情 報 学 基 礎 61−18
自 然 言 語 処 理 142−18
（2001. 3. 6）

同義テキストの照合に基づくパラフレーズに関する知識の自動獲得

村田 真樹　　井佐原 均

総務省 通信総合研究所
けいはんな情報通信融合研究センター
〒619-0289 京都府相楽郡精華町光台 2-2-2
TEL:0774-95-2424　FAX:0774-95-2429
{murata,isahara}@crl.go.jp

あらまし

近年，パラフレーズに関する知識獲得の研究が重要視されつつある．本稿では，同義のテキストを照合し，その照合結果を用いてパラフレーズに関する知識を自動獲得することを試みた．この自動獲得の実験を辞書定義文，新聞記事タイトル・本文対，講演テキストにおいて行なったところ，同義のテキストの照合による方法がパラフレーズ獲得にある程度役に立つことがわかった．

キーワード　同義テキスト，言い換え (言い替え)，辞書定義文，話し言葉，要約

Automatic Paraphrase Acquisition Based on Matching of Two Texts with the Same Meaning

Masaki Murata　　Hitoshi Isahara

Keihanna Human Info-communication Research Center,
Communications Research Laboratory
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
TEL:+81-774-95-2424　FAX:+81-774-95-2429
{murata,isahara}@crl.go.jp

Abstract

Recently, research on paraphrase acquisition has been considered to be important. We acquired paraphrases by matching two texts which have the same meaning. We did these acquisition experiments by using dictionary definition sentences, the spoken text, and pairs of headlines and texts in newspapers. We found that matching of two similar texts is useful for paraphrase acquisition to some extent.

key words　Texts with the Same Meaning, Paraphrase, Dictionary Definition Sentence, Spoken Language, Summarization

1 はじめに

ついに21世紀を迎えたが，われわれは，21世紀にふさわしい研究テーマとして，人間と同等程度の能力を有する人工知能システムの作成に向け，その第一段階として質問応答システムの研究に着手している[1, 2, 3, 4]．そのシステムでは，基本的には，与えられた質問文の答えが書いてありそうな文を探し出し，その答えが書いてありそうな文と質問文の類似度が大きくなるように双方を書き換えて照合し，答えが書いてありそうな文での，質問文の疑問詞に対応している箇所を答えとして出力するシステムである．たとえば，「日本の首都はどこですか。」という質問文があったとし，情報検索[5]などの技術により答えの書いてありそうな文として，「東京は日本の首都です。」というのが見つかったとする．このとき，「AはBです」を「BはAです」と書き換える変形などを行なって，質問文と答えを含む文をそれぞれ「日本の首都はどこですか。」「日本の首都は東京です。」と，質問文と答えを含む文の間の類似度があがるように変形する．そして，これ以上は言い換えの知識では，類似度をあげられないというところで，双方を照合し，質問文の疑問詞「どこ」に対応する「東京」を解として出すというものである．このシステムは，質問文と解を含む文の一致が大きい状態になったときに疑問詞に対応するものを答えとして出力するので高精度な質問応答を実現できると思われる．しかし，現在のシステムでは，変形規則としてはEDR辞書で同義語とされているもの(例えば，「アメリカ合衆国」と「米」など)しか用いていない．これらは，同義な言い換えを示す，パラフレーズに関する知識がきちんとした形で整備されていないことによる．本稿では，このパラフレーズに関する知識の整備に向けて行なった，同義テキストの照合に基づくパラフレーズに関する知識の自動獲得の研究について述べる．

2 パラフレーズに関する知識の自動獲得の基本的アイデア

ここで自動獲得したいものは，同義な表現対であるので，同義なテキストをもってきて，それらを照合し差分などを調査して同義な表現対を獲得していけばよい[1]

例えば，同義なテキストとして，複数の国語辞典を用意してその定義を利用するということが考えられる．ここでは「あべこべ」という語の定義文を考えてみる．大辞林では，

「順序・位置などの関係がさかさまに入れかわっていること。」

となっており，岩波国語辞典では，

「順序・位置・関係がひっくり返っていること。」

となっている．これを適当に照合すれば，

「さかさまに入れかわっている」

⇕

「ひっくり返っている」

といった言い換えの知識が得られることがわかるだろう．

本研究でのパラフレーズに関する知識の自動獲得の基本的アイデアは，基本的に上記のとおりで，同義のテキストを集めてきて，それらを照合することにより，言い換えの知識を獲得するものである．

3 辞書定義文での研究

本節の研究では，同義テキストとしては異なる辞書の定義文対を用いる．この辞書としては，岩波国語辞典と大辞林を使用した．同義テキスト対としては，二つの辞書の各見出し語の定義文同士を組にすればよいが，場合によっては一つの見出し語が複数の項目をもっている場合がある．これの対処法として，本稿ではそれぞれの定義文が，岩波国語辞典と大辞林とで一対一に対応すると仮定して，照合の度合いがよいもの同士，定義文を結び付けることにした．

まず照合のとりかたであるが，これは各定義文をJUMAN[6]をつかって形態素列に分解する．各行に形態素がくるようにしてUNIXのdiffコマンド[2]を使って，一致，不一致箇所を検出する．照合の度合いを計る式としては，以下のものを用いた．

$$\text{照合の度合い} = \frac{\text{一致文字数} \times 2}{\text{全文字数}} \quad (1)$$

ここで，一致文字数は，diffの結果一致部分と判断された部分の文字数を意味し，全文字数は，diffに与えた岩波国語辞典と大辞林の双方の定義文を合わせた文字数を意味する．この式は，0から1の値をとり，一致部分が大きいほど大きな値を持つものとなっている．

実際に上記の照合を行なった．照合は57,643個の見出し語で行なうことができた．辞書定義文の照合結果の例を表1に示す．表中で“<”,“>”で囲まれた部分は，大辞林にだけ出現したものを，また，“≦”,“≧”で囲まれた部分は，岩波国語辞典にだけ出現したものを意味する．

表をみると，「互いに」と「たがいに」や，「はかなげな」と「たよりない」や，さきほどの「あべこべ」の「さかさまに入れかわって」と「ひっくり返って」といった同義・類義表現が得られていることがわかる．しかし，「急な」と「包み隠さないで、はっきり表す」や，「数量が非常に多い」と「あわただしく動作を急

[1] この基本的アイデアの一例として，複数の国語辞典を用意し，これら複数の辞典の定義文のつき合わせにより同義語・同義フレーズに関する知識(変形規則)を獲得する案はすでに文献[2]において述べている．また，国語辞典に限らず，同義なテキスト対であればよいことについては文献[4]において述べている．

[2] diffコマンドは，sortコマンド[7]と同様に種々の言語処理で役に立つ．また機会があれば，diffと言語処理について書こうと思っている．

表 1: 辞書定義文の照合結果の例

照合の度合い	見出し語	定義文の diff の結果
0.69	あいこ	＜互いに、＞≦たがいに≧勝ち負けのないこと
0.29	あえか	＜はかなげな＞≦たよりない≧さま
0.17	あえか	＜美しくかよわげな＞≦かよわく、なよなよした≧さま
0.20	あからさま	＜急な＞≦包み隠さないで、はっきり表す≧さま
1.00	あさって	あすの次の日
1.00	あしらい	もてなし
1.00	あたふた	あわてふためくさま
0.17	あたふた	＜数量が非常に多い＞≦あわただしく動作を急ぐ≧さま
0.40	あたら	＜惜しい＞≦もったいない≧ことに
0.18	あたら	＜もったいなく＞≦おしく≧も
0.22	あっさり	＜濃かったり、くどかったり、しつこかったりせず、＞さっぱり＜としたさま＞
0.67	あっぷあっぷ	水におぼれかけて＜、もがいている＞≦苦しむ≧さま
0.54	あとり	スズメよりやや＜大形で頭と背面は黒色＞≦大形≧
0.35	あとり	＜日本へは＞秋に≦シベリア地方から日本に≧渡来＜し、全土で越冬＞する
1.00	あべこべ	反対
0.41	あべこべ	順序・位置＜などの＞≦・≧関係が＜さかさまに入れかわって＞≦ひっくり返って≧いる＜・＞こと

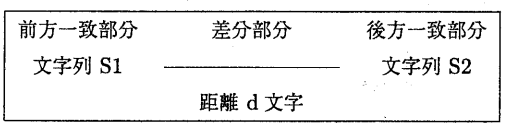

図 1: 差分の出現模式図

ぐ」といった誤った対応のものも見受けられ，この結果をそのまま用いるのは精度が悪そうである．

そこで，次に，diff の結果から，ある程度よさそうな同義・類義表現を抽出することを試みる．ここでは以下の二つの特徴を利用することにする．

- 珍しい (出現頻度の低い) 文字列に囲まれた不一致部分ほど，パラフレーズとしては確からしい．
- 複数箇所に出現した不一致部分ほど，パラフレーズとしては確からしい．

まず，一つめの「珍しい文字列に囲まれた不一致部分ほど，パラフレーズとしては確からしい」という特徴の方を考える．ここでは，差分部分 (不一致部分) が図 1のように，一致部分である文字列 S1,S2 にはさまれていて，S1 と S2 の間が d 文字[3]だけ離れているとする．このとき，S1 および S2 からみて，d 文字以内に図の方向に S2 および S1 が現れる確率を，P(S1)，P(S2) とすると，P(S1)，P(S2) は近似的に以下のように表される．

$$P(S1) \simeq (d+1) * \frac{\text{文字列}S1\text{の出現数}}{\text{文字総数}} \quad (2)$$

[3] 本稿では，この d としては，差分部分の長い方の文字数を採用している．

$$P(S2) \simeq (d+1) * \frac{\text{文字列}S2\text{の出現数}}{\text{文字総数}} \quad (3)$$

このときの差分部分が確からしい確率を P(差分,S1,S2) とすると，P(差分,S1,S2) は S1,S2 がともに図のような形であらわれにくい確率であると仮定すると，以下のようになる．(S1 と S2 が独立であることを仮定している．)

$$P(\text{差分}, S1, S2) \simeq (1 - P(S1))(1 - P(S2)) \quad (4)$$

次に，二つめの「複数箇所に出現した不一致部分ほど，パラフレーズとしては確からしい．」を考える．これは，複数箇所での確率をうまくくみあわせればよい．複数箇所のうち一か所でも正しければ，その差分部分は正しいものとして抽出できると考える．つまり，差分部分が正しい事象は，任意の S1,S2 に対して S1,S2 に囲まれる差分部分がすべて確からしくない場合の余事象なので，差分部分が確からしい確率を P(差分) とすると，それは以下の式で表される．(各差分部分が独立であることを仮定している．)

$$P(\text{差分}) \simeq 1 - \prod_{S1,S2} (1 - P(\text{差分}, S1, S2)) \quad (5)$$

実際にこの式に基づいて抽出結果をソートしてみた[4]．その結果を，表 2に示す．「・」「など」「の」を片一方では省略するなどの規則の他，主格の際の「が」と「の」や，「一〇」と「十」や，「または」と「や」や，「使う」と「使用する」などの同義の言い換え表現

[4] このソートを行なう際，各文の先頭と最後に “^”,“$” をつけてから行なっている．

表 2: 差分部分の抽出結果 (上位 35 個)

頻度	前方一致部分の例	差分部分		後方一致部分の例
408	いる		・	こと$
314	^口やかましく		、	しかりつけたり
87	アロハ		—	シャツ
115	^恐怖	・	や	疲労
102	走者が攻撃		の	資格を失うこと$
206	水		など	が
44	まばたき		を	するさま$
41	ゆとり	が	の	ないほど
30	火成岩	と	に	なる$
48	期待などのため		に	、心臓が激しく打つさま$
31	で		は	、
29	^一尺の	一〇	十	分の一
22	^振るとがらがら		と	音
21	^さえ		も	$
15	手紙を	いう	敬って言う	語$
47	^		その	時代の
10	^自分の家	に	へ	帰る
13	上下	・	または	左右に
13	など	で、	の	直線コース$
12	権利	・	と	義務
13	^また		単に	、橋$
11	原子	または	や	原子団$
14	せる		ための	装置$
9	街道	で	に	一里ごとに
178	^春の七草の一		つ	$
18	自転車		で	の遠乗り$
9	^目	が	を	はっきり
11	^病院など	で	の	、各科の首席
9	^絵	にある	の	ような
19	^ひとり		だけ	で
16	原子		」	$
27	^歯切れのよい	もの	物	をかむ
7	^また、	そういう	その	人$
7	人ずれ		が	して
9	官吏が職務上	使う	使用する	印$

も獲得されていることがわかる.

ここにあげたもの以外に得られたよさそうな同義表現を表 3にあげておく. すでにある同義語辞書にも登録されていそうな単語レベルの同義語だけでなく, 「がうまい」と「に巧みな」のようなフレーズレベルのものから, 「つつ」と「ながら」のような機能的な同義語なども獲得できている.

抽出された差分部分の総数は, 67,632 であった. また, 片一方が空欄の差分の対は, 片一方で単に詳しく述べているだけの場合や, 対応づけの誤りである場合もあり, 同義表現としてはふさわしくない対が多い. そこで, 片一方が空欄の差分を除いた差分部分の総数を調べた. それは 47,648 個であった. この中に含まれる正しい差分部分は, 大雑把に見積もると, 表 1 で両方が空欄でないパターン (表では “>” と “≦” が連続しているバータン) が, 12 個出現してそのうち 9 個が正しいと判断して良さそうであるので, 35,735 (= 47648/12 * 9) 個程度, 正しい同義表現対が得られそうな見込みとなる. 詳細な精度や獲得個数の調査は今後の課題とする.

4 講演コーパスでの研究

現在, 通信総合研究所と国立国語研究所で科学技術振興調整費開放的融合研究推進制度話し言葉の言語的・パラ言語的構造の解明に基づく「話し言葉工学」の構築の一環として開放融合プロジェクトとして日本語話し言葉コーパスを作成している[8]. 本節では, このコーパスのうち, 開放融合プロジェクトで論文の電子版を現時点で作成できている 82 編の学会講演 (全国大会, 研究会などの発表や講演) の部分を利用する.

まず, 本節では, 書き言葉データ, 話し言葉データの用語を以下のように定義する.

表 3: 獲得された同義表現の例

つつ	ながら
すべての	各
一六	十六
哺乳動物	哺乳類
中途	途中
業	職
である	となる
なる	変わる
隔たり	差
つく	到着する
で作った	の
家畜	牛馬など
がうまい	に巧みな
大事に	大切に
伝える	伝達する
ために	目的で
はずれている	合わない
食う	食べる
減少する	少なくなる

- 書き言葉データ
  - 論文データ (打ち込み, 82 編, 352,660 文字)
- 話し言葉データ
  - 開放融合のコーパスのうち，上の論文データに対応するもの (330,679 文字)

本節でのパラフレーズ獲得では，この書き言葉データと話し言葉データを照合することによって行なう．本節の場合は，同じ発表内容の，論文と講演を比較することで同義表現の獲得を目指すことになる．しかし，本節の場合は前節と違って，内容はおなじであるが，書き言葉データと話し言葉データということで，データの種類が異なることにより，書き言葉と話し言葉の違いのようなものが獲得される可能性もある．

差分部分の獲得方法は，3節とまったく同じ方法で行なう．diff をとる際には，前の場合は，見出しの項目ごとに diff をとっていたが，今回は論文ごとに diff をとった．また，話し言葉コーパスでフィラーとされている部分などは除いた[5]．

実際に抽出された結果を表 4に示す．表における「データ」「データー」の食い違いは，コーパスの定義によるもので今回は長音をなるべくつける表記にしていたことによる．その他目立つものとしては，「=」は「は」と読むということがわかったり，話し言葉では「という」をいれてやわらかくいう場合があることがわかる．

抽出結果を分析したところ主に以下のものがあった．

1. コーパスの定義によるもの (例えば先にあげた「データ」と「データー」)
2. 表記・読みを与えるもの (例えば先にあげた「=」と「は」)
3. 省略もしくは補完をしているもの (例：「スームージング処理を」と「スームージングを」など)
4. 口語調のもの (例：「という」など)
5. 類義関係のもの (例：「各」と「それぞれ」など)

このうち，「口語調のもの」と「類義関係のもの」についてはよさそうなのを表 5と表 6に集めておいた．口語調のものとしては，「。」と書いているところを「訳ですが」と文をつなげるものなど興味深いものも得られている．同義関係のものとしては，やはり論文と講演ということだけあって研究がらみの同義表現が得られている．

[5] 本研究の内容は 2001 年の 2/28,3/1 の開放融合のワークショップ「話し言葉の科学と工学」で発表する予定である．詳細はそちらを参照のこと．

表 5: 口語調のもの

書き言葉	話し言葉
	という
した。	いたしました
、	ですね
	です
られる。	られます
	っていう
や	とか
	こう
	非常にこう
いる。	います
分かった。	分かりました
ない。	ません
。	訳ですが
	っていうの
れた。	れるんですが
であり、	であって
ことである。	訳ですけれども
ある。	ありますけれども

表 6: 同義関係のもの

書き言葉	話し言葉
や	とか
論文	研究
各	それぞれ
i 番目のターム	ターム I
述語	動詞
識別	認識
異なれば、	違えば

5 新聞記事のタイトル・本文対での研究

新聞のタイトルと本文は，分量は違うが，同内容のものを記述していると考えられるので，これらを照合してもパラフレーズの知識が得られると考えられる[6]．本節で

[6] 関連研究として，新聞のタイトルと本文の照合をして，新聞のタイトルを，本文を用いてわかりやすく言い換えるというのがある(9)．

表 4: 書き言葉データと話し言葉データの照合結果の例 (上位 20 個)

前方一致部分の例	書き言葉	話し言葉	後方一致部分の例
ＩＰＡＬの形容詞	、		形容動詞の
接続型		の	テキスト音声合成システムに
ばたばた	・		ばたばた
市内料金	の		引き上げでＮＴＴ株は百万円になる
引き上げでＮＴＴ株は百万円になる	」		と予言したとか
および用言の	ガ	が	格情報を付与したコーパスの作成
本文間のハイパーリンク		を	自動生成
国際人	を		多く輩出してほしいと願わずにはいられない
Ｙは	）		Ｘと事情が違う
その結果を人手		で	修正していくプロジェクトについて
合計強度がＮ	-{		ϵ
経験的損失は全学習	データ	データー	に対する損失の
番組	を	の	最初から最後まで
割近い	クラスタ	クラスター	が意味
陳述を意味	と		する表現
神経経路は視覚	—		発話の神経経路より強く結合していると推測される
ＶＱコードブックの	２	二	種類の話者モデル
１０．９	、	と	１４．３
改善される		という	ことを
焦点	＝	は	述語または主題以外の格要素

は，新聞のタイトルと本文を照合することにより，パラフーレズの知識を獲得する試みについて述べる．

ここでは，前節までのような diff を用いた照合ではなく，どれだけ共通単語が出現するかを利用した照合を行なってみる．diff を用いた照合の場合は，出現の順序 (語順) が同じであることを前提にしていたが，どれだけ共通単語が出現するかを調べる方法では，語順が変わってもよいということで，前節までのものとは異なったものが得られると予想される[7]．

照合の仕方は，以下のとおりである．

1. 新聞タイトルとそれに対応する新聞本文を形態素解析して，形態素列に分解する．
2. 新聞本文が複数の文からなる場合は，新聞タイトルとの共通語が最も多い文を選択する．
3. 新聞タイトルと新聞本文を照合し，共通単語を対応づける．このとき，共通単語としては，簡単のため名詞のみを利用した．

この照合を実際に行なった．新聞としては，毎日新聞[11] の 91 年から 98 年のものを用いた．照合結果の例を表 7にあげる．表中の “≦”,“≧” で囲まれた部分は，タイトルと本文の共通単語を意味する．

[7] これに関連する研究として，われわれはすでに辞書の見出しと，定義文を，共通単語の出現を利用した方法で照合することで，複合語の語構成などを調べる研究を行なっている[10]．例えば，「アマチュア無線」という見出し語の定義文が，「アマチュアによる無線通信」となっているとした場合，「アマチュア」と「無線」という共通語を利用して，「アマチュア無線」という語は，「アマチュア」と「無線」という語が，「による」という意味関係で結び付いた複合語であることがわかる．

次にここから，パターンを取り出すために共通部分を変数におきかえる．このとき，連続している共通部分は連続したものをまとめて変数におきかえる．たとえば，表の “≦アジア≧新≦秩序≧≦づくり≧に≦全力≧” の場合は，“X1 新 X2 に X3” にする．次に変数をすべて含む最小の部分をタイトル，本文より取り出し，それをパターン対とする．例えば，表の 2 行目のデータは，“X1 新 X2 に X3” ⇔ “X1 の新 X2 に X3” となる．共通部分よりも外のものは対応せず，共通部分に囲まれたもののみ同士は対応すると仮定するわけである．

照合結果から上記の要領でパターンを取り出し，このパターンを頻度でソートした結果を表 8に示す．(表の［1］や［2］は変数を意味する．) パターンの数は，183,808 個であった．頻度 2 以上のパターンの数は 7,166 個であった．頻度がある程度ないとパターンの信頼度が低いが，頻度 2 以上のパターン数が七千ほどあるので，ある程度の信頼度をもったパターンが多く得られていることになる．表より，タイトルでは「の」，「を」，「が」などを削除するものや，表の最後にあげたような語順がかわるパターンが存在することがわかる．ここでは，新聞のタイトルと本文の照合ということを行なったため，新聞のタイトルの生成に役に立ちそうなパターン対が得られているように見受けられる．

6 同義テキスト以外での研究

今までは照合するもの同士が同義であることが保証されていたものであった．本節では，照合するもの同士が同義であることが保証されていない場合のものを扱う．

表 7: 新聞記事のタイトルと本文の照合結果の例

タイトル	本文
≦激動≧続くたそがれ≦年≧ ≦アジア≧新≦秩序≧≦づくり≧に≦全力≧	——≦激動≧の九〇≦年≧を受け、九一年はどんな年か。 外交では欧州の緊張緩和を受け、今年は≦アジア≧の新≦秩序≧≦づくり≧に≦全力≧を挙げる決意を示し、九日からの韓国訪問では昨年五月、盧泰愚（ノ・テウ）大統領来日時に合意した「未来志向」の日韓関係を発展させることに期待を表明した。
≦北方≧≦領土≧、≦決断≧望む	また、四月のゴルバチョフ・ソ連大統領の来日では大統領が≦北方≧≦領土≧問題で「勇気ある≦決断≧」をするよう促した。

表 8: 抽出されたパターン対の例 (上位のもの)

頻度	タイトル	本文	[1] の例	[2] の例
2101	[1] [2]	[1] の [2]	千回目	講演
1908	[1] <第 [2]	[1] 第 [2]	全日本オープン	一日
1786	[1] [2]	[1] [2]	井手	顕氏
1062	[1] < [2]	[1] ([2]	国土庁	１４日
755	[1] <最 [2]	[1] 最 [2]	優勝戦大阪大会	終日
572	[1] [2]	[1] を [2]	得点結果	公表
379	[1] [2]	[1] が [2]	円	急騰
317	[1] 、 [2]	[1] の [2]	溶岩	崩落
286	[1] [2]	[1]) [2]	PKO	関連法案
241	[1] ・ [2]	[1] の [2]	三菱石油	山田菊男
..	..	..	..	..
27	[1] [2]	[2] を [1]	ジャズ専門	チャンネル

これは，照合度が高い対ならば同じことを書いている可能性は高いと予想され，わざわざもともとの入力が同義であることが保証されていなくてもよいのではないかと考えて行なった研究である[8]．

ここでは簡単のため，前方一致部分の形態素数，不一致部分の形態素数，後方一致部分の形態素数を 3, 1, 3 に固定して抽出を試みた結果を表 9にあげる[9]．データは毎日新聞の 91 年のもののみを用いた．抽出されたパターン (ここでは前方一致部分と後方一致部分の組を単位とするものとする) は，73,168 個であった．

この結果を見ると，類義表現としては，数の集合，地名の集合，「見方」「考え」「見解」「認識」など，よさそうな結果が得られている．まわりの環境 (本研究の場合は，前方一致部分と後方一致部分) が似ているもの同士は類似しているという特徴を利用した類義表現獲得の研究は数多くあり(12)，その意味ではそれほど目あたらしい結果ではないが，類義表現の獲得には十分使えそうな結果である．

同義表現としては，獲得された「8」と「１２」と

表 10: (3)-(0/1)-(3) のパターンで抽出した結果の例

前方一致部分	不一致部	後方一致部分
姜　錫　柱	・	第一　外務　次官
臨時　行政　改革	推進	審議会　（　第
目　の　不渡り	手形	を　出し　事実
民族　解放　戦線	」	（　ＦＭＬＮ　）
舞台　に　した	巨額	不正　融資　事件

いったものがたとえ似ていたとしても異なるものであることはまちがいないことから同義表現ではないので，同義表現の獲得の研究にはあまり使えそうにない結果だと思われる．やはり，同義表現の獲得を行なおうとすると，本稿の当初の主張のように，同義テキスト対を持ってきて照合した方がよいようである．

次に，不一致部分の対の片方が空欄になるパターンで上記と同じような実験を行なった．前方一致部分の形態素数，後方一致部分の形態素数を 3, 3 にして，片方の不一致部分の形態素数を 1 にもう片方を 0 にして実験した[10]．この結果の例を表 10にあげる．表から，「・」「手形」などは場合によっては省略してもよさそうであ

[8] 本節のアイデアは，文献(2) の脚注 13 に記述している．

[9] ここでは簡単のため，3,1,3 で取り出しているが，他のパターンも含めて抽出を行ない，3節で示したような確率の評価式などを用いて抽出するなどをするとより精密な抽出，もしくは，より広範囲な抽出となる．

[10] この研究は，もともとは日本語省略現象(13, 14) をすべて洗い出すために始めた研究である．この方法では，日本語の省略現象を大量に抽出することができるので，省略現象の考察にも使えるというわけである．

表 9: 3-1-3 のパターンで抽出した結果の例

頻度	前方一致部分	不一致部分の例	後方一致部分
1207	・ 告別式 は	「8」「12」「17」「26」など	日 午後 1
1149	ため 東京 都	「新宿」「港」「文京」「渋谷」など	区 の 病院
701	的 貿易 交渉	「(」「<」「である」	ウルグアイ ・ ラウンド
557	維持 活動 (	「PKO」「JPKO」	) 協力 法案
464	、 国連 平和	「維持」「協力」	活動 (PKO
329	許 永 中	「氏」「容疑者」「被告」「役員」	(44)
304	」 と の	「見方」「考え」「見解」「認識」など	を 示した 。
287	首脳 会議 (	「ロンドン」「ヒューストン」「ベルリン」など	・ サミット)

るとわかる．これらの場合省略しても意味がかわらないであろうから，一応この手法は同義表現に関する知識の獲得には役立つと思われる．

また，片一方を省略とする表現対であるので，要約の研究のための書き換え規則としても役に立つと思われる．要約の研究のための書き換え規則の獲得の研究としては，原文と要約後のテキストを照合して，要約のための書き換え規則を獲得する研究[15]があるが，片一方を省略とする程度の書き換え規則ならば，わざわざ原文と要約後のテキスト対を持ってこなくても，原文のみからでも自動獲得できそうであることを，本節の結果は物語っている[11]．

7 おわりに

近年，パラフレーズに関する知識獲得の研究が重要視されつつある[12]．本稿では，同義のテキストを照合し，その照合結果を用いてパラフレーズに関する知識を自動獲得する試みを，辞書定義文，講演テキスト，新聞記事タイトルといったもので行ない，それぞれよさそうな結果を得た．また，同義な関係にあるかわかっていないテキストでも，同様な実験を行なったところ，同義表現ではなく類義表現が得られた．このことから，同義表現を獲得するには同義な関係にあるテキストが役にたつことがわかる．

本稿での手法では，構文的な要素をまだ使っていない．構文的な同義表現対を獲得しようと思えば，手法自体に構文的な要素を組み込んでいく必要がある．これは今後の課題である．

また，本稿では書き換え規則となるであろう，同義表現対を抽出しただけで，これをどのような場合に適用してもよいかという規則の条件にあたる部分についてはまったく語っていない．これに対する一つの答えを文献[17]で示す予定である．

[11] といっても，原文と要約後のテキスト対というものがあるのならば，それを使った方が同義性の保証が高いので使うのにこしたことはない．また，原文と要約後のテキスト対はある程度同義性の保証があると思われるので，ある程度同義表現対の獲得にも役に立つと思われる．

[12] 同義表現の知識の重要性は文献[16]においても記述してある．しかし，この文献の同義表現の扱いは，基本的に人手で書かれたルールベースに基づく方法で，手法の応用性，規模の拡大に限界があると思われる．これに対して，本稿ではある種のコーパスからの同義表現の自動獲得というアプローチをとっており，人手のコストを極力おさえるアプローチをとっていることになっている．また，文献[16]の論文表題をわかりやすくいいかえるという課題自体は，安達の研究[9]や5節の研究で，データを新聞タイトル・本文から論文タイトル・本文にかえることである程度自動的に (人手規則なしで) 論文表題をわかりやすくいいかえることができると思われる．

謝辞

本研究には6節を中心に総務省通信総合研究所内山将夫研究員との議論が役にたっている．

参考文献

(1) Masaki Murata, Masao Utiyama, and Hitoshi Isahara, Question answering system using syntactic information, (1999), http://xxx.lanl.gov/abs/cs.CL/9911006.

(2) 村田真樹, 内山将夫, 井佐原均, 類似度に基づく推論を用いた質問応答システム, 自然言語処理研究会 2000-NL-135, (2000), pp. 181–188.

(3) 村田真樹, 内山将夫, 井佐原均, 質問応答システムを用いた情報抽出, 言語処理学会第 6 回年次大会ワークショップ論文集, (2000), pp. 33–40.

(4) 村田真樹, 内山将夫, 井佐原均, 質問応答システム, 質問応答処理方法, 変形規則自動獲得処理方法およびそれらのプログラム記録媒体, (特許出願中, 2000).

(5) 村田真樹, 馬青, 内元清貴, 小作浩美, 内山将夫, 井佐原均, 位置情報と分野情報を用いた情報検索, 言語処理学会誌, Vol. 7, No. 2, (2000).

(6) 黒橋禎夫, 長尾真, 日本語形態素解析システム JUMAN 使用説明書 version 3.6, (京都大学大学院工学研究科, 1998).

(7) 村田真樹, 神崎享子, 内元清貴, 馬青, 井佐原均, 意味ソート msort — 意味的並べかえ手法による辞書の構築例とタグつきコーパスの作成例と情報提示システム例 —, 情報処理学会 自然言語処理研究会 130-12, (1999).

(8) 古井貞熙, 前川喜久雄, 井佐原均, 科学技術振興調整費開放的融合研究推進制度 — 大規模 コーパスに基づく『話し言葉工学』の構築 —, 日本音響学会誌, Vol. 56, No. 11, (2000).

(9) 安達久博, ニュース速報記事の前文情報との照合に基づく見出し文の言い替え, 言語処理学会 第 6 回年次大会, (2000), pp. 231–234.

(10) 村田真樹, 内山将夫, 井佐原均, 辞書定義文を用いた複合語分割 — 語構成情報の抽出と考察 —, 言語処理学会第 6 回年次大会発表論文集, C5-2, (2000).

(11) 毎日新聞社, 毎日新聞 1991-1998, (1998).

(12) D. Hindle, Noun classification from predicate-argument structures, *COLING'92*, (1992), pp. 658–664.

(13) Masaki Murata, Anaphora resolution in Japanese sentences using surface expressions and examples, *Doctoral dissertation, Kyoto University*, (1996), (in Japanese).

(14) 村田真樹, 長尾真, 表層表現と用例を用いた照応省略解析手法, 言語理解とコミュニケーション研究会 NLC97-56, (1998).

(15) 加藤直人, 浦谷則好, 局所的要約知識の自動獲得手法, 言語処理学会誌, Vol. 6, No. 7, (1999).

(16) 佐藤理史, 論文表題を言い換える, 情報処理学会論文誌, Vol. 40, No. 7, (1999).

(17) 村田真樹, 井佐原均, 言い換えの統一的モデル — 尺度に基づく変形の利用 —, 言語処理学会第 7 回年次大会ワークショップ論文集, (2001), (to appear).