5S−5

# 読みやすさを考慮した文生成*

柴田 昇吾，藤田 稔，柵木 孝一†
キヤノン(株) 情報システム研究所‡

## 1 はじめに

近年，文生成の重要さが認識されるようになり，単に文法的に正しい文でなく，人間にとって読みやすい文の生成を目指すようになってきている．読みやすい文を生成するためには，文生成の過程で何らかの工夫をしなければならない．文生成の過程は，伝達内容や論理組み立てといった what-to-say を決める過程と文の構成や表層語や語順といった how-to-say を決める過程とからなっている [1]．今回，what-to-say は生成システムに与えられるものとして，how-to-say を決める過程で読みやすさを考慮することにした．

本稿では，文の読みやすさと評価方法について述べ，文の読みやすさを評価して自動的に文を改良し，読みやすい文を生成するシステムを述べる．

## 2 読みやすさ

計算機が文の読みやすさを判断するためには，定量的な評価ができなければならない．このためには，文に対する評価項目と評価基準を設ける必要がある．

読みやすさに関しては，校正支援の研究を通して，評価項目が明らかになってきている．また，小学・中学・高校の教科書で用いられている文の定量的な分析が行なわれており，教科書文の長さや漢字比率などの平均値が得られている [2]．教科書文は，読みやすさという点で吟味されていると考え，この教科書の分析データを参考にして，表 1 に示す評価基準値を設けた．評価基準値は，教科書文の平均値から許容範囲を考慮して，上限と下限で表される．値は対象とする学年に応じて変化するが，表 1 では中学の教科書文の平均値を参考にした．

表 1: 評価項目と評価基準値

| 評価項目 / 評価基準値 | 文字数 | 漢字比率 | 句の数 | 用言の数 | 関係節深さ（入れ子数） |
|---|---|---|---|---|---|
| 上限 | 40 | 40% | 10 | 5 | 3 |
| 下限 | 5 | 0% | 1 | 1 | 1 |

## 3 システム構成

本生成システムの構成を図 1に示す．本システムは，生成過程を分担する生成モジュール群と，生成モジュールからの出力を評価する評価モジュール，そして，文改良のためにフィードバックを行なう制御モジュールからなる．

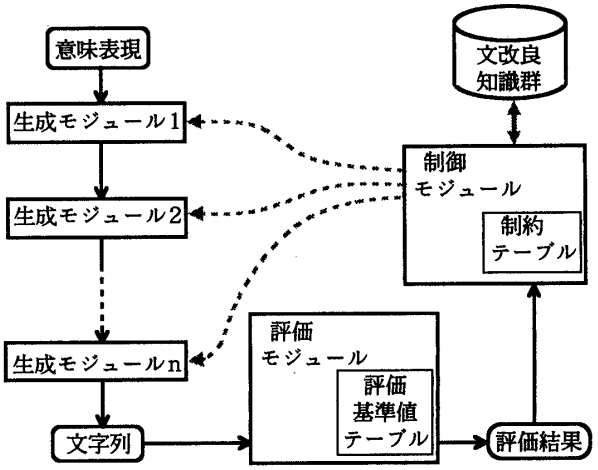


図 1: システム構成図

本生成システムでは，生成過程を 11 に分け，それぞれの過程に 表 2に示すような生成モジュールを割り当てた．本システムの入力は意味表現で，出力は文字列である．入力は，表 2 中の生成モジュール番号順に，構文木，単語リストという中間結果を経て文字列まで処理される．生成モジュールは，制約にそって生成処理を進める．制約は，制御モジュールの制約テーブルに保持される．

表 2: 生成モジュール一覧

| 入力データの構造 | No | 生成モジュール名前 | 機能概要 |
|---|---|---|---|
| 意味表現 | 1 | 単文／複合文 | 一文の範囲を決める |
| | 2 | 関係節 | 関係節を使うか否かを決める |
| | 3 | 述部決定 | 用言の表層語を決定する |
| | 4 | 連用修飾句 | 名詞句や副詞句を決定する |
| | 5 | 格順決定 | 格順を決定する |
| 構文木 | 6 | 構文決定 | 意味表現から構文木に変換する |
| | 7 | 代名詞／省略 | 代名詞や省略を使うかを決める |
| | 8 | 格順変位 | 格順の調整を行なう |
| | 9 | 線状化 | 構文木を線状化する |
| 単語リスト | 10 | 活用形決定 | 活用語の活用形を決定する |
| 文字列化 | 11 | 文字列化 | 文字列化する |

* Text generation algorithm for well-formed sentences.
† Shogo SHIBATA, Minoru FUJITA, Koichi MASEGI
‡ Canon Inc. Information Systems Research Center

さて，生成モジュール「文字列化」の出力は，評価モジュールで評価基準値と較べることによって評価される．評価結果は，制御モジュールへ送られる．

評価結果が良い場合には，生成が終了する．一方，評価結果が悪い場合には，まず，文改良知識群から評価結果に応じた文改良知識を取り出す．文改良知識には，例えば，『文が長すぎる場合には，どの生成モジュールへフィードバックをかけ，どのように制約を変更すべきか』，という文改良ルールが複数個書かれている．文改良知識の例を表 3に示す．そして，制御モジュールは，優先順に文改良ルールを適用し，指定された生成モジュールにフィードバックをかける．フィードバックによって呼び出された生成モジュールは，保存されていた中間結果と変更された制約とを用いて，再度，文生成を行なう．

生成された文の評価結果が悪い場合には，評価結果が良くなるまで文の改良を繰り返す．そして，すべての文改良ルールが尽きた場合には，how-to-say では生成できないことになる．

## 4 実行例

図 2の意味表現を入力した時の生成過程を図 3に示す．図 2の意味表現を生成していくと，図 3最上部の文字列で構成される文が生成される．これを評価モジュールで評価すると，文が長すぎることがわかる．

制御モジュールは，この評価結果から表 3に示す文改良知識を取り出し，優先順にルールを適用していく．最初に適用できるのは，『パラフレーズを用いる』ルールである．この改良によって「朝遅くまで寝る」が「寝坊する」になり[1]，文の長さは45文字になる．しかし，評価基準値である40文字以下にはならないので，再び，文が長すぎるという評価結果が出る．

評価結果から再度，文改良知識を取り出し，『文の連接の数を1減らす』ルールを適用する．この変更によって文が2つに分割され，それぞれの文の長さが21文字，26文字となり，評価基準値が満たされる．

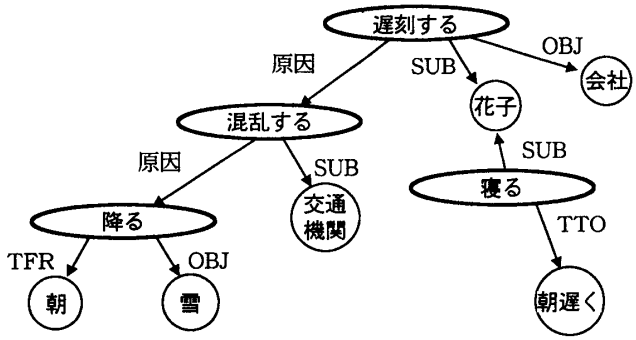


図 2: 例文の意味表現

## 5 おわりに

文生成の結果が読みやすいかを評価し，評価結果に応じてフィードバックをかけて文を改良し，読みやすい文を生成する文生成アルゴリズムを述べた．その際，文の読みやすさの定量的な評価のために，評価項目と評価基準値を示した．

今後の課題としては，次の二点がある．

- 生成モジュールの細分化を進め，文改良知識を充実することによって，よりきめの細かい制御を可能とする．
- 今回の評価は，文単位であったが，文章を単位とした時の評価と文改良方法を検討する．

表 3: 文改良知識の例

| 生成モジュール No | 名前 | 優先順位 | 制約 | 変更内容 |
|---|---|---|---|---|
| 7 | 代名詞／省略 | 1 | 代名詞／省略／何もしない | 代名詞 |
| 4 | 連用修飾句 | 2 | 表層語の文字列が長いものを使うか短いものを使うか | 短い表層語 |
| | | 3 | 表層語の漢字比率が高いものを使うか低いものを使うか | 高い表層語 |
| 3 | 述部決定 | 4 | 表層語の文字列が長いものを使うか短いものを使うか | 短い表層語 |
| | | 5 | 表層語の漢字比率が高いものを使うか低いものを使うか | 高い表層語 |
| | | 6 | 和語動詞／サ変動詞 | サ変動詞 |
| | | 7 | パラフレーズを用いるか | 用いる |
| 2 | 関係節 | 8 | 関係節の入れ子の数 | 1減らす |
| 1 | 単文／複合文 | 9 | 文中の連接の数 | 1減らす |

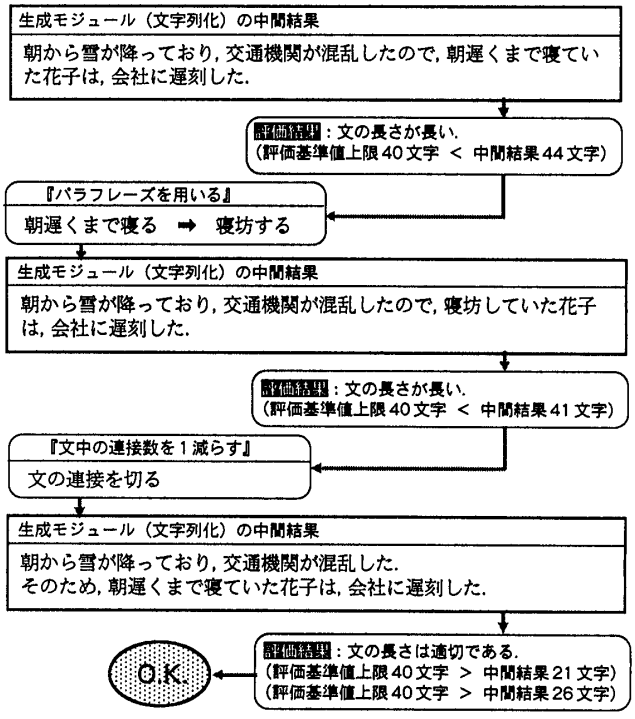


図 3: 実行例

## 参考文献

[1] McKeown, K. :Discourse Strategies for Generating Natural-Language Text, Readings in Natural Language Processing, Morgan Kaufmann Publishers, Inc, 1985.
[2] 石崎他，日本語文の複雑さの定性的・定量的特徴抽出，自然言語処理研究会67-6, 1988.
[3] 森岡，文体と表現，明治書院，1988.

[1]このパラフレーズは，新明解国語辞典の「寝坊する」の語義文から知識を獲得した．