話題に基づく語義曖昧性解消

椹澤 優希　山本 和英
長岡技術科学大学
{gumizawa, yamamoto}@jnlp.org

1 はじめに

語義曖昧性解消はコンピュータが意味理解するために必要であり，NLP のタスクにおいて大きな問題となっている．それ故に語義曖昧性解消について様々な手法が提案されており，教師あり学習や教師なし学習，半教師あり学習などの様々な手法が提案されており一定の精度が得られたことを報告している [1, 2].

しかしながら近年においては語義曖昧性解消の分野は若干下火になってきており，多くの手法が提案されているにも関わらず語義曖昧性解消を用いて後段処理を行うといった研究は非常に少ないように見える．de Lacalle らはこれらの原因について「語義曖昧性解消の精度」と「後段処理において有用な粒度」に関する問題の 2 つを上げている [3]．両者の関係は非常に密接に結びついている．より細かい粒度での語義曖昧性解消は非常に有用であるが，その一方で精度が落ちてしまい実用的ではなくなってしまう．そのため近年では粗い粒度での語義曖昧性解消の研究も行われており [4]，高い精度の語義曖昧性解消をアプリケーションに落とし込むという動きも見られる．また，これらの原因に加えて「後段処理において有用な出力形式」も問題になる．通常の語義曖昧性解消タスクは辞書の語義タグを推定する問題であり，その出力は数字であるため，それを後段処理へ活かすためには語義タグに対する情報の付与が必要となる．

我々はこれらの問題に対して「話題に基づく語義曖昧性解消」を提案する．これは文の話題を考慮し複数の話題で使われるような多義語をその話題のカテゴリへ分類するというタスクである．この語義曖昧性解消では話題を持たない語は無理に曖昧性解消を行わないため，通常の語義曖昧性解消に比べて粗いものであり高い精度が期待できる．更に出力結果として得られるものは語義タグではなくその語の話題のカテゴリであるため，その話題に関連する語を多く参照できる．

本研究の主眼は実用性のある語義曖昧性解消のツール化である．そのため，本論文では新規手法の提案などは行わず，作成した辞書資源とデータセットをもとにした本ツールについて日本語解析器雪だるま [5] に実装し検証を行う．

2 関連研究

各カテゴリへの分類を行う形での語義曖昧性解消の研究はいくつか存在する．Class に基づく語義曖昧性解消はそれらの研究の一つであり [6]，対象単語を人や場所などの意味属性に基づいて語義曖昧性解消する研究である．日本語においても村本らが意味カテゴリに基づく語義曖昧性解消について提案しており，「食べ物」や「衣類」「武器」などの様々な意味カテゴリを設け，それらのカテゴリを単語に付与する形での語義曖昧性解消を提案している [7].

我々が提案する語義曖昧性解消の粒度はこれらに非常に近いものである．一方で，我々が提案する手法において用いるカテゴリは「話題」に基づくものである．意味カテゴリにおいて「中学 3 年生」と「黒板」という語はそれぞれ「学齢」と「製品名」に分類されるが[1]，話題に基づいたカテゴリでは両方「学校」というようなカテゴリに属する．このように話題に基づく語義曖昧性解消ではその文の話題を特定するような形で曖昧性解消を行うということである．

話題に基づく語義曖昧性解消の利点としてその曖昧性解消のしやすさがある．例えば「私はタクシードライバーです」という文中の「ドライバー」という単語が「タクシー」という単語から「自動車」に属することは容易にわかる．このように話題に基づく語義曖昧性解消ではその語義を周辺の単語に結びつく話題によって曖昧性の解消が可能であると考えられる．そこで我々は単語に結びつく話題の辞書を作成し，辞書に基づき訓練データを作成，ツールの実装を行うことで精度が高く実用的な語義曖昧性解消ツールを提供する．

[1]意味カテゴリは関根の拡張固有表現に基づくため固有表現でないこれらの分類は厳密には正しいとは言えない．
https://sites.google.com/site/extendednamedentityhierarchy/

3 話題に基づいた分類辞書の作成

話題の基づいた分類辞書は以前我々が作成した辞書を更に拡張したものを使用する [8].

この話題辞書では著者らの主観によって話題となるカテゴリを以下の 3 サイトから抽出している.

1. Wikipedia に存在する記事タイトル一覧
2. Yahoo!知恵袋に存在するカテゴリ一覧 [2]
3. Yahoo!ブログに存在するカテゴリ一覧 [3]

これらの話題や場面に基づいてカテゴリの候補となる語を著者の主観で抽出し，それらの単語について同じものを指す語や関連度の高い語についてカテゴリを統合した．結果として表 1 のような形で 380 単語から 145 カテゴリを選定している.

表 1: 話題に基づく辞書のカテゴリ

対象となる単語	カテゴリ名
お笑い	お笑い・芸人
芸人	
駅	鉄道・列車
鉄道	
列車	
ゲーム	ゲーム

また，作成したカテゴリに対して Wikipedia のリンク情報や Word2Vec などを用いてカテゴリに属する語の候補を作成し，それらの候補について正しくカテゴリに当てはまるか著者らの主観によって判断し辞書への追加を行っている．カテゴリに属す対象語は名詞だけでなく動詞や形容詞も含まれる．実際に作成された辞書の一部を表 2 に示す．実際の辞書には現時点で計 12,462 語が収録されている.

表 2: 話題に基づく辞書の一部

学校	結婚・恋愛	食材・料理
ホームルーム	恋人	鍋
遠足	デート	調理する
進級	カップル	クッキング
黒板	不倫	台所
校庭	プロポーズする	流し台
入学する	許嫁	盛り付け
登校する	ブライダル	てんこ盛り
夏休み	フィアンセ	まな板
宿題	駆け落ちする	塩もみする
学生	見合い	フランベ

[2]http://chiebukuro.yahoo.co.jp/dir/dir_list.php
[3]http://blogs.yahoo.co.jp/FRONT/cat.html

4 データセットの構築

日本語語義曖昧性解消のデータセットには SemEval2010 のデータ [9] 等が存在するが，ここでは話題に基づいて語義曖昧性解消を行うため複数の話題を持つ単語に対して一からデータセットの構築を行った.

複数の話題を持つ語の選定には岩波国語辞典第五版タグ付きコーパス[4]を用いた．岩波国語辞典にて複数の語義を持つ単語について辞書の定義文と前章で作った各カテゴリに含まれる単語群について Word2Vec にて類似度を求め，類似度が高いものから順番に人手で話題の付与を行い，複数の語義を持つ単語がどのような話題に写像されるかの辞書を作成した.

データセットの構築にはクラウドソーシングを活用した．クラウドソーシングは不特定多数の人に作業依頼の募集をかけ，報酬などの条件に合意が得られた作業者に対して指示し作業を行なってもらうものである．この方法では専門的なドメインが必要な作業は不可能であるが，今回のように一般的に使われる単語を対象にした語義の付与では非常に有用である．岩波国語辞典に基づいて作成した複数の話題を持つ単語についてそれらを含む文を日本語ウェブコーパス 2010[5]から抽出し，話題カテゴリの付与を行った．なるべく高品質なテキストのみを対象にするため事前の処理として 70 字以下の文，記号や空白が多い文，似た文等を除外しなるべく多様性をもたせた形で抽出した文を対象にアノテーションを行っている．今回クラウドソーシングの対象とした単語は全部で 50 語である．対象となる文数は 28,872 文であり，1 単語あたりの平均訓練事例数は約 577 文である．語義の付与を行う人数は 2 人に設定し，それぞれ約 14,500 文ずつカテゴリを付与する作業を行ってもらった．語義曖昧性解消の対象となった単語を表 4 に示す．また，実際にクラウドソーシングを行って作成したデータの一部を表 3 に示す．実際に作成したデータにおいては各単語に付与された話題のカテゴリ以外にその他のカテゴリを設けている．これは文を読んだ時にどの話題にも当てはまらないような単語を分類するためのカテゴリである.

5 手法

日本語解析器雪だるまにツールを実装する上で用いる手法には，周辺単語の分散表現をつなぎ合わせる Sugawara らの提案手法である Context Word Embedding(CWE) を用いる [10]．この手法は窓幅 N を 5 と

[4]http://www.gsk.or.jp/catalog/gsk2010-a/
[5]http://s-yata.jp/corpus/nwc2010/

表 3: データセットの一部

対象単語	カテゴリ	文
ドライバー	工具	その後、ホームセンターに行ってラジオ・ペンチや精密ドライバーなどの工具を買って自分で修復を試みるも、フレームを傷つけるはネジはなくすわで、状況は悪くなるばかり
ドライバー	自動車	日中の最高気温が 4.3 度まで上がったこともあり、除雪作業が進まない道路では積もった雪が一気に解けだし、ドライバーは深いわだちや大きな水たまりに悪戦苦闘した
ドライバー	ゴルフ	飛ばしたいクラブ=長いクラブ (ドライバーなど) の場合は、この基本幅より大きく、飛ばしたくないクラブ=短いクラブ (アプローチウエッジなど) は、狭くします

表 4: 語義付与対象単語一覧

節	ドライブ	アーチ	銀
喰う	バック	アクセス	戦線
フラット	法	角	リード
タイトル	コート	オープン	ターミナル
パート	市場	パス	スクリーン
洗う	軌道	ウイルス	ホーム
翼	コース	鉢	クリーム
株	マフラー	洋画	ミキサー
カード	グラス	ポスト	ドライバー
台所	糸	ストック	プレーヤー
脱水	前線	カット	コード
トラック	ツアー	ベース	刑事
ラム	ショット		

し，換言対象前の単語を w_1, w_2, w_3, w_4, w_5 とし，それらの分散表現を表すベクトルを $v_{w_1}, v_{w_2}, v_{w_3}, v_{w_4}, v_{w_5}$ とした時にそれらをつなぎ合わせたベクトルを素性として用いる手法である．この素性の次元は分散表現の次元を L とすると $2 \times N \times L$ となる．

さらに我々はこの CWE の素性に対して文中に含まれる話題を持つ語の情報を Bag-of-Words のように数えて素性として加える方法 (Bag-of-Topics) も検討をする．これは 2 章で述べたように話題に基づいた語義曖昧性解消ではそのほとんどの語義を周辺の単語に結びつく話題によって曖昧性解消が可能であり，周辺単語と結びついた話題を素性として加えることが精度に大きく貢献すると考えられるからである．Bag-of-Topics の次元数は話題辞書に含まれるカテゴリと同数の 145 次元である．つまりこの素性の次元数は $2 \times N \times L + 145$ となる．

素性の具体的な例を図 1 に示す．図中の CWE での手法は周辺単語の分散表現をつなぎ合わせたものになっており，それらの手法に対して周辺単語の話題を Bag-of-Words のように表したベクトルを末尾につなぎ合わせたものが CWE と話題辞書を組み合わせた素性である．

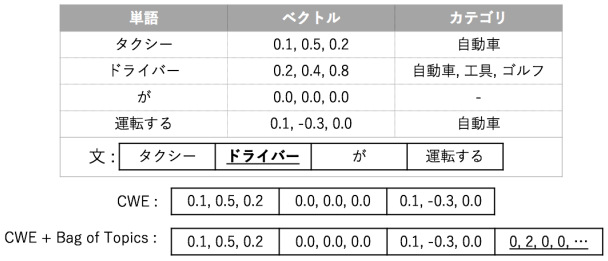

図 1: 各手法における素性

6 実験と考察

5 章で説明した素性について話題辞書の有効性と曖昧性解消の精度を検証する．データセットには 4 章で作成したものを訓練データ 7 割，テストデータ 3 割に分割したものを用いる．訓練データは 50 単語に対して 20,211 文が割り当てられるため 1 単語あたりの訓練事例数は約 404 文となる．テストデータは残った 8661 文である．分散表現の構築には 2016 年 3 月 5 日の日本語 Wikipedia のデータ[6]を wp2txt[7]で処理したコーパスを用いる．単語分割に用いる解析器には日本語解析器雪だるま [5] を用い，gensim[8]の Word2Vec で学習させ単語の分散表現として使う．Word2Vec には Skip-gram を用い，階層的ソフトマックスとネガティブサンプリング両方を使用する．学習に使用する窓幅は 5，ネガティブサンプリングは 5，ダウンサンプリングの割合は 10^{-3} に設定し各ベクトル表現は 200 次元に設定した．分類器には Scikit-learn[9]で実装されている SVM の LinearSVC を使用し，正則化パラメータとして $C = 1.0$ を使用した．また語義曖昧性解消の素性として使う窓幅の値には 5 を設定した．

表 5 に各手法における精度を示す．分散表現をつなぎ合わせる手法に比べて話題辞書を用いて単語に結び

[6] https://ja.wikipedia.org/wiki/Wikipedia:データベースダウンロード
[7] https://github.com/yohasebe/wp2txt
[8] https://radimrehurek.com/gensim/
[9] http://scikit-learn.org/

ついた話題のカテゴリを素性として加える手法では精度が約 3.5 ポイント向上している．このことから文脈が異なる場面で使われる語の語義曖昧性解消の精度向上に話題辞書が大きく貢献することがわかる．

一方で語義曖昧性解消の精度自体はあまり高くなく，実用的な精度には至っていない．日本語語義曖昧性解消タスク [9] でベースラインとなったシステムでは精度が 75.28%であり，それより粒度が粗く使われる文脈が大きく異なる本タスクとあまり精度が変わらない．この原因の一つとしてクラウドソーシングの質の問題がある．データセットの構築では各作業者に対してそれぞれ別のデータを渡し語義の付与を行ってもらったため，作業者ごとにデータの質に大きな違いがあった．例えば文を 1 回読んで文脈が読み取れない場合カテゴリ“その他”を付与する作業者や，良く文を読みなるべくカテゴリを付与する作業者などである．また，間違いも多く存在し文にそぐわない話題が付与されたデータも多い．そのため，このようなあまり高くない精度になったのだと考えられる．これに対する対策としては 3 人に同じデータをアノテーションしてもらい，確度の低いデータは自分で確認するなどの対策がある．

表 5: 手法ごとの語義曖昧性解消の精度

素性	精度
CWE	76.47%
CWE + Bag-of-Topics	79.95%

7 結論

本論文では話題に基づく語義曖昧性解消を提案し，辞書やデータセットの整備を行った．また作成したデータセット，辞書に基づき語義曖昧性解消を行うツールを実装した．結果として話題に基づく語義曖昧性解消において単語に結びついた話題の情報を使うことが精度向上に大きく寄与することがわかった．その一方で，曖昧性解消自体の精度はそのデータセットの質の問題によってあまり高くないことがわかった．今後はノイズの多いデータセットを整備し，高品質なデータを用いた際の精度を確認してより実用性の高い語義曖昧性解消ツールを作成していきたい．

本論文中で整備した辞書や語義曖昧性解消のデータセットは今後公開予定である．また作成したツールは日本語解析器雪だるまに実装し Web 上で公開する予定である．それぞれのデータ，ツールが日本語の語義曖昧性解消に役立つことを願う．

謝辞

本研究は、平成 27〜31 年科学研究費補助 基盤 (B) 課題番号 15H03216、課題名「日本語教育用テキスト解析ツールの開発と学習者向け誤用チェッカーへの展開」、及び平成 29〜31 年科学研究費助成事業挑戦的萌芽 課題番号 17K18481、課題名「やさしい日本語化実証実験による言語資源構築と自動平易化システムの試作」の助成を受けています。

参考文献

[1] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189–196, June 1995.

[2] Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. Semi-supervised word sense disambiguation with neural models. In *COLING 2016*, 2016.

[3] Oier Lopez de Lacalle and Eneko Agirre. A methodology for word sense disambiguation at 90% based on large-scale crowdsourcing. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pp. 61–70, Denver, Colorado, June 2015. Association for Computational Linguistics.

[4] Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. Semeval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 30–35, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[5] 山本和英, 宮西由貴, 髙橋寛治. 日本語解析システム「雪だるま」：単語解析部の設計思想 (言語理解とコミュニケーション) – (第 7 回テキストマイニング・シンポジウム). 電子情報通信学会技術研究報告 = IEICE technical report : 信学技報, Vol. 115, No. 222, pp. 13–18, Sep 2015.

[6] Rubén Izquierdo, Armando Suárez, and German Rigau. An empirical study on class-based word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pp. 389–397, Athens, Greece, March 2009. Association for Computational Linguistics.

[7] 村本英明, 鍜治伸裕, 吉永直樹, 喜連川優. Web テキストを対象とした語義曖昧性解消のための言語資源の半自動構築. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3338–3348, dec 2011.

[8] 桾澤優希, 山本和英. 語の話題に基づく分類辞書の作成. NLP 若手の会 第 11 回シンポジウム, Aug 2016.

[9] Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. Semeval-2010 task: Japanese wsd. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 69–74, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[10] Hiromu Sugawara, Hiroya Takamura, Ryohei Sasano, and Manabu Okumura. Context representation with word embeddings for wsd. In *International Conference of the Pacific Association for Computational Linguistics*, pp. 108–119, 2015.