複数のオンライン評価結果を利用した要約のオフライン評価法の提案と分析

難波英嗣† 奥村学‡
†広島市立大学 情報科学部
‡東京工業大学 精密工学研究所

1. はじめに

自動要約の研究分野では，コンピュータが生成した要約を評価するために数多くの方法が提案されている．それらは，人間の被験者が評価するオンライン評価法と自動評価であるオフライン評価法の大きく 2 つに分けられる．

オンライン評価法は，被験者自身が評価を行うので正確である反面，評価に非常にコストがかかるという問題がある. 他方, オフライン評価法は，ツール等を用いて評価するため，コストが小さくいつでも評価を行えるという利点があるが，オンライン評価と同等の正確さで評価するのは，現状ではかなり難しい．そこで，本研究では，両者の問題点を補う新しい評価方法について検討する．

機械翻訳の分野では，Yasuda ら[Yasuda 2003]は，人手で作成した翻訳結果をプーリングしておき，それらを用いて機械翻訳結果を自動評価する手法を提案している. 他方, 自動要約の分野では，賀沢ら[賀沢 2003]は複数の要約とそのオンライン評価結果をプーリングしておき，それらを用いて要約のオフライン評価を行う手法(以下，賀沢手法)を提案している．この手法の基本的な考え方は，「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，要約の質を推定する」というものである．

厳密な定義式は次節で述べるが，賀沢手法の大まかな評価手順は，まず評価したい要約とプーリングデータ(複数の要約)との類似度を計算し, 次にそれぞれのプーリングデータのオンライン評価値に評価したい要約との類似度に応じた重みをかけ, 最後にそれらをすべて足し合わせることで，評価したい要約の評価値を算出する．

しかし，評価値の算出にすべてのプーリングデータを用いると，適切な評価ができなくなる場合がありうる．例えば，プーリングデータの中に評価したい要約と全く同一のデータが存在していても，すべてのプーリングデータを評価値の算出に用いると，たとえプーリングデータとの類似度を考慮するにしても，他のデータがノイズとして作用し，適切な評価値が得られない．また，プールされるデータの数が増えるほど，その影響は大きくなると予想される．

もし，評価したい要約と酷似したものがプーリングデータの中にある場合には，それらデータの評価値のみを使って算出した方が，プーリングデータすべてを用いるよりも，適切な評価ができると考えられる．そこで，本研究では，プーリングデータの中で評価に用いるデータの数を制限することで，賀沢手法の有効性がどの程度変化するかについて調べる．

本論文の構成は以下のとおりである．次節では賀沢手法を中心に，要約評価に関する関連研究について述べる．3 節では，プーリングデータの数を制限した場合の賀沢手法の有効性を調べる分析手法について説明する．4 節では，分析結果について報告し，その結果を考察する．5 節ではまとめ，6 節では今後の課題について述べる．

2. 関連研究

前節で述べた賀沢手法によれば，まず自動評価のもととなるデータとして，m個の文書の各々に対してn個の要約が与えられていると仮定し，i番目の文書に対するj番目の要約をx_{ij}，そのオンライン評価値をy_{ij}と表記する．また，j番目の要約は，常に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．このとき，評価したい要約xの評価関数scr(x)は以下に示す式(1)と(2)で定義される．

$$scr(x) = \sum_{i=1}^{m}\sum_{j=1}^{n} w_j y_{ij} Sim(x, x_{ij}) + b \qquad (1)$$

$$Sim(x, x_{ij}) = \frac{|\, x_{ij} \cap x \,|}{\min(|\, x_{ij} \,|, |\, x \,|)} \qquad (2)$$

ここで，$x_{ij} \cap x$はx_{ij}とxに共通して含まれる単語の集合，$|\,x\,|$はxに含まれる単語の数をあらわす．式(1)は，xと既知の要約データx_{ij}との(正規化された)重複度(式(2))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．なお，式(1)において，w_jとbを決定する方法には様々なものが考えられるが，賀沢らは，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している．

さて，式(1)を用いて評価する際，プーリングする要約データx_{ij}をあらかじめ評価しておく必要がある．賀沢らは，被験者に要約x_{ij}を，品質に応じて”Good,” “Average,” “Poor”の 3 段階で評価してもらったデータをy_{ij}として用いている．

これに対し, 本研究では, 添削評価を利用する．添削評価とは，コンピュータの出力した要約を被験者が添削し，その添削の度合いで要約の内容に関する品質を評価するオンライン評価法で添削評価では，被験者は原文を読んだ上で，内容，可読性に関して, システムの要約を添削する. 添削は，挿入，削除，置換の 3 つの操作のみで行なう．添削の度合は，添削により削除または置換された文字数の合計を添削前の要約の総文字数で割ることで計算する．理想的な要約の場合，被験者により削除または置換される文字列がないため，評価値は 0 となる．逆に，要約の品質が低い場合には評価値は大きくなる．もし，被験者が要約の半分以上を添削しなければならない場合には，添削を行わず，評価値を 0.5 とする．

3. プーリングデータ数を制限した場合の賀沢手法の有効性の変化

1 節でも述べたように，評価値の計算にすべてのプーリングデータを用いると，適切な計算ができなくなる場合があると考えられる．そこで，本研究では，プーリングデータ中で計算に用いる要約の数を制限することで，賀沢手法の有効性がどの程度変化するかを調べる．本研究では，以下の 2 種類の方法で分析を行う．

手法 1：類似度が閾値以上のものを利用

評価したい要約と酷似したものがプーリングデータの中にある場合には，そのデータの評価値のみを使って評価値を計算した方が，プーリングデータすべてを用いるよりも，高い精度で計算ができると予想される．そこで，評価したい要約 x とプーリングデータの類似度が閾値以上のものだけを用いて評価し，要約 x のオンライン評価値とのずれを調べる．

手法 2：評価したい要約と類似度の高い上位n件を利用

手法 1 では，閾値の設定によっては，要約 x と類似度が閾値以上の要約がプーリングデータ中に存在しない場合がある．そこで，要約 x との類似度が高い上位 n 件のプーリングデータを用いて評価し, 要約 x のオンライン評価値とのずれを調べる．

4. 実験

3 節で説明した分析手法に基づき，実験を行う．

4.1. 実験に用いるデータ

本研究では, 実験に NTCIR Workshop 3 自動要約タスク（以下，TSC2）[Fukushima 2002]のデータを用いる．TSC2 はその評価方法のひとつとして 2 節に述べた添削評価を採用しているが，このデータは，以下に示す PART，FREE，SYS，BASE の 4 種類の要約と添削評価による結果，および主観評価による結果から構成されている．このうち，主観評価もオンライン評価のひとつであるが，今回の実験では用いないため，詳しい説明は省略する．

1. 人間の作成した重要個所抽出要約(PART)
2. 人間の自由作成要約(FREE)
3. 8 つのシステムが提出した結果(SYS)
4. Lead のベースラインシステムの結果(BASE)

上記の要約はすべて，1998 年，1999 年の毎日新聞から人間もしくはシステムによって作成されている．本研究では，TSC2 で行なわれた 2 つの課題のうち, 単一文書要約課題(single)のデータを用いる．この課題は，与えられた文書の 20%および 40%の長さで，要約を生成するものである．

なお，今回の実験では，添削評価値が 0.5 の要約はあらかじめ除外する．その理由は，2 節で述べたように，添削評価では要約の品質が非常に低い(被験者が要約を半分以上書き換えなければならない)場合は，評価値をすべて 0.5 にしていることによる．本来ならば 0.5 より大きな値をとるべきものを評価値 0.5 としてプーリングデータに用いると，適切な評価ができなくなるからである．

4.2. 実験条件

要約間の類似度

賀沢らは，評価したい要約とプーリングデータの類似度を測る際，式(2)を用いているが，本研究では，Lin ら[Lin 2003]が提案する ROUGE という尺度を用いる．

ROUGE とは，機械翻訳の結果を自動的に評価するために Papineni ら[Papineni 2001]が提案した BLEU という尺度を，Lin らが要約評価用に改良したオフライン評価の方法である．我々は，過去の研究において, 3 種類のオフライン評価方法, ROUGE, BLEU, content-based な評価[Donaway

2000]をオンライン評価結果と比較したとき，ROUGE がオンライン評価結果と最もよく一致することを確認している[Nanba 2004]．そこで，本研究においても，要約間の類似度の尺度としてROUGE を用いる．

評価尺度

3節で述べた分析手法の妥当性を評価するため，次に述べる尺度を用いる．評価したい要約を，3節で述べた手法で評価し，その評価値と添削評価値の差の平均を Gap とする(式(3)). Gap 値が 0 に近ければ，提案手法は添削評価との差が小さいことを意味する．

$$Gap = \frac{\sum_{k=1}^{m}\sum_{l=1}^{n} | scr'(x_{kl}) - y_{kl} |}{m \times n} \quad (3)$$

ここで，x_{kl}は，k番目のシステムのl番目の要約の添削評価結果，y_{kl}は，k番目のシステムのl番目の要約の提案手法による評価結果を示す．また，今回の実験で用いる評価関数は，賀沢手法と区別するため，scr'(x)と表記する．

さて，3 節でも述べたように，手法 1 については，もし評価したい要約との類似度が閾値以上のものがプーリングデータ中に存在しなければ，評価値を計算できない. そこで, 手法 1 については，評価したい要約のうち，どの程度実際評価することができたのかを，以下に示す式(4)を用いて計算する．

$$Coverage = \frac{\text{プーリングデータを用いて評価値が得られた要約の数}}{\text{評価したい要約の総数}} \quad (4)$$

評価関数scr'(x)のパラメータ

w_j，bの値は，今回の実験ではw_j=1/(プーリングデータ数)とし，bは$scr'(x_{ij})$とy_{ij}の二乗差の総和が最小になるように最適化する．

4.3. 実験方法

4.1 節で述べたとおり，TSC2 では，人間による 2 種類の要約(FREE，PART)，8 種類のシステム要約(SYS), ベースライン要約(BASE)の計 11 種類の要約が存在する．これらの中から任意のひとつを選択し，残りをプーリングデータとして用い，選択された要約の評価値を関数 scr'(x)で評価する．

4.4. 結果

手法 1 の実験結果

要約率 40%および 20%の時の手法 1 による結果を表 1，2 にそれぞれ示す．また，参考のため，評価値を計算する際，使用したプーリングデータ数の平均値も併せて示す．

表 1　類似度が閾値以上のものを利用した場合の評価値と添削評価値との差(要約率 40%)

閾値	Gap	平均使用データ数	Coverage
0	0.096	9.8	1.000
0.1	0.086	9.2	0.901
0.2	0.085	9.2	0.901
0.3	0.085	9.2	0.901
0.4	0.084	8.8	0.901
0.5	0.082	7.8	0.901
0.6	0.075	5.0	0.895
0.7	0.069	2.4	0.762
0.8	0.062	1.4	0.463
0.9	0.054	1.0	0.284

表 2　類似度が閾値以上のものを利用した場合の評価値と添削評価値との差(要約率 20%)

閾値	Gap	平均使用データ数	Coverage
0.0	0.139	9.8	1.000
0.1	0.128	9.0	0.906
0.2	0.127	8.6	0.906
0.3	0.124	7.5	0.903
0.4	0.118	5.6	0.887
0.5	0.110	3.7	0.858
0.6	0.103	2.7	0.706
0.7	0.099	2.1	0.557
0.8	0.096	1.5	0.453
0.9	0.084	1.4	0.353

表 1, 2 において，閾値が 0 の時は，すべてのプーリングデータを使う場合であり，賀沢手法に相当する．閾値 0 の時に，データをすべて使っているのにもかかわらず平均使用データ数が 10 になっていない理由は，4.1 節でも述べたとおり，添削評価値が 0.5 のものは今回の実験では使っていないためである．

表 1, 2 からわかるとおり，閾値を大きくするほど，評価値と添削評価値との差は小さくなり，適切な評価ができていることがわかる．他方，Coverage は低下しているため, 手法 1 では適切な評価ができるものの数は限られていると言える．

手法 2 の実験結果

手法 2 による結果を表 3 に示す．

表 3 評価したい要約との類似度が高い上位 n 件を用いた場合の評価値と添削評価値とのずれ

	要約率	
上位 n 件	40%	20%
1	0.070	0.124
2	0.080	0.126
3	0.082	0.128
4	0.085	0.131
…	…	…
8	0.096	0.140
賀沢手法	0.101	0.142

表 3 からわかるとおり，n の値が大きくなるほど，提案手法と添削評価のずれが大きくなり，すべてのデータを使った時にずれが最大になった．

以上の結果から，プーリングデータはなるべく似ているものだけ使って評価する方が，添削評価とのずれが小さいと言える．

4.5. 考察

ここでプーリングデータを用いて要約を評価する際，どの程度プーリングデータを用意しておけば十分なのかという点が問題となる．この問題に関する調査は，まだ行っていないが，今回得られた結果からわかる範囲で考察する．

4.4 節で示した実験結果から，評価したい要約との類似度の高いものがプーリングデータに含まれていれば，添削評価とのずれが小さいということがわかっている．このことから，「どの程度プーリングデータを用意しておけば十分であるか」という問題は，「評価したい要約と類似度の高いものがどれだけプーリングデータに含まれているのか」ということと関連があると思われる．これは，表 1 および表 2 の Coverage の値と関連する．閾値が 0.9 の時の Coverage が 1 になれば理想的であるが，現状ではそこまでには至っていない．

5. おわりに

本研究では，TSC2 のデータを用いて，賀沢らの提案する評価手法を分析した．その際，賀沢手法のようにすべてのプーリングデータを使うのではなく，プーリングデータ数を評価したい要約との類似度によって制限した方がよりよい評価が行われるのではないかと考え，実験を行った．その結果，プーリングデータはなるべく類似度の高いものを使って評価する方が，添削評価とのずれが小さいということがわかった．

6. 今後の課題

現状では，プーリングデータについて，どのくらいのバリエーションのものがどの程度あれば十分であるかについて，十分な調査を行っていない．今後は，この点について検討していきたい．

また，複数のオンライン評価を利用したオフライン評価が，ROUGE をはじめとする従来のオフライン評価と比べ優れた手法であるかどうかについても検討していない．今後は，今回の手法を用いて，TSC2 参加システムの順序付けを行い，従来のオフライン評価手法との比較を行う予定である．

参考文献

[Donaway, 2000] 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 2002] Fukushima, T., Okumura, M., and Nanba, H. (2002). Text Summarization Challenge 2 / Text Summarization Evaluation at NTCIR Workshop3, *Working Notes of the Third NTCIR Workshop Meeting, PART V,* pp.1-7.

[賀沢 2003] 賀沢秀人，Arrigan, T.，平尾努，前田英作 (2003). 文書要約における抽出単位と評価法についての考察，*情報処理学会研究報告 自然言語処理 NL-158*，pp.25-30.

[Lin 2003] Lin, C. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics, *Proceedings of the Human Language Technology Conference 2003.*

[Nanba 2004] Nanba, H. and Okumura, M. (2004). Comparison of Some Automatic and Manual Methods for Summary Evaluation Based on the Text Summarization Challenge 2, *Proceedings of the fourth International Conference on Language Resources and Evaluation*.

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

[Yasuda 2003] Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S. and Yanagida, M. (2003) Applications of Automatic Evaluation Methods to Measuring a Capability of Speech Translation System, *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pp.371-378.