概念辞書における子概念からの親概念の分散表現の推定

大野 達也　古宮 嘉那子　佐々木 稔　新納 浩幸
茨城大学　工学部　情報工学科
{ 12t4019g, kanako.komiya.nlp }@vc.ibaraki.ac.jp
{minoru.sasaki.01, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

1　はじめに

本論文では概念辞書の階層情報の子にあたる概念集合からその親にあたる概念の分散表現の推定を試みる。word2vec を用いて子概念の分散表現から推定した分散表現と親概念の分散表現との間でコサイン類似度を求め、手法毎のコサイン類似度を比較する。この方法で、何種類かの手法の中でより良質で精度の良い算出方法を調べた。

2　関連研究

word2vec[1]は、単語の意味を表す分散表現を算出する手法である。word2vec では、コーパスなどの大量のテキストを与えることで、テキスト中に出現する各単語をその単語の前後いくつかの単語から類推する。この時、Skip-gram モデル[1]の考えから周辺の単語のうちのいずれか一つへの重みベクトルを中間層の値として取るようにする。こうして学習したデータから、単語毎の重みを抽出することで単語に対する分散表現を得る。

ここで得た分散表現は意味空間上の一点を指しており、これを用いて意味合いの近い単語や似た使われ方をする単語を分類分けする事が可能である。

また、それぞれの分散表現のベクトル同士は足し引きを行うことができ、そうする事で擬似的に意味合いの足し引きを行うことができる[2]。

3　word2vec を用いた概念辞書上の子ベクトルによる親ベクトルの再現

本論文はコーパス上の語を概念辞書と照らし合わせてその親子関係を把握し、分散表現を使用して低次元のベクトル表現でそれらを表して、親概念を子概念の分散表現を使って表そうとするものである。

3　実験設定

ここでは以下の 4 つの手法を用いて子概念の分散表現から親概念の分散表現を推定する。コーパスの親子関係が図 1 および表 1 のようになっていた場合を例に、各手法を説明する。

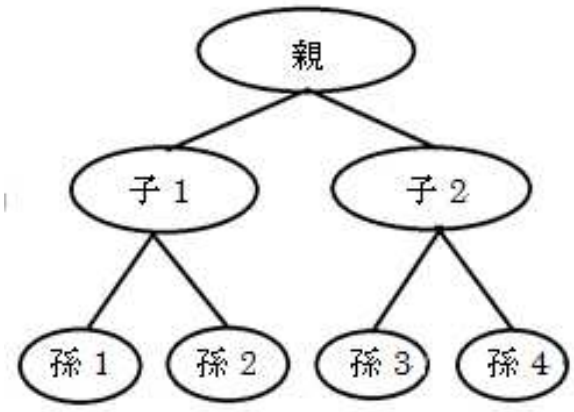

図 1　親子関係の例

[1] http://word2vec.googlecode.com/svn/trunk/

表 1　各概念の出現回数の例

概念	出現回数	概念	出現回数
親	4	孫 2	5
子 1	7	孫 3	3
子 2	2	孫 4	0
孫 1	1		

(1)単純総和

子概念の分散表現のベクトルの総和。ある概念を p と置く。その子概念が n 個あったとする。概念 p の分散表現 v_p は i 番目の子 C_i の分散表現を v_{ci} とすると、以下の式(1)であらわせる。

$$v_p=\sum_{i=1}^{n}\frac{v_{ci}}{n} \quad 式(1)$$

図 1 の例では、親概念の分散表現は子 1 の分散表現と子 2 の分散表現の和として表せる。なお、コサイン類似度を計算する際には、n で割らなくても同じ値になるため、除算は行わなくても良い。

(2)子孫数の重みづけ

子概念の各分散表現のベクトルに「下にいくつの子孫を持つか」で重みを付けた総和。ある概念を p と置く。その子概念が n 個あったとする。また、i 番目の子概念 C_i の子孫にあたる概念の総数を d_{ci} とする。概念 p の分散表現 v_p は子 i の分散表現を v_{ci} とすると、以下の式(2)であらわせる。なお、子孫数 d_{ci} に 1 を加算した上で乗算しているのは子孫数に子概念自体が含まれていないためである。

$$v_p=\sum_{i=1}^{n}\frac{d_{ci}+1}{\sum_{i=1}^{n}d_{ci}+1}v_{ci} \quad 式(2)$$

なお、この際も、(1)と同様に、除算は行わなくてもコサイン類似度は変わらない。そのため、図 1 の例では、子 1 と子 2 がともに二つの子をもつことから、親概念の分散表現は子 1 の分散表現を 3 倍したものと子 2 の分散表現を 3 倍したものの和として表せる。

(3)子概念の出現回数の重みづけ

子概念の各分散表現のベクトルに「その子概念のコーパス中での出現回数」で重みを付けた総和。ある概念を p と置く。その子概念が n 個あったとする。I 番目の子概念 C_i がコーパス中で出てきた回数を a_{ci} とする。概念 p の分散表現 v_p は子 i の分散表現を v_{ci} とすると、以下の式(3)であらわせる。

$$v_p=\sum_{i=1}^{n}\frac{a_{ci}}{\sum_{i=1}^{n}a_{ci}}v_{ci} \quad 式(3)$$

(1)、(2)と同様に、除算は行わなくてもコサイン類似度は変わらない。そのため、表 1 と図 1 の例では、子 1 の概念がコーパス中に 7 回、子 2 の概念がコーパス中に 2 回出現していることから、親概念の分散表現は子 1 の分散表現を 7 倍したものと子 2 の分散表現を 2 倍したものの和として表せる。

(4) 子孫の出現回数の重みづけ

子概念の各分散表現のベクトルに「その子概念ならびにその子孫の概念のコーパス中での出現回数」で重みを付けた総和。ある概念を p と置く。その子概念が n 個あったとする。i 番目の子概念 C_i とその子孫がコーパス中で出てきた回数を as_{ci} とする。概念 p の分散表現 v_p は子 i の分散表現を v_{ci} とすると、以下の式(4)であらわせる。

$$v_p=\sum_{i=1}^{n}\frac{as_{ci}}{\sum_{i=1}^{n}as_{ci}}v_{ci} \quad 式(4)$$

(1)、(2)、(3)と同様に、除算は行わなくてもコサイン類似度は変わらない。ここで、表 1 と図 1 の例では、コーパス中に子 1 の概念が 7 回、子 1 の子である孫 1 の概念が 1 回、孫 2 の概念が 5 回出現しており、子 2 の概念が 2 回、子 2 の子である孫 3 が 3 回、孫 4 が 0

回出現している。つまり、子 1 およびその子孫は合計 13 回、子 2 およびその子孫は合計 5 回コーパス中に出現しているといえる。そのため、親概念の分散表現は子 1 の分散表現を 13 倍したものと子 2 の分散表現を 5 倍したものの和として表せる。

4 実験

本研究の実験ではデータセットとして、EDR の日本語コーパスならびに概念体系[3]を用いる。この概念体系は、EDR 日本語コーパスに対応した語義識別子を用いてその階層関係を表した概念辞書である。この EDR コーパスの出典ごとの文書数と総単語数を表 2 に示す。また、概念辞書の総概念数は 413,153 であり、日本語コーパスから word2vec で作成した分散表現の数は 26,409 であった。

表 2 出典ごとの文書数と総単語数

出典	文書数	総単語数
アエラ	49589	1146084
朝日新聞	91454	2197506
平凡社百科辞典	10072	276703
岩波情報科学辞典	13578	357545
日本経済新聞	5029	119722
用例集	16946	356101
雑誌	21577	516817

また、分散表現の算出には word2vec を用いた。この際、ウインドウの数は±5 とし、ベクトルの要素数は 200 とした。なお、コーパスの単語には概念が付与されているため、単語ではなく、概念の連なりから分散表現の計算を行っている。

EDR の日本語コーパス中に出現するすべての概念を親とし、それぞれコサイン類似度を求めた。この際、すべての子概念が日本語コーパス中で出現して分散表現を生成できているわけではないため、それらの概念はコサイン類似度の計算に含めていない。なお、算出できた全てのコサイン類似度の平均を各手法のスコアとする。

5 結果

実験の結果を表 3 に示す。「単純総和」でのコサイン類似度が最も高く、それ以外の「子孫数の重みづけ」「子概念の出現回数の重みづけ」「子孫の出現回数の重みづけ」のベクトルでのコサイン類似度にはそれほど差は見られなかった。

表 3 各手法での親概念の分散表現ベクトルとのコサイン類似度

手法	コサイン類似度
単純総和	0.279
子孫数の重みづけ	0.242
子概念の出現回数の重みづけ	0.236
子孫の出現回数の重みづけ	0.241

6 考察

表 3 から、算出方法別の親概念の分散表現のベクトルとのコサイン類似度は、直下の子概念のベクトルの総和とのコサイン類似度のみ突出して数値が高く、その他 3 つの算出方法を用いたベクトルのコサイン類似度は軒並み数値が下がったことが分かる。また、精度は単純総和、子孫数の重みづけ、子孫の出現回数の重みづけ、子概念の出現回数の重みづけの順となった。

全体的に子概念から推定した親概念と、実際の親概念の分散表現動詞のコサイン類似度が高くならなかった。これは、必ずしもす

べての子概念の分散表現がword2vecで作成できたわけではなかったため、全ての子概念からその親概念の分散表現を作成できなかったことが原因である可能性がある。word2vecで作成した分散表現は、コーパス中に5回以上出現した概念だけであり、その結果、概念辞書中の全概念数の6.39%となっている。そのため、もっと大きなコーパスが入手できれば、さらに正確な類似度の測定が可能であると考えられる。

7　おわりに

本論文では、親概念の分散表現を子概念の分散表現から推定するために、EDRの日本語コーパスから語義識別子だけを抽出したものをword2vecにかけて、概念をベクトル情報で表現し、それらに重み付けを行い、より良い結果を得られる手法を探った。結果はすべてのベクトルの和をそのまま取る単純総和が最も優れていた。コーパスサイズをもっと大きくした実験が今後の課題である。

謝辞

文部科学省科学研究費補助金[若手B（No：15K16046）]の助成により行われた。ここに、謹んで御礼申し上げる。

参考文献

[1]　T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient Estimation of Word Representations in Vector Space. In ICLRWorkshop.2013.

[2]　吉井 和輝, Eric Nichols, 中野 幹生, 青野 雅樹. 日本語単語ベクトルの構築とその評価.第221回NL・第106回SLP合同研究発表会,pp.1-8,2015.

[3] Hideo Miyoshi, Kenji Sugiyama, Masahiro Kobayashi, and Takano Ogino. An overview of the edr electronic dictionary and the current status of its utilization. In Proceedings of the COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics, pp. 1090–1093. 1996.