単語トピック特定性を用いた文脈単語の重み付け

中山雄貴　ホー・ツー・バオ
北陸先端科学技術大学院大学知識科学研究科
{s1450021, bao}@jaist.ac.jp

1.はじめに

分布的仮説とはある単語とその他の単語の意味的な関係はそれらの単語の文脈においての出現パターンの類似性によって推定できるという仮説である。その仮説をもとにある単語とその前後 N 単語内に出現する単語との共起頻度を求め、それらをベクトルに記録し、文脈(一般的には共起単語)を次元とした高次元空間においてのベクトルとして単語の意味を表現するモデルを単語空間モデルという[1]。

このモデルによって生成された単語の共起頻度に基づくベクトルはノイズを非常に含んだものとなってしまうため、これに対処するため、一般的にはベクトルの要素に対する重み付けを行う[2]。単語の意味ベクトルに用いられる有名な重み付け手法には PMI (Point-wise Mutual Information)や T-test といった共起に基づく重み付けがある。これらはある単語ともう一方の単語が共起する期待頻度に比べてどれぐらい実際の共起頻度が高いかを推定したりすることによって重みを決定する評価指標である。しかし、共起に基づく重み付けは共起性を重視するあまり、「最近」や「いつも」といった意味比較に役立たない文脈単語であってもその共起頻度が期待頻度を大きく超えていれば、より大きな重み付けをしてしまうという問題がある。

そこで、我々は単語の意味を形成するもう一つの側面として、単語トピック特定性を提案する。単語トピック特定性とは単語がどのくらい特定のトピックに偏って分布しているかを評価する指標である。例えば、具体的な単語であれば、少数のトピックにおいてでしか出現しないが、抽象的な単語であれば、より多くのトピックに満遍なく出現するので、その指標において具体的な単語であればより大きな値を、どの単語にとっても一般的な単語にはより小さな値を与えることができる。私はその指標を利用し、より具体性をもち、単語の意味の比較に役立つ文脈単語により大きな重み付けを行えるようにした。その指標と共起に基づく重みを融合させることによって、単語同士の共起性と共起する単語のトピック特有性という 2 つの観点から単語のベクトル要素の重み付けを行うことができるようになった。既存の手法と提案手法によって生成されたそれぞれの単語の意味ベクトルの質を比較するため、単語のペアの類似度スコアが人によって評価されたデータセットとそのデータセットに存在する単語ペア間のベクトル類似度に対する Spearman の順位相関係数を求めた。その結果、我々が提案した重み付け手法は既存の重み付け手法に比べて、単語の意味比較において質の良い単語の意味ベクトルを生成するということを示した。

2.関連研究

本節では単語意味ベクトルの既存の共起性に基づいた重み付け手法の代表的な 2 つを紹介した後、その問題点について考察する。

2.1　重み付け手法

共起性に基づいた重み付け手法は、それぞれの単語が独立に出現すると仮定したときに期待される共起頻度に比べて、実際の単語同士の共起がどれくらい頻繁なのかを計算するアプローチに基づく手法である。

最も有名な手法の一つに Pointwise Mutual Information (PMI)がある。PMI は以下のように定義される。

$$\mathrm{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

この手法において、重要な性質は単語同士の偶然の共起はあまり重視されないことである。また、PMI の負の値は信頼できないため、一般的には負の PMI をすべてゼロに置き換える Positive Pointwise Mutual Information (PPMI)が使われる[1]。

$$\mathrm{PPMI}(x,y) = \begin{cases} PMI(x,y) & if\ PMI(x,y) > 0 \\ 0 & otherwise \end{cases} \quad (2)$$

もう一つの有名な手法に T-test がある。PMI は機能語などのような高頻出単語を取り除く能力は優れているのだが、低頻度単語に対しては上手く働かな

いという問題点があった。つまり、低頻度の共起情報を重視しすぎてしまう傾向がある。T-test は単語自身の頻度を考慮することによって、そういった問題を解決する。T-test は PMI のように共起の強さを測るのではなく、有意性を評価する。T-test は、以下のように定義される[3]。

$$\mathrm{T\text{-}Test}(x,y) = \frac{p(x,y) - p(x)p(y)}{\sqrt{p(x)p(y)}} \quad (3)$$

engineer	0.983976
noted	0.962621
worked	0.960127
asked	0.956972
philosopher	0.955541
prize	0.946809
nuclear	0.944733
influential	0.943037
unable	0.939655
prominent	0.937688
tried	0.937407
question	0.936846
professor	0.934604
honor	0.931251
astronomer	0.928744
artist	0.926613
joseph	0.926243
visited	0.926243
queen	0.925134
detailed	0.924579
believe	0.921537
modified	0.920157
achieved	0.919606
intellectual	0.917954
swedish	0.917404

(a)PPMI

engineer	0.027593
noted	0.015306
philosopher	0.013982
worked	0.012857
asked	0.011501
artist	0.010562
influential	0.00988
prominent	0.009484
question	0.009434
famous	0.009352
astronomer	0.008788
believe	0.008301
study	0.008266
prize	0.008199
agricultural	0.007488
numerous	0.007236
unable	0.006814
tried	0.006636
professor	0.006602
honor	0.006316
queen	0.0062
visited	0.006062
modified	0.005833
intellectual	0.005786
teacher	0.005762

(b)T-test

図 1. "scientist" の共起に基づく重み付け手法によって与えられた単語の重み

3.2 共起頻度を用いる重み付けの問題点

PMI や T-Test といった共起頻度を用いる重み付け手法は単語と単語の共起のしやすさを考慮することによって文脈単語の重要度を計算する。そのため、ある単語がある文脈単語と共起しやすければ、単語の意味分別のために重要な文脈単語としてより大きな重みを与えてしまう。図 1 は"scientist"という単語の PPMI と T-test によって与えられた重みの例である。PPMI や T-test において、"space"や"material"といった、"scientist"とより関係していそうな単語であってもそれらの単語よりも"whether"や"recently"といった単語により大きな重み付けをしてしまっていることが分かる。文脈単語の最適な重み付けが共起情報だけでは決定できないのである。

3.提案手法

単語には特定のジャンルでしか用いられないものもあれば、幅広いジャンルや文書において用いられる単語もある。前者は単語の意味分別の際に役立つ文脈単語となるが、後者の単語は特定のジャンルに存在するのではなく、大抵のジャンルに存在するため、単語の意味を分別する際の有力な文脈情報となり得ず、ほとんど役に立たない。また、LDA モデルにおいて、抽象的な単語はどの単語に対しても共起しやすい結果、ほとんどすべてのトピックに出現する傾向がある。我々はこの性質に着目し、単語のトピックの特定性から計算した文脈単語の有効性の指標を単語トピック特定性(Word Topic Specificity, WTS)を定義した。手法のフレームワークは図 2 のようになる。

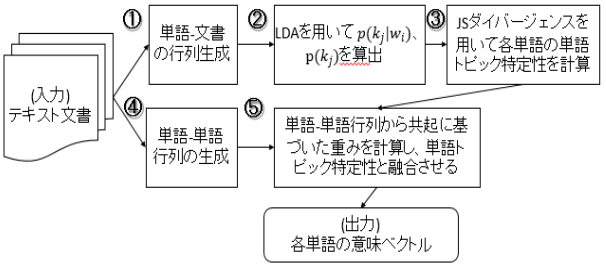

図 2.提案手法のフレームワーク

3.1 トピックモデル

潜在的ディリクレ配分法(Latent Dirichlet Allocation)は、最初の生成的トピックモデルである[4]。LDA はそれぞれの文書を潜在トピックの確率モデルとして表現することができ、すべての文書において共通の Dirichlet 事前分布トピック分布を共有していることを仮定している。LDA モデルにおけるそれぞれの潜在トピックは単語に関する確率分布として表現され、トピックの単語分布は共通の Dirichlet 事前分布を共有している。LDA モデルの生成過程は次のように記述される。

K 次元のパラメータベクトルαを含む Dirichlet 分布を与える
V 次元のパラメータベクトルβを含む Dirichlet 分布を与える
for トピック 1 からトピック K
　　パラメータβによる Dirichlet 分布からトピック k に対しての多項分布Φ_kを決定する
for 文書 1 から文書 D
　　パラメータαによる Dirichlet 分布から文書 d に対しての多項分布θ_dを決定する
　　for 文書における単語
　　θ_dからトピック z_n を決定
　　Φ_{zn}から単語 w_n を決定

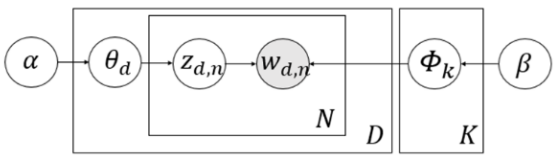

図 3.LDA のグラフィカルモデル

LDA において、トピック集合 z を推定する様々なアルゴリズムがあるが、本研究においては Gibbs サン

プリングを用いる[5]。Gibbs サンプリングは実装が簡単で、記憶容量をあまり必要としないからである。

また、本研究において最適なトピック数の選定も行う。最適なトピックの数を選定する方法はいくつか提案されているが、本研究においては他手法よりもより強健な挙動をみせる Arun によって提案された手法を用いる[6]。この手法では、ハイパラメータが一定の設定においてトピック数 K を変化させて LDA の処理を行い、それぞれの LDA の出力であるトピック-単語行列と文書-トピック行列から生成される分布を観察することによって、最適なトピック数 K^*を決定する。

3.2 単語トピック特定性

我々は、最も曖昧な単語はすべてのトピックに対して一様に分布すると仮定し、一様に分布すると仮定した際の単語のトピックに対する条件確率を次のように定義した。

$$\mathrm{P}(w_{abstract}|k_i) = \mathrm{P}(k_i)\ (\forall i \in \{1,2,\dots,K\})\ (4)$$

なお$\sum_i P(k_i) = 1$。上記の条件確率は前述の LDA によって得ることができる。上式における確率分布に類似した分布をもつ単語がより抽象的あるいは機能的な意味を持つ単語であると仮定した。この分布間の違いを求めるために、我々は Jensen-Shannon ダイバージェンスを使用した。Jensen-Shannon ダイバージェンスは以下のように定義することができる[7]。

Kullback-Leibler ダイバージェンス

$$\mathrm{KLD(P||Q)} = \sum_{x\in\chi} P(x) log \frac{P(x)}{Q(x)}\ (5)$$

Jensen-Shannon ダイバージェンス

$$\mathrm{JSD(P||Q)} = \frac{1}{2}KLD(P||\frac{1}{2}P + \frac{1}{2}Q) + \frac{1}{2}KLD(Q|\left|\frac{1}{2}P + \frac{1}{2}Q\right)\ (6)$$

KLD は非負であるので、JSD も非負である。JSD の最大値はどの基底の対数を用いるかによって変わる。自然対数、つまり、基底が e の対数を用いると、0≤JSD(P||Q)≤log2 になる。JSD は基底 2 の対数を使うと 0≤JSD(P||Q)≤1 となる。

単語トピック特定性を求めるために、トピック j の単語 i に対する条件確率を P、典型的な機能的な意味を持つ単語のトピックに対する条件確率を Q とすると、それぞれ以下のような式になる。

$$P = p(k_j|w_i)(\forall j \in \{1,2,\dots,K\})$$

$$Q = \mathrm{p}(k_j)\ (\forall j \in \{1,2,\dots,K\})$$

上式で示される確率分布同士を比較するために前節で言及した Jensen-Shannon ダイバージェンスを用いる。JSD の値は、より典型的な機能的な意味を持つ単語の分布と異なっていればいるほど、より大きな値をとり、近ければ近いほどより小さな値をとる(図 4)。

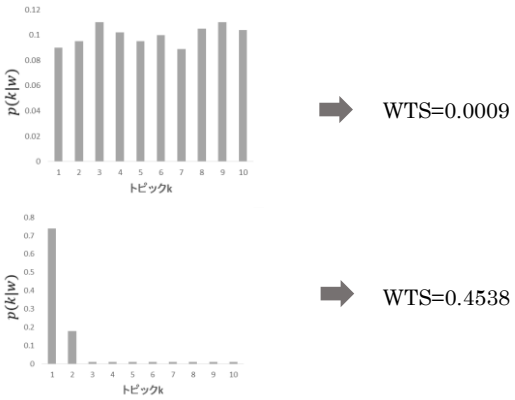

図 4.単語トピック特定性の説明

3.3 重みの融合

本研究においては単語のトピック特定性だけに重点を置くのではなく、単語同士の共起性をも重視した重み付けを行いたい。以下の式によって共起に基づく重みと単語トピック特定性に基づく重みを融合した。

$$weight(w_T, w_C) = WC(w_T, w_C) \times \left(\frac{2}{1 + e^{-\alpha WTS(w_C)}} - 1\right)\ (7)$$

上式において WC は共起に基づいた重み、WTS は単語トピック特定性に基づいた重みである。w_T と, w_C はそれぞれ目標単語、文脈単語を示す。なお本研究において、$\alpha = 2$とした。

5.評価

提案手法の単語ベクトルの質への効果を評価するために人手によってスコア付けされた単語類似度のデータセット、WordSim-353[8]、MEN[9]、MTURK[10]、Rare Word Similarity (RW)[11]を用いた。これらのデータセットはそれぞれの単語ペアにおける類似度が人間によって評価されている。例えば、WordSim-353 データセットにおいて、”computer”と”keyboard”の類似度は 7.62 である。 また、RW は他のデータセットと違い、主に珍しい単語の単語ペアに対する類似度スコアを評価している。我々はそれぞれの単語ペアの単語ベクトルにおいてのコサイン類似度を計算し、その類似度と人手のスコアを、Spearman の順位相関係数を用いることによって比較した。なお単語の意味ベクトルを学習する際のデータは 2010 年の 4 月における Wikpedia のデータから抽出した 10 万文書

を用いた。Wikipedia のデータにおいて含まれる特殊文字や記号は前処理において除去した。

表 1 に各重み付け手法において生成されたベクトル間のコサイン類似度と評価データセット間の Spearman の順位相関係数を表わす。

表 1. 各手法の Spearman の順位相関係数

	WordSim-353	MEN	MTURK	RW
Freq	0.456	0.572	0.582	0.348
PMI	0.513	0.656	0.693	0.477
Ttest	0.608	0.741	0.746	0.522
PMI+WTS	0.518	0.678	0.705	0.483
T-test+WTS	0.610	0.742	0.741	0.525

すべての評価データセットにおいて PPMI に比べ、我々が提案した重み付け手法が上回っていることが示された。しかし、WordSim-353 においては、他の評価セットと比較して、あまり PPMI の重み付けがあまり改善されていないという結果となった。これは、WordSim-353 は絶対スコアを注釈付けしていったのに対し、MEN や MTURK は 2 つのペアを比較する形で注釈付けが行われため、後者のほうがより正確な単語間類似度を示しているためだと考えられる。T-test においてはどのデータセットにおいてもあまり変わらなかった。これは T-test が PPMI に比べ、単語トピック特定性において小さい値をもつ文脈単語に重みを与えない性質があるためだとみられる。

6.おわりに

本研究では、単語トピック特定性によって、特定のトピックでしか用いられない単語により大きな値を与える重みを与え、共起性に基づく重みを調整した。前章でも説明したように PPMI の場合においては、単語トピック特定性によって文脈単語の重みを調整した場合のほうが、比較的良い結果が得られた。

本研究においてはより抽象的な意味を持つ単語は各トピックにおいて満遍なく割り当てられるということを前提にしていた。しかしながら、この前提はある意味では正しいが、ある意味では誤りである。抽象的な意味を持つ単語は様々な単語と共起しやすいが、高頻度の単語もより多くの単語と共起しやすい。そのため、同じぐらいの具体性を持つ単語であっても高頻度の単語にはより小さな重み付けしかなされないという問題が生じてしまう。これゆえ、今後の課題としては単語の頻度よりも共起の必然性によってより多くのトピックへ広がるトピックモデルを構築する必要がある。また式(7)で単語トピック特定性を調整するが、この式では、より早い段階で、単語トピック特定性が 1 に収束してしまう。よって新しい調整法を考えなければならない。

謝辞

この研究は北陸先端科学技術大学院大学(JAIST)の Data Science project において行われた。

参考文献

[1] Turney, Peter D., and Patrick Pantel. "From frequency to meaning: Vector space models of semantics." Journal of artificial intelligence research 37.1 (2010): 141-188.
[2] Clark, Stephen. "Vector space models of lexical meaning." Handbook of Contemporary Semantics, Wiley-Blackwell, à paraître (2012).
[3] Curran, James Richard. "From distributional to semantic similarity." (2004).
[4] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.
[5] Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." Proceedings of the National Academy of Sciences 101.suppl 1 (2004): 5228-5235.
[6] Arun, R., et al. "On finding the natural number of topics with latent dirichlet allocation: Some observations." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2010. 391-402.
[7] Lin, Jianhua. "Divergence measures based on the Shannon entropy." Information Theory, IEEE Transactions on 37.1 (1991): 145-151.
[8] Finkelstein, Lev, et al. "Placing search in context: The concept revisited." Proceedings of the 10th international conference on World Wide Web. ACM, 2001.
[9] Bruni, Elia, et al. "Distributional semantics in technicolor." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012.
[10] Radinsky, Kira, et al. "A word at a time: computing word relatedness using temporal semantic analysis." Proceedings of the 20th international conference on World wide web. ACM, 2011.
[11] Luong, Minh-Thang, Richard Socher, and Christopher D. Manning. "Better word representations with recursive neural networks for morphology." CoNLL-2013 104 (2013).