ＣＩＣＣインドネシア語翻訳システムの構成

2F－6

松田純一，寺田和人，村上孝也
(財)国際情報化協力センター

１．はじめに

社会の国際化に伴って、機械翻訳の必要性が高まってきている。(財)国際情報化協力センター（略称ＣＩＣＣ）では、インドネシア、タイ、中国、マレーシアと協力して、近隣諸国５ヶ国語間の機械翻訳プロジェクトを１９８７年度から６年間の予定で進めている。本報告では、このうち、インドネシア語翻訳の基本構想及びシステム構成について、言語独立性および資源の共有化の観点を中心に述べる。

２．システムの全体構成

インドネシア語（以下、イ語と略す）翻訳システムは、辞書システム、生成システム、解析システム、入出力システム、翻訳支援システムのサブシステムから構成される。これらのシステムは、翻訳対象となる他の言語とはまったく独立である。他の言語の翻訳システムとの間は、言語に独立な中間言語および概念辞書[1]を介して接続する。これにより、イ語以外の言語の知識を持たなくてもシステムの開発を行うことができる。

３．基本語辞書の構成

イ語基本語辞書は、イ語処理に必要な情報を記述し、解析や生成でも参照する。記述内容はイ語だけに依存している。イ語辞書構造を図１に示す。この辞書構造は、日本電子化辞書研究所(ＥＤＲ)の単語辞書[2]と同様の考え方に基づいている。１つの語幹が辞書中の１レコードと対応しており、１レコードには、語幹から派生したイ語の単語が全て記述されている。イ語辞書における語幹とその派生語の記述例を図２に示す。機械翻訳で利用する場合には、各派生語を別々の単語として扱う方が、辞書構造が単純になるため望ましいが、広くイ語の自然言語処理一般への適用性や辞書検索の効率化を考慮して、異なる概念を持つ語も一つのレコードに含めて関連付けることにした。実際に解析や生成で利用する場合には、各派生語を１レコードとした内部構造に変換する。

各単語に関する記述内容は以下の通りである。

(1)品詞

(2)意味情報

a.概念番号：単語が対応する概念番号を記述する。

(3)構文情報

a.品詞詳細情報：品詞ごとに、より詳細な文法情報を記述する。

b.文型情報：述語のとる構文パターンを記述する。

(4)形態素情報

a.活用情報：能動形・受動形の表記に関する情報を記述する。

b.繰返しの可否：名詞の複数形が、語の繰り返しになるかどうかを記述する。

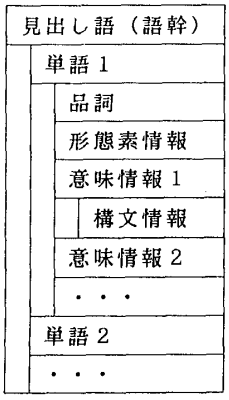
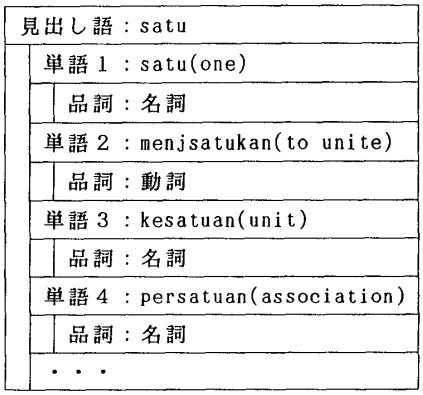


図１　インドネシア語辞書の構造



（括弧内は単語の意味を示す）

図２　インドネシア語辞書の記述例

Structure of CICC's Indonesian Machine Translation System
Jun'ichi MATSUDA,Kazuto TERADA and Takaya MURAKAMI
CENTER OF THE INTERNATIONAL COOPERATION FOR COMPUTERIZATION

各単語は、概念辞書の概念番号と対応付けられており、概念辞書中の情報を引き出すことができる。イ語では１語で表現できない事象・事物であっても、１概念とみなせる場合には、複合語として辞書に記述する。逆に、イ語に固有の概念は、新たに概念辞書に登録し、概念番号を付与する。

辞書の開発は、他の言語の影響を受けない形で進めなければならない。また、機械翻訳技術に精通していないものでも容易に記述できるようにすることが望ましい。これらのことを踏まえて、以下の手順で基本語辞書の開発を進めている。

①イ語の基本語（約５万語）をイ語独自の観点から選択する。

②イ語単語の品詞、構文情報、形態素情報を記述する。

③以下の２つの作業を並行して行なう。

③-1 概念辞書に記述されている概念が対応するイ語単語を全て取り出し、該単語の意味情報欄に概念番号を記述する。対応するイ語単語が見つからない場合は、新たに語を登録し、②の作業を行なう。

③-2 イ語単語が対応する概念を概念辞書から全て取り出し、イ語単語の意味情報欄に概念番号を記述する。対応する概念が見つからない場合は、新たな概念を概念辞書に登録する。

４．文解析および文生成

文解析は、イ語文を中間言語に変換することであり、形態素解析、構文解析、意味解析の３つのフェーズからなる。文生成は、中間言語をイ語文に変換することであり、訳語選択、構文合成、形態素合成の３つのフェーズからなる。解析･生成処理の概要を図３に示す。

解析部では、まず、辞書を参照しながら、各単語の品詞を決定する。この際、辞書検索を効率化するため、イディオム（複合語）テーブルおよび接辞テーブルを用いる。さらに句構造木を作成した後、意味解析を行って中間言語に変換する。

生成部では、中間言語から、共起辞書を用いてイ語訳語を選択し、句構造木を作る。その後、活用形処理等を行いイ語文を合成する。構文合成の前に、必要ならば、中間言語の言い替え（パラフレーズ）を行うこともある。また、イ語単語の検索を効率化するため、概念ーイ語単語対応テーブルを用いる。

解析と生成のエンジンはまったく異なるが、共通化できる部分はできるだけ共通化し、ルールの一貫性、開発の効率化、誤りの減少を図る。例えば、構文解析の結果得られる構文木と構文合成の結果得られる構文木の構造は同一である。また、解析ルールと生成ルールの中の深層格と表層語を対応付ける規則は共通である。イディオム（複合語）テーブルおよび概念ーイ語単語対応テーブルは、基本語辞書から、自動的に作成・保守できる機能を設ける。

解析・生成システムの評価を行なうために、イ語の構文パターンを網羅した３００文程度のコーパスを用意した。これらの文は、規則が網羅されているかをチェックするとともに、規則を変更したときの悪影響検出のために用いる。この他、イ語以外の言語のコーパス文から得られた中間言語も順次利用していく予定である。

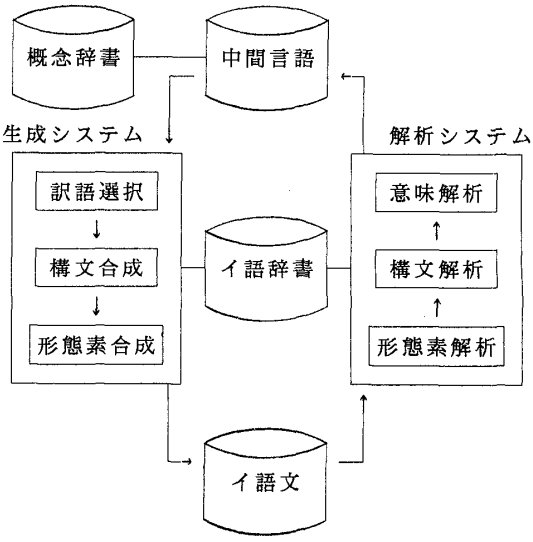


図３　解析・生成処理の概要

５．終わりに

ＣＩＣＣでは、インドネシア政府と協同で、イ語翻訳システムの研究開発を進めている。開発を進めながら、本報告で述べた方法の評価・改良を行なうとともに、種々の開発ツールや翻訳支援ツールを作成していく予定である。

最後に、本研究の機会を与えていただいた(財)国際情報化協力センターの皆様、および、日頃から活発に議論していただいている(財)国際情報化協力センタ及びインドネシア国技術応用評価庁(ＢＰＰＴ)の研究員の皆様に感謝致します。

参考文献

［１］概念辞書(第２版)：日本電子化辞書研究所

［２］単語辞書(第２版)：日本電子化辞書研究所