『分類語彙表』の類義語と分散表現を利用した all-words 語義曖昧性解消

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 ‡
{13t4039t, 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．単語の語義はその周辺の文脈によって決まることから，周辺の単語同士が類似している場合その中心にある語義曖昧性解消の対象単語同士の語義も類似していると考えられる．そこで本研究では，単語の分散表現を用いて対象単語の周辺単語群と対象単語の各語義候補における類義語の周辺単語群の間の距離を測り，その距離を用いて対象単語の語義を予測した．そして，単義語の語義と多義語の予測で得た語義を基にして『分類語彙表』の概念（語義）の分散表現を作成し，“ 単語の分散表現＋概念の分散表現 ”を用いて周辺単語群間の距離を測りなおして再び語義を予測し，さらにこれを繰り返し行った．

『現代日本語書き言葉均衡コーパス』に『分類語彙表』のコードが付与されたコーパスを用いて実験を行ったところ，単語の分散表現のみを用いた予測では 54.2%，単語と概念の分散表現を用いた予測では最大で 59.0%の正解率となった．

2 関連研究

語義曖昧性解消の手法は大きく教師ありと教師なしの二つに分けることができる．一般的に，語義曖昧性解消を教師あり学習によって解決する場合，高い精度を得ることができる．しかしその反面，十分な量の教師データが必要であるためその作成にコストがかかってしまうという問題点がある．一方教師なしの場合，コストは少ないが教師あり学習を用いる場合と同等の精度を出すことは難しい．

語義曖昧性解消に『分類語彙表』を利用する手法は数多く提案されている．中でも，教師あり学習での語義曖昧性解消において,『分類語彙表』のコードや『分類語彙表』から得られる上位概念の単語などを素性として利用することは多い．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 Komiya らの研究 [1] や新納らの研究 [2] がある。Komiya 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。

3 『分類語彙表』の類義語を利用した all-words 語義曖昧性解消

本章では，本研究の手法について説明する．

3.1 単語の分散表現を利用した手法

単語の語義は周辺の単語によって決まることから，周辺の単語同士が類似している場合，その中心にある単語同士の語義も類似している，と考えることができる．本実験の手法はこの考えをもとに以下のような手順で行う．

まず，対象単語の周辺の四つの単語（前後 2 単語ずつ）のそれぞれの単語の分散表現（word2vec : 以下 w2v）を求める．そしてこの四つの分散表現を連結し，一つの分散表現にしたものを「周辺単語ベクトル」とする．次に,『分類語彙表』から対象単語の語義候補ごとに類義語を列挙し，コーパス中に出現する類義語から周辺単語ベクトルを作成する．この際，周辺単語ベ

クトルには類義語の語義（語義候補の語義）をラベル付けしておく．最後に，対象単語と類義語の周辺単語ベクトル間の距離を測り，K 近傍法（K-NN）によって対象単語の周辺単語ベクトルと距離が近い周辺単語ベクトルのラベルを一つ求め，これを対象単語の語義と予測する．

3.2 単語と概念の分散表現を利用した手法

本研究では，3.1 の手法での結果をもとに，さらに多義語の語義の予測を繰り返し行った．本手法の繰り返し回数が n の場合の手順を以下に示す．

まず，n-1 回目の予測で得た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 で概念の分散表現（concept2vec：以下 c2v）を作成する．（繰り返し回数 0 の結果には 3.1 の手法で得た結果を利用する）次に，対象単語，対象単語の類義語の周辺単語ベクトルを作成する．この際，w2v と c2v を連結したものを単語ベクトルとし，周辺の四つの単語の単語ベクトルを連結したものを周辺単語ベクトルとする．最後に，対象単語と類義語の周辺単語ベクトルの距離を測り，3.1 の手法と同様に語義を予測する．本手法では，この操作を何度も繰り返し，精度がどこまで上昇するかを調べた．

4 実験

4.1 『分類語彙表』の類義語

本研究の手法では『分類語彙表』から類義語を求め，利用する．ここでは本研究での類義語の定義を述べる．『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]．『分類語彙表』の項目は，「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」となっている．『分類語彙表』では単語が分類番号によって単分類されており，この分類番号は単語の「類・部門・中項目・分類項目」を表したものである．例えば「犬」という単語は『分類語彙表』では 2 か所に存在し，分類番号はそれぞれ 1.2410, 1.5501 である（表 1）．また，『分類語彙表』には “ 意味的区切り ” が 240 箇所に存在し，分類番号で分類された単語をさらに細かく分類している．

[*1] https://www.ninjal.ac.jp/publication/catalogue/goihyo/

表 1: 分類語彙表の「犬」

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

本研究の 3.1 の手法では，類義語を以下のように定義した．

- 対象単語の語義（分類番号）候補と分類番号が等しく，意味的区切りがある場合は同じ区切り内の単語
- 対象単語の語義候補ごとに類義語が重複した場合，その類義語はどちらの語義の類義語からも除外する（例えば対象単語 X の語義 1 における類義語候補が A・B・C，語義 2 の類義語候補が C・D・E だった場合，どちらからも C を除外する．）

3.2 の手法において，繰り返し回数が n 回目の場合の類義語を次のように定義した．

- n-1 回目の予測において，対象単語の語義候補と等しい分類番号と予測された多義語
- 対象単語の語義候補と等しい分類番号を持つ単義語
- 対象単語の語義候補ごとに類義語が重複した場合，その類義語はどちらの語義の類義語からも除外する

4.2 実験設定

実験には『現代日本語書き言葉均衡コーパス』に『分類語彙表』のコードが付与されたコーパス [3] を用いる．このコーパスは，単語のべ数：20021, 単語異なり数：3488 からなるもので，この中に多義語は 4378 単語（のべ数）存在する．多義語の平均語義数は 3.195 であり，ランダムに語義を割り当てた場合の正解率は 31.3%である．本実験ではこの正解率を baseline とする．

単語の分散表現の作成には『国語研日本語ウェブコーパス（NWJC）』に対して word2vec というツール[*2]でアルゴリズムには Continuous Bag-of-Words(C-BoW) を利用し，次元数を 200，ウィンドウ幅を 8，ネガティブサンプリングに使用する単語数を 25，反復回数を 15，として学習を行ったベクトルファイル（NWJC2vec [4]）を使用した．概念の分散表現は，コーパスを分類

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

番号の分かち書きに変換したものを同じく word2vec で学習して作成したベクトルファイルを用いた．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．したがって，w2v のみで作成した周辺単語ベクトルは 800 次元，w2v+c2v で作成した周辺単語ベクトルは 1000 次元となる．

周辺単語ベクトルの距離を測り K-NN で分類する過程には scikit-learn[*3]の KNeighborsClassifier を使用した．ここではユークリッド距離を使用し，k=1,3,5, weight=uniform, distance（uniform=重みなし，distance=重みあり）で実験を行った．

4.3 実験結果

w2v のみを用いた手法での結果を表 2 に示す．

表 2: w2v を用いた手法の結果

	K=1	K=3	K=5
重みなし	54.0	**54.2**	52.0
重みあり	54.0	52.4	51.8

次に，c2v を組み合わせ繰り返し予測した結果を表 3 に示す．

表 3: w2v+c2v を用いた手法の繰り返し回数と正解率

K	重み	1	2	3	4	5	6
1	-	54.1	57.0	56.8	55.2	55.2	55.1
3	なし	55.3	53.0	53.6	53.3	53.2	53.2
3	あり	**59.0**	56.0	56.3	55.5	56.8	56.9
5	なし	55.1	53.0	52.4	52.5	51.7	51.9
5	あり	56.7	55.7	53.4	53.4	54.5	55.5

(一段目の数字は繰り返し回数を表す)

w2v を利用した予測では常に重みなしでよい精度が得られ，K=3, 重みなしでの 54.2%が最大となった．また，K の値，重みのあり・なしに関わらず，常に 50%上回る，baseline を有意に超す結果となった．c2v を用いた繰り返しの予測では，繰り返し 1 回目，2 回目でさらによい精度が得られることが確認でき，K=3, 重みなり，繰り返し回数 1，の時の 59.0%が最大となった．

4.4 考察

本実験の結果により，w2v のみを利用した予測で得た結果を基に作成した w2v+c2v が有効であることが確認できた．

本手法では w2v のみを利用した予測の精度が w2v+c2v の質を左右する．最初の予測の精度を向上させる工夫の一つとして，類義語をどのように定義するか，という点が挙げられる．前述の実験での類義語の定義は，「分類番号が同じ（意味的区切りを考慮），語義候補間で重複した類義語を除外」である．語義候補間で類義語が重複した場合，まったく同じ周辺単語ベクトルに異なるラベルが付与されたものが作成されてしまい，K-NN での分類の精度が低下してしまうと考え，除外した．

本手法で距離計算に用いる類義語には，現在の条件に ① 意味的区切りを考慮しない ②意味的区切りではなく段落番号を考慮する ③『分類語彙表』における単義語のみ使用する，という条件を追加するなど，様々なバリエーションが考えられる．そこで，① ② ③ を実際に一つずつ類義語の定義に追加して実験し，その結果から本手法の改善点などを考察する．

①を追加した結果，w2v，w2v+c2v のどちらの手法においても精度は低下した．①を類義語の定義に追加した場合，前述の実験と比較して，対象単語の一つの語義に対してより多くの類義語が得られることになる．このため，意味が遠い単語がより多く類義語に含まれてしまい精度が低下したものと考えられる．

②を追加した場合も、どちらの手法でも精度は低下した。段落番号は意味的区切りよりも細かく単語を分類しているため，②を類義語の定義に追加した場合対象単語により近い意味の単語を類義語として扱えることになる．しかし類義語の数は減るため，K-NN での分類の際のデータ量が減り，精度が低下したのだと考えられる．①・②を類義語の定義に追加した結果から，類義語の数は多くしても少なくしても精度が低下してしまうことがわかった．また，類義語の意味の幅と類義語の量の最適なバランスはコーパスの大きさや対象単語によって左右されるものであると考えられる．

一方で ③を類義語の定義に追加した場合，どちらの手法でも精度が向上した（表 4・表 5）．

前述の w2v のみを利用する実験では類義語が多義

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

表 4: w2v を用いた手法の結果（③を類義語の定義に追加した場合）

	K=1	K=3	K=5
重みなし	56.3	54.3	54.8
重みあり	56.3	56.6	**56.7**

表 5: w2v+c2v を用いた手法の繰り返し回数と正解率（③を類義語の定義に追加した場合）

K	重み	1	2	3	4	5	6
1	-	57.0	56.3	56.9	57.1	56.9	57.1
3	なし	56.6	57.1	57.8	57.4	57.4	57.4
3	あり	58.9	59.2	**59.6**	59.3	59.3	59.3
5	なし	57.6	57.7	57.7	57.7	57.7	57.7
5	あり	59.4	59.1	59.2	59.1	59.2	59.1

語・単義語にかかわらず類義語として利用していた．③を類義語の定義に加えることで，類義語の周辺単語ベクトルの質が向上したものと考えられる．また,『分類語彙表』について調べた結果，掲載されている単語の 7 割近くは単義語であることがわかった．このため③を類義語の定義に加えたとしてもそれほど類義語の数が減少せず，精度が向上したのだと推測できる．

5 おわりに

本稿では，対象単語の周辺単語ベクトルと対象単語の類義語の周辺単語ベクトルの距離から対象単語の語義を求める語義曖昧性解消の手法を提案した．また，語義を予測した結果をもとに概念の分散表現を作成し，周辺単語ベクトルに組み合わせ再び語義を求める，という操作を繰り返し行う実験も行った．実験の結果，baseline を超える精度を達成することができ，語義曖昧性解消において有効な手法であることが確認できた．また，概念の分散表現を組み合わせることでさらに精度が上がったが，何度も繰り返すことの効果はなく，精度が上がったのは繰り返し 1，2 回目のみだった．

『分類語彙表』の中のどのような条件の単語を本手法の距離計算に用いるのが最適であるかは，対象単語の語義やコーパスの大きさによって異なると考えられるが，本実験では『分類語彙表』中の単義語のみを使用することで精度を向上させることができた．

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-words WSD 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。

参考文献

[1] 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, (2015.10).

[2] 新納浩幸, 古宮嘉那子, 佐々木稔, 森信介，点推定による日本語 all-words WSD システム KyWSD，情報処理学会　研究報告自然言語処理（NLP），Vol.2016-NL-227 No.2，pp.1-5，(2016,07,29).

[3] 加藤祥，浅原正幸，山崎誠 ,『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号アノテーションの試行，言語処理学会第 23 回年次大会発表論文集（掲載予定）

[4] 浅原正幸，岡照晃，nwjc2vec:『国語研日本語ウェブコーパス』に基づく単語の分散表現データ，言語処理学会第 23 回年次大会発表論文集（掲載予定）