言い換えを用いたテキストの自動評価

平原一帆[†]　難波英嗣[†]　竹澤寿幸[†]　奥村学[††]

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ひと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，生成要約の評価において従来の評価手法を改善できることが確認された．

Automatic Evaluation of Texts by Using Paraphrase

Kazuho Hirahara[†]　Hidetsugu Nanba[†]
Toshiyuki Takezawa[†]　Manabu Okumura[††]

How to evaluate computer-produced abstract-type summaries has been recently recognized as one of the important research problems to solve in the field of automatic summarization. Traditionally, computer-produced extract-type summaries have been evaluated by n-gram overlap with human-produced summaries. However, these methods cannot evaluate abstract-type summaries in the same performance as extract-type summaries. In this paper, we explore the use of paraphrasing methods for refinement of automatic evaluation techniques. To confirm the effectiveness of our method, we conducted some experiments using the data of the Text Summarization Challenge 2. As a result, we found that our method could improve a traditional evaluation method when evaluating abstract-type summaries in comparison with a traditional evaluation method.

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，自動化によって行われることが求められてきた [11]．

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．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，言い換えを考慮する手法を提案する．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，本研究における提案手法について説明する．4 節では実験内容について言及し，5 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

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

に，代表的な評価手法である BLEU と ROUGE について説明する．

BLEU[12]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[8]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．しかし，要約評価の場合再現率が重要となるため，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[9]という尺度が Lin により提案されている．

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自動評価手法である．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 グラム集合を現す．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．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関連研究

同義語を自動的に抽出する研究に，海野らの研究[14]および相澤の研究[1]がある．海野は，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た[14]．彼らはアライメントのとれた二言語対訳コーパスを用意し，同じ単語とアライメントのとれた単語を言い換え表現と見なした．例えば日本語の「二酸化炭素」と「炭酸ガス」は両方とも英文中で「carbon dioxide」とアライメントがとられることが多い．このとき「carbon dioxide」をピボットとして，「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．海野らのとった言い換えの自動獲得手法は，本研究における言い換え知識獲得の一つの方法として使用している．

海野らと同様に，相澤は同義語について自動獲得と考察を行っている[1]．テキストから語の関係を自動抽出する方法として，共起語に注目しテキスト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がある[7][10]．相澤はこれについて，大規模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べた．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のもと，ノイズ低減のためにフィルタリング法，サンプリング法の 2 つの方法を提案し，提案手法の有効性を確認した．本研究では，この大規模コーパスを用いた分布類似度の使用を一つの方法として，言い換え知識の獲得を行っている．

海野らの言い換えの自動獲得手法による言い換えを用いて，テキストの自動評価する手法として，ParaEval[15]が提案されている．ParaEval は ROUGE 同様，参照要約とシステム要約を比較する自動評価手法であり，大域的には最適マッチ，局所的には最長マッチとなる探索を行うことで，言い換えマッチングを段階的に行う．すなわち，第一段階では動的計画法に基づきフレーズ対フレーズによる言い換えマッチングを行う．第二段階では，第一段階で一致しなかった語に対し，貪欲法に基づいて単一語対フレーズ, 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単語に対して，ROUGE と同様の語彙マッチングを行う．Liang らは，ParaEval の評価と人間の評価との相関が ROUGE のそれと似ているということを示し，提案手法の有効性を確認した．本研究での言い換えを用いたテキストの自動評価法の概形は，この ParaEval に準ずる形で作成している．

同様に同義語を用いて自動要約評価する研究に Kauchak と Barzilay の研究がある．この研究では機械翻訳評価の際に，文脈を考慮した言い換えの評価が行われることに着目し，自動要約評価の改善について提案した[6]．参照要約の言い換えのうち，システム要約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た．適切と判断された言い換えを用いて，複数の参照要約を生成し，自動評価における新たな参照要約としてこれを用いた．言い換えられた新しい要約を参照要約とすることで，最初の参照要約のみを用いた評価に比べ，人手により近い評価が行えることを示した．

3. 提案手法

3.1 提案手法概要

従来の自動評価手法では評価の難しい，独自の表現を含む生成に基づくテキストを評価するために，本稿では ParaEval と同様の手法を用いて言い換えを考慮する．参照要約とシステム要約を比較する際，従来手法と同様の語彙マッチングを行う前に，互いの要約の間に言い換えが含まれていれば，それを同じ単語と見なすことで言い換えを考慮する．要約の探索と単語のマッチングは，以下の手順で行う．

(1)パラフレーズ対フレーズを走査し，フレーズから成る言い換えの一致を貪欲法

に基づいて検索する．
(2) (1)で一致しなかった語に対して，単一語対フレーズ，または単一語対単一語を走査し，同義語の一致を貪欲法に基づいて検索をする
(3) (1)，(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を数え，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

Liang らによる 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を用いて同義語辞書を作成した．本稿では，さまざまな精度・規模の言い換え知識を用いて言い換えによる自動評価を行うことで，言い換えを用いた自動評価においてより有効な言い換え知識について検討する．

本稿で用いた言い換え知識の獲得法について，以下に言及する．

■統計的機械翻訳によるフレーズテーブル

Liang 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を用いて同義語辞書を作成した．複数の原言語フレーズがある一つの目的言語フレーズに翻訳されるとき，複数の原言語同士は同じ意味を持つフレーズ同士であるという考えに基づき，同義語辞書を作成した．

■分布類似度

名詞と動詞の係り受け関係，名詞句と動詞の係り受け関係を抽出することで，単名詞，名詞句，動詞に関して，類似度の分布を作成する．類似度尺度には SMART[13]を，係り受け関係の抽出には CaboCha による係り受け解析を用い，読売新聞，毎日新聞，日本経済新聞計 56 年分のデータを利用している．この分布類似度の高い単語同士を言い換え知識と見なし，同義語辞書を作成した．

■WordNet

概念辞書である WordNet[2]は，単語が 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．この WordNet において位置づけられている概念を言い換え知識と見なし，同義語辞書を作成した．

■NTT 日本語語彙大系

NTT 日本語語彙大系の単語大系の異表記項目を用いて，「１人」「一人」「独り」，「戦う」「闘う」「たたかう」などの異表記を言い換え知識と見なし，同義語辞書を作成した．

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	構築方法
フレーズテーブル	自立語・付属語を含む任意の単語列	自動
分布類似度	名詞・名詞句・動詞	自動
WordNet	名詞・動詞	手動
NTT 日本語語彙大系	名詞・動詞・形容詞	手動

4. 実験

3 節で述べた手法の有効性を調べるために実験を行った．

4.1 実験方法

実験方法として，実験に用いた要約データ・言い換え知識の作成，評価尺度，比較手法について説明する．

■要約データ

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%の要約を作成した計 600 要約から成る．

要約作成者 20 名のうち，10 名は社説原文からの抜き出しのみによる要約を作成し，10 名は自由作成による要約を作成した．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 名の評価者が採点基準に則って，全ての要約に対して 100 点を満点として要約の品質に対する評価を行った．

■実験に用いた言い換え知識

- **統計的機械翻訳におけるフレーズテーブルから作成した言い換え知識**
  統計的機械翻訳の過程で生成されるフレーズテーブルから言い換え知識を獲得した．この統計的機械翻訳については，言語モデルの作成には SRILM を，翻訳モデルの作成には Giza++を，デコーダには Moses を用いている．また，対訳コーパスとして，読売新聞 150,000 日英対訳文対とロイター通信 56,872 日英対訳文対を用い[*1]，言い換え 1,136 万対を集積した．
- **分布類似度を用いた言い換え知識**
  実験に用いた要約データ内に出現する名詞・名詞句・動詞について，分布類似度が高い単語 20 件のうち，要約データ内に出現する単語を言い換えとし，4583 対を集積した．
- **WordNet を用いた言い換え知識**
  実験に用いた要約データ内に出現する名詞・動詞について，synset である

*1 http://www2.nict.go.jp/x/x161/members/mutiyama/index-ja.html

単語のうち，要約データ内に出現する単語を言い換えとし，7873 対を集積した．

■評価尺度

実験の評価手順として，以下の手順に従って評価を行った．

3 名の評価者が決定したマニュアル評価の算術平均値と標準偏差を元に，4 段階の評価を決定した．このマニュアル評価結果と，自動評価結果とのスピアマンの順位相関係数を求めた．

ただし，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 名の評価者によるスコア付けが著しく異なっている要約は，人手による評価が難しい要約であると判断し，今回の実験データから除外した．
- 今回の要約データは，抜き出しにより作成された要約（以後，抜粋要約）と，自由作成により作成された要約（以後，生成要約）がある．言い換え知識を用いた自動評価の有効性を確認するため，実験を行う際にこれらの要約を区別して評価を行った．
- 自動評価に必要となる参照要約については，各テーマにおいて 3 名の評価者による評価平均が最も高い要約を参照要約と見なして自動評価を行った．なお，抜粋要約を参照要約と見なした場合と，生成要約を参照要約として見なした場合を区別して評価を行い，参照要約に関する検討を行う．
- マニュアル評価の階調を 4 段階に変える際，あるテーマのマニュアル評価が全て同階調だった場合，順位相関係数を求めることができない．この場合に関しては，マニュアル評価が全て同階調であるということは，評価に甲乙を付け難いと判断し，順位相関係数を 1 とした．

■比較手法

言い換え知識として，以下に示す 6 種類の言い換え知識を用いた．

(1) 統計的機械翻訳によるフレーズテーブル（表記：SMT）
(2) NTT 日本語語彙大系（表記：NTT）
(3) WordNet
(4) 分布類似度（表記：DS[*1]）
(5) (2)+(3)の統合言い換え知識
(6) (2)+(3)+(4)の統合言い換え知識

また，言い換えを用いる自動評価との比較するベースライン手法として，文字列の一致のみを評価する ROUGE-1 を用いた．

*1 分布類似度　Distributional Similarity

4.2 実験

提案手法の有効性を確認するため，以下の実験を行った．

■言い換え知識比較実験

言い換え知識を用いた自動評価について，言い換え知識の品質や規模が，評価にどのような結果を与えるかを比較検討する．

抜粋要約と生成要約それぞれに対して自動評価を行い，マニュアル評価とのスピアマンの順位相関係数を要約 30 テーマ算出した平均を表 2，表 3 に示す．ここで，表 2 については参照要約として抜粋要約を用いており，表 3 については参照要約として生成要約を用いている．

表 2　抜粋要約を参照要約とした言い換え知識比較結果

	言い換え知識	抜粋要約	生成要約
言い換え知識を用いた自動評価	(1)SMT	0.294	**0.330**
	(2)NTT	**0.375**	**0.322**
	(3)WordNet	0.350	**0.327**
	(4)DS	0.356	0.281
	(5)(2)+(3)	0.343	**0.329**
	(6)(2)+(3)+(4)	**0.361**	**0.325**
ベースライン	ROUGE-1	0.358	0.310

表 3　生成要約を参照要約とした言い換え知識比較結果

	言い換え知識	抜粋要約	生成要約
言い換え知識を用いた自動評価	(1)SMT	0.255	0.358
	(2)NTT	0.311	**0.398**
	(3)WordNet	0.311	0.374
	(4)DS	0.310	0.324
	(5)(2)+(3)	0.309	0.378
	(6)(2)+(3)+(4)	0.295	0.375
ベースライン	ROUGE-1	0.313	0.389

■閾値比較実験

本研究で用いた言い換え知識の中で，統計的機械翻訳によるフレーズテーブルから作成した言い換え知識と，分布類似度から作成した言い換え知識は，それぞれ翻訳確率と分布類似度，すなわち，同義語（翻訳語）でありやすさを数値で示すことができる．本実験ではこれを利用し，言い換え知識の同義語でありやすさを閾値によって操

作することで，精度を高めた同義語と自動評価結果との関係を調べた．

統計的機械翻訳のフレーズテーブルには，翻訳確率が付記されている．例えば，A から B の翻訳確率が 0.5 であり，B から C の翻訳確率が 0.2 であれば，A から C の同義語を作成した際の確率を 0.5*0.2=0.1 と考えることができる．これを同義語確率と定義する．なお，同義語確率は 1 以下の値を取り，数値が大きいほど精度の高い同義語であると考えられる．数値が大きくなり過ぎると，使用できる同義語がなくなり，ROUGE の値へと収束していく．分布類似度も同様に，数値が大きいほど精度の高い同義語であると考えられ，約 16 程度で ROUGE の値へと収束する．

表 4 に，統計的機械翻訳の同義語確率別の自動評価結果を示す．結果数値は参照要約に抜粋要約を用い，前項と同様に自動評価とマニュアル評価とのスピアマンの順位相関係数を要約 30 テーマに対し算出した平均値である．

表 4　同義語確率別 SMT 言い換え知識比較結果

	閾値	抜粋要約	生成要約
SMT 閾値	閾値無し	0.294	**0.330**
	0.0001	0.347	**0.322**
	0.001	**0.378**	0.276
	0.01	**0.368**	**0.331**
ベースライン	ROUGE-1	0.358	0.310

同様に，分布類似度別の自動評価の結果を表 5 に示す．

表 5　分布類似度別言い換え知識比較結果

	閾値	抜粋要約	生成要約
分布類似度 閾値	閾値無し	0.356	0.281
	2	**0.371**	0.306
	4	**0.363**	0.308
	6	**0.365**	0.305
	8	**0.359**	0.309
	10	**0.360**	0.308
	12	**0.360**	0.309
	14	**0.359**	0.309
	16	0.358	0.310
ベースライン	ROUGE-1	0.358	0.310

4.3 考察

■言い換え知識比較実験

今回使用した言い換え知識においては，(2)の NTT 日本語語彙大系を用いた言い換えを言い換え知識として用いることで，ベースラインと比較して，抜粋要約・生成要約いずれに対しても評価が改善される傾向にある．これは，NTT 語彙大系は言い換えの中でも異表記項目について言い換え知識を作成していることに依るものと考えられる．異表記項目では単語として大きく意味の変わるものが存在していないため，文章の意味合いを取り違えることなく評価を行うことができる．要約作成者がコンピュータを用いた入力により要約を作成したため，「取り組み」を「取組み」と変換したり，「ヶ月」を「ヵ月」と変換したりする場合に対応でき，抜粋による要約の評価であっても改善されたのだと考えられる．また，要約という性格上，文字数制限が設けられているため，「こと」を「事」，「さまざま」を「様々」など，漢字を用いて文の短縮を図ろうとした場合にも評価を行うことが可能になる．ただし，単語として大きく意味が変わらないため，本来の目的である同義語を用いた評価からは少々逸脱する．表記の問題が改善されるため辞書として用いる重要性は大きいが，根本的な目的の達成に直結していないとも言える．

(1)統計的機械翻訳によるフレーズテーブルからの言い換えと，(4)分布類似度による言い換えは，自動的に収集される代わりに精度が低いことが欠点であり，今回の実験ではベースラインを下回ることもままあった．単語のみを用いる辞書でなく，フレーズ単位での言い換えが豊富なメリットを生かすため，より精度の高いフレーズテーブルの作成が期待される．

(3)WordNet による言い換えは精度の高い言い換えとなるが，単語のみの言い換えであることや，多義性が多岐にわたるため，「存在」と「世界」，「市」と「フェア」など，その要約のテーマ上関係のない単語を言い換えてしまう傾向が問題点として挙げられる．要約のテーマ上や，一般的な言い換えに対する区別を可能にすることで改善が考えられる．

抜粋要約・生成要約の側面から考察を行う．今回の実験から，言い換えを用いた自動評価を行うに当たって，抜粋によって作成された参照要約を用いて，生成要約を評価する際に多くの言い換え知識で従来手法を上回っており，最も有効に働くということが確認された．

■閾値比較実験

本実験は，前項で述べた自動的に収集される言い換え知識に対する改善策として，類似度や翻訳確率を用いて精度の調整を行ったものである．今回の結果からは，一概にどの程度の閾値を設ければ良いということを決定するのは難しいが，閾値を設けることによりベースラインを超える自動評価を行うことが可能であるということが確認された．

5. おわりに

本研究では，従来手法の問題点を指摘し，表層的な文字列の一致だけでなく，言い換えを考慮することにより，従来のテキスト評価手法を改良する手法を提案した．また，この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．実験により，自動評価に用いる言い換え知識の模索を行い，自動評価に有効な言い換えを提示した．実験の結果，NTT 日本語語彙大系の異表記項目を言い換え知識として用いたときに，従来手法を平均 0.009 上回った．また抜粋要約を参照要約として生成要約を評価する際に，統計的機械翻訳に基づく言い換えを用いることで，最も高い 0.02 の改善が得られた．また全体として，従来手法に比べ，自由作成による要約に対して提案手法がより有効であるということが確認された．さらに，自動的に収集される言い換え知識の改善の可能性と，自動評価の精度が向上する傾向があることを示し，本提案手法の有効性を確認した．

6. 謝辞

言い換え知識の獲得について議論していただいた公立はこだて未来大学の藤田篤氏に感謝致します．

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., 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 Workshop3, *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) Kauchak, D., Barzilay, R: Paraphrasing for automatic evaluation. *Proc. the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp.455-462 (2006).
7) Lee, L: Measures of Distributional Similarity, Proc. 37th Annual Meeting of the Association for Computational Linguistics, pp.25-32 (1999).
8) Lin, C.-Y., 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).
9) Lin, C.-Y: ROUGE: A Package for Automatic Evaluation of Summaries. *Proc. the ACL-04 Workshop "Text Summarization Branches Out"*, pp.74–81 (2004).
10) 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).
11) 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).
12) Papineni, K., Roukos, S., Ward, T., Zhu, W.-J: BLEU: a Method for Automatic Evaluation of Machine Translation, *IBM Research Report, RC22176 (W0109-0220)* (2001).
13) Salton, G: The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River, NJ, (1971).
14) 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).
15) Zhou, L., Lin, C.-Y., Munteanu,D.S., Hovy, E: ParaEval: Using Paraphrases to Evaluate Summaries Automatically. *Proc. the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp.447-454 (2006).