4T-02 重要語の共起情報を用いた講演文の表題生成

松本 賢司　伊藤 山彦　柏岡 秀紀　浦谷 則好
(株)エイ・ティ・アール音声言語通信研究所

1. はじめに

インターネット上のWWW技術の進展によりテキスト情報の発信機会は、近年爆発的に増加している。こうしたテキストの主題を自動的に表題として生成することは、膨大なテキストの効率的処理の面からも有効である。

現在、講演文など語り口調の文の要約手法についての研究を行なっており、その研究の一環として本稿では、ニュース解説（ＮＨＫ「あすを読む」、50 件）の書き起し文を対象に表題の自動生成を試みた。

対象としたニュース解説には、すでに人手で付与された表題がある。使用されている単語に着目して、これらを以下の３つに区分した。

A	表題そのままの文字列が文中に存在	7 件
B	表題を構成するすべての自立語が文中に散在	34 件
C	文中に表れない自立語を表題として使用	9 件

今回の表題生成の試行では、人手付与の表題を評価の基準と考え、文中からの表題文字列の抽出(区分A)および文中語からの表題文字列の再現(区分B)を一応の目標とした。

2. 手法

従来から表題をテキスト中でもっとも重要な個所ととらえ文の重要個所抽出に利用する手法が知られている[1]。この考え方に基づき、最も重要と仮定した語を起点として、この基準となる語に先行または後続する適当な語句を再帰的に選択して、表題文字列を生成する。

論文の場合、表題は名詞句を構成することが多い[2]。本稿が生成の目標とした人手付与の表題についても、ほとんどが名詞句となっている（名詞句となっていない表題は全体で３件)。このことから今試行では表題として名詞句を生成する。

生成の手順は以下のようになる。

1) 起点語の決定

表題生成の起点となる語は、tf*idf 値の高位順の名詞を使用する。この語が最初の基準語になる。

2) 後続語句の取得

基準語に後続する語句のうち、「名詞」または「助詞＋名詞句」※1 が接続するものに注目する。基準語との関係において、

基準語 ＋{φ,の,や,と,...}＋{φ,連体修飾語}＋ 名詞(N)

となる N について文中での出現頻度が高い（頻度が同じ場合は文中の tf*idf 値を基準とした重要度による）N を後続語句の候補とする。基準語と N の接続関係のうち、高頻度（同一頻度の場合は基準語との距離）の語句を選択して基準語に接続する。

基準語	……	'制度'
最頻出の後続語句	……	'導入'
		'の_導入'(出現頻度 3)
		'導入'　(出現頻度 1)
文字列の接続	……	'制度の導入'
新たな基準語	……	'導入'

更に選択された N を新たな基準語として文中から後続語句を探索、取得する(最大４回繰り返す)。

3) 先行語句の取得

基準語に先行する語句のうち「名詞」または「連体修飾語」、「名詞+助詞」※1 が接続するものに注目する。基準語との関係において、

連体修飾語(A)＋ 基準語

名詞(N)＋{φ,の,や,と,...}＋ 基準語

となる A または N について文中での出現頻度が高

Automatic Construction of Lectures' Titles Using Collocational Information of Key Words

Kenji Matsumoto, Takahiro Ito, Hideki Kashioka, Noriyoshi Uratani

ATR Spoken Language Translation Research Laboratories

い（頻度が同じ場合は文中の重要度による）A またはNを先行語句の候補とする。候補語が名詞(N)の場合、基準語とNの接続関係のうち、高頻度（同一頻度の場合は基準語との距離）の語句を選択して基準語に接続する。

更に選択された語が名詞(N)ならば、これを新たな基準語として文中から先行語句を探索、取得する（最大3回繰り返す）。

4) 表題の生成

生成の起点となった基準語を中心に前後に生成された文字列を連結して表題を生成する。

※1 基準語と名詞(句)を接続する助詞には連体の「の」、並立の「と」「や」の助詞のほか、格助詞相当の句(という，といった，における，に対する，……)を対象とした。格助詞相当の句は前もって対象文中から該当語句を抽出した。

3. コーパスを利用した接続妥当性の検証

今試行では構文解析を行なわず、基準語に隣接する語句を頻度情報により接続するため、3単位以上の語句の接続については、不適当なものとなる場合がある。

基準語に接続する後続（先行）語句を選択する際に、基準語bを含む連続語句a･b と接続対象語句cについて、連続語句a･b･c が対象文中に出現していない場合は、新聞記事コーパス(1990-2000年、日本経済新聞)[3]を使用して接続の妥当性を検証する。コーパス中に連続語句 a･b･c が出現しない場合はa･bとcの接続は妥当でないと判断して、接続候補語句より除外する。

4. 考察

試行実験により生成した表題を人手付与の表題と比較したところ、以下の表に示すような結果となった。

<区分A、全7件>

人手付与の表題と(ほぼ)同様の文字列を生成	3件
人手付与の表題と一部が異なるがほぼ同様の文意	2件
文の主題を表していない、日本語として不正確	2件

<区分B、全34件>

人手付与の表題と(ほぼ)同様の文字列を生成	3件
人手付与の表題と一部が異なるがほぼ同様の文意	19件
文の主題を表していない、日本語として不正確	12件

<区分C、全9件>

人手付与の表題と(ほぼ)同様の文字列を生成	0件
人手付与の表題と一部が異なるがほぼ同様の文意	6件
文の主題を表していない、日本語として不正確	3件

実験結果の考察については現在、継続して行っている。今試行では文中に表れる語のみを連結して表題を生成するため区分Cのケースに関し表題の再現が不可能であるとしても、以下の点については、さらに検討が必要であると考える。

tf*idf 値の高位語から選んだ生成の起点となる重要語が人手付与の表題に含まれないケース(全体で7件)について、生成結果はすべて「文の主題を表していない」という判定になった。起点語の選択が適切でない場合、本手法では生成精度が低下する。

文中より名詞句を生成する際、本手法では

（連体修飾語句）＋（名詞）

の関係についての探索は行っているが、実際、人手による表題の付与の場合、

議論・が・迷走している → 迷走する・議論

のように動詞句の名詞句への言い換えによる表題生成がしばしば行われている。

5. おわりに

本稿ではニュース解説文を対象に表題の自動生成を試みた。今後、今回試行の手法を手がかりに情報発信の即時性が求められるニュース速報記事の見出しの自動生成に関しても、これに適合する生成手法を検討したい。

参考文献

[1] 吉見ほか：表題へのつながりに基づく文の重要度評価,自然言語処理,Vol.6,No.1,43-56(1999)

[2] 佐藤 理史：論文表題を言い換える,情報処理学会論文誌,40(7),2937-2945(1999)

[3] http://telecom21.nikkeidb.or.jp/