6G−4

ＥＤＲ概念関係表現からの英語生成

小松　英二　　崔　進　　安原　宏
(株)日本電子化辞書研究所

1．はじめに

ＥＤＲでは、辞書の完全性を期するために、ＥＤＲ辞書を用いた数種類の実証評価システムを作成している。以下は実証システムの１つである日英機械翻訳の構想である。本システムは、ＥＤＲが辞書開発のための基礎データとして収集した大規模なコーパス（ＥＤＲコーパス）から実例ベースを作成し、これを日本語解析と英語生成とで共有し、処理を行なうという実例ベース翻訳システムをめざしており、実例ベースの構成やその利用方法などは、日本語解析と基本的に共通にする予定である。翻訳方式としては中間言語方式を採用する。本稿では、本システムの英語生成部について述べる。

従来、英語生成の方法としては、ルールを用いた生成システムが多いことが報告されている（徳永[1],辻井[2]）。また、ルールを用いない生成システムとしては、実例ベースの翻訳システム（佐藤[3]）がある。ルールベースの翻訳システムでは大量のルールを人手で作成する必要があるため、我々は、ＥＤＲコーパスから作成した実例ベースを用いて生成を行なうことにより、ルール作成のボトルネックを解消しようと考えている。本稿では、コーパスから予め取り出した事例を用いて中間言語から係り受け表現を生成する方式について述べる。

2．ＥＤＲ概念関係表現

ＥＤＲでは、中間言語を指す用語として「概念関係表現」という言葉を用いている（EDR[4]）。図１は、説明のために簡略化した概念関係表現の例である。概念関係表現は、「概念」、概念間に成立しえる関係を識別するための「関係子」、及び、概念の範囲を限定するための「属性子」からなる。概念関係表現の中でルートである概念を「エントリーポイント」と呼ぶ。

図２は、図１に対応する係り受け表現である「構文・意味表現」の例である。係り受け関係は、ＥＤＲが文分析の結果として用いている表現であり、単語見出し、品詞、文法情報（EDR[5]）、及び、単語同士の関係を表わす合成関係子からなる。合成関係子としては、S関係子とM関係子があり、S関係子は英語の前置詞や活用語尾等、複数個の文要素が集まって１つの文要素を作る場合の関係子であり、M関係子は名詞と動詞の係り受け関係等、修飾要素と被修飾要素をまとめる関係子である。さらに、合成関係子でむすばれた単語について語順が前の単語にf、語順が後の単語にbをつけておく。

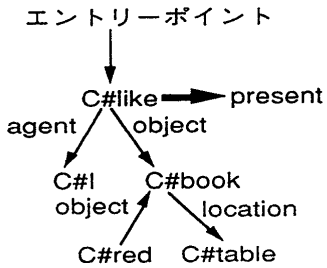
概念と単語の対応は、EDRの英語単語辞書（EDR[5]）に記述されている。

3．実例ベース

図３に英語生成部の構成を示す。ＥＤＲコーパスから抽出した実例は実例ベースに登録しておく。

English generation from EDR concept relation representation
Eiji KOMATSU　Susumu SAI　Hiroshi YASUHARA
Japan Electronic Dictionary Research Inst. Ltd.

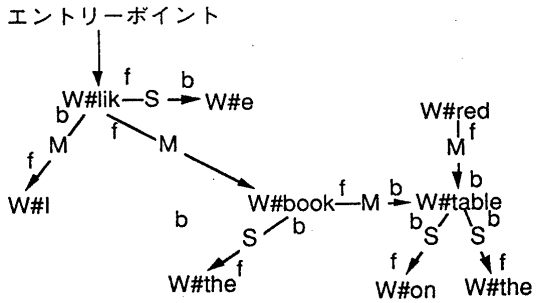
I like the red book on the table. の概念関係表現



C#like、C#book、C#table、C#red：概念
：関係子
：属性子

図１　概念関係表現の例

I like the red book on the table. の構文・意味表現



W#lik、W#e、W#book、W#table、W#red、W#on、W#the
：単語見出し＋品詞・文法情報

図２　構文・意味表現の例

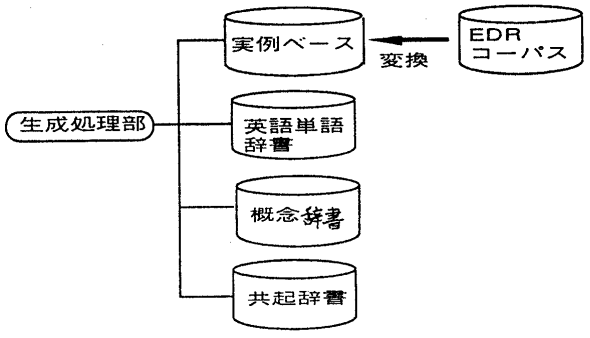


図３　英語生成部の構成

実例は、文の概念関係表現の各概念について、その概念に直接接続する概念までの範囲の概念関係表現を、対応する構文・意味表現とペアにして取り出した概念関係表現と構文・意味表現との変換例である。

図１、２のデータがペアとしてコーパスにあったとすると、各概念を中心として、計４つの実例を取り出すことができる。図４はC#bookを中心として取り出した実例である。実例は、中心にある概念の単語を決定するために用いられ、まわりの概念は制約条件として利用される。

４．生成アルゴリズム

図１のような概念関係表現を入力とし、次の(1)から(4)のような処理を行ない、図２のような構文・意味表現を出力することを目標とする。

(1) 訳語の選択
(2) 前置詞の選択
(3) 合成関係子の決定
(4) 語順の決定（修飾関係にある単語の前後関係のみ）

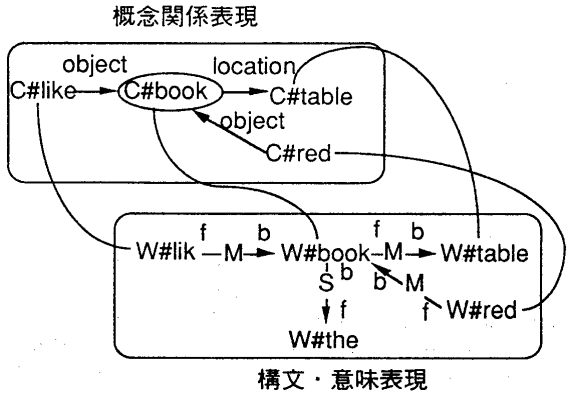
以下に生成アルゴリズムを示す。

[処理１]：単語検索
英語単語辞書を用いて、各概念に対応する単語を取り出す。
[処理２]：初期設定
エントリーポイントをカレントの概念とする。
[処理３]：実例ベース検索
カレントの概念に接続するすべての概念までの範囲の概念関係表現と類似した実例（概念関係表現の類似度は５節参照）を実例ベースから検索し、類似度が大きい順に実例を取り出す。
[処理４]：生成
取り出した実例と類似した単語（単語の類似度は５節参照）が選べるかどうかチェックする。
実例と対応するすべての概念について類似した単語が選べれば、カレントの概念の訳語、前置詞、上位ノードとの合成関係子及び語順を決定する。
類似した単語が選べない概念が１つでもあれば、実例を変えて、処理４を繰り返す。もし、適当な実例がなければ、生成処理は失敗し、終了する。
[処理５]：子ノードの再帰的処理
カレントの概念の子ノードをカレントの概念として、処理３から再帰的に処理する。上記の処理はすべての子ノードについて行なう。子ノードがなければ、再帰的な呼び出し元に戻る。呼び出し元がなければ処理６へいく。
[処理６]：出力
単語の選択結果をまとめて、構文・意味構造を出力する。

５．類似度関数

生成アルゴリズムにおいては、次の３種類の類似度関数が必要である。各類似度関数は、[0,1]の値を返す。この値を類似度と呼び、類似度が正のとき、２つの要素は類似しているといい、類似度が０のとき、２つの要素は類似していないと言うことにする。以下に類似度関数の概要を述べる。

(1) 概念の類似度関数
概念体系（EDR[4]）を用いて、概念同士の類似度を計算する。



：生成処理を行なう概念（他の概念は、制約として用いる）

図４　実例（図１と図２のペアから構成）

(2) 概念関係表現の類似度関数
概念の類似度及び関係子のパタンに従、概念関係表現同士の類似度を計算する。
(3) 単語の類似度関数
単語見出し、品詞、文法情報を用いて、単語同士の類似度を計算する。

６．おわりに

現在、システムの作成、及び、類似度関数の設計を行なっている。今後、大量のデータにより、本方式の妥当性について検証する予定である。

参考文献

[1]徳永他、1980年代の自然言語生成１－３、人工知能学会誌 Vol.6 No.3-5 1991
[2]辻井　潤一、機械翻訳における文章の生成、人工知能学会誌 Vol.4 No.6、1989
[3]佐藤　理史、実例に基づく翻訳、情報処理、Vol.33 No.6、1992
[4]日本電子化辞書研究所、TR-020 概念辞書（第３版）
[5]日本電子化辞書研究所、Tr-019 英語単語辞書（第２版再改訂）