文章の特徴を抽出するための一手法

6G−9

野崎　康夫　茂木　健　湯村　武　西田　行輝
三洋電機（株）　情報通信システム研究所

1.はじめに

文書の電子化・ネットワーク化とともに、大量の海外技術文献への高速アクセスが可能となった。これらをより迅速に入手・把握するため、必要な情報を自動的に選択・提示するシステムが求められている。われわれは現在、英文の製品マニュアルなどの要点を抜き出して高速翻訳する「抄録翻訳システム」の開発を目指しているが、その第1段階として重要文の抽出手法を構築した。

従来より、筆者の主張が明確に示される論説文などを対象とした自動文書要約の研究が行なわれている。[1] しかし実用的には、製品開発情報など、主に事実を記述した文書の抄録作成といった需要が多いと考えられる。

本稿では、これら事実を述べた文書を対象として、文章全体の特徴を表す重要語を高速に求め、次に必要な情報が記述されている重要文を抜き出す手法を提案する。

2.重要語の抽出

重要文の抽出の手がかりとして、まず文章全体や段落の記述内容を代表する重要語（テーマ語）を求める。簡易な語分割に基づいて原文中の単語の出現頻度をもとめ，これと書式情報、語彙特徴をもとに、テーマ語をいくつか選定する。処理の高速化のため，原文の形態素／構文解析を行わず，単語頻度表に現れた語に対してのみ，品詞などの語彙情報を求める処理を施している。

2.1 重要語リストの作成

語の出現頻度を求めた頻度表に，書式情報，語彙的特徴の属性を付加した重要語リストを作成する。

(1) 出現頻度表の作成

空白などを区切りとする簡易な語分割に基づき語の出現頻度統計をとり、頻度表の形の重要語リストを作成する。

(2) 書式情報[2]の付加

表題、段落タイトルに用いられている語に属性値を付加する。

(3) 特徴語抽出

固有名詞・専門用語・辞書にない語(新語)・省略語（例：GUI）は、その文書で特徴的に使用されている重要語とみなせる。重要語リストの見出し語に対して辞書引きを行ない、これら特徴語に属性値を付加する。

(4) 同内容語のまとめあげ

重要語リストに現れる見出し語は、形態素解析処理を施していないため、次のような同内容の語を別語として集計している。

① 文法的異表記語（大文字化、活用変化）
② 異表記語、表記ゆれ（U.S.A./U．S．A．/US)、
③ 同意語（America/United States)

そこで、重要語リストに現れた見出し語について辞書（品詞、類義語辞書）引きを行ない、同一内容を表す語の出現頻度を再集計する。また、

④ 省略語（GUI ⇔ Graphical User Interface)

についても、原形をあわせた再集計、原形出現位置情報の保存を行なう。

2.2 テーマ語の選定と重み付け

出現頻度と、各種属性をもとに、重要語に重み付けを行ない、重要度順にテーマ語（グローバルテーマ）を有限個選出する。また、段落情報をもとに、段落ごとのテーマ語（ローカルテーマ）も選定する。

各段落は、タイトルの特徴などをもとに、グローバル記述段落(introduction など)かローカル記述段落かの区別をし、ローカル記述段落では、グローバルテーマ語の重みは割り引くなど，重要度はローカル性を盛り込んだ値とする。

3. 事象パターン知識の利用

事実が記述された部分を抽出する手掛かりとなる典型的なパターンを、事象パターン知識として蓄え、重要文の抽出処理において，原文とのマッチングに利用する。

3.1 一般事象パターン知識

一般的に重要な事実の記述を示す、文法的なキーパターンを知識として用意しておく。このようなパターン知識の例として、つぎのようなものがある。

(1)真実事象をあらわすキーパターン
/not 〜, but 〜/,/actually,/
(2)重要性を示すキーパターン
/(most) important point is/,/be careful (not) to/
(3)記述内容紹介をあらわすキーパターン
/This document describes/,/This section describes/

A method to extract the feature of sentences.
Yasuo NOZAKI, Takeshi MOGI, Takeshi YUMURA, Yukiteru NISHIDA,
Information & Communication System Research Center, SANYO Electric Co.,Ltd.

3.2 要求別事象パターン知識

事実を記述した文書では，文の重要度を絞り込む決め手が少ない．また、重要とされる情報も、利用者や分野によって異なる．そこで、要求に応じたキーパターンを要求別事象パターン知識として用意し、利用者の好みに応じた抄録を作成するために供する．

この知識は、製品開発情報、コマンド使用法など要求に応じて任意に作成、利用できる．また、おのおのは、その下に階層化されたいくつかの要素パターン知識（時期情報、社名情報、引用情報、使用法など）を組み合わせてインクルードすることにより構成する．

個々のパターンは、マッチング条件と、抽出条件が記述できる．（表１）

要求別事象パターン
・製品開発情報:
　/〜 developed/,
　/#include 〈時期〉,〈会社名〉,〈価格〉/,
要素パターン
・使用法：
　/^To use 〜/,
　文頭にマッチ、対応文全体抽出
　/In order to use 〜,/,
　条件部分のみを抽出
・時期:
　/Jan/,/Feb/,
　/19[0-9][0-9]/,...
・不要パターン：
　/!for example〜/ 不要文指定

表１　事象パターン知識の記述例

4.重要文の抽出

4.1 重要文選択

テーマ語のリストと、一般事象パターン知識、要求別事象パターン知識をもとに、要求に応じた重要文の抽出を行なう．

(1) 無条件抽出文の決定
・タイトル行を選択マーク．
・要求別事象パターンにマッチした文を選択マーク．
(2)不要文の決定
　不要パターンにマッチした文へ抑制マーク．
(3) 条件付き抽出文の決定
　次の条件を満たす文にポイントを与える．
　・段落の先頭文
　・段落ごとのローカルテーマを含む文
　・一般事象パターンにマッチした文
　・グローバルテーマを含む文
　・省略語の原形出現文

所望の圧縮率を越えるまで、ポイントが大きい文から順に抑制マークのついてない文を選択マークする．

5.重要文抽出例

図１に，メールリファレンスマニュアルのメッセージフォルダに関する記述段落での重要文抽出例を示す．（選択マークされた文をハイライト表示、選択のためのキーとなったパターンを反転表示している）この例では，グローバルテーマのうち"mail"がローカルテーマには選ばれず，"folder"などのローカルテーマや，"To use 〜"などの事象パターンをもとに重要文が選ばれている．

3. Maintaining folders

Mail includes a simple facility for maintaining groups of messages together in folders. This section describes this facility.

To use the folder facility you must tell Mail where you wish to keep your folders. Each folder of messages will be a single file. For convenience, all of your folders are kept in a single directory of your choosing. To tell Mail where your folder directory is, put a line of the form
　set folder=letters
in your .mailrc file. If, as in the example above, your folder directory does not begin with a `/,' Mail will assume that your folder directory is to be found starting from your home directory. Thus, if your home directory is /usr/person the above example told Mail to find your folder directory in /usr/person/letters.

図１　重要文抽出例

6.おわりに

簡易に抽出したテーマ語と，要求に応じた事象パターン知識を用いることにより，事実を記述した文章から必要な情報を含む文を抽出する手法を示した．本手法では原文の形態素解析などを用いずに高速処理を実現した．出現頻度の高い「重要語」を含む文の割合は意外に多く，重要文の絞り込みは、パターン知識の記述方法による影響が大きい．さらに，より実質的な重要文を求めるためには、指示語の指示対象の認識，文間の連接関係の把握が必要と考えられる．

今後は，多分野の文章を用いて評価実験を行うとともに、文の連接関係などをも利用した高精度な重要文抽出技術を確立し、抄録翻訳システムなどへの応用を目指す考えである．

【 参考文献 】

[1] 石橋他：英文要約システム「DIET」，情報処理学会第３８回全国大会(1989)
[2] 茂木他：書式情報を利用した英日機械翻訳処理，情報処理学会第４３回全国大会(1991)