7Q-04

統計的学習モデルを利用した日本語慣用句の意味的曖昧性解消

宮田　周 †　　竹内　孔一 †

† 岡山大学大学院自然科学研究科

1　はじめに

文の意味を構造化する上で動詞の語義を同定することは必須のタスクである．動詞の語義は動詞と係り関係にある言葉との共起によって決定されるが，慣用句は同じ係り元との共起であっても意味が異なることがあり，取り扱いが特に難しい．例えば「骨が折れる」には，体の骨が折れるという字義的な意味と，苦労するという慣用句的な意味が存在する．この意味によって，文の述語が「折れる」なのか「骨が折れる」なのかが変わるため，文の構造化において慣用句の曖昧性解消が重要になる．先行研究では，SVM[1] を用いた機械学習による日本語慣用句の意味的曖昧性解消が行われている．

本研究では， 統計的学習モデルを利用した意味的曖昧性解消手法を提案する．実験の結果，提案手法が先行研究の性能を上回ることを確認した．

2　先行研究

橋本らは日本語慣用句コーパス [2] を構築している．これは人手によって意味的曖昧性を認められる慣用句を集め，各慣用句の用例を収集し整理したものである．各用例には，用例中の慣用句表現の出現位置と，それが字義的な意味か慣用句的な意味かの情報が付与されている．全体で慣用句 146 句，用例 101500 文が掲載されている．

また橋本らは日本語慣用句コーパスを利用して，慣用句の意味的曖昧性解消を行っている．教師あり学習により曖昧性解消実験を行い，結果として正解率 89.19 %を得ている．学習モデルとして SVM を利用し，素性には慣用句表現の周辺形態素，係り元形態素，係り先形態素などの表層や品詞といった情報を用いている．素性抽出のための形態素解析器には KNP[*1] を用い，KNP が出力する意味的情報も素性に用いている．

3　慣用句意味的曖昧性解消

3.1　BACT による曖昧性解消

本研究では，Boosting Algorithm for Classification of Trees[*2]（以下 BACT）を用いた曖昧性解消を提案する．BACT は Boosting アルゴリズムを用いたラベル付き順序木の分類器である．BACT は弱学習器に Decision Stump（決定株）を用いている．入力された順序木から部分木を生成し，部分木の有無を素性とした Decision Stump を弱学習器とする．

BACT は順序木を入力とするため，係り関係のような文の構造を考慮した分類が可能なことが利点として挙げられる．本研究で素性に用いた 5 種類の順序木を以下で説明する．

- N-gram 木は文中の各形態素の原型を出現順に並べた木である．つまりこの木の各部分木が各単語の N-gram になっている．解析には CaboCha[3] を用いた．
- 係り受け木は各形態素の原型を係り受け関係に沿って並べた木である．
- 品詞木は各形態素の品詞を係り受け関係に沿って並べた木である．
- 名詞カテゴリ木は係り受け木中の名詞を名詞カテゴリ [4] に置き換えた木である．
- 拡張係り受け木は，係り受け木中の慣用句表現の末尾形態素に意味的な情報を付与した木である．付与する情報には，意味役割付与システム ASA[*3]による解析結果を用いた．

Word Sense Disambiguation of Japanese Idiomatic Expressions Using Statistical Learning Models

† Shu Miyata　Koichi Takeuchi

† Okayama University

[*1] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

[*2] http://chasen.org/~taku/software/bact/

[*3] http://cl.cs.okayama-u.ac.jp/study/project/asa

3.2 Word2Vec を利用した曖昧性解消

先行研究では，単語の意味的な情報を表現する素性に JUMAN カテゴリと JUMAN ドメインを用いていた．これは KNP が出力する意味的な情報で，カテゴリは単語の上位概念を，ドメインは単語のトピックのような情報を表現する．例えば，「鶏」の JUMAN カテゴリは「動物」で，JUMAN ドメインは「料理・食事」である．しかし，これらはあらゆる単語に付与されている訳ではなく，慣用句表現の曖昧性に影響を与える単語の情報を拾えない可能性がある．また，先行研究では文中の全ドメイン・カテゴリを素性として用いているため，曖昧性とは関係ない単語の情報まで含んでいる可能性もある．

そこで本研究では，近年注目されている単語分散表現である Word2Vec[5] を素性に利用する．今回は，形態素の表層や品詞といった素性に加え，慣用句表現の前後３形態素，慣用句表現の先頭形態素の係り元形態素，慣用句表現の末尾形態素の係り先形態素の Word2Vec ベクトルを素性に用いる．これにより，慣用句表現の周辺単語の意味的な情報が学習されることを期待する．今回は，日本語 Wikipedia 全文と日本語慣用句コーパス全文を入力に，skip-gram モデルにより導出した 300 次元のベクトルを用いる．学習モデルには SVM を利用し，学習器は先行研究と同様に TinySVM[*4] を用いた．

4 実験と考察

実験データには，日本語慣用句コーパスのうち極端に用例の少ない慣用句を除いた慣用句 122 句（用例数 94650）を用いた．10 分割交差検定による評価を行った．評価指標には Precision/Recall/F1 を用いた．

また，実験環境を揃えるため，先行研究で用いられている素性を用いた実験も行い，全ての手法において同様のデータを用いた．表 1 にその結果を示す．

表 1 曖昧性解消実験結果

手法	Precisicion	Recall	F1
先行研究	0.8951	0.8842	0.8864
BACT	0.8565	0.8477	0.8389
W2V	**0.9034**	**0.8892**	**0.8920**

表 1 の「BACT」の数値は，５種類それぞれの順序木を学習させた BACT の結果を用い，さらに Boosting を行った結果である．

結果から，全指標で Word2Vec ベクトルを素性に用いた SVM が最も性能が良かった．これより，Word2Vec ベクトルが慣用句の曖昧性解消に有効な素性であることが分かる．Word2Vec ベクトルは慣用句の前後３語と慣用句と係り関係にあるもののみを用いているため，そうした周辺単語の情報が特に有効であることも分かった．

一方，BACT の結果は先行研究の性能を下回った．これは，文の構造的な情報が曖昧性解消に有効でない可能性を示している．また，Word2Vec を用いた手法で有効であった慣用句周辺単語の意味的情報を用いていない点も，性能が下回った原因として考えられる．

5 おわりに

本稿では，BACT による曖昧性解消手法と，Word2Vec ベクトルを素性に用いた曖昧性解消手法を提案した．日本語慣用句コーパスを学習データに用いて実験を行い，慣用句の意味的曖昧性解消に慣用句の周辺単語の Word2Vec ベクトルが有効であることを確認した．今後の展望として，Word2Vec ベクトルを素性に用いる単語の範囲の拡張や特徴的な単語の選択，また他の学習法において Word2Vec ベクトルを用いる方法を考えている．

参考文献

[1] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1998.

[2] 橋本力, 河原大輔. 日本語慣用句コーパスの構築と慣用句曖昧性解消の試み. 情報処理学会研究報告 2008-NL-186, pp. 1–6, 2008.

[3] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, 2002.

[4] 森安祐樹, 竹内孔一. サ変名詞を含む複合名詞の語義解析システム及び名詞辞書の構築. *NLC2011-31*, pp. 51–56, 2011.

[5] T. Mikolov, et al. Distributed Representations of Words and Phrases and Their Compositionality. *Proc. of NIPS 2013*, 2013.

[6] 池田吉優, 竹内孔一. 意味役割と述語の概念を付与するシステムの構築. *NLC2014-39*, pp. 55–60, 2014.

[7] T. Kudo and Y. Matsumoto. A Boosting Algorithm for Classification of Semi-Structured Text. *EMNLP 2004*, 2004.

[*4] http://chasen.org/˜taku/software/TinySVM/