検索結果表示向け文書要約における 情報利得比に基づく語の重要度計算

森 辰 則[†]

本稿では、情報検索の結果として得られた文書集合中の各々の文書を要約する一手法を提案する。この場合の要約の質は、検索質問-要約文書間の関連性判定が、検索質問-原文書の間の関連性判定に一致する度合で評価されるので、検索を考慮した要約が必要となる。検索質問により語の重みにバイアスを与え、語の重要度を求める従来手法とは異なり、我々の方法では、検索された文書間の表層的類似性を適切に説明する語に高い重みを付与する。具体的には、検索文書集合に階層的クラスタリングを適用することにより、文書間の類似性構造を抽出するとともに、各クラスタにおける各語の出現確率から、その構造を説明するのに寄与する単語により高い重みを与える。我々は、その重みづけに情報利得比を用いることを提案する。そして、この語の重み付けに基づき重要文抽出方式による検索文書要約システムを実装した。このシステムを評価型情報検索ワークショップである NTCIR2 における Text Summarization Challenge の情報検索タスクにより評価した結果、関連性判定において検索質問バイアス付き TF 方式、リード文方式によるベースライン手法ならびに、他参加システムよりも、良好な結果を示した。

キーワード: 文書要約, 重要文抽出, 情報利得比, 情報検索, 文書クラスタリング

A Term Weighting Method based on Information Gain Ratio for Summarizing Documents retrieved by IR Systems

TATSUNORI MORI[†]

This paper proposes a new term weighting method for summarizing documents retrieved by IR systems. Unlike query-biased summarization methods, our method utilizes not the information of query, but the similarity information among original documents by hierarchical clustering. In order to map the similarity structure of the clusters into the weight of each word, we adopt the information gain ratio (IGR) of probabilistic distribution of each word as a term weight. If the amount of information of a word in a cluster increases after the cluster is partitioned into sub-clusters, we may consider that the word contributes to determine the structure of the sub-clusters. The IGR is a measure to express the degree of such contribution. We show the effectiveness of our method based on the IGR by comparison with other systems in Text Summarization Challenge of NTCIR2.

KeyWords: *Text Summarization, Extraction of Important Sentences, Information Gain Ratio, Information Retrieval, Document clustering*

[†] 横浜国立大学大学院環境情報研究院, Graduate School of Environment and Information Sciences, Yokohama National University

1 はじめに

近年, Internet 上の検索エンジンなど, 情報検索システムが広く利用されるようになってきた. システムが提示する検索結果には, 文書の表題や URI だけではなく, 対応する文書の内容を示す短い要約文書が併せて提示されていることが多い. これは, 利用者に対して要約文書を提示することが, 原文書が実際に利用者の欲するものかを判断する際に有力な手掛かりとなるためである. この際, 情報検索結果文書に対する要約の質の良さは, 要約文書-検索質問間の関連性判定と原文書-検索質問間の関連性判定の一致の良さで測ることができよう.

しかしながら, 現在実用に供されている多くの検索エンジンでは, 原文書の最初の数バイトを出力したり, 検索要求文に含まれる語の周囲を提示するといった単純な方法が採用されている. このような単純な戦略により生成された要約の品質は関連性判定の観点からみると, 十分な品質であるとは言い難い. そのため, 多くの場合, 利用者はシステムの提示した検索結果が適切なものであるかどうかを原文書を見て判断せざるを得ない. このような状況を改善するためには, 関連性判定を重視した, より質の高い自動要約技術が必要となる.

自動要約の手法としては, Luhn(Luhn 1958)の研究に端を発する重要文抽出法が基本かつ主要な技術であり, 依然として様々なシステムで利用されている. これは, 文書の中から重要な文を, 所望の要約文書の長さになるまで順に選び, それら抽出された文を文書中での出現順に並べて出力することで, 要約とする手法である. このとき, 文の重要度は, 語の重要度, 文書中での位置, タイトルや手がかり表現などに基づいて計算している(奥村, 難波 1999, 2000). その中でも, 重要文は主要キーワードを多く含むという経験則により, 語の重要度に基づく重要文抽出が最も基本的な手法となっている. 特に, 語の出現頻度は, 簡単に求められ, 語の重要度と比較的高い相関にあるために語の重みとして広く利用されている.

語の出現頻度は個別文書によって決まる性質であるが, 一方で, 検索文書の要約においては, 原文書が検索要求の結果として得られた複数の文書であることを考慮することが要約の品質向上につながる. 例えば従来提案されている基本的な考え方として, 検索要求中の語の重要度を高くするという「検索質問によるバイアス方式」がある(Tombros and Sanderson 1998). この手法は直観的であり, かつ, 比較的良好に機能するが, 検索された文書自身の情報を考慮しないなど幾つかの欠点が存在する.

以上の点を踏まえて, 本稿では, 検索文書集合から得られる情報を語の重みづけに利用し, 検索文書の要約に役立てる新しい手法を提案する. 検索質問によるバイアス方式とは異なり, 我々の手法では語の重みづけにおいて検索質問の情報を陽に利用しない. その代わりに, 複数の検索文書の間に存在する類似性の構造を階層的クラスタリングにより抽出し, その構造を適切に説明するか否かに応じて語に重みをつける. 文書間の類似性構造を語の重みに写像する方法として, 我々は, 各クラスタ内での語の確率分布に注目し, 情報利得比 (Information Gain Ratio, IGR)(Quinlan 1993)と呼ばれる尺度を用いる. そして, この重みと従来提案されている

他の重みづけを組み合わせることにより、最終的な語の重みとし、これを用いて各文の重要度を計算する。

特に、情報利得比に基づく語の重みづけについては、次のように考えることができる。あるクラスタにおける語の情報量に注目した場合、そのクラスタを部分クラスタに分割した後のその語の持つ情報量の増分 (情報利得) が、クラスタの分割自身により得られる情報量に比して大きければ、その語は部分クラスタの構造を決定する際に役立っていると考えられる。その度合を定量化した値が情報利得比である。情報利得比自身は機械学習において属性の品質の尺度として、すでに提案されているものである。また、種々のクラスタリングアルゴリズムの過程からすれば、文書のクラスタ構造の決定に際して、各々の語の確率分布が部分的な要因となっていることは自明である。しかしながら、あるクラスタ構造が確定した時に、ある語がそのクラスタ構造の決定に際して最終的に寄与したか否かに注目し、定量化するという研究は、我々の知る限り従来存在しない。そして、本稿は、その定量化において、情報利得比が利用できることを示すものである。

2 検索文書集合中の語に対する情報利得比に基づく重みづけ

2.1 検索文書要約の特質

検索文書の要約は、以下の点で通常の文書要約と異なる。

- 検索質問文が与えられている。
- 複数の要約文書が同時に得られている。そして、ある一度の検索の結果という点において文書間に類似性が認められる。

いずれの情報も、検索文書の要約においては有効な手がかりと考えられる。ここでは、これらの手掛かりを語の重みづけに用いることを考察する。

まず初めに考えられる手法は、Tombros ら (Tombros and Sanderson 1998) が提案するように、検索質問中の語を重要語として考え、他の語よりも重みを高する方法である。これは、検索質問中の語や句は利用者の情報要求を端的に示すので、要約文書にもその語や句が含まれるべきであるという直観に基づくもので、「検索質問バイアス方式に基づく要約 (Query-biased Summarization)」と呼ばれる。

この方法は、検索質問を考慮するだけであるので、実装が簡単であり、ある程度の効果が報告されているが、次の欠点が存在する。

- 検索質問中の表現をそのまま用いるために、検索エンジンにおける工夫が要約文書に反映されない。例えば、各種フィードバックや検索質問の拡張などは、検索質問を修正/更新することによって検索効率を上げている¹。

¹ ここでは、既存の情報検索システムに対するバックエンドとして要約生成システムを利用する場合を想定している。この場合、情報検索システムを運用している組織と要約サービスを提供している組織が必ずしも一致しない。この状況に

- 検索エンジンは検索質問文に関連する文書ばかりではなく、関連性の低い文書も結果として返すことがある。検索質問との関連性が低い文書に対しては、検索質問バイアス方式は通常の文書要約になってしまう。

そこで、我々は二番目の選択肢である、検索質問文を使わずに、検索文書集合のみを用いて重みづけることを考える。検索結果の質が非常に悪くない限り、検索文書集合には検索質問に関する情報が暗に含まれていると期待できるので、その情報を引き出すのである。しかし、各文書に共通する語を抽出するといった単純な方法では精度が良くないことは容易に想像される。なぜならば、検索結果の文書集合には、もちろん検索質問との関連性が高い文書も含まれるが、関連性の低い文書も含まれるからである。しかも、その度合は検索エンジンの精度に依存してしまう。よって、単純に文書に共通する語などを取り出すだけでは達成できない。我々は以上の点を踏まえて次節に示す枠組を提案する。

2.2 提案手法の概略

我々の提案手法の概略を図1に示す。これは、次の2つの指針の組み合わせたものである。

- (1) 検索文書集合に対し階層的クラスタリングを行ない、文書間の類似性の構造を抽出する。
- (2) 文書間の類似性の構造と語の確率分布に基づいて各語の重み付けを行なう。

ステップ1においては、検索質問に関連する文書とそうでない文書がクラスタ構造の中で分離されることが期待され、なおかつ、それらの文書部分集合においても類似性に基づく細分類がなされと考えられる。我々の手法では、語を次元とする文書ベクトルの類似度による階層的クラスタリングを用い、クラスタリングアルゴリズムとしては最大距離アルゴリズムを採用した。

ここで注意すべきことは、検索されなかった文書、すなわち、文書データベース中の残りの文書の存在を類似性構造の中に組み込む必要があることである。なぜならば、クラスタ構造において、一番上位のクラスタは与えられた構造として扱う以外になく、類似性の解析の対象とならないからである。検索文書全体の類似性構造は、検索されなかった文書集合との対比によって、初めて明らかになる。この類似性構造は、全文書集合と検索質問との間の類似性を反映するので、非常に重要なものである。このため、図2に示すように検索文書集合から得られたクラスタ構造の根の上にもう一つ仮想的なクラスタを設ける。そのクラスタには、検索文書の属する部分クラスタとそれ以外の文書が属する部分クラスタが存在する。ここで、我々の方法では、検索文書の属する部分クラスタだけが、クラスタリングアルゴリズムによる部分クラスタ解析の対象となる(図2の左部分木)のに対し、検索されなかった文書の属するクラスタ(図2の右部分木)についてはそれ以上の解析が不要である点に注意されたい。後に述べる情報利得比の計

においては、情報検索システムが行なっている工夫に関する情報が利用できない。一方、もしも情報検索システムの行なうフィードバックや検索質問拡張の情報が利用できるのであれば、更新後の検索要求に基づき、検索質問によるバイアス手法を適用することも可能である。

図 1 提案手法の枠組

算においては、検索されなかった文書の属するクラスタについては、その中に存在する語の頻度だけが分かればよい。これは、あらかじめ求めておいた文書データベース中の語の頻度より、簡単に求めることができる。よって、我々の手法において実際に文書クラスタリングが行なわれるのは、検索結果として利用者に提示する文書に限定される。これは通常数十文書程度であるから、文書クラスタリングにおける計算量はあまり問題とならない。

このようにして求められた類似性構造は文書を一つの単位とする巨視的な情報であるので、要約のためには、これを文や句、単語を単位とするより微視的な情報に還元する必要がある。これがステップ2である。このステップにおいては、各クラスタが部分クラスタに分割されるにあたって、句や単語などより微視的な単位がどれ位寄与しているかを表す指標を定め、これをその語句の重要度とする。我々はその指標として情報利得比を用いる。

図 2 検索文書に対するクラスタ構造

このようにして求められた情報利得比を、既存の方法でも用いられている重みである、語の文書内頻度 (term frequency, TF), 文書頻度の逆数 (inverse document frequency, IDF)(Salton and Buckley 1988; Baeza-Yates and Ribeiro-Neto 1999) と組み合わせることにより、総合的な語の重要度とする。これら三種類の重みは、以下の通り、異なる文書情報から得られるものであることに注意されたい。よって、これらの組み合わせにより、検索文書の要約に適した総合的な重みが得られると期待できる。

- 語の文書内頻度
個別の文書における各語の分布より決まる重要度で、ある文書中でのその語の重要度を表す。
- クラスタ分割に対する語の出現確率に関する情報利得比
検索文書の類似性構造であるクラスタの分割により決まる重要度で、そのクラスタ構造におけるその語の重要度を表す。
- 文書頻度の逆数
検索対象の全文書により決まる重要度で、全検索対象文書集合におけるその語の重要度を表す。

2.3 最大距離アルゴリズムによる階層的クラスタリング

検索された文書集合の類似性の解析には、文書間の距離の定義とその距離に基づく文書集合の構造化が必要となる。これには様々な方法が考えられるが、本稿では文書間の距離として、様々な場面で利用され、かつ、簡便な TF・IDF 法ならびにベクトル空間法に基づく方法を採用する (Baeza-Yates and Ribeiro-Neto 1999)。また、文書集合の類似性に関する解析には、階層的クラスタリングを用いる。

階層的クラスタリングアルゴリズムには、併合法などが良く用いられる (Frakes and Baeza-Yates 1992)。しかし、この種の方法はクラスタ構造を無理に二分木に当てはめるため、文書間

距離の順序関係は構造に反映されるものの、その絶対値については捨象されてしまう。ここでは、文書の類似度に応じて多分岐構造を生成でき、文書間距離の絶対値情報になるべく保存されるアルゴリズムが望ましい。そのようなアルゴリズムとして、我々は最大距離アルゴリズム(長尾 1983)を採用した。この最大距離アルゴリズムは、本来、非階層的なクラスタを生成するものであるが、これを分割の結果得られた部分クラスタに再帰的に適用することにより階層構造を得る。この方法では、各クラスタが 3 以上のクラスタに分割されることもある。

2.3.1 文書間距離

ベクトル空間モデルに基づき、各文書 D_i を n 次元空間上の点 $(weight_{i1}, weight_{i2}, \dots, weight_{in})$ により表現する。 $weight_{ik}$ は文書 D_i において語 w_k に割り当てられた重みである。重み $weight_{ik}$ としては語 w_k の TF・IDF 値とする。このとき、文書 D_i と文書 D_j の距離 d を文書ベクトル間のユークリッド距離を用いて、次のように定義する。

$$\begin{aligned} d(D_i, D_j) &= \sqrt{\sum_k (weight_{ik} - weight_{jk})^2}, \\ weight_{ik} &= tf(D_i, w_k)idf(w_k), \\ tf(D_i, w_k) &= \frac{freq(D_i, w_k)}{|D_i|}, \\ idf(w_k) &= 1 + \log_2 \frac{N}{df(w_k)}, \end{aligned} \quad (1)$$

ただし、

- $freq(D_i, w_k)$: 文書 D_i での語 w_k の出現頻度
- $|D_i|$: 文書 D_i 中の形態素数
- $df(w_k)$: 検索対象の全文書集合における語 w_k を含む文書数
- N : 検索対象の全文書の数

である。

2.3.2 最大距離アルゴリズム

最大距離アルゴリズムにおいては、まず文書集合から 2 個以上のクラスタ中心を選択し、次に、残りの文書を最近のクラスタ中心と同じクラスタに配置する。その主要部分はクラスタ中心を求める部分であり、以下の手続きからなる。

- (1) 文書集合 DS から距離の最も大きい二文書を取り出し、これらを要素とする集合を作成する。これを初期のクラスタ中心の集合 C とする。
- (2) クラスタ中心集合 C において、クラスタ中心間での最大距離を求める。これを、 d_{max} とする。
- (3) DS 中の各文書 D_i について、すべてのクラスタ中心との距離を求め、その最小値を既存クラスタ中心からの距離 $d(D_i, C)$ とする。既存クラスタ中心からの距離が最も

大きい文書 D_d を DS から取り出す。

- (4) もし、 $d(D_d, C) \geq \alpha \cdot d_{max}$ ならば、その文書をクラスタ中心集合 C に追加する。そうでなければ、終了。

なお、 α は $0.5 \leq \alpha < 1.0$ なる定数であり、値が大きいほどクラスタの分割数が少なくなる。一般には 0.5 とすることが多い。

以上のアルゴリズムは、単一の文書集合を文書間距離にしたがって複数個の部分クラスタに分割する非階層的なアルゴリズムである。これを各部分クラスタに対して再帰的に適用することにより、階層的なクラスタ構造を生成する。

2.4 情報利得比に基づく語の重要度

クラスタの木における各接点 (内点) は、あるクラスタとそれを分割して得られた互いに素な部分クラスタの関係、すなわち、クラスタの分割の仕方を表現している。この分割の仕方はクラスタ内の文書の類似度に従って決定されるので、これを文書内の語の重みに反映させることができれば、複数文書間の類似性という巨視的な情報を、文書内の語の重みという微視的な情報に還元できると考えられる。

我々は、この考え方にに基づき、次の 2 つの段階から構成される方法を提案する。

- (1) 各クラスタについて、その部分クラスタの構造から、各語の重みを決定する。
- (2) 一つの文書は、クラスタの木の根接点から対応する葉接点に至るクラスタ分割の系列によって指し示される。よって、各文書における語の重みは、各分割で得られた語の重みを統合して得る。

このうち、特に重要なのは 1 である。その基本的な考え方は、クラスタの分割構造を決定することに寄与する語に高い重みを与えるというものである。例として、図 3 のように、あるクラスタ C_0 が 3 つの部分クラスタ (C_1, C_2, C_3) に分割されている場合を考える。図中、記号 $A, B, D \sim G$ は各々単語に対応するとする。さて、語 A はクラスタ C_0 における頻度が最も高いので、このクラスタの特徴を表す語と考えることができる。しかし、各部分クラスタに注目すると、いずれも均等に出現しているため、部分クラスタの選択においては役立たないことがわかる。一方、語 F はクラスタ C_0 において頻度はさほど高くはないが、部分クラスタ C_3 に集中して登場している。この場合、語 F が出現しているか否かを調べることによって、部分クラスタを言い当てることができるので、クラスタの分割構造に対する寄与度は、語 F は語 A よりも高いと考えられる。我々はこの寄与度を適切に表す尺度として、次に述べる情報利得比を用いる。

2.4.1 情報利得比

情報利得比は、決定木学習システム C4.5 において属性選択を行なうために導入された (Quinlan 1993; Mitchell 1997)。C4.5 においては、ある属性を決定木の分岐におけるテストとし

図 3 語の出現分布とクラスタの分割

たときに、その属性がどれくらい適切にクラスの出現を予測できるかを表す尺度として用いられている。我々は、表 1 に示す対応の下、クラスタの構造を決定木の構造と見なすことにより、情報利得比を用いる。C4.5 においては属性の評価値として情報利得比を用いていたが、我々の方

表 1 提案手法と C4.5 における計算方法の対応

提案手法	C4.5
クラスタの分割構造	属性によるテスト
単語の出現確率	クラスの出現確率

法においては、属性ではなくクラスに対応する単語に対する評価値として情報利得比を用いる。

クラスタ C における単語 w の情報利得比 $gain_r(w, C)$ は次の様に求められる。

$$\begin{aligned}
 gain_r(w, C) &= \frac{gain(w, C)}{split_info(C)} \\
 gain(w, C) &= entropy(w, C) - entropy_p(w, C) \\
 entropy(w, C) &= -p(w|C) \log_2 p(w|C) \\
 &\quad -(1 - p(w|C)) \log_2 (1 - p(w|C)) \\
 p(w|C) &= freq(w, C) / |C| \\
 entropy_p(w, C) &= \sum_i \frac{|C_i|}{|C|} entropy(w, C_i) \\
 split_info(C) &= -\sum_i \frac{|C_i|}{|C|} \log \frac{|C_i|}{|C|} \\
 freq(w, C) &: \text{クラスタ } C \text{ 中の語 } w \text{ の頻度} \\
 C_i &: C \text{ における } i \text{ 番目の部分クラスタ} \\
 |C_i| &: \text{クラスタ } C_i \text{ 中の総形態素数}
 \end{aligned} \tag{2}$$

情報利得 $gain(w, C)$ は、クラスタ C の分割の前後における、語 w の確率分布に関するエントロピーの減少量を表す。 $split_info(C)$ は、クラスタ C の分割に関するエントロピーである。情報利得比 $gain_r(w, C)$ は、これらの比として定義される。

例として、図 3 における各語について、上述の方法により情報利得比を計算してみると次の通りとなる。

$$\text{gain}_r(B, C_0) \simeq \text{gain}_r(F, C_0) > \text{gain}_r(E, C_0) = \text{gain}_r(G, C_0) > \text{gain}_r(D, C_0) > \text{gain}_r(A, C_0)$$

$$\text{gain}_r(A, C_0) = 0.000, \text{gain}_r(B, C_0) = 0.161, \text{gain}_r(D, C_0) = 0.031,$$

$$\text{gain}_r(E, C_0) = 0.080, \text{gain}_r(F, C_0) = 0.157, \text{gain}_r(G, C_0) = 0.080$$

B や F のようにクラスタ構造に沿って現れる語は値が大きく、語 A のように網羅的に分布する場合には値が小さいことは既に述べたとおりである。一方、語 D のように一部のクラスタに集中してはいるものの、他のクラスタにも低い確率ではあるが出現する場合には、値が低くなることがみてとれる。さらに、語 F とほぼ同じく偏りがあるが出現確率が低い語 E については、その値が相対的に低くなる。

2.4.2 情報利得比に基づく語の重要度

式 (2) に示される情報利得比は、各クラスタの分割毎に得られる。これらを、ある文書中のある語の重みとして利用する方法には、利用者向けインタフェースの設計に応じて、いくつか考えられる。

例えば、クラスタ構造を利用者に提示しながら、部分クラスタを順次利用者に選択してもらうような対話的インタフェースにおいては、各選択点において利用者が注目しているクラスタにおける情報利得比を利用し、語の重みを求めることが考えられる。

また、すべての検索文書を同時に要約し、一覧形式で利用者に提示するというインタフェースにおいては、クラスタの木の根接点からその文書に対応する葉接点に至るすべての分割で得られたの情報利得比を何らかの形で統合し、これを語の重みとすることが考えられる。この時、統合の方法には様々な方法が考えられ得る。例えば、ある階層の値を採用する、最大値を採る、すべての値を和もしくは積により統合する、などである。

本稿では、すべての検索文書を同時に要約し、一覧形式で利用者に提示するという最も基本的なインタフェースを想定し、図 4 ならびに式 (3) に示す情報利得比の重みなしの和を採用する。この方法では、すべてのクラスタ分割における情報利得比を等しく考慮することになる。

$$\text{igr}(w, D) = \sum_{C \in Cset(D)} \text{gain}_r(w, C) \quad (3)$$

$Cset(D)$: 文書 D の属するすべてのクラスタの集合

以上で定義された情報利得比 $\text{igr}(w, D)$ に基づく重みにより、文書 D 中の語 w の重要度 $\text{weight}(w, D)$ を定義する。以前述べたように、語の重要度には TF, IDF, IGR の各値の組み合

図 4 情報利得比による各文書中の語の重みづけ

わせを考えるが、各値が独立に重要度に寄与するものとし、組み合わせ方法として積を用いる。

$$weight(w, D) = igr(w, D) \cdot tf(w, D) \cdot idf(w) \quad (4)$$

なお、上記の重みを求めるにあたって、変更可能なパラメタを導入し、訓練事例などを用いて、これを適切な値に設定する(チューニングする)ことも考えられる。例えば、情報利得比の統合を、各階層の情報利得比の重み付き和により実現し、その重みをパラメタとすることが考えられる。また、上述の $weight(w, D)$ の式を積ではなく、重み付き和として定義し、重みをパラメタとすることも一案である。

しかし、本稿では以下に述べる理由によりパラメタのチューニングが難しいと判断したために素朴な統合方式を採用した。

情報検索タスクにおける要約では外的な評価(特定タスクの遂行精度により行なう要約手法の評価)によっているために、評価結果を再利用して、このタスクに適合したより良いパラメタ値を求めるといった方向に発展させることが難しい。例えば、節4で述べる評価に用いた TSC Task B では、生成された要約文章に対して被験者による適合性評価が行なわれるものの、適合性判定を適切に行なうことができる要約文章を人間に生成してもらうという段階はないために、いわゆる「正解要約」がない。また、仮にそのような「正解要約」を作成したとしても、適合性判定を正しく行なうことができる要約には様々な亜種が考えられるので、「正解要約」と同じ要約を作成することがこのタスクの本質的であるかという疑問が残る。もちろん、評価・パラメタ再設定・再評価といったチューニングのためのループに被験者の評価を組み入れることも不可能ではないが、膨大な人的資源を必要とするために、我々はこの手法を採用しなかった。

3 重要文抽出に基づく要約文書生成

語の重みは要約生成における基本要素であるから、ほとんどの要約手法に我々の手法を組み込むことができると考えられる。しかし、我々の目的は、前節で述べた語の重みづけが検索文書要約において有効であることを示すことである。そこで、次に示す、語の重要度だけによる最も基本的な要約手法を以降の評価実験で用いる。

- (1) 文書 D 中の文 s の重要度は次式の通り、文中のキーワードの重みの和を文の長さで正規化したものとする。

$$s_imp(s, D) = \frac{\sum_{w \in keyw(s)} weight(w, D)}{|s|} \quad (5)$$

$keyw(s)$: 文 s 中のキーワードのリスト
 $|s|$: 文 s の総形態素数

- (2) ある決められた要約の長さに達するまで、原文書から重要度の高い順に文を取り出ししていく。
- (3) 取り出した文を原文書における出現順に並べ変えて要約文書を得る。

4 評価

本節では、我々の要約方式について、二通りの視点から評価を行なう。

まずは、検索タスクの精度・効率の良さと言う観点から、評価型情報検索ワークショップである NTCIR2(NTCIR Project 2000) における TSC(Text Summarization Challenge) における「課題 B IR タスク用要約」(以下、TSC Task B と呼ぶ) に基づいて評価を行なう (TSC 実行委員会 2001; 難波, 奥村 2001)。

つぎに、幾つかの例により、我々の方式が各語に与える重要度を、検索質問を考慮しない語の重みづけ手法 (TF 値, TF · IDF 値) と比較することにより、我々の重要度計算手法の特徴を定性的に評価する。

なお、我々の評価実験においては、要約生成の手順に以下の条件を加えた。

- 記事の表題 (見出し文: Headline) と本文を区別せずに要約の入力とする。
- 名詞をキーワードとする。文書からの名詞抽出には形態素解析器 JUMAN3.61(黒橋, 長尾 1998) を用いた²。
- IDF 値等の計算においては、当初 TSC 実行委員会から発表された使用新聞記事データ

² 品詞細分類において、普通名詞, サ変名詞, 固有名詞, 地名, 人名, 組織名, 数詞, 名詞接尾辞, 未定義語-その他, 未定義語-カタカナ, 未定義語-アルファベットをもつ形態素を取り出した。また、これらが連続した場合には、その形態素列を複合語として認定し、単語としても取り出した。クラスタリングにおいて、複合語については、複合語自身とその構成素となった形態素の両者が文書ベクトルの成分として考慮されている。一方、各文の重要度計算においては、複合語を構成せず、各形態素の重みのみを利用している。

である毎日新聞 1994 年, 1995 年, 1998 年に加えて, 手元にあった 1997 年を全文書集合とした。

- 最大距離アルゴリズムにおけるパラメタ α は 0.5 とする。
- 要約を一覧形式で提示することを想定すると, 要約文書の長さが統一されているほうが, 見やすい。そのため, 要約文書の長さは要約率ではなく, 絶対的な長さにより決定する。具体的には, 150 形態素をしきい値とする。
- 文書の総形態素数が 150 より短い場合には要約をせずに原文書を提示する。
- 要約生成に当たって, 原文書の文が省略されている箇所には「…」を加え, 原文書の段落箇所には改行を加える。

4.1 情報検索タスクにおける要約品質の評価実験の概要

図 5 に情報検索タスクにおける要約品質の評価実験の概要を示す。TSC Task B においては, TSC 実行委員会より配布されたデータセットに, 12 のトピックがあり, それぞれ, 検索要求 1, 検索文書 50 文書から構成されている。検索文書は 1998 年の毎日新聞の記事集合から検索されたものである。

TSC への参加者は各自のシステムを用いて, これらの文書を要約し, 実行委員会に提出する。提出された要約文書に対して, TSC 実行委員会による被験者を用いた評価が行なわれた。被験者は学生 36 名で, 各検索要求につき, 3 人が割り当てられている。被験者は各要約を読むことによって, それが検索要求に適合しているか否かの判断を行う。その判断と, あらかじめ原文書に対して付与されている関連性評価³(以下, 関連度ともいう)を比較することにより, 要約の品質が評価される。すなわち, 両者が一致する度合いが高いシステムほど有効な要約を生成すると考えることができる。原文書に付与されている関連性評価は A,B,C の三段階である。ここで, A はその文書が検索要求に適合すること, B は関係のある文書であること, C は関係のない文書であることを表す。これに対して, 被験者らには関連性の有無 (YES/NO) という二段階で提示してもらう。よって, 両者の一致の判定においては, A 判定の文書だけを関連文書とする場合 (Answer Level A) と, A 判定に加えて B 判定の文書も関連文書とする場合 (Answer Level B) が考えられる。

4.2 総合評価

表 2 に個別の評価尺度についての結果を他の参加システム (8 システム) ならびに TSC 実行委員会が準備したベースラインシステム (3 システム) と比較して示す。ベースラインシステムは, 「全文提示 (Fulltext)」(要約率 100%), 「質問バイアス付き TF に基づく方式 (TF with QB)」(要約率 20%), 「文書の先頭を採るリード方式 (Lead)」(要約率 20%) である。質問バイアス付き

3 当然, 参加者には, 当初, 非公開である。

図 5 情報検索タスクにおける要約品質の評価

TF 法は、TF を語の重みとして重要文抽出を行なうものであるが、この時に、検索質問に現れる語について 2 倍の重みを与えている。これらのシステムの概略については、付録 A を参照されたい。

評価尺度には以下のものを用いた。いずれの値も、全てのトピックを通じて集計したものである。

- 被験者が 1 検索要求に関するタスク (50 文書) に要した時間 (TIME)
- タスクをどの程度適切に行なえたかを示す指標。すなわち、再現率 (Recall)、適合率 (Precision)、F 値 (F-measure)⁴。
- 要約文書の長さ (1 文書あたりの平均文字数, LENGTH)

4.3 適合性判断のための所要時間とその精度に関する評価

情報検索の結果の文書に対する要約においては、利用者が行なう適合性判断のための時間の短さと、適合性判断の正確さが共に達成されることが必要である。一方で、両者はトレードオフの関係にある。例えば、長い要約文書を提示すれば、タスク遂行の時間が長くなるが、一方で、精度は概ね向上すると考えられる。よって、両者を同時に評価する尺度が必要とされるが、まだ良いものが提案されていない。そこで、我々のシステムを含む各システムの再現率、適合

⁴ 再現率 (Recall) = $\frac{\text{被験者が正しく適合と判断した文書数}}{\text{実際の適合文書の総数}}$,
 適合率 (Precision) = $\frac{\text{被験者が正しく適合と判断した文書数}}{\text{被験者が適合と判断した文書の総数}}$,
 F 値 (F-measure) = $\frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$.

表 2 総合評価一覧

	Proposed	Sys 1	Sys 2	Sys 3	Sys 4	Sys 6	Sys 7	Sys 8	Sys 9	TF		
										Fulltext with QB Lead		
Recall												
(Ans. A)	0.907	0.833	0.899	0.793	0.818	0.858	0.831	0.824	0.849	0.843	0.798	0.740
Precision												
(Ans. A)	0.751	0.728	0.717	0.685	0.674	0.718	0.739	0.738	0.741	0.711	0.724	0.766
F-Measure												
(Ans. A)	0.808	0.761	0.785	0.715	0.718	0.763	0.766	0.749	0.768	0.751	0.738	0.731
Recall												
(Ans. B)	0.754	0.741	0.793	0.715	0.737	0.745	0.719	0.719	0.752	0.736	0.700	0.625
Precision												
(Ans. B)	0.897	0.921	0.904	0.898	0.875	0.892	0.908	0.913	0.923	0.888	0.913	0.921
F-Measure												
(Ans. B)	0.797	0.808	0.828	0.776	0.773	0.785	0.779	0.775	0.805	0.773	0.776	0.712
TIME	8:33	9:41	12:48	6:25	6:44	9:01	10:16	9:16	9:31	13:46	8:44	7:32
LENGTH	234.4	297.8	585.7	89.5	136.4	288.4	292.9	266.1	262.5	819.4	253.6	174.5

Sys 1~9: TSC 参加の他システム (付録 A 参照)

Ans. A, Ans. B: Answer Level A, Answer Level B にそれぞれ対応

Fulltext: 原文書

TF with QB: ベースラインその 1. 質問バイアス付き TF による重要文抽出手法.

検索要求中の単語に 2 倍の重み, 要約率 20%(文ベース)

Lead: ベースラインその 2. 先頭から 20%の文を抽出する手法. タイトルは出力しない.

率, F 値について, 適合性判断のための所要時間との関係をプロットした. Answer Level A,B の場合を図 6 にそれぞれ示す.

また, 判定時間と要約文書の平均文字数の間の関係を図 7 に示す.

4.4 トピック毎の適合性判断の精度に関する評価

これまで示した結果では, 全てのトピックに互る平均値を用いてタスクの遂行精度を議論してきたが, 当然, トピックによって各システムの精度が異なるはずである. そこで, 我々の手法と各ベースライン手法による要約において, トピック毎のタスク遂行精度をプロットした. Answer Level A,B の場合をそれぞれ図 8 に示す. なお, 図中 'Ave.' は全トピックに互る平均値である.

4.5 適合と判断した被験者の数による要約の質の定量的評価

要約文書の質を今少し詳細に検討するために, 各々の文書に対して, 関連性有り (YES) と答えた被験者の人数を調べる. この人数はトピックに対する生成された要約文書の関連度の高さを表す尺度と考えられる. そこで, 原文書を関連性判定 (A,B,C) によって分類し, その要約に対して YES と判定した人数毎に文書類度を集計し, プロットした. 結果を図 9 に示す.

再現率

適合率

F 値

(A) Answer Level A

再現率

適合率

F 値

(B) Answer Level B

Proposed:我々の手法, Fulltext:原文書, TF with QB:質問バイアス付き
TF による重要文抽出手法, Lead: 先頭から 20%の文を抽出する手法,
無印: 他の参加システム

図 6 判定時間と再現率, 適合率, F 値の関係

図 7 判定時間と要約文書の平均長 (文字単位) の関係

再現率

適合率

F 値

(A) Answer Level A

再現率

適合率

F 値

(B) Answer Level B

図 8 トピック毎の再現率, 適合率, F 値

図 9 原文書の関連度と要約文書に対する適合性判断の関係

4.6 語の重み付けに対する評価

語の重みづけについては正解というものが無いので、定量的にその評価をすることが難しい。ここでは、最も基本的な重みづけである、TF 値、TF・IDF 値による重みと、本手法の重みづけを実例により比較し、定性的に我々の語の重要度決定手法の特徴を検討する。

ここでは、典型的な語の重みがどの様になっているかを調べるのが主眼であるので、我々のシステムで最も F 値の高かったトピック 1027 に注目した。このトピックの内容を表 3 に示す。このトピックについて、原文書の関連度がそれぞれ、A, B, C であるものを一つずつ選択し、

表 3 TSC Task B のトピック 1027

DESCRIPTION フィールド:	ハイビジョンテレビ
NARRATIVE フィールド:	ハイビジョンテレビ (高精細度テレビ、HDTV) に関する政治、企業、ユーザーなどからの情報を含む記事。ハイビジョンの標準化、実験、放送やハイビジョンの売れ行き、国内外の動向、各種論議も含む。

各種の語の重み付けを行なった結果について、上位 10 位までを求めた。表 4,5,6 にその結果を示す。表において IGRsum は式 (3) に示される情報利得比の和であり、TFIDFIGRsum は TF・IDF 値に IGRsum を乗じた値である。

5 考察

5.1 タスク遂行の精度

TSC Task B は、利用者 (被験者) が要約文書を読むことにより、原文書のトピックに対する関連度を推定するタスクであった。よって、この場合の要約は報知的 (informative) である必要はなく、指示的 (indicative) でありさえすれば良い。

表 4 文書に対する語の重みの例 (Topic 1027, 記事番号 980822075, 関連度 A)

順位	TF		TF・IDF		IGRsum		TFIDFIGRsum	
1	放送	8	放送	1.14×10^{-4}	放送	9.52×10^{-3}	放送	1.09×10^{-6}
2	申請	6	B S	9.68×10^{-5}	デジタル	3.10×10^{-3}	B S	2.54×10^{-7}
3	番組	5	申請	9.41×10^{-5}	B S	2.63×10^{-3}	番組	1.87×10^{-7}
4	B S	4	S D T V	8.23×10^{-5}	番組	2.38×10^{-3}	デジタル	1.22×10^{-7}
5	衛星	3	番組	7.86×10^{-5}	テレビ	2.20×10^{-3}	H D T V	1.10×10^{-7}
6	1 番組	2	1 番組	7.66×10^{-5}	ハイビジョン	2.13×10^{-3}	申請	8.37×10^{-8}
7	H D T V	2	H D T V	6.44×10^{-5}	H D T V	1.70×10^{-3}	衛星	7.95×10^{-8}
8	S D T V	2	衛星	5.42×10^{-5}	郵政省	1.66×10^{-3}	郵政省	6.20×10^{-8}
9	テレビ	2	各 3 番組	4.34×10^{-5}	衛星	1.47×10^{-3}	テレビ	5.55×10^{-8}
10	デジタル	2	計 2 2 番組	4.34×10^{-5}	デジタル放送	1.37×10^{-3}	ハイビジョン	5.44×10^{-8}

表 5 文書に対する語の重みの例 (Topic 1027, 記事番号 981215216, 関連度 B)

順位	TF		TF・IDF		IGRsum		TFIDFIGRsum	
1	関西	8	関西	1.08×10^{-4}	放送	9.52×10^{-3}	放送	8.14×10^{-7}
2	放送	6	放送	8.55×10^{-5}	企業	3.15×10^{-3}	B S	1.27×10^{-7}
3	情報	4	C S	7.17×10^{-5}	情報	2.95×10^{-3}	C S	1.20×10^{-7}
4	C S	3	番組制作会社	6.59×10^{-5}	B S	2.63×10^{-3}	関西	1.20×10^{-7}
5	衛星	3	発信	5.77×10^{-5}	番組	2.38×10^{-3}	情報	1.19×10^{-7}
6	発信	3	衛星	5.42×10^{-5}	チャンネル	1.72×10^{-3}	衛星	7.95×10^{-8}
7	1 程度	2	1 程度	5.33×10^{-5}	C S	1.68×10^{-3}	番組	7.48×10^{-8}
8	B S	2	B S	4.84×10^{-5}	衛星	1.47×10^{-3}	企業	7.19×10^{-8}
9	会社	2	関西電力	4.34×10^{-5}	通信	1.40×10^{-3}	発信	3.83×10^{-8}
10	関西電力	2	関西チャンネル	4.34×10^{-5}	関西	1.11×10^{-3}	チャンネル	3.69×10^{-8}

表 6 文書に対する語の重みの例 (Topic 1027, 記事番号 981108028, 関連度 C)

順位	TF		TF・IDF		IGRsum		TFIDFIGRsum	
1	放送	7	自由ヨーロッパ	1.30×10^{-4}	放送	9.52×10^{-3}	放送	9.50×10^{-7}
2	チェコ	6	チェコ	1.26×10^{-4}	企業	3.15×10^{-3}	番組	1.12×10^{-7}
3	自由	5	放送	9.98×10^{-5}	情報	2.95×10^{-3}	チェコ	1.09×10^{-7}
4	イラク	4	ブラハ	7.95×10^{-5}	番組	2.38×10^{-3}	自由ヨーロッパ	7.03×10^{-8}
5	イラン	4	イラン	7.64×10^{-5}	局	1.21×10^{-3}	自由	5.01×10^{-8}
6	ヨーロッパ	4	ヨーロッパ	7.32×10^{-5}	開始	9.89×10^{-4}	ヨーロッパ	4.24×10^{-8}
7	政府	4	自由	7.32×10^{-5}	チェコ	8.70×10^{-4}	イラク	4.23×10^{-8}
8	ブラハ	3	イラク	7.29×10^{-5}	政府	8.05×10^{-4}	イラン	4.22×10^{-8}
9	自由ヨーロッパ	3	チェコ政府	7.10×10^{-5}	米国	7.58×10^{-4}	ブラハ	3.62×10^{-8}
10	番組	3	召	5.29×10^{-5}	自由	6.84×10^{-4}	企業	3.59×10^{-8}

このような要約においては、原文書に関する細かいニュアンスが伝達されるエレガントな要約文書が生成される必要はなく、適切なキーワードの選択とそのキーワードがどのような文脈で現れているかを説明する文書部分の抽出が適切にできればよいと考えられる。この考え方に従えば、キーワードの抽出の精度が高ければ、我々が用いた程度の文抽出による要約機構でもタスクを十分遂行できる要約文書が得られるはずである。

以下では Answer Level A,B に分けて、上記の観点からタスク遂行の精度を考察する。

5.1.1 Answer Level A

本節では検索質問に対して適合文書 (A 判定) のみを正解とした評価 (Answer Level A) について考察を行なう。表 2 によると、我々の手法は、再現率、適合率、F 値すべてにおいて、他のすべての参加システムよりも高い値を示している。

ベースラインシステムとの比較においては、我々のシステムの適合率は Lead 手法よりも 1.5 ポイント低い値を示しているものの、それ以外は勝っている。Lead 手法は、再現率が他の手法に比べて一番低いため総合指標である F 値においてはさほど高くなく、我々のシステムよりも 7.7 ポイント低い。すなわち、Lead 手法は適合率重視の手法と見なすことができる。

質問バイアス付き TF 法と比較してみると、再現率において 10.9 ポイント、適合率において 2.7 ポイント、F 値において 7.0 ポイント勝っている。このことは、検索文書の要約においては、必ずしも検索要求を直接使用しなくても、検索文書群だけで同等以上の質を持つ要約が可能であることを示している。

次に、適合性判断の所要時間とタスク遂行の精度について考える。まず、表 2 によると、所要時間単独についていえば、我々のシステムが生成した要約に対し、被験者が適合性判定に要した時間は、1 トピック (50 文書) あたり 8 分 33 秒であった。これは、TSC に参加した 9 システム中、3 番めに短いものであった。また、すべての参加システムの平均タスク時間は 1 トピック当たり 9 分 8 秒であり、我々の要約の適合性判定に要する時間はこれよりも短い。そして、所要時間と各種評価値の関係を示す図 6(A) においては、概ね、左上に位置するシステムの性能が良いと考えられるので、我々のシステムは他のシステムに勝っていると言えよう。特に、再現率は適合率に比べてシステム間での格差が大きく、我々のシステムの再現率の高さが見てとれる。

図 8(A) によれば、他のベースラインシステムはトピックによって、再現率が大きく変動しているが、我々のシステムはトピックにあまり依存せず安定して高い再現率を示している。これは、被験者が、関連文書 (関連度 A) の要約に対して適合すると概ね正しく判断したことを示す。一方、適合率についていえば、他のベースラインシステムとほぼ同様の傾向で、我々のシステムもトピックによってその精度が大きく変動している。これは、トピックによっては、我々のシステムが、非関連文書 (関連度 B,C) に対して一見すると関連性があると見誤る文を抽出し、要約の一部として提示していることを表す。以上をまとめると、まず、我々の手法は、トピックに関連する文を原文書から積極的に抽出していることがわかる。一方、そのような文の周囲の文脈については、抽出を促す戦略を採っていないことから、それらの抽出洩れにより、非関連文書に対して関連性があると見誤るような要約を生成する可能性もあることがわかる。

なお、図 7 をみると、要約文書長と適合性判定の所要時間は一定時間のオフセット (5 分 52 秒) がついているものの、システムの違いによらず、ほぼ、比例関係となっている。一方、各シ

システムの出力する要約文書については、適合性判定の精度にばらつきが見られる。よって、要約文書の適合性判定時間は適合性判定の結果によらず、要約文書の長さのみに依存していると考えられる。

5.1.2 Answer Level B

本節では B 判定まで含めた評価 (Answer Level B) について考察を行なう。他のシステムと比較して、再現率が第 2 位と高いものの、適合率は第 7 位、F 値は第 4 位と相対順位が低くなった。Answer Level B の評価においては、Answer Level A よりも正解の数が多くなるので、一般に、Answer Level A に比べて、再現率が下降し、適合率が高くなる。再現率についていえば、Answer Level A において、高い適合率となったシステムほど減少が激しくなる。一方、適合率については、B 判定のものが Answer Level A での誤判定となっているのであれば、その値の上昇が著しい。

我々のシステムの場合、再現率が 0.907 から 0.754 へと激しく低下しており、図 8(B) に示される通り、トピック毎の変動が大きくなっている。しかし、その順位について言えば、2 位であるので相対的には他のシステムよりも高いことがわかる。つまり、関連文書 (評価 A,B) に対して正しく関連性の判定が行なわれた要約文書の数他システムよりも多い。

一方で、適合率の上昇は他のシステムより低いので、被験者が関連度評価 C の文書の要約に対しても適合であると判定を下した数が多かったことになる。これは、Answer Level A の考察で述べたことを裏付けており、トピックに関連する文の抽出は成功しているものの、その文脈が脱落する場合も少なからずあることを示している。ただし、図 6(B) や図 8(B) が示すとおり、適合率のシステム間の差異は再現率ほど小さくなく、また、Answer Level A の適合率に比べてもシステム間の格差が小さくなっている。

5.1.3 提案手法とベースラインとの差異に関する検定

前節までに述べたタスクの遂行精度において、提案手法と他の手法の間に有意な差があるか否かを検証するためには、統計検定を行なう必要がある。

しかし、TSC 実行委員会が提供する結果情報において、トピック毎の個別の評価が得られるのは自システムならびにベースライン三種のみだけである。そこで、ここでは、提案手法がベースライン三種との間に有意な差があるかを検証する。

総合性能を表す F 値についてトピック毎の値の差に基づき Wilcoxon の符号順位検定を行なった。提案手法と各ベースラインを比較した時に「F 値に差が無い」という帰無仮説に対する有意確率 p を表 7 に示す。

この表によると、Answer Level A においては、提案手法がベースライン TF with QB ならびに Lead に対して、有意水準 5% の下で、差を持つことが示されている。Fulltext との比較に

表 7 本手法とベースラインとの間の差異に関する Wilcoxon 符号順位検定の結果

	対 TF with QB	対 Lead	対 Fulltext
Answer Level A	p = 0.0161 < 0.05	p = 0.0400 < 0.05	<i>p = 0.0522 > 0.05</i>
Answer Level B	<i>p = 0.4238 > 0.05</i>	p = 0.0332 < 0.05	<i>p = 0.9097 > 0.05</i>

太字は有意水準 5%の下で差があることが示されたもの。

ついては、有意水準 5%の下での差異を示すことができなかったが、帰無仮説を採択する確率が 5.2%程度で同有意水準との差は僅差である。

一方、Answer Level B においては、ベースラインとの差が Answer Level A ほど顕著ではない。有意水準 5%の下で差がある事が示されたのは Lead 手法との対比だけであった。

5.2 適合と判断した被験者の数による要約の質の定量的評価

図 9(A) についてみると、トピックに対する原文書の関連度が A の場合には、いずれのシステムにおいても、当然ながら、3 人が一致して YES とする場合が最も多く、2 人、1 人、0 人と順に頻度が低くなっていく。我々の手法についてみると、他のベースライン手法に比べて、3 人が一致して YES と答えた頻度が高くなっており、一方、1 人、2 人が YES と答えた頻度が低くなっている。すなわち、関連性のある文書については、他の手法より質の高い要約が生成されていたと考えられる。また、Lead 手法については、3 人が一致して NO とつけた件数が他のベースラインよりも多くなっている。つまり、関連判定に重要な部分は必ずしも文頭にあるわけではなく、一般的な新聞記事の要約で良い戦略の一つとされる Lead 手法が情報検索タスクにおいては、必ずしも有効ではないことがわかる。

一方、関連度 B が与えられている原文書についていえば、本来は NO であるべきであるから、YES とした人数が 0 の場合の頻度が大きく、3 人の場合の頻度が少なくなる傾向にあるはずである。ただし、関連性が一部認められる記事の要約に対する評価であるから、要約の仕方によっては 1~2 人程度の人が YES と判定する頻度も (C) の場合に比べて多くなることが予想される。図 9(B) をみると、TF 手法を除いて、この傾向がみられ、特に Lead 手法において顕著である。TF 手法においては、0 人の頻度が少なく、被験者 3 人のうち 1 人以上が関連性があると判定していることがわかる。

そして、図 9(C) についてみると、概ね、0 人、1 人、2 人、3 人と順に頻度が低くなっていく。誤った判定の箇所、すなわち 1 人、2 人、3 人の箇所をみると、ベースライン手法よりも我々の手法のほうが上方にグラフが描かれている。すなわち、他のベースライン手法よりも関連度 C の文書に対して誤って関連性があると判断した被験者が多かったことを示している。これは、節 5.1.1 ならびに節 5.1.2 で考察したことを再確認するものである。

5.3 語の重み付けに関する考察

我々の提案する語の重みづけについてその特徴を実例(検索トピック 1027,「ハイビジョンテレビ」)により考察する. 表4において, 我々が最終的に用いる語の重みである TFIDFIGRsum の列に示されるように, 関連度 A の文書においては, 検索トピックに陽に示されている語はもちろんのこと,「BS」,「番組」,「デジタル」,「申請」,「衛星」,「郵政省」など, そのトピックに関連する語も上位に重み付けられていることが分かる. その一方で, TF・IDF 値では上位にあった非関連語「SDTV」(標準テレビ)については, 上位 10 位から姿を消している. これらの効果は, TF・IDF の列と IGRsum の列を比較すれば分かるが, IGRsum の成分が主に寄与している. そして, このような重みづけは質問バイアス付き TF 手法では, 質問拡張 (Baeza-Yates and Ribeiro-Neto 1999; Salton and Buckley 1990) などを別途行なわない限り実現できないものである.

関連度 B や C の文書についても, 表 5,6 に示されるとおり, トピックに関連の深い語が上位に重み付けられていることが見てとれる. また, 関連度 C の例では, IGRsum の値が低いものも上位に見られるが, いずれもトピックとは関係の薄い語である.

5.4 語の重みづけの品質と検索文書の数ならびに品質に関する考察

本手法による重みづけは, クラスタリングの対象となる文書の件数ならびに検索結果の質と密接な関係にある.

本来ならば, TSC Task B と同一条件の下で, 要約の対象となる文書数ならびに検索結果の質を変化させての追加検証が必要であるが, TSC Task B と同一被験者による再実験が困難であるため, ここでは, 定性的な考察を行なう. 定量的な考察は今後の課題としたい.

さて, 本手法で語の重みに用いている情報利得比の和は, クラスタ構造に則した分布をしている語に高い重みを与えるものであるから, 検索文書をクラスタリングした結果, どのようなクラスタ構造が形成されるかによって, 各語の重みが決まる.

クラスタ構造は対象となる文書間の類似性により求めるので, 検索文書のうち, クラスタリングの対象となる文書群 (以下, 単に検索文書群と呼ぶ) について, その数と文書間の類似度に密接な関係がある. さらに, 本手法では, 図 2 に示すように文書データベース中の全文書が所属するクラスタを最上位に考え, 検索文書群とそれ以外の文書群の間の差異についても重みに反映しているので, 全体のクラスタ構造は情報検索の精度とも関係がある.

そこで, 以下では, 上記二点について個別に考察を行なう.

5.4.1 語の重みづけの品質と検索文書数に関する考察

採用する文書が少ない場合と多い場合について考察する．まず，採用する文書数が少ない場合について，検索文書群と残りの文書群との対比から得られる重み，ならびに，検索文書群をクラスタリングした結果から得られる重みがどのようなになるかを考える．最終的な語の重みはこれらの和である．

文書数が少ない場合には，文書集合中の総単語数ならびに個々の単語の頻度も小さくなるので，平滑化 (smoothing) をせずに単語頻度から語の出現確率を直接推定すると，語の頻度の小さな差異が語の出現確率の大きな変化となることがある．この時，残りの文書群における単語の出現確率との差が大きくなれば，その語の重みが不当に高くなる可能性がある．本稿では用いてはいないが，単語出現確率の推定においては平滑化を考慮すべきであろう．

検索文書のクラスタリングについては，クラスタ中の文書数ならびにクラスタ構造の枝分かれが少なくなるので，各文書間の類似性関係がクラスタ構造に大きな影響を与える．そして，情報利得比の計算においては，語の重要度が個別文書における語の現れ方に敏感になる．

一方，採用する文書数が大きい場合には，個々の単語の出現頻度が相対的に大きくなることと，個々の文書間の類似度がクラスタ構造に与える効果が分散・平滑化されることにより，上記と逆の傾向があると考えられる．

ただし，検索質問との関連性が低い文書が数多くなってくると，検索質問に述べられたトピックとは関連性の低い語がクラスタ形成の際に支配的になる事もあり得る．この場合は，本来のトピックに関連する語には相対的に低い重みしか与えられない．これは，検索文章群において，より重要な項目を強調するという点では正しい重みの付き方ではあるが，本来のトピックとの関連性判定という観点からは，誤らせる方向にバイアスがかかってしまう．このような場合には Tombros ら (Tombros and Sanderson 1998) の提案するような検索質問によるバイアス方式の方が適切な結果を与えるので，両者を併用する方法も検討すべきであろう．

5.4.2 語の重みづけの品質と情報検索の品質に関する考察

検索結果の質も，検索文書群と残りの文書群との対比から得られる語の重み，ならびに，検索文書群をクラスタリングした結果から得られる語の重みの両者に影響を与える．

まず，検索文書群と残りの文書群との対比から得られる語の重みについて考える．この段階で検索質問に関連する語に比重がおかれた適切な重みづけがなされるためには，残りの文書集合と比較して検索文書集合側に関連文書が多く存在し，それに伴い関連する語の出現確率が偏る必要がある．情報検索の質が非常に悪く，文書集合からほぼランダムに検索文書が取り出されるのであれば，検索文書集合とそれ以外の文書集合の間で各語の出現確率に見られる差異は小さく，対比によって得られる語の重みはいずれも小さな値に留まる．一方，検索エンジンが，検索質問に従って，ある程度分布の偏った文書集合を返すとすれば，その度合に応じて，検索

質問に関連する語の重みも高くなる。

次に検索文書群をクラスタリングした結果から得られる語の重みについて考える。節 5.4.1 でも述べた通り、検索質問との関連性が低い文書が多くなると、検索質問に述べられたトピックとは関連性の低い語が、クラスタ形成の際に支配的になる事も有り得る。この場合は、本来のトピックに関連する語とは別の語に高い重みが与えられる。この時には、検索質問によるバイアス方式の方が性能が良いと考えられる。

前項とともに以上をまとめると、我々の重み付け方式が十分な効果を発揮するためには、

条件 1 情報検索エンジンの精度が悪くない事

条件 2 検索質問に関連する文書が検索結果中にある程度存在する事

の二点を満足することが必要であると考えられる。また、これらの条件を満足しない場合、特に、条件 2 を満たさない場合にも対応できるためには、検索質問によるバイアス方式との併用を検討することも重要であろう。本節の定量的な評価とともに今後の課題としたい。

6 関連研究

節 2.1 でも述べたように、検索結果の文書要約は、通常文書要約とは次の点で異なる。

- (1) 検索要求文が与えられている。
- (2) 複数の文書が同時に得られている。そして、ある一度の検索の結果という点において文書間に類似性が認められる。

本研究においては (2) の情報を用いたが、(1) の情報を利用した手法もある。

Tombros ら (Tombros and Sanderson 1998) は、文書中のタイトル情報、文書中での位置情報、文書中での単語の出現頻度に基づいた、従来通りの文の重要度に、検索要求文中の単語が文中に出現する頻度に応じたスコアを加味することで、検索要求文に依存した重要文抽出を実現している。塩見ら (塩見, 徳田, 青山, 柿ヶ原 1998) も、文書中の単語の出現頻度に基づいた文の重要度に、検索要求中の単語が文中に出現する頻度に応じたスコアを加味する手法を提案している。しかし、これらの手法は、スコアの制御が難しいことが問題点である。また、節 2.1 で述べた通り、各種フィードバック、検索要求中の単語のシソーラスによる拡張などといった情報検索システムにおける工夫が反映されないという問題点がある。

また、(2) の情報を利用するという点では、Eguchi ら (K.Eguchi, H.Ito, A.Kumamoto and Y.Kanata 1999), Fukuhara ら (T.Fukuhara, H.Takeda and T.Nishida 1999), Radev ら (Radev, Jing and Budzikowska 2000; Radev et al. 2000) の手法に関連する。Eguchi らは、適合性フィードバックに基づく検索システムを構築している。このシステムでは、検索結果を文書間の類似度に基づいてクラスタリングし、各クラスタごとにクラスタに多く含まれる語と、そのクラスタを代表する文書のタイトルを、そのクラスタの要約として出力する。ユーザに、出力されたクラスタの中から選択してもらい、そのクラスタに含まれる文書を用いて適合性フィー

ドバックを行なっている。Fukuhara らや Radev らも、Eguchi らと同様に検索結果を文書間の類似度に基づいてクラスタリングし、各クラスタごとに要約を出力している。Fukuhara らの手法では、まず、文書中の単語の出現頻度を考慮し、クラスタごとのトピックを表す語を抽出する。そして、それらトピックを含む文を抽出し、焦点-主題連鎖を考慮して並べ替え、各クラスタの要約としている。Radev らの手法では、各クラスタについて、その重心ベクトルを TF・IDF 値を用いて計算し、その重心における各語の成分を語の重みの主要な成分としている。

これらの手法では、クラスタリングを文書のグループ分けのみに利用しており、グループ化された後では、各クラスタにおける語の分布だけを用いて重みを計算している。また、語の重要度としては単純にクラスタ内の TF や IDF を用いているだけである。直接の比較は今後の課題とするが、我々の手法においては、文書間の類似性構造の情報も取り入れて重みづけしているので、より高い分解能ならびに精度が得られていることが期待される。

TSC Task B における評価では、検索文書集合が与えられてはいるものの、最終的には個別文書の要約になっていた。一方、(Mani and Bloedorn 1999; McKeown and Radev 1999) に代表されるように、複数文書から一つの要約文書を生成するという研究も近年注目を集めている。特に Carbonell ら (Carbonell and Goldstein 1998) は、極大限界適合度 (Maximal Marginal Relevance, MMR) という概念を導入し、検索質問と検索文書の類似度ならびに、ある文書とそれよりも上位の文書との間の冗長性に基づいて、検索文書の再順位づけを行なうとともに、これを、パッセージ検索に利用することによって要約生成を行なう手法を提案している。

我々の重み付け手法は複数文書を一つの要約にする場合にでも、効果を発揮することが期待されるが、文書間の融合過程においては、冗長性の制御を陽に行なう MMR のような手法との組み合わせも必要になってくるであろう。

7 おわりに

本稿では、複数の検索文書の間に存在する類似性の構造を階層的クラスタリングにより抽出し、その構造を適切に説明する可否かに応じて情報利得比に基づき語に重みをつける手法を提案した。TSC での実験の結果、この方法に基づく重要文抽出型の要約手法は、検索文書の要約において、非常に有効であることが示された。

今後の課題としては、節 5.4 に述べたように本手法と検索文書の数ならびに品質の間の定量的な関係を明らかにすることが挙げられる。また、情報利得比に基づく語の重みを、対話型の情報検索インタフェース中で利用することを検討することも課題である。本稿では、クラスタ構造の全部分を均一に語の重みに反映させて要約を作成した。一方で、対話型のインタフェースとしては、利用者が部分クラスタを選択しながら、目的の情報に辿りつくというものも考えられる。この場合、提示された箇所のクラスタ構造のみを考慮して、要約を生成することができると考えられる。さらに、複数文書の要約において我々の枠組がどの様に役立つのかも検討

したい。

謝辞

本研究を進めるにあたり、本学大学院生であった菊池美和さん(現在, (株)NTT データ) ならびに吉田和史さん(現在, 三菱電機(株)) に多大なる御協力を頂きました。ここに感謝いたします。また、国立情報学研究所主催の NTCIR ならびに TSC1 を企画・運営し、評価用データを作成していただいた皆様に感謝致します。なお、本研究を遂行するにあたって、CD-毎日新聞 94 年版, 95 年版, 97 年版, 98 年版を利用させていただきました。使用許諾をしていただいた毎日新聞社, ならびに、同データの研究利用に対して御尽力いただいた皆様に感謝致します。最後になりましたが、数多くの有益なコメントを頂いた査読者の方に感謝いたします。

本研究の一部は文部科学省科学研究費特定領域研究「IT の深化の基盤を拓く情報学研究」(課題番号 13224041, 14019041) により支援を受けております。

参考文献

- Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. ACM press, Addison Wesley longman Ltd.
- Carbonell, J. and Goldstein, J. (1998). “The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries.” In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336.
- Frakes, W. B. and Baeza-Yates, R. (Eds.) (1992). *Information Retrieval – Data Structure & Algorithms*. Prentice Hall PTR.
- Hirao, T., Hatayama, M., Yamada, S., and Takeuchi, K. (2001). “Text Summarization based on Hanning Window and dependency structure analysis.” In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, pp. 349–354.
- K.Eguchi, H.Ito, A.Kumamoto, and Y.Kanata (1999). “Adaptive Query Expansion Based on Clustering Search Results.” *Transaction of Information Processing Society of Japanese*, **40** (5), pp. 2439–2449.
- 黒橋禎夫, 長尾真 (1998). 日本語形態素解析システム JUMAN version 3.6 使用説明書.
- Luhn, H. (1958). “The automatic creation of literature abstracts..” *IBM Journal of Research and Development*, **2** (2), pp. 159–165.
- Mani, I. and Bloedorn, E. (1999). “Summarizing Similarities and Differences among Related Documents.” *Information Retrieval*, **1** (1-2), pp. 35–67.

- McKeown, K. R. and Radev, D. R. (1999). "Generating Summaries of Multiple News Articles." In Mani, I. and Maybury, M. T. (Eds.), *Advances in Automatic Text Summarization*, chap. 24, pp. 381–389. The MIT Press.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- 長尾真 (1983). パターン情報処理. 電子通信学会大学シリーズ. コロナ社.
- Nakao, Y. (2001). "How small a distinction among summaries can the evaluation method identify?." In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, pp. 341–348.
- 難波英嗣, 奥村学 (2001). "第2回 NTCIR ワークショップ自動要約タスク (TSC) の結果および評価法の分析." 自然言語処理研究会 NL-144-20, 情報処理学会.
- Nobata, C., Sekine, S., Murata, M., Uchimoto, K., Utiyama, M., and Isahara, H. (2001). "Sentence Extraction System Assembling Multiple Evidence." In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, pp. 319–324.
- NTCIR Project (2000). "NTCIR (NII-NACSIS Test Collection for IR Systems) Project Web page." <http://research.nii.ac.jp/ntcadm/index-en.html>.
- Oka, M. and Ueda, Y. (2001). "Phrase-representation Summarization Method and Its Evaluation." In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, pp. 361–368.
- 奥村学, 難波英嗣 (1999). "テキスト自動要約に関する研究動向." 自然言語処理, 6 (6), pp. 1–26.
- 奥村学, 難波英嗣 (2000). "テキスト自動要約に関する最近の話題." Technical Memorandum IS-TM-2000, 北陸先端科学技術大学院大学情報学研究科.
- Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers.
- Radev, D. R., Jing, H., and Budzikowska, M. (2000). "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies." In *Proceedings of ANLP/NAACL Workshop on Summarization*.
- Salton, G. and Buckley, C. (1988). "Term-Weighting Approaches In Automatic Text Retrieval." *Information Processing and Management*, 24, pp. 513–523.
- Salton, G. and Buckley, C. (1990). "Improving Retrieval Performance by Relevance Feedback." *Journal of the American Society for Information Science*, 41, pp. 288–297.
- T.Fukuhara, H.Takeda, and T.Nishida (1999). "Multiple-text Summarization for Collective Knowledge Formation." In *Proceedings of Workshop on Social Aspects of Knowledge and Memory*. IEEE Systems, Man and Cybernetics Conference.
- Tombros, A. and Sanderson, M. (1998). "Advantages of Query Biased Summaries in Informa-

tion Retrieval.” In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 2–10.

塩見隆一, 徳田克巳, 青山昇一, 柿ヶ原康二 (1998). “視点を考慮した文書要約手法の提案.” 情報処理学会第 56 回全国大会講演論文集, 3 巻, pp. 104–105.

TSC 実行委員会 (2001). “NTCIR-2 自動要約タスク (automatic text summarization task)/TSC(Text Summarization Challenge).” <http://lr-www.pi.titech.ac.jp/tsc/tsc1.html>.

付録

A NTCIR2 TSC Task B に参加した他システムの概要

NTCIR2 TSC Task B に参加した他システムについて, NTCIR2 Workshop 論文集に基づきその概略を述べる (Nobata, Sekine, Murata, Uchimoto, Utiyama and Isahara 2001; Nakao 2001; Hirao, Hatayama, Yamada and Takeuchi 2001; Oka and Ueda 2001). なお, システム Sys 7 については, その詳細は不明である.

A.1 ベースライン TF with QB

TSC 実行委員会が提供するベースラインの一つで, 文抽出型でスコアの上位の文から指定の要約率になるまで抽出. 要約率は 20%(文を単位とする). 各文のスコアは, 内容語 (名詞, 動詞, 形容詞, 未定義語) の TF の和であるが, 検索トピック中の語にはバイアスが与えられ, 2 倍の重みとする. この「2 倍」という定数がどのように決定されたか, 例えば, 何らかのチューニングが施されているかなどは不明である.

A.2 ベースライン Lead

TSC 実行委員会が提供するベースラインの一つで, 文抽出型で本文の先頭から指定の要約率になるまで抽出. 要約率は 20%(文を単位とする).

A.3 ベースライン Fulltext

TSC 実行委員会が提供するベースラインの一つで, 原文書のうち, 表題 (見出し文) ならびに本文をそのまま返すもの. 要約率は 100%(文を単位とする).

A.4 システム Sys 1, Sys 2

文抽出型でスコアの上位の文から指定の要約率になるまで抽出. 要約率は 10%(Sys 1) ならびに 50%(Sys 2). 文のスコアは, 次の 3 つの値を重み付きで加算したものである. (1) 文書中

の文の位置に基づくスコアで、文書の先頭ならびに文末に高い重みを付与する。(2) 文長に基づくスコアで長いものに高い重みをつける。(3) 名詞の TF・IDF 値の和をスコアとしたもの。ただし、見出しに含まれる名詞ならびに固有表現 (Named Entity) については、その TF・IDF 値を加算し、トピック中の DESCRIPTION および NARRATIVE フィールドの名詞に対しては、さらに、TF・IDF 値を 2 倍にする。

A.5 システム Sys 3, Sys 4

語と語の間の重要な関係を見つけ出し、それを元に要約を句レベルで生成する。その関係の重みは各語の重みの和に関係の重みを乗じたものである。各語の重みは TF・IDF 値で求めるが、トピック中の語については高い重みを与える (詳細は不明)。語の間の関係については、格による依存関係には高い値を、等位接続などには低い値を与える。要約文書の長さは文字数で与えられ、100 文字以内 (Sys 3) もしくは 150 文字以内 (Sys 4) である。

A.6 システム Sys 6

文抽出型で、ハニング関数を用いた窓により各パッセージの重要度を計算し、その 中から重要文を抽出する。パッセージの重要度計算においてはトピック内の語のみに注目する。また、文書の先頭部分は無条件に加える。要約率は明示されていないが、35%程度である。

A.7 システム Sys 8, Sys 9

文抽出型で、与えられたキーワードを網羅する文を優先しつつ、全てのキーワードが要約文書に出現するまで文を抽出する。キーワードは、文書タイトル、文書の先頭段落 (Sys 8 のみ)、話題構造の境界にある文中の語 (Sys 9 のみ)、トピック中の DESCRIPTION ならびに NARRATIVE から、それぞれ、名詞、動詞、形容詞 を取り出したものである。

略歴

森 辰則: 1986 年横浜国立大学工学部情報工学科卒業。1991 年同大学大学院工学研究科博士課程後期修了。工学博士。同年、同大学工学部助手着任。同講師を経て現在、同大学大学院環境情報研究院助教授。この間、1998 年 2 月より 11 月まで Stanford 大学客員研究員。自然言語処理、情報検索、情報抽出などの研究に従事。言語処理学会、情報処理学会、人工知能学会、日本ソフトウェア科学会、日本認知科学会、ACM、AAAI 各会員。

(2001 年 9 月 28 日 受付)

(2001 年 12 月 20 日 再受付)

(2002 年 4 月 5 日 採録)