トピックモデルによる分散表現の獲得手法の提案

野沢 健人 [1,a)] 若林 啓 [2,b)]

[1] 筑波大学 情報学群 知識情報・図書館学類, [2] 筑波大学 図書館情報メディア系
a) k_nzw@klis.tsukuba.ac.jp, b) kwakaba@slis.tsukuba.ac.jp

1 はじめに

単語の分散表現は，1 単語をベクトル空間の 1 点と対応づける表現方法である．単語ベクトルを機械学習の特徴量として利用できる上に，単語間の類推など様々な性質をもつことから，分散表現の獲得手法とその応用に関する研究は，盛んに行われている [1]．しかし，分散表現のベクトル空間において単語の文法的・意味論的性質がどのように反映されているかを解釈できない課題がある．ベクトルの次元と単語の関係性について解釈を容易にすることで，分散表現の応用や性能向上に寄与できると考えられる．この課題に対して Faruqui ら [2] は，獲得した分散表現を過完備な表現へ変換する手法を提案している．変換した過完備な表現は，特徴量として分散表現よりも優れ，単語ベクトルの次元と単語の対応関係の解釈が容易になることを示している．一方で提案法は，分散表現を変換せずに分散表現以外の情報を用いることで単語の次元を解釈を容易にする手法である．

本稿では，トピックモデルを用いた単語の分散表現の獲得手法を提案する．提案法は，分布仮説 [3, 4] に基づき，単語タイプごとにその周囲に出現する単語トークンの集合を 1 文書とみなし，トピックモデルを学習する．トピックモデルの学習結果から得られる単語タイプごとのトピック分布を単語ベクトルとして扱い，分散表現を獲得する．さらにトピックモデルの学習結果から得られるトピックの単語分布を用いることで，ベクトルの 1 つの次元がどのような単語によって特徴づけられているかを求めることができる．実験では，word similarity と analogy 用いて既存手法との比較実験を行い，提案手法で獲得した分散表現の性能評価を行う．さらに獲得した分散表現の次元の解釈の方法について言及する．

2 提案手法

2.1 Latent Dirichlet Allocation

提案法においてトピックモデルは，Blei らの提案している Latent Dirichlet Allocation（LDA）[5] を使用する．LDA は，文書の確率的生成モデルとして提案された教師なし機械学習の手法である．LDA では，文書には複数の潜在的な意味（トピック）が備わっていると仮定し，文書に含まれる単語にトピックを割り当てる．トピックは文書から直接観測できないため，文書内の単語の共起情報から推定する．提案手法の説明のために以下の記号を導入する．文書集合を $\mathbf{W}$，全文書数を D，語彙数を V とし，d 番目の文書を $\mathbf{w}_d$，文書 d に含まれる i 番目の単語を $w_{d,i}$ とする．トピックは 1 から K までの整数を値域とする潜在変数 z で表し，各単語 $w_{d,i}$ にトピック $z_{d,i}$ を 1 つ割り当てる．各文書 $\mathbf{w}_d$ は，各トピックの分布を表す多項分布 $\boldsymbol{\theta}_d$ をもつ．$\boldsymbol{\phi}_k$ は，トピック k における単語タイプの多項分布を表す．また，$\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ は，それぞれ $\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ のパラメータを生成する Dirichlet 分布のパラメータのベクトルである．以上より LDA のグラフィカルモデルは図 1 で表すことができる．LDA を学習することでトピック k における単語分布 $p(w|\boldsymbol{\phi}_k)$ と文書 $\mathbf{w}$ のトピック分布 $p(\boldsymbol{\theta}|\mathbf{w})$ を求めることで分散表現の獲得と次元の解釈を行う．

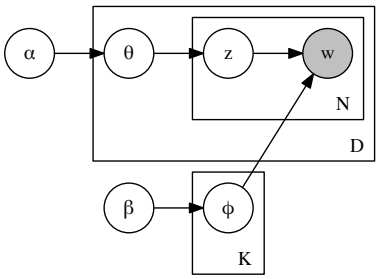

図 1: LDA のグラフィカルモデル
影の付いている単語 w のみ観測可能

2.2 LDA による分散表現の獲得手法

各文書に割り当てられた $\boldsymbol{\theta}_d$ は，K 次元の多項分布である．このため，LDA における 1 文書を単語タイプと対応づけることで，単語タイプを多項分布で表現できる．さらにこの多項分布を K 次元のベクトルとみなすことも可能である．提案法では，分布仮説に基づき単語の意味はその周囲で出現する単語によって決

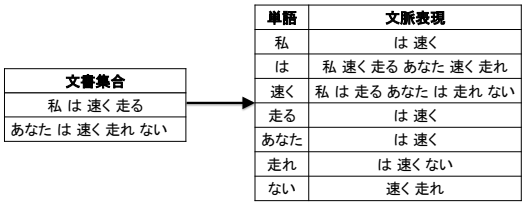

図 2: 文脈窓幅 2 における文脈表現への変換例

まるものと仮定し，単語タイプをその周囲で共起する単語を要素にもつ多重集合によって表現する．本稿では，このような多重集合で表現した単語タイプを文脈表現，対象とする単語タイプから文脈表現に含める単語トークンまでの距離を文脈窓幅と呼ぶ．図 2 に文脈窓幅を 2 としたときの文書から文脈表現への変換例を示す．変換後の 1 つの文脈表現を d 番目の文書 $\mathbf{w}_d$ とし，LDA を適用することで $\mathbf{w}_d$ のトピック分布 $\boldsymbol{\theta}_d$ を獲得し，これを分散表現とみなす．$\boldsymbol{\theta}_d$ は確率分布であることから，単語間の非類似度は Jensen–Shannon ダイバージェンス（JS ダイバージェンス）によって定義できる．

提案法では，それぞれの単語タイプを周囲の単語トークンの多重集合で表現するため，文書数と文書長ともに大きくなりやすい傾向がある．このため，学習アルゴリズムに Mimno ら [6] の提案する確率的勾配法に基づくアルゴリズムである確率的変分ベイズ法を用いる．確率的変分ベイズ法による学習アルゴリズムは，LDA の 1 回の反復計算において全文書データから十分統計量の期待値を求めるのではなく，サンプリングした文書データだけを用いて十分統計量の期待値の計算を行うことで，文書数に対してスケールせずに高速に学習を行う．このとき，サンプリングする文書数はミニバッチと呼ばれる所与の値である．

提案法では，単語タイプごとに文脈表現を生成することから，もとの文書集合に対して文脈窓幅の約 2 倍のメモリ空間を必要とするために，大規模な文書データに対して実行困難になる．これに対して転置索引を作成し，確率的変分ベイズ法でサンプリングした単語タイプについて，転置索引を用いて文脈表現を構築して空間計算量を抑える．また，文書長は (文脈窓幅) × 2 × (出現頻度) となり，大規模な文書データでは文書長が大きくなりやすい．文書長の増加を抑えるために，確率的変分ベイズ法でサンプリングした単語の転置索引から一定数の転置索引を用いて得られる文脈表現を LDA の学習に使用する．

大規模な文書集合においてストップワードのような高頻出語の影響を小さくするために，Mikolov らのモデルで行っている subsampling[7] を提案法においても適用する．subsampling は，高頻度語をその頻度に応じて確率的に文書中から取り除くことで，高頻度語彙による影響を小さくする効果がある．各単語トークンに対して式（1）で計算した確率値をもとに文書から単語トークンを削除する．ただし，t は閾値，$f(w_i)$ は単語 w_i の相対度数である．subsampling を行ってから転置索引を生成するため，本来の文脈窓幅の中で出現しない単語が文脈表現の要素となる．

$$p(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}} \tag{1}$$

3 比較実験と実験結果

3.1 データセット

本実験では，2010 年 11 月の Wikipedia の記事データ [1] を使用する．本文に対してアルファベットと数値以外の文字の削除，数字 1 桁ごとにそれぞれ対応する英語単語に変換，アルファベットを全て小文字に変換した上で実験を円滑に行うために，全記事データの中から一様ランダムに抽出した 3 割の記事を使用する．ただし，出現回数 100 回未満の単語は，低頻度語として学習には使用しない．

3.2 比較手法

比較実験では，Mikolov ら [8] の提案する CBoW と Skip–gram を比較対象とし，両モデルのオープンソース実装である word2vec[2] を使用する．今回の実験で調節するパラメータを表 1 に示す．dim は分散表現の次元，win は文脈窓幅，neg は negative sampling で使用する単語数を表す．それぞれのモデルごとにパラメータを組み合わせた全 20 通りで学習を行う．また，subsampling の閾値 t は 10^{-5} とする．

3.3 提案手法のパラメータ

提案手法で調節するパラメータを表 2 に示す．dim, win は比較手法と同様に分散表現の次元，文脈窓幅のパラメータとし，opt は，Dirichlet 分布のパラメータ $\boldsymbol{\alpha}$ の更新の有無，inv は，1 つ単語タイプで使用する転置索引の割合を表す．$dim = 1000$ において $inv = 0.5, 1.0$ に設定した場合に学習にかかる時間が増加するために $inv = 0.2, win = 2$ の場合だけで学習する．このため提案法については全 26 通りの分散表現を評価に使用する．LDA の学習における反復回数は 300 回，subsampling の t は 10^{-5}，確率的変分ベイズ法におけるミニバッチは 2000 とする．

[1] https://dumps.wikimedia.org/archive/2010/2010-11/enwiki/20101011/enwiki-20101011-pages-articles.xml.bz2
[2] https://code.google.com/p/word2vec/

表 1: CBoW と Skip–gram のパラメータ

パラメータ名	値域
dim	50, 200, 500, 1000, 2000
win	2, 5
neg	0, 15

表 2: 提案手法のパラメータ

パラメータ名	値域
dim	100, 200, (1000)
win	1, 2
opt	$true$, $false$
inv	0.2, 0.5, 1.0

3.4 評価方法

比較実験の評価指標には，分散表現の評価で広く用いられている word similarity と analogy を使用する．

3.4.1 word similarity

word similarity による評価では，ws353 (WS) [9]，ws353_similarity (WSS)[10]，ws353_relatedness (WSR) [11]，bruni_men (MEN) [12]，radinsky_mturk (TM) [13]，luong_rare (RW) [14] の 6 つのデータセットを使用する．これらのデータセットでは，単語のペアとその類似度が定義されている．文書表現を用いて単語ペアの類似度を求め，データセットの類似度との順位相関係数によって分散表現の性能を評価する．各単語ペアのうち一方でも分散表現に含まれない単語ならば，順位相関係数の評価には用いないものとする．本手法では，JS ダイバージェンスを用いて 2 単語間の非類似度を計算できるため，この場合は順位を逆にした上で順位相関係数を求める．

3.4.2 analogy

analogy による評価では，Google[8] と MSR[15] の 2 つのデータセットを使用する．データセットに含まれる 4 単語のうち 3 つの単語を用いて類推の演算を行う．類推の演算結果と演算に使用しなかった単語との一致率で評価を行う．類推の演算方法は，Levy ら [16] と同様に 3CosAdd と 3CosMul の 2 つを使用する．word similarity と同様に分散表現に含まれない単語がデータに含まれる場合は，評価に使用しない．

3.5 実験結果

3.5.1 word similarity の評価結果

各データセットごとに求めた順位相関係数の最大値を表 3 に示す．提案手法における評価結果は，類似度計算にコサイン類似度と JS ダイバージェンスを用いた場合の 2 つで順位相関係数を算出している．CBoW が 2 つ，Skip–gram が 4 つのデータセットに対してもっとも順位相関係数が高い．提案手法についてはすべてのデータセットで既存手法を下回っている．また，提案手法においてコサイン類似度よりも JS ダイバージェンスを用いたほうが順位相関係数は高いことから，word similarity では提案手法で得られた単語をベクトルではなく，確率分布として扱ったほうがよいと考えられる．

3.5.2 analogy の評価結果

各モデルごとに 2 つのデータセットの一致率の最大値を表 3 に示す．類推の演算手法である 3CosAdd と 3CosMul をそれぞれ add，mul と表記している．提案法では word similarity と異なり，確率分布として扱った場合に加算減算を定義できないことから表 3 には JS ダイバージェンスの結果を記していない．

提案法の評価結果は，word similary の結果と同様に CBoW と Skip–gram と比較して低い一致率となっており，単語ベクトルとして十分な性能が出せていない．このことから提案法は多項分布をベクトルとみなすことで分散表現の獲得を行ったが，多項分布をベクトルと扱っても，ベクトル空間が単語の意味と対応していないために，比較手法よりも十分な性能が出せないと考えられる．

4 提案手法における分散表現の次元解釈

提案手法において比較実験の評価値が高かったパラメータの組み合わせである $dim = 1000, win = 2, opt = true, inv = 0.2$ のときの分散表現を用いて次元の解釈の可能性について検討する．提案手法は LDA に基づくモデルであるため，単語タイプ d のトピック分布 $\boldsymbol{\theta}_d$ の値が高いほど，単語タイプ d にそのトピックが割り当てられている．このとき，単語タイプ $\mathbf{w}_d$ のトピック分布は，式（2）で与えられる．さらに各トピック k は，各単語の多項分布をもつために，トピック k がどの単語によって特徴付けられているのかを確率値とともに獲得できる．トピック k における単語分布は，式（3）で与えられる．

$$Multi(\theta|\mathbf{w}_d) \tag{2}$$

$$Multi(w|\boldsymbol{\phi}_k) \tag{3}$$

提案手法の適用結果から単語 *python* を例に取り上げる．$\boldsymbol{\theta}_{python}$ の確率値の高い上位 3 トピック，トピックに対応する α の値，トピック k における $\boldsymbol{\phi}_k$ の上位 5 単語を表 4 に示す．表 4 において，トピック 880 は蛇と *Monty Python*，トピック 145 はプログラミングに関する単語，トピック 732 はストップワードが含まれている．トピック 732 を除いたトピックでは *python* のもつ語義に近いトピックが抽出できている．ここでトピック 732 は，α の値が高いことからどの文脈表現に

表 3: word similarity と analogy の評価結果．最大値をボールド体で表記．

モデル	WS	WSS	WSR	MEN	MT	RW	Google add / mul	MSR add / mul
CBoW	**0.703**	0.770	0.656	0.726	0.655	**0.484**	0.546 / 0.553	0.547 / **0.585**
Skip–gram	0.702	**0.771**	**0.683**	**0.739**	**0.660**	0.461	0.624 / **0.642**	0.502 / 0.541
提案手法 + コサイン類似度	0.408	0.506	0.364	0.464	0.465	0.265	0.224 / 0.226	0.053 / 0.042
提案手法 + JS ダイバージェンス	0.437	0.563	0.377	0.508	0.516	0.266	–/–	–/–

表 4: $\boldsymbol{\theta}_{python}$ の上位 3 トピック

トピック k	θ_k	α_k	words
880	0.187	0.019	circus, snake, monty, cobra, sketch
145	0.116	0.071	archive, software, web, programming, database
732	0.074	1.127	that, i, it, you, be

も含まれやすいトピックと考えられる．このため，α に対して閾値を設けることでどの文脈表現にも現れやすいトピックを取り除くことができる．このように従来の分散表現で困難であった次元とその対応関係について，トピックを用いることで解釈可能な形で表現できる．

5 結論

本研究では，トピックモデルによる単語の分散表現の獲得手法を提案した．提案手法では，分布仮説に基づき，単語の周囲に出現する単語によって単語タイプを特徴付けた上でトピックモデルを学習し，単語の多項分布とベクトル表現を得た．既存手法の CBoW と Skip–gram の両モデルと比較実験を行い，word similarity の評価では 0.1 ～ 0.3，analogy の評価では，約 0.5 劣る結果となった．このことから確率分布をベクトルとして扱っても，十分な類推を行えないことために，3CosAdd や 3CosMul に相当する類推の演算を確率分布を用いた場合にどのように表現するかが今後の課題である．また，単語の分散表現の課題であるベクトル空間の次元の解釈について，単語に付与された確率値の高いトピックとそのトピックの単語分布を用いることで分散表現の次元をトピックとして解釈できる表現であると結論づける．

謝辞

本研究の一部は，JSPS 科研費（課題番号 25280110,25540159）および筑波大学図書館情報メディア系プロジェクト研究（Research Projects of Facalty of Library, Information and Media Science）の助成によって行われた．

参考文献

[1] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proc. ACL*, 2014.
[2] Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. Sparse overcomplete word vector representations. In *Proc. ACL*, 2015.
[3] Zellig Harris. Distributional structure. *Word*, 1954.
[4] J. R. Firth. A Synopsis of Linguistic Theory, 1933-1955. In *Studies in Linguistic Analysis*. 1957.
[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *JMLR*, 2003.
[6] David M. Mimno, Matthew D. Hoffman, and David M. Blei. Sparse stochastic inference for latent dirichlet allocation. In *Proc. ICML*, 2012.
[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, 2013.
[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proc. ICLR*, 2013.
[9] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *TOIS*, 2002.
[10] Torsten Zesch, Christof Müller, and Iryna Gurevych. Using wiktionary for computing semantic relatedness. In *Proc. NCAI*, 2008.
[11] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proc. NAACL-HLT*, 2009.
[12] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. Distributional semantics in technicolor. In *Proc. ACL*, 2012.
[13] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proc. WWW*, 2011.
[14] Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. In *Proc. CoNLL*, 2013.
[15] Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proc. NAACL-HLT*, 2013.
[16] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proc. CoNLL*, 2014.