# キーワードからのテキスト生成*

内元 清貴† 関根 聡‡ 井佐原 均†
†独立行政法人通信総合研究所 ‡ニューヨーク大学
{uchimoto|isahara}@crl.go.jp sekine@cs.nyu.edu

## 1 はじめに

テキスト生成は機械翻訳、要約、対話システムなど自然言語処理の様々な応用に利用される重要な要素技術の一つである。近年、大量のコーパスが利用可能となり、自然な表層文を生成する目的にもコーパスが利用されるようになってきた。その典型例の一つが機械翻訳に用いられる言語モデルである。統計的機械翻訳では、原言語で書かれたテキストを $S$、目的言語で書かれたテキストを $T$ として、$S$ が与えられたときに $T$ を生成する確率 $P(T|S)$ が最大になるようなテキスト $T_{best}$、つまり、

$$T_{best} = argmax_T P(T|S) = argmax_T \{P(S|T) \times P(T)\} \quad (1)$$

を最適な翻訳として出力する [1]。式 (1) の $P(S|T)$ は単語や句を原言語から目的言語に置き換えるためのモデルであり、翻訳モデルと呼ばれる。$P(T)$ が言語モデルであり、置き換えた単語や句を目的言語側で尤もらしい順序に並び替えるためのモデルである。言語モデルの入力は、一般に語の集合 (Bag of words) であり、言語モデルに要求されるのは、基本的にそれらの語の並べ換えである。ここには、与えられた語の集合を並べ換えると自然な文を生成できるという仮定がある。つまり、自然な文を生成するための語の集合は翻訳モデルにより過不足なく生成されることが前提となっている。しかし、実際に翻訳モデルでそのような語の集合を得るためには、大量の対訳コーパスが必要となる。もし、目的言語コーパスのみを利用して、欠けている語を補いながら自然な文を生成することができれば、翻訳モデルで必要となる対訳コーパスの量を軽減することができると考えられる。そこで、我々は、翻訳モデルの役割を、自然な文を生成するために必要な語の集合を与えることではなく、発話者が表現したい重要な単語 (以降で、主要語あるいはキーワードと呼ぶ) の集合を与えることであると仮定する。そして、本稿では、主要語の集合が与えられたときに、目的言語コーパスに基づき必要に応じて情報を補って自然な文を生成する枠組み及び手法を提案する。

我々は、目的言語における主要語の集合を $K$ として、式 (1) を

$$P(T|S) = P(K|S) \times P(T|K) \quad (2)$$

と考える。この式の $P(K|S)$ は、原言語で書かれたテキストが与えられたときに、目的言語の主要語の集合を生成するモデルであり、$P(T|K)$ は主要語あるいはキーワードの集合 $K$ が与えられたときに、テキスト $T$ を生成するモデルである。この $P(T|K)$ で表わされるモデルを**テキスト生成モデル**と呼ぶことにする。本稿ではこのテキスト生成モデル及びこのモデルを実装した生成システムについて述べる。我々のシステムは、キーワードの集合 $K$ が与えられたときに、生成される確率が最大となるような $T$、つまり、

$$T_{best} = argmax_T P(T|K) = argmax_T \{P(K|T) \times P(T)\} \quad (3)$$

となるようなテキスト $T_{best}$ を最適なテキストとして出力する。この式で、$P(K|T)$ を**キーワード生成モデル**と呼ぶことにする。この式は、統計的機械翻訳モデルの式 (1) において、原言語のテキスト $S$ をキーワードの集合 $K$ に置き換えたものに等しいので、この式で表わされるモデルを、キーワードをテキストに翻訳するモデル、と捉えることもできる。式 (3) の $P(T)$ は統計的機械翻訳モデルにも用いられてきた言語モデルであり、一般には n-gram モデルが用いられる。

我々はさらに、テキストが与えられると、キーワードを生成するような形態素の列 (品詞としての並び) と語と語の依存関係はほぼ一意に決まると仮定し、形態素の順序付き集合を $M$、$M$ が与えられたときの語と語の依存関係の順序付き集合を $D$ として、$P(K|T)$ を、

$$P(K|T) \approx P(K,M,D|T) = P(K|M,D,T) \times P(D|M,T) \times P(M|T) \quad (4)$$

のように表わす。この式の $P(K|M,D,T)$ をキーワード生成モデルとし、$P(D|M,T)$、$P(M|T)$ としては、解析モデルとしても用いられている係り受けモデル、形態素モデルをそれぞれ用いる。

統計的機械翻訳や実例型機械翻訳方式では対訳データの質と量が、中間言語方式や構文トランスファー方式では解析システムの精度が、翻訳精度にかなり影響する。したがって、これらの方式では、対訳データがあまりない言語との翻訳や、解析システムがないあるいは精度に問題がある言語との翻訳は難しい。原言語テキストが完全なものでなく、誤りのある OCR 認識結果や音声認識結果などの場合にはさらに難しい。このような場合でも、本稿で提案する手法により主要語を翻訳し、残りの部分を目的言語側で肉付けすれば、大まかな内容がつかめる程度の自然な文が生成できると考えている。

キーワードからテキストを生成する手法は、機械翻訳だけでなく、次のような場合にも応用できる。

- **失語症患者の文生成支援:** 失語症患者は日本全国で 30 万人いるという報告があり、患者にある情景画などを見せてそれを書字で説明させると、2、3 の単語の羅列で説明できる患者の割合が全体の 4 割、2、3 文節が 3 割、5、6 文節が 1 割とのことである。残りの 2 割の患者は単語を選ぶこともできないそうであるが、全体の 8 割の患者に対しては、不完全な語句の入力から自然な文の候補を生成して提示することにより、家族や友達とのコミュニケーションを支援できると考えられる。
- **第二言語での作文支援:** 第二言語の初学者は主要な単語は言えるけれどもなかなか文にできないという場合が多いと考えられる。このような場合に主要語の集合から自然な文の候補を生成することにより、外国人とのコミュニケーションや言語の学習を支援できると考えられる。

上記の他にも様々な応用があると考えている。

## 2 テキスト生成システムの概要

本節では、一つあるいは複数のキーワードを入力とし、テキストを生成するシステムの概要について述べる。このシステムは、図 1 に示されるように、生成規則獲得部、テキスト候補生成部、評価部から構成される。キーワードが入力されたときに、テキストが生成される手順は以下の通りである。

---

* Text Generation from Keywords
Kiyotaka Uchimoto†, Satoshi Sekine‡, and Hitoshi Isahara†
†Communications Research Laboratory
‡New York University

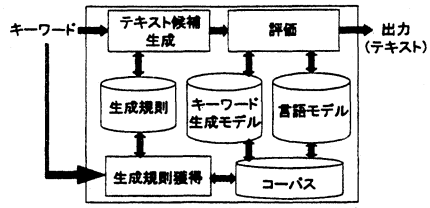

図 1: テキスト生成システムの概要

1. 生成規則獲得部でキーワードごとに規則を自動獲得する。
2. 手順 1 で自動獲得した規則を用いて、テキスト候補生成部でテキスト候補を生成する。テキスト候補はグラフあるいは木で表現する。
3. コーパスから学習した各モデルを用いて、評価部でテキスト候補を順序付けする。
4. 評価部で得れた評価値が最大あるいは閾値を超えるテキスト候補、あるいは評価値の上位 N 個を表層文に変換して出力する。

本稿では、日本語を対象とし、キーワードは文節の主辞となる語であると定義する。そして、文節の主辞となる語は、文末に一番近い内容語であるとする。ここで、内容語は、その語の品詞が、動詞、形容詞、名詞、指示詞、副詞、接続詞、連体詞、感動詞、未定義語である形態素の見出し語であるとし、それ以外の形態素の見出し語を機能語とする。ただし、サ変動詞、動詞「なる」、形式名詞「の」については、文節内で他に内容語がない場合を除いて機能語として扱う。品詞の体系は京大コーパス (Version3.0) [2] のものに従った。

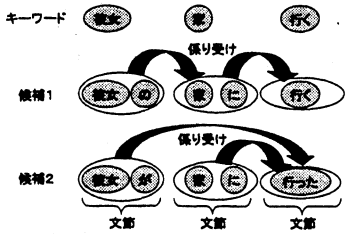

図 2: キーワードからのテキスト生成の例

例えば、キーワードとして、図 2 に示されるようなキーワード「彼女」、「家」、「行く」が与えられたとき、それぞれを含む文をコーパスから検索し、形態素解析、構文解析 (係り受け解析) をする。そして、そこからキーワードを含む文節を抽出して、キーワードから文節を生成する規則「彼女 → 彼女 の」、「彼女 → 彼女 が」、「家 → 家 に」、「行く → 行く」、「行く → 行った」などを獲得し、各キーワードに対し文節の候補を生成する。次に、これらの文節の間に係り受けの関係を仮定して、図の候補 1、候補 2 のような依存構造木の形でテキスト候補を生成する。最後に、テキスト生成モデル (キーワード生成モデルと言語モデル) で評価して候補を順序付けし、評価値が最大のものあるいは閾値を超えるものを表層文に変換して出力する。

本稿では、テキスト生成モデルのうち、式 (4) で表わされるキーワード生成モデル $P(K|T)$ のみに着目し、テキスト候補は依存構造木で表現するものとする。

## 3 テキスト候補の自動生成

### 3.1 生成規則の自動獲得

キーワードの集合を $V$ とし、キーワード $k(\in V)$ から文節を生成する規則の集合を $R_k$ とするとき、規則 $r_k(\in R_k)$ は次の形式で表現されるものと定義する。

$$k \quad \rightarrow \quad h_k m^* \qquad (5)$$

ここで、$h_k$ はキーワードを含む主辞形態素、$m^*$ は同じ文節内で $h_k$ に連続する任意個の形態素とする。ここで、主辞形態素とは、主辞となる単語を含む形態素であるとする。キーワードが与えられると、この形式を満たす規則を単言語コーパスから自動獲得する。

### 3.2 依存構造木の生成

キーワード $K = k_1 k_2 \ldots k_n$ が与えられたとき、まず、前節で述べた生成規則を用いて各キーワードを含む文節候補を生成する。次に、各文節の間に依存関係があると仮定して、日本語の係り受けの特徴である次の条件を満たすように依存構造木の候補を生成する。

(i) 係り受けは前方から後方に向いている。(後方修飾)

(ii) 係り受け関係は交差しない。(非交差条件)

(iii) 係り要素は受け要素を一つだけ持つ。

例えば、キーワードが 3 個の場合、キーワードを含む文節候補がそれぞれ $b_1$、$b_2$、$b_3$、であったとすると、順序を固定した場合には、$(b_1\ (b_2\ b_3))$、$((b_1\ b_2)\ b_3)$ の 2 通り、固定しない場合には 16 通りの候補ができる。

## 4 テキスト生成モデル

本節では、式 (4) の各確率分布で表わされる確率モデル、つまり、キーワード生成モデル $P(K|M,D,T)$ と形態素モデル $P(M|T)$、係り受けモデル $P(D|M,T)$ について述べる。このモデルの目的は、キーワードの系列から生成される語の集合のうち、自然な文を構成できる形態素の集合と係り受けの集合を持つものを選択することである。我々はこれらのモデルを最大エントロピー (ME) モデル [3, 4, 5] として実装した。

### 4.1 キーワード生成モデル

次の五種類の情報を素性として用いたモデルを考える。以下で、キーワードの集合 $V$ は、ある回数以上コーパスに出現した主辞単語の集合とし、文節は形式 (5) で表現されるものと仮定する。また、各キーワードは独立であり、与えれたテキストが単語列 $w_1 \ldots w_m$ からなるとき、キーワード $k_i$ は単語 $w_j (1 \leq j \leq m)$ に対応していると仮定する。

1. 前方の二単語を考慮 (trigram)

   $k_i$ は前方の二単語 $w_{j-1}w_{j-2}$ のみに依存すると仮定する。

   $$P(K|M,D,T) \quad = \quad \prod_{i=1}^{n} P(k_i|w_{j-1}, w_{j-2}) \qquad (6)$$

2. 後方の二単語を考慮 (後方 trigram)

   $k_i$ は後方の二単語 $w_{j+1}w_{j+2}$ のみに依存すると仮定する。

   $$P(K|M,D,T) \quad = \quad \prod_{i=1}^{n} P(k_i|w_{j+1}, w_{j+2}) \qquad (7)$$

3. 係り文節を考慮 (係り文節)

   $k_i$ を含む文節に係る文節がある場合、$k_i$ はそのうち最も文末側の文節の末尾から二単語 $w_l w_{l-1}$ のみに依存すると仮定する (図 3 参照)。

   $$P(K|M,D,T) \quad = \quad \prod_{i=1}^{n} P(k_i|w_l, w_{l-1}) \qquad (8)$$

4. 受け文節を考慮 (受け文節)

   $k_i$ を含む文節を受ける文節がある場合、$k_i$ はその文節内の主辞単語から二単語 $w_s w_{s+1}$ のみに依存すると仮定する (図 3 参照)。

   $$P(K|M,D,T) \quad = \quad \prod_{i=1}^{n} P(k_i|w_s, w_{s+1}) \qquad (9)$$

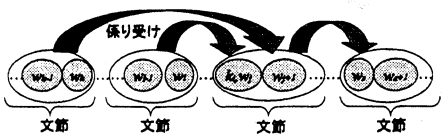

図 3: キーワードと単語の関係

5. 係り文節を最大二文節考慮 (係り二文節)

$k_i$ を含む文節に係る文節がある場合、$k_i$ は、そのうち最も文末側の文節の末尾から二単語 $w_l w_{l-1}$ と、最も文頭側の文節の末尾から二単語 $w_h w_{h-1}$ のみに依存すると仮定する (図 3 参照)。

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_l, w_{l-1}, w_h, w_{h-1}) \quad (10)$$

### 4.2 形態素モデル

テキスト $T$ が与えられたとき、順序付き形態素集合 $M$ が得られる確率は、各形態素 $m_i (1 \leq i \leq n)$ が独立であると仮定し、

$$P(M|T) = \prod_{i=1}^{n} P(m_i|T) \quad (11)$$

と表す。ここで、$m_i$ は形態素に付与するべき文法的属性を表わす。

### 4.3 係り受けモデル

テキスト $T$ と順序付き形態素集合 $M$ がが与えられたとき、各文節に対する係り受けの順序付き集合 $D$ が得られる確率は、各々の係り受け $d_1, \ldots, d_n$ が独立であると仮定し、次のように表わす。

$$P(D|M,T) = \prod_{i=1}^{n} P(d_i|M,T) \quad (12)$$

## 5 実験と考察

システムを評価するために、三つのキーワードからなるシステムの入力を 30 組作成した。これを表 1 の左欄にあげる。これらのキーワードは、京大コーパス (Version3.0)1 月 1 日分の記事に 10 回以上現われた主辞単語の集合 (約 300 個) から記事を見ずに選択した†。

モデルの評価は、出力されたテキストのうち、次の二つの基準で主観的に正しいと評価されたものの割合で行なった。

- **基準 1:** 1 位の候補が意味的、文法的に適切であればシステムの出力が正しいと判断する。
- **基準 2:** 上位 10 位に意味的、文法的に適切な候補があればシステムの出力が正しいと判断する。

キーワードつまり主辞単語の集合 $V$ としては、京大コーパス 1 月 1 日分の記事に主辞単語として現れ、かつ、1 月 1 日から 1 月 16 日分に 5 回以上現れたものを用いた。$V$ に含まれない主辞単語は未知語とし、その品詞 (大分類) を $V$ に追加した。キーワード生成モデルは京大コーパス 1 月 1 日分の記事 1,129 文を用いて作成した。形態素モデルと係り受けモデルとしては、それぞれ文献 [6] と文献 [7] のものを用いた。各モデルの学習には、京大コーパスの 1 月 1 日と 1 月 3 日から 9 日までの 8 日分 (8,835 文) を用いた。生成規則は 1 月 1 日と 1 月 3 日から 16 日までの 15 日分 (18,435 文) から獲得した‡。

† 京大コーパスの 1995 年 1 月 17 日の記事中に現れた主辞単語 (主辞形態素の見出し語) が 1 月 1 日から 1 月 16 日にも現れた割合は、83.33%(3,973/4,768) であったが、二つの主辞単語が同じ係り受けの関係として現れた割合は、21.82%(2,295/10,517) と少なかった。

‡ 生成規則により生成されたテキスト候補の数は、二つのキーワードが入力の場合、キーワード一組あたり平均 868.8 個 (26,064/30)、三つのキーワードが入力の場合、キーワード一組あたり平均 41,413.5 個 (1,242,404/30) であった。

表 1: システムの入力と出力の例

| 入力 (キーワード) | システムの出力例 |
|---|---|
| ロシア 連覇 達成 | (ロシアで (連覇を 達成した)) |
| 世界 記録 変わる | ((世界の 記録に) 変わる) |
| 課題 実現 挑む | ((課題の 実現に) 挑む) |
| チーム ともに 勝つ | (チームで (ともに 勝った)) |
| この 計画 決める | |
| 戦後 生活 支える | ((戦後の 生活を) 支える) |
| 大統領 会見 陳述 | ((大統領の 会見の) 陳述) |
| 容疑者 中国 逮捕 | (容疑者は (中国で 逮捕された)) |
| 日本 五輪 開催 | |
| 国 政策 発足 | ((国の 政策が) 発足する) |
| 課題 実現 向かう | ((課題の 実現に) 向かう) |
| 社会 働く 女性 | ((社会に 働いている) 女性) |
| 首相 政権 奪回 | (首相が (政権を 奪回する)) |
| 提案 見直し 作る | ((提案の 見直しを) 作る) |
| 政権 発足 決まる | (政権が (発足に 決まった)) |
| 香港 クルー 多い | (香港は (クルーも 多いらしい)) |
| 大会 出場 勝つ | |
| 正月 旅行 急増 | ((正月の 旅行も) 急増している) |
| 民営化 反対 人 | ((民営化に 反対している) 人) |
| 昨年 政策 認める | ((昨年の 政策を) 認めるべきだ) |
| 米国 勝つ 明らかだ | ((米国に 勝てば) 明らかになる) |
| 住民 疑い 広がる | ((住民の 疑いが) 広がる) |
| 日本 中国 近い | |
| 外国人 加入 増加 | ((外国人の 加入者が) 増加している) |
| 五輪 選手権 獲得 | ((五輪の 選手権を) 獲得する) |
| 企業 撤廃 進める | (企業が (撤廃を 進める)) |
| 将来 新進党 生まれる | (将来は (新進党が 生まれるだろう)) |
| 昨年 記録 上回る | ((昨年の 記録を) 上回る) |
| 強い チーム 目指す | |
| よい 仕事 ほしい | |

まず、表 1 の三つのキーワードのうち左から二つのキーワード、例えば「連覇 達成」を入力として、各モデルの性能を比較した。表 2 に評価結果をあげる。この表では、4.1 節で述べた五つのキーワード生成モデルをそれぞれ、KM1、KM2、KM3、KM4、KM5、形態素モデルを MM、係り受けモデルを DM と表わし、+ は各モデルの組み合わせを表わしている。MM や DM を用いていないモデルでは、それぞれ、式 (4) の $P(M|T)$ や $P(D|M,T)$ が常に 1 となると仮定している§。

表 2: 主観評価による評価結果

| モデル | 基準 1 | 基準 2 |
|---|---|---|
| KM1 (trigram) | 13/30 | 28/30 |
| KM1 + MM | 21/30 | 28/30 |
| KM1 + DM | 12/30 | 28/30 |
| KM1 + MM + DM | **26/30** | **28/30** |
| KM2 (後方 trigram) | 6/30 | 15/30 |
| KM2 + MM | 8/30 | 20/30 |
| KM2 + DM | 10/30 | 20/30 |
| KM2 + MM + DM | 9/30 | 25/30 |
| KM3 (係り文節) | 13/30 | 29/30 |
| KM3 + MM | 26/30 | 29/30 |
| KM3 + DM | 14/30 | 28/30 |
| KM3 + MM + DM | **27/30** | **29/30** |
| KM4 (受け文節) | 10/30 | 18/30 |
| KM4 + MM | 9/30 | 26/30 |
| KM4 + DM | 9/30 | 22/30 |
| KM4 + MM + DM | 13/30 | 27/30 |
| KM5 (係り二文節) | 12/30 | 26/30 |
| KM5 + MM | 17/30 | 28/30 |
| KM5 + DM | 12/30 | 27/30 |
| KM5 + MM + DM | **26/30** | **28/30** |

表 2 から分かるように、KM1 や KM3、KM5 のモデルに MM と DM を組み合わせた場合が最も良い結果となった。MM や DM を用いた場合、用いなかった場合と比べて基準 1 による評価結果が飛躍的に良くなっているが、その理由は、名詞と格の結び付きより、動詞と格の結び付きの方が強く、後者に着目して学習している

§ KM1 と KM2、KM3 と KM4 を組み合わせ、前方と後方の情報を考慮したモデルも作成し実験したが、最高でも基準 1 で 16/30、基準 2 で 24/30 と良くなかった。

KM1 や KM3、KM5 のモデルが潜在的に自然な文となる候補を上位に順序付けていたからである可能性が高いと考えている。

次に、最も良かった KM3+MM+DM と最も多くの情報を考慮している KM5+MM+DM のモデルを用いて表 1 の三つのキーワードの組を入力とし、出力を評価した。キーワードの順序は並べ替えないものとした。結果はともに、基準 1 で 19/30、基準 2 で 24/30 であった。表 1 の右欄にシステムの出力例をあげる。例えば、「(将来は (新進党が 生まれるだろう))」のように「だろう」を補って「将来」と呼応するような表現を生成したり、「((外国人の 加入者が) 増加している)」のように助詞やモダリティだけでなく、「者」などの接尾語も補った自然な表現が生成できている。適切な出力が得られなかったものについては、その原因は、候補の不足にではなく、順序付けが適切にできなかったことにある。順序付けをより適切にできるようにするためには、単語をクラスタリングした結果や類義語辞書の情報を素性として用いるなどして、モデルの汎化能力を高める必要があると考えている。

## 6 関連研究

これまで多くの統計的手法に基づく生成手法が提案されてきた。この節では、我々の手法とこれまでに提案された手法との違いについて述べる。

日本語では主格や所有格を表現するのに助詞を用いることが多いが、英語にはそれらに対応する単語がない。また逆に、英語では単語に冠詞を伴うことが多いが、日本語には冠詞に対応する単語がない。Knight や Langkilde らはこのような言語依存知識の欠落 (Knowledge Gap) が原因で欠けている情報を補う方法を提案している [8, 9, 10] が、彼らのシステムの入力は意味表現であり、我々のものとは異なる。また、彼らは意味表現から人間が作成した規則を用いてテキスト候補を生成しているが、我々はコーパスから自動的に獲得した規則を用いる。したがって、彼らの手法を我々のシステムでテキスト候補を生成するために用いるのは難しい。ただし、彼らが適切な表層文を選択する際に用いている n-gram モデルは、依存構造木から表層文を生成する際に適用できると考えている。また、Langkilde が提案しているテキスト候補の表現方法 [11] も適用できると考えている。

Bangalour と Rambow はテキスト候補を、我々と同様に構文木として生成している [12] 。彼らは、候補の木を導出する際に単語間の依存関係を考慮しているが、その依存関係は入力として与えられると仮定している。我々は、入力に依存関係が与えられるとは仮定せず、依存関係も推定する。

Ratnaparkhi は、意味属性の列からテキストを生成するためのモデルを提案している [13] 。ここで提案されているモデルでは入力が意味属性であり、我々のものとは異なる。我々のモデルは彼のモデルで意味属性をキーワードに置き換えたものに近いといえるが、彼のモデルでは、パラメータを推定するために、単語の一部が意味属性で置き換えられた文からなるコーパスが必要となる。我々の場合、単言語のコーパスと形態素・構文解析システムがあればよい。

Humphreys らは、言語モデルに構文解析用の言語モデルを用いて順序付けする方法を提案している [14] 。一方、我々は、構文解析用のモデル以外に形態素解析用に作成したモデルを用いている。また、彼らはテキスト候補を、別途生成用に作成した文法を用いて生成しており、我々のものとは異なる。

Berger は、情報検索を、入力となる質問文をそれと尤も関連する文書に翻訳する機械翻訳のタスクと捉え、そのためのモデルを提案している [15] 。このモデルの考え方は我々のものと類似しているが目的が異なる。彼らの目的は、既存の文書を検索することであり、我々の目的は、新たにテキストを産出することである。

## 7 まとめ

本稿では、一つあるいは複数のキーワードを入力とし、それらをもとにテキストを生成する枠組み及び手法を提案した。与えられた各キーワードを、句や文を構成する中心的な要素あるいは主要語と仮定し、自然な句や文を生成するために欠けている情報をコーパスから補い、依存構造木の形で出力する。このようにテキストを依存構造木で生成できれば、語順モデル [16] を用いて自然な語順で表層文を生成することもできる。

以下は今後の課題である。

- **生成規則の拡張 (テキスト候補の充実):** 文節をまたぐような規則を自動獲得して利用できるように拡張し、機能語だけでなく、欠ている内容語も補いながらテキストを生成できるようにする。コーパスから適した規則が獲得できなかった場合、デフォルトの規則で補えるようにする、あるいは、単語をクラスタリングして他の単語について獲得された規則も利用できるようにし、より柔軟な生成を実現したい。
- **生成モデルの高度化:** 非文、不自然な文を生成してしまうものをできるだけ排除できるようにする。そのために、単語をクラスタリングした結果や類義語辞書の情報を素性として用いるなどしてモデルの汎化能力を高めたい。
- **キーワードの定義の拡張:** キーワードを文節の主辞単語と定義したが、それ以外の内容語や類義語がキーワードとなっても生成できるように拡張したい。

**謝辞:** データを使用させて下さった毎日新聞社、本研究に対し、貴重なご意見を下さった、通信総研の太田公子女史、乾裕子女史、豊橋技科大の宇津呂武仁先生、東工大の奥村学先生、富士通研の潮田明氏、東大の辻井潤一先生、貴重な情報を提供して下さった、千葉労災病院の安田清先生、甲南大の太田雅久先生に心から感謝の意を表する。

## 参考文献

[1] Peter F. Brown, et. al. A Statistical Approach to Machine Translation. *Computational Linguistics*, Vol. 16, No. 2, pp. 79–85, 1990.
[2] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 言語処理学会 第 3 回 年次大会 発表論文集, pp. 115–118, 1997.
[3] Adam L. Berger, et. al. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, Vol. 22, No. 1, pp. 39–71, 1996.
[4] Eric Sven Ristad. Maximum Entropy Modeling for Natural Language. ACL/EACL Tutorial Program, Madrid, 1997.
[5] Eric Sven Ristad. Maximum Entropy Modeling Toolkit, Release 1.6 beta. http://www.mnemonic.com/software/memt, 1998.
[6] 内元清貴, 関根聡, 井佐原均. 最大エントロピーモデルに基づく形態素解析 —未知語の問題の解決策—. 自然言語処理, Vol. 8, No. 1, pp. 127–141, 2001.
[7] 内元清貴, 村田真樹, 関根聡, 井佐原均. 後方文脈を考慮した係り受けモデル. 自然言語処理, Vol. 7, No. 5, pp. 3–17, 2000.
[8] Kevin Knight and Vasileios Hatzivassiloglou. Two-Level, Many-Paths Generation. In *Proceedings of the ACL*, pp. 252–260, 1995.
[9] Irene Langkilde and Kevin Knight. The Practical Value of N-grams in Generation. In *Proceedings of the INLG*, 1998.
[10] Irene Langkilde and Kevin Knight. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the COLING-ACL*, pp. 704–710, 1998.
[11] Irene Langkilde. Forest-Based Statistical Sentence Generation. In *Proceedings of the NAACL*, pp. 170–177, 2000.
[12] Srinivas Bangalore and Owen Rambow. Exploiting a Probabilistic Hierarchical Model for Generation. In *Proceedings of the COLING*, pp. 42–48, 2000.
[13] Adwait Ratnaparkhi. Trainable methods for Surface Natural Language Generation. In *Proceedings of the NAACL*, pp. 194–201, 2000.
[14] Kevin Humphreys, et. al. Reusing a Statistical Language Model for Generation. In *Proceedings of the EWNLG*, 2001.
[15] Adam Berger and John Lafferty. Information Retrieval as Statistical Translation. In *Proceedings of the ACM SIGIR*, pp. 222–229, 1999.
[16] 内元清貴, 村田真樹, 馬青, 関根聡, 井佐原均. コーパスからの語順の学習. 自然言語処理, Vol. 7, No. 4, pp. 163–180, 2000.