言い換えを用いたテキストの自動評価

平原一帆[1]　難波英嗣[1]　竹澤寿幸[1]　奥村学[2]
1.広島市立大学大学院　情報科学研究科
2.東京工業大学　精密工学研究所

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，金銭的コストがかかり，また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るようになってきた．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，テキストに独自の表現を含む，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[5]．

しかし，このような従来の自動評価手法では，独自の表現を含む，人の手によるものに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分かっている．そこで本研究では，テキストの自動評価を行う際に，表層的な文字列の一致だけでなく，言い換えを考慮した手法を提案する．また同時に，複数の言い換え手法を比較し，テキスト自動評価に有効な言い換えの模索と検討を行うことで，従来のテキスト評価手法を改良する．

本論文の構成は以下の通りである．次節では，本研究の関連研究を示し，3節では本研究におけるテキストの評価手法を提案する．4節では実験内容について説明し，5節で考察を行い，6節で本稿をまとめる．

2. 関連研究

テキストの自動評価と同義語及び言い換え抽出の関連研究について，2.1.節と 2.2.節でそれぞれ述べる．

2.1. テキストの自動評価

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がある．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．Linにより提案されている代表的な評価手法であるROUGE[7]について説明する．

ROUGEには，様々な種類のものが存在するが，そのうちのひとつであるROUGE-Nは，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手法である．ROUGE-Nは，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．

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

ここで，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グラム集合を示す．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))の小さい方の値を採用する．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．今回の我々の比較実験のベースラインとして，N=1を用いている．

2.2. 同義語及び言い換えの自動抽出

テキストから同義語や言い換えを自動的に抽出する研究は近年数多く行われているが，本研究と関連のある代表的なものとして，統計的機械翻訳技術を用いた海野ら[11]や，分布類似度を用いた相澤[1]の研究がある．海野らは，対訳コーパスから言い換え表現を自動獲得し，これを従来の情報検索の枠組みに取り入れることによって新しいクエリ拡張手法を提案している[11]．彼らは日英対訳コーパスを用意し，同一の訳語とアライメントのとれた2つの語句を言い換えと見なしている．海野らの提案する言い換えの自動抽出手法は，本研究における言い換え知識抽出の方法のひとつとして使用する．相澤は，新聞記事を対象に分布類似度を用いた同義語抽出を行っている[1]．分布類似度を用いた同義語抽出手法とは，ある語と共起する語に注目し，テキスト中の指定した範囲内で共起する語のベクトルで各語を特徴づけ，これらの共起語ベクトル同士の類似度によって語の類似度を数値化する方法である[8]．相澤は，この手法について，大規模新聞記事コーパスを用いて，語の類似度計算する際における問題点を調査している．本研究でも，この大規模コーパスを用いた分布類似度を，言い換え知識の獲得に用いる．

上述の2つの同義語抽出手法のうち，統計的機械翻訳技術を用いた手法をテキストの自動評価に利用した研究がある[6),13]．Zhouらは，英中対訳文データから統計的機械翻訳技術を用いて言い換え表現を自動的に抽出し，テキスト要約の評価に用いるParaEvalという手法を提案している[13]．この手法では，参照要約とシステム要約の比較によりテキストを評価するが，その際，言い換えのマッチングを，(1) 大域的マッチングと(2) 局所的マッチングの2 段階的で行っている．第一段階で，動的計画法に基づき，句対句による言い換えマッチングを行った後，第二段階で，残った単語に対しGreedy法に基づいて単一語対句，または単一語対単一語による同義語マッチングを行う．第三段階では，第一段階，

第二段階で言い換えに一致しなかった残りの単語に対して, ROUGEを適用(以下, 語彙マッチング) する. Zhouらは，実験により，ParaEvalがROUGEを改良できることを確認している．ただ，言い換えのマッチングに関しては，Zhouらが行っている手順の他に，語彙マッチングを適用した後に(上述の第三段階), 言い換えを適用する(第一段階および第二段階) といった方法も考えられるが, Zhouらは, この点については検討していない. 本研究では，言い換えを用いたテキストの評価手法として，ParaEvalに基づいたものを用いるが，言い換えの適用手順に関して, この点についても調査する. また，言い換え手法として，統計的機械翻訳技術以外に利用可能な同義語抽出技術および同義語辞書を利用する.

3. 提案手法

本節では，言い換えを用いたテキスト評価手法について説明する．3.1.節で提案手法の手順について説明し，3.2. 節でテキスト評価に有効な言い換え知識について説明する.

3.1. 提案手法概要

言い換えを含むテキストを評価するために， 本研究では二種類の手順“ParaEval 手順”と“逆 ParaEval 手順”を用いる. ParaEval 手順では, ParaEval と同様に, まず, 言い換えマッチングを行い，次に語彙マッチングを行うことで，テキスト評価する．手順を以下に述べる.

［ParaEval 手順］

(1)テキストを走査し，句と句から成る言い換えの一致を Greedy 法に基づいて検索する.
(2)(1)で一致しなかった語に対して，単一語対句，または単一語対単一語を走査し，単語レベルの言い換えや, 表記の揺れによる言い換えの一致を Greedy 法に基づいて検索をする.
(3)(1)と(2)で一致しなかった語に対して，語彙マッチングを行う.
(4)(1)，(2)，(3)で参照要約に一致した語のうち，名詞・形容詞・動詞を内容語として数え，参照要約に対する再現率をスコアとして出力する.

これに対し，“逆 ParaEval 手順” では，まず，語彙マッチングを行い，次に，言い換えマッチングを行うことで，テキストを評価する．手順を以下に述べる.

［逆 ParaEval 手順］

(1)テキストを走査し，語レベルの一致を検索する.
(2)(1)で一致しなかった語に対して，句と句から成る言い換えの一致を Greedy 法に基づいて検索する.
(3)(1)と(2)で一致しなかった語に対して，単一語対句, または単一語対単一語を走査し，単語レベルの言い換えや, 表記の揺れによる言い換えの一致を Greedy 法に基づいて検索する.
(4)(1)，(2)，(3)で参照要約に一致した語のうち，名詞・形容詞・動詞を内容語として数え，参照要約に対する再現率をスコアとして出力する.

3.2. 言い換え知識

ParaEval では，英語と中国語の統計的機械翻訳により生成されるフレーズテーブル(翻訳モデル)を用いて同義語辞書を作成している. これに対し, 本研究では, 要約に用いられる言い換えの出現について，自動要約ワークショップ TSC2[3)]で用いられたデータを調査した結果 [4)]を用い, 統計的機械翻訳を用いた同義語抽出の他にも，日本語テキストの評価に利用可能な，以下の 4 種類の言い換え知識を用いて実験を行う.

- SMT(自動収集) : 統計的機械翻訳(SMT)の技術を用いて獲得した言い換え知識
- DS (自動収集) : 分布類似度に基づいて獲得した言い換え知識
- WN (手動収集) : WordNet(日本語版)を用いた言い換え知識
- NTT (手動収集) : NTT 日本語語彙大系を用いた言い換え知識

以下，各言い換え知識について述べる.

■統計的機械翻訳によるフレーズテーブル(SMT)

Zhou ら， 海野らと同様に， 統計的機械翻訳により生成されるフレーズテーブル(翻訳モデル) を用いて言い換え辞書を作成する．この手法は,「もし，X と Y の翻訳が同一であれば，X と Y は言い換えとみなすことができる」という考えに基づいている．本研究では，日英対訳文として読売新聞データベースと The Daily Yomiuri から自動的に抽出された 150,000 文対 [12)]を，また，統計的機械翻訳用のツールとして GIZA++[1] を，それぞれ用いた．ここで，得られたフレーズテーブルから，言い換え知識を獲得する際，それぞれの品詞の並びが異なっているフレーズ[2]は言い換えとして適切でないと考え，削除した．最終的に，85,858 対の言い換えを得た．これらの言い換え知識は，自立語・付属語問わず全ての品詞を含み，フレーズ長も任意である.

■分布類似度(DS)

以下，相澤 [1)]の手順に基づき，分布類似度を用いた言い換え知識を獲得した.
(1)係り受け解析器 CaboCha[3]を用い，毎日・読売・日経新聞データベース計 56 年分の記事に含まれる全ての文を構文解析する.
(2)(1)で得られた解析木から，係り受け関係のある名詞と動詞の対を抽出する.
(3)名詞ごとに，係受け関係にある動詞の頻度を数え，共起語ベクトルを作成する.
(4)与えられた名詞に対し，共起語ベクトル間の類似値が高い順に名詞を出力する．なお，共起語ベクトル間の類似度を計算する尺度として，本研究では情報検索で広く用いられている SMART[10)]を利用する.

上記(2)において，名詞の代わりに名詞句(名詞の連続)と動詞の対を抽出することにより，名詞句の言い換え知識も獲得する．また，(3)において，動詞ごとに，係り受け関係にある名詞の頻度を数えて共起語ベクトルを作ることにより，動詞の言い換え知識も獲得する.

■WordNet(WN)

WordNet は，これまで英語を対象にした言語資源として自然言語処理で広く利用されてきたが，その日本語版が 2009 年 3 月より公開されている [2)4]. WordNet に

[1] http://www.fjoch.com/GIZA++.html
[2] 例えば「名詞-動詞」から構成されるフレーズと「名詞-名詞」から構成されるフレーズ.
[3] http://chasen.org/~taku/software/cabocha/
[4] http://nlpwww.nict.go.jp/wn-ja/

は，名詞，動詞，形容詞，副詞が synset と呼ばれる同義語のグループに分類され，簡単な定義や他の同義語のグループとの関係が記述されている．我々は，日本語版 WordNet の synset を言い換え知識として利用する．

■ NTT 日本語語彙大系(NTT)

NTT 日本語語彙大系には名詞，形容詞，動詞に関して表記の揺れ(異表記)について記載された項目がある．この項目を言い換え知識として利用する．

以上 4 種類の言い換え知識を，表 1 にまとめる．

表 1 テキスト評価に用いた言い換え知識

言い換え知識	品詞	構築
統計的機械翻訳 (SMT)	自立語・付属語を含む任意の単語列	自動
分布類似度(DS)	名詞・名詞句・動詞	自動
WordNet(WN)	名詞・動詞	手動
NTT 日本語語彙大系 (NTT)	名詞・動詞・形容詞	手動

4. 実験

提案手法の有効性を調べるため，実験を行った．

4.1. 実験方法

実験に用いたデータ，言い換え知識，実験手法について，以下に説明する．

■実験データ

本研究では TSC2[3]で用いられた新聞記事の社説から，以下の手順で作成した要約データを用いた．このデータは，約 1150 字から成る新聞記事の社説 30 テーマについて，要約作成者 20 名がそれぞれ 20%の要約率で作成した計 600 要約から構成される．要約作成者 20 名のうち，10 名は原文から抜粋により要約を作成，残り 10 名は自由作成による要約(原文にない表現を使っても良い) を作成した[5]．これにより，提案手法が自由作成による要約に対して有効かどうかの比較を行うことが可能となる．この 600 要約に対して，3 名の評価者が，全ての要約を 1 から 4 までの 4 段階で評価した．

■実験に用いた言い換え知識

提案手法として，統計的機械翻訳(SMT)，分布類似度(DS)，NTT 日本語語彙大系(NTT)，WordNet(WN)の 4 種をそれぞれ組み合わせた計 15 種類を用いる．また，ベースラインとして ROUGE-1 を用いる[6]．

■実験手法

我々は，それぞれのテーマについて，抜粋要約において最も評価の高かったものと生成要約において最も評価の高かったものを参照要約として，以下の方法で実験を行った．EX-1 から EX-3 については，3.1.節で述べた ParaEval 手順を，EX-4 から EX-6 については，逆 ParaEval 手順を，それぞれ用いる．

［ParaEval 手順］

- EX-1：最も評価値の高い抜粋要約を参照要約とし，評価対象の要約として 9 個の**抜粋要約**を評価．
- EX-2：最も評価値の高い抜粋要約を参照要約とし，評価対象の要約として 9 個の**生成要約**を評価．
- EX-3：最も評価値の高い生成要約を参照要約とし，評価対象の要約として 9 個の**抜粋要約**を評価．

［逆 ParaEval 手順］

- EX-4：最も評価値の高い抜粋要約を参照要約とし，評価対象の要約として 9 個の**抜粋要約**を評価．
- EX-5：最も評価値の高い抜粋要約を参照要約とし，評価対象の要約として 9 個の**生成要約**を評価．
- EX-6：最も評価値の高い生成要約を参照要約とし，評価対象の要約として 9 個の**抜粋要約**を評価．

各実験において，参照要約と評価対象の要約との比較によって計算される評価値により，評価対象の要約を順位付けすることができる．これらの順位と，人手による評価に基づいた順位の相関を，スピアマンの順位相関係数を用いて計算し，この相関係数の値の大小により，各評価手法を評価する．

なお，実験データの都合上，以下の点に留意する．

- ある要約に対して，3 名の評価者による評価値が著しく異なっている要約は，人間でも評価が難しい要約であると判断し，評価の対象から外した．
- 人手による評価を 4 段階評価に変換する際，あるテーマの人手評価が全て同一だった場合には，順位相関係数を求めることができないため，評価の対象から外した．

4.2. 実験結果

“ParaEval 手順” を用いた実験結果

“ParaEval 手順” を用いた実験結果を，表 2 に示す．各表の数値は，参照要約と評価対象を比較した時の，提案手法およびベースライン手法に対するスピアマンの順位相関係数(30 テーマの平均値)を示している．

表 2 ParaEval 手順を用いた評価結果

組み合わせ	EX-1	EX-2	EX-3
S (SMT)	0.280	0.326	0.334
D (DS)	**0.338**	**0.379**	**0.349**
W (WordNet)	0.332	0.376	0.337
N (NTT)	0.332	0.367	0.337
SD	**0.340**	0.369	**0.348**
SW	**0.358**	0.336	0.337
SN	0.276	0.338	0.294
DW	**0.359**	0.326	0.334
DN	**0.343**	0.374	**0.357**
WN	0.332	0.376	0.337
SDW	**0.339**	0.331	0.325
SDN	**0.348**	0.356	**0.345**
SWN	**0.346**	0.350	**0.341**
DWN	**0.358**	0.327	0.326
SDWN	**0.340**	0.326	0.329
ROUGE-1(Baseline)	0.332	0.376	0.337

表 2 において，参照要約に抜粋要約を用い，評価対象の要約に抜粋要約を用いる EX-1 においてベースラインである ROUGE-1 を半分以上が上回り，言い換えを用いることが有効に機能していることが分かる. 表 2 の 15 の提案手法のうち， EX-1 において“DW”が最も有効に機能し, ROUGE-1 を 0.027(8.1%)改善している．

[5]本来ならば，要約システムが作成した要約を利用すべきであるが，数多くの異なる抜粋および生成ベースの要約システムを用意することが困難であったため，人間が作成した要約を擬似的にシステム要約と見なすことにより，実験を行っている

[6] 2.1 節で述べたとおり, ROUGE には様々な種類のものが存在するが, TSC2 のデータを用いた Nanba らの実験[9]では, ROUGE-1 を用いた時に最も高い精度が得られていることから，今回の実験では，ROUGE-1 をベースラインとして用いた．

“逆 ParaEval 手順” を用いた実験結果

“逆 ParaEval 手順” を用いた実験結果を, 表 3 に示す.

表 3 逆 ParaEval 手順を用いた評価結果

組み合わせ	EX-4	EX-5	EX-6
S (SMT)	0.265	0.373	0.308
D (DS)	**0.377**	**0.409**	0.337
W (WordNet)	**0.346**	**0.398**	0.336
N (NTT)	**0.350**	**0.398**	0.335
SD	**0.343**	**0.390**	**0.347**
SW	**0.337**	**0.382**	**0.349**
SN	0.270	**0.384**	0.310
DW	**0.348**	**0.381**	**0.349**
DN	**0.373**	**0.409**	**0.339**
WN	**0.346**	**0.398**	0.336
SDW	**0.340**	**0.380**	**0.350**
SDN	**0.335**	**0.389**	**0.342**
SWN	**0.342**	**0.383**	**0.368**
DWN	**0.345**	**0.383**	**0.351**
SDWN	**0.334**	**0.382**	**0.359**
ROUGE-1(Baseline)	0.332	0.376	0.337

表3の提案手法のうち, 参照要約に抜粋要約を用い, 評価対象の要約に抜粋要約を用いる EX-4 において “D” が最も有効に機能し，ROUGE-1 を 0.045(13.6 %)改善した．また，ParaEval 手法と比較して，EX-5 ,EX-6 について，全体的に評価値が向上している.

5. 考察

■テキスト評価における言い換えの効果

今回の実験結果では, 主に EX-1 の提案手法において, ROUGE-1 の評価を改善する傾向にあった．また，EX-3 の結果においても改善が見られたため，評価対象として抜粋要約を用いた際に，言い換えを有効に活用できていると考えられる．EX-1 は参照要約と評価対象要約がともに抜粋で作成されているが，抜粋する部分によっては言い換えを用いることで評価可能になる部分があるため，その点で機能したと考えられる．評価対象の要約が生成要約である EX-2 については, 今回実験に用いた生成要約が，非常に技巧的な表現を数多く用いて作成したため，参照要約と評価対象の要約における語の変容が大きく, 言い換えがうまく機能していない.

■ “ParaEval 手順” と“逆 ParaEval 手順” の比較

EX-1 ～EX-3(ParaEval 手順) と， EX-4 ～EX-6(逆 ParaEval 手順) の変化について考察する. 言い換え知識の適用順序を変更することにより, 特に EX-2 と比べて EX-5 が顕著に向上していることが分かる．全体的な傾向として，ParaEval 手順よりも逆 ParaEval 手順の方が改善率が高いことを考慮すると，ParaEval 手順で用いられている数多くの言い換えの中には，不適切なものが少なからず含まれていると考えられる．3.2.節でも述べたとおり，今回言い換え知識として用いた 4 手法のうち，“SMT”と“DS”は自動的に獲得されたものであるため，関連語ではあるが言い換えとしては適切でないものも含まれているが，“ParaEval 手順”では，参照要約と評価対象の要約内に不適切な言い換え知識である単語同士でも、言い換えと判断すれば適用してしまう．しかし，“逆 ParaEval 手順”では，誤った言い換えが行われる前に語彙マッチングを行うため，不適切な言い換えが適用されることが少なくなる．微小ではあるが，EX-1 と EX-4，および EX-3 と EX-6 の間においても，同様の傾向が確認された.

6. おわりに

本研究では，様々な言い換え知識を用いたテキスト評価手法を提案した．提案手法では，Zhou らが提案する統計的機械翻訳技術を用いた言い換えに基づくテキスト評価手法 ParaEval をベースにしているが，言い換え知識として Zhou らの手法の他に，分布類似度，WordNet，NTT 日本語語彙大系も利用しており，さらに，言い換えの適用手順についても工夫している点が異なる．提案手法の有効性を検証するため，TSC2 のデータを用いて実験を行った．実験の結果，分布類似度による言い換えを用いた場合に，従来手法に比べ，スピアマンの相関係数による評価値で 0.045 の改善が得られた．また，従来手法である ParaEval とは言い換えの適用順序を変えた“逆 ParaEval 手順” を用いた場合に, 提案手法の有効性が確認された.

参考文献

1) 相澤彰子: 大規模テキストコーパスを用いた語の類似度計算に関する考察，情報処理学会論文誌，Vol.49, No.3, pp.1426–1436 (2008).

2) Bond, F., et al.: Extendingthe Japanese WordNet，言語処理学会第 15 回年次大会, pp.80–83 (2009).

3)Fukushima, T., et al.: Text Summarization Challenge 2 /Text Summarization Evaluation at NTCIR Workshop 3, Proc. NTCIR-3 Workshop, PART V, pp.1–7 (2002).

4) Hirahara, K., et al.: Automatic Evaluation of Texts by Using Paraphrase. Proc. LTC'09, pp.370–374. (2009)

5) Hovy, E., et al.: Automated Summarization Evaluation with Basic Elements, Proc. LREC-2006.

6) Kauchak, D., et al.: Paraphrasing for Automatic Evaluation, Proc HLT-NAACL 2006, pp.455–462 (2006).

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

8) Lin, D: Automatic Retrieval and Clustering of Similar Words, Proc. COLING/ACL 1998, pp.768–774 (1998).

9) Nanba, H., et al.: An Automatic Method for Summary Evaluation Using Multiple Evaluation Results by a Manual Method, Proc. COLING/ACL 2006 Main Conference Poster Sessions, pp.603–610 (2006)

10) Salton, G: The SMART Retrieval System. Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River, NJ, (1971).

11) 海野裕也, 他: 自動獲得された言い換え表現を使った情報検索，言語処理学会第 14 回年次大会，pp.123–126 (2008).

12) Utiyama, M., et al.: Reliable Measures for Aligning Japanese-English News Articles and Sentences. Proc. ACL 2003, pp.72–79 (2003).

13) Zhou, L., et al.: ParaEval: Using Paraphrases to Evaluate Summaries Automatically. Proc. HLT-NAACL 2006, pp.447–454 (2006).