概念識別子の分散表現を利用した語義曖昧性の解消

情報科学科　内藤 弘章　　指導教員：山村 毅

1　はじめに

語義曖昧性の解消とは、文中などで出現したある特定の単語がその文脈の中でどのような意味で用いられているのかを判別する自然言語処理のタスクである。近年では、word embedding[1] を用いて語義曖昧性の解消の研究が行われるようになった。word embedding は、単語をベクトル表現したものであり、意味が似ていれば似ているほど近くに配置され、単語の意味表現としても用いられている。菅原ら [2] は単語の分散表現を語義曖昧性の解消の素性として用い、word embedding が語義曖昧性の解消に与える有効性を明らかにした。この研究では、素性として「単語の表層表現」の分散表現を用いているが、単語の表層表現以外についての分散表現が語義曖昧性の解消に与える影響を調べることは価値があることである、と考えられる。

本研究では、EDR 辞書データに登録されている、「概念識別子」の分散表現を用いて、語義曖昧性の解消を試みる。また、対象の単語の分散表現を求める際に用いる概念識別子の分散表現の適切な範囲も求める。

2　語義曖昧性解消手法

2.1　word2vec

word2vec は、Mikolov[1] らが提案した word embedding(単語の分散表現) の獲得手法である。テキストからニューラルネットワークを用いて学習し、入力層から中間層への重みを抽出することで単語のベクトルを求めることができる。

例えば、”王”－”男”+”女”=”女王”のように、word2vec を用いることによって、単語の意味に対する加減演算を行うことが可能になる。

2.2　提案手法

本研究では、「概念識別子」の分散表現を利用して語義曖昧性の解消を試みる。概念識別子は、文脈で用いられている概念 (意味) がわかる識別子である。そのため、この概念識別子を単語の代わりに用いて、word embedding を学習させることによって、語義曖昧性の解消の精度向上が期待できる。本研究で提案する語義曖昧性解消の流れを以下に示す。

1. コーパスの学習データを概念識別子のみで表し、「概念識別子」の分散表現を獲得する。
2. コーパスのテストデータで語義曖昧性解消の対象となる単語を定める。
3. テストデータの対象の単語以外を概念識別子で表し、対象の単語はそのまま日本語で表す。
4. CBoW モデルを用いて対象の単語のベクトルを算出する。
5. テストで算出した概念識別子が正しいものかどうか精度を比較する。

CBoW モデルは、入力層の平均を取り、それを中間層に入力しているため、文脈となる単語の平均から対象単語の語義がわかると考えた。

3　評価実験

3.1　使用データ

今回の実験では EDR 電子化辞書の「日本語コーパス」と「日本語単語辞書」を用いる。日本語単語辞書は日本語単語辞書レコードを単語見出しの読み順に並べたものであり、約 26 万語収集されている。日本語コーパスは約 20 万文収集されており、既に分かち書きも行われているコーパスである。

3.2　実験方法

日本語コーパスは、約 20 万文のうち、18 万文を学習データ、約 2 万文をテストデータとして使用する。まず word2vec を用いて、学習データで概念識別子の分散表現を学習し、テストデータを用いて語義曖昧性解消手法の評価を行う。この際、対象の単語は文の中心付近から任意に選んだ。また、次元は 200, 文脈の最大単語数は 5($N = 1 \sim 5$),6($N = 6$),7($N = 7$) とした。

4　実験結果

今回の実験結果を以下の表にまとめる。尚、合計試行回数は 23808 回である。

表 1　実験結果

N	正解確率 [%]	N	正解確率 [%]
1	44.2	5	49.7
2	47.3	6	47.0
3	48.9	7	34.5
4	49.2		

対象の単語 1 単語あたりの平均の概念識別子の個数は、3.49 個である。そのため、今回のテストデータの正解する期待値は 28.7% であるということがわかる。

今回の結果では最も精度が低い $N = 1$ のときでもその期待値を大きく上回る結果となった。表 1 を見てわかるとおり、対象の単語の前後 5 文字のときが一番精度が高かった。

5　まとめと考察

4 章より、精度が高い $N = 5$ の時でも正解確率が 50% 弱という実験結果になった。これは、前後 N 単語の概念識別子の平均ベクトルと対象の単語の概念識別子のベクトルが非線形であったからであると考える。しかし本研究の実験結果から、概念識別子の分散表現を用いて語義曖昧性を解消することの可能性を示すことができた。

参考文献

[1] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, “ Efficient Estimation of Word Represention in Vector Space ”, International Conference on Learning Representations Workshop, 2013

[2] 菅原拓夢、笹野遼平、高村大也、奥村学『単語の分散表現を用いた語義曖昧性の解消』言語処理学会第 21 回年次大会発表論文集、pp.648-651、2015