

日中機械翻訳プロトタイプ

山田節夫 林 実 片岡明 横尾昭男 大山芳史

日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

〒 619-0237 京都府相楽郡精華町光台 2-4

E-mail: {syamada,hayashi,kataoka,ayokoo,ooyama}@cslab.kecl.ntt.co.jp

あらまし

我々は以前より日英機械翻訳システム (ALT-J/E) の研究開発を進めてきている。しかし、近年インターネットの利用が増大し、英語圏以外にも情報発信する場面が増えてきたため、他の言語についても、機械翻訳が望まれてきている。そこで、我々は、先ず最初にこの ALT-J/E を日本語からマレー語への翻訳に適用するのを試みた。本稿では、さらに日本語から中国語への翻訳に適用するのを試み、中国語以外にも多数の言語が同時に扱える Unicode に対応した日中機械翻訳プロトタイプについて述べる。

キーワード 機械翻訳、日中翻訳、中国語、Unicode

A Prototype Japanese-to-Chinese Machine Translation

Setsuo Yamada Minoru Hayashi Akira Kataoka Akio Yokoo Yoshifumi Ooyama

NTT Communication Science Laboratories, NTT Corporation

2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan

E-mail: {syamada,hayashi,kataoka,ayokoo,ooyama}@cslab.kecl.ntt.co.jp

Abstract

We can send information to the world easily through the Internet. However, it is hard to translate our native language into other languages. We have been researching Japanese-to-English Machine Translation (ALT-J/E), and Japanese-to-Malay Machine Translation. Recently, we also started to work on Japanese-to-Chinese Machine Translation based on ALT-J/E. This new system uses Unicode, an encoding method which can handle many kinds of characters. This paper describes a prototype Japanese-to-Chinese Machine Translation.

key words Machine translation, Japanese-to-Chinese Machine Translation, Chinese, Unicode

1 はじめに

近年インターネットの利用が増加し、世界に情報発信するために外国語へ翻訳する機会が多くなってきている。母国語から英語への翻訳に関しては、従来から数多く提案されている機械翻訳システムを利用することができる。一方、英語以外への翻訳のニーズが高まっている中、最近では、英語以外の言語間での翻訳システムに関して、[Mit99, Sum99]などの多言語に関する翻訳システムが提案されている。しかし、英語への翻訳システムに比べて数が少ない。また、日中翻訳に関しても数が少なく、[Zon00, Yam01]などが提案されているものの、まだ検討する内容が数多くある。

我々は英語以外への翻訳に関して、以前から研究開発を進めている日英機械翻訳システム(ALT-J/E)[Ike91]のマレー語への適用を試みている[Ogu99]。本稿では、さらに中国語への適用を試み、中国語以外にも多数の言語が同時に扱えるUnicodeに対応した日中機械翻訳プロトタイプについて述べる。

今回ベースとした日英機械翻訳システムのうち、日中翻訳に適用できた部分について、最初に日英での使われ方を説明する。次に、中国語への適用例について述べ、最後に課題を示す。

2 日英翻訳

ALT-J/Eは多段階翻訳方式を利用したルールベース型の機械翻訳システム[Ike91]である。日本語が入力されると、大枠は、形態素解析、係り受け解析、意味解析、日英構造変換、英文生成と処理が行われる。日中翻訳にも適用できた以下の項目に関して、日英での使われ方を述べる。

- 日本語書き替え
- 日英対照辞書
- 日英構文意味辞書

2.1 日本語書き替え

ALT-J/Eは、人による前編集をなくすため、翻訳しにくい日本語を翻訳しやすい日本語へ自動的に書き替える処理を行っている[Shi95]。例えば、例文1の日本語文には、“乗る”、“行く”という動詞が使われている。これらをそのまま英語でも動詞として翻訳

すると不自然な英語になってしまう。そこで、述語性の意味合いを薄くし、“自転車に乗って”を“自転車”で、また、“電車に乗って”を同様に“電車”と書き替えることで、自然な英語に翻訳しやすくしている。この例では、それぞれ、“by bicycle”、“by train”と英訳され、例文1(日英)に示すような英語に翻訳される。

例文 1 (日英)

日： 彼は 自転車に乗って 学校へ行き、
英： He goes to school by bicycle and
日： 私は 電車に乗って 会社に行く。
英： I go to a company by train.

2.2 日英対照辞書

日英対照辞書には、約38万語が登録されていて、単語の意味属性、英訳、訳語選択条件等が記述されている。意味属性は日本語語彙大系[Ike97]で定義されている約3,000カテゴリに分類されている。名詞の場合、語彙大系は12レベルの階層構造で表され、その例を図1に示す。例えば、“ホテル”や“旅館”には〈宿泊施設〉等の意味属性が付与されている。

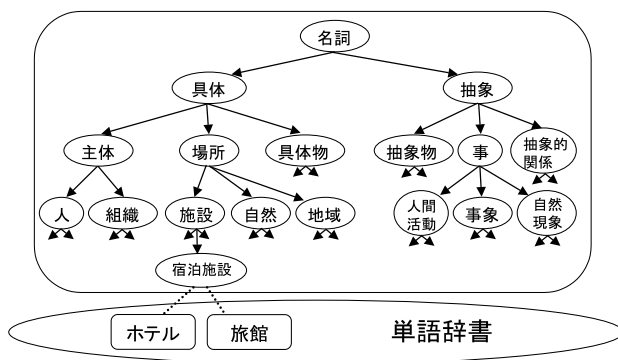


図 1: 語彙大系の例

表1に示す例のように、日英対照辞書エントリーに、単語の意味属性を記述することによって、適切な英語に訳し分けることができる。この例では、同じ“犬”でも、他の語との共起関係で、意味が〈スパイ〉なら、英訳が“spy”となり、意味が〈獣〉なら、英訳は“dog”となる。

表 1: 日英対照辞書エントリー例 1

日本語見出し	英訳	意味属性
犬	dog	〈獣〉
犬	spy	〈スパイ〉

また、他の訳し分け条件も記述されている。例えば、日本語では同じ“鼻”でも、象の“鼻”は“trunk”、豚の“鼻”は“snout”と訳し分けする必要があるが、何の“鼻”かの条件を表2のように記述しておけば、適切に訳し分けすることができる。その他一般の場合に“nose”と訳出できるよう、“nose”の訳し分け条件には“なし”と記述している。

表 2: 日英対照辞書エントリー例 2

日本語見出し	英訳	訳し分け条件
鼻	trunk	象
鼻	snout	豚
鼻	nose	なし

2.3 日英構文意味辞書

日英構文意味辞書には、約16,000パターンが登録されていて、各パターンは、日本語パターンとそれに対応する英語パターンの組から構成されている。日本語パターンには、日本語の用言（動詞、形容詞）とその用言が取り得る格要素が制約条件として記述され、英語パターンには、その日本語用言の英訳と日本語パターンの制約条件にマッチする英語の格要素とが記述されている。

図2に“取る”の日英パターンの例を示す。最初のパターンは、主格が〈主体〉で、目的格が〈許可, 賛成〉等が条件となり、日本語文がこの条件にマッチすれば、“取る”の訳は“get”となる。同様に、〈主体〉が〈休暇, 手〉等を取る場合は“take”、〈主体〉が〈宿泊施設, 席〉等を取る場合は“reserve”となり、英語パターンはそれぞれ図2に示した通りとなる。ここで、N1、N2 は変数を表し、日本語パターンと英語パターンの中で対応関係にある。例えば、日本語パターンにあるN1は、英語パターンにあるN1と対応している。

図2の日英パターンにより、例えば、「私は休暇を取る許可を取り、妻はホテルを取った。」は、例文2（日英）に示すように、適切な英語へと翻訳することができる。

例文 2 (日英)

日： 私は休暇を 取る 許可を 取り、
 英： I got permission to take a vacation and
 日： 妻はホテルを 取った。
 英： my wife reserved a hotel.

日本語パターン	⇒	英語パターン
N1: <主体> が N2: <許可, 賛成, ...> を 取る	⇒	N1 get N2
N1: <主体> が N2: <休暇, 手, ...> を 取る	⇒	N1 take N2
N1: <主体> が N2: <宿泊施設, 席, ...> を 取る	⇒	N1 reserve N2

図 2: “取る”の日英パターン対の例

3 日中翻訳

本節では、2節で挙げた、日本語書き替え、日英対照辞書、日英構文意味辞書について、日中翻訳ではどのように適用しているのかを述べる。

3.1 日本語書き替え

日英翻訳で使われている日本語書き替えルールは、主に英語への翻訳を意識して作成されているが、中国語にも適用可能なものもある。例えば、例文1に関しては、英語と同じで、日本語に出てきた動詞をそのまま動詞として翻訳すると不自然な中国語となってしまう。そこで、日中翻訳においても、この例文1に適用された書き替えルールは利用できる。

ただし、英語では書き替えた後、“～で”が交通手段の場合、“自転車でも”“電車でも”で”の訳は“by”で良いが、中国語の場合は、訳し分けが必要である。“自転車”の場合は、“骑”、“電車”の場合は、“乘”と訳す必要がある。これは、表3に示すように、助詞“で”の訳し分け条件に、何の乗り物に乗るかを記述しておけば、適切な中国語に翻訳できる。

表 3: 助詞“で”の訳し分け条件の例

助詞	中国語訳	訳し分け条件
で	骑	自転車, 馬, ...
で	乘	電車, バス, ...

例文1の日本語文は書き替え処理、中国語固有の“で”の訳し分けを適用すると適切に翻訳され、以下のようになる。

例文 1 (日中)

日：彼は 自転車に乗って 学校へ行き、
 中：他 骑 自行车 去 学校
 彼 で 自転車 行く 学校
 日：私は 電車に乗って 会社へ行く。
 中：我 乘 电车 去 公司
 私 で 電車 行く 会社

3.2 日中対照辞書

日中対照辞書は、日英翻訳と同様に単語の意味属性、中国語訳、訳語選択条件等が記述されている。ただし、訳語選択条件は、日英用とは違い、中国語の言語現象を考慮して作成しなければならない。例えば、日英翻訳では、象の“鼻”と豚の“鼻”を訳し分けする必要があるが、日中翻訳では、その必要はない。というのは、中国語でも同じ“鼻子”と訳されるからである。

一方、中国語は多様な分類辞を使う言語であり、日本語の分類辞とは違った使い方をしているので、訳し分けの必要がある。この例を例文3(日中)に示す。

例文 3 (日中)

日：一 匹 の猫が三 匹 の魚を見ている。
 中：一 只 猫 看 着 三 条 鱼
 一 匹 猫 見る 進行 三 匹 鱼

日本語では、猫も魚も分類辞は“匹”であるが、中国語では、猫のような動物を数える場合は“只”、魚のような細長いものを数える場合は“条”と訳し分けをしなければならない。これを解決するために、表4に示す通り、日中対照辞書のエントリーに分類辞を付与した。このように分類辞を登録すれば、日英翻訳と同じ枠組みで適切に翻訳が可能となる。

表 4: 日中対照辞書エントリー例

日本語見出し	中国語訳	分類辞情報
猫	猫	只
魚	鱼	条

3.3 日中構文意味辞書

日中構文意味辞書は、日英翻訳と同様に、日本語パターンとそれに対応する中国語パターンの組から構成される。日本語のパターンも日英用の辞書と同様に、

日本語の用言とその用言が取り得る格要素が制約条件として記述できる。ただし、制約条件は中国語用に作成し直す必要がある。中国語パターンは、その日本語用言の中国語訳と日本語パターンの制約条件にマッチする中国語の格要素とが記述できる。しかし、中国語の構造を考慮したパターン作成が必要となる。

図3に“取る”の日中パターンの例を示す。最初のパターンは、主格が〈主体〉で、目的格が〈許可, 賛成〉等が条件となり、日本語文がこれにマッチすれば、“取る”の訳は“得到”となる。この条件は、日英の場合と同じであるが、2番目、3番目のパターンは、日英の条件と異なり、図2の2番目の目的格の条件が分かれる形となる。つまり、中国語では〈主体〉が〈休暇〉等を取る場合は“请”、〈主体〉が〈手〉等を取る場合は“拉”となる。

また、〈主体〉が〈宿泊施設, 席〉等を取る場合は“预约”となり、それぞれ中国語パターンは図3に示した通りとなる。

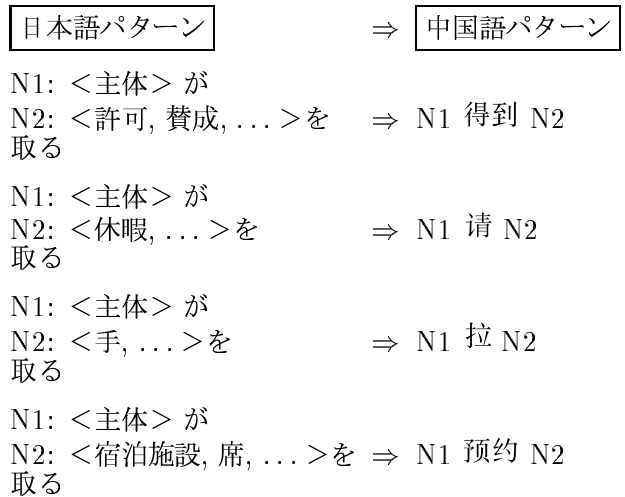


図 3: “取る”の日中パターンの例

この日中構文意味辞書を使うことにより、例文2の日本語動詞“取る”は適切に翻訳され、以下のようになる。

例文 2 (日中)

日：私は休暇を 取る 許可を 取り、
 中：我 得到 了 请 假 的 许可
 私 取 过去 取 休假 修饰 许可
 日：妻はホテルを 取った。
 中：妻子 预约 了 宾馆
 妻 取 过去 ホテル

3.4 プロトタイプ

上述したようにして、日英翻訳をベースに作成した日中機械翻訳プロトタイプ (ALT-J/C) の翻訳画面を図4に示す。Unicode を用いているので、日本語と中国語を同時に表示することができている。また、翻訳結果は、本稿で述べた例文1、2、3の順に表示している。



図4: 日中機械翻訳プロトタイプの翻訳画面

4 課題

本稿では、日中機械翻訳プロトタイプについて、日英翻訳をベースに実現できた部分について述べた。しかし、日英翻訳の枠組みでは実現が難しい問題も数多くある。特に、中国語固有の言語現象を扱うには、新たな枠組みを考える必要がある。本節では、日中機械翻訳を実現させるための課題のうち幾つかを挙げる。

動詞

中国語の動詞は英語と違い重ね型動詞、離合動詞などの言語現象が良く知られている。

重ね型動詞は、試行の様相の場合に動詞を重ねて表現する動詞である。試しに何かの動作を行う時に、中国語の動詞は繰り返される場合がある。例えば、“ちょっと見る”の場合、“見る”の中国語訳“看”が繰り返されて“看看”や“看一看”と表現される場合が多い。

離合動詞は、2文字以上から成る動詞に起こる現象で、普段は2文字で1つの動詞として扱われるが、時制や様相を表す時に、それらを表す語が2文字の

漢字の間に挿入される。例えば、“結婚したことがある”の場合、“結婚する”の中国語訳は“结婚”であるが、経験を表す場合は“结”と“婚”の間に“过”が挿入されて、“结过婚”と表現される。

上記の言語現象は、英語にはない現象なので、新たな枠組みが必要である。また、この言語現象は各動詞に依存するので、日中構文意味辞書に何らかの情報を入れる必要もある。

辞書の拡充

辞書の大枠は日英翻訳と同じなので、例えば中国語の訳語そのものは、電子化されている日中辞書等を用いて、作成することは可能である。しかし、他の内容（特に訳し分け条件や中文構造等）は、英語と同じとは限らない。

例えば、3.3節で述べたように、“取る”の訳し分け条件は、図2、3に示す日本語パターンのように、違う場合がある。そこで、日中翻訳ならではの情報を効率的に辞書に反映する手法の検討が必要である。

語彙大系

語彙大系は、英語を意識して日本語の意味を約3,000カテゴリに分類しているが、分類の粒度に関しては中国語に必ずしも適切であるとは限らない。訳し分け条件をより細かく指定したい場合や、逆に細かいのでまとめたい場合がある。

例えば、例文1の日本語文を中国語文に適切に訳出するには、乗り物によって“で”の訳を変える必要があるが、現在の語彙大系の中では〈乗り物〉という意味属性しかなく、これだけでは訳し分けの条件が記述できない。そこで、現在は、表3に示した通り、訳し分け条件を直接単語で指定している。将来は、中国語の訳し分けに必要な意味属性を新設する等の対策が必要である。

5 おわりに

我々は、以前から研究開発を進めている日英機械翻訳システム (ALT-J/E) を中国語への翻訳に適用するのを試みた。その結果、まだ多くの課題があるものの、幾つかの日英翻訳の枠組みは利用できることが分かった。また、Unicodeにも対応しているので、日中同時に取り扱うことができ、将来は中国語のみならず、他の言語でも日英翻訳をベースに発展させることができる。

今後は、日中用に辞書エントリーの修正や拡充をして、翻訳できる語彙やパターンを増やしていく予定である。また、日英ベースでは実現が難しい中国語固有

の言語現象についても、問題を整理して、翻訳できる表現の種類を増やすために新たな枠組みを検討する予定である。

参考文献

- [Ike91] S. Ikehara, S. Shirai, A. Yokoo, and H. Nakaiwa. Toward an MT system without pre-editing — effects of new methods in **ALT-J/E** —. In *the Machine Translation Summit III*, pp. 101–106, 1991.
- [Ike97] 池原他編集. 日本語語彙大系. 岩波書店, 1997.
- [Mit99] T. Mitamura. Controlled language for multilingual machine translation. In *the Machine Translation Summit VII*, pp. 46–52, 1999.
- [Ogu99] K. Ogura, F. Bond, and Y. Ooyama. **ALT-J/M** A prototype Japanese-to-Malay Translation System. In *the Machine Translation Summit VII*, pp. 444–448, 1999.
- [Shi95] 白井諭, 池原悟, 河岡司, 中村行宏. 日英機械翻訳における原文書き替え型翻訳方式とその効果. 情報処理学会論文誌, Vol. 36, No. 1, pp. 12–21, 1995.
- [Sum99] E. Sumita, S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishikawa, and S. Shirai. Solutions to problems inherent in spoken-language translation: The ATR-MATRIX approach. In *the Machine Translation Summit VII*, pp. 229–235, 1999.
- [Yam01] K. Yamamoto, S. Shirai, M. Sakamoto, and Y. Zhang. Sandglass: twin paraphrasing spoken language translation. In *ICCPOL 2001*, pp. 154–159, 2001.
- [Zon00] C. Zong, Y. Wakita, B. Xu, K. Matsui, and Z. Chen. Japanese-to-Chinese spoken language translation based on the simple expression. In *ICSLP 2000*, Vol. IV, pp. 418–421, 2000.