4K1-1

単語の分散表現を利用した文書類似度

Document similarity using distributed word representation

柳本 豪一*1
Hidekazu Yanagimoto

*1大阪府立大学
Osaka Prefecture University

I propose a new method to calculate document similarity based on distributed word representation. Neural network language models construct distributed word representation from a text corpus and the representation can capture semantical similarity. However, they are discussed in word level and it is not clear how you construct document representation from the distributed word representation. In this study I construct distributed word representation using word2vec and define document similarity using Earth Mover's Distance to consider similarity among words. In experiments the proposed method can define similarity scores considering word synonyms.

1. はじめに

ニューラルネットワーク言語モデルを用いた単語の分散表現により、類似した単語や線形演算により単語間の類似関係を類推することができると報告されている。これは従来の Bag-of-Words モデルや潜在的意味解析を行ったとしても実現することが難しいものである。しかし、これらは単語レベルのみ行われており、文を表現するために十分活用されているとは言い難い。これを実現するためには、単語間の類似性に関する情報を有している分散表現を組み合わせる方法について検討する必要がある。

本研究では単語の集合として文書を表現し、分布の距離を用いることで文書類似度を定義する。この時、単語間の意味の近さを考慮した類似度を定義する必要があるため、一般的な内積に基づいた手法を用いることはできない。したがって、分布の要素間の距離を考慮した距離を定義することができる Earth Mover's Distance[Rubner 00] を用いることとする。

単語の分散表現と Earth Mover's Distance を用いた文書類似度を用いることで、単語の同義語や類義語を考慮した類似度が計算できることが実験より確認できた。この方法では、ニューラルネットワーク言語モデルを用いて単語の分散表現を用いているため、シソーラスなどの他の言語資源を必要とせず、コーパスのみで実現している点が特徴である。

2. 単語の分散表現を利用した文書類似度の提案

ニューラルネットワーク言語モデルの一つである word2vec[Mikolov 13] を用いた単語の分散表現と Earth Mover's Distance による類似度計算を用いた手法について説明をする。

2.1 word2vec を用いた単語の分散表現

ニューラルネットワーク言語モデルを用いた単語の分散表現としては、word2vec が有名である。本研究においても、単語の分散表現を得るために word2vec を用いることとする。以下では、word2vec について説明する。word2vec では図 1 に示すような skip-gram モデルと呼ばれるニューラルネットワークを用いて単語の分散表現を作成する。

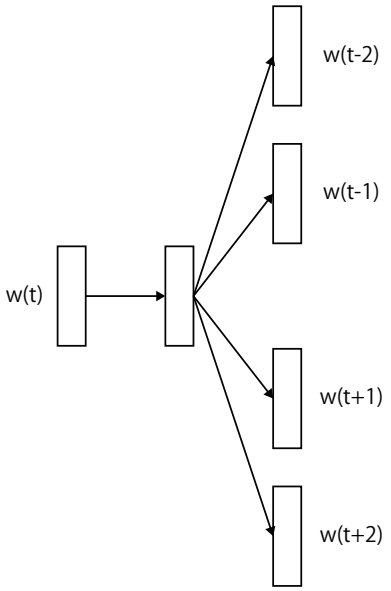

図 1: Skip-gram モデル

word2vec では単語は 1-of-N coding により表現されており、入力単語の前後の単語を予測するようにニューラルネットワークがコーパスにより学習される。学習後の入力層と隠れ層間の重みを用いて単語をベクトルとして表現する。得られた分散表現は、意味的に類似した単語は空間上の近い位置に配置されたり、線形演算により類似性の推論が可能であるという特徴を持っていることが知られている。

本研究では、このようにして得られた単語の分散表現をもとに文書を表現し、文書の類似度を計算することを目指す。具体的には、分散表現により表されている単語間の意味的な近さを考慮した類似度を提案することである。これは、単語同士が直交していないため、コサイン類似度などを用いることはできない。したがって、以下では Earth Mover's Distance を用いた類似度の計算について説明を行う。

連絡先: 柳本 豪一，大阪府立大学，大阪府堺市中区学園町 1-1，072-254-9279，072-254-9279，hidekazu@cs.osakafu-u.ac.jp

2.2 Earth Mover's Distance を用いた類似度計算

Earth Mover's Distance(以下 EMD) は 2 つの分布間の距離を輸送問題の解である輸送コストを用いて定義した距離である。一般的に分布がヒストグラムで表されていると考えると、EMD では階級間の距離をあらかじめ定義しておけば、異なる階級を持つヒストグラム間でも比較を行うことができる。本研究では、文書を単語で表されるヒストグラムとみなし、文書の類似度を EMD を用いて求める。

今、2 つの分布を $P = \{(\mathbf{p}_1, w_{\mathbf{p}_1}), \cdots, (\mathbf{p}_m, w_{\mathbf{p}_m})\}$、$Q = \{(\mathbf{q}_1, w_{\mathbf{q}_1}), \cdots, (\mathbf{q}_n, w_{\mathbf{q}_n})\}$ と荒らすとする。また、$\mathbf{q}_i$ と $\mathbf{q}_j$ 間の距離を d_{ij} とし、$\mathrm{D} = [d_{ij}]$ と表すとする。この時、EMD では以下の輸送問題を考え、最小の輸送量 F^* を用いて距離が定義される。

$$F^* = \arg\min_F \mathrm{WORK}(P, Q, F) = \arg\min_{f_{ij}} \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij} \tag{1}$$

ただし、以下の制約条件を満たすものとする。

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m, 1 \leq j \leq n \tag{2}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{\mathbf{p}_i} \qquad 1 \leq i \leq m \tag{3}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{\mathbf{q}_j} \qquad 1 \leq j \leq n \tag{4}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min(\sum_{i=1}^{m} w_{\mathbf{p}_i}, \sum_{j=1}^{n} w_{\mathbf{q}_j}) \tag{5}$$

ここで得られた最適な輸送量を用いて、EMD は以下のように定義される。

$$\mathrm{EMD}(P, Q) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} f^*_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f^*_{ij}} \tag{6}$$

本研究では、$\mathbf{p}_i$ は i 番目の単語に対応する分散表現を表し、$w_{\mathbf{p}_i}$ は i 番目の単語の出現頻度とする。これにより、単語間の近さを考慮した文書の類似度を計算することが可能となる。

3. 実験

株式ニュースをコーパスとして用いることで単語の分散表現を作成し、文を分散表現を用いて表現し、文間の類似度を求めることで提案手法の有効性を確認する。

3.1 実験環境

実験には 2010 年の T&C ニュースを用いる。これは、2010 年 1 月 1 日から 2010 年 12 月 31 日までにメールで配信された 62,378 件の記事である。ここから、自然言語で書かれた文のみを抽出したものをコーパスとして用いる。これは、チャートのアスキーアートなどが含まれているためである。この処理により 471,243 文が得られ、MeCab により分かち書きを行ったものを word2vec の入力とする。

word2vec のパラメータとしては、隠れ層のニューロン数をあらかじめ設定する必要がある。本実験では 200 として実験を行っている。このため、各単語は 200 次元のベクトルとして表現されることとなる。

表 1: 入力文とその類似度

ANA の株価が上昇した。	
全日本空輸の株価が上昇した	0.8812
JAL の株価が上昇した	0.7921
富士通の株価が上昇した	0.7717
ANA の株価が下落した	0.9410

3.2 結果

実験結果の例を表 1 に示す。これらの例文は Bag-of-Words モデルを用いてコサイン類似度を用いると全て同じ値となり、文間で差をつけることはできない。

提案手法を用いることにより、単語の分散表現が持っていた特徴である意味的な類似性を考慮した類似度が定義できていることがわかる。ANA と全日本空輸は表記は異なるが、同一の企業を表しているため、他の文に比べて高い類似度を持っているべきである。つまり、ANA の部分が JAL に置き換えられた文に比べて高い類似度になっていることは望ましい。また、JAL と ANA は共に航空会社であり、他業種の富士通に比べ高い類似度になっていることが好ましい。以上の点から結果を見ると、文の類似度に以上の観点が反映されていると言える。

一方、4 番目の文について考えると、「上昇」と「下落」のみが異なっており、ともに ANA の株価についての話題を扱っている。このため、上記の 2 つの文は高い類似度となっている。

実験結果については妥当な結果が得られているが、全ての単語に対して正しい結果が得られているわけではない。つまり、コーパスに含まれている全ての企業間の関係が正しく把握できているわではないので、どのような傾向があるかについて検討する必要がある。

4. おわりに

本研究では、ニューラルネットワーク言語モデルにより得られた単語の分散表現を用いた文書類似度の提案を行った。具体的には、word2vec を用いて単語の分散表現を作成し、その分散表現を用いて Earth Mover's Distance により類似度を計算することで、単語間の類似性を考慮した文書類似度を提案した。実験により、コーパスのみを用いることで、同義語や類義語を考慮した類似度を求められることを確認した。

今後は、コーパスに含まれる様々な単語について類似度がどう変化するか検討することで、提案手法が適用できる語彙について検討を行う。また、word2vec を改良することで、単語の分散表現の改良を目指す。

参考文献

[Rubner 00] Rubner, C., Tomasi. C., and Guibas, I.: The Earth Mover's Distance as a Metric for Image Retrieval, Journal of Computer Vision, Vol. 40, No.2, pp.99-121 (2000).

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G, and Dean, J.: Efficient Estimation of Word Representations in Vector Space, Proceedings of Workshop at ICLR(2013)