Vol.2019-DPS-179 No.21
Vol.2019-MBL-91 No.21
Vol.2019-ITS-77 No.21
2019/5/23

word2vec による類似語彙を用いたトピック抽出

伊藤晶広[†1]　白井里奈[†1]　上原宏[†1]

概要：本研究では，LDA に，特定領域の文書から得られた語彙間の類似情報を導入することで，その領域に関連するトピック抽出を試みる．具体的には，居酒屋に関する文書に word2vec を適用して獲得した語彙間の類似情報を，LDA による旅行口コミ文書のトピック分類に導入し，居酒屋に関連するトピックの抽出性能についての評価を行う．分析の結果，居酒屋に関連するトピックが抽出され，また，そのトピック語彙の文書中での意味的一貫性が認められ，提案手法の有効性が得られた．

キーワード：LDA，潜在ディリクレ配分法，word2vec，トピックモデル，事前知識

1. はじめに

LDA（潜在的ディリクレ配分法）では，文書毎のトピック分布とトピック毎の語彙分布を推定する．また，これらの分布から文書に出現する各語彙のトピック（以下，語彙トピック）を求めることができる．通常，1 つの文は，一貫した意味的表現の単位であることから，同一文中の語彙トピックは同一になると考えられる．しかし，LDA では語彙分布の推定にあたり，各文での語彙間の系列性を考慮しないため，同一文中の語彙に対してしばしば異なる語彙トピックが推定されることになる．本研究では，LDA による語彙トピック推定に，語彙間の関連性に関する事前知識を反映させることにより，語彙トピックの意味的一貫性の向上を試みる．

2. 先行研究

Word2vec[1]は語彙間の意味的類似性を当該語彙の周辺に出現する語彙の出現パターンから推定するものである．これによって得られる語彙間の関連性を LDA の事前知識として用いる研究が見られる[2,3]．[2]では，word2vec にもとづく語彙間の類似情報を用いて，意味的に類似した語彙が LDA によって同一トピックに所属する確率を向上する方法を提案している．[3]では，通常の LDA ではトピック分類が困難な短文かつ語彙頻度が小さい文書を対象として，word2vec からの語彙の類似性にもとづく分類性能の向上を図っている．本研究では，特定の話題（以下，特定ドメイン）を扱う文書に word2vec を適用することで，そのドメインに語彙間の類似性を獲得する．これを LDA の事前知識に用いることにより，特定ドメインを強く反映したトピック検出を試みるとともに，文中におけるそのトピックに関する語彙トピックの一貫性の向上を実現する．

3. LDA による語彙トピックの推定

LDA では，文書毎のトピック確率パラメータ（図 1 θ）と，トピック毎の語彙確率パラメータ（図 1 ϕ）を推定する．また，文書中に出現する語彙のトピックは，θ, ϕ に基づく確率を最大化するトピックとして求められる（図 1 z）．

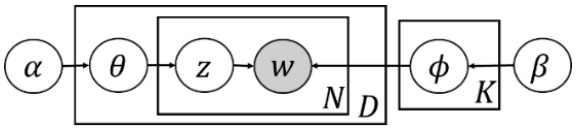

図 1　LDA のグラフィカルモデル

以下の例は，旅行サイトの口コミからの語彙トピックとする．[ ]内の番号が語彙トピック，0 は文化，1 は飲食に関するトピックであるとする．

「京都にある有名なお寺[0]でお茶[1]を頂いてきました．きれいな茶器[0]で立てた抹茶[0]はとてもおいしかったです．日本文化[0]ですね．」

上記の文書は，全体として日本文化に関するもので，文中の出現語彙も意味的に一貫して日本文化に関するものである．一方，LDA によって算出された語彙トピックは，必ずしもこの文脈と整合しない．‘お茶’は，この文書の意味上は茶道，すなわち日本文化に関する語彙であるが，飲食の語彙トピックが付与されている．LDA では，トピック分類にあたって語彙間の関連性を考慮しないため，語彙トピックが文脈と整合しないことがしばしば起きる．

4. 提案手法

本研究では特定の話題の類似語彙群が同一のトピックである確率が高くなるような事前知識を LDA に導入する．

例えば，茶道に関する文書に word2vec を適用することで‘寺’，‘茶器’と‘お茶’の語彙間の類似情報を獲得できたとする．文書語彙列を LDA に入力する際，この類似情報を反映した語彙辞書を参照することで，特定ドメイン（ここでは茶道）で高い類似性を有する語彙が同一トピックとして分類される確率を高める．具体的には，‘寺’，‘茶器’，‘お茶’に同一 ID を付与して辞書に登録する．この手法によると，‘寺’，‘茶器’，‘お茶’のトピック毎の語彙確率(ϕ)は同一になる．この手法を用いると‘寺’，‘茶器’，‘お茶’が共起する文書のように，全体として日本文化に関するトピックを表す文書（日本文化に関する θ が大きい）におい

†1 秋田県立大学　システム科学技術学部

Vol.2019-DPS-179 No.21
Vol.2019-MBL-91 No.21
Vol.2019-ITS-77 No.21
2019/5/23

て，各語彙のφが同一となるため，文中の語彙のトピック(z)も同一となることが期待される．

5. 実験データ

提案手法の有効性を確認するため，旅行口コミサイト‘fortravel’における京都府エリアの口コミ 5000 件をトピック分類の対象とし，グルメサイト‘食べログ’の居酒屋に関する口コミを事前知識として用いた．旅行サイトの口コミには，名所旧跡，宿泊，グルメ等，多様なトピックが混在している．居酒屋に関する事前知識を導入することによって，それに関するトピックが抽出され，当該文書中の語彙に対して，一貫して居酒屋に関する語彙トピックが付与されるかどうかを評価する．

6. 評価方法

トピックの意味的一貫性が高い場合，同一文中の語彙トピックの純度が大きくなる．そこで，評価対象のトピック$\mathrm{k_{target}}$の文中での純度を情報エントロピーを用いて測定することにより，意味的一貫性を評価する．

$\mathrm{k=k_{target}}$であるような文における情報エントロピーは以下の式で表される．ここで，$\mathrm{k=k_{target}}$であるような文とは，K 個のトピックのうちで$\mathrm{k_{target}}$の頻度が最も大きい文を意味する．

$$\mathrm{I_H(k_{target})} = \frac{\sum_{i=1}^{D}\left(-\sum_{k=1}^{K} \mathrm{p(k|k_{target})}\log_2 \mathrm{p(k|k_{target})}\right)}{D} \quad (1)$$

$\mathrm{p(k|k_{target})}$は $\mathrm{k=k_{target}}$の文における k の確率，D は$\mathrm{k=k_{target}}$の文の総数である．すなわち，(1)は $\mathrm{k=k_{target}}$である文の情報エントロピーの平均値を意味する．

事前知識が文中での語彙トピックの一貫性を高めるとすると，事前知識を反映したトピックを表象する文の情報エントロピーは，他のトピックのそれと比較して小さくなることが期待される．

7. 結果

従来トピックに所属する語彙の意味的な一貫性を評価する指標として coherence が用いられてきた．そこで，事前知識を導入した LDA,通常の LDA それぞれに対し coherence を求めたところ，前者が 56.8，後者が 48.6 となり，事前知識を導入した LDA の coherence が相対的に高い値を示した．

表 1 にトピック毎の情報エントロピーを示す．事前知識を導入した LDA のトピックのうち，特に事前知識を反映したトピック 3 の情報エントロピーが他のトピックのそれに比較して低い値を示していた．クチコミの文を以下に例示する．(a)は事前知識を導入したもの，(b)は通常の LDA によるものである．[ ]内の番号は語彙トピック番号である．

(a)「特筆すべきはメイン[3]の魚[3]だけではなく，炊き込みご飯[3]，貝汁も美味しいし，自家製のちりめん山椒はご飯[3]のおかず[3]にも酒[3]のアテにもなります．」
(b)「特筆すべきはメイン[1]の魚[1]だけではなく，炊き込みご飯[1]，貝汁も美味しいし，自家製のちりめん山椒はご飯[1]のおかず[1]にも酒[8]のアテにもなります．」

上記の通り，事前知識を導入した LDA では，1 つの文に対して一貫して同じトピック(3)が割り当てられているのに対して，通常の LDA では‘酒’が他の単語とは異なるトピックに割り当てられるという結果となった．

このように，事前知識を導入した LDA を用いることで通常の LDA と比べ，情報エントロピーが低く，1 つの文に対して一貫したトピックが割り当てられている．

表 1　各トピックの情報エントロピー

	Information entropy of each topic										Average
	0	1	2	3	4	5	6	7	8	9	
事前知識を導入したLDA	0.64	0.38	0.43	0.39	0.34	0.43	0.43	0.45	0.37	0.29	0.42
通常のLDA	0.61	0.40	0.46	0.44	0.49	0.51	0.48	0.48	0.36	0.31	0.45

8. 結論

本研究では話題の混在する文書から特定のトピックを検出するため，word2vec によって獲得した語彙間の類似情報を LDA に事前知識として適用する方法の提案，評価を行った．その結果，以下を含むいくつかの評価結果によって有効性が支持された．

-分類された各トピックの一貫性を示す評価指標は通常の LDA の結果より高い数値を示していた．
- 事前知識を導入することで情報エントロピーは減少した．
- 事前知識を導入した LDA を用いることで 1 つの文に対して一貫したトピックが割り当てられた．

今回使用したデータセットは旅行者のクチコミを対象にしていたが，今後は様々なデータに対してアプローチをかけ，一般的な条件下での性能の評価をすることを課題とする．

参考文献

[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013).
[2] Li, C., Lu, Y., Wu, J., Zhang, Y., Xia, Z., Wang, T., Yu, D., Chen, X., Liu, P., Guo, J.: Lda meets word2vec: A novel model for academic abstract clustering. In: Companion Proceedings of the The Web Conference 2018. pp. 1699–1706. International World Wide Web Conferences Steering Committee (2018).
[3] Moody, C.E.: Mixing dirichlet topic models and word embeddings to make lda2vec. arXiv preprint arXiv:1605.02019 (2016).