修士論文 要旨

分布変換に基づく単語分散表現

数理情報学専攻 48206205　江夏 永広
指導教員　田中 久美子 教授

1　導入

自然言語処理において単語を比較的低次元 (50-300 次元) のベクトルで表現する単語埋め込みは、2010 年代以降の自然言語処理において重要な研究課題の一つである。word2gm[2] は単語埋め込みを混合ガウス分布で表すことにより、単語の持つ多義性や曖昧性を表現することが可能となった。本研究では単語は複数のガウス分布で表せないより複雑な情報が含まれているのではないかと仮定し、ガウス分布をニューラルネットワークで変換した分布を一つの埋め込みとして、ニューラルネットワークのパラメータを変化させて単語の意味表現を学習する手法である word2flow,word2WGAN を考案した。またその手法で構成した分布に対し、多義語に対する意味表現の評価などを行った。結果、多義語に関して既存手法の word2gm では語義数を指定しなければ語義数分の分布を得られなかったのに対し、word2WGAN では指定せずとも分布が語義ごとのクラスターに分離することがあることが分かった。

2　既存研究と背景

単語埋め込みとして最も有名な word2vec[1] は一つの単語を一つのベクトルで表現するものである。『ある単語 w の類似単語はコーパス中で w の前後に位置することが多い』という分布仮説に基づく手法により、コーパスから単語ベクトルを 50-300 次元に埋め込み、かつ似た意味を持つ単語同士のベクトルの cos 類似度 (ベクトル間の角度 θ に対し $\cos\theta$ の値のこと) が大きくなる方向に位置するように学習することを可能にした。特に word2vec のモデルの一つである skipgram は、単語同士の類似度を表すエネルギー関数 E に対して $E(w, w_c)$ が大きくなるように (ただし w_c は w のコーパス中の周辺語)、$E(w, w_n)$ の値が小さくなるよう (ただし w_n は w と関係のないランダムな単語) に学習するものであった。特に後者の $E(w, w_n)$ を小さくする学習のことをネガティブサンプリングと呼ぶ。そして、$E(w_1, w_2)$ には単語 w_1, w_2 の単語ベクトルの内積を使用していた。

word2gm[2] は一つの単語を混合ガウス分布で表現するものである。学習方法は skipgram と非常に類似している。ただし、エネルギー関数の計算に二つの単語ベクトルの内積を使えない代わりに、二つの単語埋め込み分布の積を空間全体で積分したもので代用して計算するという点で違いがある。この学習手法により、似た分布ほど似た形になるように学習することが可能となった。本分布は rock や bank などの複数の意味を持つ単語に対して、その意味に即した分布を獲得できていることが実験により確認された。

3　提案手法

混合ガウス分布を超える表現力を持つ分布を考えたい。多変量標準正規分布に従う確率変数 Z をニューラルネットワーク f_θ で変換した $f_\theta(Z)$ に対し、$f_\theta(Z)$ が従う分布 p_θ を埋め込みとして用いることを考える。ただし、意味を学習できるための条件として変換後の p_θ と他の分布の類似度、すなわちエネルギー関数を計算できる必要がある。これは word2vec/gm で使用されているエネルギー関数を使用する skipgram を元にした意味の学習法を使用するためである (ただし周辺語分布、関係ない単語の分布は学習済み word2vec から構成して使う)。その上で word2flow と word2WGAN という二つの手法を考えた。

word2flow は f_θ として、逆関数 f_θ^{-1} とヤコビアン $|f_\theta|$ が計算可能という特徴を持つニューラルネットワークを使用したものであり (f_θ^{-1} を flow と呼び、本関数を変分推定などに使う手法を Nomalizing flow[3] と呼ぶ)、これらの特徴から p_θ と任意の分布との KL ダイバージェンスが容易に計算できる。KL ダイバージェンスは分布間の距離 (非対称性により厳密な距離とは言えないが) のように使えるため、これの符号を反転させたものを類似度、エネルギー関数として分布を学習するのが word2flow である。

word2WGAN は f_θ には特に制約のないニューラルネットワークを使用し、分布 p_θ と分布 q のエネルギー関数に Wasserstein 距離の符号を反転させたものを用いるものである。ただし、Wasserstein 距離の学習にはさらに別の Critic と呼ばれるニューラルネットワークで loss 関数を計算する [4] 必要があり、計算負荷の高いものとなっているが、その分 Wasserstein 距離特有の θ に対する勾配の安定性から、安定した学習が可能となっている。

4 実験結果

多義語に関して word2WGAN が適切に意味を反映した分布となっているかを調べたい (loss の不安定性により word2flow は考案に留め実験には用いない)。そのため、word2WGAN から 10000 ベクトルサンプリングし、DCSCAN で形成したクラスターと、word2gm から形成されるクラスター (ガウス分布) の最類似単語を比べる。表 1 には、word2gm のクラスターの中心 (ガウス分布の期待値) ベクトルとの cos 類似度が高い順の単語と、word2WGAN のクラスター中の各ベクトルと cos 類似度が最大となる単語 (最大となった回数の降順に並べる) を載せている。ただし word2WGAN は”the”などの頻出語のクラスターが 2 つほど形成されているがそれは表に載せていない。word2gm では意味の数 (ガウス分布の数) を 2(元論文の実験設定) と設定しているため取得できていない「燃料」を意味するクラスターが学習できていることが分かる。

	word2gm	word2WGAN
クラスター 1	cytoplasm vesicle cytoplasmic macrophages secreted membrane mitotic endocytosis	cell cells membrane glomeruli function processes proteins genes
クラスター 2	cellular Nextel 2-line Sprint phones pda handset handsets pushbuttons	phone phones pager phreak cellphones walkietalkie typewriter tivo switchboard
クラスター 3		gasification photovoltaic fuel alumina wastetoenergy transpiration refining minings leaching

表 1. word2gm,word2WGAN の”cell”のクラスターの比較

次に、形成したクラスターの形状について考察する。図 1 左は”rock”という単語の word2WGAN の分布埋め込みの内「音楽」の意味合いを持ったクラスターを取得し、PCA によって 2 次元にしてプロットしたものである。右は、クラスターと同一の期待値と分散を持つガウス分布からサンプリングしたベクトルを、左図と同様の空間に射影しプロットしたものである。図 1 左は album(クラスター中、右上) から song(クラスター中、下) からロックバンドの固有名詞 (クラスター中、右側) の意味を表す空間に連続的に分布しているのに対して、図 1 右のクラスターは woomble や granz など、rock と関わりはするが明確にその意味を表しているとは言えない空間に多く分布している。この結果から、word2WGAN によって形成されたクラスターはガウス分布とは異なる形状をしており、かつその形状は意味を反映したものとなっている場合があると言える。

5 結論

本研究で考案した word2WGAN は、不要な頻出語のクラスターを作ってしまうという欠点はあるが、word2gm で語義数を指定する必要があった多義語の語義について、指定せずともその語義ごとのクラスターを学習できた。また、クラスターの形状についても分析の結果、通常のガウス分布では表現できない意味を反映したものとなっていることが判明した。

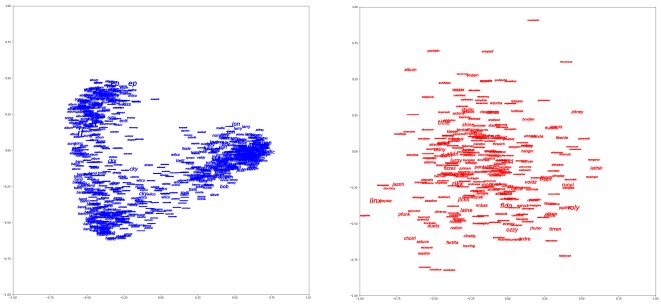

図 1. クラスターを PCA により 2 次元にしてプロットしたもの (左), クラスターと同様の期待値と分散を持つガウス分布をサンプリングしてプロットしたもの (右)

参考文献

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. Proceedings of ICLR.

[2] Ben Athiwaratkun and Andrew Gordon Wilson. 2017. Multimodal word distributions. In ACL.

[3] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In: International Conference on Machine Learning. 1530 – 1538

[4] Martin Arjovsky, Soumith Chintala, and Leon Bottou. Wasserstein generative adversarial networks. In ´ ICML, pp. 214 – 223, 2017.