漢字情報を利用した評価表現辞書の拡張法[1]

乾 孝司　村上 浩司　橋本 泰一
東京工業大学 統合研究院
{inui,murakami,hashimoto}@iri.titech.ac.jp

1　はじめに

近年，評判分析に関する技術が高い関心を集めている [12]．評判分析では,「良い – 肯定」や「悪い – 否定」のような，ある単語が肯定／否定のどちらの評価をあらわしやすいかに関する知識が集積され，利用される．このような知識集合は評価表現辞書と呼ばれる．評価表現辞書では，いわゆる評価を表す表現の他に,「渋滞 – 否定」など，ある特定の評価が導かれやすい事柄全般が含まれることもある．

我々は，現在，評価表現辞書を社会課題抽出に利用することを検討している．社会課題抽出 [16] とは，自然災害，少子高齢化，個人情報保護と言った，現代における社会的な課題群を整理，把握する取り組みの総称である．この取り組みの一環として，社会課題，および，その課題の現れた経緯や課題解決の過程など，社会課題に関連する一連の社会変化を表す用語を，評価極性付きで新聞記事や特許文書から収集する技術の開発を行っている．

これまで，評価表現辞書を（半）自動構築する技術 [4, 14, 5, 6, 13] が一定の成果をあげているとは言え，実用に耐える大規模な辞書を構築することは以前として難しく，既存辞書には被覆率の点で問題がある．さらに，評判分析等で利用される評価表現辞書は，Web 掲示板や Blog 等のテキストを想定して構築されており，既存辞書のエントリと我々が対象データとして想定している新聞記事や特許文書に現れる語は必ずしも一致しない．例えば，主要かつ典型的な評価表現は形容詞である [10]．また，Web 掲示板や Blog 等のテキストでは表記が揺れやすいため，既存辞書では，代表的な用語表記に加え，平仮名や片仮名表記を追加する等の工夫がなされる．しかしながら，一方の社会課題抽出では，上記のような用語ではなく，むしろ「豪雨災害」や「医療過誤」といった漢字熟語をより多く収集することが望まれるが，このような語は既存の辞書構築手法では収集が困難である．また，漢字熟語は他の語との結合性が高く，その組合せ数の多さを考慮すると，人手で網羅的な辞書構築を行うには限界がある．

以上の背景から，本稿では，漢字熟語（漢字列）の評価極性を判定する手法を提案する．提案手法では，既存の評価表現辞書内の語が含む漢字の情報を利用して，コーパスから得られた未知漢字列（評価表現辞書にない漢字列）の評価極性を判定する．

[1]Augmenting Sentiment Polarity Dictionary Using Kanji-letter Information.

2　準備

2.1　前提

提案手法では，手元に，既存手法によって構築された小中規模の評価表現辞書があることを前提とする．文脈から明らかな場合，以降，評価表現辞書のことを単に辞書や初期辞書と呼ぶ．また，以降，肯定極性を「p」，肯定極性を「n」，肯定否定のどちらでもないものを「e」で参照することがある．同様に，コーパスから抽出されたある程度の規模の漢字列集合があることを前提とする．この漢字列集合のうち，提案手法によって評価極性が与えられ，かつ初期辞書に登録されていない語を新たに辞書登録することを狙う．

2.2　基本的な考え方

漢字列の評価極性を判定する際の基本的な考えについて述べる．本研究では，漢字が表意文字である点に着目し，**「同じ漢字を含む語の極性は一致しやすい」**という仮定に基づいて判定モデルを構築していく．ただし，我々がおこなう言語活動において，明示的／暗示的を問わず，何かを評価するという活動は，活動全体の一部分であることから，評価極性に関連する漢字も全漢字中の一部分であり，残りの大部分の漢字は評価極性と関連を持たないと考えるのが自然であるだろう．この考えに従うならば，ある漢字列があった場合，その漢字列の評価極性に影響を与えるのは，その構成素の一部分（以下，**核要素**と呼ぶ）であると考えられ，極性判定モデルを構築する際は，漢字列の中から核要素を見つけ，それらを適切に扱うことが重要となる．

3　基本法

基本法では，2 節で述べた仮定に基づいて，漢字列の極性判定モデルを構築する．具体的には，初期辞書に含まれる語から p か n のいずれかの極性を表す際に使われやすい漢字，すなわち核要素を決定リスト学習を用いて獲得する．昨今では，高性能な学習アルゴリズムが多く知られているが，学習されたモデルの視認性，保守性を重視し，決定リスト学習を採用した．ただ，もともとの決定リスト学習では初期辞書に存在しない漢字を考慮することができない．また，評価表現辞書には，通常，評価極性が p か n の語を登録し，e の語を登録することはないことから，決定リストの学習において，e の情報を不当に過小評価する可能性が高い．そこで，これらの問題に対応するため，ここではさらに，Yarowsky [11] のように，bootstrapping 法で決定リストを補強することを考える．

3.1 節で決定リスト学習について述べ，3.2 節で bootstrapping 法について述べる．

3.1 決定リスト学習

決定リスト学習（decision lists learning，DL）[9] とは分類問題を扱うための古典的な教師あり学習アルゴリズムであり，図 1 のような構造をもつ，信頼度つきの規則の群（決定リスト）を生成する．ある事例のクラスを判定する際は，決定リストの中で適用可能，すなわち条件部が事例と照合する規則のうち，もっとも高い信頼度をもつ規則に従ってクラスが割り当てられる．以下，決定リストの学習について詳細を述べる．

凡例	evi	$\rightarrow$	$class$	$(conf)$
具体例	好	$\rightarrow$	p	(0.98)
	無，益	$\rightarrow$	n	(0.67)

図 1: 規則の凡例と具体例

3.1.1 証拠の生成

初期辞書内の各語から，漢字ユニグラム，漢字バイグラムを抽出し，規則の証拠として利用する．例えば，「災害」に対しては「“災”，“害”，“災害”」の 3 つの証拠を生成する．ある事例が，証拠となる漢字ユニグラムあるいは漢字バイグラムを含む場合，その事例はその証拠に照合したことになる．

3.1.2 証拠に対するクラスの決定と信頼度の見積り

これまでに，数多くの信頼度指標が提案されている．各種指標およびそれらの間の関係は Fürnkranza et al.[3] によって簡潔にまとめられている．本研究では，一般的な指標の一つである条件つき確率 $P(class|evi)$ を信頼度として採用する．以下に具体的な計算法を示す．

ある k 番目の証拠 evi_k と照合する事例集合を $M^k = M^k_p \cup M^k_n \cup M^k_e$ とする．M^k_j（$j \in \{p, n, e\} = C$）は，M^k の部分集合で，クラス j の事例のみで構成されているとする．ここで，各クラスに対して，

$$f_j \quad = \sum_{m \in M^k_j} occur(m) \tag{1}$$

を求める．関数 $occur$ は，次節以降で述べる信頼度計算の説明のために導入された関数であり，事例の出現を表す．通常，$occur(m) = 1$ であり，この時，「事例 m が 1 度現れれば頻度を 1 カウントする」ことを意味する．

以上の準備のもと，evi_k に対応するクラス $class_k$，およびクラスの信頼度 $conf_k$ を次式で求める：

$$class_k \quad = \quad \arg\max_j \frac{f_j + \delta}{\sum f_j + \delta|C|}, \tag{2}$$

$$conf_k \quad = \quad \max_j \frac{f_j + \delta}{\sum f_j + \delta|C|}. \tag{3}$$

ただし，$|C|$ はクラス数，δ はスムージング・パラメータであり，評価実験では $\delta = 0.5$ とした．

3.1.3 デフォルト規則

信頼度の低い規則はクラス予測能力が低く，全体的な判定性能の劣化を招く．そこで，信頼度に対する下限 λ を設け，λ 未満の信頼度をもつ規則はリストから削除する．さらに，決定リストの末尾には以下のデフォルト規則を設ける．今回，評価実験では $\lambda = 0.5$ とした．

$$true \quad \rightarrow \quad e \ (\lambda).$$

3.2 Bootstrapping

決定リストを補強するために，bootstrapping 法 [11] を併用する．まず，基本的な手続きを以下に示す．

1. 初期辞書を利用して決定リストを学習する．
2. 学習された決定リストに従って，コーパスから抽出した漢字列の評価極性を判定する．
3. 判定結果を疑似教師ありデータとみなし，これと初期辞書を新たな教師ありデータとして，決定リストを再学習する．
4. 学習の結果，以前と同じ決定リストが学習されれば手続きを終了する．そうでなければ 2. へ戻る．

上記の手続きにおいて，学習された決定リストの誤り率に比例して，誤ったクラスをもつ疑似教師あり事例も増加する．このことから，疑似教師あり事例は，初期辞書に含まれていた教師あり事例に比べ，クラス情報に対する信頼性が劣る．これを踏まえ，以下の 2 点の操作変更を行う．

規則数の制限 Collins et al. [2] の Yarowsky-cautious アルゴリズムでは，繰り返し過程の中で学習される規則数に制限を設け，疑似教師あり事例に起因する誤りの影響を回避している．この考え方に従い，先の手続き 1. では，各クラスにつき最も信頼度の高い規則を一つづつ獲得し，以降，繰り返し毎に，学習される規則の最大数を各クラスに対して 1 規則づつ増やすこととした．これにより，疑似教師あり事例のクラス情報は繰り返し過程を進めることで，徐々にモデルに取り込まれる．

頻度のディスカウント 式 (4) に示す関数 $occur$ を採用することで，信頼度計算時の疑似教師あり事例の影響を減じる．ここで，$conf$ は，事例 m のクラス判定の際に適用された規則の信頼度であり，また，Dic は初期辞書内の語集合，$Corpus$ はコーパスから得られた漢字列の集合である．

$$occur(m) = \begin{cases} 1 & (m \in Dic) \\ 0.1 \times conf & (m \in Corpus) \end{cases} \tag{4}$$

4 基本法の改良

前節で基本法について述べたが，そこでの設定は，3.1.1 節での証拠生成の箇所を除けば，一般的な設定であると言える．本節では，漢字列データの持つ特性を考慮して改良を施す．

2.2 節の考えに従うならば，複合語を含めたすべての語の集合において，p か n のいずれかの評価極性をもつ語は一部分であり，残りの大部分の語の極性は e となる．これは，典型的なクラス不均質性 [7] の問題に該当する．つまり，単純に 3 値の評価極性を学習すると，多数派クラスとなる e に関し，偏った学習がなされてしまう．このクラス不均質性の問題に対し，基本法に次の改良を施す．

表 2: 漢字列の長さの影響（値は正解率，括弧内の値は β）

	漢字列の長さの上限					
DL+bootstrapping	1	2	3	4	5	6
基本法	0.555	**0.752**	0.587	0.482	0.480	0.483
改良法/改良A有	0.561 (1.0)	0.783 (1.0)	**0.798** (0.2)	0.663 (0.5)	0.588 (0.2)	0.559 (0.2)
改良法/改良A無	0.555 (1.0)	0.752 (1.0)	**0.795** (0.2)	0.745 (0.1)	0.727 (0.1)	0.730 (0.1)
考慮する漢字列の数	191	7,181	30,906	100,161	152,443	200,000

表 1: 対象から除外する名詞

名詞-固有名詞-一般	名詞-固有名詞-人名-名
名詞-固有名詞-人名-姓	名詞-固有名詞-人名-一般
名詞-固有名詞-組織	名詞-固有名詞-地域-一般
名詞-固有名詞-地域-国	名詞-副詞可能
名詞-代名詞-一般	名詞-数
名詞-接尾-人名	名詞-接尾-地域

表 3: 各手法の正解率

Baseline1	0.598
Baseline2	0.577
DL	0.654
DL+bootstrapping（基本法）	0.752
DL+bootstrapping（改良法）	**0.798**

改良A 学習によってクラス e の規則が多く得られることは，クラス判定時にクラス e の規則が過適用されることに繋がる．そこで，学習された規則のうち，クラスが e となる規則をすべて削除する．今回の場合，クラス e の規則を削除することは，その規則の信頼度をデフォルト規則の信頼度 λ に変更させたことと実質的に同じである．

改良B クラス e の事例の出現を過小評価することで，クラス不均質性に対処する．具体的には，規則の信頼度計算において，クラス e の事例のみ，式 (4) の代わりに式 (5) に示す関数 $occur$ を採用することで，クラス e の事例の影響を減じる．ここで，β は $0 \leq \beta \leq 1$ の定数である．

$$occur(m) = 0.1 \times \beta \times conf \tag{5}$$

5 評価実験

5.1 利用するデータ

高村ら [13] の手法の後，人手による修正を加え，評価表現辞書を作成した．評価実験では，この内，以下の条件 1. を満たす 573 語（p：318，n：255）を初期辞書として用いた．手法 [13] の制約により，初期辞書内に複合語は含まれない．語の平均文字数は 2.0 である．

次に，bootstrapping 法で利用する漢字列を新聞記事データから抽出した．条件 1. に加え，次の条件 2. および条件 3. も導入し，３つの条件すべてを満たした漢字列から無作為に 200,000 件を抽出し，実験に利用した．条件 3. は，固有表現や数（すう）は評価極性を持たないという考えに基づき，対象とする漢字列を絞り込むために導入したヒューリスティックである．なお，品詞情報は ChaSen[8] を用いて取得した．

条件 1. すべての文字が常用漢字 [15] である．

条件 2. 文字数が６以下である．

条件 3. すべての形態素の品詞が表 1 に列挙した細分類以外の名詞か，未知語である．

また，以下の手続きで評価セットを作成した．まず，先の 200,000 件の漢字列の一部を無作為抽出した後，第一著者が評価極性を人手で与えた．この時，単純に無作為抽出をすると，評価極性が e となる事例が大部分を占めた．そこで，評価極性が p か n となる事例がある程度確保できるまで，無作為抽出と極性付与を繰り返し，その後，e となる事例数をある程度削除することで 1,000 件（p：220，n：332，e：448）の評価セットを得た．この評価セットに対し，第一著者とは異なる２名の被験者による一致度を測定したところ，単純な一致率は 0.81，κ 値は 0.67 となることを確認した．

5.2 漢字列の長さの影響について

手法の性能比較を行う前に，bootstrapping 法に利用する漢字列の長さについて検討する．経験的に，漢字列の長さに限らず，漢字列内において核要素となる漢字の数は一定である．そのため，長い漢字列ほど，評価極性が p でも n でもない漢字を多く含んでおり，学習時にクラス不均質性の影響を受けやすくなると考えられる．そこで，利用する漢字列の長さに上限を設け，上限値を変化させることで複数の漢字列セットを作成し，各セットを利用して学習させた複数の決定リストの性能比較を実施した．

各セット毎の正解率を表 2 に示す．行が手法，列が長さの上限値である．表の各セルの値は，改良Bの β の値を $\beta = \{0.0, 0.1, 0.2, 0.5, 0.7, 1.0\}$ の間で変動させた際の最良値であり，括弧内はその時の β の値である．

まず，表 2 の各列の比較から，長い漢字列を利用することの効果がないことがわかる．今回の場合，長さ３までの漢字列を利用する場合が最良となった．また，考慮する漢字列が長くなるにつれ，最良な β の値は低くなる傾向がある．ここから，長い漢字列を考慮することでクラス不均質性の影響を受け，かつ，その影響を改良Bが軽減していることが確認できる．

5.3 手法の性能比較

各手法の正解率を表 3 に示す．Baseline1 は，初期辞書内の語との単純な照合に従って評価極性を判定した結果である．Baseline2 も同様であるが，ここでは，初期辞書内の語をすべて一文字づつ漢字に分割したもの

表 4: Baseline1 の結果

正解＼出力	p	n	e	Recall
p	47	11	162	0.214
n	2	107	223	0.322
e	3	1	444	0.991
Precision	0.904	0.899	0.536	

表 5: Baseline2 の結果

正解＼出力	p	n	e	Recall
p	133	80	7	0.605
n	72	253	7	0.762
e	137	120	191	0.426
Precision	0.389	0.558	0.932	

表 6: DL の結果

正解＼出力	p	n	e	Recall
p	140	72	8	0.636
n	29	296	7	0.892
e	135	95	218	0.487
Precision	0.461	0.639	0.936	

表 7: DL+bootstrapping（改良法）の結果

正解＼出力	p	n	e	Recall
p	120	60	40	0.545
n	17	286	29	0.861
e	41	15	392	0.875
Precision	0.674	0.792	0.850	

と判定事例を照合させる．DL は 3 節で述べた決定リスト学習単体の結果である．4 行目，5 行目は，表 2 で掲載した基本法，改良法の正解率の中で，最良である値を再掲している．また，基本法を除く各手法の結果の混合表を表 4 から表 7 に示す．

まず，表 3 の結果から，改良法が他の手法を上回っていることがわかり，提案手法の有効性が確認できる．次に，混合表から，Baseline1 は語そのものが証拠となっているため汎化性が低く，p および n の適合率が高い反面，再現率が著しく低い．その他の手法は，Baseline1 ほど顕著な傾向はなく，総合的に性能が改善されている．

表 7 において，p を n と誤ることで，p の再現率が下がっていることがわかる．事例分析から,「災害復興」のように，漢字列の構成素となる語が p と n の両方の評価極性をとる（「災害」は n,「復興」は p）場合にこのような誤りが多く起こることがわかった．また，本研究で敷いた仮定の基では,「偽造困難」のように，構成素の評価極性と漢字列全体の評価極性が一致しない事例の扱いは困難である．

6 おわりに

本研究では，漢字列の評価極性を判定する手法を提案し，正解率で 0.798 の結果を得た．実用面を考えば，正解率をさらに向上させることが要求される．今後，以下のような話題を検討する予定である．

- 今回はすべて漢字からなる語のみを扱ったが，仮名漢字混じり語についても検討する．
- 語境界，文字位置情報などを証拠に取り込む．
- 提案手法を Abney [1] のような数学的枠組みの基で捉え直すことで，現行モデルの動作原理の説明，およびモデル拡張の見通しを立てる．

参考文献

[1] Steven Abney. Understanding the yarowsky algorithm. *Computational Linguistics*, Vol. 30, No. 3, pp. 365–395.

[2] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[3] Johannes Fürnkranz and Peter Flach. An analysis of rule learning heuristics. Technical report, Technical Report CSTR-03-002, Department of Computer Science, University of Bristol, 2003.

[4] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th ACL*, 1997.

[5] Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. Using wordnet to measure semantic orientations of adjectives. In *Proceedings of the 4th LREC*, 2004.

[6] Hiroshi Kanayama and Tetsuya Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, 2006.

[7] Miroslav Kubat, Robert C. Holte, and stan matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, Vol. 30, pp. 195–215, 1998.

[8] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara. *Japanese Morphological Analyzer ChaSen Users Manual version 2.0.* Technical Report NAIST-IS-TR990123, Nara Institute of Science and Technology Technical Report, 1999.

[9] Ronald L. Rivest. Learning decision lists. *Machine Learning*, Vol. 2, No. 3, pp. 229–246, 1987.

[10] Janyce M. Wiebe. Learning subjectives adjectives from corpora. In *Proceedings of the 17th AAAI*, 2000.

[11] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of The 33rd ACL*, pp. 189–196, 1995.

[12] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, 2006.

[13] 高村大也, 乾孝司, 奥村学. スピンモデルによる単語の感情極性抽出. 情報処理学会論文誌, Vol. 47, No. 2, 2006.

[14] 那須川哲哉, 金山博. 文脈一貫性を利用した極性付評価表現の語彙獲得. 情報処理学会自然言語処理研究会 (NL-162-16), pp. 109–116, 2004.

[15] 内閣告示第一号. 常用漢字表, 1981.

[16] 大熊和彦. 新しい大学研究 -「ソリューション研究」の意義と課題. 研究・技術計画学会第 21 回年次学術大会, pp. 88–91, 2006.