E-033

類似文書群特徴に基づく重要文抽出方式の実験評価
Experimental Evaluation of A Sentence Extraction based on Similar Documents

川越 智吏† 絹川 博之†
Satoshi Kawagoe† Hiroshi Kinukawa†

1. はじめに

テキスト自動要約技術とは，膨大な情報の概要を短時間に把握する一つの方法として注目されている技術である．

従来の要約の手法は，統計的な単語重み付け法と，文書中の文間関係を解析したテキスト構造情報等の，文書に付随する表層情報を組み合わせ，重要と判断される文を抽出する，重要文抽出型の手法である．一般に，これら表層情報は，文書の性質に大きく依存しており，また，そのシステムの構築と維持，管理に多大な労力を必要とする．

この問題点を解決するために，本研究では，統計的な単語重み付け法を拡張し，要約対象文書中において，類似文書群を特徴付ける単語と，類似文書群との違いを特徴付ける単語の重みをその他の単語より大きくする手法を提案する．

この手法の特徴は，要約対象文の単語重み付けに，それに類似した複数の類似文書を集めた，類似文書群を用いるということである．本研究ではこの手法を検証し，この手法の効果的な利用法を考察する．

2. 類似文書群特徴に基づく重要文抽出方式

2. 1 類似文書群

ある文書dに含まれる単語 $\{w_1, w_2, \cdots, w_n\}$ の重みを語の出現頻度に基づき算出する場合，従来では，ある文書集合全体Dにおける $w_i\ (1 \leqq i \leqq n)$ の出現頻度と，ある文書dにおける w_i の出現頻度をもとに重みが算出される．この方法で算出された重みは，文書dの主題内容が的確に反映されていないものとなってしまう．

文書dの主題内容を的確に反映させるために，要約対象文書dと類似した文書の集合 D_S に着目し，dと D_S との共通点，相違点が見られる単語により大きな重みを与える単語重み付け方式が提案された [1]. D_S を導入することにより，

D： 全文書集合
d： 要約対象文書
D_S： dの類似文書群

という3つの集合（$d \subset D_S \subset D$）が定義される．

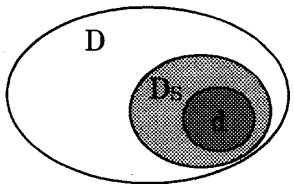

図1 集合関係

2. 2 類似文書群特徴重み

要約対象文書dの類似文書群 D_S の特徴を利用した単語重み付け方式を，以下に示す [1].

Pをある集合とし，Sをその部分集合としたとき，ある単語重み付け法 Method による，SのPにおける特徴を表す単語wの重みを，$W_{Method}(w|S|P)$ と表すものとする．

(1) 類似文書群共通特徴重み

要約対象文書dにおいて，類似文書群 D_S を共通的に特徴付ける単語wの重み．

$$W^{C}{}_{Method}(w) = W_{Method}(w \mid D_S \mid D)$$

(2) 文書個別特徴重み

要約対象文書dと類似文書群 D_S との違い，すなわち要約対象文書d独自の話題や情報を特徴付ける単語wの重み．

$$W^{D}{}_{Method}(w) = W_{Method}(w \mid d \mid D_S)$$

(3) 合成特徴重み

類似文書群共通特徴重みと文書個別特徴重みとの双方を併せ持つ重み．

$$W^{C+D}{}_{Method}(w) = W^{C}{}_{Method}(w) + W^{D}{}_{Method}(w)$$

(4) 類似文集合を用いない従来の重み

類似文書群を用いずに，要約対象文書dにおいて，全文書集合Dとの違いを特徴付ける単語wの重み．

$$W^{U}{}_{Method}(w) = W_{Method}(w \mid d \mid D)$$

3. 単語重み算出法

本研究で用いた単語重み算出法は以下のものである．

(1) tf-idf法

より少ない文書に偏って出現する単語が多く出現するときに大きな重みを与える方式．

$$W_{tf-idf}(w|S|P) = tf(w|S) \times \log\left(\frac{N(P)}{N(P|w)}\right)$$

(2) tf/TF法

wの，Sにおける出現確率と，Pにおける出現確率の比を求める方法．

$$W_{tf/TF}(w|S|P) = \frac{tf(w|S)}{tf(w|P)}$$

(3) SMART法

tf-idf法に文書長による正規化を施し，精緻化した方式．

$$W_{SMART}(w|S|P) = \left\{\sum_{t \in S} \frac{\log(tf(w|t))+1}{Ave_{u \in t}\{\log(tf(u|t))+1\}}\right\} \times \log\left(\frac{N(P)}{N(w|P)}\right)$$

†東京電機大学大学院工学研究科

(4) HGS 法

超幾何分布を応用した確率計算に基づく方式であり，高頻度語や低頻度語に偏らない公正な重み付けが高速に行えるとされる方式.

$$W_{HGS}(w \mid S \mid P) = -\log\left(\sum_{l \geq k} hg(N,K,n,l)\right)$$

$$hg(N,K,n,l) = \frac{C(k,l)C(N-K,n-l)}{C(N,n)} = \frac{n!K!(N-K)!(N-n)!}{N!l!(n-l)!(K-l)!(N-K-n+l)!}$$

$$(\min\{0, N+K-n\} \leq l \leq \max\{n,K\})$$

$N = P$の単語数　　$K = w$のP中での頻度

$n = S$の単語数　　$k = w$のS中での頻度

4. 重要文抽出型単文書要約への適用

提案方式を適用し，以下の処理手順で新聞記事の重要文抽出型単文書要約システムを試作した.

(1) 重み付け対象単語の選出

d の見出しと本文を JUMAN Version 3.61 を用い，内容語として，名詞，動詞，形容詞，未定義語を選出した.

(2) Ds 作成

Ds の作成には，見出しの類似度による上位①300 件，②500 件，③800 件，④1000 件，⑤2000 件，⑥3000 件の選出を行った後，本文の類似度による上位の選出を行い，①は上位から 50 件ずつ増やし 150 件まで計 3 個の Ds を，②は上位から 50 件ずつ増やし 250 件まで計 5 個の Ds を，③は上位から 100 件ずつ増やし 400 件まで計 4 個の Ds を，④は上位から 100 件ずつ増やし 500 件まで計 5 個の Ds を，⑤は上位から 100 件ずつ増やし 1000 件まで計 10 個の Ds を，⑥は上位から 100 件ずつ増やし 1500 件まで計 15 個の Ds を作成した [2].

(3) 単語重み付け

重み付け対象用語 w に対し 3 章で述べた 4 方式を適用し，2.2 章で述べた重みを算出した.

(4) 文重みの計算

ある文 s に含まれる重み付け対象語$\{w_0, w_1, \ldots, w_n\}$の重みの総和を文重みとした.

(5) 重要文抽出

文重みの上位から，指定された要約率を超えるまで抽出を行い，出現順に接続して要約文として出力する.

5. 評価と考察

試作システムを，NTCIR Workshop2 TSC1 の重要文抽出型要約タスク（Task A1）の Formal run データにより評価した．その結果，HGS 法－W^C を組み合わせた要約がもっとも高い要約精度を示した．特に，要約率 50% における要約精度は，テストコレクション参加システムの最良値 61.2% を越える 63.5% となり，本提案手法がその他の手法に比べ同程度以上の要約精度を得られることを確認した．各 Ds における HGS 法と W^C の組み合わせによる，要約率 50%における要約精度をグラフに示す(図 2).

また，要約率 10%では，試作システムの要約精度が，テストコレクション参加システムの最良値に比べやや劣る結果となり，要約率 30%では，試作システムの要約制度が劣る結果となった.

この結果から，試作システムにおける要約は要約率が元のテキストの 50%程度のテキストを作成する際に有効な手法であると考えられる.

6. おわりに

本研究では，類似文書群の情報を用いる単語重み付け方式において，HGS 法－W^C の組み合わせの有効性を確認し，有効的な要約率について検証した．今後の課題としては，類似文書群の作成手法を，現在の件数を指定する手法から，一定の類似度以上のものを取り出す手法に変更することを考えている．また各単語重み算出法と，類似文書群特徴重みの関係における理論的な考察が必要である.

参考文献

[1] 木村 誠，絹川 博之，""新聞記事のテーマ指向性要約における各種単語重み付け方式の定量的評価，"情報科学技術フォーラム一般講演論文集，Vol. 2，No. E-4，2002.

[2] 渡辺修司，"類似文書群特徴に基づく重要文抽出方式の提案とその実験評価，"東京電機大学大学院，修士論文，2004.

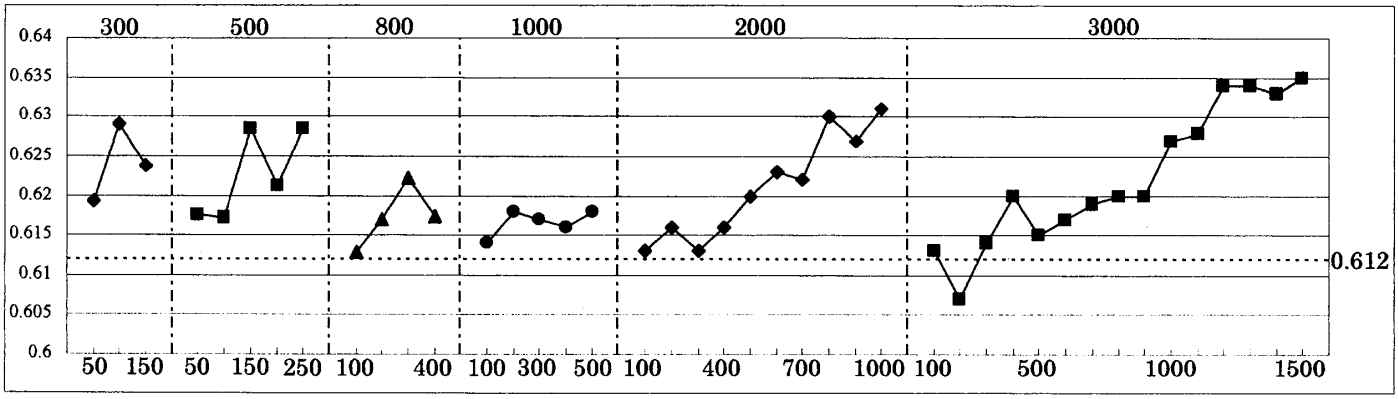

図 2　HGS 法－類似文書群共通特長重みを利用した要約率 50%における要約精度

※縦軸の値は要約精度．横軸の値はグラフの上段にあるものが，見出し類似度による選出数，グラフ下段にあるものが，本文類似度による選出数である．また NTCIR-2 参加システムの最良結果は 0.612 である.