情報処理学会論文誌　データベース　Vol. 3　No. 2　91–101 (June 2010)

言い換えを用いたテキスト要約の自動評価

平原一帆[†1]　難波英嗣[†1]
竹澤寿幸[†1]　奥村学[†2]

コンピュータにより生成された要約の評価は，近年の自動要約研究における重要な研究課題の 1 つと認識されている．従来は，抜粋形式の要約を，正解要約との文字列の一致度により評価するのが一般的であった．しかしながら，生成要約には原文にはない独自の表現が含まれることがあるため，従来の評価手法では，生成要約を抜粋要約と同様の精度で評価することができないという問題があった．そこで，本稿では，言い換え技術を用いることにより，従来の評価手法の改善を試みる．提案手法の有効性を確認するため，テキスト自動要約タスク TSC2 のデータを用いて実験を行った．実験の結果，提案手法は，抜粋要約と生成要約のいずれの評価においても従来の評価手法を改善できることが確認された．

Automatic Evaluation of Text Summaries by Using Paraphrase

Kazuho Hirahara,[†1] Hidetsugu Nanba,[†1]
Toshiyuki Takezawa[†1] and Manabu Okumura[†2]

The evaluation of computer-produced summaries has been recognized as an important research problem for automatic text summarization. Traditionally, computer-produced summaries were evaluated automatically by n-gram overlap with human-produced texts. However, these methods cannot evaluate summaries correctly, if the n-grams do not overlap between computer-produced and human-produced summaries, even though the two summaries convey the same meaning. We explored the use of paraphrases for the refinement of traditional automatic methods for summary evaluation. To confirm the effectiveness of our method, we conducted some experiments using the data from the Text Summarization Challenge 2. We found that the use of paraphrases created using a statistical machine translation technique improved the traditional evaluation methods.

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るうえに，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，自動化によって行うことが求められるようになってきた[13]．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[4),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分かっている．これらを解決するために，表層的な文字列の一致だけでなく，言い換えを考慮したテキストの自動評価手法が提案されている[18)]が，言い換えの適用順序や，テキスト自動評価に有効な言い換えにおける議論は尽くされていない．そこで本研究では，従来の言い換えの適用手法を検討し，また，複数の言い換え手法を比較することで，テキスト自動評価に有効な言い換えの模索と検討を行い，従来のテキスト評価手法を改良する．

本稿の構成は以下のとおりである．次章では，本研究の関連研究を示し，3 章では，テキスト評価における言い換えの必要性について述べる．4 章では本研究におけるテキストの評価手法を提案する．5 章では実験内容について説明し，6 章で考察を行い，7 章で本稿をまとめる．

†1 広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University
†2 東京工業大学精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology

2. 関 連 研 究

テキストの自動評価と同義語およ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．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[14] 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，近年では，テキスト要約のための自動評価の手法としても注目を集めている．BLEU はシステム要約と 1 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[9] が参照要約中にどの程度出現するかにより測定する．しかし，要約評価の場合，原文中の重要な情報がどの程度要約に含まれるか，すなわち再現率による評価が重要となるため，精度を評価する BLEU は馴染まないこと，要約はできるだけ短いほうが望ましいため，要約が短い場合に補正を行う BLEU は要約評価には適さないなどの問題点がある．これらの問題点を要約評価用に改良したものとして，ROUGE[10] という尺度が Lin により提案されている．ROUGE には，様々な種類のものが存在するが，そのうちの 1 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．

$$ROUGE\text{-}N = \frac{\sum_{S \in \{Reference\ \ Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Reference\ \ Summaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

ここで，n は N グラムの長さを意味し，$gram_n$ と $Count_{match}(gram_n)$ は，参照要約のすべてとシステム要約において一致している N グラムの最大数である．

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 同義語およ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[6]，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[16] 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[16]．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．たとえば日本語の「二酸化炭素」と「炭酸ガス」は両方とも英文中で “carbon dioxide” と対応付けがとられる確率が高い．このとき “carbon dioxide” をピボット（軸）として，「二酸化炭素」と「炭酸ガス」が言い換え表現になっていると見なすことができる．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 1 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どうし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),11]．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，語の類似度計算をする際における問題点を調査している．この調査では，広範囲の語と共起する語が類似度計算におけるノイズとなるという前提に基づき，ノイズ低減のために「フィルタリング法」と「サンプリング法」という分布類似度を改良する 2 つの方法を提案し，実験により提案手法の有効性を確認している．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[7),18]．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[18]．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 1 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 2 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 3 段階では，第 1 段階，第 2 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ほかに，まず語彙マッチングを適用した後に（上述の第 3 段階），言い換えを適用する（第 1 段階および第 2 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の手法も利用する．

言い換え技術をテキスト評価に用いたこのほかの研究に Kauchak らのものがある[7]．この研究では，機械翻訳の評価において，文脈を考慮した言い換え手法を提案している．参照テキスト（人間の被験者が作成した正解翻訳）の言い換えのうち，システムテキスト（システムが自動的に生成した翻訳）に現れている語のみを言い換え候補とし，言い換え候補を参照要約に適用する際に文脈的に適切かどうかを判断している．適切と判断された言い換えを用いて，複数の参照テキストを生成し，自動評価における新たな参照テキストとして用いている．この手法により，最初の参照テキストのみを用いた（言い換えを用いて生成された参照テキストを用いない）評価と比べ，人手により近い評価が行えることを示している．しかしながら，Kauchak らの手法でも，同義語抽出技術や同義語辞書などの言語リソースは数多く存在するにもかかわらず，それらを網羅的に利用するまでには至っていない．本研究では，テキストの自動評価に，日本語用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. テキスト評価における言い換えの必要性

我々は，同一テキストから生成された複数の要約を比較することにより，実際のテキスト評価にどのような言い換えが必要であるのか，調査を行った．

調査用データ

調査には，自動要約ワークショップ TSC2[3] で用いられた毎日新聞のデータベースの社説 30 記事中の 10 記事を用いた．各記事について，要約率 20%，原文にない表現を自由に用いてよいという条件で人間の要約作成者 10 名が作成した 100 要約について調査を行った[⋆1]．これらの要約および原文（記事）を比較した結果，計 318 個の言い換えが存在することが分かった．

テキスト評価に出現する言い換え

上述の 318 個の言い換えを，以下の 5 種類に分類した．

（A）表記の揺れレベルの言い換え表現（64 件）20.1%
（B）語レベルの言い換え表現（78 件）24.5%
（C）句レベル言い換え表現（38 件）11.9%
（D）節・付属語レベル言い換え表現（39 件）12.3%
（E）その他の言い換え表現（99 件）31.2%

これらの言い換えについて，以下に説明する．

（A）表記の揺れレベルの言い換え表現

「まじめ」と「真面目」，「3ヶ月」と「三ヶ月」など，漢字表記と仮名表記の両方が自然に用いられる単語において，表記の揺れによる言い換えが行われている．これは仮名表記を漢字表記にすることでの文の短縮を狙ったものや，変換ミスなどが要因であると考えられる．この場合は，表記に関する言い換えを用いることで，対処が可能となると考えられる．

（B）語レベルの言い換え表現

「歳月」と「時間」，「オリンピック」と「五輪」など，名詞・動詞・形容詞における単語どうしで似た表現において言い換えが行われている．これは要約作成時に，短縮やより一般的な語の選択を目的とする場合，あるいは，より洗練された文章を作成する目的などが要因であると考えられる．この場合は，一般的な同義語辞書による言い換えを用いることで対処が可能となると考えられる．

（C）句レベルの言い換え表現

「理想の形」と「理想形」など，名詞句の短縮や複合表現などを用いた表現で言い換えが行われている．これは要約という性質上，文短縮を目的とした言い換えが用いられる傾向にあるためである．この場合は，句に対応した言い換えが必要となる．

（D）節・付属語レベルの言い換え表現

「X がなければ Y はできなかった」と「X により Y できた」，「飛べる」と「飛ぶことができる」など，節間・節内にわたって変化する言い換えや，可能・使役・態・文体などにより発生する付属語の変化による言い換えが行われている．この場合は，付属語を考慮しない評価を行うことや，態や使役，自他の交替も含めた慣用的な言い換えを考慮することで対処が可能となると考えられる．

（E）その他の言い換え表現

「アメリカに対する見方は余り好意的でない」と「嫌米ムードがある」，「社会不安を押さえ込むべき」と「安定追求を求める」などが，この分類に含まれる．このような言い換えは，ある行為が文章中で特別な意味を持っている場合や，要約として簡潔な表現への転換，文章全体からの筆者の総合的な判断による詳述などが，その要因であると考えられる．この場合は，言い換えの考慮のみでの評価は難しいと考えられる．

以上の調査結果から，本研究では前述する言い換え A〜D について，言い換え知識を用

⋆1 5 章で示す実験データと同じ要約のうち，記事テーマ 1–10 を自由作成により要約した 100 要約．

いて自動評価を行うものとした．

4. 提 案 手 法

本章では，パラフレーズを用いたテキスト評価手法について説明する．4.1 節で提案手法の手順について説明し，4.2 節で上述のカテゴリ A〜D に有効な言い換え知識について説明する．

4.1 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 2 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，テキストを評価する．手順を以下に述べる．

[ParaEval 手順]

(1)　テキストを走査し，句と句からなる言い換え（カテゴリ C と D）の一致を検索する．
(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（カテゴリ B）や，表記の揺れによる言い換え（カテゴリ A）の一致を検索をする．
(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．

ここで，(1)，(2) の言い換えの一致を検索の際に複数の候補があがる場合，形態素数の大きさを評価値と見なした Greedy 法により言い換えを決定する．また，手順 (4) において，すべての単語を使うのではなく，内容語に限定しているのは，Lin [10] の報告に基づいたものである．

これに対し，“逆 ParaEval 手順” では，まず，語彙マッチングを最初に行い，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，語彙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，句と句からなる言い換え（カテゴリ C と D）の一致を検索する．
(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（カテゴリ B）や，表記の揺れによる言い換え（カテゴリ A）の一致を検索する．
(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．

4.2 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が，本研究では，統計的機械翻訳を用いた同義語抽出のほかにも，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 は言い換えと見なすことができる」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[17] 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・付属語問わずすべての品詞を含み，フレーズ長も任意である．したがって，3 章で述べた言い換えの分類すべてを含む．

■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すべての文を構文解析する．

⋆1 http://www.fjoch.com/GIZA++.html
⋆2 たとえば「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/

表 1　テキスト評価に用いた言い換え知識
Table 1　Paraphrasing methods for text evaluation.

言い換え知識	品詞	構築方法
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

(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，係り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 [15] 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),⋆1]．WordNet には，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．また，**表 2** は，3 章で述べた 4 つの言い換えカテゴリと，4 つの言い換え知識との関係について示したものである．ここで，◎は，言い換えカテゴリに最も対応する言い換え知識であることを，○は，言い換えカテゴリに対応する言い換えが言い換え知識に含まれていることを，△は言い換えカテゴリに対応する言い換えが，言い換え知識に含まれる可能性があることを，それぞれ示している．

⋆1 http://nlpwww.nict.go.jp/wn-ja/

表 2　言い換えの分類と言い換え知識の対応
Table 2　Correspondence of the classification of paraphrases and paraphrasing methods.

分類	言い換えの粒度	出現頻度	SMT	DS	Word Net	NTT	必要となる技術
A	文字表記	20.1% (64/318)	△	△	△	◎	文字表記の言い換え
B	単語	24.5% (78/318)	○	◎	◎		単語単位の言い換え
C	句	11.9% (38/318)	○	△	△		句単位の言い換え
D	節	12.3% (39/318)	△				節単位の言い換え
E	その他	31.2% (99/318)					構文解析による言い換え

5. 実　　験

4 章で述べた手法の有効性を調べるため，実験を行った．

5.1 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 1,150 字からなる新聞記事の社説 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よい）を作成した[⋆2]．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 名の評価者が評価基準に即して評価を行い，それぞれ 1–10 の 10 段階の評価値を決定している．評価基準については，8 点を基本点とし，「要約が記事中の重要な内容をどの程度の割合で含んでいるか」という観点から，評価すべき点がある要約には 1 点の単位で加点を，逆に減点すべき点がある要約は 1 点の単位で減点するとして評

⋆2 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．

表 3　3 名の評価者による評価結果の分布
Table 3　The distribution of evaluation results by three human subjects.

評価値	評価者 A	評価者 B	評価者 C
4 以下	0	0	29
5	53	0	67
6	89	38	180
7	143	372	200
8	165	170	93
9	89	20	27
10	61	0	4

価を行った[⋆1]．この 3 名の 10 段階評価の結果を用いて，最終的にすべての要約について，A–D の 4 段階の評価値を決定した[⋆2,⋆3]．

■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[⋆4]．以上を**表 4** にまとめる．

■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[⋆5]．EX-1 から EX-4 については，4.1 節で述べた ParaEval 手順を，EX-5 から EX-8 については，逆 ParaEval 手順を，それぞれ用いる．

表 4　提案手法とベースライン手法
Table 4　List of 15 proposed methods and a baseline method.

手法		SMT	DS	WordNet	NTT
提案手法	S	○			
	D		○		
	W			○	
	N				○
	SD	○	○		
	SW	○		○	
	SN	○			○
	DW		○	○	
	DN		○		○
	WN			○	○
	SDW	○	○	○	
	SDN	○	○		○
	SWN	○		○	○
	DWN		○	○	○
	SDWN	○	○	○	○
ベースライン	ROUGE-1				

[ParaEval 手順]

- EX-1：参照要約として最も評価値の高い抜粋要約を用い，評価対象の要約として 9 個の**抜粋要約**を評価．
- EX-2：参照要約として最も評価値の高い抜粋要約を用い，評価対象の要約として 9 個の**生成要約**を評価．
- EX-3：参照要約として最も評価値の高い生成要約を用い，評価対象の要約として 9 個の**抜粋要約**を評価．
- EX-4：参照要約として最も評価値の高い生成要約を用い，評価対象の要約として 9 個の**生成要約**を評価．

[逆 ParaEval 手順]

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

⋆1 このときの 3 名の評価結果について言及する．評価者 A の平均点は 7.55 点，評価者 B の平均点は 7.29 点，評価者 C の平均点は 6.58 となった．また，実際の評価の分布について，**表 3** に示す．

⋆2 3 名の評価者につけられた 10 段階評価を 1 つの 4 段階評価にするために，3 名の評価者による評価値の算術平均と標準偏差を用いた．各要約について，算術平均 + 標準偏差より高いものを評価 A（最高評価），算術平均よりも高いものを評価 B，算術平均よりも低いものを評価 C，算術平均 − 標準偏差よりも低いものを評価 D（最低評価）としている．たとえば，評価値の平均が 5.5，標準偏差が 1 だった場合を考えると，ある要約の 3 名の評価者による平均評価値が 6 だった場合，4 段階評価値は B とする（6.5 未満 5.5 以上であるため）．同様に評価値が 3 だった場合，4 段階評価値は D とする（4.5 未満であるため）．

⋆3 本研究では，「要約の内容」に関する評価に焦点を当て，この側面からのみの評価を用いている．「要約の読みやすさ」に関する評価もテキスト要約の分野における重要な研究課題の 1 つとして認識されつつあるが，その方法論については現状では十分に議論されておらず，今後，検討していく必要があると考えられる．

⋆4 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，Lin [10] の実験において，ROUGE-1 または ROUGE-2 を用いたときに最も高い精度が得られていること，また，TSC2 のデータを用いた Nanba らの実験[12] では，ROUGE-1 を用いたとき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．

⋆5 最も評価の高かったものだけを参照要約として使う方法のほかに，上位 2 件以上を使う方法も考えられるが，実際に作成された要約を見たところ，1 名だけ他の被験者のものと比べ，突出して質の高い要約を作成していたため，今回は参照要約を上位 1 件のみ利用した．

- EX-7：参照要約として最も評価値の高い生成要約を用い，評価対象の要約として 9 個の**抜粋要約**を評価.
- EX-8：参照要約として最も評価値の高い生成要約を用い，評価対象の要約として 9 個の**生成要約**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．本実験では，3 名の評価者が決定した評価値の標準偏差が 1.5 以上の要約 39 件を評価対象外としている．
- 人手による評価を 4 段階評価に変換する際，あるテーマにおける人手評価がすべて同一だった場合[⋆1]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．本実験では，30 テーマのうち 1 テーマが当項目に該当し，4 件の要約を評価対象外としている．

5.2 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 5** と**表 6** に示す．各表の数値は，参照要約と評価対象を比較したときの，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．また，表 5 は参照要約を抜粋要約とした場合，表 6 は参照要約を生成要約とした場合の結果を，それぞれ示している．

表 5 において，EX-1 においてベースラインである ROUGE-1 を半分以上が上回っており，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 5 の 15 の提案手法のうち，抜粋要約を評価した結果における “DW” が最も有効に機能し，ROUGE-1 を 0.027（8.1%）改善している．また，表 6 に示す 15 種類の提案手法のうち，抜粋要約を評価した結果における “DN” が最も有効に機能し，ROUGE-1 を 0.020（5.9%）改善している．

⋆1 たとえば，あるテーマの 9 つの要約すべてが人手評価値「B」と判断される場合．本実験における評価対象外の 4 件は，あるテーマの評価対象 9 要約のうち 5 件が前項の「評価の難しい要約」に該当したため，そのテーマの評価対象が 4 要約となり，その 4 要約の人手評価値がすべて同一だったことによる．

表 5　抜粋参照要約を用いた評価結果（ParaEval 手順）
Table 5　Evaluation results using an extract-type reference summary (ParaEval method).

	組合せ	抜粋（EX-1）	生成（EX-2）
提案手法	S（SMT）	0.280	0.326
	D（DS）	**0.338**	**0.379**
	W（WordNet）	0.332	0.376
	N（NTT）	0.332	0.367
	SD	**0.340**	0.369
	SW	**0.358**	0.336
	SN	0.276	0.338
	DW	**0.359**	0.326
	DN	**0.343**	0.374
	WN	0.332	0.376
	SDW	**0.339**	0.331
	SDN	**0.348**	0.356
	SWN	**0.346**	0.350
	DWN	**0.358**	0.327
	SDWN	**0.340**	0.326
ベースライン	ROUGE-1	0.332	0.376

表 6　生成参照要約を用いた評価結果（ParaEval 手順）
Table 6　Evaluation results using an abstract-type reference summary (ParaEval method).

	組合せ	抜粋（EX-3）	生成（EX-4）
提案手法	S（SMT）	0.334	0.364
	D（DS）	**0.349**	0.421
	W（WordNet）	0.337	0.448
	N（NTT）	0.337	0.428
	SD	**0.348**	0.374
	SW	0.337	0.435
	SN	0.294	0.352
	DW	0.334	0.420
	DN	**0.357**	0.403
	WN	0.337	0.448
	SDW	0.325	0.412
	SDN	**0.345**	0.374
	SWN	**0.341**	0.424
	DWN	0.326	0.416
	SDWN	0.329	0.400
ベースライン	ROUGE-1	0.337	0.448

表 7　抜粋参照要約を用いた評価結果（逆 ParaEval 手順）
Table 7　Evaluation results using an extract-type reference summary (Reverse ParaEval method).

	組合せ	抜粋（EX-5）	生成（EX-6）
提案手法	S（SMT）	0.265	0.373
	D（DS）	**0.377**	**0.409**
	W（WordNet）	**0.346**	**0.398**
	N（NTT）	**0.350**	**0.398**
	SD	**0.343**	**0.390**
	SW	**0.337**	**0.382**
	SN	0.270	**0.384**
	DW	**0.348**	**0.381**
	DN	**0.373**	**0.409**
	WN	**0.346**	**0.398**
	SDW	**0.340**	**0.380**
	SDN	**0.335**	**0.389**
	SWN	**0.342**	**0.383**
	DWN	**0.345**	**0.383**
	SDWN	**0.334**	**0.382**
ベースライン	ROUGE-1	0.332	0.376

表 8　生成参照要約を用いた評価結果（逆 ParaEval 手順）
Table 8　Evaluation results using an abstract-type reference summary (Reverse ParaEval method).

	組合せ	抜粋（EX-7）	生成（EX-8）
提案手法	S（SMT）	0.308	0.352
	D（DS）	0.337	0.420
	W（WordNet）	0.336	0.440
	N（NTT）	0.335	0.437
	SD	**0.347**	0.389
	SW	**0.349**	0.377
	SN	0.310	0.349
	DW	**0.349**	0.375
	DN	**0.339**	0.424
	WN	0.336	0.440
	SDW	**0.350**	0.380
	SDN	**0.342**	0.394
	SWN	**0.368**	0.383
	DWN	**0.351**	0.367
	SDWN	**0.359**	0.373
ベースライン	ROUGE-1	0.337	0.448

“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 7** と**表 8** に示す．表 7 は参照要約を抜粋要約とした場合，表 8 は参照要約を生成要約とした場合の結果を，それぞれ示している．表 7 の提案手法のうち，抜粋要約を評価した結果における “D” が最も有効に機能し，ROUGE-1 を 0.045（13.6%）改善した．一方，表 8 においては，抜粋要約を評価した結果における “SWN” が最も有効に機能し，ROUGE-1 を 0.031（9.2%）改善した．抜粋参照要約を用いて抜粋要約を評価する際の評価値は，ParaEval 手法と比較して，全体的に向上している．また，抜粋参照要約を用いて生成要約を評価する際の評価値も向上していることが分かる．

6. 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，特に評価対象要約として，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，たとえば，問題文原文で「震災地」と「被災地」という語が両方用いられている場合，要約作成時に「震災」という単語を用いる筆者と「被災」という単語を用いる筆者が現れる可能性がある．このような，同じ問題文原文を抜粋するうえで発生する言い換えは，要約評価における問題として従来から指摘されてきたが，言い換えによる評価を行うことでこの問題に対応することが可能になり，精度の向上につながったと考えられる．

参照要約と評価対象要約がともに生成要約で作成されている EX-4 については，参照要約と生成要約との間の語の変容が大きく，言い換えがうまく機能していない．その要因の 1 つは，今回参照要約として実験に用いた生成要約の作成者が，原文にはない非常に技巧的な表現を数多く用いて作成したため，今回実験に用いた言い換え知識だけでは対処できないような高度な言い換え知識が評価に必要となるケースが非常に多かったことによる．

■ “ParaEval 手順” と “逆 ParaEval 手順” の比較

EX-1〜EX-4（“ParaEval 手順”）と，EX-5〜EX-8（“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-6 が顕著に向上していることが分かる．実際に，言い換え知識の適用順序の変更により，評価の際に利用される言い換えにどのような変化があったのか調査した．

この調査では，4.1 節で述べた “ParaEval 手順” と “逆 ParaEval 手順” において，それぞれ，句レベル（“ParaEval 手順” (1) と “逆 ParaEval 手順” (2)）および語レベル（“Para-

Eval 手順”（2）と“逆 ParaEval 手順”（3））でマッチングした 1 要約あたりの平均言い換え個数を調べる．この際，“ParaEval 手順”と“逆 ParaEval 手順”の違いにおける変化のみを調査するため，“ParaEval 手順”として EX-1 から EX-4，“逆 ParaEval 手順”として EX-5 から EX-8 までの抜粋・生成要約の組合せすべてについて，言い換え知識を SDWN（すべての言い換え知識）としたうえで，1 要約あたりの平均言い換え個数を調べた．調査の結果，“ParaEval 手順”では句レベル 5.2 個，語レベル 49.1 個であったのに対し，“逆 ParaEval 手順”では句レベル 0 個，語レベル 4.6 個となった．“逆 ParaEval 手順”においては，句レベルの言い換えは小数点以下にしか現れないほど使用数が少ない．これは先に語彙マッチングを行うことで，助詞などの付属語が一致するため，「学校に行く」の「に」が一致済みとなり「学校」と「行く」のような語レベルの部分ばかりが残り，適用可能な句が消滅してしまうことによると考えられる．なお，全体的な傾向として，“ParaEval 手順”よりも“逆 ParaEval 手順”の方がより改善された結果となることを考慮すると，“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．

4.2 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，たとえば，「生徒」と「登校」のように，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．しかし，“ParaEval 手順”では，参照要約と評価対象の要約内に，それぞれ「生徒」および「登校」という語が含まれていれば，この不適切な言い換え知識を用いて言い換え関係にあると認定されてしまう[⋆1]．他方で，もし，参照要約と評価対象の要約の両方に「生徒」という言葉が含まれていれば，“逆 ParaEval 手順”では，先に「生徒」がマッチングするため，後で「生徒」と「登校」が誤ってマッチングされることがなくなる．また，微小ではあるが，EX-1 と EX-5，および EX-3 と EX-7 の間においても，同様の傾向が確認された．

一方で，EX-8 においては，EX-4 に比べ，評価を下げる結果となった．これは，先に語彙マッチングを適用することにより，適用可能な言い換えが消失してしまうという問題に影響された結果である．たとえば，ある要約において，言い換えを優先することで出現する，「クリントン大統領」と「米大統領」という言い換えは，先に語彙マッチングを行うと「大統領」で語彙レベルで一致が発生してしまい，「クリントン」と「米」という言い換えが存在しないため，正しい言い換えであるにもかかわらず言い換えが行われない．今回の実験においては，言い換えの精度が十分とはいえなかったため，語レベルのマッチングを先に行うことにより，不要な言い換えに左右されることなく評価できたが，今後，言い換えの自動抽出精度が向上すれば，“ParaEval 手順”による評価が向上する可能性もある．

■言い換え知識

4 つの言い換え知識とその組合せを用いた実験の結果，“D”（分布類似度）と“W”（WordNet），あるいはその組合せを用いた場合に，特に改善が見られた．これらの共通点は，語レベルの名詞や動詞の言い換えが中心的であり，必要な言い換えとして数が多かったことがあげられる．今回あまり改善につながらなかった“S”（統計的機械翻訳）と“N”（NTT 日本語語彙大系異表記項目）においては，“S”は大規模に大量収集できる反面，精度の点で十分とはいえないことがあげられる．一方，“N”は精度の点では十分な反面，数が少ないことと，求められる言い換えとして，表記の揺れの需要が他の言い換えよりも少なかったことに起因すると考えられる．また，これらの組合せにおいては，複数の言い換え知識を複合することで，互いに適用される言い換えが阻害されることがあり，一概に言い換えの個数が多ければ良いとはいえない．精度の点，また，言い換えを要約に適用するアルゴリズムの点からも改善が望まれる．

■トピック別・手順別の言い換え適用数

次に，本実験において適用された言い換えについて，トピックごとに傾向がないか調査を行った．調査の結果，トピックによっては出現しやすい句レベルの言い換えが存在することが確認された．実際，「阪神大震災」に関するトピックにおいて，「阪神大震災」と「震災」，「震災犠牲者」と「被災者」などの言い換えがほとんどの要約に出現していることが分かった．また，「自民党による改正法案」に関するトピックで，「自民党は」と「党が」，「改正法案」と「改正案」などの言い換えが出現していた．今回実験に用いた記事は 30 トピックしかないため断定的な結論は下せないが，全体的な傾向として，単名詞（単一語）で一般名詞で和語が言い換えられるケースが多く，逆にカタカナ語などの外来語が言い換えられるケースはそれほど多くはなかった．将来的に，より多くのトピックを対象に分析することで，言い換えが発生しやすいトピックを自動判定し，より効果的に評価を行うことができる可能性も示唆される．

⋆1 この問題に対し，Kauchak ら[7] は，言い換え候補になっている 2 つの単語それぞれの周囲に出現する単語を素性とし，言い換えるべきかどうかを判定するフィルタを機械学習により獲得している．たとえば，“place”といった多義語の言い換えを適用する場合，語義が異なればその周囲に出現する単語も異なるため，この手法は有効に機能すると思われる．しかし，上述のような「生徒」と「登校」の場合，類義語である「生徒」と「登校」それぞれの周辺に出現する単語も類似するため，Kauchak らの手法では解決できないと考えられる．

7. 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要約の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 らの手法のほかに，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., Isahara, H., Uchimoto, K., Kuribayashi, T. and Kanzaki, K: Extending the Japanese WordNet，言語処理学会第 15 回年次大会，pp.80–83 (2009).
3) Fukushima, T., Okumura, M. and Nanba, H: Text Summarization Challenge 2/Text Summarization Evaluation at NTCIR Workshop 3, *Working Notes of the 3rd NTCIR Workshop Meeting*, PART V, pp.1–7 (2002).
4) 平尾　努，奥村　学，磯崎秀樹：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 (2006).
5) Hovy, E., Lin, C.-Y., Zhou, L. and Fukumoto, J: Automated Summarization Evaluation with Basic Elements, *Proc. 5th Conference on Language Resources and Evaluation* (2006).
6) 乾健太郎，藤田　篤：言い換え技術に関する研究動向，自然言語処理，Vol.11, No.5, pp.151–198 (2004).
7) Kauchak, D. and Barzilay, R: Paraphrasing for Automatic Evaluation, *Proc. 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp.455–462 (2006).
8) Lee, L: Measures of Distributional Similarity, *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, pp.25–32 (1999).
9) Lin, C.-Y. and Hovy, E: 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 (2003).
10) Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, *Proc. ACL-04 Workshop “Text Summarization Branches Out”*, pp.74–81 (2004).
11) Lin, D.: Automatic Retrieval and Clustering of Similar Words, *Proc. 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pp.768–774 (1998).
12) Nanba, H. and Okumura, M: 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).
13) 難波英嗣，平尾　努：テキスト要約の自動評価，人工知能学会誌，Vol.23, No.1, pp.10–16 (2008).
14) Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation, *IBM Research Report*, RC22176 (W0109-0220) (2001).
15) Salton, G.: *The SMART Retrieval System. Experiments in Automatic Document Processing*, Prentice-Hall, Inc., Upper Saddle River, NJ (1971).
16) 海野裕也，宮尾祐介，辻井潤一：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).
17) Utiyama, M. and Isahara, H.: Reliable Measures for Aligning Japanese-English News Articles and Sentences, *Proc. 41st Annual Meeting of the Association for Computational Linguistics*, pp.72–79 (2003).
18) Zhou, L., Lin, C.-Y., Munteanu, D.S. and Hovy, E.: ParaEval: Using Paraphrases to Evaluate Summaries Automatically, *Proc. 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp.447–454 (2006).

(平成 21 年 12 月 20 日受付)
(平成 22 年　4 月　7 日採録)

(担当編集委員　相良 毅)

平原　一帆

2008 年広島市立大学情報科学部知能情報システム工学科卒業．2010 年広島市立大学大学院情報科学研究科博士前期課程修了．同年株式会社日立システムアンドサービス入社．現在に至る．修士（情報工学）．

難波　英嗣（正会員）

1996 年東京理科大学理工学部電気工学科卒業．1998 年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．2001 年同大学情報科学研究科博士後期課程修了．同年日本学術振興会特別研究員．2002 年東京工業大学精密工学研究所助手．同年広島市立大学情報科学部講師．2010 年広島市立大学大学院情報科学研究科准教授．現在に至る．博士（情報科学）．テキストマイニング，情報検索，自動要約，特許情報処理に関する研究に従事．言語処理学会，人工知能学会，ACL，ACM 各会員．

竹澤　寿幸（正会員）

1984 年早稲田大学理工学部電気工学科卒業．1989 年同大学大学院博士後期課程修了．同年（株）国際電気通信基礎技術研究所入社．2007 年広島市立大学大学院情報科学研究科教授．現在に至る．工学博士．音声対話翻訳の研究開発に従事．平成 18 年度電子情報通信学会 ISS 論文賞受賞．電子情報通信学会，人工知能学会，日本音響学会，言語処理学会各会員．

奥村　　学（正会員）

1962 年生．1984 年東京工業大学工学部情報工学科卒業．1989 年同大学大学院博士課程修了．同年東京工業大学工学部情報工学科助手．1992 年北陸先端科学技術大学院大学情報科学研究科助教授，2000 年東京工業大学精密工学研究所助教授，2009 年東京工業大学精密工学研究所教授，現在に至る．工学博士．自然言語処理，知的情報提示技術，語学学習支援，テキスト評価分析，テキストマイニングに関する研究に従事．人工知能学会，AAAI，言語処理学会，ACL，認知科学会，計量国語学会各会員．