

極性を考慮した同義語による文書分類 Document Classification based on of Synonyms

逆瀬川 滉大[†]
Kota Sakasegawa

廣川 佐千男[‡]
Sachio Hirokawa

1. はじめに

同じ意味の内容を表すとしても、その文書を書く人により異なる単語で表現することがある。また、同じ単語でも、文脈で異なる意味を持つことがある。前者について同じ意味を持つ異なる単語は、同義語と呼ばれる。WordNet [Miller 90] は、このような単語とその意味について、同義語や上位・下位概念をまとめたシソーラスである。文書をベクトル化するとき、このような同意語や上位語を考慮することで、検索性能や文書識別性能の向上が期待される。[Scott1998]では、WordNet における単語の上位下位に基づき上位語密度という概念を導入し、識別困難な課題について上位語密度を用いたベクトル化で識別性能が向上する場合があることを示している。ただし、一般的な場合に常に性能向上するとは限らないことを示している。[Hotho2003]では、文書を単語のベクトルとして表現する代わりに、WordNet の synset のベクトルとして表現することで、クラスタリングの性能が向上することが述べられている。[福本 2002]では WordNet の 25 種類の各意味クラスについて限定し、同意語と上位関係をもちいたベクトル化により、Reuters21578 のデータセットで識別性能が向上するか否かを調べている。その結果、10 種類の意味クラスでしか向上がなく、正例の数が比較的少ないカテゴリについて有効だが、一般的には WordNet を利用しない場合より精度が下がることを示している。[上嶋 2004]では、各単語についての意味(synset)の頻度を用いることで、分類性能が向上することが述べられている。これらの先行研究では、WordNet の利用が必ずしも識別性能の向上に寄与するわけではないことが知られている。本稿では、同義語を考慮してベクトル化したとき、識別性能向上が損なわれる理由を、識別に対する単語の極性にあるのではないかと考えた。つまり、同義語であっても、識別対象文書群の正例の特徴語であるものと、負例の特徴語であるものを同一視すれば、逆に識別性能は低下するのではないかと考えた。本稿では、実験データセットとして 20 newsgroups を使い、この仮説を検証した。

2. 文書分類のための特徴語と単語識別重要度

本稿では、できるだけ少ない属性で文書分類を行うことを最終的な目的とした。そのために、正例の特徴語であるか負例の特徴語であるかといった識別極性を考慮したうえで、同義である単語をまとめ、属性選択を用いて文書分類を行った。本章では、この識別極性の考慮と属性選択を行う際に用いた SVM スコアについて説明する。

SVM スコアとは、[Sakai&Hirokawa2012] らによって提案された、適切な属性選択を行うための単語重み付け手法である。SVM スコアは、分析対象の文書を使った学習により得られる SVM モデルから、各単語に対応する重みベクトルを抽出することで得られる。具体的には、ある単語 w の SVM スコアは、1 つの単語 w のみで構成された仮想文

書を SVM モデルに適用したときの出力と等しい。SVM スコアが正の単語は正例の特徴語であり、SVM スコアが負の単語は負例の特徴語というように、割り当てられた SVM スコアの符号でその単語の識別極性が得られる。また、SVM スコアが大きいものは正例の特徴語としてより重要であり、SVM スコアが小さいものは負例の特徴語としてより重要というように、SVM スコアの絶対値で特徴語としての重要度が得られる。

3. 同義語の識別極性

同義である単語を集めて同義語集合を作ったときに、その集合内の単語の識別極性の正負は混在するのかを調べた。文書データとしては 20newsgroup データセットを用いた。まず、データセットの文書群を学習した SVM を作成し、文書群中に出現する各単語についての SVM スコアを得た。次に、WordNet を用いて文書群中に出現する単語で同義であるものを集め、同義語集合(synset)を作成した。さらに、その同義語集合に属する単語の SVM スコアの符号、つまり識別極性を調べた。その結果、集合内で識別極性が異なる単語が混在しているような同義語集合があることがわかった。紙面の都合上、5 クラス分の結果を表 1 に示す。

表 1. 20newsgroup における synset の識別極性

クラス名	synset 数	識別極性が混在していた synset 数
alt.atheism	49980	9270
comp.graphics	49980	7399
comp.os.ms-windows.misc	49980	7186
comp.sys.ibm.pc.hardware	49980	8019
comp.sys.mac.hardware	49980	8419

なお、WordNet の synset を利用する場合、単語に対する synset は複数あるが、それらは頻度順に並んでいる。synset を使ったクラスタリングなどの論文では、最初の synset を利用するのが全部を使うより分類性能が良いことが知られている[Hotho2003]。本稿の実験でも、単語に対する synset は WordNet で各単語についての synset で最初に現れるものを使った。

4. 20newsgroup データセットでの実験

20 newsgroups データセットに対し、同義語集合の識別極性を考慮する効果について確認した。3 通りの実験を行った。1 つ目は同義語をまとめる操作を行わず、SVM スコアによる属性選択のみを行って識別性能を評価するもので

[†]九州大学 Kyushu University

[‡]九州大学情報基盤研究開発センター

Research Institute for Information Technology
Kyushu University

ある。2つ目は WordNet で同義である単語をまとめ、SVM スコアによる属性選択を行って識別性能を評価した。3つ目は、同義語をまとめる際に、識別属性が正のものと負のものそれぞれまとめて、SVM スコアによる属性選択を行って識別性能を評価した。なお、文書のベクトル化には bag-of-words を、文書分類のモデルには線形 SVM を用いた。また、属性選択の際の単語数は 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000, 20000 で行った。

紙面の都合上、alt.atheism クラスと comp.graphics クラスの結果を図1, 2に記載する。結果として、同義語をまとめた方が性能は若干ながら低下した。また、同義語の識別極性を考慮することの効果は確認できなかった。同義語を考慮して性能が低下したことに対しては、同義語のまとめる精度が低かったことが考えられる。ある単語に対して複数ある候補 synset に対して、それらが頻度順に並んでいることから最初に出現する synset を用いたが、必ずしもそれが文脈上の意味と一致しているとは言えない。この方法では文書に本来存在しなかった意味の単語を入れてしまう可能性がある。続いて、識別極性を考慮しても性能にほぼ変化がなかったことに対して考察するため、表1に示したような識別極性が混在した synset について、synset 内の識別極性の割合と、SVM スコアの絶対値の2点をさらに調査した。その結果を表2に示す。識別極性の割合については、synset 内の単語の極性が混在していても、正負どちらかの数が極端に多いならば、同義語の識別極性を考慮する意味が薄いと思われる。逆に正：負=5:5 というように均衡して混在しているならば、同義語の識別極性考慮に期待できる。これに関しては、各クラスとも比較的均衡していることが確認できた。しかし、synset 内の識別極性の割合が均衡して混在していたとしても、そのような synset 内の単語がそもそも識別にあたっての重要度が低いものばかりならば、それらの単語について考慮する意味があまりない。表2を見ると、識別極性が混在する synset 内の単語の SVM スコアの絶対値は小さく、識別にあたってあまり重要でない単語が多かったことがわかる。これが原因で、識別極性を考慮しても効果が得られなかったことが考えられる。

表2. synset 内の識別極性の割合と平均 SVM スコア

クラス名	識別極性の割合	synset 内平均 SVM スコア
'alt.atheism'	0.4515	0.0156
'comp.graphics'	0.4532	0.0337
'comp.os.ms-windows.misc'	0.4546	0.0281
'comp.sys.ibm.pc.hardware'	0.4530	0.0362
'comp.sys.mac.hardware'	0.4519	0.0319

5. まとめと今後の課題

精度を保ちながらより少ない属性で文書分類を行うために、識別極性を考慮して同義語をまとめることを試みたが、識別極性を考慮した効果は確認できなかった。同義語をまとめるにあたって文脈上の意味の考慮が不足していたことと、識別極性が混在しているような synset 内の単語の、識別の際の重要度が低かったことが原因だと思われる。今後

は、文脈上の意味を考慮して同義語をまとめることが課題となる。

参考文献

- [1] S. Scott, S. Matwin, Text Classification Using WordNet Hypernyms, 1998, in: COLING/ACL Workshop in Usage of WordNet on NLP Systems, pp.45-51
- [2] A. Hotho, S. Staab, G. Stumme, 2003, Ontologies improve text document clustering, Proceedings - IEEE International Conference on Data Mining, ICDM, pp.541-544
- [3] 福本文代, 鈴木良弥, 2002, WordNet の同義語クラスとその上位関係を利用した文書の自動分類, 情報処理学会論文誌, Vol.43, No.6, pp.1852-1865
- [4] 上嶋宏, 三浦孝夫, 塩谷勇, 2004, 同義語, 多義語の考慮による文書分類の精度向上, 電子情報通信学会論文誌. D-I, 情報・システム, I-情報処理, Vol.2, pp.137-144
- [5] Tom Mitchell, Machine Learning, McGraw Hill, 1997.
- [6] T. Sakai, S. Hirokawa, Feature Words that Classify Problem Sentence in Scientific Article, Proc. iiWAS2012, 360-367, 2012

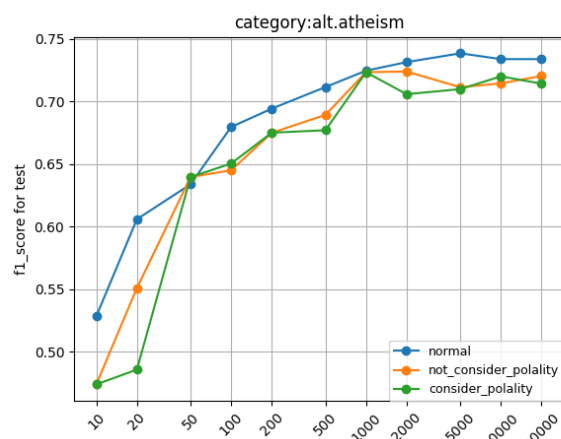


図1 alt.atheism クラスにおける識別極性考慮の効果

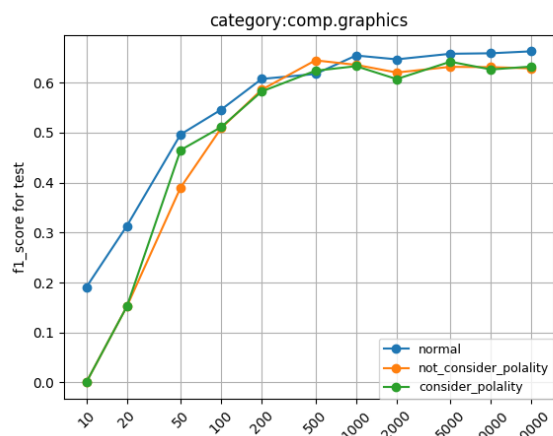


図2 comp.graphics クラスにおける識別極性考慮の効果