ガウス埋め込みに基づく単語の意味の史的変化分析

時武　孝介　　村脇　有吾　　黒橋　禎夫
京都大学工学部　京都大学大学院情報学研究科
{tokitake, murawaki, kuro}@nlp.ist.i.kyoto-u.ac.jp

1　はじめに

時代の変化に伴い、単語の意味は大きく変化したり新たな意味が加わったりすることが知られている。例えば、英単語 gay は 1920 年代以前には主に、「陽気な、楽しい」という意味で使われていたが、1920 年代を境に、徐々に「男性間の同性愛」を表す文脈において使われるようになった [4]。また単語 monitor は 1930 年代を境に「監視する」という意味に加えて「ディスプレイ、スクリーン」という意味を新たに獲得した [4]。

単語の意味の史的変化を分析する際、word2vec[5] に代表される単語埋め込みを用いることで、定量的な分析が可能となった。単語埋め込みの手法は、単語の主要な意味をベクトルで表し、線型空間上の点として表現する。従来研究では年代ごとに単語の分散表現を獲得し、その変化を見ることで意味の変化を捉えてきた [4]。例えば、空間上で類似度の高い単語を年代別に見ると、broadcast は 1850 年代には sow, seed など「ばらまく、タネをまく」といった意味に近かったが、テレビなどの普及により 1900 年代には television, radio などの単語の意味に近くなった。

この手法の問題点として、様々な文脈において現れる単語の意味の広がりを知ることができないことが挙げられる。単語の意味の広がりを知るためには、年代を経るにつれて、より広い文脈で使われるようになった単語があるかどうか、また使用される文脈が特定化される単語があるかどうかを分析して、主要な意味だけでなく、その単語の使われ方や意味の広がりを捉えたい。

本研究では Vilnis et al.[6] が提案した単語のガウス埋め込みに着目する。この手法は、単語の主要な意味を表現空間上の点として表現するだけでなく、その単語の意味の広がりをガウス分布の分散として表現する。例えば、広い意味を持つ man という単語のガウス埋め込みには、その大きな分散が作る分布の広がりによって、作曲家を表す composer という単語のガウス埋め込みを包含し、composer にはより具体的な Bach、Beethoven といった単語が包含されることが期待される。

ガウス埋め込みを単語の意味の史的変化の分析に用いるためには、年代ごとに学習したガウス埋め込みの統合が必要となる。　年代ごとにばらばらに学習した場合、ほとんど意味の変化しない単語に対しても、異なる年代では全く異なるガウス埋め込みが学習されるからである。

Hamilton et al.[4] は年代ごとに単語の分散表現を獲得したのち、隣り合う年代のそれらの分散表現を、同一空間上で表現するための直交行列を学習することで、隣り合う年代の分散表現を同一空間上で表現した。この手法は分散の対角成分のみを考えるガウス埋め込みのモデルには応用することができない。また Yao et al.[7] は分散表現の学習時に、隣り合う年代の分散表現を近づけるためのペナルティ項を損失関数に付け加えることで、分散表現を同一空間上で表現している。そこで本研究では、Yao et al. の手法をガウス分布の統合の手法として採用する。

本研究で提案する手法は単語の意味の史的変化を分析する際に、単語の主要な意味を分散表現で表す従来の分析手法を再現することができ、加えて史的変化する単語の意味の広がりを、分散を用いて分析する新たな視点を提供する。

2　ガウス埋め込み

単語のガウス埋め込み手法として Vilnis et al.[6] の手法を採用する。ガウス埋め込みでは各単語に対し、平均ベクトルと分散行列を割り当てる。ガウス埋め込みの分散が単語の意味の広がりに対応する。分散行列は対角成分のみを考えることでベクトルとして表現する。各単語に対応する平均、分散ベクトルを用いて対応するガウス埋め込みを計算することができる。単語同士の近さに基づいて、エネルギー関数をガウス分布の内積の対数を用いた以下の式で定義する。

$$E(P_i, P_j) = \log(\int_{x \in R^n} \mathcal{N}(x; \mu_i, \Sigma_i)\mathcal{N}(x; \mu_j, \Sigma_j)dx) \\ = \log N(0; \mu_i - \mu_j, \Sigma_i + \Sigma_j) \tag{1}$$

この式は 2 つのガウス分布の内積は別のガウス分布の $x = 0$ における値と等しいことを意味する。2 単語の平均ベクトルが近いほど、また分散の和が大きいほどエネルギー関数は大きくなる。ある単語（学習語) が特定の単語と共起しやすい場合、それらの単語のガウス埋め込みの平均は近くなる。また特定の単語のガウス埋め込みとの内積が大きければ良いので分散は小さくなる。このことは意味の広がりが小さな単語に対しては分散が小さいことに対応する。一方で、学習語が様々な単語と共起する場合、そのガウス埋め込みの平均は特定の単語と近くなることはなく、様々な単語のガウス埋め込みとの内積が大きくなってほしいということから、学習語のガウス埋め込みの分散は大きくなる。このことは意味の広がりが大きな単語に対しては分散が大きいことに対応する。

右辺のガウス分布の対数は以下のように計算できる。

$$\log N(0; \mu_i - \mu_j, \Sigma_i + \Sigma_j) = -1/2 \log \det(\Sigma_i + \Sigma_j) \\ -1/2(\mu_i - \mu_j)^\top (\Sigma_i + \Sigma_j)^{-1}(\mu_i - \mu_j) \\ -\frac{d}{2}\log(2\pi) \tag{2}$$

ベースとなる学習手法は word2vec の skipgram である。学習対象の単語に対して指定された窓枠内の単語を文脈語とする。各文脈語に対して語彙分布からランダムにサンプリングされた負例語を対応させる。次に学習対象語と文脈語から得られるエネルギー関数、学習対象語と負例語から得られるエネルギー関数を用いて hinge 損失関数を定義する。

$$L_m(w, c_p, c_n) = \max(0, m - E(w, c_p) + E(w, c_n)) \tag{3}$$

w は minibatch 内の学習対象の単語、c_p は文脈語、c_n は負例語, m はマージンである。この損失関数を最小化することは正例のエネルギー関数を大きく、負例のエネルギー関数を小さくすることを意味し、学習語の埋め込みと、それとよく共起する単語の埋め込みを近づけることを目的としている。これを最小化するように各単語の平均ベクトル、分散ベクトルを学習することで単語のガウス埋め込みを獲得できる。

3 年代ごとのガウス埋め込みの統合

次に史的変化を分析するために年代ごとのガウス埋め込みを統合する。統合とは年代ごとのガウス埋め込みを、同一空間上で表現することを意味する。そのために同じ単語の、隣り合う年代のガウス埋め込みの平均ベクトルがほぼ同じものなるようなペナルティ項を, 損失関数に加える。隣り合う年代において単語の意味はほとんど変化しないと考えられるからである。よって統合時の学習における損失関数は、各年代の損失関数の和に penalty 項を加えたものになる。加えるペナルティ項は以下のものである。

$$penalty = \lambda \times \Sigma_i ||W_i - W_{i+1}||^2 \tag{4}$$

λ はペナルティ項の係数、W は全単語の平均ベクトルを行にもつ行列、Σ_i は全年代で足し合わせることを表す。毎回の学習において、各単語のガウス埋め込みが学習されつつ、隣の年代のガウス埋め込みが似たものになるようにパラメータの更新が行われる。

4 実験

4.1 ガウス埋め込み獲得の予備実験

4.1.1 実験設定

先行研究 [6] で提案されたガウス埋め込みの性能を確認する予備実験を行った。学習には Copus of Historical American English (COHA)[2] を用いた。COHA はアメリカ英語のコーパスで 1810 年 2009 年の文書を年代ごとにまとめた genre balanced のコーパスである。語彙サイズは 28 万 、出現頻度 5 以下の単語は unknown で、アラビア数字は全て N で置き換えた。大文字を含む単語は出現頻度 200 以下の単語については全て小文字に置き換えた。予備実験では 2000 年代の文書のみを使用した。含まれる単語数は 3300 万である。

4.1.2 モデル設定

ガウス埋め込みの次元は 50 次元とした。optimizer は Adagrad とし、学習率は 0.5。minibatch を 64、窓枠を 10、 epoch を 20 とした。単語間の頻度の偏りをならすために subsampling を行った [5]。

4.1.3 評価方法

評価のためのデータセットとして Wordsim353 を用いる [3]。これはあらかじめ決められた同義語、反意語などの単語間の関係に対して、各単語ペアがどの

程度当てはまるかを 0-10 の数値で人が評価したデータセットである。評価時には、まず Wordsim353 で用意された単語対の近さを計算する。ここで単語対の近さを、学習されたガウス分布の平均ベクトルのコサイン類似度で定義する。全単語対のコサイン類似度を計算し、データセット内で人手で与えられた評価値との Spearman の ρ をモデルのよさと考える。

4.1.4 結果

結果の最良値は 2 節で紹介したモデルは $\rho = 0.59$ であった。エネルギー関数を定義する際、KL divergence の負で定義するモデルもある [6]。こちらの最良値は $\rho = 0.36$ であり、史的変化の分析においては後者のモデルを用いた。比較対象として、word2vec[5] で評価を行ったところ、$\rho = 0.58$ であった。多峰性ガウス分布を用いた埋め込み手法 [1] では同様のデータセットで $\rho = 0.76$ が報告されており、これと比べると値が低めであるが、分散表現を獲得するための代表的な手法である word2vec と同等の値であった。

4.2 ガウス埋め込みの統合実験

4.2.1 実験設定

史的変化を分析するための埋め込みの統合実験をする。COHA の 1900 年代以降の文書を 10 年ごとに区切り、それぞれを一つの年代とした。実験設定値は予備実験の値と同じにした。ペナルティ項の係数は λ =0.2 とした。

4.2.2 評価方法

統合後のモデルの評価を次のように行う。先行研究 [4] において意味の変化の開始時期と、変化前後の意味が明らかになっている単語がある。それらの単語に対して期待される変化が生じているかどうかを見るために、変化後の意味に近い単語とのコサイン類似度を計算し、その年代変化を見る。またガウス埋め込みの分散は多次元ガウス分布の 95%区間に当たる hyper-volume に換算でき、新たな意味が加わった単語に対しては、hyper-volume の年代変化を見る。1900 年代以降で変化が期待される単語は gay, broadcast, monitor, record の 4 単語である。gay、broadcast は意味が劇的に変化した単語、monitor, record は新たな意味が加わった単語であり、表 1 のような意味の変化があったことがわかっている。

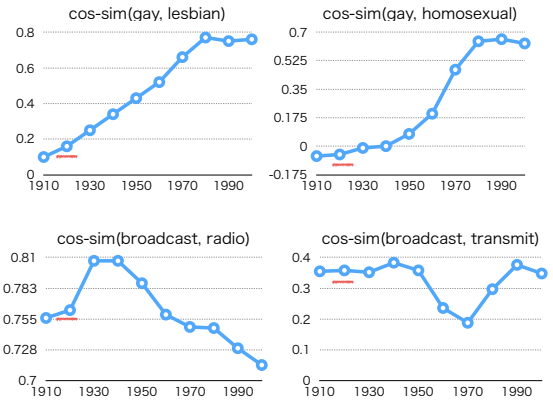

図 1: コサイン類似度の変化 1

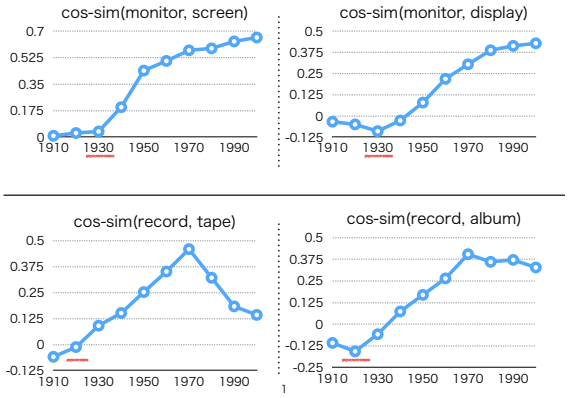

図 2: コサイン類似度の変化 2

4.2.3 結果

図 1 の赤線は先行研究で意味の変化が開始するとされる年代を示す。(gay,lesbian), (gay,homosexual) のコサイン類似度は共に赤線の 1920 年代付近から上昇している。これは、1920 年代から徐々に単語 gay の意味が lesbian, homosexual に近づいていることを表す。一方で (broadcast,radio), (broadcast,transmit) のコサイン類似度の変化では期待される変化を見てとることはできない。新たな意味が加わった単語である monitor, record では、それぞれ (monitor,screen), (monitor,display), (record,tape), (record,album) のコサイン類似度は赤線付近から上昇している。これらのことから、提案手法を用いることで先行研究において明らかになった単語の意味の変化を 4 単語中 3 単語において捉えることができたと言える。次に hyper-volume の変化を図 3 に示す。新たな意味が加わった単語である monitor は hyper-volume が上昇している。一方で gay, broadcast, record では全体的な変化を見て取る

単語	変化後	変化前	変化開始年代	出典
gay	homosexual, lesbian	showy, happy	1920	Kulkarni et al., 2014
broadcast	radio, transmit	seed, scatter	1920	Jeffers and Lehiste, 1979
monitor	display, screen	–	1930	Simpson et al., 1989
record	album, tape	–	1920	Kulkarni et al., 2014

表 1: Hamilton et al.[4] が調査したうち 1900 年以降で意味が変化した単語

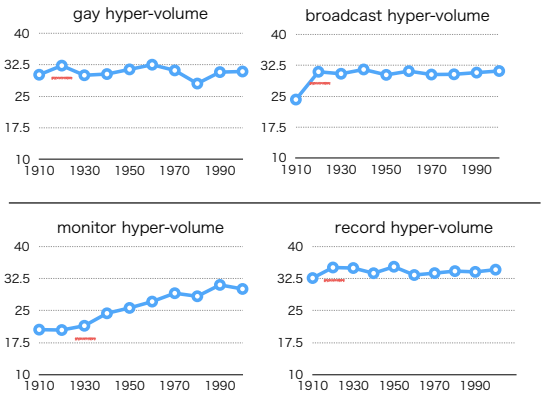

図 3: hyper-volume の変化

ことはできない。gay,broadcast は新たな意味が加わった単語ではなく、結果は妥当であるとも考えられるが、新たな意味が加わった record においても上昇がみられなかった。

4.2.4 議論

損失関数に加えたペナルティ項は、学習時に毎回全単語の平均ベクトルを近づける。このため出現頻度の高い単語に対しては学習がよく進む一歩で、出現頻度の低い単語に対してはガウス埋め込みの学習よりも隣り合う年代の平均ベクトルを近づけるペナルティの効果が大きくなり、意味の変化が適切に捉えられない可能性がある。そこでペナルティ項の改善案として下式のように、学習対象の単語に制限する方法が考えられる。

$$penalty = \lambda \times \Sigma_i ||W_i[w] - W_{i+1}[w]||^2 \qquad (5)$$

w は毎回の学習において学習対象となる単語である。

5 おわりに

本研究では、従来分散表現を用いて行われていた単語の意味の史的変遷の分析に対して、ガウス埋め込みを用いた手法を提案した。今回明らかになった問題点は改善していきたい。また今後の発展として多峰性のガウス分布 [1] を用いて単語の埋め込みを行うことによって、多義語の単語の意味をよりはっきりと捉えられるようにすること、また多言語コーパスを用いることで意味の借用による変化が捉えられるようにすることなどが考えられる。

参考文献

[1] Ben Athiwaratkun and Andrew Gordon Wilson. Multimodal word distributions. *ACL*, 2017.

[2] Mark Davies. The corpus of historical american english: 400 million words, 1810-2009. 2010.

[3] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, pp. 20(1):116–13, January 2002.

[4] William L. Hamilton, Jure Leskovec, and Dan Jurafs. Shift or linguistic drift? comparing two computational measures of semantic chang. *EMNLP*, 2016.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dea. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing system*, pp. 3111–3119.

[6] Vilnis and McCallum. Word representations via gaussian embedding. *ICLR*, 2015.

[7] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Discovery of evolving semantics through dynamic word embedding learning. *arXiv*, 2017-03-02.