

文末語を用いた社説内容の判別

石井 真人

Classification of Editorial Content using Sentence-final Words

Masato ISHII

Abstract

Editorials and commentaries clearly advocate the pros and cons of a particular topic. However it is difficult to find time to read these many useful documents.

The purpose of this study is to develop a method to determine the pros and the cons of a topic easily and quickly from documents such as editorials.

The sentence-final word used at the end of a sentence is a word that expresses the author's intention. In other words the author's claim, the pros and cons, can be determined by investigation of the sentence-final words frequency of the document.

Therefore, in this study, we have focused on sentence-final words, which are used at the end of sentences, and then a method of determining the content of a document has been investigated using the frequency of the sentence-final words.

The documents to be tested are editorials from the following newspapers: Asahi, Yomiuri, Mainichi, Sankei, and Tokyo from September to December 2022.

Three evaluators classified the content of the editorials into three categories: "proposal", "affirmative" and "negative". If the content of an editorial denies the subject matter, the evaluator judged it as "negative".

To classify the editorials automatically to the three categories 46 sentence-final words of which frequency are high have been selected from other editorials. These have been named to common sentence-final words.

Editorial contents have been determined to three categories using the common sentence-final words by the Quantification2. To calculate the discrimination rate, the values classified by evaluators have been set to the external reference values of the Quantification2.

The obtained results are shown below.

① Editorials have been discriminated by the common sentence-final words with a frequency of 8 or more (Group A) and the common sentence-final words with a frequency of 10 or more (Group B). But the overall average discrimination rate of Group A has been higher than Group B.

② The negative-discrimination rate for all editorial categories have been in the range of 30-35%. By newspaper, Asahi and Tokyo had a low negative-discrimination rate for "affirmative" category. Mainichi and Sankei also had low scores for "proposal". On the other hand, "negative" category in each newspaper has been a high negative-discrimination rate.

From these results, it is clear that this method is effective in determining the content of documents easily and quickly.

Keywords: editorial, newspaper, sentence-final word, Quantification2, discrimination rate

1 はじめに

筆者らは既報^[1]で、日本の代表的な4紙である朝日、読売、毎日そして東京新聞に関して、新聞の主張欄である社説の文末語が読者に与える印象を報告した。その結果、要望や提案する印象を与える文末語が多用されていることが分かった。たとえば毎日新聞は、情報提示型の文末語が多い。東京新聞は、否定の文末語も多いが、提案を促す文末語の頻度も高い。そして4紙間の共通した文末語による類似性は、読売、東京と毎日新聞の距離が遠く、3紙の類似性が低いことを示した。朝日新聞は、これら3紙の中間の距離に位置した。

以上の文末語使用頻度の研究結果から、社説は新聞社の主張を、読者に“報告・否定・提案”などの印象を与える文末語を用いて論じていることが分かった。そして、それらの使用頻度は、新聞社で異なることが明らかになった。しかし、社説の主張内容との関連性に関しては検討されなかった。そこで本稿では、社説の主張内容を文末語から簡便に推測する方法について検討した。

個人が特定のテーマに関する情報を全て網羅することは、限界がある。主張部分の詳細は不明でも、知りたいテーマに関する文献で「否定」あるいは「肯定」などの情報が分かれば、時間やエネルギーロスを軽減できる。社説や論説の主張する内容の多くは、タイトルから判別できることが多いが、できない場合や、タイトルが主張を直接的に示しているとは限らないこともある。

本稿では、社説の文末語使用状態が新聞紙間で特徴があることや、主張する分野で使用状態が異なることから、社説の主張する内容が文末語で判別可能と考えた。そして、全紙共通に使用されている文末語の使用頻度から社説内容の判別を検討した。文末語で内容の判別が可能であれば、個人が多く社説などを含む文章の内容判別を簡易に素早くできる。したがって、文章の詳細な主張内容は不明でも、多量の情報の判別が短時間で可能になる。

判別対象とした新聞は既報の4紙に産経新聞を加えた5紙である。以下2章では、文末語の抽出方法を含む実験方法を述べる。3章では、実験結果と考察を行う。そして4章で結論を述べる。

2 実験方法

本章では、文末語抽出の手法とそれを用いた社説内容の判別実験方法を述べる。はじめに、評価者による判別対象とする社説内容の判別を行う方法を述べる。つぎに一定頻度以上の文末語を抽出し、それを共通文末語とする手法を述べる。最後に共通文末語を用いて判別対象の社説の文末語と比較して社説内容の自動判別を行い、評価者との判別率を計算する方法を述べる。

2.1 判別対象とする新聞と刊行期間

本稿では、発行部数の多い朝日、読売、毎日と産経を選択した。また既報と同様の理由^[2]により、東京新聞を加えた。期間は、2022年9月1日から同年12月31日までに刊行された社説である。

2.2 評価者による社説内容の分類

自動判別との比較のため、3人の評価者が、5新聞の判別対象の社説を表-1の3種の判別項目に分類した。

〔表-1〕評価者による社説内容の判別項目と基準

判別項目	判別基準
A 提 案	新たな知見から主題に対する解決案として提案している内容が強い社説。
B 否 定	主題に対して異なる意見を述べておりジャーナリズム的色彩が強い社説。
C 肯 定	主題をほぼ肯定し、さらにそれを支援する内容が強い社説。

評価者は30～60歳代であり、特定の政党、宗教団体には属していない。社説の判別に費やした時間、場所、読み方は各人の自由である。

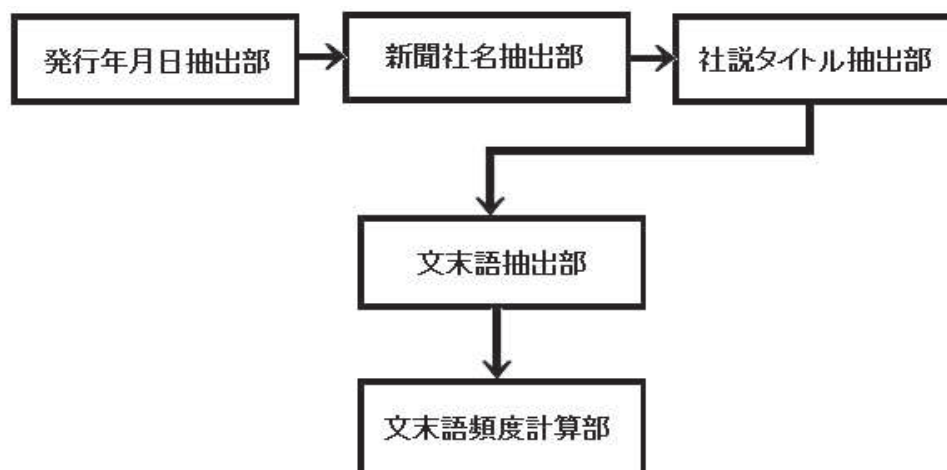
2.3 共通文末語の抽出

判別対象の5社の社説から、判別対象以外の期間で刊行した全社説の文末に用いられた文字列を抽出した。この文字列を文末語とする。文末語の内容は既報^[1]にならない文字数は日本語文字で5文字とした。つまり抽出する文末語は、本文の文末句読点を除く5文字である。抽出例を表-2に示す。なお、タイトルは含まれない。

〔表-2〕 文末語抽出例

文末語	文
加に転じた	武力紛争は、2010年を境に増加に転じた。
われている	尊重する社会をいかに作り出すかが問われている。
ありません	子孫の視点に立って考えるなら、こんな理不尽な話はありません。

近年はテキストマイニングが発展し、ネット上のテキスト分析やビジネスへの応用としてセンチメント分析で、顧客が書き込んだ文章の感情分析を行っている。しかし本稿では、特定文字から一定数の文字を抽出加工するため使いにくい。そこで文末語の抽出は、Perlで新たに記述した。抽出過程のブロック図を図-1に示す。抽出対象となる新聞の発刊日は、2022年1月1日から同年2月31日である。



〔図-1〕 文末語抽出の流れ

5紙から抽出した全文末語より、頻度8以上の46文末語を共通文末語とした。抽出結果の一部を表-3に示す。

〔表-3〕 抽出した共通文末語の一部と頻度

頻度	共通文末語例			
高	につながる	少なくない	されている	もらいたい
中	とみられる	めるべきだ	きではない	られている
低	ではないか	ばならない	としている	れなかった

2.4 共通文末語による社説内容の判別

2.4.1 文末語と社説内容

既報^[1]にて社説は新聞社の主張を、読者に“報告・否定・提案”などの印象を与える文末語を用いて論じていることを報告した。この結果から、社説の内容は多用される文末語の出現状態で判別できる可能性がある。本稿で論じる判別対象の社説内容とは、社説の主題に対して“提案・否定・肯定”の内容を示す。つまり、社説のテーマに新聞社が肯定的な内容を示しているならば、支援を意味する

単語と文を用いるとともに、それに適切な文末語を付けると推測される。たとえば、表-2では「もらいたい」は、否定的な文にも使うが肯定的や提案する場合に用いることがある。一方テーマに否定的な場合は、「きではない」や「ばならない」などが使用される可能性がある。また、「としている」などは、事実を報告している場合に使うことが多く、否定と肯定の両方に使用される。

これらのことから、文末語の出現状態は社説のテーマに強く関連しているから、この出現状態を数量化することで、社説の内容や主張がそのテーマに肯定的か否定的か判別できる。

2.4.2 共通文末語の使用状態の分類

共通文末語は、判別対象社説に使用されている可能性が高い。そこで共通文末語の使用状態で内容を分別するため、判別対象社説の共通文末語使用状態を使用頻度で3分類した。結果を表-4に示す。

〔表-4〕 共通文末語使用状態と使用頻度

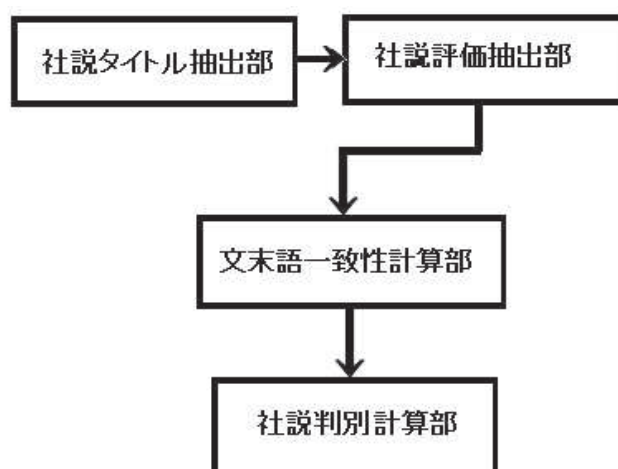
共通文末語の使用状態コード	使用頻度
1	1 社説中に対応する共通文末語が全く使用されていない。
2	1 社説中に対応する共通文末語が1回使用されている。
3	1 社説中に対応する共通文末語が複数回使用されている。

2.4.3 共通文末語による社説内容の自動判別

文章の自動解析は文章理解を始め多く研究されてきた。近年は、情報処理システムの飛躍的な向上により、テキスト文書の自動分類は、テキストマイニング（TM）の分野で重要な研究分野として多くの方面で国際的に発展してきた^{[3][4][5][6]}。一方個人レベルでは、データ処理に融通がきいて、小規模計算で簡易に設定できる手法で処理したい場合が少なくない。

本稿では、後者による手法で手軽にデータを加工でき、問題解決に効果的な手法について検討している。これらの手法で社説内容の判別を行う手法として、広く応用できる判別分析などがあげられるが、共通文末語データの連続性が無く、一社説中に含まれない語も多いため利用しにくい。そこで本稿では、数量化二類を用いて社説内容の判別を試みた。この方法であれば、共通文末語の使用状態コードで区分した量的な要因により、2022年9月1日から同年12月31日まで発刊した5新聞社説を質的な“提案・肯定・否定”の3外的基準に判別する方法として応用しやすい。

社説タイトル抽出から社説と共通文末語一致性までを計算し、最後に数量化二類による社説判別計算部までの一連の処理を図-2に示す。



〔図-2〕 数量化二類による社説自動判別の流れ

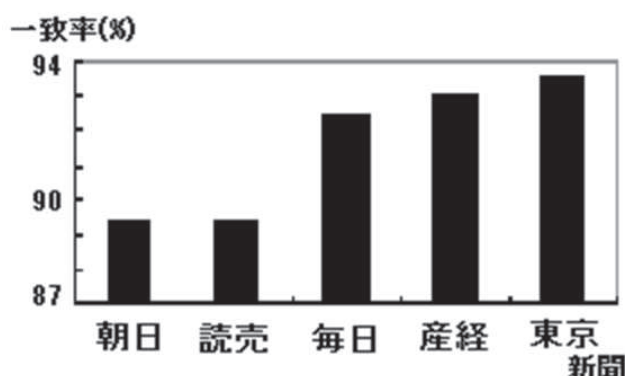
3 実験結果と考察

本章では、2章の方法に従って社説内容の判別を行った結果とそれについての考察を述べる。

3.1 評価者による一致性と判別対象社説

3人の評価者による社説内容の一致率の結果を図-3に示す。なお、全員が全ての社説を読んだわけではないため、2人以上読んだ社説を一致性の対象とした。そして、3人または2人のうち1人でも異なる判別があれば不一致とした。一致率は新聞により異なるが、89.3～93.6%の範囲であり全新聞に関して外的基準値としての使用は可能と判断した。結果を図-3に示す。

外的基準値として使用する条件は、2人以上の一致である。1人しか判別しなかった社説や3人の判別結果が全部異なった社説は自動判別対象から外した。この結果、判別対象とした社説数は、朝日、読売、毎日、産経、東京の順に、198、214、211、202、171である。



〔図-3〕 評価者による各新聞社説の一致率

3.2 共通文末語群と判別率の関係

共通文末語は対象とする同時期の社説から抽出した語では無い。それゆえ、46語のなかには判別対象社説で使用されない語が含まれている。そこで、共通文末語数と判別率の関連性を検討するため、頻度を変えた2つの共通文末語群を作成した。この共通文末語群を表-5に示す。

〔表-5〕 2群の共通文末語の頻度と語数

共通文末語群	頻 度 と 語 数
A	頻度 8 以上の46語
B	頻度10以上の30語

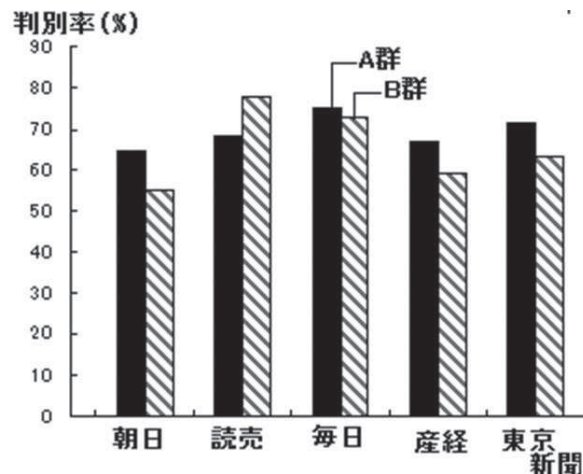
そして、評価者による判別項目と自動判別の一致性を検討するため、判別率を式-1で求めた。

判別率 (%) = 評価者による判別項目と数量化二類の結果が一致した社説数 / 社説数 (式-1)

また、誤判別率は式-2で求めた。

誤判別率 (%) = 100 - 判別率 (式-2)

数量化二類による判別率は、共通文末語A群では、64.7～75.4%の範囲にあり、平均値は68.8%であった（誤判別率は31.2%）。共通文末語B群では、55.0～77.6%の範囲にあり、平均値は65.0%であった（誤判別率は35.0%）。この結果から、共通文末語の多いA群が判別率は高いが、読売のようにB群の方が高い結果もあった。新聞別では、読売と毎日の判別率が両群とも高く70%超である。朝日、産経、東京はこの2紙よりも2～4%低い。この結果を図-4に示す。

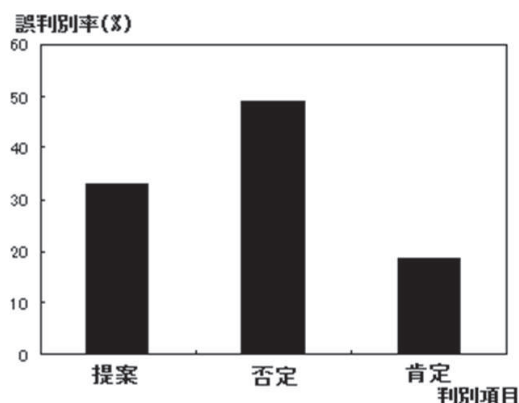


〔図-4〕 数量化二類による共通文末語群別判別率

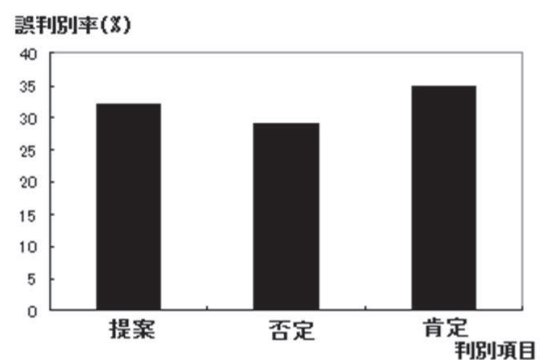
本実験ではA群の判別率が高く出たが、B群に対して平均値3.8%の差であり、両群で明確な有意差があると断定することは難しい。また新聞間での判別率の差の要因としては、共通文末語は対象とする期間外から抽出した語であるから、判別対象期間には使われなかった可能性もある。なお、以降は文末語数46語（A群）を用いた結果について考察する。

3.3 判別項目と誤判別率の関係

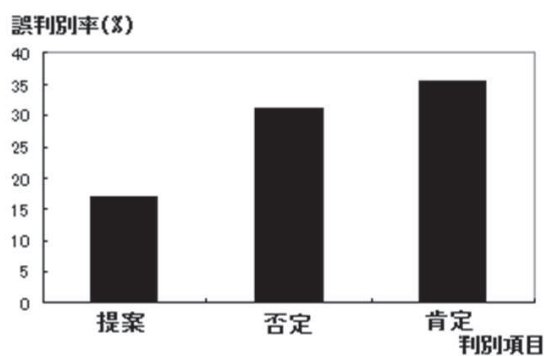
新聞ごとに3つの判別項目と各項目の誤判別率の結果を図-5～10に示す。



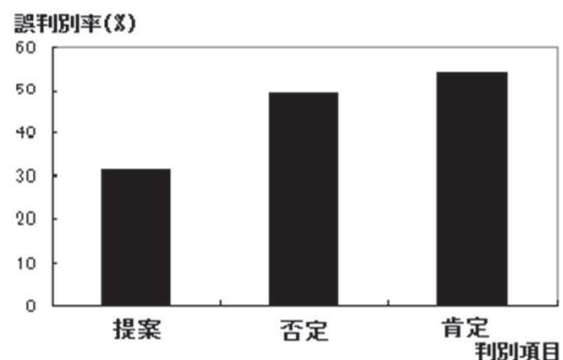
〔図-5〕 判別項目と誤判別率（朝日）



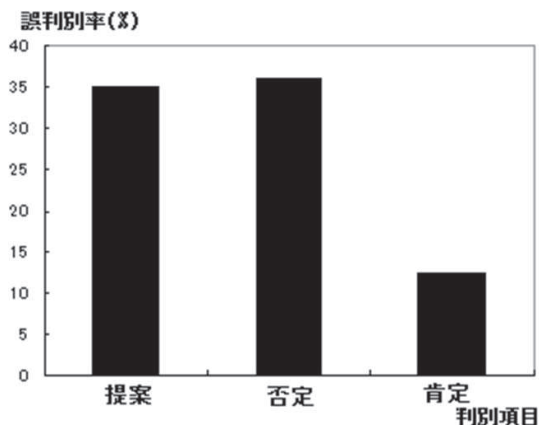
〔図-6〕 判別項目と誤判別率（読売）



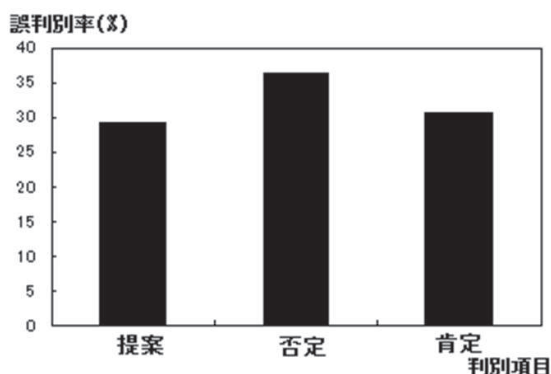
〔図-7〕 判別項目と誤判別率（毎日）



〔図-8〕 判別項目と誤判別率（産経）



〔図-9〕 判別項目と誤判別率（東京）



〔図-10〕 判別項目と誤判別率（全紙）

誤判別率は、図-10より「否定」が一番高いが全紙各項目とも30～35%の範囲であり、項目間の差は小さい。新聞別では、朝日と東京の「肯定」項目が低い（図-5、9）。「提案」項目では、毎日と産経が低い（図-7、8）。「否定」項目は、各紙とも高い誤判別率を示しており、「肯定」と「提案」とは異なっている。これは、「否定」内容に他項目とは異なる文末語を多用しているよりも、「肯定」と「提案」項目にも使われている共通文末語も多く使われている可能性が考えられる。

4 結論

メディアから流される情報のうち、個人が取り込んで理解する情報はごく一部である。知りたい情報の内容を短時間で判断したい場合は、ネットワークによる学習などが考えられるが、無料でユーザが自由に入力データを加工できるケースは少なく、得た大量の情報を理解する時間も限られる。そこで本稿では、情報も基幹的部分である主題に対する筆者の主張だけを取りあげ、文末語を用いて多量の文書から主張を簡易に短時間で判別する方法を検討した。

文の終わりで使われる文末語は作者の意志を表現するため、文の内容を左右する語が使われる場合が多い。それゆえ、文末語の使用状態を調べることは文章全体の主張を判別できる可能性がある。本稿では、文の終端に使用されている5日本語文字列を文末語として、文中の文末語使用頻度の状態から内容の判別を試みた。

文章の主張を「提案」、「肯定」、「否定」の3つに分類して、文末語の使用状態と比較して判別した。文末語の使用状態は、判別対象以外の文章から高頻度で使われる文末語を46語選択して、これを共通文末語とした。共通文末語は、2022年1月～2月の朝日、読売、毎日、産経、東京の5新聞の社説から抽出した。そしてこの共通文末語を用いて、同新聞の同年9月～12月の社説の判別を試みた。

文章内容の自動判別は、共通文末語の使用頻度に応じ3つのカテゴリに分類して数量化二類を用いて判別率を計算した。判別の外的基準は、3人の評価者により判別対象の社説を同じく「提案」、「肯定」、「否定」に判別した。得られた結果を以下に示す。

- ①評価者による判別は、2人以上読んだ社説を実験対象の社説とした。それらの一致率は新聞により異なるが、89.3～93.6%の範囲であり全新聞に関して外的基準値として高かった。
- ②頻度8以上の46共通文末語と頻度10以上の30共通文末語の2群で自動判別を行ったが、46語の方が判別率は高かった。
- ③自動判別による誤判別率は、全紙としては各項目とも30～35%の範囲である。新聞別では、朝日と東京の「肯定」項目が低い。「提案」項目では、毎日と産経が低い。「否定」項目は、各紙とも高い誤判別率を示しており、「肯定」と「提案」とは異なっている。これは、「否定」内容に、他項目にも使われている共通文末語も使われている可能性が考えられた。

以上の結果から、共通文末語を用いた文章の内容判別の手法は、判別率が64.7～75.4%の範囲にあり効果的であると考ええる。さらなる判別率の向上に向け、今後の課題として以下の点があげられる。

- ①適切な共通文末語数の決定法。
- ②適切な共通文末語の文字数の決定法。

参考文献

- [1] 石井真人 “新聞社説の類似性Ⅱ 文末語の印象”、メディア情報研究 5、p1-9、2020
- [2] “スペシャル・レポート 2 外国人記者が評価「日本のマスコミ」信頼度ランキング”、プレジデント、2016年 7 月18日号
- [3] 渡部勇、三木和夫 “単語の連想関係によるテキストマイニング”、情報処理学会 情報学基礎 55-8、p57-64、1999
- [4] P.V. Arivoli “Document Classification Using Machine Algorithms”, International Journal of Scientific Engineering and Research、2015
- [5] 八軒浩子、松岡有紗他 “医療薬学研究のテキストマイニングによる計量的分析—「医療薬学」と「日本病院薬剤師会雑誌」の比較” 医薬品情報学13-4、p152-159、2011
- [6] GUO Dui-ming, LI Guo-qing 他 “Big data analysis and visualization of potential hazardous risks of the mine based on text mining”, Chinese Journal of Engineering, p328-338、2022