新聞記事の要約のためのテンプレートの自動抽出

吉田 和広　徳永 健伸　田中 穂積
東京工業大学 大学院情報理工学研究科
{yoshida,take,tanaka}@cs.titech.ac.jp

1 はじめに

近年，利用可能な電子化されたテキストの増加に伴い，それらから必要な情報を効率良く入手する技術が望まれている．このような技術として従来から抄録や要約に関する研究 [1, 2, 3] が行なわれているが，これらの多くは，主に一つの記事（文書）の内容を要約することに重点を置いているため，新聞やニュースのような一つのイベントに対して複数の記事が存在するテキストに適用した場合，複数の記事によって表される情報（時間的なイベントの変化など）をうまく要約することができないという問題がある．そこで，ある話題の記事に対して重要と考えられる項目をテンプレートとし，それを用いて抽出した各記事のデータから要約を作り出すことが考えられる．このような考え方をもとにした要約処理の研究としては [4] などがあげられるが，現段階では，テンプレートの作成は話題ごとに人手によって行なわれている．

そこで，本研究では新聞記事を対象とし，要約のためのテンプレートを，与えられた話題に関する記事集合から自動的に抽出することを目的とする．本研究で提案するテンプレートの自動抽出手法では，「話題で重要な表現＝話題に固有な表現」という考えのもとに，対象の話題の記事の出現頻度とランダムに集めたサンプル記事の出現頻度をもとに計算した**重要度**という値を用いてテンプレートを抽出する．

2 システム概要

ここでは本手法により抽出するテンプレートの形式と，処理手順について説明する．

2.1 テンプレートの形式

本研究では，テンプレートを“動詞＋格要素”の意味的に類似した集合という形で表現する（図1）．格要素にはその格に望ましいと考えられる意味カテゴリ名を記述する（図中，中括弧で記した部分）．

{固有地名}	で	{数値}	を記録する
{固有地名}	で	{数値}	を観測する

図 1: テンプレートの例（地震の話題）

2.2 処理手順

本手法では以下の手順でテンプレートを抽出する．

1. 重要な動詞の抽出
   話題に固有な動詞を重要度をもとに抽出する．この際，類似した動詞のグループ情報をもとに重要度の再計算を行なうことにより，表現上の違いによる重要度の差を少なくする．
2. 重要な動詞組（動詞＋格助詞）の抽出
   抽出した重要な動詞に対して，その動詞の格助詞を伴った動詞組を収集し，重要な動詞の抽出と同じ手順で話題に固有な動詞組を抽出する．
3. テンプレート形式への変換
   抽出した動詞組の格要素を抽象化し，意味的に類似したグループに分けテンプレートとする．

3 重要な動詞の抽出

(i) 動詞の収集

動詞の収集では，形態素解析器 JUMAN [5] を利用して形態素解析を行ない，その結果から動詞を抽出する．

(ii) 重要度の計算

記事から収集した動詞に対して話題における固有性を考慮した重要度を求める．

重要度の計算には，各動詞の出現頻度を用いる．この場合，対象の話題の記事の出現頻度のみを利用したのでは，どの話題にも出現する動詞も，話題に固有な動詞も，出現頻度が同じ場合に，重要度に差がでないという問題がある．

そこで，本研究では，一般の記事からランダムに集めたサンプル記事の出現頻度を用いることによりこのような問題を解消する．具体的には，記事 A_i 中の動詞 $v_k(k=1,2,\ldots,Nv_i)$ の重み w_{ik} を次の式で表し，

$$w_{ik} = \frac{f_{ik}/(n_k+1)}{\sqrt{\sum_{j=1}^{Nv_i}(f_{ij}/(n_j+1))^2}}$$

f_{ik} ： 単語 T_k の記事 A_i での出現頻度
n_k ： 単語 T_k がサンプル記事で出現する記事数

ある話題の記事 $(A_1,\ldots,A_{Nd})$ の単語 T_k の重要度 i_k をこの重みの総和で表す．

$$i_k = \sum_{j=1}^{Nd} w_{jk}$$

重み w_{ik} の計算では，分子の部分で各記事での出現頻度をサンプル記事で出現する記事数で割っており，これにより，サンプル記事で多く出現するものには小さな重みが，サンプル記事でほとんど出現しないものには大きな重みが与えられ，その総和である重要度もそれに応じて変化する．

地震の話題に対してこの方法により計算した重要度の例を以下に示す．

動詞	重要度
推定する	20.68
観測する	13.52
発生する	7.97
ある	1.57

(iii) 分類語彙表による動詞のグループ化

前述の重要度計算では頻度情報を利用しているため，高い重要度をもつ動詞と同じような意味を表す動詞であっても，出現回数が少ないものには低い重要度が与えられ，重要な動詞として判定されないことがある．そこで，同じような意味をもつ動詞をグループ化し，その情報をもとに重要度の再計算を行なう．

グループ化の方法として，ここでは分類語彙表を用いる．分類語彙表 [6] では，7桁の分類コードが振られており，一致する桁数が多いほど意味的に類似しているような尺度づけになっている．そこでこの情報を用いて以下の手順で動詞をグループ化する．

1. 出現記事中の全ての動詞を分類語彙表の分類コードに変換する．サ変動詞は名詞部分を分類コードに変換しそれを利用する．
2. これらの動詞を6桁の分類コードの一致でグループ化する．

(iv) グループ情報を用いた重要度の再計算

グループ情報を用いた重要度の再計算方法として，同一グループに属する動詞の重要度を足し合わせ，その値をグループの重要度とする方法を提案する．例えば，「起きる, 起こす, 起こる, 引き起こす」がグループを形成しており，それぞれ図2の左のような重要度が割り当てられている場合，再計算後は，「起きる, 起こす, 起こる, 引き起こす」を1つのグループとして，それに対して重要度の総和 7.34 を割り当てる．

起きる	3.64	→	起きる	7.34
起こる	2.21		起こる	
引き起こす	1.07		引き起こす	
起こす	0.42		起こす	

図 2: グループ情報を用いた再計算の例

この方法により，各々が小さな重要度をもつものであっても，動詞の意味的に類似したグループ全体で重要度が高ければ，重要と評価されるようになる．

(v) 重要度リストからの動詞の抽出

以上の手順により計算された重要度の振られた動詞のリストから，上位のものを重要な動詞として抽出する．

地震の話題の場合の抽出例を以下に示す．

推定する, 推測する, 推す, にらむ, 観測する,
測定する, 測量する, 発生する, 出す, 示す, ···

4 重要な動詞組の抽出

重要な動詞に対して，その格助詞を伴った動詞組（動詞＋格助詞）を収集し，さらにその中で話題に重要な動詞組を推定，抽出する．基本的には先ほどの重要な動詞の抽出と同じ手法を用いる．

(i) 動詞＋格助詞の抽出

重要な動詞に対して，その動詞の格助詞を抽出する．

（例） [記録する, で, を]
[崩れ落ちる, が]

(ii) 重要度の計算

前述の重要度の計算方法により，動詞組に対して重要度を計算する．

（例）

動詞組	重要度
[記録する, で, を]	17.05
[推定する, と]	12.07
[襲う, を]	8.05
[発生する, に]	7.89

(iii) 類似度計算による動詞組のグループ化

動詞組のグループ化では，重要な動詞の抽出のときと異なり，動詞組間の類似度をもとにグループ化を行なう．この類似度は，ベクトルモデルにおける文書の類似度 [7] を応用し，動詞組の取り得る格要素のベクトルの類似性として計算する．

1. 動詞組の格要素を分類語彙表の分類コード（5桁）に変換する．
   八戸で 六を 観測する → 99999 で 11950 を 観測する
2. 動詞組 V_iごとに格要素（格助詞＋分類コード）C_k の出現頻度 f_{ik}をもとめる．
   （例） [観測する, で, を]　を:11950　8
   で:99999　7
   を:31120　3
3. 動詞組 V_iごとに格要素 C_kの重み w_{ik} を計算し，これを V_iのベクトル $(w_{i1}, w_{i2}, \ldots, w_{it})$ とする．
   $$w_{ik} = f_{ik} \times \log(N_d/n_k)$$
   (N_d : 総動詞組数, n_k : C_kが出現する動詞組数)
4. 動詞組 V_iと V_jの類似度を，ベクトル間の角度を θとしたときの $\cos\theta$ とする
5. 動詞組 V_iに対してある閾値以上の類似度をもつ動詞組を同一グループとする

地震の話題の場合のグループ化の例を以下に示す．

[記録する, で, を], [観測する, で, を]
[発生する, が, に], [発生する, が], [起きる, が], [起こる, が], [多発する, が]

(iv) グループ情報を用いた重要度の再計算

類似度計算によりもとめたグループ情報を用いて，重要な動詞の抽出で提案した重要度の再計算を行なう．

(v) 重要度リストからの動詞組の抽出

重要度順に並べた動詞組のリストから，上位のものを重要な動詞組として抽出する．

地震の話題の例を以下に示す．

[記録する, で, を], [観測する, で, を],
[発生する, が], [起きる, が, に], [発生する, が, に],
[起きる, が], [多発する, が], [起こる, が], ···

5 テンプレート形式への変換

重要な動詞組に対して，格要素の抽象化とグループ化を行なう．

(i) 格要素の抽象化

動詞組の格要素を分類コード（5桁）に変換し，格ごとに出現頻度の高いものからいくつかをその格要素の望ましい格要素とする．

(ii) 動詞組のグループ化

類似度計算による動詞組のグループ化を行ない，グループ内で，格要素の共通のものがあった場合，それらをマージする．

（例）「11950 と 15260 で 推定する」「11950 と 推定する」
↓ マージ
「11950 と 15260 で 推定する」

(iii) 項目名への変換

分類コードを分類語彙表の項目名に変換し，テンプレートとする．テンプレートの作成例を図 3に示す．中括弧は意味カテゴリを表し，上位のものほどその格に望ましいことを表す．***は分類語彙表で未定義である固有地名などを表す．

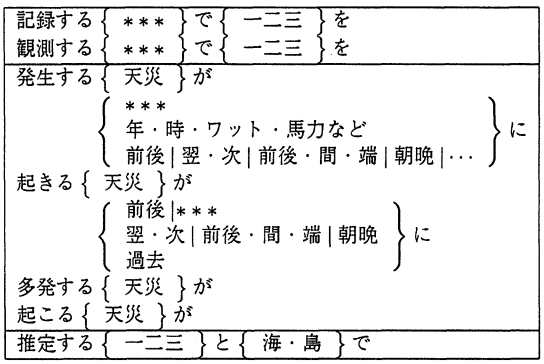

図 3: テンプレートの例（地震の話題）

6 実験

本手法を用いてテンプレートを作成し，簡単な適用実験を行うことにより，本手法により作成されたテンプレートの特徴の分析を行なう．以下その方法と結果について説明する．

6.1 実験方法

実験では，CD-ROM 版日本経済新聞 94 年度版 [8] から，「地震」，「交通事故」，「火災」，「記者会見」の記事を，キーワードを用いた検索により収集し，各話題の記事として利用する．また，本手法の重要度計算に用いるサンプル記事には，ランダムに抽出した 1000 記事を用いる．実験には，上記の記事のうち本文のあるもののみを利用する．

実験として，各話題に対して本手法を適用し作成したテンプレートを用いて，データ抽出を行なう．データの抽出方法は，各記事に対してテンプレートに記された格要素の分類コードと一致した動詞組のみを抽出する．結果を表 1に示す．

表 1: テンプレートの適用結果

	交通事故	地震	火災	記者会見
記事数	419	379	478	876
抽出したテンプレート数	30	202	124	140
総動詞組数 (Z)	5159	7116	8273	17691
テンプレートの適用できた動詞組数 (Z')	925	1368	1347	3518
テンプレートの適用できた動詞組の比率 (Z'/Z)	0.179	0.192	0.162	0.198

6.2 結果と考察

テンプレートの抽出能力

表 1よりテンプレートの適用できた動詞組の比率 (Z'/Z) が，ほぼ一定 (0.162〜0.198) であることから，本手法により作成されたテンプレートは，同定度の動詞組の抽出能力を持つことがわかる．

また，各話題のテンプレートの動詞組の抽出能力が同程度であることから，テンプレート数の少ない話題ほど，テンプレート 1 つあたりの抽出能力が高いと考えられる（例 交通事故）．

テンプレートの適用傾向

今回の実験において，テンプレートが全く適用できない記事がいくつかあった．テンプレートは話題で重要な情報を表す表現であるため，これらの記事は話題で重要な情報を持たない記事と考えられる．そこで，テンプレートの適用できなかった記事に対して，見出しや，本文を見て，一般的にその話題で重要かどうかの判定を行なった．結果を表 2に示す．

表 2: テンプレートの適用傾向の分析

	交通事故	地震	火災	記者会見
全体の記事数	419	379	478	876
適用できない記事数	108	23	77	65
(重要でないもの)	99	23	74	27
(重要でないものの比率)	0.91	1.00	0.96	0.41

- フィルタとしての可能性
  表 2よりテンプレートを適用できない記事に対する重要でない記事の比率が，平均 0.82 であることから，本手法により作成されたテンプレートは，話題で重要でない記事を取り除くフィルタとして利用可能であると考えられる．
- 事故など状況の変化を含む話題に有効
  重要でない記事の比率や，実際に内容を調査した結果から，本手法は，事故の状況など記事の内容がある程度パターン化して表せるような，報道的な内容の場合に有効と考えられる．

7 おわりに

本研究では，話題における固有性を表す重要度の計算法を提案し，それを用いた新聞記事の要約のためのテンプレートの自動抽出法を提案した．また，4 つの話題に対して本手法を適用し，作成したテンプレートを用いて簡単な適用実験を行なった．実験により，本手法により作成されたテンプレートは，重要でない記事を除くフィルタ的な利用が可能であることが確認された．

参考文献

[1] 田村俊哉, 田村直良．文章の表現形式に基づいた要約文章の生成について．情報処理学会 自然言語処理研究会, Vol. NL92-11, 1992.

[2] 原正巳, 木谷強, 江里口善生．特徴的表現を利用した特許抄録作成法の検討．情報処理学会 自然言語処理研究会, Vol. NL100-14, 1994.

[3] 山本和英, 増山繁, 内藤昭三．文章内構造を複合的に利用した論説文要約システム GREEN．言語処理学会, Vol. 2, No. 1, pp. 39–52, 1995.

[4] Kathleen McKeown and Dragomir R.Radev. Generating summaries of multiple news articles. In *SIGIR'95*, pp. 74–82, 1995.

[5] 松本裕治ほか．日本語形態素解析システム JUMAN 使用説明書 version 2.0, 1994.

[6] 国立国語研究所（編）．分類語彙表．秀英出版, 1964.

[7] Gerard Salton, James Allan, Chris Buckley, and Amit Singhal. Automatic analysis, theme generation, and summarization of machine-readable texts. *SCIENCE*, Vol. 264, pp. 1421–1426, 1994.

[8] 日本経済新聞社（編）．日本経済新聞 CD-ROM 版 1994 年版．日本経済新聞社, 1995.