『岩波国語辞典』の語義タグを用いた all-words の語義曖昧性解消

平林 照雄　鈴木 類　古宮 嘉那子　浅原 正幸　佐々木 稔　新納 浩幸
茨城大学工学部情報工学科
人間文化研究機構 国立国語研究所
{14t4051n, 17nm709g, kanako.komiya.nlp, minoru.sasaki.01,
hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp
masayu-a@ninjal.ac.jp

1　はじめに

all-words の語義曖昧性解消とは、文章中の全多義語の語義を一意に決定するタスクである。鈴木ら [1] は、コーパス中の全多義語に、シソーラス『分類語彙表』で定義されている分類番号を一意に付与した。

本研究では、鈴木らの手法 [1] に基づき、『岩波国語辞典』の語義が一部付与されたコーパスに対して、『分類語彙表』の分類番号を対象に all-words の語義曖昧性解消の実験を行った。この際、『岩波国語辞典』の語義を『分類語彙表』の分類番号の代わりに疑似的な正解データとして利用する。その上で、『岩波国語辞典』の語義の、『分類語彙表』の分類番号を付与する語義曖昧性解消における影響を調査した。

2　関連研究

語義曖昧性解消の手法は、大きく教師あり学習、教師なし学習、半教師あり学習の三つに分けることができる。教師なし学習における all-words の日本語語義曖昧性解消の関連研究には古宮らの研究 [2] や新納らの研究 [3] がある。古宮らの研究では、多義語の周辺に現れる語義の分布を利用する周辺語義モデルを提案している。また、新納らの研究では、単語分割をするテキスト解析のツールキットを応用し、all-words の日本語後語義曖昧性解消を簡易に行えるシステムを提案している。本研究は、『分類語彙表』のタグ付きコーパスを利用していないという意味では、教師なし学習である。一方、『岩波国語辞典』の語義を疑似正解データとして利用しており、疑似正解データを与えた半教師あり学習とみることもできる。（我々の知る限り、本論文は語義曖昧性解消に別の辞書の語義を疑似正解データとする初の論文である。）

3　『分類語彙表』と『岩波国語辞典』

3.1　『分類語彙表』

『分類語彙表』とは、語を意味によって分類・整理したシソーラス（類義語集）である。[1] 一つのレコードは「レコード ID 番号／見出し番号／レコード種別／類／部門／中項目／分類項目／分類番号／段落番号／小段落番号／語番号／見出し／見出し本体／読み／逆読み」という要素から構成される。分類番号は「類／部門／中項目／分類項目」を表す 5 桁からなる数字である。例えば「犬」という言葉は、『分類語彙表』では 2 箇所に登録されている多義語である。それぞれの分類番号は 1.2410、1.5501 であり、表 1 のように分類されている。

表 1: 『分類語彙表』における「犬」

分類番号	類	部門	中項目	分類項目
1.2410	体	主体	成員	専門的・技術的職業
1.5501	体	自然	動物	哺乳類

3.2　『岩波国語辞典』における語義タグ

『岩波国語辞典』では、単語の語義を特定するために “2712-0-0-1-0” のような語義タグが付与されている。これは [見出し ID]-[複合語 ID]-[大分類 ID]-[中分

[1] http://pj.ninjal.ac.jp/corpus_center/goihyo.html

類 ID]-[小分類 ID] を並べたもので、分類に対応がないとき ID の値は 0 である。例えば、犬は “2712-0-0-1-0”、“2712-0-0-2-0”、“2712-0-0-3-0” の三つの語義が存在し、語義の粒度は必ずしも分類語彙表とは対応しない。

4 提案手法

本研究では、鈴木らによる語義曖昧性解消を行った後、『分類語彙表』の分類番号と『岩波国語辞典』の語義タグが複数回対応がとれ、かつ設定した閾値以上の対応がとれた時、語義タグを基に分類番号の更新を行い、再び鈴木らによる語義曖昧性解消を行った際の正解率の変化を調べた。また、岩波国語辞典の語義タグは、コーパスの一部に付与されていたため、残りの単語はツールを使って自動的に付与した。

4.1 鈴木らの手法

鈴木らの手法では「周辺単語ベクトル」と「類義語」を以下のように定義する。

- 対象単語の周辺単語ベクトル:word2vec [2] を用い、対象単語の前後 2 単語の分散表現を求め、連結したもの。
- 類義語:語義曖昧性解消を行う多義語と分類番号が等しい単義語。

これを基に、1 周目では、以下のように語義曖昧性解消を行う。

1. 語義曖昧性解消の対象単語の類義語をコーパス中から探し、その周辺単語ベクトルを求める。
2. 1 の類義語の周辺単語ベクトルと、語義曖昧性解消の対象単語の周辺単語ベクトルとの距離を測り、K 近傍法 (K-NN) により『分類語彙表』の分類番号をラベルとして多義語に付与する。

n 周目 ($n \in \mathbb{N} \wedge\ n > 1$) は以下のように語義曖昧性解消を行う。

1. (n-1) 周目の結果を基にコーパスを概念（分類番号）の分かち書きに変換し、word2vec を用いて概念の分散表現 (concept2vec) を作成する。
2. n 周目では、対象単語の前後 2 単語の concept2vec を対象単語の「周辺単語ベクトル」とする。
3. 語義曖昧性解消の対象単語の類義語をコーパス中から探し、その周辺単語ベクトルを求める。
4. 3 の類義語の周辺単語ベクトルと、語義曖昧性解消の対象単語の周辺単語ベクトルとの距離を測り、K 近傍法 (K-NN) により『分類語彙表』の分類番号をラベルとして多義語に付与する。

4.2 コーパスの増補

『分類語彙表』の分類番号と『岩波国語辞典』の語義タグの対応をとるためには双方のタグが付いたコーパスが必要であるが、これらのコーパスを大量に用意するのは困難である。そこで、分類番号が付与された『現代日本語書き言葉均衡コーパス』[4] に、新納ら [3] が提案したシステム、all-words の語義曖昧性解消モデル KyWSD を適用し、コーパスの増補を行った。

5 実験

5.1 実験設定

コーパスとして使用する分類番号が付与された『現代日本語書き言葉均衡コーパス』[4] は、単語のべ数 22,568、単語異なり数 3,790 からなるもので、この中に多義語は 4,760 単語（のべ数）存在する。多義語の平均語義数は 3.16 であり、ランダムに語義を割り当てた場合の正解率は 31.7%である。KyWSD によるコーパスの増補においては、平文を短単位にあらかじめ分割し、KyWSD を適用した。また鈴木らの設定と同様に、単語の分散表現は NWJC2vec[5] を用いた。

概念の分散表現 (concept2vec) は、コーパスを分類番号の分かち書きに変換したものを word2vec で学習して作成したベクトルを用いた。その際、アルゴリズムは C-BoW を利用し、次元数を 50、ウィンドウ幅を 5、ネガティブサンプリングに使用する単語数を 5、反復回数を 3、min-count を 1 として学習を行った。また周辺単語ベクトルを作成する際、周辺に単語が四つない場合（対象単語が文頭や文末にある場合など）や、word2vec で学習されていない単語の分散表現などは、同じ次元の零行列を用いた。

周辺単語ベクトルの距離を測り K 近傍法で分類する過程では、sckit-learn [3] の KNeighborsClassfier を使用した。ここではユークリッド距離を使用し、かつ最良の実験結果が得られた k=3、weight=uniform

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

[3] http://scikit-learn.org/stable/

(uniform=重みなし) で実験を行った。

鈴木らの語義曖昧性解消は 2 周行った際が最良であるため、鈴木らの語義曖昧性解消を 2 周行った。この時の状態を with c2v とする。この後、『分類語彙表』の分類番号と『岩波国語辞典』の語義タグが、2 回以上、かつ閾値 80 %または 90 %対応が取れた際、語義タグを基に分類番号の更新を行った。この時の状態を with iwanami 1 とする。with iwanami 1 から再び同じ実験設定で鈴木らの語義曖昧性解消を 1 周回した。この時の状態を with c2v&iwanami 1 とする。この後、再び分類番号と語義タグの対応をとり、分類番号の更新を行う。この時の状態を同様に with iwanami 2 とし、同様に with c2v&iwanami 2 等を定義する。これを with c2v&iwanami 5 を定義するまで行い、このサイクルを 10 回行った時の平均をとった。

また、参考として、鈴木らの手法において、語義曖昧性解消の対象コーパスの一部に正解の分類番号を付与した場合の実験を行った。この際、1 周目については、鈴木らの手法と等しい手順とし、2 周目以降で正解の分類番号を一部与えた。具体的には、手順 1 において (n-1) 周目の結果を基にコーパスを分類番号の分かち書きに変換する際に、一部（コーパス中の単語の 1/5 または 4/5）を正解の分類番号に書き換えた。提案手法は正解データを利用しない手法であるが、比較手法では正解データ利用している点に留意されたい。なお、評価の際は正解データを与えていない部分のみの正解率を求めて比較した。

5.2 実験結果

KyWSD によるコーパスの増補では 6,515 単語に付与された。またそのうちの 315 単語について人手でつけたデータと比較すると、正解率 59.0 %であった。

鈴木らの語義曖昧性解消を行った後、閾値 80 %で分類番号の更新を 5 回行った時の結果を表 2 に示す。表 2 の更新平均分類番号数はのべ数である。

同様に鈴木らの語義曖昧性解消を行った後、閾値 90 %で分類番号の更新を 5 回行った時の結果を表 3 に示す。表 3 の更新平均分類番号数はのべ数である。

比較実験として、コーパスの一部を教師データとして与えた半教師あり学習の結果と教師なし学習の結果を、表 4 に示す。

表 2: 閾値 80 %での分類番号の更新結果

状態	正解率	更新平均分類番号数
w/c2v	58.2	
w/iwanami 1	58.2	45.2
w/c2v&iwanami 1	56.9	
w/iwanami 2	56.9	52.7
w/c2v&iwanami 2	58.0	
w/iwanami 3	58.0	46.1
w/c2v&iwanami 3	57.4	
w/iwanami 4	57.3	47.9
w/c2v&iwanami 4	57.5	
w/iwanami 5	57.5	46.4
w/c2v&iwanami 5	56.8	

表 3: 閾値 90 %での分類番号の更新結果

状態	正解率	更新平均分類番号数
w/c2v	58.3	
w/iwanami 1	58.3	3.0
w/c2v&iwanami 1	57.5	
w/iwanami 2	57.5	2.9
w/c2v&iwanami 2	57.3	
w/iwanami 3	57.3	3.1
w/c2v&iwanami 3	57.2	
w/iwanami 4	57.2	2.8
w/c2v&iwanami 4	57.5	
w/iwanami 5	57.5	3.2
w/c2v&iwanami 5	56.5	

5.3 考察

表 2 表 3 より、『分類語彙表』の分類番号を『岩波国語辞典』の語義タグと対応させて、分類番号の更新を行った場合、閾値 80 %、90 %いずれの場合でもほとんど変化しなかった。その結果から concept2vec を作成すると、語義タグを使用しなかった場合の正解率を下回ったことがわかる。一方、表 4 から、比較手法ではわずかに正解率が上昇していることがわかる。

これらの原因として、以下の三点が考えられる。ひとつは、疑似正解データの正解率の低さである。今回、KyWSD により作成した疑似正解データの正解率をタグ付けされている部分だけで求めたところ、59 %であった。この正解率を上げられれば、もっと正確な疑似正解データを与えられることができると考えられる。現在は、岩波国語辞典の単語の単位とコーパスの単語の単位にズレがあることを考慮して、分かち書きされた文書を KyWSD の入力としたために正解率が低くなっている。そのため、今後は平文で KyWSD に入力した後、うまく単語の単位に合わせて疑似データを与えることができるようにする必要がある。

ふたつめは、閾値の妥当性である。閾値 80 %、90

表 4: 半教師あり学習と教師なし学習の正解率

教師データ	あり	なし	あり	なし
繰り返し回数	教師データがコーパス全体の 1/5		教師データがコーパス全体の 4/5	
1	56.0%	54.8%	57.5%	55.0%
2	57.8%	58.4%	57.5%	58.8%
3	57.6%	57.1%	57.6%	57.5%
4	57.4%	58.0%	57.7%	57.6%
5	58.1%	56.8%	58.0%	57.4%
6	57.5%	57.1%	57.8%	58.0%

%での分類番号の更新を行うには、語義タグと分類番号の対応が完全に取れた時、分類番号の更新を行う必要がないことから、それぞれ語義タグと分類番号の対応が 4 回以上、9 回以上取れる必要があった。そのため、出現回数の多い単語のみ、書き換えが行われた。疑似正解データがもっと正確になってから、どのような閾値がよいか分析する必要がある。

みっつめは、分類番号と語義タグの対応関係である。閾値 80 %の時、平均して 47.7 の分類番号の更新が行われたが、正解率はほとんど変化しなかった。これは更新された分類番号には正解と不正解のデータがそれぞれ同程度存在したためであると考えられる。

また比較手法でも鈴木らの手法を平均 0.3 %上昇させることにとどまった。このことから少量の正解のデータを与えるだけでは正解率の明確な向上が見られないことがわかる。したがって今後の研究では、疑似的な正解を与えるだけではなく効率的にその知識を利用するアルゴリズムが必要であると考えられる。

6 おわりに

本研究では、鈴木ら [1] の『分類語彙表』の類義語と分散表現を利用した all-words の語義曖昧性解消に基づき、『岩波国語辞典』の語義タグを利用した all-words の語義曖昧性解消の実験を行った。またコーパスの一部を教師データとして与えることによる正解率の変化を実験して比較した。

その結果、提案手法では、正解率が上がらなかった。原因として、疑似正解データの正解率の低さが考えられる。また、正解データを一部与えた比較手法でも正解率の上昇はわずかだったことから、少量の正解データによる正解率の向上のためには疑似的な正解タグの精度だけではなく、効率的に正解データを利用できるアルゴリズムが重要であるということが考察された。

謝辞

本研究の一部は国立国語研究所の共同研究プロジェクト「all-wordsWSD システムの構築及び分類語彙表と岩波国語辞典の対応表作成への利用」の研究成果を報告したものである。また、本研究は、茨城大学の女性エンパワーメント支援制度補助金および JSPS 科研費 15K16046 の助成を受けたものである。

参考文献

[1] Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki and Hiroyuki Shinnou, All-words Word Sense Disambiguation Using Concept Embeddings, LREC 2018,(to appear), (2018).

[2] Kanako Komiya, Yuto Sasaki, Hajime Morita, Minoru Sasaki, Hiroyuki Shinnou, and Yoshiyuki Kotani, Surrounding Word Sense Model for Japanese All-words Word Sense Disambiguation, PACLIC 2015, pp. 35-43, (2015).

[3] Hiroyuki Shinnou, Kanako Komiya, Minoru Sasaki and Shinsuke Mori, Japanese all-words WSD system using the Kyoto Text Analysis ToolKit, PACLIC 2017, no 11, (2017).

[4] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, Yasuharu Den, Balanced Corpus of Contemporary Written Japanese,Language Resources and Evaluation, Vol.48, pp.345-371, (2014).

[5] 新納 浩幸，浅原 正幸，古宮 嘉那子，佐々木 稔, nwjc2vec:国語研日本語ウェブコーパスから構築した単語の分散表現データ, 自然言語処理, Vol. 24, No. 5，pp. 705-720, (2017).