固有表現抽出を用いた, Word2vec による語義曖昧性解消の精度向上

情報科学科　宮崎 智也　　指導教員：山村 毅

1　はじめに

語義曖昧性解消とは, 自然言語処理において, 文書に現れたある特定の単語やフレーズが，その文脈の中で, どのような語義を持っているか判断するための，一つのタスクである.

EDR 日本語コーパスは, 辞典や新聞記事, 雑誌などから収集した解析済み文データで, およそ 20 万の文が含まれている. 形態素情報, 構文情報に加えて, 各単語に意味を表す概念識別子が付けられている. 太田ら [1] は, この概念識別子を単語の代わりに用いて，文を (概念識別子の列として) 表現し, これを, Word2Vec で学習させて, 各概念識別子の分散表現を求めた. そして, この分散表現を利用して, 語義曖昧性解消の対象単語の概念識別子をその周辺の単語の概念識別子から求めることで, 語義曖昧性の解消を行った. 語義推定精度は 59 %程度であり, EDR 日本語コーパス中の単語で概念識別子が付けられていないものがあることが大きな問題であった.

本研究では, 太田らの研究における課題点であった『省いた素性に対しての処理』について, 『固有表現の抽出』を用いることで, 語義曖昧性解消の精度向上を目指すこととする.

2　提案手法

前節で述べたように, EDR 日本語コーパスは, 各単語に意味を表す概念識別子が付けられているが, 一部付けられていないものもあった. このため, 語義推定に用いる周辺情報が不足することになり, 結果として語義曖昧性解消精度の低下を招いたと考えられる. したがって, もし, それらに対し概念を表すラベルを新たに与えることができれば, 語義曖昧性解消の精度を向上させることができると考えられる.

そこで, 日本語自然言語処理ツール GiNZA の固有表現抽出機能を利用して, 概念識別子が与えられてない単語に固有表現ラベルを与え, これをこれまでの概念識別子に加え, 新たな概念識別子として捉え, 太田らの方法で語義曖昧性解消を行う. 具体的には, 概念識別子が与えられていなかった単語に対し, 『Person, Company, Country, Province, City, Island, Time, Date, Money, Percent, Position_Vocation 』の 11 個の固有表現ラベルを与えた.

3　評価実験

EDR 日本語コーパス中の概念識別子のない名詞 241,641 個に固有表現ラベル与え, それを実験データに用いて，太田らの方法で語義曖昧性解消実験を行う. ただし, EDR 日本語コーパスによる単語の取り扱いと GiNZA による単語の取り扱いとに差があるため, EDR 日本語コーパスの文に直接 GiNZA で固有表現抽出処理を行なった場合, 概念識別子が付けられていない単語に固有表現ラベルが正確に付けられないことがあった. そこで, 次の 3 つの方法を試みた.

(方法 1) 正確に固有表現ラベルが付けられたものだけ使用する.

(方法 2) 概念識別が付けられていない単語に単独で GiNZA で固有表現ラベルを付与する.

(方法 3) 方法 1 に加えて, 正確に付けられなかった単語に単独で GiNZA で固有表現ラベルを付与する. ここで , 「単独」というのは, 単語一つだけで固有表現判定するものである (通常は周辺の単語を用いて固有表現判定をするので, 単独の場合, 固有表現判定精度は悪くなる).

Word2Vec のパラメータは次元数 200，学習の窓幅 2，学習の反復回数 10 とし，3 回未満しか登場しない概念識別子については学習しないこととした. 曖昧性解消の対象単語は, 文の中心に最も近い多義語の名詞とした. 対象とした単語は 167,801 個であった. なお, 精度の検証方法としては 10 分割交差検証を用いた.

4　結果

表 1 に実験結果を示す. この表で『与えた数』は, 固有表現ラベルを概念識別子として付与した個数を示す. 従来の手法と比較して, 約 1 ポイントの上昇が見られた.

手法	与えた数 [個]	正解率 [%]
従来手法	ー	59.44
方法 1	27,154	60.33
方法 2	50,281	60.45
方法 3	54,485	60.49

表 1　固有表現ラベルの付与数と正解率

従来研究と本研究の正解率に差があるかどうかを，フィッシャーの正確検定を用いて検定したところ, 有意水準 5 %で差があるという結果が得られた.

5　考察と今後の課題

本研究では，固有表現ラベルを新たな概念識別子として用いることで，太田らの研究における語義曖昧性解消手法の精度向上を試みた.

表 1 に示すように, 固有表現抽出により与えるラベルの数が増えるにつれて, 少しづつながら精度が上昇している. このことから, 固有表現に対して与えるラベルの数が多いほど, より高い精度を得られると考えられる.

本研究では, 先行研究に対して, 有意差はあったが, 正解率は 60 %程度で, まだ十分とは言えない. これは, 固有表現 241,641 個のうち, 54,485 個に対して新たにラベルを与えたが, まだラベルを与えることのできていない固有表現が多くあるためだと考えられる.

今後の課題として, より多くの概念識別子を持たない単語に対しラベルを与えることで更なる精度の向上ができるのではないかと見込んでいる.

参考文献

[1] 太田 剛貴, 加藤 敦也, 山村 毅 : ” Word2vec を用いた非線形的に予測する語義曖昧性解消”, 電気・電子・情報関係学会東海支部連合大会講演論文集, K3-3, 2020.