国語辞典の語義掲載順を利用した新語義検出

赤崎 智† 乾 孝司‡ 山本 幹雄‡
†筑波大学情報学群情報科学類 ‡筑波大学大学院システム情報工学研究科
s.akasaki@mibel.cs.tsukuba.ac.jp {inui,myama}@cs.tsukuba.ac.jp

1 はじめに

ソーシャルメディアの発達によって，Web を通して情報発信することが容易となり，Web 上には個人が記したテキストが多く流通している．これらのテキストは個人が自由に記述しているため，この中には，新しい単語や既知単語であるが新しい意味で使用されている単語がしばしば含まれている．それゆえ，テキスト中の新語や新語義を検出する技術への関心も高まっていると言える．

本稿では新語義検出の手法について論じる．語の意味を自動処理する研究としては語義曖昧性解消（WSD）[4] が以前から中心的な課題であった．WSD は語の語義クラスが事前に与えられる典型的な分類問題である．そのため，教師あり学習に基づく手法が採用されることが多く，学習データも SemEval-2 Japanese WSD Task の活動 [5] などを通して整備されている．

WSD 用学習データでは，語義を推定したい事例（単語）に対して語義 ID を示すラベルが付与されている．我々は現在，この WSD 用に整備された学習データを最大限に活用できる新語義検出手法を検討している．本稿では，これまでに検討した提案手法の概要を述べたあと，簡単な評価実験を実施したので，その結果について述べる．

2 提案手法

提案手法では国語辞典に記述されている語義の掲載順序に注目する．国語辞典には各語の語義が記述されているが，ある見出し語が複数の語義をもつ場合は編纂者が定めた掲載順序方針に従って順に語義が記述される．掲載順序方針は国語辞典によって異なるが，岩波国語辞典 [7] では「出来るだけ現代語として最も普通に行われている意味から始める」とされている．我々はこれを「よく使われる」順と解釈した．この解釈が正しいと仮定するなら，岩波国語辞典の語義の掲載順序を参照することで，語毎に語義の使われやすさを求めることが出来ることになる．以下，掲載順序と呼ぶ場合は，岩波国語辞典における語義の掲載順序を指す．

次に，語義の掲載順序と新語義の関係について考える．新語義とは，ある語におけるこれまで使われていなかった新しい意味であることから，その語の新語義の用例はすべてのどの既知語義の用例よりも出現しないと仮定できる．つまり，我々が検出したい新語義を国語辞典に記述することを考えた場合,「よく使われる」順に掲載するなら，新語義はすべての既知語義のあとに続く形で掲載することになる．

以上を踏まえると，次のような新語義検出手法を検討することができる．

- WSD 用に整備された既知語義データを利用して，語義の掲載順序をモデル化する．ここでは，掲載順序として，絶対的な順位ではなく，相対的な順序関係を想定している．
- ある語に対して，新語義検出対象の用例（n 個）と既知語義用例からなる事例集合（適当な m 個）を準備し，これに上記の順序モデルを適用することで，用例間に順序を与える．
- ここまでの仮定が全て正しいなら，順序付けられた事例群の末尾側に新語義用例が集まることになるので，末尾部分の n 個以下の用例を新語義と判定する．

3 評価実験

3.1 実験の設定

SemEval-2 Japanese WSD Task（以下，JWSD Task）[5] の設定に準拠して，提案手法の有効性

表 1: 実験データ

見出し語数	39
上記の総用例数	2,785
うち，新語義用例数	147

表 2: 実験データにおける既知語義数の分布

既知語義の語義数	単語数	単語例
2	14	電気，カバー
3	11	意味，求める
4	8	時代，カード
5 以上	6	手，落とす
計	39	

を検証する．JWSD Task は WSD が主課題であるが，新語義の扱いも含まれている．ただし，JWSD Task で正式に採用されているデータ内には新語義データはわずかしか存在しない．そこで今回は，JWSD Task と同じ BCCWJ コーパス [3] に対して，新語義の事例数を増やしたデータを用いることにした．実験で使用したデータの詳細を表 1 に示す．新語義を含む異なり単語（国語辞典の見出し語）は 39 件で，単語あたりおよそ 72 件の用例をもち，そのうち平均 4 件が新語義となるデータである．

見出し語 39 件がもつ既知語義の語義数の分布および単語例を表 2 に示す．既知語義の分類粒度は中分類である．手法の制約から，既知語義の語義数が 2 以上のデータを使用している．

次に，語義の掲載順序モデルの学習について述べる．モデルは見出し語ごとに構築する．ある見出し語の既知語義用例に対して，まず，語義掲載順位が異なる用例を対にする．これを $\mathbf{x}$ とおく．そして，この用例対間での順序関係（どちらが語義掲載順位が高いか）を求める．これを y とおく．このとき $(\mathbf{x}, y)$ をモデル学習に用いるひとつの事例とすることで順序モデルを構築する．モデル学習アルゴリズムには Ranking SVM を採用し [1; 2]，実装物として SVM-Rank [1] を用いた．

ここまで，国語辞典の語義はリスト構造を仮定し，このリスト内での並び方を掲載順序と読んでいた．しかし，実際の語義構造は階層構造を持っている．例えば，以下の表 3 の左の例は語義の大分類は 2 つであるが，それぞれ 2 つの中分類をもつ例である．この構造をリスト化するにあたり，表 3 の右のように，大分類を優先するか，中分類を優先するかの選択肢がある．今回は大分類を優先させ，リスト化処理を施した．

表 3: 語義定義のリスト化

階層構造（大-中）	大分類優先	中分類優先
1-1	1	1
1-2	2	3
2-1	3	2
2-2	4	4

各用例は次の素性からなるベクトルとして表現される．

- 対象単語の前後 4 単語の基本形の bag-of-words．ただし，助詞等の機能語はカウントから除外する．要素値には出現頻度を用いる．

検出時は，まず，テスト用例と学習用例の集合をモデルへ入力し，用例間での順序関係のついた用例リストを得る．この用例リストに対して，次の新語義判定規則をテスト用例ごとに適用することで，新語義用例を検出する．

- 新語義判定規則：用例リストの中で，注目しているテスト用例以外のテスト用例をリストから取り除く．このリストにおいて，注目しているテスト用例がリストの下位 k% 以内に含まれていれば，新語義用例と判定する．そうでなければ既知語義用例と判定する．

今回の実験では $k = 5$，10，15，20 を試した．

上記の設定のもと，データセットに対する 3 分割交差検定を実施した．

3.2 実験結果

実験結果は見出し語ごとに得られるが，以後の評価はすべて，見出し語で平均した評価値である．

まず，モデル適用の後に得られる用例リストを評価する．結果を表 4 に示す．見出し語ごとに用例数が異なるため，表 4 では，用例リストの末尾をリスト内での用例順位を 1 位とし，先頭を最下位として，新語義／既知語義の平均用例順位を求めている．

[1] https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

表 4: 用例リストの評価

新語義用例の平均順位	18.25
既知語義用例の平均順位	51.28

表 4 から，新語義用例の方が平均順位が高いことがわかる．この結果，平均的には新語義用例の方が既知語義用例よりも末尾側に偏って配置されていることが確認できる．

次に検出性能を評価する．評価指標は，新語義に対する Precision，Recall，および F 値である．結果を表 5 に示す．表には補助情報として，新語義検出総数と正解総数も示している．

表 5: 新語義検出性能の評価

	P	R	F	検出数	正解数
$k = 5$	0.037	0.014	0.020	54	2
$k = 10$	0.035	0.027	0.031	114	4
$k = 15$	0.065	0.238	0.102	537	35
$k = 20$	0.074	0.354	0.122	702	52

表 5 から，まず，k を大きくすることで検出数が増えるが，これに従って正解数も増え，Recall が向上することが確認できる．このことからもある程度妥当な用例リストが出力できていることがわかる．Precision は Recall ほど規則的な変化が得られていないが，k が大きいほど良い評価値を示しており，その結果，F 値もそれに従うように変化している．なお，一見，評価値全体を通して，低い値を示しているように思われるかもしれない．データセットが異なるため厳密な比較はできないが，表 5 の評価値は新語義検出の最新の研究結果（例えば，文献 [6]）と比べても遜色ない結果であると言える．

最後に，本手法の出力結果を例示する．用例の後ろに，既知語義／新語義の別および実験結果の正誤を示す．

見出し語：カバー
- ... 美しい長椅子 カバー が出来上がった．（既，正）
- ... 1 月に施行した拉致被害者支援法で カバー できない...（既，正）
- ... 貿易の自由化をどの程度 カバー するかにある．（新，正）

見出し語：以前
- ... 善悪 以前 に，これ以上怒られる...（既，誤）
- ...以前 宝塚の方が出ていた時...（既，正）
- ... 言葉 以前 で命題にしていない．（新，正）

見出し語：サービス
- ... 味と サービス のバランスが良い...（既，正）
- サービス 残業しているわけではない．（既，誤）
- ... 携帯電話の サービス など...（新，誤）

見出し語：生命
- 遺伝子制御から 生命 解明へ（既，誤）
- ... 我が国の 生命 線ともいうべき...（既，誤）
- ...oo生命 の研修経験者です．（新，誤）

4 おわりに

本稿では国語辞典の語義掲載順を利用した新語義検出手法について述べた．用例リストの実験結果から，既知語義と新語義の区別はある程度できているが，現状では，新語義判定の精度は高くない．今後，共通データセットの元での先行研究との比較を通して，問題点の整理を進めていきたい．

参考文献

[1] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceeding of the 8th KDD*, pp. 133–142, 2002.

[2] Thorsten Joachims. Training linear svms in linear time. In *Proceeding of the 12nd KDD*, pp. 217–226, 2006.

[3] Kikuo Maekawa. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources*, pp. 101–102, 2008.

[4] Roberto Navigli. A quick tour of word sense disambiguation, induction and related approaches. In *Proceedings of The 38th Conference on Current Trends in Theory and Practice of Computer Science*, pp. 115–129, 2012.

[5] 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, 2010.

[6] 新納浩幸, 佐々木稔. 外れ値検出手法を利用した新語義の検出. 自然言語処理, Vol. 19, No. 4, pp. 303–327, 2012.

[7] 西尾実, 岩淵悦太郎, 水谷静夫（編）. 岩波国語辞典. 岩波書店, 1994.