品詞情報とルールベースによる否定表現有無の判定

山下 紗苗 † 上 泰 † 奥村 紀之 ‡
† 明石工業高等専門学校 電気情報工学科 ‡ 大手前大学 現代社会学部
e1437@s.akashi.ac.jp, kami@akashi.ac.jp, noriyuki@otemae.ac.jp

1 はじめに

本研究では，日本語文章に否定表現が含まれるかどうかを，品詞情報のみに頼らず判定する方法について検証している．否定表現についての研究は主に言語学の分野でなされているが，文を構成する単語の品詞情報が必須となっている．品詞情報を得るためには形態素解析器を用いることが多いが，辞書にない否定表現は検出不可能であり，常に文法的に正しい品詞推定や単語区切りができるとは限らない．

本研究では，形態素解析で得られる品詞情報以外の情報も併用することで，辞書にない形の否定表現が文中に含まれていても正しく判定できる手法の開発を目指す．実験として，品詞情報，代表的な否定表現の出現位置，代表的な否定極性表現の出現位置を素性として組み合わせた 3 つの手法を比較検討する．対象となる日本語文章は Lang-8 と Twitter である．

2 関連研究

言語学の分野では，日本語の否定表現に関する様々な書籍や論文が公開されている．日本語には，「ちっとも」のように文末に否定表現を要求する副詞がある [1]．このような語句を否定極性表現（Negative Polarity Items）と呼ぶ．郡司はこの否定極性表現について韻律と意味の関係を考察している [2]．

自然言語処理の分野では，何をもって否定表現とするか厳密な決まりはなく，各々で定義している．菅沼らは論文やレポートなどの文章を対象に，1 文字前，2 文字後などの細かいルールを設定した文字列マッチングで「ない」「ぬ」「まい」の活用形を抽出し，字面のみから否定表現の検出を試みている [3]．Precision は 9 割以上と報告している．松吉は否定の焦点検出システムを構築するための基盤として，日本語における否定の焦点をテキストにアノテーションする枠組みを提案し，楽天トラベルのレビューと新聞の文章からコーパスを構築している [4]．否定表現の検出は単語の字面と品詞情報から行っている．

3 否定表現有無の判定手法

字面のみによる判定手法，品詞情報も用いた判定手法の 2 つを参考に，否定表現有無の判定手法を検討する．まず，既存手法の問題点を 3.1 節で検討する．次に，本研究において否定表現と判定する語句，しない語句を 3.2 節で定める．3.3 節では 3.1 節の問題点をふまえて新たな否定表現有無の判定手法を提案する．

3.1 既存手法の問題点

菅沼ら [3] は字面のみから否定表現を検出するために細かいルールを設定したが，ルールから少しでも外れると全く検出ができないという問題がある．例えば「見ない」は検出できても，「見ねぇ」「見るもんか」「見やん」「見ん」などの砕けた表現や方言を検出するにはルールの追加が必要となる．また，全ての表現をルールにすることは現実的でない．

松吉 [4] は字面と品詞情報から否定表現の検出を行った．品詞情報の取得には形態素解析器を用いることが多いが，菅沼らの手法同様，形態素解析器の辞書にない単語は検出できない．さらに，形態素解析器がいつも文法的に正しい品詞推定や単語区切りをするとは限らない．これらのことから，形態素解析器から得られる品詞情報のみに頼らず，より少ないルールで否定表現の有無を判定できることが望ましい．

3.2 否定表現の定義

本研究では，松吉 [4] の提案した否定要素を基準とし，表 1, 2 のように否定表現を定義する．複合語については，肯定形が一般的に使用されないものは否定表現としない．

3.3 提案手法

形態素解析器を用いて単語の品詞情報と基本形を取得し，これをもとに否定表現の有無を判定したものをベースラインとする．ベースラインでは，3.2 節で述べた否定表現の定義より，表 2 の例にある「物足り」「仕方」「思わ」「なければなら」「ないといけ」「かもしれませ」「にもかかわら」「だけで」から次に出てく

表 1： 否定表現と判定する語句の定義

否定辞	助動詞「ない」「ぬ」「ん」，接尾辞「ない」，接頭辞「非」「不」「無」「未」「反」「異」
非存在の内容語	形容詞「無い」，名詞「無し」

表 2： 否定表現と判定しない語句の定義

複合語	「物足りない」「仕方がない」「思わず」など
否定以外の意味を持つ複合辞	「なければならない」「ないといけない」「かもしれません」「にもかかわらず」「だけでなく」

る否定表現までの文字列を削除したものが表 1 のいずれかの語句を含むとき，その文章は否定表現を含むと判定する．例を (1) (2) に示す．

(1) それに，道路で色々な状況がありますので，まだ運転に慣れていない私は，相当集中しないといけないです．（否定表現を含む）

(2) 目を見てお話をしないといけないので，面接は怖いなぁ（否定表現を含まない）

提案手法では，ナイーブベイズと SVM を用いて，文中に否定表現が含まれているかを判定する．3.1 節で挙げた問題点をふまえ，形態素解析器を用いずに少ないルールで否定表現有無を判定できるような素性を 3 つ提案する．素性 1 では，3.2 節で述べた否定表現の基準から，「ない」「ぬ」「ん」「非」「不」「無」「未」「反」「異」「無い」「無し」「物足り」「仕方」「思わ」「なければなら」「ないといけ」「かもしれませ」「にもかかわら」「だけで」の 19 項目について，それぞれの否定表現が文中のどのあたりで初めて出現するかを素性とする．具体的な値は (出現位置) / (文の文字数) で求め，出現しない場合は 0 とする．例えば項目「ない」について，例 (3) の場合は 0.9，例 (4) の場合は 0 となる．

(3) 意味がよくわからない（9/10 = 0.9）

(4) 我々は本当に雑談 bot が欲しいのか？(0)

素性 2 では，素性 1 に，ベースラインでの否定表現有無の判定結果を追加する．素性 3 では，素性 2 に，否定極性表現の「しか」「も」が文中のどのあたりで初めて出現するかを追加する．素性の一覧を表 3 に示す．

4 否定表現の検出実験

本節では，3.3 節で提案した手法を用いて日本語文章から否定表現の有無を検出し，それぞれ評価を行う．本実験に使用したデータセットなどを 4.1 節で示す．4.2 節ではベースラインの評価，4.3 節では各素性の評価を行う．

4.1 データセット

Lang-8 が公開している日本人による添削文章 (Lang-8 Corpus of Learner Japanese[1]) および筆者の Twitter ログから，筆者自身が否定表現の有無をアノテーションした 3 種類のデータセットを用いた．内訳を表 4 に示す．なお，All は Lang-8 と Twitter のデータセットを合わせたものである．学習にはデータセットの 3/4，評価には 1/4 のデータを充てた．

また，形態素解析には MeCab[2]を，システム辞書には NEologd (v.0.0.6)[3]を使用した．

4.2 ベースラインの評価

ベースラインの評価値を表 5 に示す．本実験では，否定表現を含むと判定された事例にそうでない事例が混ざらないようにするため，Precision を重視する．

否定表現を含むがシステムによって含まないと判定された事例（FN）には，表 1 のルールにない平仮名の形容詞「ない」や助動詞「やん」「へん」「ひん」を含むもの，例 (5) のように形態素解析器の辞書が原因で，期待通りに単語区切りができない事例などがあった．

(5) あれ/は/私/の/髪/では/ありません/，/かつら/です/.（名詞：有馬線）

否定表現を含まないがシステムによって含むと判定された事例（FP）には，例 (6) のような「なければならない」のバリエーション，例 (7) のような「かもしれません」のバリエーションが見られた．その他，例 (8) のような疑問，提案，依頼を表す「ない」や例 (9) のような同意を求める「ない」なども見られた．

(6) (少し富山弁) いやはや，もうちょっと寝ないと…

(7) やば　シンカリオンめっちゃ好きかもしれん

(8) 皆さん，もし 3D カメラを知っていたら教えてくださいませんか．

(9) 制限行為能力者って名前格好良くないですか

[1] http://cl.naist.jp/nldata/lang-8
[2] http://taku910.github.io/mecab
[3] https://github.com/neologd/mecab-ipadic-neologd

表 3： 素性の組み合わせ

	ベースライン	素性 1	素性 2	素性 3
品詞情報と基本形によるマッチング	○		○	○
否定表現の出現位置		○	○	○
否定極性表現の出現位置				○

表 4： データセットの内訳

データセット	否定表現あり [件]	否定表現なし [件]	合計 [件]	文字数の平均値	文字数の中央値
All	888	1982	2870	32	26
Lang-8	569	703	1272	25	22
Twitter	319	1279	1598	37	31

4.3 各素性の評価

素性 1〜3 の評価値を表 6〜8 に示す．本節では，誤判定された事例の特徴を素性ごとに調査していく．データセット All を対象として，誤判定された事例の特徴と事例数を表 9 に示す．なお，一つの事例が複数の特徴を持っていることもある．ナイーブベイズと SVM の評価データサイズはデータセット全体の 1/4 であるため，データ全てを用いたベースラインと比較するために事例数を 4 倍にして表示してある．

素性 1 では，ベースラインで見られたような「なければならない」のバリエーションや，疑問，提案，依頼の「ない」を含む文章の誤判定がほとんどなくなった．しかし，形容詞「ない」の活用形である「なかっ」「なく」「なけれ」，例 (5) に挙げた「ありません」など助動詞「ん」を含む文章の判定結果はベースラインよりも劣った．ナイーブベイズ，SVM ともこの傾向は同じであった．

素性 2 では，ナイーブベイズにおいて助動詞「ん」の誤判定がなくなり，品詞情報から判断した結果を素性に加えることの有用性が示された．一方で，SVM では「なければならない」のバリエーションと，疑問，提案，依頼の「ない」についての誤判定がなくなったが，ナイーブベイズの誤判定は素性 1 のときより増加した．

素性 3 では，他の素性と比較して形容詞「ない」とその活用形の誤判定が減り，特にナイーブベイズは誤判定の数がベースラインを下回った．このことから，否定極性表現を素性に加えることで，形容詞「ない」とその活用形の誤判定を軽減できることがわかった．助動詞「ん」やバリエーション「なければならない」，バリエーション「かもしれません」の傾向は，ナイーブベイズ，SVM とも素性 2 と同じであった．

5 おわりに

本研究では，日本語文章に否定表現が含まれるかどうかを，品詞情報のみに頼らず判定する方法について検証した．結果，単語の品詞情報と基本形によるマッチング，代表的な否定表現の出現位置，否定極性表現の出現位置を素性とした素性 3 が有用であるとわかった．誤判定された事例の特徴を調査した結果，どのタイプの誤判定を減らしたいかでナイーブベイズと SVM を使い分ければよいと考えられる．形容詞「ない」とその活用形，助動詞「ん」のように，否定表現が含まれるが含まれないと誤判定される事例（FN）を減らしたい場合は，ナイーブベイズを採用するとよい．対して，バリエーション「なければならない」，疑問，提案，依頼の「ない」のように，否定表現は含まれないが含まれると誤判定される事例（FP）を減らしたい場合は，SVM を採用するとよい．

誤判定が多かった形容詞「ない」とその活用形を正しく判定できるような素性を検討することが今後の課題である．

謝辞

本研究は JSPS 科研費 18K11455 の助成を受けたものである．

参考文献

[1] 中島平三, 外池滋生. 言語学への招待. 大修館書店, 1994.

[2] 郡司隆男. 日本語の NPI の韻律と意味. 神戸松蔭女子学院大学研究紀要 言語科学研究所篇, Vol. 9, pp. 17–30, 2006.

[3] 菅沼明, 倉田昌典, 牛島和夫. 日本語文章推敲支援ツール『推敲』における否定表現の抽出法. 情報処理学会論文誌, Vol. 31.6, pp. 792–800, 1990.

[4] 松吉俊. 否定の焦点情報アノテーション. 自然言語処理, Vol. 21.2, pp. 249–270, 2014.

表 5： ベースラインの評価値

	データセット	Recall	Precision	F_1	TP	FN	FP	TN
	All	0.832	0.932	0.879	739	149	54	1928
	Lang-8	0.861	**0.957**	0.907	490	79	22	681
	Twitter	0.781	0.886	0.830	249	70	32	1247

表 6： 素性 1 の評価値

学習手法	データセット	Recall	Precision	F_1	TP	FN	FP	TN
ナイーブベイズ	All	0.617	0.923	0.739	156	97	13	452
	Lang-8	0.561	**0.925**	0.698	74	58	6	180
	Twitter	0.686	0.892	0.776	83	38	10	269
SVM	All	0.490	0.932	0.642	124	129	9	456
	Lang-8	0.212	**0.966**	0.348	28	104	1	185
	Twitter	0.512	0.899	0.653	62	59	7	272

表 7： 素性 2 の評価値

学習手法	データセット	Recall	Precision	F_1	TP	FN	FP	TN
ナイーブベイズ	All	0.838	0.938	0.885	212	41	14	451
	Lang-8	0.955	**0.947**	0.951	126	6	7	179
	Twitter	0.793	0.897	0.842	96	25	11	268
SVM	All	0.530	0.971	0.685	134	119	4	461
	Lang-8	0.258	**1.0**	0.410	34	98	0	186
	Twitter	0.537	0.942	0.684	65	56	4	275

表 8： 素性 3 の評価値

学習手法	データセット	Recall	Precision	F_1	TP	FN	FP	TN
ナイーブベイズ	All	0.854	0.939	0.894	216	37	14	451
	Lang-8	0.894	**0.952**	0.922	118	14	6	180
	Twitter	0.793	0.897	0.842	96	25	11	268
SVM	All	0.526	**0.971**	0.682	133	120	4	461
	Lang-8	0.727	0.950	0.824	96	36	5	181
	Twitter	0.537	0.942	0.684	65	56	4	275

表 9： 誤判定された事例の特徴と事例数（NB：ナイーブベイズ）

		素性 1		素性 2		素性 3	
	ベースライン	NB	SVM	NB	SVM	NB	SVM
ルールにない形容詞「ない」とその活用形	132	224	364	152	332	**128**	324
ルールにない助動詞「やん」「へん」「ひん」	3	4	4	4	4	4	4
ルールにある助動詞「ん」	11	140	152	**0**	128	**0**	132
バリエーション「なければならない」	20	4	**0**	16	**0**	16	**0**
バリエーション「かもしれません」	6	4	4	4	4	4	4
疑問，提案，依頼の「ない」	9	**0**	**0**	12	**0**	12	**0**
同意を求める「ない」	6	12	12	12	8	12	8