自 然 言 語 処 理 148－5
（2002．3．4）

# コーパスから自動抽出した表現パターンを用いる日本語文生成

西村仁志　坂本仁

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所

音声翻訳において、原言語と目的言語の双方の換言処理を行うことで変換処理の負荷を最小限にしようという試みがなされている。そこで、本稿では、目的言語の換言処理として、訳語選択、単語の並び替え、削除および補完の必要な入力を、単言語コーパスから取得したパターン用いて自然な文にする方法を提案する。本手法では、プレインなテキストコーパスからパターンを自動的に取得し、それらのパターンを用いて、入力に近く自然な文であると判断した文を生成する。提案手法の有効性を確かめるために実験システムを構築し、評価実験を行った。その結果、入力全体の57～77%の文について効果があることが分かった。

# Japanese automated generation based on the corpora

Hitoshi Nishimura,　Masashi Sakamoto

Advanced Telecommunications Research Institute International
ATR Spoken Language Translation Research Laboratories

Native speakers can paraphrase the natural sentence from the sentence with some errors.　This paper gives a brief introduction to the method that we extract the pattern from the plain text corpora　and generate the natural sentence from outputs of machine translation. In order to confirm the effectiveness of our method, we developed a prototype Japanese generation system and evaluated its effectiveness. From the result of the experiment, we found that our method is effective on the 57%-77% of whole sentences.

## 1　はじめに

近年、音声翻訳システムが盛んに研究されており、例えば、音声翻訳システムにおいて翻訳の知識を対訳コーパスから自動的に取得する手法[1]が提案されている。しかし、日本語と英語との間の翻訳であれば対訳コーパスを集めることは比較的容易であるが、言語間によっては対訳コーパスを集めることが容易でない場合がある。一方、単言語のコーパスを集めることは、対訳コーパスに比べればまだ容易である。そこで、両言語の対訳知識を必要とする変換処理にはできるだけ少ない知識だけをもたせ、原言語および目的言語の単言語知識を利用した換言処理に重点を置く手法[2]が提案されている。

換言に関する従来の研究の内、誤りを含む入力を誤りのないものにする手法としては、変換処理

で出力された結果と正解をルールとして使用する手法[3][4][5]や、シソーラスを利用しコーパスから類似したものを検索することにより音声認識誤りを訂正する手法[6]がある。

これに対し、我々は、翻訳システムにおける目的言語の生成処理において、入力と正解のペアをあらかじめ用意したりシソーラスを使ったりせずに、コーパスから自動的に得られる情報だけをパターンとして抽出し、抽出したパターンから自然な文を生成する方法を提案する。

## 2 システム概要

提案手法のシステムの構成は、パターン抽出部とパターン適用部に分かれる（図1）。

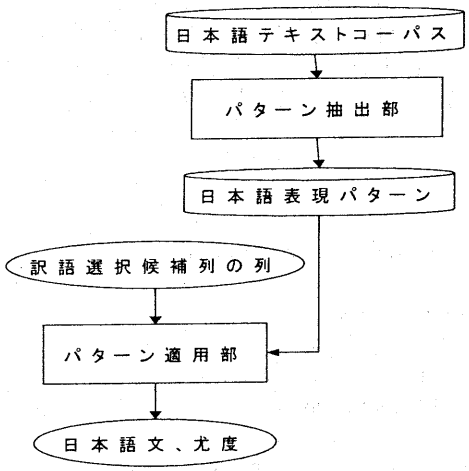


図1システム構成

パターン抽出部は、入力を日本語テキストコーパス（以降、コーパス）とし、テキストを形態素解析し、形態素解析結果から日本語表現パターン（以降、パターン）を構成し、出力する。

パターン適用部は、入力を訳語選択候補列の列とし、パターンを組み合わせて得られる日本語文の中から、入力との近さ（忠実さ）と文としての自然さとを備える文を選び、その選択基準となった尤度とともに出力する。

## 3 パターン抽出部

入力は、一文一行に区切られただけのプレインな日本語テキストである。

処理としては、まず、テキストを形態素解析し、形態素列をいくつかに区切る。次に、区切られた形態素列に対し、他の区切られた形態素列に組み合わせるための位置情報や生成される文の自然さを求めるための情報（3-gram確率）を付加し、これをパターンとして出力する（図2）。ここで、3-gram確率とは、ある形態素が連続する2つの形態素の次に現れる確率をコーパスから求めたものである。

本提案では、パターンを抽出するときの形態素列の区切り方は特定していないが、例えば、構文解析を用いる方法やn-gramを用いる方法がある。

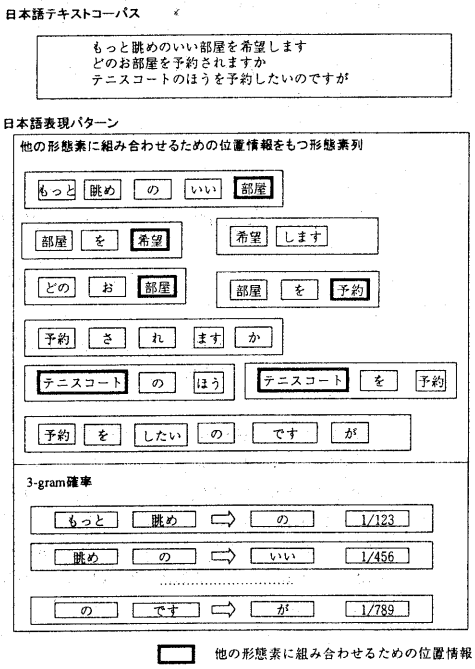


図2 パターン抽出の例

## 4 パターン適用部

入力は、機械翻訳システムの変換処理部が出力するような訳語選択候補列の列である。形式は、

例えば、「((眺め)(予約します)(は)(いい)(部屋 ルーム))」であり、訳語選択候補は形態素解析されていない。

処理としては、まず、訳語選択候補を形態素解析する。次に、パターンを組み合わせることにより、訳語候補の選択、並び替え、不要な形態素の削除、必要形態素の補完を行い、日本語文を得る。得られた日本語文について、生成文としての申し分なさを示す尤度を求め、最も尤度の高い文を出力する。

本提案では、尤度の求め方について特定していないが、現在、4.1節と4.2節で示す「形態素解析済み訳語選択候補列の列と生成される文との距離」と「生成される文の自然さ」を元に求めている。

### 4.1 形態素解析済み訳語選択候補列の列と生成される文の距離

形態素解析済み訳語選択候補列の列（I）とパターンにより生成される文（S）との距離d(I,S)を次のように求める。

$$d(I,S) = (w1 * \sum_{i} \text{iとパターンとの距離} + w2 * \sum_{p} \text{入力とpとの距離}) / I\text{の長さ}$$

ここで、$i$は入力に含まれる形態素、
$p$はパターンに含まれる形態素、
$w1, w2$は重み、
iとパターンとの距離
$= i$の重み$* \min_{\text{パターンに含まれる形態素}\ p}$($i$と$p$の距離)、
入力と$p$との距離
$= p$の重み$* \min_{\text{入力に含まれる形態素}\ i}$($i$と$p$の距離)
iとpの距離 = (if i = p then 0 else 1)
とする。

例えば、図3に示す形態素解析済み訳語選択候補列の列の長さが7であるので、

$d(I,S) = (w1 *$(「もっと」の重み　+「の」の重み
+「を」の重み　+「の」の重み
+「です」の重み　+「が」の重み　)
$+ w2 *$(「します」の重み　+「は」の重み　)
$/ 7$

となる。

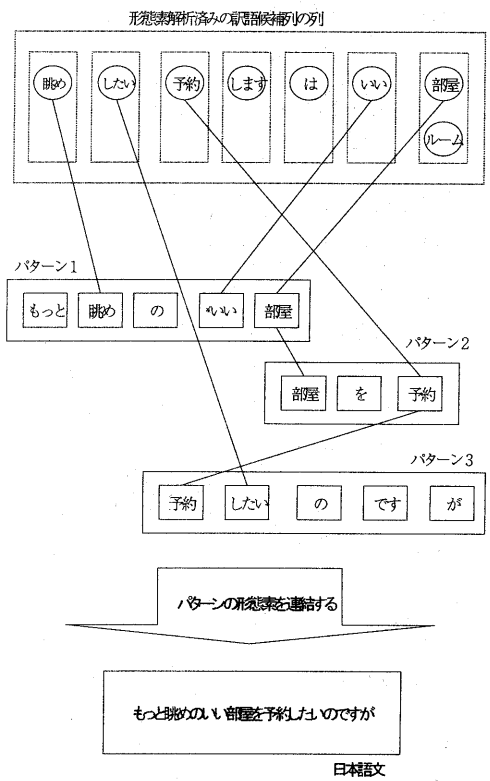


図3 形態素解析済み訳語選択候補列の列と生成される日本語文

### 4.2 生成される文の自然さ

生成される文(S)の自然さN(S)を次のように求める。まず、形態素単位の3-gram確率

$P(W_i | W_{i-2} W_{i-1})$を用いて文のn-gram確率を

$P(W_1^n)$求める。そのあと、相乗平均を取り、計

算を容易にするために対数を取り、コーパス中の文の自然さが0.5~1になるように正規化する。式で、書くと次のとおりである。

形態素$w_1,\ldots,w_n$からなる文$\mathbf{W}_1^n$に対し、

$$N(S)=1-\frac{\log\left(\sqrt[n]{P(\mathbf{W}_1^n)}\right)}{Min}$$

$$\cong 1-\frac{\log\left(\sqrt[n]{\prod_{i=1}^{n}P(w_i|w_{i-2}w_{i-1})}\right)}{Min}$$

$$=1-\frac{\sum_{i=1}^{n}\log\left(P(w_i|w_{i-2}w_{i-1})\right)}{n*Min}$$

ここで、

$w_{-1}$=文頭の単語より2つ前の架空の形態素

$w_0$=文頭の単語より1つ前の架空の形態素

$$Min=\frac{\min_{\substack{3\text{-}gramを取得したテキスト\\ コーパスの文\mathbf{W}_1^m}}\left(\log\left(\sqrt[m]{P(\mathbf{W}_1^m)}\right)\right)}{2}$$

となる。

## 5 実験方法

### 5.1 実験対象

次の4つのコーパスを使用する。

（コーパス1）ホテル担当者と旅行者との対話の実収録を書き起こした日本語と中国語の対訳テキストコーパス

（コーパス2）コーパス1以外のホテル担当者と旅行者との対話の実収録を書き起こした日本語テキストコーパス

（コーパス3）内省により作成した旅行会話の日本語テキストコーパス

（コーパス4）コーパス3の規模を拡大したもの

まず、コーパス1の中国語コーパスを入力とし、中日の変換処理を機械翻訳で行い、その出力である訳語選択候補列の列の中で、日本人であれば容易に完全な生成文を作れると考えられるもの220個を対象として選んだ。なお、選んだ訳語選択候補に対応する日本語テキストの平均形態素数は11.1であり、標準偏差は5.8であり、コーパス1の標準偏差は7.1である。

また、コーパス1、2、3、4から、訳語選択候補を得るために使用した文を削除し、パターン抽出の入力であるコーパスを得る。

取得した訳語選択候補列の列とコーパスを用いて、5.2節で示す条件で、パターン抽出処理とパターン適用処理を行い、日本語文を生成し、それを評価文とする。

表1コーパスののべ文数と平均形態素数

| コーパス | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| のべ文数 | 15810 | 32715 | 19400 | 162281 |
| 平均形態素数 | 12.7 | 11.6 | 9.1 | 6.9 |

### 5.2 実験条件

実験条件として、次に示す5通りを設定した。

（条件1）訳語選択候補の先頭を選びそれらを連結するだけの場合（単純生成）

（条件2）コーパス1を入力とし構文解析し、パターンを抽出した場合

（条件3）コーパス1、2、3を入力とし構文解析し、パターンを抽出した場合

（条件4）コーパス1、2、3、4を入力とし構文解析し、パターンを抽出した場合

（条件5）コーパス1、2、3、4を入力とし6-gram以下の形態素をすべてパターンとして抽出した場合

なお、条件2～5ではパターンを適用して文を生成する。

### 5.3 評価方法

評価者は2人とし、条件1～5で日本語生成された評価文5文と評価文の前後2文づつの計9文（1セット）を同時に提示し、それぞれの評価文に対する絶対評価と、5文を比較する相対評価をしてもらった。

絶対評価は、申し分ないか、そうでないかをA、Bで判定してもらい、相対評価は、5文を申し分なさで、最も良いものを1とし、順位付けをしてもらった。この時、同順位があってもよいとした。

### 5.4 実験結果

実験の結果を図4と図5に示す。

図4は、絶対評価の結果であり、全評価文に対する申し分ないと判定された文の割合である。

また、図5は、相対評価の結果であり、判定結果の順位で、正規化（順位の合計が15になるように順位を配分）し、それをスコア化し平均したものである。

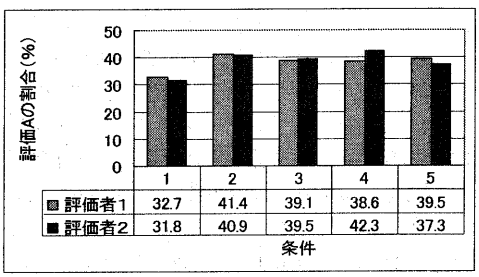


図4 絶対評価

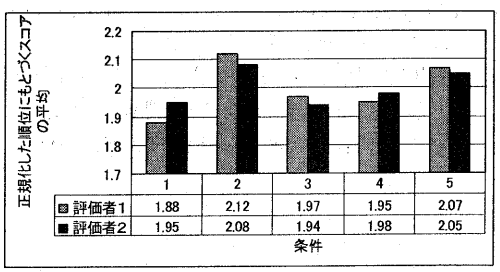


図5 相対評価

絶対評価の結果をAの割合で$\chi^2$検定すると、次のことが言える。

評価者1からは、危険率10%以下で条件1より条件2は良い。

評価者2からは、危険率2.5%以下で条件1より条件4は良い。危険率5%以下だと、さらに、条件1より条件2は良い。危険率 10%以下だと、さらに、条件1より条件3が良い。

相対評価の結果を順位差がある文の割合で$\chi^2$検定すると、次のことが言える。

評価者1からは、危険率2.5%以下で条件1より条件5の方が良い。危険率10%以下で、さらに、条件1より条件2の方がよい。

評価者2からは、何も言えない。

## 6 考察

生成文とともに出力された尤度が大きいほど生成文の評価は良好であった（図6）。なお、図6中の大、中、小で分けた文の割合は条件2～5で異なるが、大：38～43%、中：15～39%、小：23～43%、である。また、訳語選択候補列の列は、中日機械翻訳の変換結果を使用しており、単純生成するだけで絶対評価がAになるものがある。

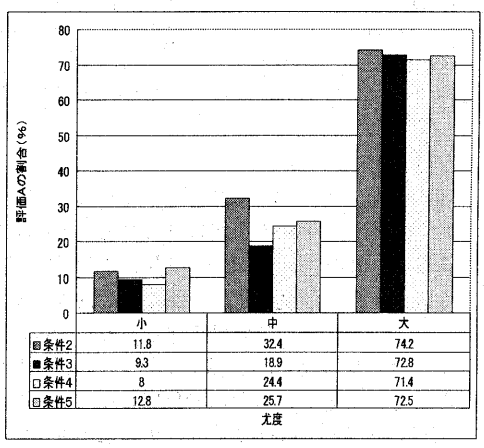


図6 尤度と評価者による評価との関係

そこで、単純生成したときと本手法で生成したときで、文頻度に対しては絶対評価がAである割合を尤度の大きさごとに比べる（図7）。

すると、尤度が小さい文（下位23～43%）については単純生成するほうが、よいことが分かる。

つまり、閾値を設定し閾値より小さい尤度である文に対しては単純生成し、大きい文（上位57～77%）に対しては本手法による生成をするようにシステムを構成すれば、現段階でもさらによい結果が得られる。

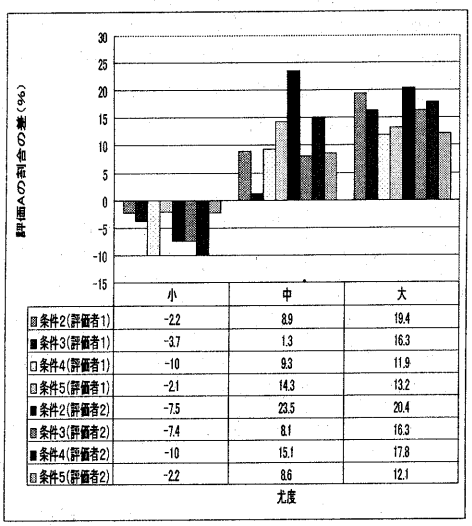


図7 提案手法の生成と単純生成との評価Aの差

## 7 終わりに

本稿では、コーパス中の言語表現を表現パターンして抽出し、それらを組み合わせることにより、訳語選択、並び替え、単語の削除および補完が必要な入力を自然な文に生成する方法を提案した。さらに、提案手法の有効性を確認するために、実験システムを作成し、小規模ながら評価実験を行った。その結果、システムが算出した尤度の小さい文（43～23%）に対しては、訳語選択で先頭を取り、つなげるだけの処理（単純生成）をし、尤の大きい（57～77%）文については本手法による生成を行うことで単純生成に比してより妥当な生成文が得られることが分かった。

今後は、生成方法を切り換える閾値の設定方法について検討するとともに、尤度の精度を上げ、人間の判断に近づけるよう研究を進めたい。

なお、本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

## 参考文献

[1] 隅田 英一郎: An Example-based Machine Translation Using DP-matching between Word Sequences; ACL-2001 (39th Annual Meeting of the Association for Computational Linguistics) Workshop on Data-driven MT (2001)

[2] 山本和英、白井諭、坂本仁、張玉潔：Sandglass:両言語換言機構を基軸とする音声翻訳、言語処理学会第7回年次大会発表論文集、A4-1(2001)

[3]尾崎正行、荒木健治、栃内香次：帰納的学習を用いた自然な日本語文生成手法の評価、情報処理学会、研究報告 「自然言語処理」No.141-10 (2001)

[4] 山本和英：機械翻訳における自動校正と日中翻訳への適用、言語処理学会第5回年次大会発表論文集、pp21-24 (1999)

[5] Kaki, S., Sumita, E., and Iida, H.： A Method for Correcting Errors in Speech Recognition Using the Statistical Features of Character Co-occurrence. In Proceedings of COLING' 98, pp. 653-657 (1998)

[6] 石川開、隅田英一郎：テキストデータを使った音声認識誤りの訂正、言語処理学会会誌、自然言語処理、Vol.7, No.4, pp.205-227, 2000