2E3-1

単語の分散表現における関係抽出

Extraction of relations in distributed representation of words

蛭子 琢磨[*1, 2]
Takuma Ebisu

市瀬 龍太郎[*2, 1]
Ryutaro Ichise

[*1] 総合研究大学院大学
SOKENDAI

[*2] 国立情報学研究所
National Institute of Informatics

Distributed vector representation of words has played a wide range of roles in natural language processing and become more important because of its ability to capture a large number of syntactic and lexical meaning or relationship. Relation vectors are used to represent relations of words but it has some problems; some of relations cannot be explained, for example sibling-relation, parents-relation and many-to-one-relation. To deal with the problems we created novel representation of relation. Relations are represented by planes instead of vectors in representation space and it enables us to predict relation more accurately.

1. はじめに

知識をどのようにコンピュータ内で表現するかということは，人工知能を作る上での重要な課題であり，最近ではニューラルネットワークによる単語の分散表現が様々なタスクに応用され，注目が集まっている．分散表現とは，対象を比較的低次元なベクトル空間である表現空間内の密な点として表現することである．この単語分散表現の便利な特徴として，単語間の関係をベクトルで表現できるということがある．例えば，英単語の現在形と過去形の関係を表すベクトルrが存在して，run の分散表現にrを足したものに一番近くに ran の分散表現が存在する．しかし，関係には様々なものがあり，原理的にこの手法では表現できないようなものも存在する．

本研究では，そのような問題を解決しうる新たな方法として，単語間の関係を表す方法を拡張し，関係を 1 つのベクトルでなく，より幅をもった面として表現することを提案する．この手法が従来の手法と比べてよりよく関係を表すことを，関係の予測精度を実験で計ることで示す．

2. ニューラルネットワーク言語モデル

ニューラルネットワーク言語モデルは，ニューラルネットワークを用いて，単語の分散表現の学習を行うモデルである．ラベル付けがされていない大量の文章を用いて単語の分散表現を学習する．初期のニューラルネットワーク言語モデルである Bengio らのモデル [Bengio 03]や，モデルの単純化と計算量の軽量化を行ったものである Mikolov らのモデル[Mikolov 13]が知られている．これらは，文章内の単語をその周辺の単語から予測するように学習され，Mikolov らのモデルでは，共起する単語の分散表現の内積がより大きくなるように学習する．この結果，意味的や文法的に近い単語は，表現空間内の近い位置に写像される．

最近では，複数の意味を持つ単語を意味ごとに分離して扱うような手法もある．例えば，Jauhar らは，分散表現を学習する際に平文とともにオントロジーを用いて意味ごとの分離を行っている[Jauhar 15]．具体的には，オントロジーとして WordNet を用いて多義語を意味ごとに分離し，WordNet 上で関係している単語同士を表現空間で近くなるように学習する．Neelakantan らは，分散表現を学習しながら，周辺にある単語の分散表現の平均を文脈と定義し，その文脈の違いにより多義語を分類する[Neelakantan 15].

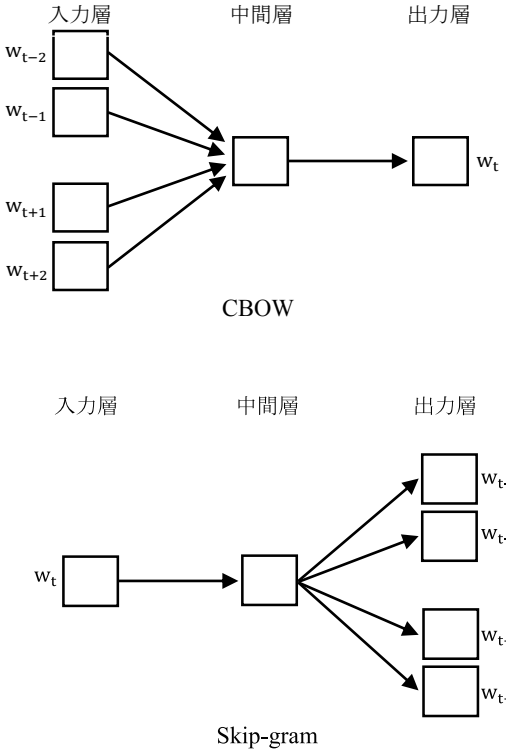

図 1:Mikolov らが提案した 2 つのニューラルネットワーク言語モデル[Mikolov 13]．CBOW は周辺の単語から現在の単語を予測できるように学習する．Skip-gram モデルは現在の単語から周辺の単語を予測できるように学習する．

また，単語を表現空間内の点ではなくガウス分布で表現して，曖昧な意味をもつ単語に対応した手法[Vilnis 14] もでてきている．曖昧な単語の分布は分散が大きく，表現空間内に広がっており，逆に意味が狭い単語の分布の分散は小さく，局所的に存在するように表される．

連絡先：蛭子 琢磨，総合研究大学院大学複合科学研究科情報学専攻，〒101-8430 東京都千代田区一ツ橋 2-1-2，E-mail:takuma@nii.ac.jp

このように，単語の分散表現の研究は活発であり，様々な手法が考案されている．しかし，単語間の関係をどのように表現するかについての研究は乏しい．

3. 単語間関係の表現

Mikolov らの研究[Mikolov 13]において，単語間の関係をベクトルで表すことができることが示されている．例えば，(Tokyo, Japan)，(Paris, France)，(Beijing, China)の単語の組は，それぞれ首都とそれが位置する国という同じ関係にある．ここで，v_wで単語 w の分散表現を表すと，

$$\begin{aligned} v_{Japan} - v_{Tokyo} &\fallingdotseq v_{France} - v_{Paris} \\ &\fallingdotseq v_{China} - v_{Beijing} \end{aligned}$$

となる．つまり，(Paris, ?)は(Tokyo, Japan)と同じ関係であるとすると$v_{Japan} - v_{Tokyo} + v_{Paris}$に近い分散表現を持つ単語を探すことで，？がフランスであることを予測することが可能となる．

しかし，上式において，≒の代わりに＝が成り立つような理想的な状況を想定した場合に，sibling 関係などの反射律が成り立つ関係や，is-a 関係などの推移律が成り立つ関係，そして，多対1関係などを表現できないという問題点がある．たとえば，(A, B)が対称律の成り立つ関係にあるとすると，(B, A)も同じ関係にある．したがって，

$$\begin{aligned} v_B - v_A &= v_A - v_B \\ \therefore v_A &= v_B \end{aligned}$$

A と B は同一のものを表している単語ではないのにもかかわらず，分散表現が完全に一致してしまうという問題が生じる．また，関係を表すベクトルも 0 になってしまう．これらの問題を解決する，新しい関係の表現方法を本研究では提案する．

4. 提案手法

4.1 提案手法の概要

本研究では，関係をベクトルではなく，広がりを持つ面で表すことを提案する．つまり，分散表現空間Vのあるベクトルrと部分空間Uの和として関係を表現する．数式を用いると，関係を表す集合をRとして，

$$\begin{aligned} R &= r + U \\ &= \{r + u \mid u \in U\} \end{aligned}$$

と表される．

次に R の決定方法について述べる．同一の関係にある subject と object の組からなる訓練集合$S_{training} = \{(s_i,\ o_i) \mid i = 1,\ \dots,\ \mathrm{N}\}$が与えられた時に，$s_i + R$と$o_i$のユークリッド距離の二乗和である目的関数$f$が最小になるように定める．

$$\begin{aligned} f &= \sum_{\mathrm{i}=1}^{\mathrm{N}} d\left(v_{s_i} + R, v_{o_i}\right)^2 \\ &= \sum_{i=1}^{N} \min_{x \in \mathrm{v}_{s_i} + R} |x - v_{o_i}|^2 \end{aligned}$$

この関係と同じ関係にある組(s, ?)の予測は，

$$\operatorname*{argmin}_{w} d(v_s + R,\ v_w)$$

を計算することで行われる．

4.2 計算手法

上記の手法を実装する際の計算方法を示す．

fを変形すると，

$$f = \sum_{\mathrm{i}=1}^{\mathrm{N}} d\left(R, v_{o_i} - v_{s_i}\right)^2$$

$x_i = v_{o_i} - v_{s_i}$とおくと

$$r = \overline{x} = \text{ } x_i\text{の平均}$$

としてよいことを示す．Vの次元をD，Uの次元をE，$(b_1, \dots, b_D)$をVの正規直交基底で，b_1，b_2，…，b_EがUを張るものとすると，

$$\begin{aligned} d(R, x_i)^2 &= d(U, x_i - r)^2 \\ &= |x_i - r|^2 - \sum_{j=1}^{E} \left(b_j^{\mathrm{T}}(x_i - r)\right)^2 \\ &= \sum_{j=1}^{D} \left(b_j^{\mathrm{T}}(x_i - r)\right)^2 - \sum_{j=1}^{E} \left(b_j^{\mathrm{T}}(x_i - r)\right)^2 \\ &= \sum_{j=E+1}^{D} \left(b_j^{\mathrm{T}}(x_i - r)\right)^2 \\ \therefore f &= \sum_{i=1}^{N} d\left(R, v_{o_i} - v_{s_i}\right)^2 \\ &= \sum_{j=E+1}^{D} \sum_{i}^{N} \left(b_j^{\mathrm{T}}(x_i - r)\right)^2 \end{aligned}$$

これを解くと，$\mathrm{r} = \overline{\mathrm{x}}$で fが最小となることがわかる．したがって，$\hat{x}_i = x_i - \overline{x}$とおくと，

$$\begin{aligned} f &= \sum_{\mathrm{i}=1}^{\mathrm{N}} d(U, \hat{x}_i)^2 \\ &= \sum_{j=E+1}^{D} \sum_{i}^{N} \left(b_j^{\mathrm{T}} \hat{x}_i\right)^2 \end{aligned}$$

を最小とするUを見つければ良い．fを変形すると，

$$\begin{aligned} f &= \sum_{j=E+1}^{D} \sum_{i}^{N} \left(b_j^{\mathrm{T}} \hat{x}_i\right)^2 \\ &= \sum_{j=E+1}^{D} \sum_{i}^{N} b_j^{\mathrm{T}} \hat{x}_i \hat{x}_i^{\mathrm{T}} b_j \\ &= \sum_{j=E+1}^{D} b_j^{\mathrm{T}} \sum_{i}^{N} \hat{x}_i \hat{x}_i^{\mathrm{T}} b_j \end{aligned}$$

$\sum_i^N \hat{x}_i \hat{x}_i^{\mathrm{T}}$は対称行列であるので，固有ベクトルを正規直交基底となるように取ることができる．つまり，固有値が大きい順にD個の固有ベクトルを選び，それらが張る部分空間をUとすればよい．

Uと直交する正規化された固有ベクトルを並べた行列をBとすると，ユークリッド距離関数dは次のようにかける．

$$d(v_s + R, v_w) = |B^{\mathsf{T}}(v_w - v_s)|$$

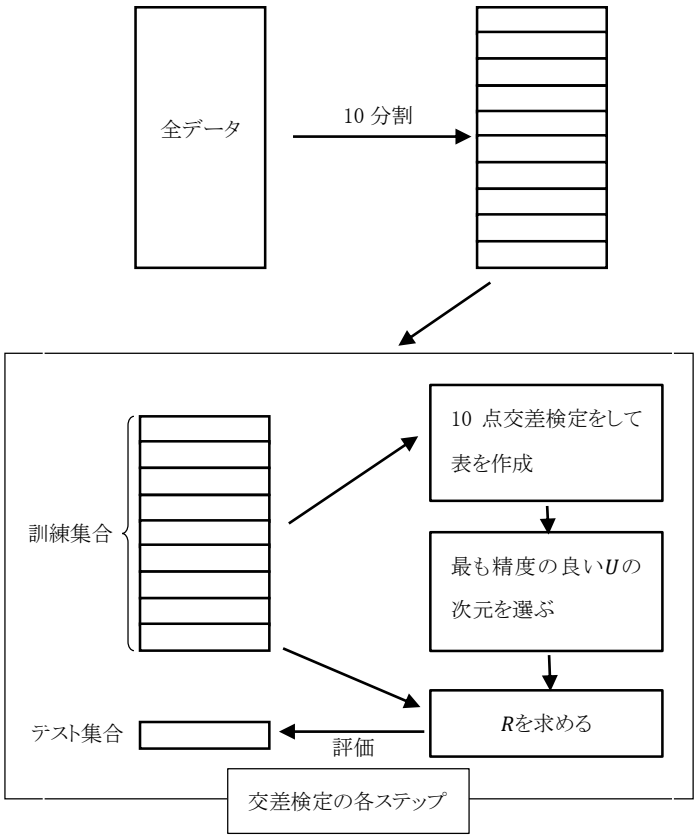

図 2:実験 2 のプロセス図 ．交差検定の各ステップの訓練集合に対して再度交差検定を行い，パラメータであるUの次元を決定する.

4.3 提案手法の別の解釈

この手法には別の解釈が存在する. $\sum_i^N \hat{x}_i \hat{x}_i^{\mathsf{T}}$は$\{x_i\}$の分散共分散行列の定数倍であるので，第 4.2 節から導かれる通り，この手法は，基底を$\{x_i\}$の共分散が 0 になるような正規直交基底に取り直し，そのあと$\{x_i\}$の分散が大きい成分から順にD個減らして関係を表すベクトルとして単純に平均$B^{\mathsf{T}}\overline{x}$を取ることを意味する. これは関係を表すのに必要である成分だけを抜き出すことに対応していると考えられる.

5. 実験

5.1 実験設定

単語の分散表現を得るために，word2vec[1]を用いた. 表現空間の次元は 300 で CBOW と Skip-gram によって単語の分散表現をそれぞれ作成した. その際のコーパスとして，英語版 wikipedia を使用した. 総単語数は約 18 億で，出現数が 5 以上のユニークな単語数は約 170 万で，この単語の分散表現を学習させた. Mikolov らの研究[Mikolov 13]で，関係の表現能力の実験に用いられた時と同様に，この分散表現を正規化して実験に使用する.

関係の表現能力を調べるために，都市－国の関係を持つ単語の組を用いた. このデータは GeoNames より，各国から人口の多い順に都市を 10 個ずつ取ってきたものである. ただし，単語の分散表現は熟語に対応していないため，1 つの単語からなる都市名と国名の組だけを使用した. その結果，合計 599 組のデータが得られた.

これらのデータを用いて，提案手法を評価するための実験を 2 つ行った. 実験 1 は，Uの次元Dが 1 から 160 まで動かした場合にそれぞれ 10 点交差検定を行い，都市－国関係の予測精度について実験した. 実験は$v_{subject} + R$に正しい単語が最も近い場合を正解とする場合と，正しい単語が 100 番目以内に出現する場合を正解とする場合の両方の場合について行われた. なお，比較対象として，単純に関係ベクトルを，次のように平均で定めた場合[市瀬 15]も同様に実験を行った.

$$R = \overline{x} = \frac{\sum_{i=1}^{N}(v_{o_i} - v_{s_i})}{N}$$

また，提案手法を実際に予測器として用いるためには，Uの次元を前もって定めておく必要がある. そこで，交差検定のそれぞれのステップにおいて，訓練集合に対して 10 点交差検定を再び行い，そこで最も精度のよいUの次元を用いて評価を行うようにした. これが実験 2 である. 実験プロセスを図 2 に示した. なお，正解率の突発的なぶれに対応するため，次元を定める際に用いるための精度を示す評価値として，前後それぞれ 2 つの次元も合わせた 5 つの次元の正解率の和を用いた. 例えば，次元 10 の評価値は，次元が 8, 9, 10, 11, 12 の場合の正答率の和である.

提案手法の実装の際の固有値分解のアルゴリズムとして，Jacobi eigenvalue algorithm を用いた.

5.2 実験結果

実験 1 の結果を表 1 に示す. 比較対象は，Uの次元が 0 の場合に相当する. CBOW1，では，Uの次元を 50 にすると正解率が約 11%上昇している. CBOW100 でも 5%以上の上昇が見られる. Skip-gram1 ではUの次元を 120 としたときに約 12%の改善がみられ，Skip-gram100 でもUの次元を 70 か 120 にしたときに 6%以上の改善がみられる. このことから，関係を面として表現することは従来の手法より有効であることが示された. 全体的に，CBOW を使用するよりも Skip-gram を使用した場合のほうが正解率はよくなったが，どちらのモデルにおいても提案手法は正解率を改善した.

比較対象も，それなりに高い正解率が得られていることがわかる. これは，分散表現において都市周辺の単語の密度が国周辺の単語の密度より十分に低いことが理由であると考えられる.

また，正解率はUの次元を増やすと急激に増加したあと，安定し緩やかに下降していっている. Uの次元を増やすということは，分布表現の無駄な成分をなくして次元を減らすことである. このことから，多くの成分は，特定の関係の表現に，逆に悪影響を与えており，一定数の無関係な要素を提案手法により削除することにより，正解率が上がったと考えられる.

次に，実験 2 の結果を表 2 に示す. Uの次元を訓練集合から定めた場合にも，CBOW1，Skip-gram1 では 10%程度の改善，CBOW100，Skip-gram100 では 5%程度の改善が見られ，十分

[1] https://code.google.com/archive/p/word2vec/

表1:実験 1 の結果. 値は、正解率である. CBOW, Skip-gram のあとの数字nは、正しい単語が、予測ベクトルからn番目以内に存在する場合に正解としたものを表す. 例えば, CBOW 1 は予測ベクトルから1番近い単語が正しい場合に正解とする. 各列において, 最もスコアの良い物を太字で示した. 紙面の都合上, 次元は 10 刻みで示した.

Uの次元	CBOW 1	CBOW 100	Skip-gram 1	Skip-gram 100
0	0.600653	0.905951	0.546843	0.882874
10	0.639042	0.942344	0.606313	0.940493
20	0.662119	0.946226	0.594920	0.934761
30	0.677540	0.948149	0.616038	0.940530
40	0.693033	0.953919	0.621843	0.940530
50	**0.710305**	0.953882	0.616001	0.942453
60	0.698803	0.948149	0.641001	0.946299
70	0.696807	0.953919	0.623730	**0.948186**
80	0.704463	0.953882	0.639151	0.944340
90	0.687192	**0.959652**	0.648766	0.946263
100	0.681459	0.955806	0.648766	0.944376
110	0.693033	0.955806	0.658382	0.946263
120	0.681495	0.946190	**0.662192**	**0.948186**
130	0.687228	0.942380	0.654499	0.946263
140	0.691074	0.940457	0.656495	0.942417
150	0.691074	0.936575	0.648839	0.945417
160	0.666074	0.928919	0.635377	0.940493

表 2: 実験 2 の結果. 訓練集合を交差検定し、そこで最も精度のよかった次元数をテスト時のパラメータとして用いた場合の結果である.

CBOW 1	CBOW 100	Skip-gram 1	Skip-gram 100
0.704415	0.953935	0.639155	0.944338

な効果がある. このことから, 提案手法におけるUの次元を訓練集合から定め, 予測器として用いることも有効であることが示された. ただ, Skip-gram で得られた分散表現を用いた場合の結果は, 表 1 の最も良い正解率と比べて幾分低い正解率となっている.

6. おわりに

単語の分散表現において, 関係をベクトルとして表現することについての問題点と, それに対処するための手法について研究した. 関係を面として表す方法は, 多対1関係である, 都市―国関係の予測について, 精度の大きな改善をもたらした. しかし, 単語同士の関係は他にも多数存在する. それらについても, この提案手法がどこまで上手く動作するのかを確かめることが今後の課題である. また, 異なる設定や, 他のニューラルネットワーク言語モデルから得られた分散表現についても調査する必要がある.

参考文献

[市瀬 15] 市瀬 龍太郎, 荒川 直哉:分散表象とオントロジーの関係, 人工知能学会全国大会(第 29 回), 2I4-OS-17a-5, 2015.

[Bengio 03] Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C.: A Neural Probabilistic Language Model, *Journal of Machine Learning Research*, Vol. 3, pp. 1137-1155, 2003.

[Jauhar 15] Jauhar, K., Dyer, C. and Hovy, E.: Ontologically Grounded Multi-sense Representation Learning for Semantic Vector Space Models, *in Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 683-693, Association for Computational Linguistics, 2015.

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *in Proceedings of Workshop at International Conference on Learning Representations*, 2013.

[Neelakantan 15] Neelakantan, A., Shankar, J., Passos, A. and McCallum, A.: Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space, *arXiv preprint arXiv:1504.06662* 2015.

[Vilnis 14] Vilnis, A. and McCallum, A.: Word Representations via Gaussian Embedding, *arXiv preprint arXiv:1412.6623*, 2014.