要約の内的 (intrinsic) な評価法に関するいくつかの考察 – 第 2 回 NTCIR ワークショップ 自動要約タスク (TSC) を基に –

難波　英嗣[†]　奥村　学[†]

システムの出力した要約そのものを評価する方法は，一般に内的な評価と呼ばれている．これまでの典型的な内的な評価の方法は，人手で作成した抜粋と要約システムの出力との一致度を，F-measure 等の尺度を用いて測ることで行われてきた．しかし，F-measure は，テキスト中に類似の内容を含む文が複数存在する場合，どちらの文が正解として選択されるかにより，システムの評価が大きく変化する，という問題点がある．本研究では，この問題点を解消するいくつかの評価方法をとりあげ，その有用性に関する議論を行う．F-measure の問題点を解消する評価方法の 1 つに utility に基づく評価があるが，この方法では評価に用いるデータ作成にコストがかかるという問題がある．本研究では，あるテキストに関する複数の要約率のデータを用いることで，疑似的に utility に基づく評価を実現する方法を提案する．提案する評価方法を，第 2 回 NTCIR ワークショップ自動要約タスク (TSC) のデータに適用し，有用性に関する調査を行った結果，提案方法は，F-measure の問題点をある程度改善できることが確認された．次に，F-measure の問題点を解消する他の評価方法の一つである content-based な評価を取り上げる．content-based な評価では，指定された要約率の正解要約を一つだけ用意すれば評価可能であるため，utility に基づく評価に比べ，被験者への負荷が少ない．しかし，この評価方法で 2 つの要約を比較する場合，どの程度意味があるのかについては，これまで十分な議論がなされていない．そこで，pseudo-utility に基づく評価と同様に TSC のデータを用い，content-based な評価の結果を被験者による主観評価の結果と比較した結果，2 つの要約が content-based な評価値で 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ことが分かった．

キーワード: *pseudo-utility* に基づく評価，*F-measure*，*content-based* な評価，テキスト自動要約，内的な評価，*TSC*，*NTCIR*

Some Examinations of Intrinsic Methods for Summary Evaluation Based on Text Summarization Challenge(TSC), a Subtask of NTCIR Workshop 2

Hidetsugu Nanba[?] and Manabu Okumura[†]

Evaluation methods whose targets are system outputs (summaries) themselves are often called “intrinsic methods”. Computer-produced summaries have been traditionally evaluated by comparing with human-written summaries using the F-measure. But, the F-measure has the following problem: the F-measure is not appropriate

when alternative sentences are possible in a human-produced extract. For example, when there are two sentences 1 and 2, and sentence 1 is in a human-produced extract, if a system chooses sentence 2, it obtains lower score, even if sentences 1 and 2 are interchangeable. In this paper, we examine some of the evaluation methods devised to overcome the problem. Several methods that devised to overcome the problem have been proposed. Utility-based measure is one of them. However, the method requires a lot of effort for humans to make data for evaluation. In this paper, we first propose pseudo-utility-based measure that uses human-produced extracts at different compression ratios. In order to evaluate the effectiveness of pseudo-utility-based measure, we compare our measure and the F-measure using the data of Text Summarization Challenge(TSC), a subtask of NTCIR workshop 2, and show that pseudo-utility-based measure can resolve the problem. Next, we focus on content-based evaluation. Though it is reported that content-based measure is effective to resolve the problem, it has not been examined from a viewpoint of comparison of two extracts that are produced from different systems. We evaluated computer-produced summaries of the TSC by the content-based measure, and compared the results with a subjective evaluation. We found that the evaluation by the content-based measure matched those by humans in 93% of the cases, if the gap in the content-based scores between two abstracts is more than 0.2.

KeyWords: *pseudo-utility-based evaluation, the F-measure, content-based evaluation, automatic text summarization, intrinsic method, TSC, NTCIR*

1 序論

近年，テキスト自動要約の研究が活発化するとともに，要約の評価方法が研究分野内の重要な検討課題の一つとして認識されてきている．これまで提案されてきた要約の評価方法は，内的な (intrinsic) 評価と外的な (extrinsic) 評価の 2 種類に分けることができる (Sparck-Jones and Galliers 1996)．内的な評価とは，システムの出力した要約そのものを，主に内容と読みやすさの 2 つの側面から評価する方法である．一方， 外的な評価とは，要約を利用して人間がタスクを行う場合の，タスクの達成率が間接的に要約の評価となるという考え方に基づいて評価を行う方法である．本研究では，近年活発にその評価方法が議論され，改良が試みられている内的な評価，特に内容に関する評価方法に焦点を当てる．

これまでの要約の内容に関する評価は，人手で作成した抜粋と要約システムの出力との一致の度合を，F-measure 等の尺度を用いて測るのが典型的な方法であった．しかし，Jing ら (Jing, Barzilay, McKeown, and Elhadad 1998) は，要約の F-measure による評価と外的な評価を分析し，F-measure には「テキスト中に類似の内容を含む文が複数存在する場合，どちらの文が正解として選択されるかにより，システムの評価は大きく変化する」という問題があることを指摘している．

この問題点を解決する方法がこれまでにいくつか提案されている．Radev ら (Radev, Jing,

† 東京工業大学精密工学研究所, Precision and Intelligence Laboratory, Tokyo Institute of Technology

and Budzikowska 2000) は，文の utility という概念を用いた評価方法を示している．文の utility とは，そのテキストの話題に対する各文の適合度 (重要度) を 10 段階で表したものであり，正解の文の utility にどのくらい近い utility の文を選択できるかで評価を行なう．しかし，このような適合性の評価は被験者への作業負荷が大きいという問題がある．

Donaway ら (Donaway, Drummey, and Mather 2000) は，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コサイン距離で評価する content-based な評価を提案している．content-based な評価では，指定された要約率の正解要約を一つだけ用意すれば評価可能であるため，utility に基づく評価に比べ，被験者への負荷が少ない．しかし，この評価方法で 2 つの要約を比較する場合，どの程度意味があるのかについては，これまで十分な議論がなされていない．

そこで，本研究では，まず，utility に基づく評価の問題点を改良する新しい評価方法を提案する．一般に低い要約率の抜粋に含まれる文は高い要約率の抜粋中の文よりも重要であると考えられる．このような考えに基づけば，あるテキストに関して複数の要約率のデータが存在する場合，テキスト中の各文に重要度を割り振ることが可能であるため，utility に基づく評価を疑似的に実現することができる．これまでの要約研究において，1 テキストにつき複数の要約率で正解要約が作成されたデータは数多く存在する (例えば，(Jing, et al. 1998)) ことから，提案する評価方法に用いるデータの作成にかかる負荷は決して非現実的なものではなく，utility を直接被験者が付与するより負荷は小さいと考えられる．

本研究では，評価型ワークショップ NTCIR 2 の要約サブタスク TSC(Text Summarization Challenge)(Fukushima and Okumura 2001:a, 2001:b) で作成された 10%，30%，50% の 3 種類の要約率の正解データを用いて，提案方法により評価を行う．この評価結果を F-measure による結果と比較し，提案方法が F-measure による評価を改善できることを示す．

次に，本研究では，content-based な評価を取り上げる．同様に TSC のデータを用いて，人間の主観評価の結果と比較し，これまで十分議論されていないその有用性に関する議論を行う．

本論文の構成は以下のとおりである．次節では，まず，これまで提案されてきた内的な評価方法，特に F-measure の問題点の解消方法について述べる．3 節では，本研究で提案する評価方法について説明する．4 節では，F-measure と提案する評価方法を比較し，結果を報告する．また，content-based な評価に関する調査についても述べる．最後に結論と今後の課題について述べる．

2　関連研究

要約の内容に関する評価について，Jing ら (Jing, et al. 1998) は，典型的な評価方法の 1 つである F-measure をとりあげ，その問題点をいくつか指摘している．Jing らは，システムの要約と人間の被験者の作成した抜粋との比較による評価と，要約を利用して人間がタスクを行な

う場合のタスクの達成率による評価の 2 つの評価方法を分析し，評価結果に影響を与える要因を同定することを試みているが，その結果少なくとも次の 2 つの点において，これまでの人間の抜粋を用いた評価方法は問題であるとの知見を得ている．

- **問題点 1(要約率の変化に伴う評価値の変化):**
  人間の抜粋との比較による評価では，要約率を変化させると，システムの評価がかなり変化する．このため，特定の要約率でシステム間の性能の比較をする意味がどの程度あるのかは疑問が残る．
- **問題点 2(テキスト中の複数類似個所の選択問題):**
  テキスト中に類似の内容を含む文が複数存在する場合，どちらの文が正解として選択されるかにより，システムの評価は大きく変化する．

これまで，問題点 1(要約率の変化に伴う評価値の変化) を解消するいくつかの方法が提案されている．Mittal ら (Mittal, Kantrowitz, Goldstein, and Carbonell 1999) は，要約率の違いによるシステムの評価の違いに関して，さまざまな要約率における精度を求めた上で，情報検索の評価で用いられている 11 点平均精度 (11 point average precision) のように，複数の要約率での精度の平均として結果を示すべきであるとしている．

また，コーパスとするテキスト集合の違いが精度に影響を与えることから，コーパスの要約のしやすさを計る指標として，ランダムに文を選択して要約を作成した場合の精度をベースラインとして示すべきであると主張している．そして，システムの性能を評価する場合，

$$p' = \frac{p - b}{1 - b}$$

(ここで，p，b，p' はそれぞれシステム，ベースライン，補正後のシステムの精度) のように，ベースラインを用いて補正した精度を用いるべきであるとしている．

一般に，F-measure で要約の精度を評価する場合，ベースライン値＝要約率と考えることができるため，要約率が大きくなるにつれ，F-measure 値は大きくなる傾向にある．従って，ベースラインを用いて評価値を補正する上記の評価方法は，Jing らの指摘する問題点 1 の解消には有用であると考えられる．

一方，被験者間の一致の度合を J とすると，J は要約システムの精度の上限と考えられ，また，ランダムに選択した時の精度 R は下限と言える．そのため，Radev ら (Radev, et al. 2000) も，Mittal らと同様に，システムの性能を計る値を示す際，普通に計算された値 S を単に用いるのではなく，これらの値で正規化した値

$$S' = \frac{S - R}{J - R}$$

を示すべきであるとしている．

問題点 2(テキスト中の複数類似個所の選択問題) を解消する方法もいくつか提案されている．Jing ら (Jing, et al. 1998) は，人間が選択した重要文を用いて評価を行なう際，正解と一致し

た場合正解数 1，一致しない場合 0 として再現率，精度を計算するのではなく，正解数を被験者間の一致の度合として計算する方法を提案している．たとえば，5 人の被験者中 3 人，2 人がそれぞれ一致して選択した文が存在する場合，これまでの評価方法では，前者をシステムが選択した場合正解数 1(過半数以上の被験者が選択しているので)，後者では 0 となるが，提案する方法では，システムの正解数は，前者では 3/5，後者では 2/5 となる．

Radev ら (Radev, et al. 2000) は，文の utility という概念を用いた評価方法を示している．文の utility は，文がそのテキストの話題に対してどの程度適合した内容であるかを示す尺度であり，[0-10] の値をとる[1]．人間が選択した重要文を用いたこれまでの評価方法は，正解と一致した場合正解数 1，一致しない場合 0 として再現率，精度を計算していたが，utility に基づく評価値は，システムが選択した文に対して人間が割り当てた utility の総和を，正解の文の utility の総和で割った値として計算する．これまでの評価方法では，システムが選択した不正解の文は，全く評価が得られなかったのに対し，utility に基づく評価の場合，Jing らの方法と同様に，たとえ不正解でもその文がある程度の重要度を持つ場合，その重要度に対する部分的な評価が得られる点が異なる．ただ一つ正解が存在し，それとまさに一致することを要求されていたこれまでの評価に比べ，正解の文の utility にどのくらい近い utility の文を選択できるかで評価を行なう．

Donaway ら (Donaway, Drummey, and Mather 2000) は，2 種類の評価方法を提案している．一つは，人間にも，システムにも，テキスト中の文にすべて順位をつけさせるようにして，その文の序列を比較して評価を行なう方法である．これは，これまでの方法がテキスト中の文を重要/非重要の 2 つに分類して評価に利用していたのに対し，テキスト中の文数に分類して利用することに相当する．Donaway らが提案するもう一つの評価尺度は，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コサイン距離で評価する方法 (以後，content-based な評価) である．

Donaway らは，この 2 種類の評価尺度にこれまでの評価方法である再現率に基づいた評価を加え，これらを実験により比較，検討している．正解の抜粋に含まれる個所が要約作成者毎に異なっていても，内容の類似した個所を抜き出しているのであれば，どの要約作成者の抜粋を用いても似たような評価値が得られる必要がある．Donaway らは，4 人の要約作成者の作った抜粋を用いて，上で述べたいくつかの尺度で要約を評価し，尺度毎に評価値の相関を調べている．その結果，content-based な評価が人間の要約との比較による評価方法としては，Jing らの指摘する問題点 2 に対する解決策ともなっており，もっとも優れていると結論づけている．

1 generic な要約を考えた場合，テキスト中での文の重要度を示していると考えて良い．

3　pseudo-utility に基づく評価

本研究では，あるテキストに関する複数の要約率の正解データを用いることで，utility に基づく評価を疑似的に実現する方法 (以後，pseudo-utility に基づく評価) を提案する．

例えばあるテキストに，要約率が r_1%，r_2%，r_3%$(r_1 < r_2, r_2 < r_3)$ の 3 種類の正解データが存在する場合，テキスト中の各文は (1)r_1%の要約に含まれる文，(2)r_1%には含まれないが r_2%には含まれる文，(3)r_2%には含まれないが r_3%には含まれる文，(4)r_3%に含まれない文の 4 種類に分けられる[2]．これらは，テキストの話題に対する各文の適合度が 4 段階で表されたデータと考えることができるため，4 段階の疑似的な utility に基づく評価が実現できる．

表 1に示す例を用いて，pseudo-utility の計算方法を説明する．表 1は，要約率 10%，30%，50%の要約データを用いた場合について述べている．表 1では，S1–S10 の 10 文からなるテキストについて，要約率毎に，要約作成者と 2 つの要約システム (*System 1* と *System 2*) が選択した重要文を '+' で示している．また，ここでは各文の重要度 w を '1/要約率' として計算する．

表において，例えば System 1 の要約率 50%の要約において，System 1 が重要文として選択した 5 文 (S3，S4，S7，S9，S10) のうち 3 文 (S4，S7，S10) が一致するため，F-measure 値は 0.6(3/5) となる．一方，System 1 が選択した 5 文 (S3，S4，S7，S9，S10) の重要度はそれぞれ 0，1/30，1/50，0，1/30 であるため，重要度の総和は $0 + 1/30 + 1/50 + 0 + 1/30 = 13/150$ となる．また，要約作成者は要約率 50%では S1，S4，S7，S8，S10 の 5 文を選択している．この場合の重要度の総和は $1/10 + 1/30 + 1/50 + 1/50 + 1/30 = 31/150$ となる．pseudo-utility 値は，システムの選択した文の重要度の総和を要約作成者の選択した文の重要度の総和で割って正規化した値であり，この例の場合 $\frac{13/150}{31/150} = 0.419$ となる．

表 2に，System 1，2 の F-measure 値と pseudo-utility 値を示す．表 2において，要約率 10%における F-measure 値と pseudo-utility 値を比較すると，どちらのシステムも 10%要約の正解である S1 ではなく S4 を選択しているため，F-measure 値は 0 になる．ここで，S4 は 30%要約の正解に含まれているため，S1 よりも重要度は低いが，ある程度重要な情報を含んだ文であると考えられる．この例の場合，要約率 10%では F-measure 値は 0 か 1 しか取り得ないが，pseudo-utility に基づく評価では，このような文も評価の対象とすることで，より適切な評価が可能になる．

また，要約率 50%における System 1 と System 2 の結果を比較すると，どちらも選択した 5 文のうち 3 文が 50%要約の正解データに含まれているため，F-measure 値は共に 0.6 である．この 3 文のうち S4 と S10 は System1，2 が共通して選択しているが，他の 1 文は System 1 は S7(重要度 1/50)，System 2 は S1(重要度 1/10) を選択している．2 システムが同数の正解文を抽出している場合でも，特に重要と考えられる文 (この場合 S1) が抽出されている場合とそうでない場合との区別が F-measure ではできない．一方，pseudo-utility に基づく評価では，この場

2 ただし，r_1%の要約は r_2%の要約に $(r_1 < r_2)$，r_2%の要約は r_3%の要約に $(r_2 < r_3)$ 含まれていることが前提となる．

表 1　pseudo-utility に基づく評価の例

	正解データ			重要度	*System 1*			*System 2*		
	10%	30%	50%	(重要度)	10%	30%	50%	10%	30%	50%
S1	+	+	+	1/10	-	-	-	-	+	+
S2	-	-	-	0	-	-	-	-	-	-
S3	-	-	-	0	-	-	+	-	-	-
S4	-	+	+	1/30	+	+	+	+	+	+
S5	-	-	-	0	-	-	-	-	-	-
S6	-	-	-	0	-	-	-	-	+	+
S7	-	-	+	1/50	-	-	+	-	-	-
S8	-	-	+	1/50	-	-	-	-	-	-
S9	-	-	-	0	-	+	+	-	-	+
S10	-	+	+	1/30	-	+	+	-	-	+

表 2　F-measure と pseudo-utility に基づく評価によるシステムの比較例

	System 1		*System 2*	
	F-measure	pseudo-utility	F-measure	pseudo-utility measure
10%	0.000 ($\frac{0}{1}$)	0.333 ($\frac{1/30}{1/10}$)	0.000 ($\frac{0}{1}$)	0.333 ($\frac{1/30}{1/10}$)
30%	0.667 ($\frac{2}{3}$)	0.400 ($\frac{2/30}{1/10+2/30}$)	0.667 ($\frac{2}{3}$)	0.800 ($\frac{1/10+1/30}{1/10+2/30}$)
50%	0.600 ($\frac{3}{5}$)	0.419 ($\frac{2/30+1/50}{1/10+2/30+2/50}$)	0.600 ($\frac{3}{5}$)	0.806 ($\frac{1/10+2/30}{1/10+2/30+2/50}$)

合 System 1，2 それぞれにおいて 0.419，0.806 と評価値に開きがあり，この例では両者の区別がつけられることが分かる．

4　評価方法の分析

本研究では，pseudo-utility に基づく評価の有効性を調べるために，TSC のデータを用いて評価を行う．また，TSC では content-based な評価がシステムの評価方法の一つに採用されているが，この評価結果を用い，content-based な評価の有効性についても検討する．

本節では，まず，4.1 節で TSC の課題および評価方法について説明する．次に，4.2 節で TSC のデータを用いた本研究の分析について述べる．

4.1　TSC における評価

TSC とは，要約研究における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に行われた，第 2 回 NTCIR ワークショップのタスクである．TSC では 3 種類の課題が設定されているが，本節ではそのうち内的 (intrinsic) な評価を適用している 2 つの課題 A-1「重要文抽出型要約」と A-2 「人間の自由作成要約と比較可能な要約」について述べる．なお，結果に関する詳細およびこの他の課題については，

(Fukushima and Okumura 2001:a, 2001:b; 難波，奥村 2001) を参照されたい．

4.1.1　課題

・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%で重要文を抽出する．

・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する．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．

4.1.2　要約対象テキスト

毎日新聞 94 年および 98 年から 15 記事づつ，計 30 記事が選ばれている．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が，98 年からは 1200，2400 文字以上の 2 種類の長さの社説が選ばれている．

4.1.3　評価方法

課題 A-1

課題 A-1 のシステムの提出結果は，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-measure 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

(i) 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．

(ii) 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 (FREE と PART) 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 (Donaway, Drummey, and Mather 2000)．ベクトルの要素は，各内容語の tf*idf 値とし，idf 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，課題 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．

4.2　評価方法の分析

pseudo-utility に基づく評価の有効性を示すためには，pseudo-utility に基づく評価と utility に基づく評価の比較，および，F-measure と pseudo-utility に基づく評価の比較を行う必要があると考えられる．しかし，先にも述べたとおり，utility に基づく評価に用いるデータの作成にかかる負荷は非常に高く準備が困難であったため，本研究では pseudo-utility に基づく評価方法を課題 A-1 に適用し，F-measure との比較のみを行った．4.2.1 節では，まずこの結果につい

て報告する．

次に，課題 A-2 の結果を用いて，content-based な評価と主観評価の比較を行った．比較結果について 4.2.2 節で述べる．

4.2.1 F-measure と pseudo-utility に基づく評価の比較 (課題 A-1)

まず，実際にどの程度 pseudo-utility に基づく評価が有効に機能しているか，いくつかの事例にあたって調べてみた．図 1は，pseudo-utility に基づく評価が有効に機能した典型例である．2 文は,「アジアにおけるエイズ感染」に関する報道記事から，要約率 10%(1 文) で重要文を選択したシステムの出力結果と正解の要約である．この 2 文は，どちらも「アジアにおいてエイズ患者が急増している」ことを示した個所である．F-measure による評価では，システムは正解文を選択していないので，F-measure 値は 0 となる．一方，システムの選択した文は 30%要約には含まれているため，pseudo-utility 値は 0.333($\frac{1/0.3}{1/0.1}$) となる．一般に，報道記事 1 記事に含まれる文数は 10 文-20 文が中心的であり，この場合，要約率 10%の時は正解文が 1-2 文しかない[3]．このような場合，システムがある程度重要な情報を含んだ文を抽出していても，最重要文が抽出されなければ F-measure では全く評価に反映されない．一方，pseudo-utility に基づく評価では，図 1の例のようにある程度評価値に反映されるため，より適切なシステムの評価が行なえると考えることができる．

別の例を図 2に示す．記事 940715208 において，要約率 10%では正解要約文数は 3 文である．システムが出力した 3 文のうち第 1 文目が正解の要約に含まれているため，F-measure 値は 0.333 となっている．一方，システムの出力した 3 文のうち，正解に含まれていない残りの 2 文の一方は 30%の正解に，もう一方は 50%の要約に含まれているため，pseudo-utility 値は 0.511($\frac{1/0.1+1/0.3+1/0.5}{3/0.1}$) となっている．正解とシステムの出力を比較すると，正解の 2 文目にある「大学や教育施設一体となった動き」の具体例がシステムの要約の 2 文目と 3 文目になっていることがわかる．つまり，システムの抽出した 2 文は正解文 (2 文目) の部分情報となっている．このような個所をシステムが抽出できたことを pseudo-utility では適切に評価できていることは妥当であると思われる．

これらの調査から，pseudo-utility に基づく評価が，Jing らの指摘する問題点 2(テキスト中の複数類似個所の選択問題) をある程度解消できていると考えられる．

次に，F-measure と pseudo-utility に基づく評価を適用した結果をシステム別にまとめた．結果を表 3および表 4 に示す[4]．課題 A-1 には 7 団体 10 システム参加しており，表中の I–IX は各システムの ID を，また，同団体の異なるシステムはダッシュで示してある．

F-measure と pseudo-utility に基づく評価の各システムの順位を比較すると，F-measure で

3 TSC データにおいて，要約率 10%の時の報道記事の正解文数は過半数が 1, 2 文であった．

4 なお，評価用のデータは，脚注 1 の条件を満たさない 4 記事 (940701189, 940702187, 940716331, 980203053) を除く 26 記事を用いている．

記事番号: 940702171，要約率: 10%(1文)
見出し: エイズ感染「アジア、２０００年には４倍」－－来日のＷＨＯ局長警告
F-measure 値: **0.000**，pseudo-utility 値: **0.333**

- (正解)
  世界のエイズ患者は推計で約四百万人に達し、特にアジアではこの一年間で八倍にも急増して約二十五万人になったと、世界保健機関（ＷＨＯ）＝ＮＥＷＳのことば参照＝世界エイズ対策プログラム局長のマイケル・マーソン博士が一日、発表した。
- (システム)
  八月に横浜市で開かれる第十回国際エイズ会議を前に、来日中の同局長は厚生省で会見し「アジアの累積感染者数は二百五十万人以上だが、二〇〇〇年には四倍増の一千万人以上になると見込まれる」と警告した。

図 1　pseudo-utility に基づく評価がうまく適用された例 (換言)

記事番号: 940715208，要約率: 10%(3文)
見出し: 止まるか「理工系離れ」－－大学・文部省など“あの手この手”
F-measure 値: **0.333**，pseudo-utility 値: **0.511**

- (正解)
  技術立国ニッポンが危ない－－理科嫌いの子供の増加や大学の理工系志願者の伸び悩みなど「理工系離れ」が深刻になっている。
  こうした傾向にストップをかけようと、大学や教育施設一体となった動きが出ている。
  こうした動きの背景にあるのが、若者の理工系離れ。
- (システム)
  技術立国ニッポンが危ない－－理科嫌いの子供の増加や大学の理工系志願者の伸び悩みなど「理工系離れ」が深刻になっている。
  大学側などは、この夏、子供向けに科学の面白さをＰＲするプログラムを続々登場させた。
  文部省も十四日、理数系に強い高校生への支援策を開始する一方、専門家の懇談会からの報告を受け、魅力ある理工系大学作りに乗り出した。

図 2　pseudo-utility に基づく評価がうまく適用された例 (例示)

はそれぞれ 1 位，2 位であるシステム II，I が，pseudo-utility に基づく評価では順位が逆転している．また，多くのシステムは順位が 1 位か 2 位程度変動しており，中でもシステム V は，F-measure では 9 位なのが pseudo utility では 5 位になっている．そこで，これらの順位の変動が適正であるかどうかを調べるため，システム I と II の出力結果を比較した．

システム I と II が出力したそれぞれ 90 個の要約 (30 テキスト× 10%, 30%, 50%) のうち，システム I と II で F-measure 値は同じだが pseudo-utility 値の異なる 16 組の要約について調査した．16 組のうち，システム II よりも I の方が pseudo-utility 値が高くなる場合は 10 組，II が高い場合が 6 組であった．表 5にシステム I と II の出力例を示す．表 5は，記事 980500136 における要約率 10%での例で，原文中の文 ID，pseudo-utility に基づく評価に用いた重要度，シ

ステムIとIIが選んだ文，および文の内容を示している．重要度 1/10 の文が要約率 10%の正解である．システムIが選択した 5 文のうち要約率 10%の正解に含まれるもの (重要度 1/10) が 2 文 (S44 と S52) あるため，F-measure 値は 0.4 になる．システムIはこの他に重要度 1/30 の文を 1 文 (S30)，重要度 1/50 の文を 2 文 (S3 と S4) 選択しており，結果として，このテキストにおいては pseudo-utility 値 0.547 を得ている．

一方，システムIIもシステムIと同様に要約率 10%の正解に含まれる文 (重要度 1/10) を 2 文 (S26 と S43) 選択しているため，F-measure 値ではシステムIと同じく 0.4 になる．システムIIが選んだ残りの 3 文のうち，重要度 1/50 の文の 2 文 (S3 と S4) はシステムIと共通であるが，残りの 1 文 (S31) は重要度が 0 であり，pseudo-utility 値はシステムIよりも低い 0.480 に留まっている (表 6).

この記事の主題は「定年制　高齢者に多様な働き方を　６５歳現役社会の道も開け」であり，S22(重要度 1/10) はその問題提起になっている．システムIが選んだ S50 は S22 の一つの解決方法であり，ある程度重要な情報を持った文であるため，システムIとIIでこの文が選択できたかどうかで，pseudo-utility 値に差ができることは妥当であると考えられる．

4.2.2　content-based な評価の考察　– 主観評価との比較 – (課題 A-2)

まず，content-based な評価の比較対象となる主観評価の結果について簡単に述べる．次に，主観評価と content-based な評価の結果を比較し，考察する．

主観評価の結果

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 7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た割合を示している[5]．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，FREE や PART に比べ，SYS と BASE の品質は僅差であると言える．4 種類の要約の品質を大小関係で示すと大まかに次のようになる．

(1)FREE > (2)PART > (3) システム要約とベースライン

このようなデータの性質をふまえ，次節では主観評価と content-based な評価の比較を行う．

主観評価と content-based な評価の比較

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

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

表 3　課題 A-1 における各システムの F-measure 値

SYSTEM	10%	30%	50%	total (順位)
I	0.363	0.435	0.589	0.463 (2)
II	0.337	0.452	0.612	0.467 (1)
V	0.251	0.447	0.574	0.424 (9)
VI	0.305	0.431	0.568	0.435 (6)
VI'	0.282	0.435	0.572	0.429 (8)
VII	0.305	0.474	0.586	0.455 (3)
VII'	0.241	0.497	0.578	0.439 (5)
VIII	0.199	0.399	0.590	0.396 (11)
IX	0.358	0.420	0.571	0.450 (4)
IX'	0.268	0.409	0.570	0.416 (10)
TF	0.284	0.433	0.586	0.434 (7)
Lead	0.276	0.367	0.530	0.391 (12)
Ave.	0.289	0.433	0.577	0.433

表 4　課題 A-1 における各システムの pseudo-utility 値

SYSTEM	10%	30%	50%	total (順位)
I	0.518	0.559	0.664	0.581 (1)
II	0.450	0.603	0.673	0.569 (2)
V	0.410	0.546	0.641	0.527 (5)
VI	0.444	0.537	0.608	0.521 (8)
VI'	0.420	0.516	0.607	0.504 (9)
VII	0.433	0.560	0.651	0.541 (3)
VII'	0.401	0.556	0.636	0.525 (6)
VIII	0.330	0.515	0.654	0.495 (11)
IX	0.463	0.544	0.616	0.535 (4)
IX'	0.388	0.509	0.612	0.498 (10)
TF	0.406	0.526	0.657	0.525 (6)
Lead	0.401	0.481	0.549	0.468 (12)
Ave.	0.422	0.537	0.630	0.530

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[6]．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．また，content-based の評価値は，TSC における評価では FREE を基準にした場合と PART を基準にした場合の 2 種類で計算しているが，主観評価との比較も，それぞれの場合毎に分けて行っている．

6 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 5　記事 980511036 におけるシステム I と II の要約結果 (要約率 10%)

見出し: 定年制　高齢者に多様な働き方を　６５歳現役社会の道も開け

文 ID	重要度	I	II	文
S3	1/50	+	+	東京都武蔵野市にある「横河エルダー」の最高齢者、菅野清治さん（７９）は今も現役時とほぼ同じ週４０時間のフルタイムで元気いっぱいに働き続ける。
S4	1/50	+	+	「横河エルダー」は１９７５年に工業計器メーカー「横河電機」（従業員６３１１人）を定年退職した人たちのための受け皿会社として設立された。
S22	1/10			一律にではなく高齢者のニースに合わせ、多様なメニューをどう用意するか。
S26	1/10		+	年金支給開始年齢まで働きたくとも働く場がない、という切実な雇用問題が起きるおそれが多分にある。
S31	0		+	今年３月ごろから、６０歳定年制の見返りに、退職金や賃金をダウンさせたという訴えが連合東京をはじめ、全国の労組や労働相 談窓口などに相次いで寄せられている。
S43	1/10		+	約２０年前には２０歳代の若者は５人に１人、６５歳以上は１０人に１人だったのが、２０１５年には２０歳代は１０人に１人足らずとなり、逆に６５歳以上の人口比率は４人に１人を占める、世界に例のない高齢社会となる。
S44	1/10	+		意欲はあっても働けない高齢者が多くなるほど、年金や医療などの社会保障負担はより若い世代にしわ寄せされるのは明らかだ。
S50	1/30	+		それまでのキャリアを生かす継続雇用を基本に据え、職種によっては高齢者向けの職域拡大を図り、短時間勤務も認める。
S52	1/10	+		２１世紀の初めには「６５歳現役」が当たり前となる社会にしたい。

表 6　記事 980511036 におけるシステム I と II の F-measure 値および pseudo-utility 値 (要約率 10%)

	I	II
F-measure	0.400	0.400
pseudo-utility	0.547	0.480

表 7　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 8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 8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 9に示す．表 9において，主観評価と content-based な評価との相関は，表 8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

表 8　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 を基準	PART を基準
20%要約	91.4%(1371/1500)	88.6%(1329/1500)
40%要約	89.3%(1339/1500)	90.0%(1350/1500)

平均: 89.8%

表 9　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 の比較)

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

平均: 61.2%

表 10　content-based の評価値と主観評価の順位との大小関係が一致する事例の割合

content-based の評価値の差	主観評価と一致する事例の割合 (%)
0.0 - 0.1	0.578(718/1242)
0.1 - 0.2	0.771(916/1188)
0.2 - 0.3	0.928(966/1041)
0.3 - 0.4	0.959(805/839)
0.4 - 0.5	0.964(588/610)
0.5 - 0.6	0.988(589/596)
0.6 - 0.7	0.994(336/338)
0.7 - 0.8	0.990(103/104)
0.8 - 0.9	1.000(26/26)
0.9 - 1.0	1.000(16/16)

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 の評価値の差と信頼度 (精度) に関する調査を行なった．content-based の評価値の差に着目し，値の差 0.1 毎に 2 つの要約の content-based の評価値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調べた．結果を表 10に示す．表より, content-based の評価値で 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．

表 10から得られた知見を元に，表 9に示したシステムの要約とベースライン (Lead) の要約の比較結果のうち，content-based の評価値の差が 0.2 以上の場合について調べてみた．表 9の中で content-based の評価値の差が 0.2 以上となる事例の割合を表 11に示す．表 10において，評価値の差が 0.2 以上になる場合は全体の $59.5\%(1 - \frac{1242+1188}{6000})$ であるのに対し，システムの要約とベースラインの要約の間では全体の 14.5% しかない．このことからも，先にも述べたようにシステムの要約とベースラインの要約は，全体的に品質の近い要約であることが分かる．

次にシステムの要約とベースラインの要約との間の content-based の評価値の差が 0.2 以上

表 11 content-based 値の差が 0.2 以上ある事例の割合 (SYS と BASE の比較)

	FREE を基準	PART を基準
20%要約	17.0%(51/300)	23.0%(69/300)
40%要約	10.0%(30/300)	8.0%(24/300)

平均: 14.5%

表 12 content-based 値の差が 0.2 以上ある場合に主観評価による順序と一致する事例の割合 (SYS と BASE の比較)

	FREE を基準	PART を基準
20%要約	74.5%(38/51)	73.9%(51/69)
40%要約	60.0%(18/30)	70.8%(17/24)

平均: 71.3%

となる場合に，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を調べた．結果を表 12に示す．表 10における調査 (事例数:6000) と比べると事例数が少ない (174) ので，あまり厳密な値であるとは言えないが，表 12 の値 (71.3%) と表 9の 61.2%とを比較すると，評価値の差が 0.2 以上の場合 content-based な評価の信頼度が 10%以上高くなることが確認できる．しかし，この結果からは表 10における評価値の差 0.2 における一致度 92.8%までには至っていない．

5 結論と今後の課題

本研究では，要約の評価方法について，pseudo-utility に基づく評価方法を提案し，F-measure との比較を行った．また，content-based な評価と被験者による主観評価との結果を比較し，結果について検討した．

F-measure と pseudo-utility に基づく評価の比較では，要約システムの出力をいくつか調べたところ，正解には含まれていないが正解文と類似する内容の文をシステムが抽出した場合，pseudo-utility に基づく評価では評価値にそれが反映されていることが確認された．すなわち，pseudo-utility に基づく評価は，F-measure がかかえる 2 つの問題点のうち「(2) テキスト中に類似の内容を含む文が複数存在する場合，どちらの文が正解として選択されるかにより，システムの評価が大きく変化する」が解消できていることがわかった．

次に，content-based な評価と被験者による主観評価との比較の結果，2 つの要約が，content-based 値で 0.2 以上の開きがあれば，93%以上の割合で人間の主観評価の結果と一致することがわかった．

本研究では，複数の要約率のデータを用いることで，Radev らの提案する utility に基づく評価を疑似的に実現できることを示した．本研究は TSC で作成された 10%, 30%, 50%の 3 種類の要約データを用いたが，今後は，この他の要約率の組合せについても調べる必要がある．

また，本研究では pseudo-utility に基づく評価において，文の重要度を '1/要約率' として計

算したが，この他にも様々な重要度を設定することが可能である．重要度をどのように設定すればより良い評価が可能になるかについても調べる必要があると考えられる．

本研究では扱っていないが，Jingらの指摘する問題点 1(要約率の変化に伴う評価値の変化)を解消する評価方法についても今後検討していく必要がある．

参考文献

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

Fukushima, T. and Okumura, M. (2001). "Text Summarization Challenge Text Summarization Evaluation at NTCIR Workshop2.", *Proceedings of the Second NTCIR Workshop Meeting*, pp. 45–51.

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

Jing, H., Barzilay, R., McKeown, K., and Elhadad, M. (1998). "Summarization Evaluation Methods: Experiments and Analysis.", *Technical Report SS-98-06, Intelligent Text Summarization, AAAI Press*, pp. 51–59.

Mani, I. (2001). "Automatic Summarization.", *John Benjamins Pub Co.*

Mittal, V., Kantrowitz, M., Goldstein, J., and Carbonell, J. (1999). "Selecting Text Spans for Document Summaries: Heuristics and Metrics.", *Proceedings of the 16th National Conference on Artificial Intelligence*, pp. 467–473.

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.", 情報処理学会研究報告，NL-144, pp. 143–150.

Radev, D.R., Jing, H., and Budzikowska, M. (2000). "Centroid-base Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies.", *Proceedings of the ANLP/NAACL2000 Workshop on Automatic Summarization*, pp. 21–29.

Sparck-Jones, K. and Galliers, J. (1996). "Evaluating Natural Language Processing Systems: An Analysis and Review.", *Lecture Notes in Artificial Intelligence 1083*, Berlin: Springer.

略歴

難波 英嗣 (正会員): 1996年東京理科大学理工学部電気工学科卒業．1998年北陸先端科学技術大学院大学情報科学研究科 博士前期課程修了．2001年北陸先

端科学技術大学院大学情報科学研究科 博士後期課程修了．同年 4 月 日本学術振興会 特別研究員．2002 年 東京工業大学精密工学研究所 助手，現在に至る．博士 (情報科学)．自然言語処理，特にテキスト自動要約に関する研究に従事．情報処理学会，人工知能学会 ACL，ACM 各会員．nanba@pi.titech.ac.jp.

奥村 学 (正会員)： 1984 年東京工業大学工学部情報工学科 卒業．1989 年東京工業大学大学院理工学研究科 博士課程 修了．同年，東京工業大学工学部情報工学科 助手．1992 年北陸先端科学技術大学院大学情報科学研究科 助教授．2000 年東京工業大学精密工学研究所 助教授，現在に至る．工学博士．自然言語処理，知的情報提示技術，語学学習支援，語彙的知識獲得に関する研究に従事．情報処理学会，人工知能学会，AAAI，ACL，認知科学会，計量国語学会各会員．oku@pi.titech.ac.jp.

(2001 年 11 月 7 日 受付)
(2002 年 1 月 13 日 再受付)
(2002 年 4 月 5 日 採録)