中日・日中機械翻訳システム

Chinese-to-Japanese / Japanese-to-Chinese Machine Translation System

出羽　達也　　熊野　明
■ IZUHA Tatsuya　■ KUMANO Akira

中国進出企業による情報収集・発信などのビジネス活動支援を主な目的として，日中・中日機械翻訳システムを開発した。高い翻訳精度で定評のある英日・日英機械翻訳ソフト The翻訳™シリーズ[1]に搭載されている翻訳エンジンとモジュールの共有を図ることにより，中日・日中翻訳でもきめ細かな訳し分けを可能にした。同時に，統計ベースの形態素・構文解析技術，及び中国語固有の言語現象である離合詞の解析技術を新規に開発し，高い翻訳精度を実現した。

現在，インターネット上での翻訳サービス実験を通じて，商品化に向けた性能強化を図っている。

Toshiba has developed a Chinese-to-Japanese / Japanese-to-Chinese machine translation system to facilitate the collection and distribution of information by Japanese businessmen in China.

The system incorporates key components of the well-established translation engine used in the English-to-Japanese / Japanese-to-English machine translation system, which is providing efficient, high quality translations.　In addition, we have developed technology for statistics-based parsing and detachable verb analysis specifically for the Chinese language.

The system is now in trial service on the Internet to fine-tune its performance through real use by real users.

1　まえがき

近年の急速な経済発展を背景に，日本企業にとって中国の重要性が非常に高まっている。安価な労働力を活用した調達・生産拠点という従来からの位置づけに加えて，膨大な人口と購買力の向上が巨大市場を形成しつつある。更に，高度な技術力を持つ中国企業が増えてきており，ビジネスパートナーあるいはコンペティターとしても存在感を高めている。このように，マーケティングや研究開発，調達，製造，販売，保守など，ビジネス活動のあらゆるフェーズで言語の壁を超えた日中間の情報発信・収集やコミュニケーションの需要が増している。

これらの状況を背景に，中国進出企業のビジネス活動支援を主な目的として，中日・日中機械翻訳システムを開発した。ここでそのシステムの概要と特徴について述べる。

2　システムの概要

開発したシステムにおける，中日翻訳と日中翻訳の処理の流れを**図1**に示す。機械翻訳の基本方式の主なものには，規則ベース機械翻訳，例文ベース機械翻訳[2]，統計的機械翻訳[3]などがあるが，現在実用に供されている機械翻訳システムのほとんどは規則ベース機械翻訳であり，その中でもトランスファ方式と呼ばれる方式がよく用いられる。このシステムも規則ベースのトランスファ方式を採用している。

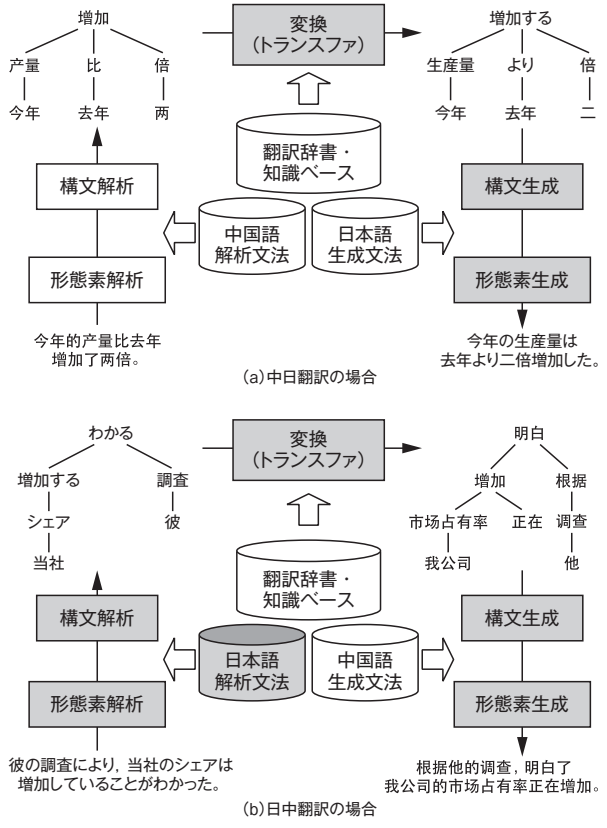

図1．中日・日中翻訳システムの処理概要 — 網掛け部分は，英日・日英翻訳システムと共有している。

Overview of the machine translation system

トランスファ方式による機械翻訳処理は，原文解析，変換（トランスファ），及び訳文生成の3ステップから成る。

原文解析のステップでは，解析文法を参照して処理を行う。中日翻訳の場合は中国語解析文法，日中翻訳の場合は日本語解析文法である。まず入力文を単語に分割して各単語に品詞を割り当てる形態素解析を行った後，単語間の関係，すなわち文の構造を決定する構文解析を行う。

変換のステップでは，翻訳辞書を参照して入力言語の構造を出力言語の構造に変換する。翻訳辞書には，中日・日中翻訳ともに約30万語の見出し語が登録されており，各見出し語は原語表記や，文法属性（品詞ほか），訳語，訳し分け規則などの情報を持っている。

訳文生成のステップでは，生成文法を参照して処理を行う。中日翻訳の場合は日本語生成文法，日中翻訳の場合は中国語生成文法である。まず語順を決定する構文生成を行った後，語尾変化などを処理する形態素生成を行い，最終的な訳文が出力される。

3　システムの特徴

3.1　きめ細かな訳し分け―英日・日英翻訳とのモジュール共有

このシステムは，図1の網掛けの部分で示すように，英日・日英翻訳システムとモジュールを共有している。すなわち，変換と訳文生成のモジュールは，中日・日中翻訳ともに英日・日英翻訳と同じものを使用している（日中翻訳では更に，原文解析モジュールと日本語解析文法も共有している）。既存のモジュールに対して，中国語の文字コードへの対応など若干の拡張を行うことで共有が可能となった。これにより，開発・保守コストが低減できただけでなく，英日・日英翻訳ソフト The翻訳$_{TM}$シリーズに搭載され高い翻訳精度を実現した実績を持つ変換・生成モジュールが中日・日中翻訳でも利用可能となり，これまで培った翻訳知識や生成文法の構築ノウハウの活用と併せて，きめ細かな訳し分けが可能となった。

多くの語は複数の訳語を持つため，適切な訳文を得るには訳し分けが不可欠である。**図2**は，依存関係を参照して“わかる”を“明白”，“懂”，“知道”に訳し分ける日中翻訳規則の例を示している。

図2（a）は，図1（b）の例文に適用される訳し分け規則である。任意の名詞から“手段”の関係で修飾され，かつ任意の用言（動詞，形容詞など）を目的語に取るとき，“わかる”の訳語は“明白”になる。同時に，“手段”の関係でつながっていた名詞との間に“根据”という訳語を持つノード（節点）が追加されている。このように訳し分け規則は，訳語の決定だけでなく構造の変換も同時に行うことができる。

図2（b）は，“～語がわかる”という表現(注1)に適用される訳し分け規則である。表記の末尾が「語」である名詞を目的語に取るとき，“わかる”の訳語は“懂”になる。

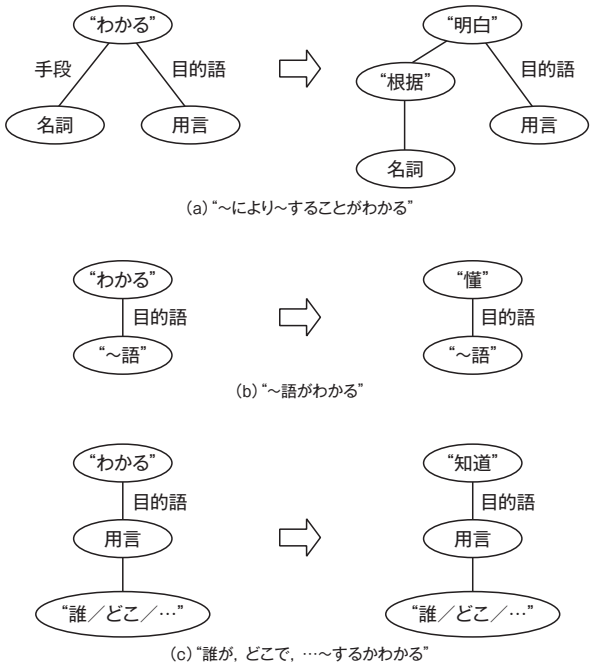

図2. 日中訳し分け規則の例 ― 依存関係を参照して“わかる”を“明白”，“懂”，“知道”に訳し分ける。

Example of target word selection

図2（c）は，“誰が，いつ，どこで，どのように～するかわかる”という表現(注2)に適用される訳し分け規則である。任意の用言を目的語に取り，かつその用言が“誰”や“どこ”などにより修飾されるとき，“わかる”の訳語は“知道”になる。

3.2　中国語解析技術の開発―統計ベース

中日翻訳の原文解析（中国語解析）には，英日・日英翻訳で使われている既存技術を活用し，中国語解析文法だけを新規開発するという選択肢もあったが，統計ベースの原文解析技術を新規に開発することにした。既存の原文解析は規則ベースであるため，専用の文法規則を作り込む必要があり，文法開発のコストが大きくなってしまう。一方，統計ベースの手法では，既存の言語データから文法を自動抽出するため，文法開発のコストを低く抑えることができる。

中国語は語の間の切れ目が明示されない膠着（こうちゃく）語であるため，計算機で単語分割を行う際にあいまい性が生じる。**図3**は“不断开发引领世界的新技术（世界をリードする新技術を絶え間なく開発する）”という中国語例文を形態素解析する過程で生じるあいまい性を示しており，実線で結ばれた単語列が正しい結果である。多数の候補の中から正しいものを選ぶ（あいまい性を解消する）ために，この

（注1）　例えば，“日本語がわかりますか？（你懂日语吗?）”
（注2）　例えば，“どのように洗濯機を使うかわかる（知道怎么用洗衣机）”

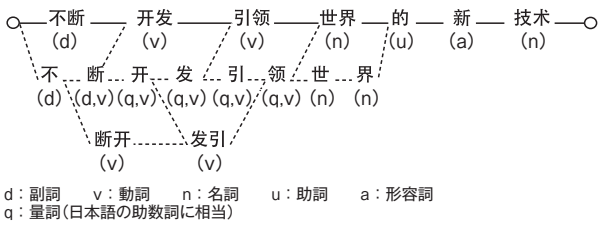

図3. 形態素解析のあいまい性の例 — “不断开发引领世界的新技术”という例を形態素解析する過程で，単語分割と品詞付与の複数の候補が得られ，あいまい性が生じる。実線が正しい結果である。

Example of ambiguity in morphological analysis

システムは接続コスト最小法という方法を採用している。

形態素解析結果(単語列)の候補の一つを$W=w_1 \cdots w_n$と表し，そのときの品詞列を$T=t_1 \cdots t_n$としたとき，Wが得られる確率$P(W)$は式(1)で与えられ，もっとも大きい$P(W)$を与える候補が正しい結果であると推定される。

$$P(W)=\prod P(t_i \mid t_{i-1}) P(w_i \mid t_i) \qquad (1)$$

式(1)において$P(t_i \mid t_{i-1})$は品詞の連接確率，$P(w_i \mid t_i)$は品詞別の単語出現確率である。このような確率を得るには，正しく単語分割され品詞が付与された大量のテキスト(品詞タグ付きコーパスと言う)が必要になる。このシステムでは，中華人民共和国教育部 言語文字応用研究所が開発した品詞タグ付きコーパスの一部(約1,700万語)を利用している。

構文解析では，形態素解析結果に対し，チャート法というアルゴリズムを用いて文脈自由文法(注3)(CFG：Context Free Grammar)をボトムアップに適用する。**図4**は，“送朋友一件礼物(友達にプレゼントをあげる)”という例にCFGを適用して得られた構文森である。意味的に正しい解釈は実線で示された動詞句②であるが，構文的には動詞句①や名詞句③の解釈も可能であり，あいまい性解消が必要となる。形態素解析同様，構文解析のあいまい性解消にも確率モデルを用いる。

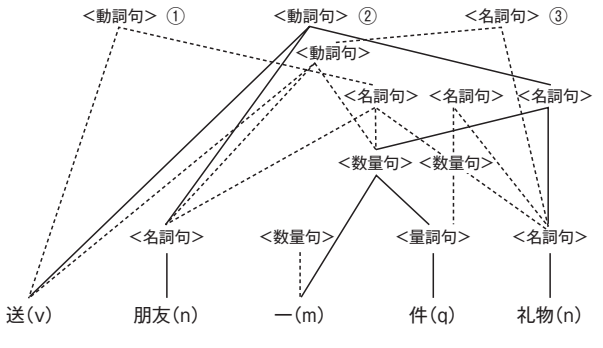

図4. 構文解析のあいまい性の例 — “送朋友一件礼物”という例を構文解析する過程で，複数の構文候補が得られ，あいまい性が生じる。実線が正しい結果である。

Example of ambiguity in parsing

l個の部分木を組み合わせてできる構文木$T=\{t_1,\cdots,t_l\}$の生成確率は，各部分木の独立を仮定して次式により求める。

$$P(T)=\prod P(t_i) \qquad (2)$$

部分木はCFG規則に対応しているため，部分木の確率を求めるためにはCFG規則の確率が必要になる。確率を伴ったCFGをPCFG (Probabilistic CFG)と呼ぶ。近年は，動詞や名詞句といった文法カテゴリーだけでなく，そこに含まれる語いを考慮して確率を求める語い化(Lexicalized) PCFG[4, 5]が主流となっており，このシステムも同様のアプローチを取っている。このような確率を求めるには，大量の構文解析正解データ(構文タグ付きコーパスと言う)が必要になる。このシステムでは，ペンシルバニア大が開発したPenn Chinese Treebank 5.0[6](約50万7千語)を利用している。

図4に示すような構文木は句構造と呼ばれるが，変換以降の処理がしやすいように，図1に示すような依存構造へ変換する後処理を行っている。

3.3 中国語固有の言語現象への対応―離合詞の処理

前節で述べた解析アルゴリズムは特定の言語に依存したものではないが，中国語独特の言語現象である離合詞[7]を適切に処理するためには，特別な処理が必要である。

離合詞は2文字以上から成る単語で，途中にほかの成分を挿入することができる。例えば，“生气(腹を立てる)”という単語を用いて“私に腹を立てる”という表現をするときには間に“我的”を挿入して“生我的气”となる。

“生”と“气”の個々の意味を単純に組み合わせても“腹を立てる”という意味にはならないため，“生气”を2語とみなすのは適当ではない。

形態素解析結果に対して，あらかじめ用意したパターンにマッチする成分を読み飛ばすことにより離合詞を認識するというアプローチも考えられるが，非常に複雑なパターンを用意する必要があるうえに，読み飛ばした成分のその後の処理が難しい。そこで，このシステムでは構文解析の後処理として離合詞処理を実現した。

“生我的气”の例に対する離合詞処理の概要を**図5**に示す。まず，構文解析結果は図5 (a)のようになる。“生”は，“生气”のほか“生病”や“生锈”などの離合詞の前方成分になるため，これらの離合詞の後方成分のリスト(“气”“病”“锈”)をあらかじめ持っている(注4)。図5 (b)では，“生”の子ノードの中からこのリストの要素にマッチするものを探し，見つかった“气”のノードを統合する。これに伴い“气”の子ノードであっ

(注3)　句の構成規則を，＜動詞句＞→＜動詞＞＋＜名詞句＞のような形式で定義する。“→”の右辺の語句の並びが左辺の句を構成する。

(注4)　北京大学の中国語辞書[8]には約3,400個の離合詞が登録されている。離合詞の後方成分のリストはこのデータを利用して作成した。

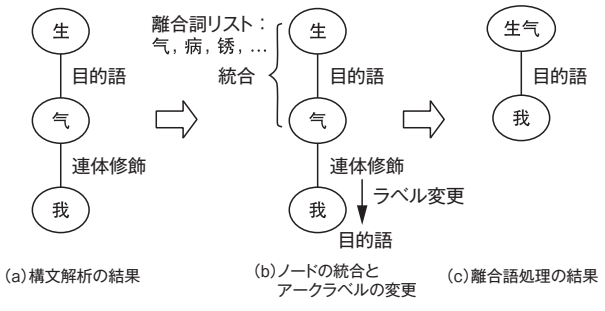

図5. 離合詞処理の概要 — 中国語離合詞“生气”の用例である“生我的气”に対する処理は，まず構文解析を行い，次に離合詞リストを元に“生气”を統合し，最後に“我”を目的語に変更して，その結果が得られる。

Overview of Chinese detachable verb processing

た“我”は統合してできる“生气”の子ノードになるが，その際ノードとノードを結ぶ線のアークのラベルを“連体修飾”から“目的語”に変更する。図5 (c)は離合詞処理の結果である。図5 (a)から“私に腹を立てる”という訳を得るためには，変換で複雑な訳し分け規則を記述する必要があるが，図5 (c)であれば一般的な翻訳規則だけで十分である。

4　あとがき

中国進出企業による情報収集・発信などのビジネス活動支援を主な目的として開発した中日・日中機械翻訳システムについて述べた。

現在，インターネット上での翻訳サービス実験[9]を通じて，商品化に向けた性能の強化を図っているところである。今後は更に，英日・日英翻訳システムに先行搭載されている技術の導入を進め，翻訳精度向上に注力していく。

文　献

(1) 東芝ソリューション(株). 英日／日英翻訳ソフト The翻訳TMシリーズ. <http://hon-yaku.toshiba-sol.co.jp/>, (参照2006-12-19).
(2) Nagao, M. "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle." Artificial and Human Intelligence. Alick Elithorn and Ranan Banerji. Elsevier, 1984, p.173－180.
(3) Brown, P.F., et al. A Statistical Approach to Machine Translation. Computational Linguistics. **16**, 2, 1990, p.79－85.
(4) Collins, M. "Three Generative, Lexicalised Models for Statistical Parsing". In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. Madrid, 1997-07, the Association for Computational Linguistics. p.16－23.
(5) Bikel, D.M., et al. "Two Statistical Parsing Models Applied to the Chinese Treebank". In Proceedings of Second Chinese Language Processing Workshop. Hong Kong, 2000-10, the Association for Computational Linguistics. p.1－6.
(6) University of Pennsylvania. "Penn Chinese Treebank Project". <http://www.cis.upenn.edu/~chinese/ctb.html>, (参照2006-12-19).
(7) 中山時子, ほか. 中国語離合詞500. 東京, 東方書店, 1990, 239p.
(8) Yu, S., et al. The Grammatical Knowledge-base of Contemporary Chinese - A Complete Specification. Beijing, Tsinhua University Press, 2002, 958p.
(9) (株) ニューズウォッチ. フレッシュアイ 中日・日中翻訳サービス. <http://mt.fresheye.com/>, (参照2006-12-19).

出羽　達也　IZUHA Tatsuya
研究開発センター 知識メディアラボラトリー主任研究員。主に自然言語処理の研究・開発に従事。情報処理学会，電子情報通信学会，言語処理学会会員。
Knowledge Media Lab.

熊野　　明　KUMANO Akira
研究開発センター 知識メディアラボラトリー主任研究員。主に機械翻訳システム及び電子化辞書の研究･開発に従事。情報処理学会，人工知能学会，言語処理学会会員。
Knowledge Media Lab.