単語の分散表現を用いた語義曖昧性解消

菅原 拓夢† 笹野 遼平‡ 高村 大也‡ 奥村 学‡
†東京工業大学 総合理工学研究科 ‡東京工業大学 精密工学研究所
suga@lr.pi.titech.ac.jp, {sasano,takamura,oku}@pi.titech.ac.jp

1 はじめに

多義語の存在は計算機による自然言語文書の意味解釈において大きな障害となる．このため，語義曖昧性解消 (WSD) は，機械翻訳，文書要約，意見抽出など自然言語処理の様々なタスクにおいて重要であるといえる．たとえば，“His performance was so cool.” という文を日本語に翻訳する場合を考えると，“cool” は「涼しい」や「かっこいい」など複数の語義を持つ多義語であるため，文中の語義の曖昧性解消を行わないと，“cool” の訳として「涼しかった」などの誤った訳が選択される可能性がある．

語義曖昧性解消において，周辺の文脈情報は大きな手がかりとなることが知られている [11]．このため分類器を学習する際には，一般的に周辺に出現した単語の情報を基本的な素性として用いる．しかし，これらは 0 または 1 で素性を表現する離散的な表現手法であるため，学習データに出現しなかった単語を分類の手掛かりとして利用できないという問題がある．

この解決策として各単語をシソーラスなどのクラスで表現する方法，各単語を複数の連続値で構成されるベクトルで表現する方法などが考えられる．単語をベクトルで表現する研究は従来から行われてきたが，近年分散表現 (distributed representation, word embeddings) などと呼ばれる，単語間の特徴をより良く表した連続値のベクトル表現を獲得する手法が提案されている [3]．そこで，本研究では，単語の分散表現を語義曖昧性解消の基本的な素性として用い，その有効性を明らかにする．

2 関連研究

Yarowsky らの報告 [11] のように，機械学習に基づく語義曖昧性解消では，多義語と共起する単語が語義の決定の大きな手がかりとなる．また，多義語の周辺文脈に出現する単語の他にも通常様々な素性が用いられる [9]．

単語のベクトル表現を語義曖昧性解消に用いる研究は過去にいくつか行われている．Agirre ら [1] は LSA(潜在意味解析) による単語間類似度を用いる手法を提案し， 語義曖昧性解消におけるドメイン適応に利用した．また，Cai ら [4] は，潜在ディリクレ配分法 (LDA) をモデルに組み込み，ナイーブベイズ分類器を作成し，高い精度を達成した．

また，Feedforward Neural Network Language Model [3] や Recurrent Neural Network Language Model [7] などに代表されるような，ニューラルネットワークに基づく言語モデルの学習の過程で獲得された word embeddings と呼ばれているベクトル表現は，単語の意味的な特徴をよく表していることが報告されている [2]．なかでも，Mikolov ら [8] は言語モデルを単純化し，ベクトル表現の学習を高速に行う skip-gram モデルと CBoW モデルを提案した．さらに Chen ら [5] は skip-gram モデルを拡張し，WordNet の語義ごとの embeddings の学習と同時に語義曖昧性解消を行った．Chen らのモデルは all-words WSD，domain-specific WSD において，state-of-the-art の性能を示している．本研究では word embeddings を教師付き語義曖昧性解消タスクにおいて素性として用い，その効果の分析を行う．

3 分散表現を用いた語義曖昧性解消

3.1 タスク設定

語義曖昧性解消タスクでは，文書中の多義語について，その語が取りうる語義の中から最もふさわしいものを選ぶ．この際，語義は一般的に辞書や WordNet などの外部知識に基づき定義される．

教師あり語義曖昧性解消タスクでは，一部の単語に対し人手で正しい語義が付与されたデータが与えられたという条件の下で，高い精度でその語義を推定できる語義分類器を構築することが目的となる．

分類器を作成する際には，Naive-Bayse 分類器や Support Vector Machine (SVM) などが用いられることが多く，高精度な分類器を構築するためには素性設計が重要となる．本研究では，分類器として SVM を用いた上で，分散表現に基づく素性を新たに導入する．

3.2 素性

本研究では，提案する分散表現を用いた素性（CWE）に加え，比較のため 3 つの素性（BoW，posBoW，SumWE）を使用する．以下では，各素性の詳細を説明する．

Bag-of-Words (BoW) 対象単語の前後 N 単語に出現した単語の種類を表す．辞書に登録された単語数を V とすると V 次元の二値ベクトルとなる．

たとえば，“I have not prepared for the **meeting** at all.” という文が与えられ，meeting が対象単語だとする．このとき素性になる単語は前後 5 単語の $[\mathrm{have}, \mathrm{not}, \mathrm{prepare}, \mathrm{for}, \mathrm{the}, \mathrm{at}, \mathrm{all}]$ であり，このとき素性はこれらの単語に対応するインデックスが 1 となった (0,1,1,1,1,1,0,0,1,0,1,0,...,0) というベクトルで表現される

position-Bag-of-Words (posBoW) 位置の違いを考慮した BoW である．対象単語のインデックスのみ 1，残りが 0 の二値ベクトルで単語を表現し，二値ベクトル表現を出現順に繋げた素性とする．一つの単語が V 次元の二値ベクトルで表現されるため，この素性表現の次元は $2 \times N \times V$ となる．前述した BoW は前後 N 単語に現れる単語のみを表し，出現位置を考慮していないが，posBoW は位置の情報が入った素性である．以下の CWE が位置を考慮した素性であるため，比較のために用いた．

Average-Word-Embeddings (AveWE) 単語の表現として二値ベクトルではなく，学習された単語の実数値ベクトルを用いる．以下の CWE は BoW と異なり位置を考慮した素性であるため，位置の情報を持たない素性も提案し，比較する．対象単語の前後 N 単語の $2N$ 単語に対応した $2N$ 個のベクトル表現を平均した一つのベクトルとする．この素性表現の次元数は単語の分散表現の次元数 L と等しい．

Context-Word-Embeddings (CWE) 素性として対象単語の前後 N 単語の各単語の実数値ベクトルを与える．たとえば，$N = 5$ であり，前 5 単語が w_1, w_2, w_3, w_4, w_5 であるとき，前 5 単語を表すベクトルは v_{w_1},v_{w_2},v_{w_3},v_{w_4},v_{w_5} を並べたベクトルとなる (v_w は単語 w の分散表現を意味する)．後ろ 5 単語も同様である．単語の分散表現の次元数を L としたとき，この素性の長さは $2 \times N \times L$ となる．

4 実験

4.1 実験設定

実験には SemEval2007 task17 [10] の Lexical Sample データを用いる．100 個の異なる多義語で構成されるデータであり，各単語ごとに訓練データとテストデータが含まれている．単語ごとの事例数にばらつきがあり，1 単語あたりの平均事例数は訓練データは 222，テストデータは 48 である．

前述した各素性作成の前処理として，各単語の見出し語化を行った．見出し語化には，NLTK(Natural Language Toolkit)[1] を用いた．ベクトル表現に基づく素性を作成する際の分散表現として，1000 億単語からなるニュース記事から skip-gram モデルを用いて学習された単語の分散表現を使用した．このデータは Mikolov らにより公開されており [2]，各単語は 300 次元のベクトルで表現されている．

分類器の作成には SVM のツールとして LIBLINEAR [6] を用い，SVM の正則化パラメータ C は 0.1 から 1 の間 0.1 刻み，1 から 10 の間 1 刻みに変化させ，訓練データにおいて 5 分割交差検定を行った結果，正解率の高いものを選択した．多値分類器を構築する際は one-versus-rest 法を用いた．

4.2 実験結果

3 節で説明した各素性とそれらを組み合わせた素性集合を用いて実験を行った．各素性の分類結果の正解率を表 1 に示す．3 節で紹介した略記で各素性を表している．また，各素性を用いて構成した分類器の分類結果に統計的に有意な差が存在するか，マクネマー検定を用いて確認した．検定の結果は表 1 に合わせて記載した．

CWE 素性が BoW に基づく素性を大きく上回り，BoW と CWE の組み合わせを用いた結果が最も良い結果となった．一方，BoW と比べ，posBoW が良い

[1] http://www.nltk.org/
[2] https://code.google.com/p/word2vec/

素性集合	正解率
BoW	84.72%
posBoW	85.53%
AveWE	84.56%
CWE	87.51% * †
posBow+CWE	87.18% * †
BoW+CWE	87.80% * †

表 1: 素性による分類結果の違い
*は BoW と統計的有意差があった結果であることを示し，
† は posBoW と有意差があった結果であることを示す．

結果を示した．このことから，単語の出現位置の情報は重要だと考えられるものの，posBoW に含まれる情報は CWE にも含まれているため，posBoW と CWE の組み合わせが CWE とほぼ変わらない結果となったと考えられる．また，BoW 素性が単語の出現位置の微細な変化に影響を受けない素性であるため，BoW と CWE の組み合わせが最も良い結果を示したのだと考えられる．

また，分散表現に基づく素性である CWE を用いることで正解率が向上したのは，教師データに出現しなかった単語の情報を分類に利用できたためと考えられる．このことを確かめるため，各テスト事例の対象単語の前後 5 単語において，訓練において対象単語の前後 5 単語に出現していない単語の数を事例ごとに調べ，訓練事例に未出現の単語数と，各テスト事例の正解率との相関を調べた．図 1 に訓練事例に未出現の単語数ごとの正解と不正解の割合を示す．それぞれ，縦軸が正解率，横軸が訓練データに未出現の単語数を表している．また表 2 に未出現の単語数ごとのテスト事例の数を示す．

各グラフの横軸は訓練事例に未出現の単語数を表し，右に行くほど，訓練事例で未出現の単語が多いことを表す．左端の棒は前後 5 単語がすべて訓練事例に出現しているテスト事例における正解等の割合を示し，右端の棒は前後 5 単語がすべて訓練事例に未出現の単語であるテスト事例における割合を表している．すなわち，図 1 の横軸は訓練事例とテスト事例とで BoW 素性でどれだけ情報の共有が行われているかを表し，右端はまったく共有されていない事例である．右に行くほど BoW 素性での不正解を表す灰色の割合が増えているのが確認できる．また，CWE 素性を用いることにより改善したことを表す斜線の棒に着目すると，訓練事例に未出現の単語の数が増加しても，単語の分散

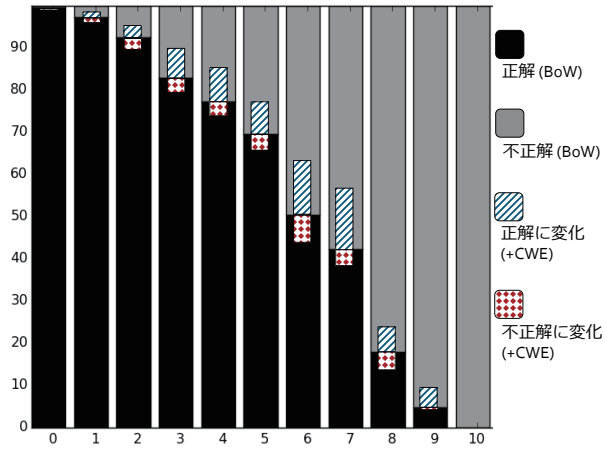

図 1: 訓練事例に未出現の単語数の影響

0	1	2	3	4	5	6	7	8	9	10
564	765	872	846	766	525	321	109	50	21	12

表 2: 各未出現の単語数ごとの事例数

表現を用いた素性を用いたシステムは正解の割合が減りにくく，教師データに出現していない単語にもうまく適応できたと考えられる．

4.3 ベクトル構築法の比較

WSD の精度向上には，単語の分散表現の構築方法が大きく関係していると考えられる．どのような手法が効果的であるかを調査するため，word2vec により学習されたベクトル表現と特異値分解 (SVD) を用いた方法により得られたベクトル表現の比較を行った．

ベクトル表現の比較を行う際には同一のコーパスでベクトル表現の学習を行う必要がある．学習には，英語の Wikipedia (EnWiki) の記事データを用いた．記事データは 2014 年 8 月時点でのデータであり，約 17 億単語からなる記事データである．特異値分解を用いる方法では，Baroni ら [2] と同様に出現位置が一定距離以内の単語同士を共起したとみなし，コーパスを用いて単語間の共起行列を作成し，この行列を特異値分解することで得られる単語ベクトル (EnWiki SVD) を用意した．word2vec を用いる方法 (EnWiki w2v) では skip-gram モデルを使用し，負例サンプリングの数を 10，ダウンサンプリングの割合を 10^{-5} に設定した．また，各ベクトル表現はすべて 300 次元に設定した．この結果を表 3 に示す．

英語の Wikipedia で skip-gram を用いて学習されたベクトル表現を用いた場合，CWE 単体のみで BoW を用いた結果を大きく上回っている．一方，SVD を用

	CWE	BoW + CWE
EnWiki SVD	83.54%	86.15%
EnWiki w2v	86.45%	86.58%
Google w2v	87.51%	87.80%
(BoW)	84.72%	

表 3: ベクトル表現の構築法による分類結果の違い

いた構築方法で得られたベクトルでは，CWE 単体を用いた場合は BoW を用いた結果に正解率で負けているが，CWE と BoW を組み合わせることで BoW 素性を大きく上回る結果を示している．SVD に基づくベクトル表現を用いた CWE では，単語の特徴を十分に捉えられていないと考えられるが，組み合わせで上昇したことから，BoW と異なる観点を表現できていると考えられる．一方，skip-gram モデルに基づくベクトル表現を用いた CWE では，BoW より有用な情報を含んでいると考えられる．また，BoW と組み合わせた際の正解率の上昇率が少ないことから，得られたベクトル表現が十分に単語の特徴を表していると考えられる．

最後に CWE 素性を用いることで正解に変化したテスト事例の例と，類似する訓練事例の例を表 4 に載せる．これらは，management の “経営者” を意味する同じ語義の例である．青のイタリック体の単語が訓練事例において前後 5 単語に出現しなかった単語を表し，赤字で太字の単語が対象の多義語である．対象語の周囲に未知語の企業名が出現しているが，企業名を表すベクトル表現を手がかりに学習できたため，正解に変化したと考えられる．

5 結論

本研究では語義曖昧性解消タスクにおいて単語の分散表現を分類器の作成に利用することの有効性を確認した．実験の結果，分散表現を用いた表現手法が二値ベクトルに基づく表現手法を大きく上回ることが確認され，分散表現を用いることで，訓練データと表層的な類似度が低い事例にも対応することが可能になることを明らかにした．今回用いた素性は他の素性との組み合わせも可能であり，教師あり学習に基づく語義曖昧性解消システムの性能向上に寄与することが期待できる．

対象語	事例
訓練事例	But the government ’s action , which caught Jaguar **management** flat-footed , may scuttle the GM minority deal by forcing it to fight for all of Jaguar .
	Claude Bebear , chairman and chief executive officer , of Axa-Midi Assurances , pledged to retain employees and **management** of Farmers Group Inc.
テスト事例	*Younkers* **management** is likely to buy a 10 % to 20 % interest in the chain in January ,
	New **management** at *Kentucky Fried Chicken*, a unit of PepsiCo Inc. , has fought back with new medium and large chicken sandwiches for the lunch crowd .

表 4: CWE を用いることで正解に変化した例

参考文献

[1] E. Agirre and O. L. de Lacalle. Supervised domain adaptation for WSD. In *Proc. of EACL 2009*, pages 42–50, 2009.

[2] M. Baroni, G. Dinu, and G. Kruszewski. Don’t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proc. of ACL 2014*, pages 238–247, 2014.

[3] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[4] J. F. Cai, W. S. Lee, and Y. W. Teh. NUS-ML:Improving word sense disambiguation using topic features. In *Proc. of SemEval-2007 ACL 2007*, pages 524–531.

[5] X. Chen, Z. Liu, and M. Sun. A unified model for word sense representation and disambiguation. In *Proc. of EMNLP 2014*, pages 1025–1035, 2014.

[6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[7] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048, 2010.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. *NIPS*, 26:3111–3119, 2013.

[9] R. Navigli. Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2):10:1–10:69, 2009.

[10] S. S. Pradhan, E. Loper, D. Dligach, and M. Palmer. SemEval-2007 task 17: English lexical sample, SRL and all words. In *Proc. of SemEval-2007 ACL 2007*, pages 87–92.

[11] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of ACL 1995*, pages 189–196, 1995.