Category2Vec: 単語・段落・カテゴリに対するベクトル分散表現

丸井 淳己 *
東京大学大学院工学系研究科
marui@ipr-ctr.t.u-tokyo.ac.jp

萩原 正人
楽天技術研究所 New York
masato.hagiwara@mail.rakuten.com

1 はじめに

従来、文や文書などの表現には、BOW (bag-of-words) 表現が広く用いられてきた。BOW では、対象を単語の集合とみなし、各次元が異なり語に対応するような疎ベクトルを用いて表現し、分類やクラスタリング等の各種タスクに用いる。しかしながら、BOW は、語順の違いを考慮できない、同義語等など関連する語であっても異なる次元に割り当てられてしまう、などの問題点がある。

そこで近年、Bengio ら [1] がニューラルネットワーク言語モデルを提案したのを皮切りに、単語や文などに対してベクトル分散表現を割り当てるモデルが提案されている。特に、Skip-gram と CBOW (continuous bag-of-words) モデル [4] では、構文解析などに頼らずコーパスから高品質のベクトルを学習することに成功している。しかしながら、これらのモデルを用いて文や文書などより大きな言語的構造をどのように表現するかは必ずしも自明ではない。

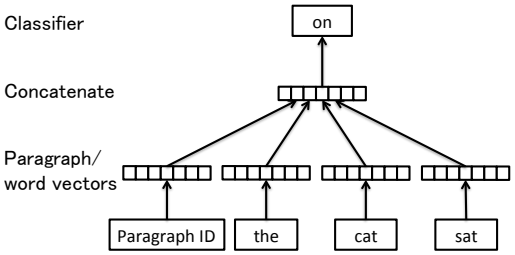
そこで、Le and Mikolov [3] は、Skip-gram と CBOW の拡張として、段落ベクトル (paragraph vector) モデルを提案した [1]。単語ベクトルと同様に、類似した意味を持つ段落に対しては類似した段落ベクトルが学習される。具体的には、PV-DM (paragraph vector with distributed memory) および PV-DBOW (paragraph vector with distributed bag of words) の２つのモデルが提案されている（図 1）。これらのモデルは、単語ベクトルや段落ベクトルから、対象の単語（群）を推定するというニューラルネットワークモデルであり、訓練時には、段落ベクトルと単語ベクトルを同時に推定する。テスト時に新しい段落が与えられた場合、単語ベクトルを固定しながら段落に対応する段落ベクトルを推定することができる。この段落ベクトルは、分類やクラスタリングなどのタスクにおいて、固定長の素性ベクトルとして使うことができる。実際、Le and Mikolov は、分類・感情極性推定タスクにおいて、段落ベクトルモデルが従来の BOW やその他のニューラルネットワーク言語モデルを上回る性能を上げることを示した。

このように、段落ベクトルモデルでは、何らかのまとまりを持つ単語列として段落や文などの言語的構造を扱うことができる。しかしながら、実際に我々が扱うテキストデータは、さらに高次の構造を持つことが多い。例えば、Web 上のテキストデータに注目した場合、ニュースサイトやショッピングサイトでは、文書や商品がそのトピック（政治、経済、スポーツ等）やジャンル（ファッション、電化製品、食品等）に従って分類されているのが一般的である。また、Wikipedia では、各記事について、そのトピックや特徴等を表すカテゴリが付与されている。例えば「アラン・チューリング」の記事には、「イングランドの数学者」「コンピュータ関連人物」「1912 年生」などのカテゴリが付与されている。以上のような、同じトピックや特徴を共有する文書の集合を本稿ではカテゴリと呼ぶ。同じカテゴリに属する文書は、何らかの構造的・意味的類似性を有しているため、その特徴をニューラルネットワークによって捉えることにより、さらに正確なベクトル分散表現を学習できると考えられる。

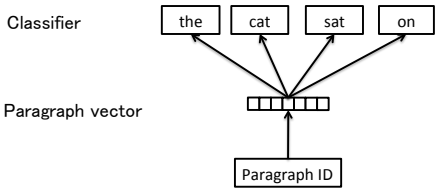
そこで本稿では、段落ベクトルモデルの拡張として、カテゴリベクトルモデルを提案する。提案するモデルでは、単語ベクトル、段落ベクトルに加え、「カテゴリベクトル」を陽に与え、対象の単語（群）を推定する。このカテゴリベクトルを用いることにより、同じカテゴリに属している文書群の類似性を捉えられると考えられる。実験では、ショッピンサイトの商品データおよび日本語 Wikipedia データを用いたカテゴリ推定のタスクにおいて、提案手法が従来手法よりも高い精度を上げることを示した。

*本研究は、楽天技術研究所 NY でのインターンシップの成果である。

[1] ここで言う段落とは、文、段落、文書など、何らかの構造的まとまりを持った単語列一般を指す言葉であり、厳密な意味での「段落」とは限らないことに注意されたい。

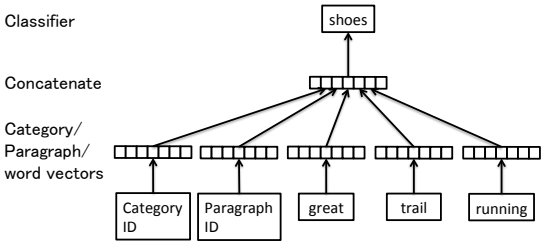


(a) PV-DM モデル

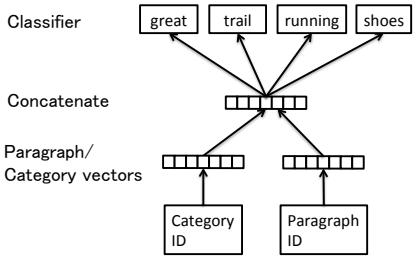


(b) PV-DBOW モデル

図 1: 段落ベクトルモデル



(a) CV-DM model



(b) CV-DBOW model

図 2: カテゴリベクトルモデル

2 単語および段落ベクトルモデル

本節では、従来手法である単語ベクトルモデルの CBOW モデルおよび Skip-gram モデル、および段落ベクトルモデルの PV-DM モデルおよび PV-DBOW モデルについて説明する。

2.1 単語ベクトルモデル

CBOW モデル CBOW モデルは、対象の単語を文脈ウインドウ内の単語ベクトルから推定する。入力となる単語ベクトルは、中間層において合計される。推定された確率を最大化するように SGD を用いて単語ベクトルを学習させることにより、単語の意味的な類似度を捉えた固定長実数値ベクトルが学習できる。訓練時の学習を効率的にするために、出力層では、階層的ソフトマックスもしくは負例サンプリングを使う。

Skip-gram モデル Skip-gram モデルでは、CBOW モデルとは逆に、文脈ウインドウ内の単語を、対象の単語から推定する。CBOW モデルと同様に、出力層では、階層的ソフトマックスもしくは負例サンプリングを使う。その他学習の詳細については、文献 [4] 等を参照されたい。

2.2 段落ベクトルモデル

PV-DM モデル 段落ベクトルモデルでは、単語ベクトルとは別に段落ベクトルを考える。PV-DM モデルは、単語ベクトルモデルの CBOW モデルに着想を得たモデルであり、対象の単語を段落ベクトルおよび文脈ウインドウ内の単語ベクトルから推定する。この際、入力の各ベクトルを中間層において結合もしくは平均する。その他の学習の詳細については CBOW モデルと同様である。訓練時の学習を効率的にするために、出力層では、階層的ソフトマックスもしくは負例サンプリングを使う。

PV-DBOW モデル 一方、PV-DBOW モデルは、単語ベクトルモデルの Skip-gram モデルに着想を得たモデルであり、段落ベクトルから文脈ウインドウ内の単語ベクトルを推定する。その他の学習の詳細については、Skip-gram モデルと同様である。

3 提案手法

段落ベクトルモデルと同様、提案手法であるカテゴリベクトルモデルは 2 種類のモデルからなる。具体的には、図 2 に示した、CV-DM (category vector with distributed memory) と CV-DBOW (category vector with distributed bag of words) の 2 つのモデルを考える。どちらのモデルにおいても、各カテゴリに対してカテゴリベクトルを対応させ、それらを学習する。

CV-DM モデル CV-DM モデルは、段落ベクトルモデルの PV-DM モデルを拡張したものであり、カテゴリベクトル、段落ベクトル、および単語ベクトルを中間層において結合もしくは平均し、対象の単語を予測する。その他の学習の詳細については、PV-DM モデルと同様である。

CV-DBOW モデル　CV-DBOW モデルは、段落ベクトルモデルの PV-DBOW モデルを拡張したものであり、カテゴリベクトルと段落ベクトルを中間層において結合し、文脈ウインドウ内の単語を予測する。訓練時の学習を効率的にするために、出力層では、階層的ソフトマックスもしくは負例サンプリングを使う。学習およびベクトル推定のアルゴリズムは従来手法と類似しているが、初期化の方法を工夫し、SGD に加え、更新式に AdaGrad [2] を用いることにより、ベクトルの収束速度を改善し、高い頻度で更新されるベクトルが発散するのを防いだ点が異なる。

テスト時には、単語ベクトルを固定し、カテゴリベクトルと段落ベクトルを求めた。この際、変数の自由度があるため両ベクトルの和しか求めることができないが、この和ベクトルをカテゴリ推定に使うことができる。この詳細については次節にて述べる。

4　実験

本節では、段落ベクトルモデルとカテゴリベクトルモデル (SGD, AdaGrad) を比較するため、楽天市場データと Wikipedia 記事を用いて評価実験を行う。評価には、カテゴリが未知の商品や記事に対して正しいカテゴリを割り当てられるか、というカテゴリ推定のタスクを用いる。

4.1　データセット

実験のためのデータセットとして、楽天市場 [2] および日本語 Wikipedia[3] の記事を用いた。本稿ではそれぞれ楽天市場データ、Wikipedia データと呼ぶ。学習およびテストに使ったデータは、楽天市場データが 760 万商品／ 15,110 商品、Wikipedia データが 29 万記事／ 9,667 記事である。

楽天市場ではカテゴリは木構造の集合として構築されており、各商品に対して 1 つのカテゴリが振られている (図 3)。商品が所属するのはすべて末端カテゴリである。カテゴリは 4 万以上存在するが、サイズやメーカーなど、商品の属性によって分類されていることも多く（例：「エアコン＞おもに 6 畳用＞パナソニック」)、本実験で扱いたいカテゴリとしては粒度が細かすぎる。カテゴリの性質を人手でラベル付けしたデータを用いて、ある商品が属性を表すカテゴリに属している場合、親を辿り最初に見つけた属性ではないカテゴリをその商品のカテゴリとする。この操作により商品が属するカテゴリの総数は 6,000 程度となる。商品のタイトルと説明文を連結し、モデルへの入力とした。

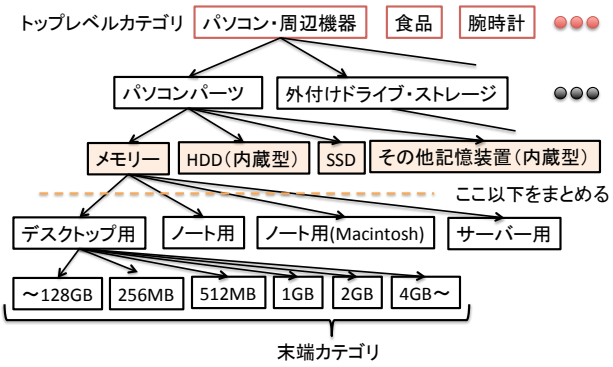


図 3: 楽天市場でのカテゴリ構造

Wikipedia データでは、1 節において述べたように、各記事について、そのトピックや特徴等を表すカテゴリが付与されている。一つの記事に複数のカテゴリが付与されていることが多い。日本語記事からカテゴリ、記事名、記事の最初の段落を抽出し、Wiki 表現などを除いて取り除いたテキストをモデルに与えた。上記いずれのデータも、日本語の分かち書きには MeCab[4] と UniDic[5] を用いている。

4.2　手順

段落ベクトルモデルの場合は、学習時に段落ベクトルおよび単語ベクトルを同時に学習し、テスト時には単語ベクトルを固定し、段落ベクトルのみを推定する。カテゴリベクトルモデルの場合は、学習時にカテゴリベクトル、段落ベクトル、単語ベクトルを同時に学習し、テスト時には単語ベクトルを固定し、カテゴリベクトルと段落ベクトルの和を推定する。得られたベクトルからカテゴリを推測するため、学習データを対象に k 近傍法を用いる。具体的には、計算されたベクトルの k 近傍（類似度が高い順から k 個の商品もしくは記事）を計算し、テストデータの正解カテゴリが、この k 個のカテゴリ候補の中に存在すれば正解、なければ不正解として精度を測る。Wikipedia データでは、1 記事にある複数のカテゴリと最大 k 個のカテゴリが重複した場合に正解、それ以外は不正解としている。また段落ベクトルは記事同士の類似性を見るため、Wikipedia データではカテゴリ推測に用いない。楽天市場データでは 30 回同じ学習データを与え、Wikipedia データでは 15 回与える。いずれのモデルも出力層の近似に階層的ソフトマックスを用いる。

[2]http://www.rakuten.co.jp/
[3]http://ja.wikipedia.org/
[4]https://code.google.com/p/mecab/
[5]http://sourceforge.jp/projects/unidic/

k-NN	段落ベクトルモデル		カテゴリベクトルモデル			
	SGD		SGD		AdaGrad	
	PV-DM	PV-DBOW	CV-DM	CV-DBOW	CV-DM	CV-DBOW
$k=1$	64.6%	67.1%	**75.2%**	71.0%	66.3%	68.5%
$k=3$	79.4%	79.1%	**87.1%**	82.6%	76.1%	78.7%

表 1: 楽天市場データを用いたカテゴリ推定性能

k-NN	カテゴリベクトルモデル			
	SGD		AdaGrad	
	CV-DM	CV-DBOW	CV-DM	CV-DBOW
$k=1$	25.1%	22.2%	33.3%	**35.2%**
$k=3$	39.5%	38.8%	48.1%	**53.1%**

表 2: Wikipedia データを用いたカテゴリ推定性能

4.3 結果

表 1 と表 2 に、楽天市場データと Wikipedia データに対する結果を示した。楽天市場データにおいて、カテゴリベクトルモデルが段落ベクトルモデルを上回っていることが分かる。カテゴリベクトルモデルの中では、楽天市場データにおいては、SGD を用いた CV-DM モデルが、Wikipedia データにおいては、AdaGrad を用いた CV-DBOW モデルの精度が最も良いことが分かる。具体的には、楽天市場データにおいては、$k=3$ 個の近傍により、約 6,000 個のカテゴリ候補の中から正解カテゴリを精度 85%で発見できることになる。ショッピングサイトの類似商品は同一のカテゴリに属することが多いので、カテゴリベクトルモデルを用いることにより、同一のカテゴリの商品のベクトルを類似したベクトルとして学習できる。学習したカテゴリベクトルは、分類やクラスタリングなどの他のタスクに用いることができる。

5 おわりに

本稿では、単語・段落・カテゴリに対してベクトルを学習するニューラルネットワークモデルである、カテゴリベクトルモデルを提案した。提案手法では、段落ベクトルおよびカテゴリベクトルの両者を用いて対象の単語（群）を推定する。ショッピンサイトの商品データを用いた実験により、提案手法が高い精度で未知商品のジャンルを推定できることを示した。カテゴリ構造が付与されたテキストにおいては、提案手法がベクトル分散表現をより正確に学習できると言える。

なお、本手法は、ショッピンサイトにおける購買履歴など他のデータに応用することもできる。例えば、あるユーザーを、そのユーザーの購入した商品のテキストの集合によって表現した場合、段落ベクトルは商品ベクトルに、カテゴリベクトルはユーザーベクトルに対応する。これにより、商品タイトルや説明文中の単語を推測するかたちで商品ベクトルとユーザーベクトルを学習することができ、各ユーザーの特徴を捉えることができる。このようにして学習されたユーザーベクトルは、商品の推薦やターゲティング広告等において有用なツールとして使える可能性がある。

謝辞

本研究に対してご助力いただいた村上浩司, Javier Artiles, 関根聡の各氏に感謝する。

参考文献

[1] Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pp. 137–186. Springer, 2006.

[2] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, Vol. 12, pp. 2121–2159, 2011.

[3] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.