複数の要約率の重要文データを用いた要約評価方法

難波 英嗣
日本学術振興会
特別研究員
nanba@lr.pi.titech.ac.jp

奥村 学
東京工業大学
精密工学研究所
oku@pi.titech.ac.jp

1 はじめに

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．これまでの要約の評価は，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 [2] 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」という問題点を指摘している．

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 [2, 4]．その一つに文の utility という概念を用いた Radev ら [4] の評価方法がある．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，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．

そこで，本研究では 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，[2]) 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)[1] 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．

以下では，まず，本研究で提案する評価方法について説明する．次に，F-measure と提案する評価方法を比較し，結果を報告する．

2 pseudo-utility に基づく評価

本節では，まず，Radev ら [4] の提案する utility に基づく評価について述べる．次に，本研究で提案する評価方法を説明する．

2.1 utility に基づく評価方法

Radev ら [4] 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．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，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．

2.2 utility に基づく評価方法の改善

utility に基づく評価は，より柔軟で自然な評価方法と言えるが，このような 10 段階での重要性 (適合性) 評価を複数の被験者がゆれなく一致して行なえるか，その作業負荷は大きくないかという問題が存在する．

そこで，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．

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

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 ($\frac{1/10+2/30}{1/10+2/30+2/50}$)

に，要約作成者と 2 つの要約システム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．

3 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．本節では，まず，3.1 節で TSC の課題および評価方法について説明する．次に，3.2 節で TSC のデータを用いた本研究の分析について述べる．

3.1 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，本節ではそのうち課題 A-1「重要文抽出型要約」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，[1, 3] を参照されたい．

課題 A-1 では，新聞 30 記事から，要約率 10%, 30%, 50%で重要文を抽出する．この 30 記事は毎日新聞 94 年および 98 年から 15 記事づつ選ばれている．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．

また，課題 A-1 では，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，TF と Lead の 2 種類を用いる．

3.2 分析

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

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)

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．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 5 および表 6 に示す[1] ．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 3 にシステム I と II の出力例を示す．表 3 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．

[1] 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．

表 3: 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ズに合わせ、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 4: 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 5: 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 6: 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

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 4)．

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　おわりに

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．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．

参考文献

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

[2] 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*, 51–59.

[3] 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, 143–150.

[4] 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*, 21–29.