講演文を対象にしたトピックフレーズの抽出

1Q−3　伊藤 山彦† 谷田 泰郎‡ 松本 賢司† 柏岡 秀紀† 田中 英輝†

†(株)エイ・ティ・アール音声言語通信研究所 ‡(株)国際電気通信基礎技術研究所

1．はじめに

近年、大容量の回線が整備される中、インターネットを介した情報の配信も、テキストのみの形態から、映像や音声を扱う形態にまで広がりつつある。大量の情報が入手可能となる中で、自動要約技術に代表される情報を効果的に整理する技術は、ますます重要性を増している。

我々は、音声を用いた情報形態として、講演を対象にした自動要約の研究を進めている。従来の自動要約の研究は、利用者に依存せず、文書中で計算機が重要と判定した部分を抽出する手法が主流であった。しかし、重要性の基準は人間でも曖昧であり、機械による重要個所の判定は信頼性が低いという問題がある[1]。これに対し、本研究では、個々の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行うことを目的とする。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。利用者は、興味のあるトピックフレーズを指定して、それを更に詳しく記述した文書の範囲(パッセージ)を参照する。実現には、(1)トピックフレーズの抽出と、(2)トピックフレーズに対応したパッセージの抽出が必要である。本稿では(1)について述べる。また、本手法を NHK のニュース解説番組「あすを読む」に適用した結果について報告する。

2．従来のトピック抽出の研究と問題点

文書からのトピックの抽出に関連する研究として、文献[2][3]がある。文献[2]では、主題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。

上記研究では、トピックを単語または単語の集合としており、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。また、文献[3]では、文書内における単語の出現傾向の変化からテキスト分割を行う。しかし、我々が対象とした「あすを読む」は 10 分間の短い番組であり、番組中の話題は、番組全体のテーマから大きく外れることはない。そのため、話題ごとの単語の出現傾向の変化は少なく、正確なテキスト分割ができない[4]。

上記の問題に対し、本稿では、主として表層の手掛かりを利用して、利用者が理解可能な句の形でトピックを提示する手法を提案する。

3．トピックフレーズの抽出処理

図1に、本手法の概要を示す。以下、図に示した各処理について説明する。

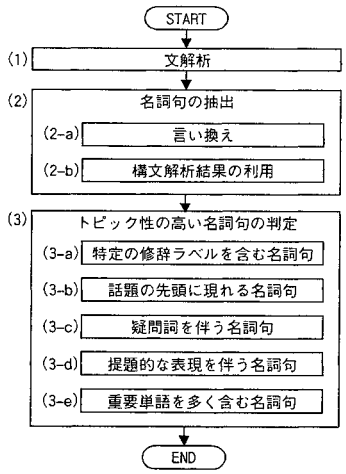

図1 トピックフレーズの抽出処理の概要

(1)文解析

形態素解析に Juman を、構文解析に KNP を用いた。

(2)名詞句の抽出

上記(1)の文解析結果を基に、以下の 2 つの処理によって文書から名詞句を抽出する。

(2-a)「なぜ〜のでしょうか」は「〜理由」のように、文書中で名詞句に言い換え可能な表現は、規則を用いて名詞句に変換する。実験では、24 個の言い換え規則を用いた。

(2-b)KNP が出力した構文木を基に、図 2 に示すように、木の中に現れる全ての体言を起点とし(図 2 では「内容」を例とする)、その直接の子孫となる形態素を連結して、名詞句を抽出する。

Extraction of Topic Phrases for Lecture Sentences

Takahiro ITO†, Yasuo TANIDA‡, Kenji MATSUMOTO†, Hideki KASHIOKA†, Hideki TANAKA†

†ATR Spoken Language Translation Research Laboratories.

‡ Advanced Telecommunications Research Institute International.

2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, JAPAN.

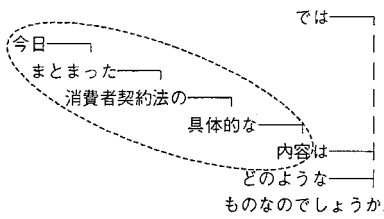

図 2 構文解析結果を利用した名詞句の抽出

(3) トピック性の高い名詞句の判定

以下の条件に従って、上記(2)で抽出された名詞句の中からトピック性の高い名詞句を判定する。

(3-a)「～の目的」「～の課題」「～の背景」など、特定の語(修辞ラベル)を含む名詞句。

(3-b)テキスト分割処理[1]によって分割された各話題の先頭に現れる名詞句。

(3-c)「なぜ」「どうして」のように疑問詞を伴う文に現れる名詞句。講演では疑問詞を伴う文は問題を提示する表現であり、トピック性が高い。

(3-d)「～という」「～でありますけれども」など提題的な表現を伴う名詞句。

(3-e)重要単語を多く含む名詞句。単語の重要度は tf*idf 値によって判定し、抽出された名詞句中に含まれる単語(内容語)の tf*idf 値の合計に従って順位付けを行う。

4．実験

上記手法に基づき、「あすを読む」の書き起こし原稿 25 件を対象に実験を行った。トピック性の判定は以下の優先順位に従った。

(1)(3-a)～(3-d)に該当する条件の数が多い名詞句

(2)(1)が同じときは条件(3-a)を含む名詞句

(3)(1)と(2)が同じ場合、条件(3-e)が上位の名詞句

上記に従い、1 文書について 10 個の名詞句を抽出した。図 3 は抽出結果の例である。正解の判定は、「出力された名詞句を詳しく記述したパッセージが文書中に存在するか否か」という基準により行った。図 3 の例では(3)(4)(5)(7)(8)が正解となる。(2)のように、文書全体のテーマとなる句は正解から外した。また、評価のため、文や句の重要性の判定に一般的に用いられる基準である tf*idf 値の合計が高い名詞句上位 10 個に対して同じ判定基準を適用し、本手法と比較した。結果を表 1 に示す。

番組の表題「死刑適用の基準」

(1)検察側が死刑を求めて上告をしておりました強盗殺人事件
(2)死刑適用の課題
(3)検察当局が死刑の適用を求めて上告をした五つの事件
(4)国立の事件の一審二審の判断の中身
(5)一審死刑にいたしました理由
(6)性犯罪を繰り返すといった反社会性が著しい点などでありましてこれが死刑を選択をした理由
(7)死刑と無期をわけたもの
(8)死刑適用について今日の判決が基準にいたしましたこと
(9)犯行の内容のうち計画性この態様のうちの中に入ると思いますけれどもこれについては計画的ではなく用意周到なものではなかったと認定をいたしまして死刑をためらった様相
(10)どのような場合に死刑を適用するのか大変重い課題

図 3 抽出したトピックフレーズの例

表 1 実験結果

	本手法	tf*idf
正解数(1 文書平均)	3.56	0.92

実験結果より、tf*idf による順位付けに比べ本手法は大幅に良好な値を示すことが確認され、本手法の有効性が示された。抽出された名詞句を分析したところ以下の問題が見られた。今後、改良のための検討が必要である。

(1)**名詞句の抽出の問題:**抽出結果には、図 3 の(9)のような不自然な名詞句が多く見られた。長さの制限やパターンの併用により、適切な名詞句を抽出するための工夫が必要である。

(2)**トピック性の判定の問題:**「～という」を伴う表現は、第 3 節(3-d)の条件では、提題的な表現と判定されるが、「～ということはない」のようにあてはまらない場合もある。トピック性の判定について、更に詳細な検討が必要である。

(3)**照応の問題:**「これらの出来事の背景」のように、代名詞の照応先が不明ため、それだけを見ても利用者にとって興味の判定の手掛かりとならない句が見られた。照応処理の導入が必要である。

5．おわりに

本稿では、講演文を対象にトピックフレーズを抽出する手法を提案し、実験を試みた。今後、実験結果からの考察を基に抽出精度を高めると共に、パッセージの抽出についても検討を行う予定である。

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,情処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,情処学会 自然言語処理研究会 NL-139, pp.61-68(2000).
[4]伊藤ほか:単語の共起知識を利用した講演文のテキストセグメンテーション,情処学会第 61 回全国大会(2), pp.159-160(2000).

[1]表層のパターンから以下を話題の境界と判定した。
(a)話題の転換となる接続詞(例:さて)、宣言的表現(例:～てみましょう。)、及び問題提示表現(例:なぜ～でしょうか。)を含む文から始まる位置。
(b)主題導入表現(例:～についてお伝えします。)、及び総括的説明表現(例:～わけです)を含む文で終わる位置。