自然言語処理　144－20
(2001. 7. 17)

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

難波 英嗣† 奥村 学‡
† 日本学術振興会 特別研究員
E-mail: nanba@lr.pi.titech.ac.jp
‡ 東京工業大学 精密工学研究所
〒 226-8503 横浜市緑区長津田 4259
E-mail: oku@pi.titech.ac.jp

本稿では, 2000 年 5 月から 2001 年 3 月まで開催された第 2 回 NTCIR ワークショップサブタスク TSC(Text Summarization Challenge) の結果および評価方法の分析結果について報告する.

キーワード: テキスト自動要約, TSC, NTCIR

Analysis of the Results and Evaluation Methods of Text Summarization Challenge(TSC), A Subtask of NTCIR Workshop 2

Hidetsugu NANBA† Manabu OKUMURA‡
† Research Fellow of the Japan Society for the Promotion of Science
E-mail: nanba@lr.pi.titech.ac.jp
‡* Precision and Intelligence Laboratory, Tokyo Institute of Technology
4259 Nagatsuta, Yokohama, 226-8503, Japan
E-mail: oku@pi.titech.ac.jp

Abstract: In this paper, we report the result of TSC, a subtask of NTCIR Workshop 2, conducted from May of 2000 to March of 2001. We also discuss their evaluation methods.

key words: automatic text summarization, TSC, NTCIR

1 序論

本稿では, 2000 年 5 月から 2001 年 3 月まで開催された第 2 回 NTCIR ワークショップサブタスク TSC(Text Summarization Challenge) の結果および評価方法の分析結果について報告する.

テキスト自動要約は 1950 年代から研究されている研究分野であるが, 1990 年代後半になって, 急速に活発化し, 21 世紀を迎えた今日, 更にその勢いは増しているといえる. 自動要約研究では, システムによって良い要約を作成することを目的としているわけであるが, システムの出力である要約をどのように評価するかという問題に関しては, 分野内においても明確な基準がなく, これまでの長い間, 要約の評価は難しい問題とされてきた. しかし, 自動要約の研究が活発化するに伴い, 評価方法を広く議論し, 基準を明確にしていこうという動きも活発化してきた. 既に行なわれた要約の評価に関する重要な活動として, 1998 年 5 月の DARPA Tipster プロジェクト (Phase III)[8] の一貫として, 要約の評価を行なう会議,「SUMMAC」[4] がアメリカで開催された. 現在は TIDES プロジェクトと名前が変わり, 趣旨も若干変更はされるものの, アメリカにおいては, 今後も要約の評価を行なっていく動きに変化はないものと思われる. 日本でも, これまで多くの優れたテキスト自動要約の研究が行なわれてきたが, 主に個々の研究機関, 企業や大学等が独自の基準で評価を行なってきており, 共通の評価基準や評価方法に関する議論がなされ

てこなかったため, システムを比較することが難しい状況にある. また, 要約研究に共通に使用できる人間の作成した要約などの資源が不足しているという問題もある. そこで, このような問題を認識する研究者, 開発者を募り, 資源の共有や日本語テキストの要約に関する共通の評価方法や評価基準の明確化を本格的に推進させるため, NTCIR Workshop2 のタスクとしてテキスト自動要約が行なわれた.

本稿では, 2001年3月に開催されたNTCIR Workshop2 における成果報告を踏まえ, TSC の結果を分析し, また, 評価方法について検討する.

本稿の構成は以下のとおりである. 次節では, TSC で行なった3種類の課題について説明する. 3節では, 課題に用いたテキスト (新聞記事) について説明する. 4節では, TSC における要約の評価方法について述べる. 5節では, 参加システムの特徴について簡単に述べる. 6節では, 結果を分析し, また評価方法について検討する.

2 課題

TSC では3種類の課題を設定した.

課題 A-1: 重要文抽出型要約

新聞30記事から, 要約率10%, 30%, 50%で重要文を抽出.

課題 A-2: 人間の自由作成要約と比較可能な要約

新聞30記事を対象に, 要約率20%, 40%を越えない文字数で要約を作成. なお, 要約部分が plain text であり, 指定文字数以内に納まっていれば, どのような要約でも構わないため, 課題 A-1 と同じシステムの出力からタグを取り除いて, plain text にすれば, 課題 A-2 にも参加できる.

課題 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. 要約の長さは自由とするが, 要約は plain text で提出. なお, 要約は, 各テキストに対して1つずつ作成し, 複数テキストに対する要約を作成するのではない. また, 検索結果のテキストは検索要求に適合しているものばかりではなく, 適合しないものも含まれている.

3 要約対象テキスト

課題 A-1, A-2:

毎日新聞94年および98年から15記事づつ, 計30記事を抽出. 記事は94年から600, 900, 1200文字以上の3種類の長さの報道記事を, 98年からは1200, 2400文字以上の2種類の長さの社説を抽出.

課題 B:

毎日新聞98年から, 12トピックに関する記事を抽出. 1トピックにつき50記事. また, 各記事はトピックに対する適合性が, A,B,C の3段階で評価されている. このうち A 判定だけを正解とした場合 (Answer Level A) と, B 判定も正解 とした場合 (Answer Level B) の2種類を結果として提示.

4 評価方法

課題 A-1 の評価方法

課題 A-1 の提出結果は, 重要文抽出に基づいて作成された要約であり, 人間が選択した重要文との間の一致度を元に評価を行なう. 評価尺度としては, 以下の3つを用いる.

- 再現率 $= \frac{\text{システムが選んだ文の内で正解の文の数}}{\text{人間が選んだ正解の文の総数}}$
- 精度 $= \frac{\text{システムが選んだ文の内で正解の文の数}}{\text{システムが選んだ文の総数}}$
- F 値 $= \frac{2*\text{再現率}*\text{精度}}{(\text{再現率}+\text{精度})}$

これらの値を要約率ごとに求めた後, 平均したものを最終的な結果とする.

また, ベースラインシステムとして, 以下の2種類を用いる.

- Lead:
  本文の先頭から要約率 (文) として指定された文数だけ出力する.
- TF:
  本文の各文ごとに内容語の TF の和を計算し, このスコアの高い文を要約率 (文) として指定された文数だけ選択する. 選択した文を元の文の出現順に戻して出力する.

課題 A-2 の評価方法

1. 主観評価

まず,

- 人間の作成した重要個所抽出要約 (PART)
- 人間の自由作成要約 (FREE)
- 1 システムが提出した結果 (SYS)
- Leadのベースラインシステムの結果 (BASE)

の 4 種類の要約を用意する. 同時に元テキストも用意しておく.

要約評価者 (1 名) に元テキストと各要約結果を読んでもらい, 次に「テキストとして読みやすいかどうか」の観点と,「元テキストの重要な内容を不足なく記述しているかどうか」の観点の 2 点から要約を評価をしてもらう. 評価は, 読みやすいものから, 1, 2, 3, 4 となり, 同様に内容の点で見て良いものから, 1, 2, 3, 4 となる.

2. content based な評価

人間の作成した要約およびシステムの作成した要約をともに, Jumanで形態素解析し, 名詞, 動詞, 形容詞, 未定義語を抽出する. そして, 人間の作成した正解要約の単語頻度ベクトルとシステムの要約の単語頻度ベクトルの間の距離を計算し, どの程度内容が単語ベースで類似しているかという値を求める [3].

なお, ベクトルの要素は, 各内容語の tf*idf 値とし, df の計算には, 課題と同じ年の毎日新聞 CD-ROM('94 or '98) の全記事を同じく形態素解析した結果を用いる.

なお, タスク A-2 において, 人間の作成する要約は, (1) 人間が自由作成した要約 (2) 人間が重要個所抽出により作成した要約, の 2 種類があり content based な評価は, この両方に対して行なった.

課題 B の評価方法

被験者に, 検索要求とその検索結果としてテキストの要約を提示する. 被験者は各要約を読むことによって, そのテキストが検索要求に合っているかどうか (適合性) の判断を行う.

評価基準として以下の 3 種類を用いる.

- タスクに要した時間

  50 テキストを処理するのにかかった時間

- タスクをどの程度うまく行なえたかを示す指標 (再現率, 精度, F 値)

$$再現率 = \frac{被験者が正しく適合と判断したテキスト数}{実際に適合するテキストの総数}$$

$$精度 = \frac{被験者が正しく適合と判断したテキスト数}{被験者が適合と判断したテキストの総数}$$

$$F\ 値 = \frac{2 * 再現率 * 精度}{(再現率 + 精度)}$$

- 要約の長さ (LENGTH)

  1 テキスト (要約) あたりの平均文字数

5 参加システム

formal run が終了した後, 各参加団体にシステムの特徴についてアンケートで答えていただいた. その中から, システムを比較する上で重要と思われるいくつかの特徴に着目し, 表 1 にまとめた.

6 結果および評価方法の分析

TSC の結果の詳細についてはすでに文献 [1] や文献 [2] で述べられているので, 詳細はそちらを参照されたい. 本節では, 分析結果を中心に報告する.

課題 A-1 の結果分析

一般に, 報道記事と社説はテキストの特性が異なる. そこで, まず, 報道記事と社説それぞれにおける各システムの重要文抽出の F 値を調べた. 結果を表 2 に示す. 表からわかるとおり, 多くのシステムは, 社説よりも報道記事の F 値の方が 5%から 10%高い. 一般的に, 新聞記事は Lead 文が重要であると言われているが, 社説よりも報道記事においてその傾向が強いと考えられる. 多くのシステムは文の位置情報 (Lead) を用いており, この情報が報道記事の要約において有効であったと推測される.

一方, システム I だけは, どの要約率においても社説の F 値の方が報道記事よりも高かった. システム I は, 対象記事のジャンルによって重要文抽出の方法を変えている. 文の位置情報について, 報道記事の場合は記事の先頭ほど重要度の重みを高くしているが, 社説の場合は, 記事の先頭だけでなく末尾にも重要な記述があると考えている.

同様な考え方はシステム II でも取り入れられている. システム II は, システム I と異なり, 報道記事の方が F 値が高いが, 他のシステムと社説の F

表 1: 参加システムの特徴および課題別参加状況

SYS	(1) 内容語 (重要語)			(2) 手がかり語	(3) 構文解析		(4) 談話構造			(5) ジャンル	参加タスク		
	固有表現	タイトル	idf	文末表現	A-2	B	文の位置	接続詞	照応省略	の区別	A-1	A-2	B
I	○	○	○	○			○	○		○	○	○	○
II	○	○		○	○		○	○		○	○	○	
III			○		○	○						○	○
IV			○										○
V		○	○				○				○		
VI			○	○	○		○	○	○		○	○	○
VII		○		○	○		○				○	○	
VIII				○	○						○	○	
IX		○					○				○	○	○

表 2: 課題 A-1 における各システムの報道/社説記事における F 値

SYSTEM	10%		30%		50%		total
	報道	社説	報道	社説	報道	社説	
I	0.324	0.354	0.395	0.444	0.566	0.609	0.463
II	0.378	0.185	0.474	0.417	0.622	0.593	0.467
V	0.243	0.185	0.500	0.380	0.622	0.532	0.424
VI	0.297	0.215	0.509	0.326	0.617	0.510	0.434
VI'	0.297	0.154	0.518	0.310	0.622	0.516	0.429
VII	0.324	0.154	0.518	0.390	0.633	0.542	0.454
VII'	0.270	0.185	0.518	0.417	0.612	0.545	0.434
VIII	0.216	0.154	0.395	0.374	0.617	0.561	0.396
IX	0.405	0.246	0.447	0.337	0.592	0.561	0.449
IX'	0.216	0.215	0.447	0.358	0.617	0.545	0.416
Ave.	0.297	0.205	0.472	0.375	0.612	0.551	0.437

値を比較してみると, 要約率 30%と 50%において, F 値がシステム I の次に高い値が得られているため, ジャンルの区別が有効に働いていると考えられる.

システム I とシステム II は, 課題 A-1 で最も精度が高い 2 システムである. この 2 システムは, 先に述べたように記事のジャンルを考慮している他に, 共に固有表現抽出 (あるいはそれに準ずる処理) を行なっているという点でも似ており, これも重要文抽出の精度に影響を与えたのではないかと推測される.

課題 A-2 の結果分析

課題 A-2 においても課題 A-1 の分析と同様に, 報道記事と社説で, 人間による主観評価の結果を分けて調べてみた (表 3). その結果, 読みやすさに関しては要約率 20%, 40%共に報道記事の方が社説よりも平均順位を上回っていたが, 内容に関しては要約率 20%, 40%共に社説の方が報道記事よりも平均順位が高かった.

課題 A-2 の評価方法の分析

1. 主観評価について

主観評価に用いた 4 種類の要約 (FREE, PART, SYS, BASE) と順位の関係を表 4 に示す. 表は, FREE, PART, SYS, BASE の 4 種類の要約が, 内容 (CONT) および読みやすさ (READ) の観点において, 1 位, 2 位, 3 位, 4 位それぞれにランクされ

表 3: 課題 A-2 の結果 (主観評価における平均順位)

SYSTEM	20% 内容		20%読みやすさ		40% 内容		40 %読みやすさ	
	報道	社説	報道	社説	報道	社説	報道	社説
I	3.5	3.2	3.1	3.1	3.3	2.9	2.7	2.5
II	3.0	2.9	2.7	2.3	2.9	2.6	2.9	2.5
III	3.7	3.3	4.0	3.5	3.8	3.2	4.0	3.8
III'	3.3	3.2	3.7	3.7	3.5	3.1	3.9	3.5
VI	3.7	3.1	3.3	3.5	3.1	3.2	3.1	3.5
VII	3.2	3.2	2.4	3.0	3.0	3.1	2.3	2.9
VIII	3.3	3.2	3.3	3.2	3.3	3.1	2.9	3.2
IX	3.0	3.2	2.6	3.3	3.3	3.0	2.8	3.4
IX'	3.3	3.1	3.0	3.0	3.1	3.0	2.6	2.9
Lead	3.3	3.2	3.1	3.3	3.0	3.1	2.8	2.7
Ave.	3.3	3.2	3.1	3.2	3.2	3.0	3.0	3.1

た割合を示している[1]．表より，FREE は 1 位を占める割合が一番高く (73.5%)，次いで PART, SYS, BASE の順になっているが，大まかには以下に示すような大小関係があると考えられる．

(1)FREE>(2)PART
>(3) システム要約とベースライン

FREE や PART は，システムの要約と比べると，要約の品質 (CONT, READ) に少し開きがあり，システムを評価する上で，少し厳しい比較対象であった．すなわち，4 種類の要約のランキングによる主観評価は，システム間の比較という意味では十分な意味を持たせられていないと言える．

2. content based な評価について

まず，content based な評価結果と主観評価の結果の相関について調査した．調査は，主観評価に用いた 4 種類の要約の中から任意の 2 つを選び，主観評価による順序と content based な評価の大小関係が一致する割合を調べた．4 種類の要約の組合せは「FREE-PART」「FREE-SYS」「FREE-BASE」「PART-SYS」「PART-BASE」「SYS-BASE」の 6 通りあるが，FREE と PART は共に content based な評価で評価基準として用いており，どちらも人手で作成した理想的な要約であるため，6 通りの組合せから「FREE-PART」の組合せだけ除外した[2]．また，主観評価は内容と読みやすさの 2 つの側面から行なったが，content based な評価は，要約間の内容の類似度を測るために用いられた指標であるため，主観評価結果は内容による比較のものを用いた．

表 5 は，その結果である．表から，要約率が 20% と 40%の両方において，主観評価の結果と content based な評価が，高い割合 (約 90%) で一致していることが分かる．

一方，先にも述べたように，主観評価で比較した 4 種類のうち，システムの要約とベースライン (Lead) の要約は，FREE や PART と比べると平均的に同程度の品質の要約であると考えられる．そこで，表 5 の中でも特にシステムの要約とベースラインに着目し，比較を行った．結果を表 6 に示す．表 6 において，主観評価と content based な評価との相関は，表 5 の場合ほどはっきりとは現れていない．このことから，content based な評価は，品質に大きな違いのある 2 つの要約を比較する上では，よい指標であるが，品質が僅差な 2 つの要約を比較する上では，それほど有用な指標ではないと考えることができる．

そこで，さらに，content based のスコアの差と信頼度 (精度) に関する調査を行なった．content based のスコアの差に着目し，スコアの差 0.1 毎に 2 つの要約の content based のスコアと主観評価の順位との大小関係が一致する事例の割合について

[1] (要約率 20%, 40%) × 30 テキスト × 10 システム=600

[2] すなわち，5 通りの組合せ × 30 テキスト × 10 システム =1500 通りの組合せについて調べた．

表 4: 主観評価に用いた 4 種類の要約と順位の関係

		1 位	2 位	3 位	4 位
FREE	CONT	69.8%(419/600)	28.7%(172/600)	1.5%(9/600)	0.0%(0/600)
	READ	77.7%(466/600)	19.0%(114/600)	3.2%(19/600)	0.2%(1/600)
	TOTAL	73.5%(885/1200)	23.8%(286/1200)	2.3%(28/1200)	0.1%(1/1200)
PART	CONT	49.0%(294/600)	49.0%(294/600)	1.8%(11/600)	0.2%(1/600)
	READ	40.6%(244/600)	47.5%(285/600)	8.5%(51/600)	3.0%(18/600)
	TOTAL	44.8%(538/1200)	48.3%(579/1200)	5.3%(64/1200)	1.6%(19/1200)
SYS	CONT	2.3%(14/600)	3.3%(20/600)	68.0%(408/600)	26.3%(158/600)
	READ	11.2%(67/600)	10.3%(62/600)	43.3%(260/600)	38.8%(233/600)
	TOTAL	6.6%(79/1200)	6.8%(82/1200)	55.7%(668/1200)	32.6%(391/1200)
BASE	CONT	0.0%(0/600)	0.8%(5/600)	38.2%(229/600)	61.0%(366/600)
	READ	6.5%(39/600)	8.0%(48/600)	52.7%(316/600)	32.8%(197/600)
	TOTAL	3.2%(39/1200)	4.4%(53/1200)	45.4%(545/1200)	46.9%(563/1200)

表 5: 主観評価による順序と content based な評価の大小関係が一致する事例の割合 (全データ)

	FREE	PART
20C	91.4%(1371/1500)	88.6%(1329/1500)
40C	89.3%(1339/1500)	90.0%(1350/1500)

表 6: 主観評価による順序と content based な評価の大小関係が一致する事例の割合 (SYS と BASE)

	FREE	PART
20C	64.3%(193/300)	58.0%(174/300)
40C	58.7%(176/300)	63.7%(191/300)

表 7: content based のスコアと主観評価の順位との大小関係が一致する事例の割合

CB スコアの差	主観評価と一致する事例の割合 (%)
0.0 - 0.1	0.718(1784/2484)
0.1 - 0.2	0.829(1967/2374)
0.2 - 0.3	0.931(1938/2082)
0.3 - 0.4	0.960(1613/1681)
0.4 - 0.5	0.950(1157/1218)
0.5 - 0.6	0.966(1150/1190)
0.6 - 0.7	0.965(656/680)
0.7 - 0.8	0.966(201/208)
0.8 - 0.9	0.980(50/51)
0.9 - 1.0	0.969(31/32)

調べた. 結果を表 7 に示す. 表より, content based のスコアで 0.2 以上の開きがあれば, 93%以上の割合で主観評価の結果と一致する, すなわち, 93%以上の信頼度で要約を評価することが可能になると思われる.

課題 B の結果

課題 B では, 要約率を特に指定せずに要約を作成してもらった. まず, 被験者が検索クエリと要約とのレレバンスの判定にかかった時間と平均要約長との関係について調べた. 結果を図 1 に示す. 図において平均要約長の一番短いシステム III でも, 判定にある一定の時間 (約 400 秒/50 要約) が必要であることがわかる. 課題 B の評価は, 被験者に計算機を用いて行なってもらったが, 400 秒の時間の中には, 計算機の操作に要する時間等が含まれていると考えられる.

次に判定時間と各システムの F 値との関係について調べた. 結果を図 2 と図 3 に示す. Level A に

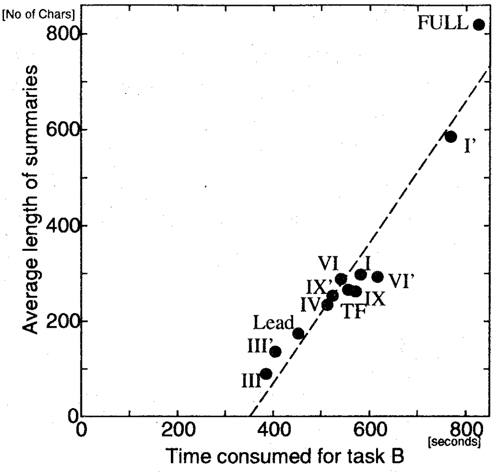

図 1: 課題 B における判定に要した時間と平均要約長との関係

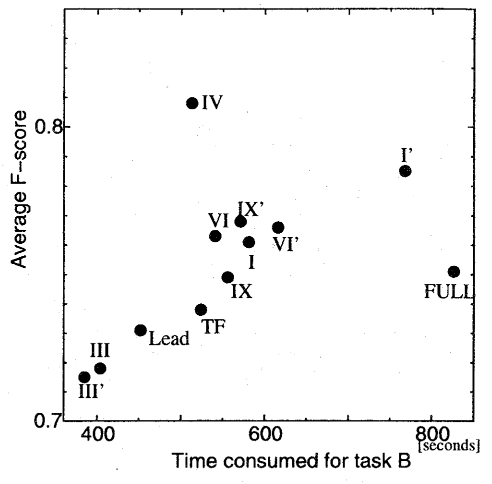

図 2: 課題 B における判定に要した時間と F 値との関係 (Level A)

関して, 全体的な傾向としては, 判定時間の長い要約 (平均要約長の長いもの) は平均 F 値が高くなる傾向にある. この中で, 特にシステム IV は平均要約長が短めであるにもかかわらず, 非常に高い F 値が得られていることがわかる.

課題 B の結果の分析

課題 B ではシステム III と III' を除き, 多くのシステムが重要文抽出により要約を作成している. そこで, 望月ら [5] の分析手法を用い, システムが出力した要約の類似性に基づいて要約を分類し, 結果を考察する. まず, 提出された要約毎に, 要素を元テキスト (全文) の各文とし, 値をその文が重要文として選択されれば 1, されなければ 0 としたベクトルを用意する. 次に 2 つのベクトル間のコサイン距離を計算し, 各要約間の類似度とする. 最後に, 最短距離法と平均距離法の 2 種類のクラスタ間の距離によって, 要約作成手法の階層型クラスタリングを行なう.

結果を図 4 に示す. 図はクラスタリングに平均距離法を適用した場合のクラスタおよびクラスタ間の類似度を示している. 最短距離法の場合も, クラスタ間の類似度は異なるが, 同一の階層型クラスタが得られている.

課題 B に参加した団体のうち, 4 団体がそれぞれ 2 種類の要約システムをエントリしたが, 図 4 から分かるように, 同一団体の 2 システムはクラスタが形成される初期の段階で融合されている. 一

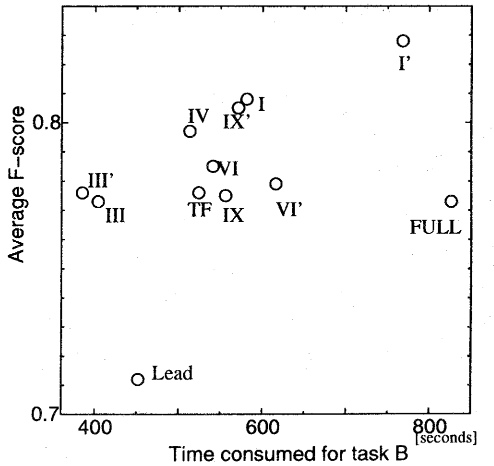

図 3: 課題 B における判定に要した時間と F 値との関係 (Level B)

方, Level A で最も F 値の高かったシステム IV はクラスタの一番上位で融合されている.

実際, システム IV の要約作成手法は, 他のシステムと大きく異なる. 多くのシステムは, 課題 B を単一テキスト要約の枠組でとらえ, 従来の要約手法に加え, 検索クエリに含まれる語の重要度の重みを増やすという方法で要約を作成している. これに対し, システム IV では, 課題 B を複数テキスト要約の枠組でとらえており, ある検索クエリに対するテキスト集合のテキスト間の類似性を考慮している. システム IV はテキスト集合をテキス

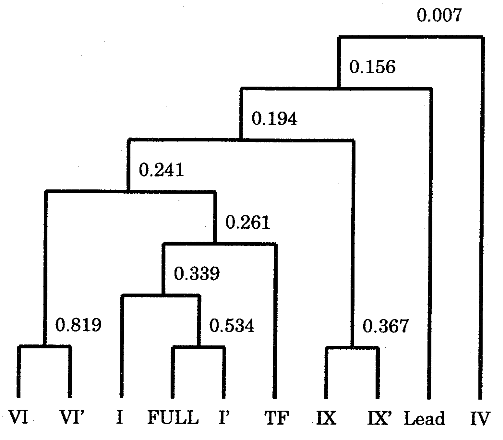

図 4: 課題Bにおける要約間の類似度

ト間の類似度に基づいて階層的にクラスタリングする．次に，情報利得比を用いてクラスタの特徴を良く表す語に重み与え，その重みに基づき要約を作成する．近年，複数テキスト要約研究において，同様の考え方に基づいた要約手法が着目されつつあるが [7]，課題Bの評価においてその有効性が部分的に確認されたと考えることができる．一方，formal run 後のアンケート結果によれば，すべてのシステムの中でシステム IV は要約にかかる計算コストが高い (100 秒/記事).

課題Bは，要約作成の実時間性というのが本質的な課題であり，評価の一指標として要約作成時間も考慮に入れる必要がある．しかし，formal run 後のアンケート結果における計算時間の数値は，必ずしも信憑性が高くないので，ここでは詳しく述べない．

7 結論

本稿では，2000 年 5 月から 2001 年 3 月まで開催された第 2 回 NTCIR ワークショップサブタスク TSC の結果を分析し，また評価方法について検討した．

今後，以下の 2 点について分析を行なう予定である．

- 課題 A-2 について

  課題 A-1 では，システム全体の F 値の平均は，報道記事の方が社説よりも上回っていた．一方，課題 A-2 では逆に社説の F 値の方が高かった．この結果について，今後分析する必要がある．

- 課題 B について

  課題Bの評価において，検索クエリと要約のレレバンスの判定を行なった被験者の，被験者間の判定精度 (F 値) のばらつきが大きいと予想される．今後は，望月らが同様のタスクの分析に取り入れている手法を参考に [5]，F 値が極端に異なる被験者は評価に用いないといった処置も必要であると考えられる．

参考文献

[1] Fukushima, T. and Okumura, M., "Text Summarization Challenge Text Summarization Evaluation at NTCIR Workshop2", Proceedings of the Second NTCIR Workshop Meeting, 2001. (to appear)

[2] Fukushima, T. and Okumura, M., "Text Summarization Challenge Text Summarization Evaluation in Japan", Proceedings of NAACL 2001 Workshop Automatic Summarization, pp.51–59, 2001.

[3] Donaway, R.L., Drummey, K.W., and Mather, L.A., "A Comparison of Rankings Produced by Summarization Evaluation Measures", Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization, pp.69–78, 2000.

[4] Mani, I., et al., "The Tipster SUMMAC Text Summarization Evaluation", Technical Report, MTR 98W0000138, The MITRE Corp, 1998.

[5] 望月源, 奥村学, "語彙的連鎖に基づく情報検索タスクを用いた評価", 自然言語処理, Vol.7, No.4, pp.63–77, 2000.

[6] 奥村学, 難波英嗣, "テキスト自動要約に関する研究動向", 自然言語処理, Vol.6, No.6, pp.1–26, 1999.

[7] 奥村学, 難波英嗣, "テキスト自動要約に関する最近の話題", 北陸先端科学技術大学院大学 情報科学研究科 Research Report, IS-TM-2000-001, 2000. (http://galaga.jaist.ac.jp:8000/pub/papers/oku/summarize2000.ps.gz)

[8] Proceedings of The Tipster Program Phase III, Morgan Kaufmann, 1999.