重要名詞の共起情報を利用した表題生成

松本 賢司[†]　伊藤 山彦[†]　谷田泰郎[‡]　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

あらまし

文書中の相対的に重要であると仮定した名詞を起点とし、これに先行、後続する語句を再帰的に接続して表題句を生成する手法を提案する。複数の起点から複数生成した表題候補句について、単語の重要度と文節のつながりの良さで順位付けを行ない最適な表題句を選択した。

1. はじめに

講演文など独話文を対象として表題を自動生成する手法についての研究を行なっている。

文書における表題は本文の内容を簡潔に表現している。その意味で表題の自動生成は文書の非常に簡潔な要約の生成と捉えることが出来る。従来、文書の要約手法としては何らかの選定基準に従い重要文を抽出する手法が広く用いられている。新聞記事・放送ニュース記事のようにリード文に文書の主題が重点的に記述されている特別な分野では、先頭の1文またはそれに続く数文を抽出し、その情報を元に表題生成する手法は有効である[1]。

しかし、一般に文書の主題を表す表現は文書中に散在しており、1文のみからの表題生成では文書の主題を表現出来ない可能性がある。また簡潔な表現からなる表題句を生成するには文より小さな単位、形態素あるいは文節を単位とした表題句の生成手法が有効だと考える。要約手法においても、文抽出型の要約に対して、要約スコアが最大になるよう単語を接合することで要約文を生成する手法や文節重要度と係り受け整合度に基づいて文を要約する手法など非抽出型とも言える要約手法が提案されている[2][3]。本稿の提案する表題生成手法も、形態素や文節を生成の単位としている。これらを原文書中の出現順に拘束されずに接合して簡潔な表題句を生成する。

すでに報告した表題生成手法では単一の生成起点から単一の表題候補句を生成するのみであったため、生成した表題の精度に問題があった[4]。本稿では、表題生成の起点を複数とした上で、ひとつの起点から複数の表題候補句を生成する。複数の生成句から最良句を選択することにより表題生成の精度の向上を図った。

提案手法について、NHK「あすを読む」[1]の書き起しテキストを対象に表題の生成実験を行なった。

2. 手法

本稿は表題として名詞句を生成する。名詞句は名詞を含む文節が連続する単純な形式とし、句の先頭にのみ形容詞、連体詞を認める。

例）ヨーロッパ．の．右翼．勢力．の．台頭
新しい．世紀．に対する．希望．と．不安

生成は文書中の重要名詞を起点に接続可能な文節を連続して接続することによって行なう。

生成の流れは以下のようになる(次頁図1参照)

(a) 対象文書から文節2-gramを抽出する。
(b) 対象文書のtf・idf値の上位の名詞(複数)を起点に(a)を探索して複数の表題候補の文節連続を生成する。
(c) 表題としての表現の妥当性を検証する。
(d) 表題候補句を順位付けして最良句を選ぶ。

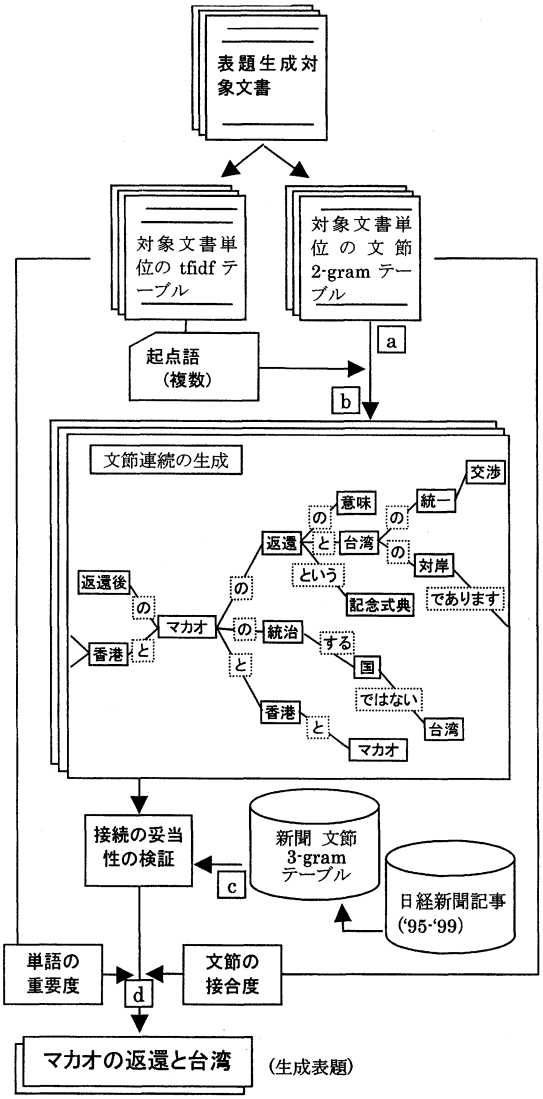
Automatic Construction of Titles Using Co-occurrence Information of Key Noun Words

Kenji Matsumoto, Takahiro Ito, Yasuo Tanida, Hideki Kashioka, Hideki Tanaka

ATR Spoken Language Translation Research Laboratories

[1] 「あすを読む」の書き起しテキスト使用に関してはNHKの許諾をいただいた。

図 1 表題生成の流れ



2.1. 文節 2-gram の抽出

重要名詞に接続する文節連続の要素となる文節2-gramを抽出ルールを適用して対象文書から抽出する(表3)。ここでいう文節2-gramは表1の形式をとる。1文節が含む自立語は1個で名詞が連続する複合語は異なる文節とした。

表 1　文節 2-gram

自立語部＝[接頭辞]*[自立語][接尾辞]*
第 1 文節＝[自立語部][付属語部]*
第 2 文節＝[自立語部]
文節 2-gram＝[第 1 文節] [第 2 文節]
* ： 0回以上繰り返し

抽出する文節 2-gram が名詞または形容詞・連体詞と名詞を接続する形式になるように以下の抽出条件を適用した。

- 第1文節の自立語は名詞または形容詞、連体詞のみ。
- 第2文節の自立語は名詞のみ。
- 第1文節の付属語部分を構成する最後尾の付属語については表2のいずれかに適合するもののみ。

表2　付属語部分の最後尾の付属語

品詞	その他の条件
助詞･連体化	
助詞･並立助詞	
助詞･接続助詞	出現形「および」のみ
助詞･格助詞･連語	出現形[う、る、た]で終わる う:という、とかいう、など る:に対する、に関する、など た:といった
助動詞	体言接続または基本形
動詞･非自立	基本形
動詞･接尾	基本形

- 名詞(多くはサ変接続名詞)の直後に続く場合の動詞は自立語部(名詞)間を接合する付属語部分と同様の扱いする。

例) 山積. する. 課題
　　運行. さ. れる. 便

※ 本稿での形態素の品詞区分は「茶筌」[5]の品詞体系に従った。

表3　文節2-gramの抽出例

.....行政による被災者支援の制度を見直すだけではなく選択するような工夫が必要です。
↓
. 行政. . による. . 被災. 者
. 被災. 者. . . 支援.
. 支援. . の. . 制度.
. 選択. . するような. . 工夫.

2.2. 表題候補の文節連続の生成

表題候補の文節連続を生成する際に処理の起点となる語を起点語とし、対象文書中の tf･idf 値の高位順の名詞を複数個、選択する。

文節2-gramテーブル中で第1文節の自立語が起点語と等しい文節2-gramを検索し取り出す。これらの第2文節自立語部と第1文節自立語部が等しい文節2-gramを文節2-gramテーブルから検索し、合致した文節2-gramを表題候補の文節連続として接続する。接続可能な文節2-gramがある限り処理を繰り返して起点語に後続する文節連続を生成する。起点語に先行する文節連続も同様にして生成する。

ただし起点語から文節連続の末端までのパスで同一文節2-gramを2回以上使用しない。これは生成処理が無限に連続するのを回避するためである。

起点語の後方に接続する文節連続 *w* を深さ優先で生成するアルゴリズムの概要を示す。*T* は文節2-gramテーブルとし *lastT* はテーブルの大きさとする。各テーブル内の *1st(bgm)* は第1文節自立語部、*fuzok(bgm)* は第1文節付属語部、*2nd(bgm)* は第2文節自立語部である。

```
1   s <- 起点語
2   w <- s
3   proc(w, s, T)
4    for j=1 to lastT {
5      if ( s = 1st(bgm[j]) ) {
6          w <- w + fuzok(bgm[j]) + 2nd(bgm[j])
7          T' <- T-bgm[j]
8          proc( w, 2nd(bgm[j]), T' )
9      }
10   }
11   write( w )
12   return
13  }
    7 :無限連続回避のため
```

起点語を中心に前後の文節連続を接続して表題候補となる文節連続を生成する(図 1-「文節連続」)。

2.3. 新聞 3-gram による検証

文節2-gramを再帰的に接続して得られる文節連続は対象文書中に出現するとは限らず、日本語として不適当な可能性もある。接続で得られた文節連続の妥当性を新聞記事(1995-1999年、日本経済新聞)コーパスから抽出した文節3-gramで検証する。文節3-gramは文節2-gram抽出条件に準じた基準により抽出し、得られた文節3-gram集合を新聞3-gramテーブルとした。

2.2.で作成した文節連続が新聞3-gramテーブルで被覆できる場合のみ接続が妥当であるとする。接続妥当性の検証は起点語から末梢方向に行ない、妥当性が検証された部分までを表題候補句とする。

文節間の接続妥当性の検証は、新聞3-gramテーブルを用いる他に、表題生成対象の「あすを読む」そのものに出現する文節3-gramを使うことも考えられる。しかし「あすを読む」には言いよどみ、言い直し表現がある他、語り口調の文特有の冗長な表現(「発表いたしました内容」など)も多く見られる。これらの表現を含む接続が、妥当とされてしまうことを避け、「あすを読む」を書き言葉的な観点でチェックするために新聞3-gramテーブルのみを用いて検証した。

2.4. 最適表題句の選択

2.3.で残った表題候補句を対象に表題としての良さを示すスコアを計算し最適な表題句を選定する。

計算は以下の通り、N 個の文字列からなる表題候補句 $W=w_1, w_2, \ldots, w_N$ について隣接する 2 文節 w_i, w_{i+1} が対象文書中に出現する回数を $FRQ(w_i, w_{i+1})$ とする。文節 w_i に含まれる自立語(名詞、表題候補句の先頭に限り形容詞, 連体詞も含む) v_i の tf･idf 値を $TFIDF(v_i)$ とする。表題候補句 W の表題として良さのスコアを以下のように計算する。

$$S(w) = \frac{1}{n-1}\sum_{i=1}^{N-1} FRQ(w_i, w_{i+1}) + \frac{1}{n}\lambda\sum_{i=1}^{N} TFIDF(v_i)$$

すなわち、文節のつながりの良さを第1項で、文節中の自立語の重要度を第2項で計算し、その合計値を

表題の良さの指標とした。本稿では、予備実験により係数 λ を0.1とした。また対象とした「あすを読む」の既存の番組表題がすべて3文節以上であることから、上記評価式により順位付けを行なう対象は3文節以上の表題候補句とした。

3. 実験と考察

「あすを読む」(50 件)の書き起しテキストを対象に、各文書の tf･idf 値の上位5名詞を起点語として、表題の生成実験を行なった。

表題句としての評価で上位1,2位となった生成句について以下の 3 段階の評価を行なった。

(1)適切な表題である。
(2)表題として許容できる。
(3)不正な表題である。

表4 生成表題に対する評価

	適切	許容	不適
1位	26%	34%	40%
2位	12%	36%	52%
1位 or 2位	30%		
前回実験	10%	56%	34%

評価式の適用により1,2位となった表題間で、「適切」の評価に関して、10%程度の差が見られる。前回実験[4]とは、「許容」の基準が異なるため、比較は難しいが、今回1位と判定された表題は「適切な表題」の割合が15%程度向上している。また順位2位までを加えると20%の向上が見られた。

表題の良さで1位とされた13件の生成表題句のうち、10 件については結果的に、対象文書中に出現する文字列を表題句にしており、残り3件は、対象文書中に出現しない文字列を生成している。

本手法を用いた表題生成とその評価については、今後さらに詳細に行なうことを予定しているが、表題生成実験の過程の検証により、以下の点についての検討が必要であると考える。

評価テキストと文節3-gramテーブル間の用語選択の差異

今回行なった表題生成実験では、評価用テキストとして「あすを読む」を使用し、接続妥当性検証に新聞記事より抽出した新聞 3-gram テーブルを用いた。このように、異なるテキストを利用する場合、両者間の用語選択の差異、表記法の差異による処理精度低下の問題が考えられる。

「エヌティーティー」(評価用テキスト)、「NTT」(3-gram テーブル)のような単純な表記の違いの他、「あすを読む」では「ヨーロッパ」(全体で 21 件出現)が使われ、「欧州」は使われていないのに対し、新聞記事では、「欧州」がより多く使われる傾向にある(2000 年 1 月以降で「ヨーロッパ」を含む記事数 782 件に対し、「欧州」8972 件)。このような場合を考慮した同義語展開処理を検討する必要がある。

4. おわりに

本稿では文節を単位とした名詞句表題生成の手法について、「あすを読む」の書き起しテキストを対象に実験を試みた。

今後、考察で言及した点についても検討しつつ、表題生成の対象を拡張する予定である。更に大規模な生成実験を行ない評価について報告したい。

参考文献

[1] 畑山ほか :日本語記事の重要情報に基づく英文ヘッドライン生成法,言語処理学会第 5 回年次大会発表論文集,17-20(1999)

[2] 堀 智織　古井 貞熙:話題語と言語モデルを用いた音声自動要約法の検討,音声言語情報処理,29-18(1999)

[3] 小黒ほか :文節重要度と係り受け整合度に基づく文要約アルゴリズム, 言語処理学会第 5 回年次大会発表論文集,133-136(2000)

[4] 松本ほか : 重要語の共起情報を用いた講演文の表題生成,情報処理学会第 61 回全国大会(2),161,(2000)

[5] 日本語形態素解析システム「茶筌(ChaSen) version 2.0 for Windows」,1999