Vol.2016-NL-226 No.7
Vol.2016-SLP-111 No.7
2016/5/16

分散表現による語義曖昧性解消の領域適応

鈴木 翔太[1,a)]　古宮 嘉那子[1,b)]　佐々木 稔[1,c)]　新納 浩幸[1,d)]　奥村 学[2,e)]

概要：語義曖昧性解消における領域適応において、これまでの手法である、巨大なコーパスデータの分散表現を用いる方法は行われていたが、ソースデータの分散表現やターゲットデータの分散表現など、様々な分散表現を組み合わせることで、語義曖昧性解消の性能が向上するかを検証した。

SUZUKI Masaya[1,a)]　KOMIYA Kanako[1,b)]　SASAKI Minoru[1,c)]　SHINNOU Hiroyuki[1,d)]
Okumura Manabu[2,e)]

1. はじめに

複数の語義を持つ単語を多義語といい、その多義語が文章中でどの語義で使用されているかを判定するタスクを語義曖昧性解消 (Word Sense Disambiguation, WSD) という。また、自然言語の処理などを行う際、テストの対象となるドメインではなく、異なるドメインのデータ (ソースデータ) で学習を行い、それをターゲットとなるドメインのデータ (ターゲットデータ) に適応することを領域適応 (Domain Adaptation) と言い、近年様々な手法が研究されている。一方、単語の意味をベクトルで表現したものを分散表現といい、近年、この素性が語義曖昧性解消において有効であることが報告された ([1])。

本稿では語義曖昧性解消の領域適応に、文章から作成された分散表現を利用する。特にこれまでは、大規模データから作成された分散表現を素性として追加する手法は行われていたが、ターゲットデータやソースデータから作成された分散表現を付加するというような、領域適応に特化した手法は行われていなかった。そのため本研究では、どのようなデータの分散表現を利用する手法が領域適応において効果的であるかを検証する。

[1] 茨城大学
Ibaraki University
[2] 東京工業大学
Tokyo Institute of Technology
[a)] 12t4042a@vc.ibaraki.ac.jp
[b)] kanako.komiya.nlp@vc.ibaraki.ac.jp
[c)] minoru.sasaki.01@vc.ibaraki.ac.jp
[d)] hiroyuki.shinnou.0828@vc.ibaraki.ac.jp
[e)] oku@pi.titech.ac.jp

2. 関連研究

領域適応は、学習に使用するデータの種類により教師あり学習 (supervised)、 半教師あり学習 (semi-supervised)、教師なし学習 (unsupervised) の 3 種類に分けられる。本研究は semi-supervised な領域適応であり、ラベル付きのソースデータとラベルなしのターゲットデータを利用するものである。

本実験にもっとも近い研究は、Sugawara らの手法 ([1]) である。彼らはその研究において、語義曖昧性解消を行う際のテストデータ、ターゲットデータの双方に Wikipedia [*1] のダンプデータによる分散表現を追加し、追加しなかったときに比べて性能が向上したことを報告している。ただし、これは語義曖昧性解消自身の研究であり、領域適応についての実験は行っていない。また、用例ベクトルにおける対象単語において、Sugawara らは分類語彙表 [3] による素性を含めていない。

また、山木らの手法 ([2]) は、Sugawara らの手法の改良を行っている。彼らは、単なる分散表現ではなく、分散表現から一歩工夫した素性を利用して、Sugawara らの手法を改良した。しかし、[2] もまた語義曖昧性解消の実験であり、その点において、領域適応という本手法の目的とは異なる。

また、Taghipour ら [5] は分散表現を学習して語義曖昧性解消の性能を向上させている。しかし、彼らも一般的なドメインに対する実験は行っているが、領域適応に特化して、ターゲットドメインの分散表現を利用するという視点では

[*1] https://ja.wikipedia.org/wiki/%E3%83%A1%E3%82%A4%E3%83%B3%E3%83%9A%E3%83%BC%E3%82%B8

実験を行っていない。そのため、本研究では、Wikipedia などの巨大なコーパスから作成した分散表現をただ利用して性能を向上させるだけではなく、領域適応という設定の上で、ターゲットデータのジャンルに特化した分散表現を利用することで、どのような効果があるかに着目して実験を行った。

なお、日本語の語義曖昧性の領域適応の研究として、小林らの手法 ([4]) などがある。[4] では、語義曖昧性解消の精度を上げるために、訓練事例を反復的に選択していく手法がとられている。本稿では、素性などに関してはこの論文を参考にしている。

3. 分散表現を利用した語義曖昧性解消の領域適応

本稿では、ソースデータとターゲットデータの双方に、語義曖昧性解消の対象単語の前後の語の分散表現を追加することで，語義曖昧性解消の領域適応を行う．具体的には，以下の組合せで分散表現を付加した。

Add Target　ターゲットデータの分散表現を利用する手法
Add Source　ソースデータの分散表現を利用する手法
Add Wiki　巨大なコーパス (Wikipedia) による分散表現を利用する手法
Add Target & Source　ソースデータの分散表現とターゲットデータの分散表現をそれぞれ追加する手法
Add Target & Wiki　ソースデータの分散表現と Wikipedia の分散表現をそれぞれ追加する手法
Add Target Large　より大きなコーパスである Large データによる分散表現を利用する手法

なお、[1] の報告に従い、分散表現は、単語ごとの分散表現を足して利用するのではなく、それぞれの単語の分散表現を素性として連結する形で利用した。また、複数コーパスの分散表現を利用する際には、それぞれのコーパスのそれぞれの単語の分散表現を素性として連結して利用した。

4. 実験

4.1 データセット

コーパスのデータとして、現代日本語書き言葉均衡コーパス ([6]) の非コアデータのうち、YAHOO!知恵袋 (YAHOO) と白書 (BCCWJ) の 2 種のデータ、および毎日新聞のデータである RWC コーパス ([7]) を用いた。これらのコーパスには、岩波国語辞典 ([8]) での語義が付与されている。なお、対象となる単語は BCCWJ において用例ベクトルが 50 個以上ある 36 種類の単語のデータを対象とした。語義の個数ごとの単語の内訳は、1 語義 (新語義を入れると 2 語義) : 可能、2 語義 : 生きる、一般、生まれる、書く、考える、技術、経済、現在、現場、子供、自分、情報、高い、作る、強い、電話、場合、早い・速い、文化、ほか、見せる、3 語義 : 相手、与える、言う、今、入れる、大きい、教える、買う、関係、聞く、市場、市民、社会、進む、地方、出来る、出る、入る、初め・始め、始める、場所、開く、前、求める、訴える、4 語義 : 時間、時代、出す、乗る、計る、一つ、見える、認める、持つ、進める、5 語義 : やる、良い、6 語義 : 合う・会う、立つ・建つ、見る、もの、7 語義 : 手、8 語義 : する、取る、上げる となる。

これらのデータに加え、3 種類のコーパスのより大きなデータを使用した (Large). YAHOO 知恵袋および白書のための大きなデータは，現代日本語書き言葉均衡コーパス中の YAHOO ！知恵袋および白書のデータ（可変長の文書構造タグつき xml）すべてである．また，新聞のための大きなデータは，毎日新聞のデータである RWC コーパスの 1991 年〜2005 年までのデータである [*2].

また、ラベルなしの巨大なコーパスデータとして、Wikipedia のダンプデータを利用した。

4.2 語義曖昧性解消に使用する素性

先行研究である [4] では、語義曖昧性解消の素性として、

(1) 対象単語と前後 2 単語の表記 (1〜5 個目)
(2) 対象単語と前後 2 単語の品詞 (6〜10 個目)
(3) 対象単語と前後 2 単語の品詞の細分化 (11〜15 個目)
(4) 係り受け (16 個目)
(5) 前後 2 単語の 5 桁の分類コード (17〜20 個目)
(6) 前後 2 単語の 4 桁の分類コード (21〜24 個目)

の 6 種類の素性を用いており、本研究においても同じ素性をもちいた。なお，分類コードは分類語彙表のコードである．また，形態素解析には MeCab[*3] を利用した．

本実験では、この 24 個の素性に加え、分散表現を末尾に追加していくことで語義曖昧性解消の正解率が上昇するかを調査した。

4.3 実験設定

分散表現の作成には、word2vec [*4] ([9], [10]. [11]) を使用し、size は 200、window を 5、cbow を利用して分散表現を作成した。なお、コーパスごとの分散表現の語彙数は **表 1** のようになった。もっとも大きなサイズの Wikipedia のデータにより作成した分散表現の語彙数をもとにした割合も参考として示す。

語義曖昧性解消の分類器としては、libsvm[12] というサ

[*2] 語義タグつきコーパスは 94 年の新聞からなる.
[*3] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html?sess=3f6a4f9896295ef2480fa2482de521f6
[*4] https://code.google.com/archive/p/word2vec/

表 1: 分散表現の語彙数と WIKI の語彙数に対する割合

コーパス	分散表現の語彙数	WIKI に対する割合
BCCWJ	13,336	1.24%
YAHOO	22,709	2.10%
RWC	11,685	1.08%
BCCWJ Large	14,303	1.33%
YAHOO Large	37,893	3.51%
RWC Large	201,421	18.67%
WIKI	1,078,930	100.00%

表 2: 手法ごとの正解率

手法	マクロ平均	マイクロ平均
ベースライン	77.90%	79.79%
Add Target	78.35%	79.92%
Add Source	73.48%	75.98%
Add Wiki	78.60%	79.73%
Add Target & Source	78.13%	79.77%
Add Target & Wiki	76.99%	78.45%
Add Target Large	78.64%	80.02%
アッパーバウンド	89.88%	89.90%

ポートベクターマシンを利用した.

本稿では、前述した Add Target, Add Source, Add Wiki, Add Target & Source, Add Target & Wiki, Add Target Large の六手法を実験する．また、分散表現を追加していないもとの素性ファイル同士で libsvm によって語義曖昧性解消を行ったものをベースラインとした。一方、BCCWJ に BCCWJ による分散表現を追加するといった、素性データに同じドメインのコーパスデータによる分散表現を追加したものを、libsvm によって 5 分割交差検定を行ったものを、アッパーバウンドとする。

5. 結果

実験の結果、**表 2** のような正解率となった。なお、マクロ平均は語彙ごとの正解率の平均であり、マイクロ平均は全体の用例数に対する正解率となっている。

6. 考察

表 2 より、マクロ平均がベースラインより上がったのは、Add Target と Add Wiki、Add Target & Source、Add Target Large の四手法であることが分かる。このうち、最も良いのは Add Target Large の 78.64%、次は Add Wiki の 78.60%であった。このことから、大きなコーパスから作成した分散表現は、様々な単語タイプの語義曖昧性解消に対して有効であることが読み取れる。次に、マイクロ平均がベースラインより上がったのは、Add Target および Add Target Large の二手法であることが分かる。マイクロ平均は、一用例をひとつの重みとして計算する平均値であるため、この結果は、ターゲットデータから作成した分散表現を利用すると、（ターゲットデータに）より多くの用例を持つ単語タイプの正解率が上昇していることを意味する。語義曖昧性解消の対象単語の文脈に出てくる語が互いに似通っていると仮定すると、Add Target および Add Target Large の二手法は、用例がたくさんある単語タイプの周辺語もコーパス中に頻出することになるため、ターゲットデータの素性として利用される分散表現が、より多くの用例から分散表現が作成できた可能性が高いと考えられる。そのため、よりドメインに特化した分散表現を作成できていると考えられる。

また、マイクロ平均で最も正解率が高かったのは、Add Target Large の 80.02%、次は Add Target の 79.79%であった。特に、Add Target Large とベースラインの差は、カイ二乗検定において、有意水準 0.05 で有意であった。また、これらのことから、マクロ平均およびマイクロ平均両方において、Add Target Large が最も正解率が高いことが分かる。

次に、Wikipedia から作成した分散表現を用いた場合（Add Wiki）について見る。今回の実験では、Add Wiki は、マイクロ平均に関してベースラインを下回った。この結果は Sugawara の手法 [1] の結果と異なっているが、これは (1)Sugawara の手法 [1] は英語を対象言語にしており、利用している Wikipedia のデータがさらにずっと大きいためと、(2) 我々のベースラインが分散表現と同様に意味のスパースネスを解消する目的で取り入れられた、分類語彙表のコードを素性として利用しているためであると考えられる。

マクロ平均とマイクロ平均の双方がベースラインが下降したものは Add Source, Add Target & Wiki の 2 つの手法となった。Add Source 手法の正解率が下降した原因としては、ソースデータの分散表現は、ターゲットデータと比べて単語の意味合いが異なってしまうためと考えられる。Add Target & Wiki は、Add Target はマクロ平均およびマイクロ平均の両方がベースラインを上回っており、さらに Add Wiki に関してもマクロ平均はベースラインを上回っているのにも拘わらず、ベースラインを下回っている。この原因を究明するためにはさらなる実験が必要である。

次に、素性として利用した分散表現の語彙数と語義曖昧性解消の正解率について考察する。表 1 を見てみると、BCCWJ、YAHOO、RWC の語義曖昧性解消用のコーパスから作成された分散表現の語彙数は、Wikipedia から作成された分散表現に比べて、それぞれ 1.24%、2.10%、1.08%にとどまっている。また、BCCWJ、YAHOO、RWC のより大きなデータから作成された分散表現の語彙数の割合も、それぞれ 1.33%、3.51%、18.67%である。しかしながら、マイクロ平均については、Add Target および Add Target Large は、Add Wiki を上回っており、マクロ平均においても、Add Target Large は、Add Wiki を上回っている。こ

のことから、語義曖昧性解消の素性に利用する分散表現は、ターゲットデータの用例数が少ない場合には、Add Wiki の方が有効であるが、ある程度用例数がある場合には、作成に利用するコーパスの大きさよりも、よりドメインに特化した分散表現を利用した方がよいことが読み取れる。

7. まとめ

本稿では、語義曖昧性解消の semi-supervised な領域適応において、分散表現を利用してその性能を向上させる研究を行った。特にこれまでは、大規模データから作成した分散表現を素性として追加する手法は行われていたが、領域適応に特化してターゲットデータやソースデータから作成した分散表現を利用する実験は行われてこなかった。そのため、本研究では、このようなデータの組合せに着目して、どのようなデータから作成した分散表現を利用するのが領域適応において効果的であるか検証した。その結果、Wikipedia から作成した分散表現を素性として追加するよりも、分散表現の語彙サイズがその 20%に満たなくても、ターゲットデータのドメインのコーパスから作成した分散表現を利用した方が、語義曖昧性解消の正解率が上昇することを示した。

参考文献

[1] Sugawara, H., Takamura, H., Sasano, R. and Okumura, M.: Context Representation with Word Embeddings for WSD, *Proceedings of PACLING 2015*, pp. 1–9 (2015).
[2] 山木翔馬，新納浩幸，古宮嘉那子，佐々木稔：分散表現を用いた教師あり機械学習による語義曖昧性解消，研究報告自然言語処理 (NLP)，Vol. 2015-NL-224, No. 17, pp. 1–8 (2015).
[3] 国立国語研究所：分類語彙表，集英社 (1964).
[4] 小林優稀，古宮嘉那子，佐々木稔，新納浩幸，奥村　学：領域適応のためのサポートベクトルを用いた訓練事例の反復的選択，コーパス日本語学ワークショップ予稿集，pp. 129–136 (2015). http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no7_papers/JCLWorkshop_No.7_16.pdf.
[5] Taghipour, K. and Ng, H. T.: Semi-supervised Word Sense Disambiguration Using Word Embeddings in General and Specific Domains, *Proceedings of the 2015 Annual Conference of North American Chapter of the ACL (NAACL 2015)*, pp. 314–323 (2015).
[6] Maekawa, K.: Balanced Corpus of Contemporary Written Japanese, *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pp. 101–102 (2008).
[7] Hashida, K., Isahara, H., Tokunaga, T., Hashimoto, M., Ogino, S. and Kashino, W.: The RWC text databases, *Proceedings of the First International Conference on Language Resource and Evaluation*, pp. 457–461 (1998).
[8] Nishio, M., Iwabuchi, E. and Mizutani, S.: *Iwanami Kokugo Jiten Dai Go Han*, Iwanami Publisher, In Japanese (1994).
[9] Mikolov, T., tau Yih, W. and Zweig, G.: Linguistic Regularities in Continuous Space Word Representations, *Proceedings of NAACL 2013*, pp. 746–751 (2013).
[10] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *Proceedings of ICLRWorkshop 2013*, pp. 1–12 (2013).
[11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *Proceedings of NIPS 2013*, pp. 1–9 (2013).
[12] Chang, C.-C. and Lin, C.-J.: *LIBSVM: a library for support vector machines* (2001). Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.