係り受け共起確率を利用した文生成

藤尾 正和　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{masaka-h,matsu}@is.aist-nara.ac.jp

Abstract

コーパスからの統計情報を利用した係り受け解析の研究が盛んに行われている。本稿では、係り受け解析モデルをそのまま利用し、単語による依存構造木から表層文を生成するモデルを提案する。提案する統計モデルおよび本モデルの利用目的について述べ、文生成実験の結果についても報告する。

1 はじめに

文の統語解析に比較して、文生成はその利用法や前提によって様々な形態があり、入力として何を仮定するかによってその方式も異なる。従来の規則主導の文生成では、与えられた入力に対して生成可能なすべての文を出力する能力をもつことが多く、したがって、適格な文候補には曖昧性があるうる。例えば、意味主辞駆動文生成 [1] では、意味主辞からの生成により、文法規則の適用順序にある程度の制約はかかるものの、直接依存関係のない構成素間の生成順序には自由度がある。もちろん、生成される文はどれも文法的には適格な文であるが、文の自然さという意味では許容度に違いがありうる。

本稿では、我々がこれまで提案してきた共起確率に基づく統計的係り受け解析システム [2] の処理過程を逆に用いて、文生成システムとして利用する方法について検討する。統計的な文生成 (特に語順の決定) については、Shaw ら [3] の名詞に係る形容詞や名詞などの語順の決定に関する統計的な研究や、内元ら [4] の最大エントロピー法による語順決定規則の学習の研究がある。これらはいずれも語順に特化してコーパスから規則を学習しようという研究である。これに対して、本生成システムは、我々の研究室で行っている依存構造間の対応に基づいた日英両方向の機械翻訳システム [5] の生成部として用いることを前提として考えられた。その際に、文の係り受け解析に用いられる統計パラメータを生成部でもそのまま流用することを前提とした。次節で、本システムの利用の前提について述べ、その後、文生成モデルおよび評価実験について述べる。

2 統計的文生成の利用

本稿で述べる統計的文生成は、現在構築中の統計的な依存構造解析および対訳コーパスから抽出された翻訳パターンに基づく機械翻訳システムの文生成部として位置付けられる。本節ではこのための機能として考えている要件について概観する。

図1に、機械翻訳システムの概念図を示す。図では英語文から日本語文の翻訳過程が示されているが、言語対に関しては双方向に利用可能である。統計的な係り受け解析システム [2] は、現在、日英両言語で動いている。源言語の解析結果に対して、依存構造に基づいた翻訳パターンを適用し、その結果、目標言語における依存構造木が生成される。本稿で提案するシステムは、この依存構造木を入力として、目標言語の表層文を生成するために用いることを想定している。生成システムに与えられる目標言語の依存構造木において考えられる特徴について整理しておく。

- 依存構造木の木構造は曖昧性を含まない
- 各依存関係は、一意に決められているとは限らないが、依存関係の候補がある程度限定されている。

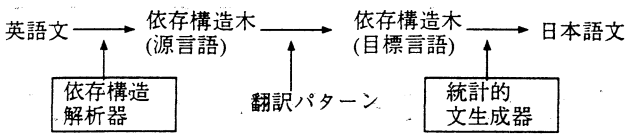

図 1: 依存構造パターン翻訳の流れ

- 各依存関係には、源言語中の対応する部分に関する係り受け距離の情報を含めることができる

これらの前提を考慮して、統計的文生成モデルを設定する。ただし、後述する実験では、予備実験として特に翻訳環境は想定せず、日本語のみを対象として、より簡単な前提で実験を行った。

3 統計モデル

本節では、まず基礎となる係り受け解析モデルについて述べ、次に、同じ統計パラメータを利用した文生成モデルについて説明する。

3.1 統計的係り受け解析モデル

入力文字列をS、分かち書きされてタグ付けされた単語列$< w_1, t_1 >, \ldots, < w_n, t_n >$を$T$、文節にまとめられ属性付けされた文節列$< b_1, \mathbf{f_1} >, \ldots, < b_m, \mathbf{f_m} >$を$F$、文節区切りに対する係り受けパターンの組$\{Dep(1), Dep(2) \ldots Dep(m-1)\}$を$L$とする。但し$Dep(i)$は文節$b_i$の係り先の文番号を表す。$w_i, t_i, b_i, m$はそれぞれ単語、タグ、文節、文節数を表す。また$\mathbf{f_i}$は文節$b_i$の持つ属性の集合を表すものとする。係り受けの構造には、次の2つの制約を仮定する。

1. 文末を除き各文節は文末側に必ず一つの係り先を持つ
2. 係り受けは非交差

係り受け解析は、条件付き確率$P(L.F.T|S)$が最大になる$L.F.T$を求めることであり、次のように定式化する。

$$\begin{aligned} P(L.F.T|S) &= P(L|F.T.S)\ P(F|T.S)\ P(T|S) \\ &= P(L|F)\ P(F|T)\ P(T|S) \end{aligned}$$

実際には、文節区切りは分かち書きと品詞タグから決定でき、係り受けは文節の属性のみで決定できると考え、次の式の最大化のみを対象とする。

$$L_{best} = \arg\max_{L} P(L|F)$$

さらに、各係り受けは独立であると仮定すると、次のように展開できる。

$$P(L|F) = \prod_{i=1}^{m} P(Dep(i){=}j \mid \mathbf{f_1} \ldots \mathbf{f_m})$$

$P(Dep(i){=}j \mid \mathbf{f_1} \ldots \mathbf{f_m})$は、文節区切りと属性集合$\mathbf{f_1} \ldots \mathbf{f_m}$が与えられた時に、文節対$\mathbf{f_i}, \mathbf{f_j}$が、係り受け関係にある確率を表わしている。

確率$P(i \stackrel{\mathrm{rel}}{\rightarrow} j \mid \mathbf{f_1} \ldots \mathbf{f_m})$は、語彙共起確率と文節間確率の積で定義するが、これらの定義のために以下の属性を使用する。

文節属性 (係り側文節、受け側文節)

- 文節の主辞 h_i, h_j
- 係り関係名 r_i, r_j
- 句読点の有無 p_i, p_j

文節間属性

- 文節間文節情報 d_{ij}

文節間属性とは、係り受け関係にある2つの文節の間に存在する文節に関する情報で、おもに文節間の距離等を係り受け解析の優先度に反映させるために導入する。なお、d_{ij}が具体的にどのような属性をとるかは、次節で説明する。

以上の定義により、語彙共起確率(P_h^{ij})と文節間確率(P_d^{ij})は、それぞれ次のように定義できる。なお、d_{ij}が具体的にどのような属性をとるかは、次節で説明する。また、s_iは、翻訳文生成として用いる際に与えられる源言語文での係り

関係の距離に関する情報であるが、今回の実験のモデルではこのパラメータは使用しなかったので、以下では省略する。

$$P_h^{ij} = P(i \overset{\text{rel}}{\rightarrow} j \mid h_i, r_i, p_i, \quad h_j, r_j, p_j)$$
$$P_d^{ij} = P(i \overset{\text{rel}}{\rightarrow} j \mid r_i, p_i, d_{ij}, s_i)$$

それぞれの確率は、最尤推定により、解析済みコーパスから求めることができる。

3.2 統計的文生成モデル

文生成モデルにおいても、解析モデルと同じものを用いる。ただし、係り受けの依存構造木は入力として得られるので、文節間の係り受け関係の存在は決定している。そのため、生成に本質的に影響を与えるのは、文節間確率 (P_d^{ij}) だけである。

$$P_d^{ij} = P(i \overset{\text{rel}}{\rightarrow} j \mid r'_i, p'_i, d_{ij})$$

ここで、r'_i, p'_i と書いたのは、生成システムへの入力として、文節がもつ係り関係や読点情報が一意に決まっているとは限らないからであり、それぞれ係り関係と読点情報の部分的な情報である。

文節間属性 d_{ij} については、いくつかのバリエーションが考えられる。ここでは、d_{ij} が、文節対 b_i, b_j の間に存在し、文節 b_j に直接係る文節 (b_i と兄弟の文節) の係り関係および読点情報からなるとし、さらに、それらのうちどれだけの文節を考慮するかによってモデルを分類した。繁雑さを避けるために、r'_i, p'_i の組を改めて r'_i と書く事にし、$d_{ij} = r'_h \cdots r'_{j-1}$ と書く。(文節 b_i が文節 b_j に係る時は必ず文節 b_{j-1} も b_j に係ることに注意。ただし、b_i と b_j の間に文節が存在しない場合は、d_{ij} は空列)

$$P_d^{ij} = P(i \overset{\text{rel}}{\rightarrow} j \mid r'_i, r'_h, \cdots, r'_{j-1})$$

なお、データの過疎性の問題を避けるため、ここでは次の 2 つのモデルを採用して、実験を行った。

(1) $P_d^{ij} = P(i \overset{\text{rel}}{\rightarrow} j \mid r'_i, r'_h)$

(2) $P_d^{ij} = \prod_{k \in \{h, \cdots, j-1\}} P(i \overset{\text{rel}}{\rightarrow} j \mid r'_i, r'_k)$

生成モデル	完全一致による正解率
モデル (1)	65.08 %
モデル (2)	65.61 %

表 1: 文生成実験の結果

(1) は、各係り文節が自分の兄弟のうち右側に存在する最初の兄弟の情報のみを文節間属性として利用する。(2) では、各係り文節が自分の右側に存在するすべての兄弟の情報を文節間属性として利用するが、兄弟すべてを同時に見るのではなく、自分と相手一つずつとの統計を利用する。

文生成のアルゴリズムは、一般的にはに、文解析と逆方向に Chart 法を利用することで可能であるが [6][7]、ここでは係り受けが対象であり、二分木に相当する規則しか存在しないので、CKY アルゴリズムを生成用に変更した方法を用いた。

4 文生成実験

4.1 実験の概要と結果

前節までは、我々が想定している一般的な統計的文生成手法について説明した。本節では、一部限定されたモデルを対象に行った生成実験とその結果を示す。

実験データとしては、京大コーパス [8] の係り受け解析済み文を用い、これを 10 個の部分コーパスに分離して、10 回繰り返しの交差検定を行った。実際に用いた文は、係り文節を複数もつ文節の部分であり、受け側の文節数にして 19,664 個、総文数は 19,956 文であった。

学習および生成モデルとして、3.2節で示した 2 つの文節間属性モデルを用いた。生成のための入力情報は、係り受け依存構造木とし、係り関係は元の文の情報をそのまま渡した (したがって、係り関係の曖昧性については今回は考慮していない)。評価は、受け側の文節ごとに行い、その文節に係るすべての文節が元の順番に一致して生成された場合のみを正解とした。それぞれのモデルによる正解率を表 1 に示す。同じ文節に係る兄弟の中で、自分より右側に現れるすべて

の兄弟を対象にするモデル (2) の方が、わずかであるが高い正解率が得られた。

4.2 考察

本節では、関連研究との比較および結果に関する考察を行う。Shaw ら [3] は英語の名詞に前方から係る修飾語 (形容詞、名詞など) を対象に語順の学習を行っている。形容詞や名詞の修飾は、日本語の文節がもつような係り関係を持たないので、利用できる情報は単語そのものあるいは単語のクラスであり、それらに基づくいくつかのモデルで実験を行っている。学習データに単語の組合せそのものが存在し、直接的な証拠がある場合には、専門分野でも Wall Street Journal でも高い精度 (95%以上) を得ているが、そうでない場合は、特に WSJ では、70%程度の精度になっている。彼らの方法とは対象とする現象がかなり異なるので、比較を行う事は難しい。特に我々の方法では、係り関係のみを扱い、文節の内容語の情報は用いていない。

内元ら [4] の研究は、我々と対象が近い。彼らは、語順の学習に特化したモデルを提案し、完全一致率で 75.2% の正解率を報告している。彼らのモデルは、係り側および受け側の主辞の単語、品詞、意味素性、さらに、文脈指示語の存在を考慮に入れていることなど、我々のモデルに比較してはるかに多くの素性を利用している。また、同じ文節に係る兄弟間のすべての対について語順の優先確率を考慮に入れている。素性の多さやモデルの細かさから見ると、約 10%の差は妥当な数値といえるのではないだろうか。

正解が得られなかったデータについての詳細な考察はまだ行っていないが、主なエラーの原因としては、慣用的な表現など、係り側と受け側の内容語を考慮にいれなければならない場合が多い。また、従属節に関係するエラーが多数みられた。今回のモデルでは、従属節に関して特別な扱いをしていないが、従属節とそれ以外の修飾要素に関しては、例えば、それを考慮した解析モデル [9] と同様に、分離したモデルを考えるべきかも知れない。

5 おわりに

統計的係り受け解析における文節間の共起確率を利用した文生成について述べた。今回は予備的な実験に留まったが、利用対象と考えている翻訳文の生成により忠実な設定での実験、すなわち、係り関係に曖昧性がある場合や、源言語文の語順情報が利用可能な場合を考慮したシステムの作成と実験を行いたい。

参考文献

[1] Shieber, S., et al., "Semantic Head-driven Generation," Computational Linguistics, Vol.16, No.1, pp.30-42, 1990.

[2] 藤尾正和, 松本裕治, "語の共起確率に基づく係り受け解析とその評価," 情報処理学会論文誌, Vol.40, No.12, pp.4201-4212, 1999.

[3] Shaw, J. and Hatzivassiloglou, V., "Ordering Among Premodifiers," 37th ACL, pp.135-143, 1999.

[4] 内元清貴 他, "コーパスからの語順の学習," 情報処理学会 自然言語処理研究会, 2000-NL-135, pp.55-62, Jan. 2000.

[5] 北村美穂子, 松本裕治, "対訳コーパスを利用した翻訳規則の自動獲得," 情報処理学会論文誌, Vol.37. No.6, pp.1030-1040, 1996.

[6] Kay, M., "Chart Generation," 34th ACL, pp.200-204, 1996.

[7] Haruno, M., et al., "Bidirectional Chart Generation of Natural Language Texts," AAAI-93, pp.350-356, 1993.

[8] 黒橋禎夫, 長尾真, "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, pp.115-118, 1997.

[9] 宇津呂 他, "コーパスからの日本語従属節係り受け選好情報の抽出およびその評価," 自然言語処理, Vol.6, No.7, pp.29-60, 1999.