2003－NL－158 (5)
2003／11／6

文書要約における抽出単位と評価法についての考察

賀沢秀人　Thomas Arrigan *　平尾努　前田英作
NTT コミュニケーション科学基礎研究所
〒619-0237 京都府相楽郡精華町光台 2-4
{kazawa,arrigan,hirao,maeda}@cslab.kecl.ntt.co.jp

概要

近年，自動要約研究の分野では，共通のデータセットを用いて互いの技術を比較的に評価する動きが盛んである．しかし，現状では主に人間の主観評価にもとづいて比較をおこなっているため，追加実験をおこなっても以前の結果と比較することが困難であるという問題点がある．そこで，本研究では，人間による要約とプーリングデータを用いて要約の自動評価を行う方法を提案し，疑似データによる精度評価を行った結果について報告する．実験の結果，提案手法は，従来用いられてきた正解要約との重複度にもとづく方法より，高精度な評価ができることがわかった．また，精度向上にはプーリングデータが重要な役割を果たすこともわかった．

キーワード: 自動要約，extract，自動評価，プーリングデータ，重複度

An Automatic Evaluation Method of Machine-generated Extracts

Hideto Kazawa, Thomas Arrigan, Tsutomu Hirao and Eisaku Maeda
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
{kazawa,arrigan,hirao,maeda}@cslab.kecl.ntt.co.jp

Abstract

Recently in the automatic summarization community, there have been many research initiatives in which machine-generated summaries are compared using the same documents. In most of these initiatives, the summaries are evaluated by humans and thus it is difficult to compare these results with new methods once the evaluation was finished. In this paper, we propose a new automatic evaluation method which combines the information from the human-created summaries and the machine-generated summaries evaluated in the past(pooling data). The experimental results show that out method is better than existing evaluation methods and the pooling data improves the performance significantly.

Keywords: automatic summarization, extract, automatic evaluation, pooling data, overlap measure

1 はじめに

近年，自動要約研究の分野では，DUC*1やTSC*2のように，共通のデータセットを用いて，互いの技術を比較的に評価する動きが盛んである．しかし，これらの会議では，主に人間の主観評価にもとづいて比較をおこなっている．そのため，追加実験をおこなっても以前の結果と比較することは困難であり，蓄積されたデータ（プーリングデータ）が有効に活用できないという問題点がある．また，評価者にとっても大量の評価作業は大きな負担となり，一貫性などに問題が生じる可能性が高くなる．このような状況を鑑みると，何らかの自動評価法を確立することは自動要約研究にとって，重要な課題である．

要約の自動評価としては，人間が作成した要約との重複度（意味的なものではなく表層的なもの）が用いられる場合が多い [1]．これは，（人間の作った）「良い」要約と似た要約は「良い」，という自然な感覚に基づいたものであり，ある程度

* 本研究は第二著者がNTTにて研修中におこなったものである．現在の所属は Department of Engineering, Trinity College Dublin.

*1 Document Understanding Conference

*2 Text Summarization Challenge

有効な指標と考えられる．しかし，実際に，主観評価と重複度の間にどの程度定量的な相関があるのかは明らかではない．

一方，良いものと似たものは良い，という発想を一般化すると，評価したい要約と似たプーリングデータがあった場合，そのプーリングデータの評価から，要約の質を推定できるのではないか，というアイデアにつながる．しかし，今のところプーリングデータを用いて要約の自動評価を行った研究は，筆者の知る限り見当たらない．

そこで，本研究では，人間による要約とプーリングデータの両者との重複度にもとづいて要約の自動評価を行う方法を提案し，実際どの程度の予測が可能か，また，プーリングデータはどの程度予測に有効なのか，について検討する．なお，本研究では「重複度」が客観的に定まるように，何らかの言語単位を抽出し並べることで要約とする手法を対象とする自動評価法について考察する．したがって，以下で言う「要約」は extract を指す．

本稿の構成は次の通りである．まず，第 2 章で，人間による要約とプーリングデータを用いた評価関数を提案する．次に，第 3 章において，評価関数がどの程度人間による評価を近似しているかを測定する方法について説明する．第 4 章では，提案した評価関数について，疑似プーリングデータを用いた実験にもとづき，予測精度やプーリングデータの有効性についての議論を行う．最後に，第 5 章でまとめと今後の方向性について述べる．

2 評価関数

以下では，自動評価のもととなるデータとして，m 個の文書の各々にたいして n 個の要約が与えられていると仮定し，i 番目の文書にたいする j 番目の要約を x_{ij}，その人間による評価を y_{ij} と書くことにする．ただしここでの「要約」は人間が作成したものと自動で作成されたもの（プーリングデータ）の両方を含み，評価 y_{ij} については適当に数値化されているものとする．また，j 番目の要約は，常にある一定の人物もしくは要約手法によって作成されたものとする．

以上の記法のもとで，本研究では，次の要約評価関数 $\mathrm{scr}(x)$ を考える．

$$\mathrm{scr}(x) = \sum_{i=1}^{m}\sum_{j=1}^{n} w_j y_{ij} \frac{|x_{ij} \cap x|}{\min(|x_{ij}|, |x|)} + b \quad (1)$$

ここで $s \cap t$ は要約 s と t に共通して含まれる言語単位の集合，$|s|$ は s に含まれる言語単位の数をあらわす．

式 (1) は，既知の要約データ x_{ij} との（正規化された）重複度にその評価 y_{ij} と各要約手法の信頼性 w_j をかけたものを，各データから与えられる部分点として考え，その総和で未知の要約 x を評価するというアイデアに基づいたものである．また，x_{ij} を人間の要約に限定し，かつ w_j を一定値とすると，scr は，単に人間の要約結果との重複度を平均したものと一致する．したがって，式 (1) は，従来用いられてきた重複度に基づく評価関数を，プーリングデータと手法の信頼性を考慮するように拡張したものになっている．

w_j, b の決定法については様々な方法が考えられるが，本研究では，$\mathrm{scr}(x_{ij})$ と y_{ij} の二乗差の総和が最小になるように最適化するものとする．これにともない，以下では $(x_{ij}, y_{ij})(1 \leq i \leq m, 1 \leq j \leq n)$ を訓練データと呼ぶことがある．

3 近似精度の測定

一般に，人間の評価は順序尺度であることが多く，それを数値化した y は $1, 2, 3, \ldots$ のような離散値を取ることが普通である．それにたいして，scr の出力は実数値であるため y と直接比較することはできない．そこで，本研究では scr と y がどれくらい近いか（近似精度）を測定するために，修正残差と順序誤り率の二つを提案する．

3.1 修正残差

実際に評価したい要約を $x'_1, x'_2, \ldots, x'_k$，真の評価結果を $y'_1, y'_2, \ldots, y'_k$（これらをテストデータと呼ぶ）としたとき，評価関数 scr の修正残差 MR を次の式で定義する．

$$\mathrm{MR} = \min_{\lambda,\mu} \frac{1}{k} \sum_{i=1}^{k} |y'_j - \lambda \mathrm{scr}(x'_i) - \mu|^2 \quad (2)$$

式 (2) では，$\mathrm{scr}(x'_i)$ と y'_i の残差が最小になるように，scr のスケール（λ）とオフセット（μ）を修正した上で，その修正残差を近似精度としている．これは，実際に評価関数に用いる際には，

スケールやオフセットの違いは重要な意味を持たないため，あらかじめ最も都合良く修正した上で比較するのが妥当である，という考えに基づいている．

修正残差は，評価関数が真の評価値から大きく離れた値を取ることにたいして，大きなペナルティを与える指標である．

3.2 順序誤り率

評価関数 scr の順序誤り率 OE を次の式で定義する．

$$\mathrm{OE} = \frac{1}{N} \times \sum_{1 \le i < j \le k} \theta(-(y'_i - y'_j)(\mathrm{scr}(x'_i) - \mathrm{scr}(x'_j))) \quad (3)$$

ただし N は $i < j$ で $y'_i \neq y'_j$ となる i, j の組合わせの数であり，$\theta(\cdot)$ は引数が正の時 1，それ以外で 0 となる階段関数である．

順序誤り率は純粋に順序関係のみに依存し，評価関数の値自体の大きさには依存しない指標である．

4 実験と考察

4.1 実験データの準備

実験には Linguistic Data Consortium から発行されている RST Discourse Treebank[2] (以下，RSTtrees) を使用した．RSTtress には retorical structure theory[3] にもとづいた修辞構造がアノテーションされた Wall Street Journal の記事が 385 含まれており，そのうち 150 記事について，二人の人間が作成した discourse unit(DU) 単位の extract がある．今回，その 150 記事のうちランダムに取り出した 38 文書 (表 1) を使用し，次の手順で人間による要約データと疑似プーリングデータを作成した．

人間による要約データ (ha, hb, hc)

1. 各文書にたいして，英語のネイティブスピーカーが作成した要約 (ha)
2. RST corpus に含まれる，人間による要約 2 種 (hb, hc)

文書名	#DU	#EXT
file1	87	10
wsj_0601	75	9
wsj_0609	185	12
wsj_0624	24	6
wsj_0632	62	9
wsj_0664	196	15
wsj_0677	71	9
wsj_0689	132	12
wsj_1101	88	10
wsj_1107	51	11
wsj_1120	136	12
wsj_1125	57	8
wsj_1137	141	14
wsj_1146	304	17
wsj_1156	57	8
wsj_1160	96	10
wsj_1171	114	12
wsj_1187	124	11
wsj_1193	73	10
wsj_1305	115	13
wsj_1314	141	12
wsj_1320	156	12
wsj_1330	102	12
wsj_1349	83	10
wsj_1367	202	13
wsj_1375	157	13
wsj_1387	134	12
wsj_1391	75	12
wsj_1970	88	10
wsj_1998	19	5
wsj_2309	50	8
wsj_2317	34	6
wsj_2325	79	8
wsj_2338	87	10
wsj_2347	90	8
wsj_2360	85	11
wsj_2366	119	12
wsj_2386	127	13

表 1: 実験に用いた文書．#DU は文書中の discourse unit(DU) の総数，#EXT は extract に含まれる DU の総数．

疑似プーリングデータ (lead, tfidf, clead, ctfidf, crand1〜6)

以下では，ha，hb，hc に含まれる DU からなる集合を cheating set と呼ぶ．

1. 文書の先頭から DU を抜き出したもの (lead)
2. TF*IDF 値の大きなものから順に DU を抜き出したもの (tfidf) IDF 値は Penn Tree Bank 全てを用いて計算した．
3. cheating set から，文書中の出現位置が早いものから順に DU を取り出したもの (clead)
4. cheating set から，TF*IDF 値の大きなものから順に DU を抜き出したもの (ctfidf)
5. cheating set から，ランダムに DU を抜き出したもの×6．(crand1〜6)

次に，各要約データの内容について，ha の作成者が次の三段階で評価した．(以下は，実際に作業者に与えた指示)

Good

Extract contains all the important points from the important points from the document. Content is similar to human-created extracts in quality.

Average

Some details are missing, but the reader can grasp the main points of the article just by reading the extract.

Bad

The extract is missing important details. It is not possible to say what the article is about from reading the extract.

作成したデータについての評価結果を表 2 に示す．

4.2 疑似プーリングデータの妥当性

ここでは，実験で用いた疑似プーリングデータが，実際のプーリングデータの代替物として妥当かどうか調べるために，

1. 疑似データが不自然に良い要約となっていないか
2. 疑似データは不自然に似通っていないか

の二点について考察する．

要約データ	Good	Average	Bad
ha	89%	11%	0%
hb	89%	11%	0%
hc	79%	21%	0%
clead	76%	23%	0%
ctfidf	50%	45%	5%
lead	26%	71%	3%
crand × 6	32%	45%	22%
tfidf	13%	37%	50%

表 2: 実験データの評価 (crand は crand1〜6 の平均)

	tfidf	clead	ctfidf	crand
lead	0.09	0.57	0.28	0.29
tfidf		0.15	0.32	0.17
clead			0.51	0.49
ctfidf				0.50
crand				0.48*

表 3: 疑似データ間の重複度 (*は自分自身以外の crand との重複度の平均)

疑似データは不自然に「良く」ないか

表 2 からわかるように，crand は lead と tfidf の中間ぐらいの評価を得ている．一方，実験で用いた文書 (新聞記事) にたいしては lead が比較的有効な手法であり，現時点での自動要約手法が lead の評価を大きく超えることは少ない．したがって，crand の評価は実際の自動要約の評価と comparable なものと考えられる．

一方，clead と ctfidf の評価は lead よりもやや高めになっており，既存手法の代替物としては若干不適切と思われる．しかし，疑似データ中の割合が少ないこと (10 データ中の 2 データ)，将来的に要約精度が上がった場合に到達不可能な評価 (特に ctfidf) とは言えないことから，疑似データで全体としては，それほど不自然な「良さ」ではないと考える．

疑似データは不自然に「似通って」ないか

疑似データ同士の重複度を次の式で計算した結果を表 3 に示す．

$$\frac{|x \cap y|}{\min(|x|, |y|)} \tag{4}$$

「自然」な重複度を一般的に定義するのは難しいため，ここでは一つの目安として次のような例

人手＼プーリング	無し	P3	P5
無し	0.532	0.474	**0.463**
H1	0.497	0.475	0.467
H2	0.489	0.477	0.470

表 4: 平均修正残差

人手＼プーリング	無し	P3	P5
無し	1.000	0.358	**0.343**
H1	0.372	0.363	0.352
H2	0.376	0.367	0.358

表 5: 平均順序誤り率

	P5 H1	P5 H2	P3	P3 H1	P3 H2	H1	H2
P5	++	++	++	++	++	++	++
P5H1		++	++	++	++	++	++
P5H2			00	++	++	++	++
P3				++	++	++	++
P3H1					++	++	++
P3H2						+0	++
H1							−+

表 6: Wilcoxon signed rank test の結果

を考える.

まず，新聞記事のようなデータでは，現状の自動要約で，人間の作った要約（正解要約）の約60%程度が抽出できることが知られている [4]. そこで，仮に抽出率 60%の手法が 2 つ存在し独立に要約を作成したとする．すると，互いに相関がなくランダムに抽出を行ったとしても，正解要約中の 36%が共通に抽出されることになる．実際には，要約手法間に全く相関がないことは稀であり，共通に抽出される部分はこれよりも多い．また，正解要約以外での重複もあることを考慮すると，表 3 の重複度は不自然に高いとは言えない.

4.3 近似精度の評価

準備した実験データから，次のように 6 種類の訓練データと 1 種類のテストデータを作成した.

1. 疑似データ 10 種 (lead, tfidf, clead, ctfidf, crand×6) の中から，ランダムに取り出した 5 種と hc を合せてテストデータ T とする.
2. 残りの疑似データ 5 種を訓練データ P5 とする．また P5 からランダムに 3 種取り出したものを，訓練データ P3 とする.
3. 訓練データ P3，P5 に，ha を加えたものを訓練データ P3H1, P5H1 とする．さらに P3H1, P5H1 に hb を加えたものを訓練データ P3H2, P5H2 とする.

以上の訓練・テストデータの組を 1 セットとして，計１００セットのデータを作成した.

実験では，最初に，各訓練データについて scr(x) と y の二乗差の総和が最小になるように，式 (1) 中の w_j を決定した．次に，テストデータ T を用いて，修正残差（式 (2)）および順序誤り率（式 (3)）を計算した．また，比較のため，各テストデータについて ha（要約評価者）の作成した extract との重複度と，ha, hb の作成した extract との平均重複度をスコアとしたときの修正残差および順序誤り率も計算した.

修正残差および順序誤り率の１００セットについての平均値を，それぞれ表 4，表 5 に示す．各表において「プーリング無し」に相当する部分は，上から順番に，「単に一定の値を返す関数（式 (1) で b のみ残した関数)」「ha との重複度」「ha,hb との平均重複度」でスコアリングしたときの値である．また，太字は各表において最も精度が高かった（残差，誤り率が小さかった）組合せである.

次に，表 4, 5 に現れた精度の差が，統計的に有意であるか調べるために，１００セットのテスト結果を対にして，Wilcoxon signed ranks test[5] を行った（表 6）．表において，左側の文字は修正残差に対する結果，右側は誤り誤差に対する結果をあらわし，「＋」は行側の手法が列側の手法が５％水準で有意に精度が高く，逆に「−」は有意に低いことを示す．また「0」は統計的に有意な差が認められなかったことを示す.

表 4，5，6 から次のことがわかる.

1. プーリングデータを利用することで，単に人間の要約結果との重複度を測るよりも，高精度な評価ができる.
2. 近似精度に最も大きく影響するのは，訓練データ中のプーリングデータの数であり，その数が大きいほど精度は高くなる.
3. 人間の要約を訓練データに含めても近似精度は向上せず，逆に僅かではあるが精度が低下する.

一般に，訓練データが増えれば精度は向上するため，1,2 番目の結果は自然である．一方，3 番目

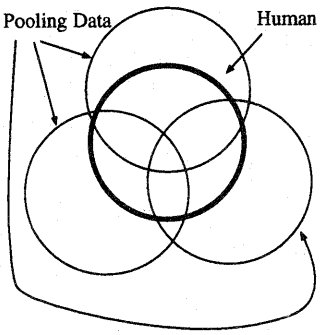

図 1: 人間の要約データの影響

の結果に関しては次のような説明が考えられる．

一般に，十分多くのプーリングデータが集まれば，正解要約 (ha) に含まれる DU が，そのプーリングデータのどれかに含まれる可能性は高いと考えられる（図 1）．したがって，そのようなデータに正解要約を追加することは，式 (1) に冗長な次元を追加することになる．一方，冗長な次元は，パラメータ (w_j) の推定結果におけるばらつきを増やし，過学習などの問題を起こすことが知られている．そのため，今回の実験においても，人間の要約結果をデータを追加することは，冗長な次元を追加することになり，むしろ精度精度を下げる結果になっていると考えられる．

5 まとめ：自動要約テストコレクションに向けて

実際に自動要約の研究を行うと，大量の要約結果にたいして評価を行いたくなることが多い．しかし，実際には同一の評価者をいつも用意することは困難である．さらに，もし可能だとしても評価者への負担を考えると，望むときに好きなだけ評価を行うことは現実には不可能である．

また，言うまでもなく研究を健全に進めるためには，過去の研究との比較はかかせない要素である．しかし，人間による評価しか比較手法がないとすると，事実上，過去の研究との比較は不可能ということになる．

このような状況を鑑みると，たとえ１００％の精度でなくても，ある程度の精度を持った自動評価法を確立することは，自動要約研究を推進する上で非常に重要である．

本研究の結果は，プーリングデータの活用により，従来用いられてきた手法（正解要約との重複度）よりも，有意に高精度な評価関数が構成できることを示している．また，実験に対する考察から，従来自動評価に用いられてきた正確要約だけではなく，できるだけ多様な自動要約にたいする評価結果を集めることが，有用なプーリングデータを作る鍵であることも示唆された．

情報検索の分野においては，TREC のプーリングデータが事実上の標準テストコレクションとして，技術の発展に多いに役立っている．TREC データには，過去に参加した（必ずしも優秀とは限らない）システムの結果が混在しており，それがデータとしての価値を高めている．同様に，もし自動要約における「TREC データ」が存在すれば，非常に有益なものになると考えられる．

本研究の結果は疑似データに基づいたものであり，実際のプーリングデータでも同様の結果が得られるかは，確認が必要である．また，予測精度に関しても決して十分とは言えないため，その点についても検討が必要である．今後は，以上のような点を考慮しつつ，自動要約テストコレクション構築に向けた研究を行っていく予定である．

参考文献

[1] 平尾努，賀沢秀人，磯崎秀樹，前田英作，松本裕治．機械学習による複数文書からの重要文抽出．自然言語処理，vol.10，no.1，pp.81–107．(2002)

[2] Rhetorical Structure Theory Discourse Treebank. www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T07. Linguistic Data Consortium. (2002)

[3] W. C. Mann, S. A. Thompson. "Rhetorical Structure Theory: A Theory of Text Organization." ISI Reprint Series, ISI/RS-8-190. Univ. of South Calfornia. (1987)

[4] 賀沢秀人，平尾努，前田英作．Order SVM: 一般化順序統計量に基づく順位付け関数の推定．信学論 vol.J86-D-II, no.7, pp.926–933. (2002)

[5] S. Siegel, N. J. Castellan, Jr. Nonparametric Statistics Second Edition, pp.87–94. McGraw-Hill, Inc. (1998)