複数の評価尺度を統合的に用いた重要文抽出システム

野畑 周† 関根 聡‡ 村田 真樹† 内元 清貴† 内山 将夫† 井佐原 均†
† 総務省通信総合研究所　‡ ニューヨーク大学
{nova,murata,uchimoto,mutiyama,isahara}@crl.go.jp　sekine@cs.nyu.edu

1 はじめに

重要文抽出は、自動要約を行う有用な手法の一つである [8]。文章から重要文を抽出するためには、各文がどの程度重要であるかを比較する評価尺度を用意する必要がある。評価尺度を求める方法には、さまざまな手法がある。文の位置情報、たとえば文章の先頭にあるものほど重要だとみなす手法は、単純ではあるが現在でも情報抽出の主要な手法である。他にも記事中の単語の頻度などの統計的な情報や、文書構造を示す表現などの手がかりなどが用いられている。これらの評価尺度を統合的に用いる手法も研究されている。Edmundson[2] は、人手で重み付けの値を与えることによって、また野本 [10]、Watanabe[7], Aone[1] は学習によって複数の情報を統合的に用いている。

本システムは、複数の評価尺度を重み付けを用いて統合し、文の重要度を求めて重要文抽出を行う。各重み付けの値は、各要約率、記事の種別ごとに訓練データから最適値を求める。以下、本システムが参加した自動要約の評価ワークショップ TSC について述べ、次いで本システムの手法を説明し、評価結果の提示とその考察を行う。

2 TSC

TSC(Text Summarization Challenge) は、国立情報学研究所によって主催された自動要約の評価ワークショップである [5]。TSC では、以下の課題が提示された。

A1: 重要文抽出型要約 テキスト中から文を要約率に応じて抽出する。要約率は、文数を元にした原文との割合とし、対象テキストごとに要約として選択できる文数の上限が与えられる。

A2: 人間の自由作成要約と比較可能な要約 要約をプレインテキストで作成し提出する。要約率は、文字数を元にした原文の割合とし、対象テキストごとに要約の上限となる文字数が与えられる。

B: IR タスク用要約 提示した検索要求と、その検索結果としてのテキストを元に要約を作成し提出する。

課題 A1 では、人間が選択した重要文との間の一致度を元に再現率・精度・F 値の 3 種類の評価が行われた。課題 A2 では、人間の作成した正解要約との類似度を単語頻度ベクトルを用いて評価したもの (Content-based) と、要約に熟練した人間による内容・読み易さの順位付けとの 2 種類の評価が行われた。課題 B では、被験者に検索要求とその検索結果としてのテキストの要約を提示し、被験者が行った情報検索タスクに基づく評価が行われた。

3 手法

本節では、重要文抽出で用いた手法を紹介する。まず評価尺度を与える関数について説明し、次にしきい値、規則、パラメータ等その他の部分について説明する。

3.1 評価尺度

3.1.1 文の位置情報

本システムでは、文の位置情報に基づく関数を 3 種類用いている。一つ目の関数は、出力すべき文が N 文であると指定されたときに、記事の先頭から N 文目までにスコア 1 をつけ、それ以外は 0 とするものである:

$$\begin{aligned} Score_{\mathrm{loc}}(S_i)(1 \le i \le n) &= 1(\text{if} \quad i < N) \\ &= 0(\text{otherwise}) \end{aligned} \quad (1)$$

ここで n は記事中の文の数を示す。2 つ目の関数は、文の位置の逆数を与えるものである。つまり i 番目の文に対するスコアは、

$$Score_{\mathrm{loc}}(S_i) = \frac{1}{i} \quad (2)$$

となる。これら 2 つの関数は、先頭に近い文ほど重要である、という仮定に基づいたものである。

3 つ目の関数は、2 つ目の関数に手を加え、先頭からの文の位置と末尾からの文の位置を共に用いるものである。この関数は、先頭か末尾に近い文ほど重要である、という仮定に基づいている。

$$Score_{\mathrm{loc}}(S_i) = \max(\frac{1}{i}, \frac{1}{n-i+1}) \quad (3)$$

3.1.2 文の長さ

文の長さを用いる関数では、短い文ほど重要な文になりにくい、と仮定している。本システムでは、各文の文字数に基づく2種類の関数を用いている。一方は、文の長さをそのままスコアとして与える。もう一方は各文の長さ (L_i) が一定の値 C より短ければペナルティとして負の値を与えるものである：

$$\begin{aligned} Score_{\mathrm{len}}(S_i) &= L_i \quad (\text{if} \quad L_i \geq C) \\ &= L_i - C \quad (\text{otherwise}) \end{aligned}$$

3.1.3 TF/IDF

この関数は、記事中の単語の頻度 (*tf*) と、その単語がある文書群の中で現れた記事の数 (*df*) を用いてTF/IDF値を計算し、文のスコア付けを行う。本システムでは、単語の切り分けにはJUMAN[12]を用い、TF/IDF値を与える単語を、時相名詞や副詞的名詞を除いた名詞に限定した。記事群には、1994年と1995年の毎日新聞の記事を用いた。

TF/IDF値は以下の式によって求めた。ここで DN は与えられた記事群の記事数である[4]：

$$\mathrm{TF/IDF}(w) = \frac{tf}{1+tf} \log \frac{DN}{df}$$

3.1.4 見出し

本システムは、対象記事の見出しに含まれる単語に対するTF/IDF値を用いて文のスコア付けを行なう。注目する単語は、前節のTF/IDFを用いた関数と同様に、時相名詞や副詞的名詞を除いた名詞に限定している。文中の全名詞について、その名詞が見出しに含まれていればそのTF/IDF値を文のスコアに加算する。我々は、名詞の代わりに固有表現を用いてこの関数を計算することも行った。固有表現の抽出には、最大エントロピー法を用いたシステムを使用した[6]。固有表現を用いる際には、TF/IDF値ではなく記事中の単語の頻度のみを用いた。この理由は、固有表現が記事群中に現われる頻度は通常低く、TF/IDF値では固有表現間の違いが明確にならないと考えたからである。

3.1.5 検索要求

課題Bでは、対象記事に加えて検索要求が提示される。本システムは、前節の見出しを用いた評価尺度と同様に、名詞について、その名詞が見出しに含まれていればそのTF/IDF値を文のスコアに加算する。

課題Bでは、我々は二通りの要約結果を提出した。以下、これらをSum1、Sum2と呼ぶ。Sum1は、課題Bにおいて精度の評価を上げることを重視して作成した。要約率は基本的に10%としたが、最低でも記事ごとに3文を出力する。各スコア関数の重み付けは、課題A1で要約率を10%としたときのものを用いた。検索要求に対する重みの値の変更は、見出しの情報を利用して行った。すなわち、検索要求にある名詞がその記事の見出しにも現れるならば、検索要求の重みを倍にした。さらに文のスコアに対するしきい値も設定した。このしきい値以下の文は課題Bの要約結果には出力されない。

3.2 しきい値

本システムでは、重要文抽出を行う際に、与えられた記事中の全文にスコア付けを行い、その結果を元にスコアの良い順に各文を順位付けする。これらの文のうち何文まで出力するかを決定するのに、本システムでは文数・文字数・スコアの3種類のしきい値を用いることができる。どのしきい値を用いても、出力される文の順番が元の記事のまま保たれることは変わらない。

文の数 N がしきい値として与えられたならば、システムは順位付けされた文の上位 N 文までを重要文として抽出する。文字数が与えられたときには、システムはこれを文数のしきい値に変換する。すなわち、与えられた文字数を越えるまで順位の小さい順に各文の文字数を加算し、文字数に納まる文の数を求める。文の数が求まれば、それをしきい値として用いることができる。スコアがしきい値として与えられたならば、システムはそのしきい値より大きい値をもつ文のみを出力する。

3.3 変換規則

本システムでは、課題A2において文を短縮する変換規則を用いている。これは、できるだけ多くの文を要約に含めたいという意図に基づく。このような文短縮を行う変換規則の作成にはいくつかの先行研究がある。若尾ら[9]は、ニュース番組の字幕を生成するための規則を人手で作成し、その評価を行なっている。加藤・浦谷[11]は、ニュースの番組原稿と文字放送用の原稿から変換規則を自動的に作成する手法を提案している。本システムについては、TSCの予備実験で用いられたデータを元に人手で文を短縮する規則を作成したが、このような規則を自動的に生成することも今後の研究課題として行いたいと考えている。

3.4 重みづけ

本システムでは、各スコア関数のスコア付け結果に重み付けをしたものの総和をとり、各文の重要度を与える。

この重み付けの最適値は、TSCの予備試験のデータを用いて求めた。このデータは、30の新聞記事と、各

要約率 (10%, 30%, 50%) ごとに作成されたそれらの記事の要約とから成る。重要文抽出の課題では、我々はこれらの新聞記事を更に社説 15 記事とそれ以外の 15 記事とに分けて、それぞれについて最適な重み付けを求めた。文の位置情報や長さの関数については、その種類の選択も重み付けとともに行なわれる。

一方、自由作成要約の課題では、要約率は 20%と 40%の 2 種類が設定されている。この課題では、重要文抽出の課題で求めた重み付けをそのまま用いた。すなわち要約率 20%の課題には 10%の、要約率 40%の課題には 30%の重み付けを用いた。

4 結果と考察

本節では、本システムが TSC の各課題で得た評価結果を示し、その考察を行う。

4.1 A1: 重要文抽出型要約

表 1 は、本システムの重要文抽出での評価結果を、ベースラインシステムの結果とともに示したものである。本システムは、平均·各要約率の各評価においてベースラインシステムの結果を上回る成果を得た。システムが正しく出力できなかった原因を調べたところ、正解に含まれない文には短い文が多くあった。また、TF/IDF や見出しを用いた関数は事実を具体的に記述している文に高いスコアを与える傾向があるが、正解として与えられた要約では、短いより抽象的な表現が多く選択されていた。各スコア関数のうち一つを無効にしたときの要約性能の変化を表 2 に示す。文長·TF/IDF·見出しを用いた関数に対する結果をみると、これらは性能にそれほど貢献していなかったことが分かった。このことが先にふれた正解とシステムの出力との違いを示しているといえる。

社説とそれ以外の記事を分割したことは、性能の向上に意味があった。特に文の位置情報を用いた関数に使い分けに拠るものが大きかった。表 3 に、記事の種類別に見た文の位置情報の評価結果を示す。各関数の種類は式の番号に対応している。すなわち、Loc.1 は式 1 で示された関数を表す。Loc.1 と Loc2 は、文の位置情報を単独で用いたときには同じ値を返すので、ここでは一つにまとめた。表 3 に示されるように、Loc1,Loc2 は社説以外の記事で Loc.3 より高い結果を示し、Loc.3 は社説において Loc1,Loc2 より高い結果を示した。これらを適切に使い分けたのが位置情報を用いた本システムの評価尺度である。

表 1: 課題 A1 の評価結果

要約率	10%	30%	50%	Ave.
本システム	0.363	0.435	0.589	0.463
Lead-based	0.284	0.432	0.586	0.434
TF-based	0.276	0.367	0.530	0.391

表 3: 記事の種類別に見た位置情報の評価結果

Loc.1, Loc.2				
要約率	10%	30%	50%	Ave.
社説	0.158	0.256	0.474	0.293
その他	0.394	0.478	0.586	0.486
全体	0.276	0.367	0.530	0.391
Loc.3				
社説	0.323	0.360	0.557	0.413
その他	0.356	0.436	0.544	0.445
全体	0.339	0.398	0.550	0.429
本システム				
全体	0.359	0.419	0.572	0.450

4.2 A2: 人間の自由作成要約と比較可能な要約

課題 A2 では、我々は重要文抽出の結果に文短縮規則を適用した結果を提出した。Content-based の評価結果を表 4 に示す。「正解要約 (FREE)」は、文字数の制約以外は自由に作成された正解要約、「正解要約 (PART)」は、元の記事から重要な語句を抽出することで作成された正解要約である。Content-based の評価結果では、本システムはベースラインシステムとそれほど差がなく、要約率 40%の場合においてはベースラインシステムよりも悪い。この原因としては、各関数に対する重み付けが適していなかったことが考えられる。

文短縮規則を適用した目的は、できるだけ多くの文を要約に含めることであった。表 5 は、規則適用前と適用後での一記事あたりの平均文数の変化を示している。要約率 20%においてはパターン適用の効果はほとんどないが、要約率 40%においては、3 分の 1 の記事で文の数が増えたことが分かる。

4.3 B: IR タスク用要約

課題 B では、先に述べたように、我々は 2 種類の結果 Sum1, Sum2 を提出した。表 6 に課題 B の評価結果を示す[1]。本システムの評価結果は、A,B 判定を正解とした場合の方がベースラインシステムとの差が大きい。我々の提出した要約は、検索要求に何らかの関連があ

[1]情報検索の課題設定は IREX[3] に基づいている。記事が検索要求にどの程度適合しているかという判断を示す判定の基準には A,B,C の 3 段階があり、それぞれ、A: 記事の主題が検索課題に関連している、B: 主題ではないが記事の一部が関連する、または何らかの関連がある, C: 関連しないという判断を示す。

表 2: 各尺度を除いたときの課題 A1 の評価結果

要約率	10%	30%	50%	Ave.
全評価尺度 (ALL)	0.363	0.435	0.589	0.463
(ALL)− 文の位置	0.326(−.037)	0.394(−.041)	0.575(−.014)	0.432(−0.031)
(ALL)− 文長	0.372(+.009)	0.472(+.037)	0.600(+.011)	0.481(+0.018)
(ALL)−TF/IDF	0.372(+.009)	0.439(+.004)	0.582(−.007)	0.464(+0.001)
(ALL)− 見出し:単語	0.403(+.040)	0.449(+.014)	0.589(±.000)	0.480(+0.017)
(ALL)− 見出し:NE	0.381(+.018)	0.438(+.003)	0.589(±.000)	0.469(+0.006)

表 4: 課題 A2：Content-based の評価結果

正解要約 (FREE) との比較			
要約率	20%	40%	Ave.
本システム	0.452	0.566	0.509
TF-based	0.437	0.596	0.516
Lead-based	0.383	0.580	0.481
正解要約 (PART) との比較			
本システム	0.507	0.611	0.559
TF-based	0.476	0.622	0.549
Lead-based	0.421	0.605	0.513

表 5: 課題 A2 での一記事あたりの平均文数

要約率	20%	40%
規則適用あり	4.53 (136/30)	8.93 (268/30)
規則適用なし	4.63 (139/30)	9.27 (278/30)

る記事を関連しない記事と区別する情報を他システムよりも多く提供していたと考えられる。

5 まとめ

複数の評価尺度を統合的に用いた重要文抽出システムを用いて、自動要約の評価ワークショップの全課題において要約の評価を行った。重要文抽出の課題では、本システムの手法が有効であることが示された。自由作成要約の課題では、評価結果はそれほど良くなかったが、変換規則によって文短縮を行うことで、要約に含まれる情報量を増すことができた。情報検索のための要約課題では、提出した双方の要約結果が共に適合度を広くとった場合に良い評価を示した。

表 6: 課題 B の評価結果

A 判定のみを正解とした場合			
Measurement	再現率	精度	F
Sum1	0.833	0.728	0.761
Sum2	0.899	0.717	0.785
Full text	0.843	0.711	0.751
TF-based	0.798	0.724	0.738
Lead-based	0.740	0.766	0.731
A,B 判定を正解とした場合			
Sum1	0.741	0.921	0.808
Sum2	0.793	0.904	0.828
Full text	0.736	0.888	0.773
TF-based	0.700	0.913	0.776
Lead-based	0.625	0.921	0.712

参考文献

[1] C. Aone, M. E. Okurowski, and J. Gorlinsky. Trainable, Scalable Summarization Using Robust NLP and Machine Learning. In *Proc. of COLING-ACL'98*, pp. 62–66.

[2] H. Edmundson. New methods in automatic abstracting. *Journal of ACM*, Vol. 16, No. 2, pp. 264–285, 1969.

[3] IREX. http://cs.nyu.edu/cs/projects/proteus/irex. Information Retrieval and Extraction Exercise, 1999.

[4] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retreival. In *Proc. of SIGIR'94*.

[5] TSC. http://galaga.jaist.ac.jp:8000/tsc/. Text Summarization Challenge, 2001.

[6] K. Uchimoto, Q. Ma, M. Murata, H. Ozaku, and H. Isahara. Named Entity Extraction Based on A Maximum Entropy Model and Transformation Rules. In *Proc. of ACL2000*, pp. 326–335.

[7] H. Watanabe. A Method for Abstracting Newspaper Articles by Using Surface Clues. In *Proc. of COLING'96*, pp. 974–979.

[8] 奥村学, 難波英嗣. テキスト自動要約に関する研究動向 (巻頭言に代えて). 自然言語処理, Vol. 6, No. 6, pp. 1–26, 1999.

[9] 若尾孝博, 江原輝将, 白井克彦. テレビニュース番組の字幕に見られる要約の手法. In *IPSJ-NL 122-13*, pp. 83–89, July 1997.

[10] 野本忠司, 松本祐治. 人間の重要文判定に基づいた自動要約の試み. In *IPSJ-NL 120-11*, pp. 71–76, July 1997.

[11] 加藤直人, 浦谷則好. 局所的要約知識の自動獲得手法. 自然言語処理, Vol. 6, No. 7, pp. 73–92, 1999.

[12] 黒橋禎夫, 長尾真. 日本語形態素解析システム JUMAN version 3.61. 京都大学, 1999.