情 報 学 基 礎 63−15
（２００１．７．２５）

情報検索タスクに基づく自動要約手法の評価

仲尾 由雄
富士通研究所
〒 211-8588 川崎市中原区上小田中 4-1-1
ynakao@jp.FUJITSU.com

本稿では、NTCIR-2 ワークショップの要約評価タスクで得られた結果に基づき、情報検索タスクベースの評価手法の有効性について議論を行う。NTCIR-2 に提出した２種類の要約と、主催者が基準として用意した３種の要約との計 5 種類に対する評価結果を、統計的手法で検定した結果、提出した２種の要約の違いは微妙すぎて評価不能であることがわかった。この結果は、情報検索タスクというタスクが、通常の状況では要約作成にとってやさしすぎることに由来すると見られ、例えば、より短い時間で多量の検索結果を判定するというような形にタスク設定を改める必要があることを示唆している。

Effectiveness of an IR-based Evaluation Method for Text Summarization

Yoshio Nakao
Fujitsu Laboratories Ltd.
4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa, 211-8588 Japan

This paper examines the effectiveness of an evaluation method used for Japanese summary evaluation tasks of the NTCIR-2 workshop. The discussion focuses on the effectiveness of an extrinsic evaluation based on relevance assessment in information retrieval. Statistical analysis of the evaluation results obtained for five types of summaries comprising two types of submitted summaries and three types of official baseline summaries suggests that the difference between the former two was too small for the evaluation to identify. This suggests that the task might better be performed under the condition that only a few subjects could complete the given task.

1 はじめに

本稿では、第2回NTCIR[1] ワークショップにおける要約評価タスクの結果に基づき、情報検索タスクに基づく要約手法の評価の有効性の議論を行う[2] 。

いずれの技術の研究開発においてもそうであるように、要約技術の研究開発においても、技術の評価方法の確立はひとつの大きな課題となる。すなわち、現状の要約技術の有効性を正しく評価する方法があれば、よりよい技術を開発する指針を示すことができ、また、要約技術を組み込んだ実用性の高いシステムを設計するための大きなヒントを得ることができる。この点において、要約技術の評価方法は、現状では未成熟であり、よりよい評価方法を求めて、様々な試み [3, 4, 5, 等] が行われている。

要約技術の評価手法は、要約の品質を直接的に評価する方法 (instrinsic method) と、要約を利用して何らかの課題を被験者に解かせ、要約の性質の違いによる作業効率の向上率などを測定する間接的評価方法 (extrinsic method) に大別される。実用的な要約システムの構築を目的とする場合、システムの用途に対して要約がどれだけ有効に働くかという観点からの評価が最も重要と考えられ、従って、後者の評価手法の確立が強く望まれる。

現在試みられている間接的評価方法に、文書検索における適合判定作業の支援効果に基づくものがある。これは、文書検索結果を要約して提示し、被験者に個々の文書の適合度を判定してもらい、判定精度および判定時間で要約の有効性を評価しようというものである。本稿の目的は、このような評価方法によって、どの位細かい要約の違いまで評価可能なのかを調べ、よりよい評価方法とするために、何をなすべきかを明らかにすることにある。

情報検索タスクに基づく要約手法の評価に関し、Tombros ら [3] は、利用者の情報要求を考慮して作成した要約を提示した方が、情報検索システムで一般的に用いられてきた文書の先頭部を要約として提示するより効果的であることを、関連度判定実験により示している。すなわち、情報要求 (検索キーワード) を考慮した要約を使った場合の方が、判定精度が高く、判定時間が短く、原文参照回数も少なく、アンケート調査においても役に立つという回答が多かったと報告している。

Tombros らの研究は、綿密な計画に基づいて客観的な評価を行ったという点、特に、先行的な評価研究 [6] に比較すると、検索キーワードを考慮して要約するという観点を取り上げて比較を試みている点と、記事全文の参照回数の減少という文書選別過程における具体的操作と結びついた評価結果が示されている点は注目される。

しかしながら、Tombros らが比較している要約手法には、検索キーワードの出現箇所を考慮するか以外の違いが見られ、評価結果は必ずしも質問文を考慮することの効果だけを表しているわけではない。すなわち、評価対象とした要約手法では、抜粋すべき文の重要度の評価において、検索キーワードの出現に加え、記事見出し・記事内の小見出しに含まれる語や高頻度語の出現や文の出現位置も考慮しているため、比較手法に比べ適切に記事の主題を示唆する内容を要約に含めることができ、それが評価結果に表れた可能性も高い。

また、SUMMAC プロジェクト [4] の情報検索タスクに基づく評価 (ad-hoc task) では、参加システムによる有効性違いはほとんど見られなかったという報告もある。

そこで、本研究では、要約手法の話題検出機能に関する要約手法の違いが、情報検索タスクに基づく評価手法でどの程度区別できるのかを見るという趣旨に基づき、NTCIR-2 の要約評価タスクに 2 種類の異なる手法により作成した要約を提出し、その結果を分析することとした。今回の目的は、話題階層に基づく話題検出機能 [7] の効果を確かめることにあり、そのため、比較対象として、重要文抜粋型の 2 種類の手法 (文の変形等の操作は全く行わない手法) を用意した。

以下、2 章で、2 種類の要約手法の概要を紹介してから、3 章で、各タスクで用いた要約手法と評価結果について報告し、4 章で、情報検索タスクに基づく要約手法の評価の有効性について考察する。

2 比較対象要約手法

2.1 キーワードベースの要約手法 (基準要約)

キーワードベースの要約手法 [8](以下図表では “Baseline” と表記) は、与えられたキーワード (以下「要約の核」と称す) を全て含み, かつ, なるべく少ない文数からなる要約を作成するという趣旨で考案した手法である。この手法は、与えられたキーワードを幅広く含む文を 1 つ選択・抽出し、選択した文に含まれるキーワードをキーワードリストから取り除くという操作を、キーワードリストが空になるまで繰り返すというものである。直観的にいえば、全てのキーワードが 1 回ずつ現れるような要約を作成する手法である[3] 。

今回のタスク A(指定サイズの記事要約の品質を人間の要約との比較による直接的評価タスク) するにおいては、記事見出しとリードパラグラフ[4] る内容語[5] を要約の核とした。そして、指定サイズの要約を作成するため

[1] NII-NACSIS Test Collection for IR Systems [1].
[2] 本稿は、[2] の内容を整理・縮小し、考察を補ったものである。
[3] 見出しに含まれる名詞をキーワードとして与えた場合、新聞や雑誌の記事に対して原文の約 1〜3 割程度の量に相当する 1〜3 文程度の要約を作成できることが確認されている [9]。
[4] フォーマルランの場合。ドライランでは記事見出しのみを手がかりとした。
[5] ドライランにおいては名詞、フォーマルランにおいては、名詞・動詞・形容詞。

に、文選択処理は、選択・抽出した文の合計サイズが、指定サイズを超える直前まで、要約の核とするキーワードリストを再構成しながら繰り返した。すなわち、一端全てのキーワードを含む文の集合が抽出された時点で、まだ指定サイズに余裕がある場合には、その時点までに抽出された文に含まれる内容語をキーワードリストに追加してから、文抽出処理を繰り返した。タスクB(関連度判定に基づく間接的評価タスク)においては、質問文から抽出した内容語もキーワードリストに加え、全てのキーワードが1回以上現れる要約を作成した(詳細は、タスクBの評価の節参照)。

2.2 話題階層ベースの要約手法

話題階層ベースの要約手法 [7](以下図表では "TH-based" と表記)では、まず、TextTilingアルゴリズム [10] をベースにした手順によって、文書の話題階層を認定する。ここで、話題階層とは、大きさの異なる複数の話題区画が2段以上の階層構造を成していることを意味する。話題区画とは、文書中である粒度の話題に関して記述している一続きの部分のことである。本稿では、このような話題区画の集合で、階層構造をなしているものを、話題階層と称する。

次に、各話題区画の開始位置付近から、2〜3文ずつ、話題区画の内容を端的に示していそうな文(以降「境界文」)を抽出する。境界文は、典型的には、(小)見出しと話題を導入する役割を持つ文の組である[6]。

本手法の特徴は、要約として出力すべきサイズに応じて、適切な粒度の話題を検出し、要約に取り込めること、また、文抽出において話題の立ち上がり位置に相当する狭い文書中の部分から2文以上まとめて抽出していることにある。例えば、10文からなる要約を作成する場合には、文書全体を5個程度に分割するほぼ同じ大きさの話題区画に分割し、各区画の開始位置付近から2〜3文ずつ文を抽出することになる。

今回のタスクAにおいては、話題階層のルートノードに相当する話題区画(文書全体に対応)から、最下層に位置する最小話題区画(大きさは40語程度)まで、要約のサイズに関する制約が許す範囲で、順々に境界文を抽出した。すなわち、まず、文書の冒頭付近から文書全体に対する境界文を抽出し、次にその直下の層に属する話題区画の境界文を抽出する、というように、階層の高い順[7](階層が同じ場合は文書における出現順)に、境界文を抽出した[8]。

タスクBにおいては、上記手法で認定した境界文から要約に含めるべきキーワード抽出し、実際の文抽出は、キーワードベースの要約手法により行った(タスクBの評価の節参照)。

3 評価結果

NTCIR-2の要約評価タスクは、要約そのものの品質の(intrinsicな)評価を行うサブタスク(タスクA1、A2)と、情報検索タスクに基づく外的(extrinsic)評価を行うサブタスク(タスクB)が実施された [11]。いずれの評価も、テストデータは、新聞記事(事件記事/特集記事および社説)である。

タスクA1、A2は、人間の作成した要約との比較による評価である。具体的には、30の記事と数種類の要約サイズに対して作成した要約を、人間が抽出した文との一致率(タスクA1)、人間の作成した2種類の要約および1種類の公式ベースライン要約との比較による相対順位(タスクA2)による評価が行われた。

タスクBは、SUMMACプロジェクト [4] のad-hoc taskにおける評価と類似の手法による評価が行われた。具体的には、まず、参加要約システムは、検索トピック(質問データ: ドライランでは10トピック、フォーマルランでは12トピック)と検索結果に相当する記事群(ドライランでは30記事、フォーマルランでは50記事)に対して要約を作成して提出する。次に要約を使って、3人ずつの被験者がトピックに対する記事の関連度判定を行う。最後に、参加要約システムを、それを使った判定作業を行った被験者に判定精度および判定時間によって評価される。という手順で評価が行われた。

3.1 タスクAの結果

表1、2は、それぞれ、タスクA1とA2の評価結果である。これらの表には、主催者が用意し2種類のベースライン要約手法に関する結果も含めて示してある。TF-basedとは、高頻度語を多く含む順[9]に文を抜粋する手法(以下「高頻度語ベースの要約」)であり、Lead-basedとは、記事の先頭から指定サイズ分だけ文を抜粋する手法(以下「記事冒頭部抽出要約」)である。

表1は、人間による文抜粋結果との一致率を、要約手法毎に示したものである。ドライランにおいては、話題階層ベースの要約は基準要約より高い値を示していた。

[6] 比較的大きな文書を使った実験 [7] では、各話題区画から抽出した最初の境界文の約半分は、文書中の(小)見出しに対応していた。
[7] 結果として通常は区画サイズの大きい順になる。
[8] 出力サイズを微調整する目的で、実際には、境界文抽出は指定サイズの8割までとし、残りの部分は、記事見出しと境界文候補中の内容語を手がかりに、てキーワードベースの要約手法(文抽出処理)で補う形をとった。
[9] 正確には、文中の語のTFの和の順

表 1: タスク A1 の結果

Summary type	F-score	
	Formal run	Dry run
TH-based	.416	.540
Baseline	.449	.536
TF-based	.391	.525
Lead-based	.434	.554

表 2: タスク A2 の結果

Summary type	Impression score				Cosine	
	20R	20C	40R	40C	free	ext.
Formal run						
TH-based	3.00	3.17	2.73	3.03	.526	.561
Baseline	2.97	3.10	3.10	3.13	.522	.552
TF-based	3.20	3.27	2.77	3.07	.516	.549
Lead-based	–	–	–	–	.481	.513
Dry run						
TH-based	2.63	3.00	2.63	2.97	–	.586
Baseline	3.23	3.60	2.73	3.03	–	.546
TF-based	3.37	3.70	3.27	3.40	–	.569
Lead-based	–	–	–	–	–	.582

しかし、話題階層ベースの要約には、記事の先頭付近の文を含むという性質があり、それがこの結果をもたらしている可能性がある。そこで、フォーマルランでは、基準要約にもこれと類似の性質をもたせるため、リードパラグラフから抽出した内容語も手がかりに加えて、基準要約を作成した。その結果、フォーマルランでは、基準要約が話題階層ベースの要約より高い評価値を示すようになった。

表 2 の、*cosine* 欄は、人間による文選択 (ext 欄)/自由要約結果 (free 欄) と、それぞれの自動要約結果との内容の類似性を tf・idf 法によるコサイン値により示している。この値に関しては、話題階層ベースの要約の方が、基準要約よりよい値を示す傾向がみられた。ただし、フォーマルランにおける 20%要約率においては、基準要約の方がよい値を示していた。

これらの結果は、話題階層ベースの要約は、主要な話題 (端的には重要語) を適切に取り入れることができているが、人間とは異なる文の選択をしていることを示唆している。話題階層ベースの文選択アルゴリズムが、記事中の (小) 見出しを選択しやすいという性質に由来すると考えられる。

表 2 の、*impression-score* 欄は、要約品質の相対的主観評価の結果を示している。これは、それぞれの自動要約結果を、2 種類の人間による要約結果 (自由要約および重要文抽出) および公式ベースライン要約 (Lead-based) と比較し、好ましいと思える順に与えた順位点 (1〜4) の平均値である。*20R* 欄は、要約率 20%と要約の読み易さに関する評価値を、*20C* 欄は、要約率 20%の要約の内容の適切さに関する評価値を示している。*40R* と *40C* は、要約率 40%の要約に関するそれらの値である。これらの値も、コサイン値と同様、少なくとも 40%の長めの要約に関しては、話題階層ベースの要約の方が、基準要約より適切に主要な話題を抽出していることを示している。また、読み易さに関しても、話題階層ベースの要約は比較的良好な評価値を得ているが、これは、話題の開始付近と推定された比較的狭い範囲から集中的に文を抜粋し、また、話題毎に区切り (空行) をいれて出力するという戦略が、有効に機能したことを示唆している。ただし、話題階層ベースの要約の最良の評価値でも 2.73(*40R*)[10] にすぎず、人間の作成した 2 種の要約を上回る品質の要約は作成できなかったことになる。

3.2 タスク B の結果

タスク B に提出した話題階層ベースの要約 (TH-based)・基準要約 (Baseline) は、どちらもキーワードベースの文抽出処理を基本に、検索用質問も考慮して作成した。両者の違いは、要約の核とするキーワードの選択にある。基準要約は、記事見出し・リードパラグラフおよび検索用質問データ中の description と narrative のフィールドに含まれる内容語を核として作成した。一方、話題階層ペースの要約は、記事見出しと検索用質問データ中の description フィールド、および、話題階層に基づく文選択処理によって認定された、最も大きな話題の区切りに関する境界文に含まれる内容語を核として作成した。端的にいうと、話題階層ベースの要約は、リードパラグラフの代わりに境界文を用い、また、質問内容の詳細説明に相当する narrative フィールドを用いなかった点が違いである。この設定にあたっては、要約サイズ[11] がほぼ同じになることを考慮した (表 4)。この表には、主催者側で用意した 3 つのベースライン要約のサイズも含めている。

表 3 は、タスク B における評価結果の概要を示している。*level A* 欄は、それぞれの要約を使って被験者が判定した結果を、A 判定 (記事の主題が質問と一致するか) を基準に評価した再現率 (*rec.*) と適合率 (*prec.*)、F-score(F: $\frac{2*recall*precision}{recall+precision}$) の平均値を列挙している。*level B* 欄は、同様に B 判定 (部分的にでも質問文と関連しているか) を基準に評価した結果である。*time* 欄は、各被験者が判定に要した時間の平均値を示している。1 つの質問

[10] 内訳は、1 位 6 回、2 位 2 回、3 位 16 回、4 位 6 回。

[11] 要約のサイズは、特に制約を設けず、キーワードベースの文抽出処理によって、要約の核となる単語を含む文が一通り抽出した時点の抽出結果を要約としている。

表 3: 情報検索タスクに基づく評価結果の概要

(a) Formal run

Summary type	A 判定			B 判定			平均判定時間	
	F	rec.	prec.	F	rec.	prec	質問当たり	記事当たり
TH-based	.768	.849	.741	.805	.752	.923	9'31"	11.4"
Baseline	.749	.824	.738	.775	.719	.913	9'16"	11.1"
Full-text	.751	.843	.711	.773	.736	.888	13'46"	16.2"
TF-based	.738	.798	.724	.776	.700	.913	8'44"	10.5"
Lead-based	.731	.740	.766	.712	.625	.921	7'32"	9.0"

(b) Dry run

Summary type	A 判定			B 判定			平均判定時間	
	F	rec.	prec.	F	rec.	prec	質問当たり	記事当たり
TH-based	.838	.915	.796	.857	.869	.864	5'40"	11.3"
Baseline	.822	.840	.838	.814	.786	.891	6'01"	12.0"
Full-text	.842	.913	.796	.867	.878	.874	8'46"	17.5"
TF-based	.794	.804	.827	.802	.757	.895	5'12"	10.4"
Lead-based	.773	.781	.813	.781	.744	.883	4'25"	8.8"

表 4: 要約のサイズ

Summary type	Avg. length in characters (condensing rate)			
	Formal run		Dry run	
TH-based	263	(35%)	178	(38%)
Baseline	266	(35%)	153	(33%)
Full-text	819	(108%†)	463	(100%)
TF-based	254	(33%)	–	
Lead-based	175	(22%)	–	

† フォーマルランにおける全文 (Full-text) には記事見出しを含む。

に対する一連の記事 (検索結果に相当: フォーマルランでは 50 記事、ドライランでは 30 記事) を評価するのに要した時間の平均値である。

4 タスク B の結果の分析と考察

4.1 統計分析

本節では、タスク B の結果について、記事の冒頭部を抽出するベースライン要約 (記事冒頭部抽出要約: Lead-based) を基準に吟味し、情報検索タスクに基づく要約評価の有効性について考察する。

表 5 は、繰り返し[12] のある二元配置分散分析法 (two-factor factorial ANOVA) による F-score の分析結果で示したものである[13] 。表は、要約の種類に関する要因と、質問に関する要因との間の交互作用を示している。下線を付与した数値 (パーセントポイント) は、有意に大きな交互作用が観察された部分である。有意に大きな交互作用は、要約の種類による F-score の大小関係 (の程度) が、質問毎にはっきりと異なっていることを示すので、有意に大きな交互作用が観察された要約の間には、大きな性質の違いがあると推定される。すなわち、記事冒頭部抽出要約との性質の差異は、話題階層ベース・高頻度語ベースの要約の方が、基準要約・全文より大きいことになる。今回の場合、基準要約も、記事の冒頭部の内容を中心に抽出したものであることを考慮すると、話題階層ベース・高頻度語ベースの要約は、記事の冒頭部には見られない情報を含んでいることが、上記の差異につながっていると考えられる。

表 7 は、表 3 に概要を示した F-score の違いが有意であるかを、順位和検定 (Wilcoxon test) によって調べた結果である。順位和検定というノンパラメトリック検定手法を用いたのは、有意な交互作用が検出されたこと (表 5)、F-score の分布が正規分布を仮定することができないたことによる。表によれば、判定 A を基準にした F-score の違いで有意なものはなく、判定 B を基準にした F-score では、話題階層ベースの要約・基準要約・全文のいずれかと、記事冒頭部抽出要約との違いが有意である。特に、話題階層ベースの要約と記事冒頭部抽出要約との差異の有意水準は 1%以下のはっきりとした違いである。

以上の結果は、話題階層ベースの要約は、記事冒頭部抽出要約・基準要約より多くの話題を含んでおり、かつ、それらの話題は関連度判定に有効であったことを示して

[12] 本稿では、同じ質問-記事要約の対を評価した 3 人の被験者の評価値 (F-score) を、繰り返しデータと見なして分析した。

[13] Hull[12] は、複数の要約に関する評価結果を一度に比較できる手法として ANOVA を紹介しているが、後述の交互作用によって、比較結果の有意性の判定ができなくなるため、今回は、要約手法の対毎に、分析を行った

いる。

順位和検定結果に比べると、ANOVA による要約の種類に関する主効果の検定結果 (表 6) では、高頻度語ベースの要約と記事冒頭部抽出要約との F-score の差が有意性が高くなっている。この検定結果の違いは、高頻度語ベースの要約と記事冒頭部抽出要約の F-score は、大きく違う振る舞いをしているが、中央値にはほとんど差が無いことを示している。このことは、高頻度語ベースの要約は、記事冒頭部抽出要約とは、かなり異なった情報を含んでいるが、その情報は必ずしも関連度判定に有効ではないことを示していると考えられる。

よって、話題階層ベースの要約と基準要約は、高頻度語ベースの要約より正確に主要な話題を抽出できていると判断される。判定 A を基準にした場合の F-score において、両要約が、高頻度語ベースよりよい値を示していることもこの判断の妥当性を示唆している。

しかしながら、話題階層ベースの要約と基準要約との差異については、判定 B の F-score に関する記事冒頭部抽出要約との比較における有意水準の違い (表 7) からすると、話題階層ベースの要約の方が関連度判定には有効そうではあるものの、統計的に検定した範囲では、その差は有意でないという結果となった[14] 。

表 5: 要約種-トピック間の交互作用

Summary type	F-value (percentage points)			
	Lead-based		Full-text	
TH-based	2.1	$(p < .05)$	2.2	$(p < .05)$
Baseline	1.4	$(p > .05)$	.59	$(p > .05)$
Full-text	.57	$(p > .05)$	–	
TF-based	2.9	$(p < .05)$	1.6	$(p < .05)$

表 6: 要約の種類に関する主効果

Summary type	F-value (percentage points)			
	Lead-based		Full-text	
TH-based	17	$(\mathbf{p < .01})$	1.6	$(p > .05)$
Baseline	7.1	$(p < .05)$	.004	$(p > .05)$
Full-text	5.5	$(p < .05)$	–	
TF-based	7.9	$(\mathbf{p < .01})$	.009	$(p > .05)$

表 7: 順位和検定の結果

Summary type	Z-value (percentage points)			
	Lead-based		Full-text	
TH-based	2.8	$(\mathbf{p < .01})$	1.1	$(p > .05)$
Baseline	1.9	$(p < .05)$	.005	$(p > .05)$
Full-text	1.8	$(p < .05)$	–	
TF-based	1.5	$(p > .05)$	.38	$(p > .05)$

4.2 被験者間の判定の一致状況と判定時間

今回の評価タスク B では、各々の質問-記事 (要約) に対して、3 人ずつの被験者が関連度判定をしているので、(1) 全ての被験者が一致して同じ判定をする、(2)2 対 1 に判定が分かれる、という 2 つのケースが考えられる。表 8 は、全員の判定が一致した記事数の割合を、それぞれの要約の種類毎に示したものである。表 9 は、全員一致判定に関する判定精度 (B 判定基準の F-score) を示したものである。

表 8 によれば、全員一致判定の割合の全体平均 (Overall) は、5 種類の要約とも 7 割程度と、TREC-5[13] で報告された一致率[15] とほぼ等しい。また、表 9 に見られるように、全員一致判定の精度は、.9 程度以上の高い値を示しているものが多いので、全般的に見れば、いずれの要約も関連度判定に十分に役に立っているといえる。

しかしながら、T1014(不良債権の処理)・T1015(携帯電話、簡易型携帯電話のサービス)・T1026(年金改革の財源) の 3 つのトピックについては、一致率が低い。すなわち、表 8 に見られるように、質問毎の平均で 5 割を下回る一致率を記録したのはこれら 3 つのトピックであった。以降、判定時間と判定精度の関係を、判定の分かれた質問-記事対に関するデータを基に議論するが、それらは、主にこれらのトピックに関わるものである。

表 10 は、判定の分かれた質問-記事対に関し、正しく判定した場合と間違った判定をした場合との判定時間の差を、次のような手順で集計したものである。まず、それぞれの記事に対して、正しい判定をした場合・間違った判定をした場合の判定時間の平均をそれぞれ求め、それら 2 つの平均値の差を計算した。次に、このように求めた判定平均時間の差を、質問に対する記事の 2 つの判定カテゴリ (A 判定の記事と C 判定の記事[16]) 毎に集計し、平均値と標準偏差 (表 10 の括弧内の数値) を求めた。また、Wilcoxon 符号付き順位和検定により、それぞれの判定カテゴリ内の判定時間の差を分析し、正/負の偏りが有意 (両側 5%水準) であった場合に、下線を施した。

表より、関連記事に対応するカテゴリ (TP-FN) では、誤った判定の方が、正しい判定より判定時間が長くかかっている傾向が見て取れる。特に、話題階層ベースの要約においてはその傾向は有意である。対象記事は同一であるので、この傾向をもたらした要因は、被験者の違い、例えば、与えられたトピックに関する知識レベルや読解

[14] 表に挙げた要約の種類の対以外の対の検定結果で、F-score の差が有意 (5%水準) であるという結果が得られたのは、話題階層ベースの要約と高頻度語ベースの要約を順位和検定した場合だけだった (ANOVA ではいずれの対の差も有意でないという結果が得られた)。

[15] TREC-4 の正解文書のプーリング作業における関連度判定の整合性 (consistency) を評価した実験における一致率 (71.7%= 3 人の被験者全員の一致率に相当)。

[16] B 判定 (部分的に関連) の記事は、関連/非関連のいずれと判定されても、ある意味で正解であるので、はずして集計した。

表 8: 全員一致判定の比率の詳細

Topic	TH	B	Full	TF	Lead	Avg.
T1009	.68	.88	.60	.86	.64	.73
T1012	.86	.76	.76	.70	.88	.79
T1014	**.46**	.58	.50	**.48**	**.42**	**.49**
T1015	.80	.56	.78	.58	**.46**	.64
T1021	.94	.82	.92	.84	.84	.87
T1022	.90	.82	.78	.82	.78	.82
T1025	.82	.82	.76	.86	.94	.84
T1026	**.28**	.56	**.46**	**.36**	.62	**.46**
T1027	.82	.84	.82	.76	.84	.82
T1034	.76	.78	.78	.70	.86	.78
T1035	.84	.70	.80	.80	.88	.80
T1036	.76	.62	.74	.82	.70	.73
Overall	.74	.73	.73	.72	.74	.73

表 9: 全員一致判定の精度

Topic	TH	B	Full	TF	Lead	Avg.
T1009	.96	**.32**	**.46**	.68	**.15**	.51
T1012	1	1	1	1	1	1
T1014	.94	.73	**.50**	.93	**.18**	.66
T1015	.95	.93	.79	.93	1	.92
T1021	.95	1	.96	1	.80	.94
T1022	.96	1	1	.81	.95	.94
T1025	.88	.96	.96	.73	.75	.86
T1026	1	.94	.97	1	.84	.95
T1027	.78	**.57**	.78	.64	**.64**	.68
T1034	.97	.97	1	1	.88	.96
T1035	.98	.92	.93	.96	.86	.93
T1036	**.35**	.86	.69	.52	**.40**	.57
Overall	.91	.86	.87	.83	.78	.85

下線は、再現率 5 割以下だった部分

表 10: 正しい判定と誤り判定に関する判定時間の差

Summary type	Avg. time difference (σ) [sec]			
	TP - FN (A)		TN - FP (C)	
TH-based	**-9.92**	(15.44)	-0.5	(18.29)
Baseline	-2.67	(14.88)	-0.35	(12.41)
Full-text	-4.87	(19.56)	-3.38	(29.19)
TF-based	-1.66	(9.12)	1.06	(10.12)
Lead-based	-1.05	(8.46)	**1.47**	(5.45)

のスキル等の違いであると考えられる。

非関連記事に対応するカテゴリ (TN-FP) では、要約の種類毎に違った傾向が現れており、要約の品質の違いを示している可能性がある。話題階層ベースの要約・基準要約・全文の場合には、関連記事同様、誤った判定の方が判定時間が長くかかっている傾向がある。ただし、その傾向は有意でなく、平均時間差も、関連記事の場合より小さい。高頻度語ベースの要約・記事冒頭部抽出要約の場合には、逆に、正しい判定に関する判定時間の方が長くなる傾向がある。特に、記事冒頭部抽出要約では、その傾向は有意である。この原因は不明であり、また、平均時間差も僅かであるが、両方の要約とも固定の要約率 (文数で 20%) で作成されたことに由来している可能性がある。例えば、要約が短すぎるため、要約にたまたま現れた質問に関係しそうなキーワードにより、一見関連記事と思えてしまうものが多くあった可能性がある。そして、比較的慎重な (かつ読解のスキルが高い) 被験者が、要約に含まれている範囲の文脈をよく分析して要約に含まれていない記事内容を推測した結果、記事内容は質問と関連しないと正しく判定できたのかもしれない。

4.3 考察

上記の統計分析結果によれば、話題階層ベースの要約手法は、基準要約手法よりややよい結果を残したとはいえるが、その差は有意ではない。5 種類の要約に関する統計分析において有意差が確認できたのは、1 組 (話題階層ベースの要約-高頻度語ベースの要約) の例外を除けば、質問を考慮しないで作成した記事冒頭部抽出要約と比較した場合に限られる。よって、情報検索タスクにおいて、どのような要約が適するのかという問いに対しても、Tombros ら [3] で既に確かめられていたこと以外には、ほとんど情報を得られなかったことになる。

要約の優劣をはっきりと判定できなかった理由の一つとして、今回の関連度判定タスクの設定が、自動要約技術にとってやさし過ぎたことが考えられる。すなわち、いずれの要約を使っても、8 割以上の精度 (A 判定の再現率と B 判定の適合率[17]) が出ているため、精度だけで判断する限り、被験者の個人差等による違い (実験誤差に相当) を上回る差異を観察することは困難になっている。

よって、話題階層に基づく話題検出機能の効果を確かめるという筆者の目的に限れば、より多くの話題を含む長く複雑な文書を要約対象とすれば、よりはっきりとした評価結果が得られた可能性がある。ただし、長い文書と質問文との関連度判定においては、質問文に関連する内容を抽出するだけでは十分でなく、抽出内容が文書全体においてどれくらいの重要度をもつかを表現することの重要性も増すと考えられるので、それに関する評価が得られるよう、例えば、関連度の高そうな順に文書を並べるというタスクとし、要約を使った場合と、全文を使った場合の比較するなどという形に、タスクを改める必要があるかもしれない。

判定時間に関して見ると、今回の設定では、被験者は自由に時間を使える状況になっていたため、数値の意味づけがはっきりしないという問題点がある。すなわち、長い判定時間が必ずしもよい精度に結びつかないという

[17] この 2 つの値は、安定性が高いという意味でとりあげた。A 判定の再現率は、必ず関連と判定すべき記事に対する再現率、B 判定の適合率は正解としてもよい記事に対する適合率という意味を持つ。

結果が得られており、被験者は、要約から瞬時に読み取れる以上に推論しようした形跡が見られる。関連して、フォーマルランにおいて、記事全文を使った関連度判定結果の精度が比較的低い値に止まったことは注目される。すなわち、B 判定を基準にした F-score において、全文を利用した判定より悪い精度だったのは、記事冒頭部抽出要約だけであった。このことは、全文によって関連度を判定した被験者が、記事全文の全てを詳細に吟味して判定したわけではなく、記事見出しやリードパラグラフなどから素早く読み取れる情報を主な手がかりとして、判定したことに由来すると考えられる[18] 。

この点に関する改善策として、短めの判定時間の制約を設定することが考えられる。検索結果を素早く選別するために要約を使うことを考えると、長い時間をかけて分析しなければわからないような情報を出力しても意味がないので、判定時間に短めの制約を加えることは自然であろう。その意味で、まず、利用者が許容できる判定時間の範囲を割り出し、その時間範囲内で、どの位の精度がでるのかを測定することなども考えられる。あるいは、Tombros ら [3] が全文参照回数を測定したように、要約内容が十分でないと利用者が判断したことを、客観的に測定できる手段を用意することも有効であろう。

5 まとめ

本稿では、NTCIR-2 において筆者が提出した 2 種の要約と 3 種の公式ベースライン要約の評価結果の分析を行い、情報検索タスクによる評価手法により、どの位細かい要約の違いまで評価可能なのかに関して議論を行った。その結果、提出した 2 種の要約の違いは微妙すぎて評価不能であることがわかった。また、被験者の意見が分かれた記事に関して、関連度判定に要した時間を分析した結果、長い判定時間が必ずしもよい精度に結びつかないという結果が得られた。

これらの結果は、今回の評価プロジェクトにおける情報検索タスクの設定では、要約作成にとってやさしすぎることに由来すると見られ、例えば、より短い時間で多量の検索結果を判定するというような形にタスク設定を改める必要があることを示唆している。

その他、多くの要約対に関して、質問・被験者の違いに関わる有意な交互作用 (表 5) が観察されているが、今回の実験は、同じ被験者は、ある質問に関して、1 種類の要約のみを使って関連度を判定する形で行われたので、この要因が何であるかを特定できなかった。改善策としては、例えば、同一の被験者が同じ質問に対して 2 種類以上の要約を使って判定するようにすることなどが考えられる。ただし、この場合、評価作業に要するコストが増大するので、効率化するためには、何らかの工夫が必要であろう。

[18] フォーマルランの方がドライランより平均記事長が長かった (約 760 字対約 460 字) にも関わらず、平均判定時間は、フォーマルランの方が約 1 秒短かくなっていることもこの見方を支持している。

参考文献

[1] Kando, N., Koyama, T., Oyama, K., Kageura, K., Yoshioka, M., Nozue, T., Matsumura, A. and Kuriyama, K.: NTCIR: NACSIS Test Collection Project, in *[Poster] IRSG98*the British Computer Society (1998), (http://research.nii.ac.jp/ntcir/).

[2] Nakao, Y.: How small a distinction among summaries can an IR-based evaluation method identify?, in *Proc. of Workshop on Automatic Summarization (WAS2001)*, pp. 69–78Association for Computational Linguistics (2001).

[3] Tombros, A. and Sanderson, M.: Advantages of Query Biased Summaries in Information Retrieval, in *Proc. of SIGIR'98*, pp. 2–10the Association for Computing Machinery (1998).

[4] Mani, I., House, D., Klein, G., Hirschman, L., Obrsi, L., Firmin, T., Chizanowski, M. and Sundheim, B.: The TIPSTER SUMMAC Text Summarization Evaluation (Final Report), Technical Report MTR 98W0000138, MITRE Corporation, Virginia (1998), (http://www.itl.nist.gov/div894/894.02/ related_projects/tipster_summac/final_rpt.html).

[5] Marcu, D.: The Document Undefstanding Conference: A New Forum for Summarization Research and Evaluation, in *Proc. of Workshop on Automatic Summarization (WAS2001)*Association for Computational Linguistics (2001).

[6] 住田一男, 知野哲朗, 小野顕司, 三池誠司：文書構造解析に基づく自動抄録生成と検索提示機能としての評価, *Transactions of the Institute of Electronics, Information and Communication Engineering*, Vol. J78-D-II, No. 3, pp. 511–519 (1995).

[7] Nakao, Y.: An Algorithm for One-page Summarization of a Long Text Based on Thematic Hierarchy Detection, in *Proc. of the 38th Annual Meeting of Association for Computational Linguistics*, pp. 302–309 (2000).

[8] 仲尾由雄：自動要約技術を利用した文書選別支援機能の試作,「知識発見のための自然言語処理」シンポジウム Institute of Electronics, Information and Communication Engineering (1999), (http://www.etl.go.jp/etl/nl/nlsympo/ よりオンライン論文集にアクセス可).

[9] 仲尾由雄：見出しを利用した新聞・レポートからのダイジェスト情報の抽出, 情処研報 NL-117-17, 情報処理学会 (1997).

[10] Hearst, M. A.: Multi-paragraph segmentation of expository text, in *Proc. of the 32nd Annual Meeting of Association for Computational Linguistics*, pp. 9–16 (1994).

[11] Fukushima, T. and Okumura, M.: Text Summarization Challenge: Text summarization evaluation at NTCIR Workshop 2, in *Proc. of the NTCIR-2 Workshop meeting*, pp. 4–9–13National Institute of Informatics, Japan (2001), (http://oku-gw.pi.titech.ac.jp/tsc/ に関連情報).

[12] Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments, in *Proc. of SIGIR'93*, pp. 329–338the Association for Computing Machinery (1993).

[13] Voorhees, E. M. and Harman, D.: Overview of The Fifth Text REtrieval Conference, in *NIST Special Publication 500-238: The Fifth Text REtrieval Conference (TREC-5)*, pp. 1–28National Institute of Standards and Technology (1996), (http://trec.nist.gov/pubs/trec8/t5_proceedings.html).