# 決定木学習を利用した講演文からのトピックフレーズの抽出

伊藤 山彦　谷田 泰郎　柏岡 秀紀　田中 英輝
ＡＴＲ音声言語コミュニケーション研究所

## 1 はじめに

近年、大容量の通信環境が整備され、大量の情報が入手可能となる中、入手可能な大量の情報の中から個々の利用者に有用な情報を抽出する技術は、ますます重要性を増している。こうした必要に対し、利用者が文書の概要を速やかに把握するための技術として、従来から自動要約技術が研究されている。自動要約技術では、重要文抽出や文圧縮の手法により、文書から相対的に重要な箇所を抽出して要約文を生成する。しかし、重要性の基準は人によって曖昧であり[1]、機械的に生成された要約文は、必ずしも個々の利用者に有用な情報を提供するとは限らない。この問題に対し、我々は利用者が文書中で興味のある個所を参照するための支援を目的とした理解支援技術の研究を進めている。

本研究で検討中の理解支援技術は、(1)計算機は、文書から利用者が興味のある個所を参照するための手掛かりとなる句(トピックフレーズ)を抽出して利用者に提示する、(2)利用者が興味のあるトピックフレーズを指定したら、それを更に詳しく記述した文書の範囲(パッセージ)を提示する、という機能によって実現する。実現のためには、(a)トピックフレーズの抽出と、(b)トピックフレーズに対応したパッセージの抽出、が必要となる。本稿では(a)について述べる。我々が提案する手法では、対象文書から名詞句を抽出し、それぞれの名詞句に対して、トピック性の高さを決定すると考えられる複数の言語的特徴を抽出し、それらを属性とした決定木学習を行うことにより、トピックフレーズの判定を行う。本稿では、提案手法の詳細を述べるとともに、NHK のニュース解説番組「あすを読む」の書き起こし原稿に適用して行った実験結果について報告する。

## 2 トピック抽出の研究におけるアプローチ

トピック抽出の研究において、近年多く見られるのは、単語または単語の集合をトピックとして捉え、対象文書中に出現する単語の統計的な情報からトピックの同定を行う研究である(文献[2][3])。文献[2]では、トピック抽出を一種のテキスト分類として捉えている。文書内に出現した単語を分類項目とし、文書と各分類項目との関連度を、統計的な手法によって推定する。文献[3]では、テキスト内の単語の分布から、テキスト分割とトピックの同定を行う。これらの手法では、文書の種類に依存せずトピックを抽出できるが、トピックを利用者が理解可能な意味を持つ句として提示することはできない。

文献[4][5]では、文書中に現れる手掛かり表現を利用して、トピックとなる名詞句を抽出している。文献[4]では、話題展開の手掛かりとなる表現を抽出し、文頭表現を手掛かりに話題スコープ(範囲)の認定を行い、更に付属語表現を手掛かりに各話題のラベルとなる名詞句の抽出を行う。文献[5]では、名詞句のトピック性を決定すると考えられる複数の条件を定め、当てはまる条件の数を加算してトピック性の高さを判定する。これらの手法は、対象の文書の種類に応じて、手掛かり表現を１つ１つ手作業で抽出する必要があるため、汎用性に問題がある。

本研究では、上記のような従来研究の問題点に対し、理解可能性と汎用性を両立させるために、まず名詞句を抽出し、次に各名詞句のトピック性を機械学習によって判定するアプローチを採用した。

## 3 トピックフレーズ抽出処理

### 3.1 トピックフレーズ

本研究で呼ぶトピックフレーズとは、利用者が興味のある箇所を参照するための手掛かりとなる名詞句である。トピックフレーズを更に詳しく記述した文書の範囲をパッセージと呼ぶ。トピックフレーズは、対応するパッセージの見出しの役割も果たす。

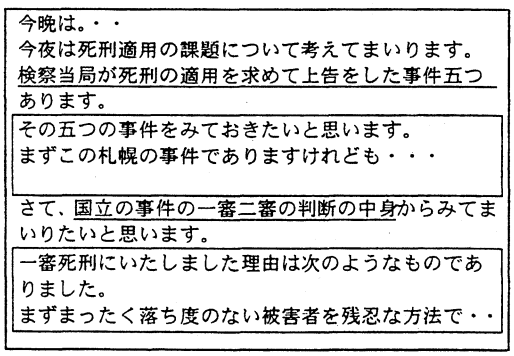


図１ トピックフレーズとパッセージ

図１に、トピックフレーズとパッセージの例を示す。下線を引いた部分がトピックフレーズ、四角で囲った部分がパッセージである。下線部に示された

名詞句を提示し、利用者が興味を持つ句を指定したら、その句に対応したパッセージを提示することにより利用者の理解支援を行うことが、本研究の目的である。

### 3.2 処理概要

トピックフレーズ抽出処理の流れを図2に示す。

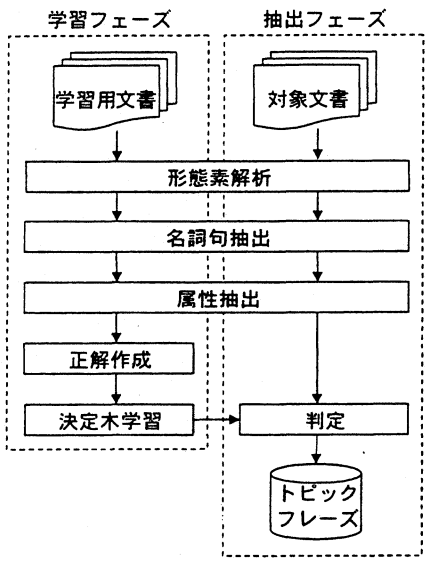


図2 処理の流れ

処理は大きく学習フェーズと抽出フェーズに分かれる。それぞれのフェーズにおいて、文書を形態素解析し、形態素解析結果を利用して名詞句を抽出する。次に、抽出した名詞句のトピック性に関係すると考えられる言語的特徴を学習の属性として抽出する。学習用文書から抽出した各名詞句に対して、トピックフレーズであるか否かの判定を人が行い、決定木学習のトレーニングデータとして利用する。学習で作成された決定木を利用し、対象文書から抽出した名詞句に対しトピックフレーズの判定を行う。以下、図2の形態素解析、名詞句抽出、属性抽出、及び正解作成の各処理について説明する。

### 3.3 形態素解析

文書はまず形態素に分割する。実験ではATRの「変換主導型翻訳システム(TDMT)」[6]の体系に基づいて、人手で形態素付与作業を行った結果を利用した。

### 3.4 名詞句抽出

形態素解析結果を利用し、以下の2つの処理によって名詞句を抽出する。

**(1) 形態素パターンを利用した抽出**

文書中に出現する以下の条件の名詞句を抽出する。

**(a) 文節内の名詞句**

単独の体言、体言の連続、または接頭辞+体言+接尾辞のように、全体として1語の体言の役割を果たす表現を文節内の名詞句として抽出する。

**(b) 連体文節または並立文節を伴う名詞句**

上記(a)で抽出した文節内の名詞句に、自立語+連体助詞[1]、自立語+並立助詞[2]、または活用語の基本形[3]の何れかの文節が前接するとき、前接の文節を含めた全体の範囲を1つの名詞句として抽出する。この様な文節が複数前接するときは、全体を抽出する。ただし、「・・の課題ということ」のように体言に引用的な表現が続く場合、引用的な表現を除き「・・の課題」までを抽出範囲とする。

**(c) 連体節を伴う名詞句**

上記(a)または(b)で抽出した名詞句の前に、文献[7]の節境界判定規則によって「連体節」と判定された表現が接続するとき、文頭からの全体の範囲を1つの名詞句として抽出する。ただし、節境界判定規則で判定された表現が名詞句の一部になりにくい表現(例:「が」で区切られる並列節)を含む場合、文頭からではなく、その表現より先を抽出範囲とする。

上記(a)～(c)で抽出された名詞句のうち、複数の形態素からなり、かつ5文字以上の名詞句をトピックフレーズの候補とし、正解作成の対象とした。

**(2) 言い換えによる抽出**

文書中に名詞句として現れない表現でも、名詞句に言い換え可能な表現は、言い換え規則を用いて名詞句に変換する。

```
(defrule rule1
 :in
 ((:surf "なぜ")
  (:id ?idA :surf ?surfA :pron ?pronA :org ?orgA :pos ?posA +)
  (:or (:surf "の") (:surf "ん"))
  (:surf "でしょう")
  (:surf "か"))
 :out
 ((:id ?idA :surf ?surfA :pron ?pronA :org ?orgA :pos ?posA)
  (:id 0 :surf "理由" :pron "リユウ"
   :org "理由" :pos "普通名詞"))
)
```

図3 言い換え規則

言い換え規則は、形態素の並びと、and、or、繰り返しの演算を用いて記述する(図3)。形態素は、表記(:surf)、読み(:pron)、原型(:org)、品詞(:pos)

[1] 「の」や「という」のような、体言に接続する助詞。
[2] 「と」や「や」のような、並立関係を表す助詞。
[3] TDMT体系では、活用語の連体形は基本形と表される。

のパターンで指定できる。

図 3 の言い換え規則は、「(表記が"なぜ"の形態素) (任意の形態素の 1 回以上の繰り返し) (表記が"の"または"ん"の形態素) (表記が"でしょう"の形態素) (表記が"か"の形態素)」のパターンに一致した表現を、前記「任意の形態素の 1 回以上の繰り返し」に「理由」を連結した名詞句に変換する。例えば、「なぜ今消費者契約法が必要とされているのでしょうか」という表現を、「今消費者契約法が必要とされている理由」という名詞句に変換する。

特にトピック性が高いと考えられる「なぜ〜か」「どう〜か」のような疑問詞的な表現を含む文に対し、人手で 42 個の言い換え規則を作成した。

文書中の同じ箇所に対し、形態素パターンによる抽出と言い換えによる抽出のそれぞれによって異なる名詞句が抽出された場合は、言い換えによる抽出を優先させる。また、文書中の同じ箇所に対し、複数の言い換え規則が適用された場合は、長く一致する方を優先させる。

### 3.5 属性抽出

名詞句のトピック性の高さに関連すると考えられる以下の言語的特徴を、学習の属性として抽出する。

**(1) 抽出した名詞句の最後の形態素**

「〜の目的」「〜の背景」のような名詞句はそれを更に詳しく説明したパッセージが存在する可能性が高く、名詞句の最後の形態素はトピック性に関連すると考えられる。抽出した名詞句の最後の形態素のタイプ(異なり)を学習の属性値とした。

**(2) 抽出した名詞句に後続する付属語列**

「〜という」「〜でありますけれども」のような提題的な表現を伴う名詞句はトピック性が高いと考えられるため、名詞句に後接する付属語列のタイプを学習の属性値とした。付属語列の中には、助詞や助動詞だけではなく、「こと」や「ある」のような形式的な自立語も含めた。漢字と平仮名の違いによる表記の揺れや、「では」と「じゃ」のような音韻の揺れに関する違いは吸収した。

**(3) 抽出した名詞句を含む文の文頭の形態素**

「さて」「ところで」のような語で始まる文は、話題を導入する文であり、このような文に含まれる名詞句は、トピック性が高いと考えられる。文頭に出現する品詞が接続詞、接続副詞、または副詞である形態素を抽出して学習の属性とし、それぞれの形態素の文頭における出現の有無を属性値とした。

**(4) 抽出した名詞句を含む文の文末表現**

「〜でしょうか」「〜ましょう」のように呼び掛ける文は話題を導入する表現になり得るため、文末表現は文中の名詞句のトピック性に関係すると考えられる。文末から最長 5 形態素の付属語列に対し、上記 (2) と同様な方法で属性値を設定した。

**(5) 抽出した名詞句を含む文の前の文の文末表現**

「〜というわけです」のように話題を総括する表現で終わる文の次の文は、新たな話題を導入する文になり得るため、前の文の文末表現は名詞句のトピック性に関係すると考えられる。上記 (4) と同様な方法で属性値を設定した。

**(6) 列挙表現**

「第一に」「第二に」のように、数え上げる表現を伴う名詞句は話題のポイントを示す可能性が高く、列挙表現の有無は、名詞句のトピック性に関係すると考えられる。名詞句を含む文の文頭における列挙表現の有無を属性値とした。

**(7) 疑問詞的表現**

「なぜ」や「どう」のような疑問詞的表現を伴う文は問題を提示する文である可能性が高く、疑問詞的表現の有無は名詞句のトピック性に関係すると考えられる。名詞句を含む文における疑問詞的表現の有無を属性値とした。

### 3.6 正解作成

名詞句抽出処理で抽出した名詞句に対し、人がトピックフレーズであるか否かの判定を行い、学習のトレーニングデータとした。判定は、「抽出された句をさらに詳しく記述したパッセージが文書中に存在するか否か」という基準に従って行った。同じパッセージを指す複数の名詞句が存在する場合は全て正解とした。

## 4 実験

### 4.1 実験方法

実験の対象としたデータは、NHK 番組「あすを読む」の書き起こし原稿 50 文書である。「あすを読む」は、主として時事問題をテーマとした 10 分間の番組であり、書き起こすと約 3000 字のテキストとなる。

実験対象の 50 文書に対して、まず、名詞句抽出処理により 5168 個の名詞句を抽出した。次に、人手で正解を付与し、651 個をトピックフレーズと判定した。作成した正解データに対して、45 文書を学習用データ、5 文書を評価用データとし、決定木学習 (C4.5) による 10 分割の交差検定を行った。

### 4.2 実験結果

実験結果を表 1 に示す。決定木学習によって、ト

ピックフレーズであると判定された名詞句は、494個であった。表1に、下限値として、651個の正解を持つ5168個のデータから494個をランダムに抽出した場合の再現率と適合率の値を付す。

表1 実験結果

| | 再現率 | 適合率 |
|---|---|---|
| 本手法 | 31.98 | 46.52 |
| 下限値 | 9.56 | 12.60 |

実験の結果、本手法による抽出の再現率・適合率は、下限値に比べ大幅に良好な値であり、本手法の有効性が確認された。

## 5 考察

実験の結果得られた再現率・適合率は、現段階で実用のために十分高い値とは言えない。しかし、実験データが50文書と、学習を利用した手法としては少なかったことを始め、名詞句抽出処理の精度向上や、表記揺れへの対処にも改善の余地があり、今後これらの点への検討を進めることにより、精度向上は期待できると考えている。

その他、今後検討すべき点について以下に考察を述べる。

### (1) 人による判定の揺れについて

トピックフレーズであるか否かの判定は、重要文抽出における重要性の判定と異なり、人の価値観に依存する相対的な尺度ではないため、判定の基準を細かく定めることにより、人による判定の揺れを小さくすることは可能と考えられる。

しかし、抽出すべきトピックの粒度など、判定の揺れに関係する問題は残る。例えば、「現在の状況については不明である」のような簡潔な表現においても、「現在の状況」をトピックフレーズと判定すべきか否かは、利用者の要求にも関わるものである。

また、名詞句抽出処理によって抽出された名詞句が人に理解可能な表現として不完全である場合、不完全さをどこまで許容するかに関しても、判定に揺れが生じる可能性がある。

今後、判定の基準を詳細に定めた上で、人による判定の揺れがどの程度生じるか検証する必要がある。

### (2) 話題の網羅性について

抽出したトピックフレーズは、全ての文書の範囲をカバーするとは限らないため、トピックフレーズから参照できないパッセージが生じる可能性がある。そのようなパッセージの参照を如何に実現するかについて、今後検討を行う必要がある。

### (3) 照応について

本稿で述べた実験では、名詞句に照応表現が含まれている場合には、照応が解消されたものとしてトピックフレーズであるか否かの判定を行った。しかし、実際にトピックフレーズを利用者が興味のある箇所を参照するための手掛かりとして提示するためには、照応を解消した形で提示する必要がある。照応の問題も今後の検討課題である。

## 6 まとめ

本稿では、決定木学習を利用した講演文からのトピックフレーズの抽出について述べた。文書から名詞句を抽出し、各名詞句に対してトピック性に関係すると考えられる言語的特徴を抽出して学習の属性とし、人による判定を正解データとして実験を試みた結果、本手法の有効性を確認した。

今後、考察した検討課題を踏まえ、さらに大規模なデータに対しても実験を試みるとともに、トピックフレーズに対応したパッセージの抽出についても、検討を行う予定である。

## 謝辞

本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

## 参考文献

[1] 伊藤ほか: "講演文を対象にした重要文抽出実験", 話し言葉の科学と工学ワークショップ講演予稿集, pp. 157-164 (2001).

[2] 野本: "確率モデルによる主題の自動抽出", 情報処理学会自然言語研究会 NL-108, pp. 1-6 (1995).

[3] 李ほか: "線形結合モデルを用いたトピック分析", 情報処理学会自然言語処理研究会 NL-139, pp. 61-68 (2000).

[4] 竹下ほか: "モノローグに対するブラウジング支援のための話題構造抽出", 情報処理学会論文誌, Vol. 37, No. 11, pp. 1919-1927 (1996).

[5] 伊藤ほか: "講演文を対象にしたトピックフレーズの抽出", 情報処理学会第63回全国大会講演論文集(2), pp. 151-152 (2001).

[6] 古瀬ほか: "構成素境界解析を用いた多言語話し言葉翻訳", 自然言語処理, Vol. 6, No. 5, pp. 63-91 (1999).

[7] 丸山ほか: "日本語における独話の特徴と文分割", 言語処理学会第7回年次大会発表論文集, pp. 429-432 (2001).