原文からの抜粋度合を考慮した要約の自動評価法

平原一帆[1]　難波英嗣[2]　竹澤寿幸[2]　奥村学[3]　平尾努[4]
1.広島市立大学　情報科学部
2.広島市立大学大学院　情報科学研究科
3.東京工業大学　精密工学研究所
4.NTT コミュニケーション科学基礎研究所

1　はじめに

近年，ウェブページの検索結果として表示されるスニペットや，インターネットで配信されるニュースの要約など，電子化された文書の要約を求められる場面が増えている．このような状況にあって要約の自動生成の研究が活発化する一方，自動生成される要約を評価する手間やコストが問題となっている．人間の手による評価（以下，マニュアル評価）は正確である反面，時間，金銭的コストが多大にかかる上に，評価を繰り返し行うことが困難である．こうしたことを背景として，自動生成されるテキスト要約の評価もまた，自動化されることが求められてきた．

近年のテキスト要約研究は，テキスト内の重要箇所を抽出するものから，テキストから独自の表現を含む，テキスト要約を生成するものへと主流が移行しつつある．これまで提案されてきた自動評価手法は，抽出に基づく要約を評価するために，精度や再現率といった尺度を用いて，人間が作成した要約（以下，参照要約）と，コンピュータの作成した要約（以下，システム要約）の一致度を測る手法が一般的であり，単文，単語列，単語など，様々な言語単位で比較を行う手法が提案されている．(平尾, 2006, Hovy, 2006, Lin, 2003)

しかし，このような従来の自動評価手法では，独自の表現を含み，人の手によって書かれたものにより近い生成に基づく要約に対しては，抜粋に基づく要約に対する評価ほど十分な精度が得られないことが，今回我々の行った実験から分かった．本研究では，原文からの抜粋度合を考慮して要約を自動評価することで，従来の自動評価手法の問題点を改善する手法を提案する．

本論文の構成は以下の通りである．次節では，従来の自動評価手法および評価結果を示し，その問題点を指摘する．3 節では，2 節で指摘した問題点を改善する抜粋度合を考慮した評価法と，その有効性を示す実験結果について述べる．4 節で本稿をまとめる．

2　従来の自動評価手法とその問題点

2.1　従来の自動評価手法

従来の自動評価手法として，参照要約との類似性による自動評価手法について説明する．この手法は，参照要約とシステム要約との間の一種の類似度を計算するものであり，参照要約との類似度が高いほどより良い要約であるという考えに基づく．以下に，代表的な評価手法である BLEU と ROUGE について説明する．

■BLEU　(Papineni et al., 2001)

BLEU は，機械翻訳の評価尺度として開発された自動評価手法であり，要約の自動評価のための尺度としても注目を集めた．BLEU はシステム要約と一つ以上の参照要約とを比較し，システム要約中の N グラム（単語ベース）(Lin and Hovy, 2003)が参照要約中にどの程度出現するかを，精度 P を用いて測定する．ここで，一度マッチングしたものは再度マッチングをしないという補正と，システム要約が極端に短い場合の補正を行う．r を参照要約の長さ，c をシステム要約の長さとして，以下の補正項を用いる．

$$BP = \begin{cases} 1 & if(c > r) \\ e^{(1-r/c)} & if(c \le r) \end{cases}$$

ユニグラムにおける単語の一致度を精度 P_1，バイグラムを P_2，…，n-gram を P_n とし，最終的に BLEU は以下の式を用いて計算する．

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log P_n\right)$$

なお，要約の評価尺度として BLEU を使う問題として，

- 要約評価の場合，再現率が重要となるため，BLEU のような精度による評価は馴染まない．
- 要約はできるだけ短いほうが望ましいため，要約が短い場合にかかる補正項 BP は要約評価には適さない．

などの点が挙げられている．これらの問題点を要約評価用に改良したものとして，ROUGE という尺度が Lin により提案されている．(Lin, 2004)

■ROUGE-N (Lin, 2004)

現在，要約システムの自動評価法として最も広く用いられている．参照要約と，システム要約の間で一致するNグラムの割合を以下の式を用いて計算する．

$$ROUGE(C,R)=\frac{\sum_{e\in n-gram(C)} Count_{clip}(e)}{\sum_{e\in n-gram(R)} Count(e)}$$

n-gram(C)は，システム要約に含まれるNグラム，n-gram(R)は，参照要約に含まれるNグラム集合をあらわす．Count(e)は，あるNグラムの出現頻度を数える関数であり，$Count_{clip}$(e)は，システム要約に含まれるNグラムのシステム要約における出現頻度 Count(e∈n-gram(C))と参照要約における出現頻度 Count(e∈n-gram(R))の小さいほうの値を採用する．Linらは，Nを1～4まで変化させ，マニュアル評価結果との相関を調べた結果，N＝1, 2が最も高い相関であったと報告している．

■ROUGE-S (Lin, 2004)

ROUGE-Nは，隣接という強い制約にある単語共起のみに着目し，スコアを計算する．よって，隣接してないが，係り受け関係にあるような単語の共起を考慮することができない．このような問題を解決するため，Linらは，以下の式で定義されるスキップを許したバイグラムを考慮して一致率を計算するROUGE-Sを提案している．

$$ROUGE-S(C,R)=\frac{(1+\gamma^2)Ps(C,R)Rs(C,R)}{Rs(C,R)+\gamma^2 Ps(C,R)}$$

ここで，Ps(C,R)，Rs(C,R)は，それぞれ以下の式で定義される．

$$Ps(C,R)=\frac{\sum_{e\in bigram(C)\cup skip-bigram(C)} Count_{clip}(e)}{\sum_{e\in bigram(C)\cup skip-bigram(C)} Count(e)}$$

$$Rs(C,R)=\frac{\sum_{e\in bigram(C)\cup skip-bigram(C)} Count_{clip}(e)}{\sum_{e\in bigram(R)\cup skip-bigram(R)} Count(e)}$$

■ROUGE-SU (Lin, 2004)

さらに，ROUGE-Sに対しユニグラムを素性として追加した，ROUGE-SUも提案されている．

ただし，ROUGE-S，ROUGE-SUでは，スキップを許したトライグラム(3-gram)を扱うことはできない．さらに単語の組み合わせが参照要約，システム要約のどちらか一方ではバイグラム，もう一方ではスキップバイグラムとして出現した場合，スキップの有無を区別せずに一致数を計算するという問題がある．さらに，ROUGE-Nも含め，単語の表記での一致しか見ておらず，単語の言い換えがあった場合には一致数が著しく低下するという問題もある．

2.2 従来の自動評価手法による評価

■実験データ

本研究では株式会社リクルートマネジメントソリューションズから提供された，以下の手順で作成したデータを実験に用いた．このデータは「明治時代の日本における現代化」に関する10文から成るテキストについて，問題文の最後にある「近代化」について100字以内の一文で説明せよという小論文の課題である．

この課題に対して，評価の基準としてあらかじめ複数名の評定者らによる議論の後，採点基準の擦り合わせが行われ，さらに評定者により，模範解答8編が作成されている．採点基準に則って，全回答のうち109回答を評価者らが質を判断し，A（申し分ない）～E（問題外）の五段階に評価済みである．一つの回答には，評価者によるひとつの評価がつけられている．全109の回答の内，A：32，B：10，C：7，D：18，E：42の割合で五段階評価されている．これを二段階で評価するために，五段階でのマニュアル評価のうちA，B，Cを「優」(49回答)，DとEを「劣」(60回答)として二段階に分けた．

■評価尺度

二段階評価済みのマニュアル評価結果と，自動評価による二段階評価との一致の割合である一致率を指標とする．

■自動評価手法比較実験

各自動評価手法に対して，自動評価結果とマニュアル評価結果との一致率を求める．自動評価スコアは，複数の模範解答と被験者の回答とを比較して算出したスコアのうち，最も高かったものを選んだ．また，自動評価スコアの二段階判定は，一致率が最良になる閾値を選択している．自動評価とマニュアル評価の一致率を表1に示す．

表 1　従来の自動評価手法別の一致率

自動評価手法	一致率
BLEU	0.872
ROUGE-1	0.872
ROUGE-2	**0.899**
ROUGE-3	0.872
ROUGE-4	0.872
ROUGE-S1	0.743
ROUGE-S4	0.734
ROUGE-S9	0.734
ROUGE-SU1	0.881
ROUGE-SU4	0.862
ROUGE-SU9	0.826

ROUGE-N の評価結果が比較的高く，また，一致率が最も高かったのはROUGE-2 で, これはLin らの報告と一致する．(Lin, 2004)

■不一致回答の分析

従来の手法で評価した被験者の回答と，参照要約との一致を求めた際に，マニュアル評価と自動評価結果が不一致だった回答について分析を行った．以下に分析結果をまとめる．

- 自動評価手法で結果が「優」でマニュアル評価が「劣」の場合：正答例と一致する文章を抜粋して回答に使用している部分が多いが, 文末や係り受けが, 正解とは反対の意味を持つなど, 回答として題意にそぐわなくない.
- 自動評価手法で結果が「劣」でマニュアル評価結果が「優」の場合：参照要約との語の一致度により, 被験者の要約のスコアを算出しているため, 内容的には正答に近いが, 独自の表現を含んだ自由作成で回答しているために参照要約との語の一致度が低い.

■従来の自動評価手法の問題点

要約の自動評価に対して一般的に用いられるこれらの指標はいずれも，参照要約と被験者の要約との単語列の重なりを調べることで，その精度や再現率を求める．しかし，人手により作成された要約や，近年のテキスト要約研究で生成される要約は, 原文にない独自の表現を用いたものである．これまでの自動評価手法は，言い換えなどの独自の表現を含む生成に基づく要約に対しては，参照要約の単語と被験者の要約の単語との一致が悪いため，十分な精度が得られないと考えられる．

3　原文からの抜粋度合の考慮を利用した，従来自動評価手法の改善

3.1　原文からの抜粋度の考慮

参照要約との類似性による自動評価手法は，コストのかからない評価手法であるが，評価に対する信頼性は十分に高いとは言えず，マニュアル評価の代替として利用するまでには至っていない．本研究は，人間による評価作業の一部を，自動評価が担うことにより，評価にかかる負荷を軽減しようと言うものである．

前節の実験結果から, 従来の自動評価手法では, 抜粋に基づく要約は十分高い精度で評価できるが, 独自の表現を多く含む要約の評価に対しては必ずしも十分な精度で評価できるとは限らない，ということが分かっている．従って，あらかじめ被験者の回答が抜粋に基づいて作成されたものであるか, 独自の表現を多く用いたものであるかを調べ, 抜粋に基づいて作成された要約にのみ自動評価手法を適用することにより，部分的にマニュアル評価の代替とすることができる．

被験者による要約の抜粋度合いを調べるため，要約を作成する前の原文と被験者の要約との比較を行う．原文からの抜粋度合いを測る尺度として，ROUGE-4 と，DP マッチング（動的計画法による文字列類似度の計算)を用いた．

•ROUGE-4

2 節で前述した ROUGE-N における，4 グラムを使用したものである．原文と被験者の要約とを比較した際に，4 グラムの長い単語列の一致が多く見られる回答というのは，原文からの抜粋を用いて要約を作成していると考えられる．

•DP マッチング（Needleman et al., 1970）

動的計画法を用いた，ペナルティを課すやり方で点数付けを行い, 文字列の類似度を求める方法.

抜粋度合の高いものに対して従来の自動評価手法を適用することで，その範囲において正確な結果を求めることができる．部分的に，十分な精度での自動評価が可能になる．

3.2　抜粋度合別比較実験

■原文からの抜粋度合を考慮した自動評価

抜粋度合を考慮した自動評価結果とマニュアル評価結果との一致率を示す．本手法は，ベースラインの結果が良かった ROUGE-1，ROUGE-2 に対して適用した．被験者の要約全体に対する一致率から，抜粋度合の低いものを除外し，徐々に抜

粋度合の高い要約だけを残し，その一致率の変化を調べた．ROUGE-1，ROUGE-2 に対する抜粋度合別一致率を表 3.1，表 3.2 にそれぞれ示す．

表 2　ROUGE-1 における抜粋度合別一致率

上位%	DP	ROUGE-4
100%	0.872 (95/109)	0.872 (95/109)
上位 90%	0.880 (88/100)	0.869 (86/99)
上位 80%	0.899 (80/89)	0.852 (75/88)
上位 70%	0.897 (70/78)	0.844 (65/77)
上位 60%	0.910 (61/67)	0.833 (55/66)
上位 50%	0.911 (51/56)	0.836 (46/55)
上位 40%	0.911 (41/45)	0.886 (39/44)
上位 30%	0.941 (32/34)	0.970 (32/33)
上位 20%	0.913 (21/23)	**1.000 (22/22)**
上位 10%	**1.000 (12/12)**	**1.000 (12/12)**

表 3　ROUGE-2 における抜粋度合別一致率

上位%	DP	ROUGE-4
100%	0.899 (98/109)	0.899 (98/109)
上位 90%	0.890 (89/100)	0.909 (90/99)
上位 80%	0.899 (80/89)	0.920 (81/88)
上位 70%	0.910 (71/78)	0.922 (71/77)
上位 60%	0.896 (60/67)	0.924 (61/66)
上位 50%	0.893 (50/56)	0.909 (50/55)
上位 40%	0.867 (39/45)	0.909 (40/44)
上位 30%	0.853 (29/34)	0.909 (30/33)
上位 20%	0.826 (19/23)	0.909 (20/22)
上位 10%	0.917 (11/12)	**1.000 (11/11)**

各表から，特に ROUGE-1 において，抜粋度合が高い回答の割合を増加させると，自動評価とマニュアル評価の結果が一致しやすくなるということが読み取れる．ROUGE-2 については，一概に抜粋度合の高まりが一致率を高めるとは言えないが，ごく上位の抜粋度合に対しては同様に，一致が向上することがわかる．抜粋度合が高いもののみを評価対象とすることで，自動評価手法の一致度を上げることができると考えられる．

4　おわりに

本研究では，従来手法の問題点を指摘し，原文からの抜粋度合をあらかじめ考慮することで，従来手法を改善する手法を提案した．また，この提案手法を用いた実験により，抜粋度合の高い回答に対して自動評価の精度が向上する傾向があることを示した．特に，ごく抜粋度合の高い上位 10%～20%の回答に対して，マニュアル評価と自動評価との評価が極めて高い割合で一致することを示し、本提案手法の有効性を確認した．

謝辞

本研究で用いた要約データを提供していただいた，株式会社リクルートマネジメントソリューションズの鷺坂由紀子氏と入江崇介氏に深く感謝いたします．

参考文献

平尾 努，奥村 学，磯崎秀樹 (2006). 拡張ストリングカーネルを用いた要約システムの自動評価法，情報処理学会論文誌, Vol.47, No.6, pp.1753-1766

Hovy, E., Lin. C., Zhou, L. and Fukumoto, J. (2006). Automated summarization evaluation with basic elements, Proc. 5th Conference on Language Resources and Evaluation.

Lin, C. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. Proceedings of the ACL-04 Workshop “Text Summarization Branches Out”, pp.74–81.

Lin, C. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics, Proc. 4th Meeting of the North American Chapter of the Association for Computational Linguistics and Human Language Technology, pp.150-157.

Needleman, S. B. and Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. Journal of Molecular Biology, Vol. 48, pp. 443–453.

Papineni, K., S. Roukos, T. Ward, W.-J. Zhu. (2001). BLEU: a Method for Automatic Evaluation of Machine Translation, IBM Research Report, RC22176 (W0109-022).