類義語を利用した単語の分散表現から語義の分散表現の構築

大内 克之　新納 浩幸　古宮 嘉那子　佐々木 稔
茨城大学 工学部 情報工学科
11t4010l@hcs.ibaraki.ac.jp, hiroyuki.shinnou.0828@vc.ibaraki.ac.jp,
kanako.komiya.nlp@vc.ibaraki.ac.jp, minoru.sasaki.01@vc.ibaraki.ac.jp

1　はじめに

本論文では語義の分散表現の構築方法を提案する。

単語の分散表現とは、その単語の意味を低次元の密なベクトルで表現したものである。従来の bag of words による高次元の疎なベクトルで表現するよりも、よりよく意味を表現できていると考えられる。そのため様々な自然言語処理のタスクに利用され、多くの成果を出している。

語義曖昧性解消のタスクに対しては、通常、教師付き学習手法が用いられる。しかし教師付き学習手法の場合、訓練データの作成コストが高いことから対象とする単語が限定されてしまい、実用的ではないという問題がある。一方、語義の分散表現を求めることができれば、対象単語の文脈のベクトルとどの語義の分散表現が類似しているかを調べることで語義曖昧性解消が実現できる。単語の分散表現はタグなしコーパスから構築できるため、語義の分散表現も同様の手法から構築できれば教師なしの語義曖昧性解消が実現できることになる。このような背景から語義曖昧性解消に関しては語義の分散表現を構築する試みがなされている [3][1]。ここでは語義の分散表現を構築するために、多義語の各語義の分散表現の和が、多義語の分散表現になっていると考える。つまり多義語の分散表現を $\boldsymbol{v}$ とし、その多義語の各語義 $s_i(i = 1 \sim K)$ の分散表現を $\boldsymbol{v}_i$ とするとき、

$$\boldsymbol{v} = \sum_{i=1}^{K} \boldsymbol{v}_i$$

が成立していると考える。そして本論文ではこの関係式を利用して $\boldsymbol{v}$ から $\boldsymbol{v}_i$ を構築する方法を提案する。具体的には s_i の語義を持つ類義語 w_i の分散表現 $\boldsymbol{u}_i$ を利用する。$\boldsymbol{u}_i \approx \boldsymbol{v}_i$ と考えられるため $\boldsymbol{v}_i = \alpha_i \boldsymbol{u}_i$ とし、

$$\boldsymbol{v} = \sum_{i=1}^{K} \alpha_i \boldsymbol{u}_i$$

から最小二乗法により α_i を求めることで $\boldsymbol{v}_i$ を構築する。

実験では BCCWJ コーパス [2] から分散表現を求め、単語「意味」が持つ 3 つの語義の分散表現を構築した。この構築した語義を利用して SemEval-2 の日本語辞書タスク [4] における「意味」のテスト用例の語義曖昧性解消を行い、56.4% の正解率を得た。

2　語義の分散表現による語義曖昧性解消

2.1　提案手法

本論文で提案する手法は、類義語の分散表現を語義の分散表現として利用する手法となる。

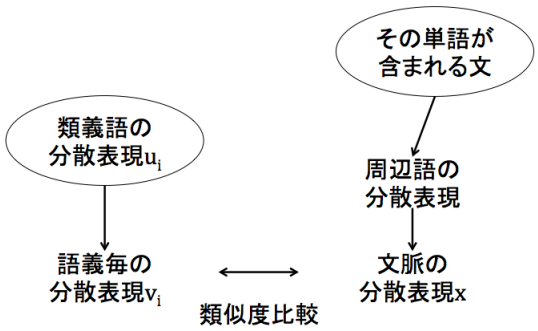

図 1: 本手法イメージ

対象単語の分散表現を $\boldsymbol{v}$ とし、その単語の各語義 $s_i(i = 1 \sim K)$ の分散表現を $\boldsymbol{v}_i$ とするとき、

$$\boldsymbol{v} = \sum_{i=1}^{K} \boldsymbol{v}_i$$

が成立しているとする。この関係式と s_i の語義を持つ類義語 w_i の分散表現 $\boldsymbol{u}_i$ を利用して、$\boldsymbol{v}$ から $\boldsymbol{v}_i$ を

求めていく。

$\boldsymbol{u}_i \approx \boldsymbol{v}_i$ と考えられるため $\boldsymbol{v}_i = \alpha_i \boldsymbol{u}_i$ とし、

$$\boldsymbol{v} = \sum_{i=1}^{K} \alpha_i \boldsymbol{u}_i$$

から最小二乗法により α_i を求めることで $\boldsymbol{v}_i$ を構築する。

2.2 語義と類義語

ここからは「意味」という単語を例にとって、具体的に語義の分散表現を構築していく過程を示す。

岩波辞書において「意味」は以下の 3 つの語義を持つ。

2843-0-0-1 その言葉の表す内容。意義。「辞書を引けば–がわかる」

2843-0-0-2 表現や行為の意図・動機。「どういう–でそんなことをしたのか」

2843-0-0-3 表現や行為のもつ価値。意義。「そんな事をしても–がない」

各語義に対する類義語を以下のように定めた。

2843-0-0-1 「趣旨」「内容」

2843-0-0-2 「目的」「含意」

2843-0-0-3 「価値」「重要性」

2.3 類義語の分散表現

語義曖昧性解消を行うために、語義毎の分散表現を求める必要がある。そのために、先程集めた類義語である、「意味」「趣旨」、「目的」「含意」、「価値」「重要性」の分散表現を利用する。

単語の分散表現は、BCCWJ コーパスから word2vec[1]を用いて構築しておく。次元数は 100 とする。

2.4 語義の分散表現の重み

図 2 は、単語「意味」の分散表現 $\boldsymbol{v}$ とその語義の分散表現 $\boldsymbol{v}_i$、類義語の分散表現 $\boldsymbol{u}_i$ の関係を、便宜的に

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

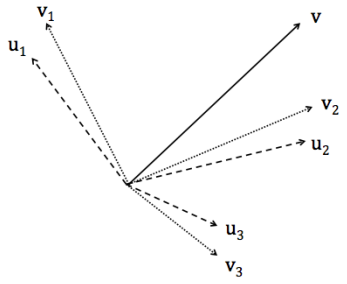

図 2: $\boldsymbol{v} \approx \boldsymbol{u}_1 + \boldsymbol{u}_2 + \boldsymbol{u}_3$

2 次元のベクトルで表したものである。$\boldsymbol{v}$, $\boldsymbol{u}_1$, $\boldsymbol{u}_2$ 及び $\boldsymbol{u}_3$ が既知のベクトルであり、これらから $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ 及び $\boldsymbol{v}_3$ を構築する。

「はじめに」で述べたように、

$$\boldsymbol{v} = \sum_{i=1}^{K} \alpha_i \boldsymbol{u}_i$$

といえる。実際には、ベクトルの大きさが揃っていなければならないため、正規化を行う。ベクトル $\boldsymbol{u}_n$ の要素を $\boldsymbol{u}_{ni}$ のように表すと、次元数は 100 なので、式は

$$u'_{ni} = \frac{u_{ni}}{\sqrt{\sum_{i=1}^{100} u_{ni}}}$$

となる。さらに、「意味」の語義は三つであるため $K = 3$となる。それを踏まえると

$$\boldsymbol{v} = \sum_{i=1}^{3} \alpha_i \boldsymbol{u'}_i$$

のようになる。この α は、実際に単語が使用された際のそれぞれの語義の重みとなる。

この α を求めるために、最小二乗法を用いる。まず、

$$\alpha_1 \boldsymbol{u'}_1 + \alpha_2 \boldsymbol{u'}_2 + \alpha_3 \boldsymbol{u'}_3 - \boldsymbol{u'} = 0$$

といえる。この式を変形して二乗すると、

$$(\alpha_1 \boldsymbol{u'}_1 + \alpha_2 \boldsymbol{u'}_2 + \alpha_3 \boldsymbol{u'}_3)^2 - \boldsymbol{u'}^2 = 0$$

となり、上の式の $\alpha_1, \alpha_2, \alpha_3$ に対して偏微分を行うことで、

$$\begin{pmatrix} |\boldsymbol{u'}_1{}^2| & |\boldsymbol{u'}_1 \boldsymbol{u'}_2| & |\boldsymbol{u'}_1 \boldsymbol{u'}_3| \\ |\boldsymbol{u'}_1 \boldsymbol{u'}_2| & |\boldsymbol{u'}_2{}^2| & |\boldsymbol{u'}_2 \boldsymbol{u'}_3| \\ |\boldsymbol{u'}_1 \boldsymbol{u'}_3| & |\boldsymbol{u'}_2 \boldsymbol{u'}_3| & |\boldsymbol{u'}_3{}^2| \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} |\boldsymbol{u'}_1 \boldsymbol{u'}| \\ |\boldsymbol{u'}_2 \boldsymbol{u'}| \\ |\boldsymbol{u'}_3 \boldsymbol{u'}| \end{pmatrix}$$

という式が得られる。これを解くことで、$\alpha_1, \alpha_2, \alpha_3$ を求めることができる。

2.5 文脈の分散表現

実際に意味という単語が含まれる文脈の分散表現を求める。

文脈の分散表現を得るために、周辺語の分散表現を利用する。そのため、まず、対象単語「意味」の周辺の自立語を取り出す。

以下の例で考える。

「そんなことに意味は無い。」

上記の文から「そんな」「事」「無い」の三つの自立語を取り出せる。これらの単語の分散表現を $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}$ とすると、文脈の分散表現 $\boldsymbol{x}$ を周辺語の分散表現の平均に設定し、$\boldsymbol{x}$ は以下から得られる。

$$\boldsymbol{x} = \frac{\boldsymbol{x_1} + \boldsymbol{x_2} + \boldsymbol{x_3}}{3}$$

2.6 語義曖昧性解消

ここまでに求めた値を用いて、語義曖昧性解消を行う。まず、類義語の分散表現と文脈の分散表現の類似度をを求める。ここではコサイン類似度を用いるため、分散表現 $\boldsymbol{x}$ を正規化しておく。次元数は 100 なので、式は

$$x'_i = \frac{x'_i}{\sqrt{\sum_{i=1}^{100} x'_i}}$$

となる。新しく出来た文脈の分散表現と、類義語の分散表現をそれぞれ $\boldsymbol{x'}$ とすると、コサイン類似度は、

$$cos(\boldsymbol{x'}, \boldsymbol{u'}_n) = \sum_{i=1}^{100} x'_i u'_{ni}$$

となり、さらに

$$c_n = cos(\boldsymbol{x'}, \boldsymbol{u'}_n)\alpha_n$$

として重み付けをする。この c_n が最も大きい n が識別結果となる。

3 実験

実験として、単語「意味」を使った語義曖昧性解消を行う。テストデータとして、SemEval-2 の日本語辞書タスクの、「意味」のテストデータを用いた[2]。テストデータ 50 個のうち、文脈の分散表現が構築できたのは 39 個であり、これらに対して語義曖昧性解消を行った。

[2]baseline システム（教師あり手法）での単語「意味」の正解率は約 38 %である。

実験では、重み付けをした場合としていない場合で行った。また全ての類義語を用いるだけではなく、類義語の組み合わせ全てを試し、最も結果の良かった組み合わせと、最も悪かった組み合わせのそれぞれの正解率を求めた。

重み付けをして語義曖昧性解消を行った結果は、

2843-0-0-1 「趣旨」
2843-0-0-2 「目的」
2843-0-0-3 「価値」「重要性」

の組み合わせで最高の 56.4% 、

2843-0-0-1 「内容」
2843-0-0-2 「含意」
2843-0-0-3 「価値」

の組み合わせで最低の 23.1% となった。

重み付けをせず語義曖昧性解消を行った結果は、

2843-0-0-1 「趣旨」
2843-0-0-2 「目的」「含意」
2843-0-0-3 「重要性」

の組み合わせで最高の 38.5% 、

2843-0-0-1 「趣旨」「内容」
2843-0-0-2 「目的」「含意」
2843-0-0-3 「価値」「重要性」

の組み合わせで最低の 17.9% となった。

4 考察

ここで構築した語義の分散表現が適切であるかどうかの評価は難しい。ただし実験では重み付けが有効であった。重み付けを行わないというのは単に類義語との類似性から語義を判定していることに対応する。また重み付けを行うというのは語義の分散表現を求めていることに対応する。このことから考えると、得られた語義の分散表現が類義語の分散表現以上には適切であったと考えられる。

本手法の問題点としては類義語を見つけることの困難性がある。ここで題材とした単語「意味」は各語義に対して類義語が見つけられたが、このような単語は

稀である。また、今回集めた類義語においても、似ているだけで完全に一致しているとは言い難い。「内容」にしても、「目的」にしても多義語なので、「意味」と共通する語義を持ってはいるが、全く同じであるとは言えない。

語義の分散表現を見つける場合、語義の類義語が見つかれば、それは大きな手がかりとなるが、それは困難である。そのため、辞書の例文から文脈の分散表現を構築していく方向が現実的な手法と考えている。今後は辞書の例文を利用する方法を検討したい。

5 おわりに

本論文では、類義語を利用した語義の分散表現の構築方法を提案した。そして構築した分散表現を利用して、教師なしの語義曖昧性解消を行った。

実験では SemEval-2 の日本語辞書タスクでの単語「意味」のテストデータを用いて、重み付けをした場合（語義の分散表現を利用）としていない場合（類義語の分散表現を利用）との正解率を求めた。それぞれの最高値は、重み付けをした場合 56.4%、しなかった場合 38.5%となり、語義の分散表現が適切に構築されたと考えられる。

本手法の問題点は語義の類義語を見つけることが困難なことである。今後は辞書の例文を利用する方法を検討したい。

参考文献

[1] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A Unified Model for Word Sense Representation and Disambiguation. In *EMNLP-2014*, pp. 1025–1035, 2014.

[2] Kikuo Maekawa. Design of a Balanced Corpus of Contemporary Written Japanese. In *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58, 2007.

[3] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In *EMNLP-2014*, pp. 1059–1069, 2014.

[4] Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. On SemEval-2010 Japanese WSD Task. 自然言語処理, Vol. 18, No. 3, pp. 293–307, 2011.