画像検索エンジンを用いた概念知識の獲得

村岡 雅康　　那須川 哲哉　　金山 博
日本アイ・ビー・エム株式会社 東京基礎研究所
{mmuraoka, nasukawa, hkana}@jp.ibm.com

1　はじめに

グローバリゼーションの進展に伴い、多言語テキストマイニングの需要が高まっている [11, 12]。多言語テキストマイニングを行う上では、多言語のテキストから、分析対象となる個々の概念を示す表現（文字列）を正しく同定することは極めて重要である。本稿では、「概念を表す表現」として、何らかの実体を示す名詞・複合名詞や固有名詞を対象とする。例えば、

(1) A baseball player loves a newly released Big Mac.

という文において、“baseball palyer” や “Big Mac” を 1 つの概念を表す表現（以下、概念表現）とみなす。これらが正しく 1 つの概念として抽出できれば、love（baseball player, Big Mac）といった関係を取ることができ、頻出するパターンの分析が可能となる。

このような概念表現の同定を行うためにパイプライン処理と end-to-end 処理の 2 つのアプローチが考えられる。パイプライン処理は、前処理として品詞タグ付けや構文解析等の言語処理を行い、その後チャンキング、あるいは、固有表現抽出を行うものである。しかし、各処理を行うシステムは大抵の場合、統計的機械学習を用いた手法で構築されており、それらの手法の学習にはそれぞれ正解が付与された大量の学習データが必要となる。一方の end-to-end 処理は、与えられた入力文の表層形のみから直接チャンキング、あるいは、固有表現抽出を行う方法である。これは前述の前処理を必要としない点で優れており、特に近年はニューラルネットワークによる手法が高成績を収めている [1, 2, 5, 10]。しかしながら、大量の学習データが必要という点で上記のパイプライン処理と同様、学習データが存在しない言語や異なるドメインの文書に適用することが難しい。

この問題に対処するために、本稿では、画像検索エンジン、および、その結果としての画像集合に注目する。画像検索エンジンにある文字列をクエリとして入力すると、その文字列に対応した画像の集合を出力するようになっている。例えば、“Big Mac” というクエリを入力すると、検索結果の上位の画像は背景の色や写っている内容物の大きさや角度などに差異はあるものの、ほとんどすべての画像は同じ内容物の “Big Mac” を表している（図 2（a））。この性質を利用すれば、言語によらず、あるテキストに対応する画像集合を得ることができる。つまり、テキストから画像集合への変換器として画像検索エンジンを用いることで、様々な言語で書かれたテキストを画像集合を介して統一的に処理できる。

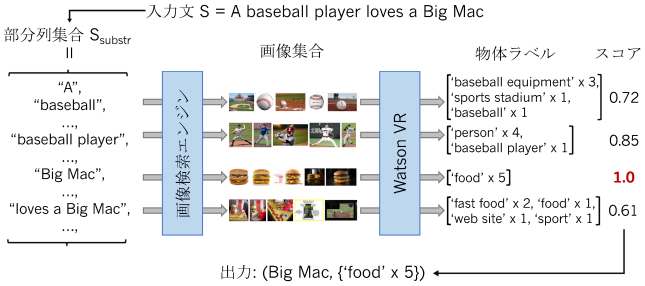

図 1: 画像検索エンジンを用いた概念表現獲得

また、検索結果としての画像集合を用いれば、クエリのテキストが、ある 1 つの概念を表すかどうかを判定できる可能性がある。上記では、画像検索エンジンに “Big Mac” と投げれば、検索結果上位に “Big Mac” の画像が集中したが、一方で、“Big Mac pro” や “free Big Mac” などで検索すると、上記の結果とは異なり、また、検索結果の画像に写る内容物にもばらつきが見られる（図 2（b）、（c））。言い換えれば、ある概念表現がクエリの場合はその時のみそれに対応する画像集合が得られ、そうでない場合は一貫性のない画像集合が得られる。

本稿では、これらの性質を利用することで言語やドメインに非依存の概念表現の同定手法を提案し、画像検索エンジン、および、画像集合を用いた概念表現の同定の可能性について調査する。

（a）“Big Mac” の画像検索結果

（b）“Big Mac pro” の画像検索結果

（c）“free Big Mac” の画像検索結果

図 2: 画像検索エンジンのクエリによる結果の比較[1]

2 関連研究

本節では本研究の取り組むタスクと関連のある研究について述べる。

2002 年、2003 年に固有表現抽出の CoNLL Shared Task[8, 9] が行われて以降、これらのデータセットが固有表現抽出のベンチマークとして用いられるようになり、様々な手法が提案されている。現時点での固有表現抽出の state-of-the-art は Yang らの手法 [10] である。Yang らは文字ベースの GRU（Gated Recurrent Unit）、および、単語ベースの GRU を組み合わせたニューラルネットワークを考案し、複数のタスクや言語のデータで学習する際、一部のパラメータを共有することで学習データが少ないタスクにおける精度向上を実現させている。しかしこの手法の学習は言語に依存したデータセットで行う必要があるという課題が残る。

一方で Mayhew らは、対訳辞書の集合である Panlex を用いて英語の固有表現抽出のデータセット (CoNLL2003-en) を他の言語に翻訳して、多言語の固有名詞抽出の擬似学習データを作成している [6]。しかし、固有表現抽出のモデルは素性として、Wikipedia で学習した Brawn cluster や wikifier という Wikipedia 由来の言語依存の素性を使っている。すなわち、Wikipedia が潤沢にない言語に関して良い結果が出る保証がない。対して提案手法は、画像検索エンジンを用いるため、対象を Wikipedia に限定せず、テキストおよびそれと対応する画像が存在する Web ページ全体とすることでこの問題に対処する。

我々の手法に最も近い手法が、Esteves らが提案したものである [4]。彼らも検索エンジンから得られる画像とテキストを用いて固有表現抽出のための素性を作っている。しかし、これも言語別に学習データが必要であるため、多言語にスケールしない。

3 提案手法

本節では、正解データを必要としない概念表現の同定方法について説明する。まず、画像検索エンジンを用いてテキストを画像へ変換する方法について説明し、続いて、得られた画像集合から概念表現らしさの判定方法について説明する。本手法の概要を図 1 に示す。

3.1 検索エンジンによるテキストから画像集合への変換

本手法は単語分割や品詞タグ付けの手段がない言語にも適用できるよう、そのような前処理を仮定しない設計とした。n 個の単語からなる入力文 $S = w_1, w_2, ..., w_n$ が与えられたとき、任意の部分列が概念表現となり得るため、文 S から得られる全ての部分列 $S_{substr} = \{s_{i,j} = w_i, ..., w_j | i < j, 1 \leq i, j \leq N\}$ について、画像検索エンジンを用いて上位 N 枚の画像 $\{img_{k,s_{i,j}} | 1 \leq k \leq N\}$ をそれぞれ保存する。本研究では、画像検索エンジンとして Google 画像検索エンジンを採用し、$N = 5$ とした。Google 画像検索エンジンでは URL にクエリパラメタを埋め込むことで、検索の設定を行うことができる。本研究では以下のクエリパラメタおよび値を明示的に指定した。

as_epq=$s_{i,j}$：部分列 $s_{i,j}$ を完全一致検索
tmb=isch：サーチタイプを画像検索に設定
pws=0：パーソナライズサーチを無効化
oe=utf-8：検索結果を utf-8 でエンコード
hl=en：インターフェイスの表示言語を英語に設定
gws_rd=cr：リダイレクトを無効化

例えば、“Big Mac” をクエリとした場合の URL は次のようになる：`https://www.google.com/search?as_epq=Big+Mac&tbm=isch&pws=0&oe=utf-8&hl=en&gws_rd=cr`

3.2 画像集合を用いた概念表現の判定

あるテキストに対応する画像集合を用いてそのテキストが概念表現かどうかを判定するために、本研究で

[1] （a） https://www.google.com/search?as_epq=Big+Mac&tbm=isch&pws=0&oe=utf-8&hl=en&gws_rd=cr
[1] （b） https://www.google.com/search?as_epq=Big+Mac+pro&tbm=isch&pws=0&oe=utf-8&hl=en&gws_rd=cr
[1] （c） https://www.google.com/search?as_epq=free+Big+Mac&tbm=isch&pws=0&oe=utf-8&hl=en&gws_rd=cr

は画像中の物体に着目する。3.1 節で獲得した画像に IBM Watson^{TM}Visual Recognition [2]（以下、Watson VR）を適用し、物体認識を行う。Watson VR は入力画像に対して物体認識を行い、1000 種類以上のラベルから確信度が高い上位 10 件を確信度付きで返す。本研究では、1 画像について最も確信度が高いラベルのみ考慮する。その結果、ある部分列 $s_{i,j}$ について最大 N 個のラベル $L_{s_{i,j}} = [l_{1,s_{i,j}}, l_{2,s_{i,j}}, ..., l_{m,s_{i,j}}](m \leq N)$ を得る。このラベルのリストに対して概念表現らしさを表すスコアを以下のように定義する：

$$score(L_{s_{i,j}}) = \frac{2.0}{(1 + \exp(-\sum_{i=1}^{3} w_i x_i)} \quad (1)$$

$$x_1 = m - N \quad (2)$$

$$x_2 = 1 - |set(L_{s_{i,j}})| \quad (3)$$

$$x_3 = \sum_{k=1}^{|set(L_{s_{i,j}})|} p_k \log p_k \quad (4)$$

$$p_k = \frac{\#(l_k)}{m} \quad (5)$$

ただし、$|set(L_{s_{i,j}})|$ はラベルのリスト $L_{s_{i,j}}$ に含まれるラベルの異なり数を表し、$\{w_i | i = 1, 2, 3\}$ はハイパーパラメータである[3]。また、画像が得られなかった場合、スコアは 0.0 とする。上記のスコア関数は以下のヒューリスティックスに基づいて設計した。

- 部分列 $s_{i,j}$ がある概念表現ならば、画像検索によってその部分列に対応する画像が一定数得られるため、得られるラベル数 m もより大きい方が良い
- 部分列 $s_{i,j}$ が曖昧性のない概念表現[4]ならば、得られる画像集合の内容物は同じものになると期待できるため、ラベルの異なり数 $|set(L_{s_{i,j}})|$ はより小さい方が良い
- ラベルのリストに含まれるラベルの異なり数 $|set(L_{s_{i,j}})|$ が同じ場合、ラベルの分布がより偏っていた方が 1 つの概念表現を表すと期待できるため、ラベルのリストに対する Negative entropy $\sum_{k=1}^{|set(L_{s_{i,j}})|} p_k \log p_k$ はより大きい方が良い

また、経験的に以下に示すラベルが出現した時はノイズの可能性が高いため、スコアにペナルティを課す。

$$L_{noisy} = \{\text{computer, newspaper,}$$
$$\text{machine, map, autoradiograph, window,}$$
$$\text{apparatus, print media, lines, comic book}\}$$

$$score(L_{s_{i,j}}) = score(L_{s_{i,j}}) * 10^{-penalty} \quad (6)$$

$$penalty = \sum_{l \in L_{s_{i,j}} \cap L_{noisy}} \#(l) \quad (7)$$

最終的に、上記のスコアが最大となる部分列、および、対応するラベルのリストを出力する：

$$output : (s^\star, \{l_{k,s^\star}\}) \quad (8)$$

$$s^\star = \underset{s_{i,j} \in S_{substr}}{\arg\max}\ score(L_{s_{i,j}}) \quad (9)$$

[2]https://visual-recognition-demo.ng.bluemix.net/

[3]4 節の実験では全てのデータセットにおいて $w_1 = 0.5, w_2 = 0.1, w_3 = 0.4$ とした。

[4]ある部分列 $s_{i,j}$ が概念表現だったとしても曖昧性がある場合（例えば “Apple” は果物の林檎と企業の Apple 社の少なくとも 2 通りの意味が考えられる）、本手法では獲得することが難しいため、本稿では取り扱わない。曖昧性のある概念表現については今後の課題とする。

表 1: 各評価データの文数

データセット名（言語）	設定	文数	評価文数	分類結果
CoNLL2002（スペイン語）[8]	上位 100 文	100	15	3
CoNLL2003（英語）[9]	$R = 10, r = 10$	133	40	22
Twitter NER（英語）[7]	全て	40	10	1
NERIL（ヒンディー語）[3][5]	$R = 100, r = 10$	16	3	0
NERIL（マラヤーラム語）[3][5]	$R = 100, r = 10$	13	2	0
NERIL（ベンガル語）[3][5]	$R = 100, r = 10$	17	4	2

4 実験

本節では、異なる言語、または、ドメインで書かれたテキストを用いて、提案手法の適用可能性を示す。

4.1 評価データ

データセットとして表 1 に示す 5 言語、2 ドメインからなる計 7 つのデータセットを用いる。本研究では、Watson VR の API の制限の都合上、次の方法で抽出した評価データで実験を行う。共通の前処理として、それぞれのデータセットのテストデータにおいて、重複する文、固有表現を含まない文、および、1 語のみからなる固有表現のみを含む文を除外し、単語数に関して昇順に並べる。CoNLL2002 のスペイン語に関しては上位 100 文、Twitter コーパスに関しては全文、それ以外のデータセットに関しては R 文ごとに r 番目の文を抽出する。抽出された各評価データの文数は表 1 の「文数」に示す。

4.2 結果

抽出された概念表現について定性的な調査を行なった。1 文内に固有表現が 1 つのみの文を対象とし、獲

[5]http://au-kbc.org/nlp/NER-FIRE2013/

表 2: 抽出された概念表現例

正誤	言語	部分列
正	スペイン語	“Palacio Municipal”
誤	スペイン語	“superviviente de la”
正	英語	“John Lewis UK”
正	英語	“chancellor of the exchequer”
誤	英語 (Twitter)	“this time”
誤	ヒンディー語	“उथल पुथल”
誤	マラヤーラム語	“എക്കാലത്തയും മികച്ച ഒരു”
正	ベンガル語	“জগদীশচন্দ্র বসু”

得した概念表現 78 個について全て人手で確認し、概念表現として正しいか否か分類した。それぞれのデータセットにおける評価文数、および、正しいと分類された概念表現の数を表 1 の「評価文数」「分類結果」に示す。結果として合計 78 個中 28 個が正しいと判定された。表 2 に具体例を示す。

スペイン語の “Palacio Municipal” や英語の “John Lewis UK”、表の最後のベンガル語は、それぞれ固有名詞であり、対応する画像が存在し、概念表現としても正しく抽出できている。また、“John Lewis UK” は画像のラベル [building × 4, retail store × 1] から、これは何らかの建物、もしくは、小売店であることが推測できる。一方で、“chancellor of the exchequer”（大蔵大臣の意）のような 1 種の定型表現も獲得できることを確認した。

しかし、スペイン語の “superviviente de la”(〜の生存者) や英語（Twitter）の “this time” のように、明らかに概念表現ではないものも多数確認された。このような事例において画像検索エンジンによって取得された画像集合にはそれぞれ人物が写っていたが、どれも違う人物、状況を表し、一貫性がないものであった。しかし Watson VR による物体ラベルは全て person となり、最終的なスコアとして高くなってしまうため、このような誤りが起きたと考えられる。これに対処するには、より高度な素性を考慮する必要がある。例えば、1 画像について 1 つの物体ラベルのみ考慮するのではなく、物体ラベルの集合や物体間の空間的、もしくは、semantic な関係まで考慮する必要がある。

5 おわりに

多言語や複数ドメインのテキストマイニングへ向けて、画像検索エンジン、および、その結果の画像集合を用いて単一のモデルで概念表現を獲得する手法の適用可能性を示した。5 言語、2 ドメインのデータセットにおいて、獲得された表現 78 個を人手で調査し、内 28 個が正しい概念表現であると確認できた。今後は精度向上に向けて、言語非依存なテキスト素性や、画像中の物体間の関係等より高度な素性の設計、考案を目指す。また、言語非依存である特徴を生かして、単一言語でモデルを学習させた時の、他言語、他ドメインへの適用可能性を調査することも考える。

参考文献

[1] Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 357–370, 2016.

[2] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, Vol. 12, No. Aug, pp. 2493–2537, 2011.

[3] Sobha Lalitha Devi., C.S Malarkodi Pattabhi RK Rao, and R Vijay Sundar Ram. Indian language ner annotated fire 2013 corpus (fire 2013 ner corpus). In *Named-Entity Recognition Indian Languages FIRE 2013 Evaluation Track*, 2013.

[4] Diego Esteves, Rafael Peres, Jens Lehmann, and Giulio Napolitano. Named entity recognition in twitter using images and text. *CoRR*, Vol. abs/1710.11027, , 2017.

[5] L. Liu, J. Shang, F. Xu, X. Ren, H. Gui, J. Peng, and J. Han. Empower Sequence Labeling with Task-Aware Neural Language Model. In *AAAI*, 2018.

[6] Stephen D. Mayhew, Chen-Tse Tsai, and Dan Roth. Cheap translation for cross-lingual named entity recognition. In *EMNLP*, pp. 2536–2545. Association for Computational Linguistics, 2017.

[7] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1524–1534. Association for Computational Linguistics, 2011.

[8] Erik F. Tjong Kim Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pp. 155–158. Taipei, Taiwan, 2002.

[9] Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 142–147, 2003.

[10] Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. Transfer learning for sequence tagging with hierarchical recurrent networks.

[11] 海野裕也, 那須川哲哉. 言語横断テキストマイニング. JSAI, 人工知能学会 (3A4-1), pp. 1–4, 2010.

[12] 那須川哲哉, 吉川克正, 鈴木祥子, 森田千明. 画像を介した自然言語表現の同義性判別. 人工知能学会全国大会論文集, Vol. 28, pp. 1–4, 2014.