を選ぶ形で評価を行った。

a: 正しい
b: 正しいと思うが、確証は持てない
c: 正しくないが、関連はしている
d: 間違い
e: 判断できない

評価に際し、抽出名称候補の表現を評価者が知っていると限らない点を考慮した。例えば、footwear（履物）ラベルの画像に対する名称候補表現として、「靴」「長靴」「サンダル」といった表現の他に「アシックス」「NIKE」といった靴のブランド名が抽出され、知らないブランド名などが提示されるケースがある。その際、一般的に通用する名称と判断できるかどうかを検討するうえで、Wikipediaのみ参照しても良いという条件を付け、評価用シートからWikipediaの情報を簡単に参照できるようにした。そのため、評価結果として「e:判断できない」が選択された名称候補の大半はWikipediaの見出しに存在しない表現であった。

実際の評価結果の一部を表3に示す。5名の評価者全員が同じ値を選択したのは、131件中20件であった。abcdeの5つの選択肢のうち、abは正解とし、それ以外のcdeを不正解とする2択評価に置き換えると、5名の評価者全員が同じ値を選択したのは、131件中78件であった。5名の評価者の間のカッパ値を表4に示す。2択評価の場合は全員が中程度以上の一致度であったものの、5択の場合、著者2名の間のみが中程度の一致であり、その他の評価者の間の一致度は若干低かった。この違いの一つの理由は本タスクの背景を知らせずに評価依頼したためと考えられる。特に、評価データ提示の際、限定性を考慮し、

「名称候補」は「画像ラベル表現」(の一種)。

という形式で提示したため、例えば、下記のデータに対しては、両著者がaとしたのに対し、非著者3名の評価結果はabcに分かれていた。

「カバン」は「bag」(の一種)。
「木」は「tree」(の一種)。

名称獲得タスクの精度評価という観点から、正解とした評価者が3名以上のケースを正解、それ以外を不正解とすると、131件中70件が正解であり、61件が不正解であった。従って、131件の候補の正解率は53.4%となるが、この正解率は、名称候補で検索した100件の画像に対する画像ラベルの確信度の総和の閾値によって変化し、今回の評価データにおいては、表5のようになった。また、「c:正しくないが，関連はしている」と判断されたケースの具体例として下記などがあった。

「満開」は「plant（植物）」(の一種)。
「コーディネート」は「garment（衣類）」(の一種)。
「大会」は「sport」(の一種)。

今回の試行手法では、画像との単純な共起関係で名称候補の名詞を抽出しているため、plant画像と「満開」、garment画像と「コーディネート」といった表現は共起し易く、候補に含まれる。さらに、これらの表現で画像検索すると、「満開」からはplant画像が、「大会」からはsport画像が多く得られる傾向があるため、名称候補に残ることになる。こういった点で、名称獲得タスクには、今後の改善の余地があると考えられる。

サンプル評価を行った131件中、評価者全員がdもしくはeと判断したのは下記の3件のみであった。

「タケダ」は「food」(の一種)。
「漁業」は「vehicle」(の一種)。
「kazuharukina」は「garment」(の一種)。

この3件に関しては、名称候補で検索した100件の画像に対する画像ラベルの確信度の総和が、上から順に、33.606、25.145、21.884といずれも低く、閾値調整により、画像と全く無関係な表現を名称として抽出する確率は低く抑えることができそうである。

表 3:名称候補の評価結果の例 （右端の列の値は、名称候補で検索した 100 件の画像に対する画像ラベルの確信度の総和）

評価対象（名称候補と画像ラベル）	5 名による評価結果	精度評価用の判定	確信度の総和
「チンチラ」は「animal」(の一種)。	aaaaa	正解	91.151
「サンゴ礁」は「nature」(の一種)。	abaaa	正解	70.483
「車両」は「train」(の一種)。	accca	不正解	62.112
「宿泊」は「bedroom」(の一種)。	ccccc	不正解	40.37
「トンネル」は「road」(の一種)。	bcbab	正解	21.703

表 4: 5 名の評価者間の評価結果の一致度 （右上の数値が 5 択のカッパ値、左下の斜体の数値が 2 択のカッパ値）

	著者 1	著者 2	非著者 1	非著者 2	非著者 3
著者 1		0.454	0.292	0.379	0.264
著者 2	*0.784*		0.279	0.385	0.279
非著者 1	*0.568*	*0.561*		0.367	0.242
非著者 2	*0.721*	*0.663*	*0.623*		0.221
非著者 3	*0.517*	*0.535*	*0.488*	*0.530*	

表 5:名称候補で検索した 100 件の画像に対する画像ラベルの確信度の総和の閾値と名称獲得精度

確信度の総和の下限	正解数	不正解数	精度
80	24	0	100.0%
60	46	16	74.2%
40	60	34	63.8%
20	70	61	53.4%

得られた名称を活用する上では限定性も考慮する必要が出てくる。今回の試行手法では、画像ラベルの対象概念（例えばネコ）よりも上位概念（例えば動物）の名称を候補として抽出すると、その候補表現で検索した画像には対象概念以外（例えばイヌなど）の画像が含まれるため、画像ラベルの確信度の総和が低くなり、名称候補から外すことができる。獲得される名称のうち、基本的には、最も限定性が低い名称が画像ラベルの概念そのものの名称に近くなると考えられる。そのため、得られた名称のうち、最も限定性が低く上位概念を示すものを選べることが望ましい。与えられた表現のどちらが上位概念を示すかを判断する取り組みは既に存在[3-7]し、公開された評価データも存在する。我々は、別途、表現と紐付いた画像を利用することで上位—下位関係の判別精度を上げる取り組みを進めており、公開された評価データでは90%を超える精度で判断できるようになっている[8]。

6. おわりに

視覚情報と言語情報を結び付ける取り組みの一環として、画像認識対象の名称を獲得するタスクを提案し、12%弱の割合で画像が付いている約百万件のツイートデータから実際に名称を獲得した結果を示した。本稿の試行手法では、名称候補で検索した画像に対する画像ラベルの確信度の総和の閾値を上げることで高い精度で名称を抽出することができた。

既に、日本語に加え、英語やインドネシア語など多言語の解析も進めており、言語に依存せず同じ仕組みで名称を獲得できることが確認されている。現在、より大規模なデータでクラウドソーシングを利用した評価実験の取り組みを進めようとしており、評価データに関しては、より大規模なデータが揃った段階で、公開することを検討している。

謝辞

本稿の取り組みに関し、画像認識器の活用で IBM Research - Thomas J. Watson Research Center の Bishwaranjan Bhattacharjee 氏、試行実験で日本アイ・ビー・エム株式会社東京基礎研究所(当時)の Khan Md. Anwarus Salam 氏、結果評価で日本アイ・ビー・エム株式会社の石井旬氏、日本アイ・ビー・エム・システムズ・エンジニアリング株式会社の髙谷尚子氏、日本アイ・ビー・エム・サービス株式会社の小林容子氏からの多大なご支援をいただきました。ここに記して感謝致します。

参考文献

[1] 松名隆, 塩谷亨, アイヌ語とハワイ語の気象語彙に関する対照研究について. 室蘭認知科学研究会, 第44回大会プロシーディングズ, pp.12-17, 2004.

[2] Wei-Dong Zhu, et al. IBM Watson Content Analytics: Discovering Actionable Insight from Your Content. An IBM Redbooks publication. ISBN-10:0738439428. 2014.

[3] Kiela, Douwe, Laura Rimell, Ivan Vulić, and Stephen Clark. "Exploiting image generality for lexical entailment detection." In ACL-IJCNLP, vol.2, pp.119-124. 2015.

[4] Vulić, Ivan, and Nikola Mrkšić. Specialising word vectors for lexical entailment. In NAACL (Vol. 1), pp.1134–1145, 2018.

[5] Santus, E.; Lenci, A.; Lu, Q.; and Schulte im Walde, S. Chasing hypernyms in vector spaces with entropy. In EACL (Vol. 2), pp.38–42, 2014.

[6] Nguyen, K. A., Köper, M., Walde, S. S. I., & Vu, N. T. Hierarchical embeddings for hypernymy detection and directionality. In EMNLP, pp.233–243, 2017.

[7] Geffet, M., and Dagan, I. The distributional inclusion hy- potheses and lexical entailment. In ACL, pp.107–114, 2005.

[8] 村岡雅康, 那須川哲哉, 画像認識器の物体ラベルを活用した単語の特徴表現, 言語処理学会第 25 回年次大会予稿集, 2019.