1V-2

組み合わせ確立モデルを用いた文書の自動要約の評価

藤沼　卓也† 岸　義樹‡

†茨城大学大学院理工学研究科情報工学専攻　‡茨城大学工学部情報工学科

1　はじめに

近年、コンピュータの急速な発展と普及に伴い電子化された文書の数が膨大なものになった. それゆえに、人間が膨大なテキストデータの中から重要な文書を見つけ出す際に多くの時間を消費することになり、性能の良い自動要約システムへの必要性が高まっている.

このような状況を踏まえ、本研究では、大量のテキストデータを扱うのを容易にかつ比較的短時間で行えるように文章の自動要約システムの開発を行った. 要約の方法として、各単語に重み付けを行うことによる文書中から文書の主な単語を表す特徴単語抽出方式と、単語の文書の位置を考慮する Lead 法を組み合わせて [1] 適用し、再現率や適合率、F 値及び人間によるアンケートを用いて評価した.

2　要約手法

2.1　特徴単語抽出方式

2.1.1　単語重み付け法

本研究では、2 種類の単語重み付け法を用いて、単語抽出を行った.

2.1.2　tf-idf 法

より少ない文書に偏って出現する単語に高スコアを与える方法. 式 (1) で定義される.

$$tf/idf(v|D(\omega)) = tf(v|D(\omega)) \times log(N_{all}/N(v)) \quad (1)$$

2.1.3　HGS 法

超幾何学分布を応用した確率計算に基づく方式 [2] 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。式 (2)〜(4) で定義される.

$$W(N,K,n,k) = -log(hgs(N,Kn,k)) \quad (2)$$

$$hgs(N,K,n,k) = \sum_{l \geq k} hg(N,K,n,l) \quad (3)$$

$$hg(N,K,n,l) = \frac{C(K,l)C(N-K,n-l)}{C(N,n)} = \frac{N!K!(N-K)!(N-n)!}{N!(n-l)!(K-l)!(N=K-n+l)!} \quad (4)$$

A proposal of method for extracting a image of word from a great deal of sentence

Takuya Fujinuma†, Yoshiki Kishi‡

†‡Ibaraki University

4-12-1 Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan

2.1.4　HGS 法の改良

本研究では、上記の HGS 法を利用するが、この方法は複数の文書中から特徴単語を抽出する方法のため単一の文書において単語を抽出するには HGS 法を改良する必要がある. そのため、本研究で HGS 法を利用するために行った改良を以下で説明する.

HGS 法で定義されている N は全文書中 D_0 の単語数を示す変数であり、n は一文書中に含まれる単語数を示す変数である. この 2 つの変数について改良した方法では、N は変わらず文書中に含まれる単語数を示す変数とし、n を各段落が持つ単語数と見なす. これは HGS 法における文書集合数 D_0 を、改良法では段落の数と見立てるということである. これにより、式の意味を破壊することなく特徴単語を抽出するための式を利用することが可能となる.

2.2　Lead 法

本研究では Lead 法としての処理を以下のように行った。

1. 第一パラグラフに含まれる文では単語重み付け法で算出されたスコアを 2 倍にする.
2. 各パラグラフに内の第一文では単語重み付け法で算出されたスコアを 2 倍にする.

3　要約文の評価

各単語重み付け法による要約文を評価するために、システムを構築し要約文を作成して再現率、適合率、F 値による評価行い. さらに、アンケートを実施した.

3.1　要約文の作成

要約文書として「平成 13 年度　情報通信白書 [3] 」のデータを利用し文書群ではなく単一文書の要約を行った. tf-idf 法と HGS 法の 2 つの特徴単語重み付け法を用いて 5〜35%の要約率で 5 つの要約文を作成し全部で 10 の要約文を作成した.

要約率	再現率	適合率	F 値
8%	0.978	0.957	96.785
12%	0.931	0.911	92.142
20%	0.873	0.855	86.428
25%	0.837	0.819	82.857
35%	0.711	0.696	70.357

表 1: HGS 法

要約率	再現率	適合率	F 値
8%	0.978	0.957	96.785
12%	0.967	0.946	95.714
20%	0.844	0.826	83.571
25%	0.819	0.802	81.071
35%	0.700	0.685	69.285

表 2: tf-idf 法

3.2 評価手法

3.2.1 再現率 (Recall)

全正解中で、システムがどれだけ正解を見つけられたかを示す割合. 以下の式 (5) で定義される.

$$Recall = \frac{\text{抽出された正しい固有表現数}}{\text{抽出された固有表現数}} \quad (5)$$

3.2.2 適合率 (Precision)

システムが生成した結果の中でどれだけが正解だったかを示す割合. 以下の式 (6) で定義される.

$$Precision = \frac{\text{抽出された正しい固有表現数}}{\text{正解の固有表現数}} \quad (6)$$

3.2.3 アンケートの内容

各重み付け法で作成した 5 つの要約文を要約率が同等のもので互いに比較し、「可読性」、「意味の保全」、「文法の正しさ」、「受容性」の 4 つの設問に対して「tf-idf 法」、「HGS 法」、「両方良い」、「両方悪い」のうち 1 つを選んでもらう.

評価はあらかじめ原文を読んだ 9 人に行ってもらった.

3.3 結果

再現率、適合率、F 値の結果を表 1〜表 2 に示し, アンケートの結果をまとめたものを表 3〜表 7 に示す.

4 考察

要約率が低い時は tf-idf 法、高い時は HGS 法のが再現率、適合率、F 値が良いことを示した. そして、アンケートによる結果も同様であり、上記の評価法が人間の判断と近いと考えられる. この結果より単一の文書を要約するのに要約率が低いときは tf-idf 法のが適しており高いときは HGS 法のが適した方法であると考えられる.

評価	可読性	意味	文法	受容性
tf-idf	6	1	3	3
HGS	1	0	1	0
両方良い	2	8	5	5
両方悪い	0	0	0	1

表 3: 要約率 8%の要約文

評価	可読性	意味	文法	受容性
tf-idf	5	3	4	5
HGS	0	0	1	0
両方良い	4	6	3	4
両方悪い	0	0	1	0

表 4: 要約率 12%の要約文

評価	可読性	意味	文法	受容性
tf-idf	4	2	3	4
HGS	2	4	4	1
両方良い	3	3	0	4
両方悪い	0	0	2	0

表 5: 要約率 20%の要約文

評価	可読性	意味	文法	受容性
tf-idf	2	2	2	1
HGS	6	5	5	4
両方良い	0	0	0	2
両方悪い	1	2	2	2

表 6: 要約率 25%の要約文

評価	可読性	意味	文法	受容性
tf-idf	0	0	0	0
HGS	4	3	1	3
両方良い	0	0	0	0
両方悪い	5	6	8	6

表 7: 要約率 35%の要約文

5 まとめ

本研究では組み合わせ確立モデルを用いた特徴単語抽出方式である HGS 法と tf-idf 法の単一文書の要約性能を再現率、適合率、F 値及びアンケートを用いた評価実験を通して比較、評価した. その評価結果から HGS 法の最適な要約率が判明した.

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」, FIT2002,E-4

[2] 久光　徹, 丹羽　芳樹:「組み合わせ確立モデルに基づく特徴単語選択方法-超幾何学分布の応用」, 情報処理学会自然言語処理研究会,140-12,2000.

[3] 総務省編: 「平成 13 年度　情報通信白書」, ぎょうせい,2001