Internal Use Only（非公開）

TR－SLT－0064

日中／中日機械翻訳システム

Machine translation system between Japanese and Chinese

坂本　仁

Masashi SAKAMOTO

2004年3月31日

2000年5月から2004年3月まで、ほぼ4年間にわたって研究開発を行なってきた日中機械翻訳システムについて説明する。本システムは中国語の音声認識技術、音声合成技術と共に研究開発を進め、日中間の音声対話翻訳を実現しようとしたものであり、実際に2003年11月の「ＡＴＲ研究発表会2003」において、日本人と中国人との音声対話翻訳実験に成功した。本稿では特に、トランスファー(言語変換)部を中心に説明する。また、研究開発の成果として、本技術の実用化に有用と考えられる発明を提案する。

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

©2004（株）国際電気通信基礎技術研究所
©2004 Advanced Telecommunication Research Institute International

目　次

1.開発経緯

2000 年 5 月 21 日から 2004 年 3 月 31 日までの研究開発活動全体を概観できるように、年度別にトピックを挙げる。

1.1. 2000 年度

翻訳辞書開発

- 学研アンカー和英辞書の日本語見出しの内、日中 TDMT にないもの
- ATR コーパスに現れる語（多くは固有名詞）

コーパス整備

- TDMT 用訓練文（SLDB、LDB/JEK の抜粋）
- LDB/JEK のパラフレーズ、2 万文に各 2 文の換言文（中国科学院）
- SLDB 全部、フレーズブック 20 万文の約 1/3 を中国語翻訳（インターグループ）

パターン翻訳

- パターン翻訳のランタイム処理系を検討
- 翻訳パターンを人手で記述（TDMT 用訓練文を対象）

中国語パラフレーズ

- 基本検討に着手

日本語パラフレーズ

- 中日翻訳（日本語生成）用に着手

1.2. 2001 年度

翻訳辞書開発

- 学研を通じて学習者向け日中／中日辞書の電子データを購入
- 上記データ用に、コード変換や日本語ひらがな表記の漢字化等のソフトウェア開発

コーパス整備

- フレーズブック 20 万文の 2/3 を中国語翻訳（インターグループ）

中日パターン翻訳

- 中国語基本動詞辞書の整備
- 中国語基本名詞辞書の整備

中国語パラフレーズ

- 中国科学院の中英翻訳との共通化も視野に入れて検討、プロトタイプ開発に着手

日本語パラフレーズ

- 日中翻訳用には待遇表現等を検討
- 中日翻訳（日本語生成）用にはコーパスベースで試作

1.3. 2002 年度

翻訳辞書整備

- 学研を通じて入手の中日辞書の日本語を語釈、説明ではなく「訳語」に書き直し

コーパス整備

- SDB/TRA を中国語翻訳（インターグループ）
- SLDB のパラフレーズ、1 万 6 千文に各 2 文の換言文（中国科学院）

中日パターン翻訳

- 対訳コーパスからの翻訳パターン自動生成により、完全なコーパスベースにシフト

日本語パラフレーズ

- 多言語翻訳を効率よく実現するため、日本語処理重視
- 完全なコーパスベースに方式シフト

1.4. 2003 年度

翻訳辞書整備

- 中国語基本用言辞書例文（中国語 700 動詞、2000 語義、5 万文）を日本語翻訳
- IPAL 日本語動詞辞書・形容詞辞書の例文 9000 文を中国語翻訳

コーパス整備

- フレーズブック 50 万文へ対象拡大、BTEC3,Telephon を中国語翻訳（コングレ）

日中／中日双方向翻訳

- 対訳コーパスからの翻訳パターン自動生成を日中翻訳にも適用可能とし双方向化

日本語パラフレーズ

- 日中翻訳でも音声認識と言語変換との間に位置して機能することを確認

中国語パラフレーズ

- コーパスベースの中国語構文解析等、検討再開

2.設計の狙い

採用した方式の背景となる考え方について説明する。

2.1.モジュール化

考え方

多言語間の翻訳へ発展させやすいように、日本語、中国語という言語の関わりで以下のように３タイプ、４つのモジュールに分割した。

- 日本語自動換言部（日本語→日本語）
- 中国語自動換言部（中国語→中国語）
- 日中言語変換部（日本語→中国語／中国語→日本語）

日本語自動換言部と中国語自動換言部との２つのモジュールは、日中言語変換部とは独立である。つまり、日本語自動換言部は、例えば日英言語変換部や日韓言語変換部の前に構成しても、英日言語変換部や韓日言語変換部の後に構成しても問題なく動作する。これによって、日本語を中心に多言語との翻訳へ展開する場合に、日本語自動換言部に開発資源を集中し、各言語毎に開発せざるを得ない言語変換部に要する開発資源を低減することで全体として効率の良い研究開発を可能にする。また、中国語自動換言部は、中国語と英語等との翻訳を目指す中国の研究機関との間で、研究プラットホームの共通化等の連携、交流を容易にする狙いがある。

2.2.コーパスベース化

考え方

人間に「航空券を予約する」は中国語で「預訂飛机票」であり、「航空券」は中国語で「飛机票」であることを教え、「部屋」が中国語で「房間」であることを教えれば、「部屋を予約する」は中国語で「預訂房間」であろうと推測できる。この場合、人間は「航空券」と「飛机票」という特定の語句を一般化（または汎化）することで、日本語「～を予約する」は中国語で「預訂～」であるという翻訳知識を学習し、汎化した部分に「部屋」と「房間」という特定の語句の対応を適用することで、日本語「部屋を予約する」を中国語「預訂房間」に翻訳したと考えるのが自然である。この推測機能を直接的に計算機システム上に実現することが狙いである。

本手法は統計的な翻訳手法等とは異なり、コーパスから自動的に獲得した翻訳知識を人間が読める形式にして、人間が翻訳知識を確認したり直接記述したりすることが可能になる。これによって、コーパスから自動的に獲得した翻訳知識と人間が直接記述した翻訳知識とを同じ機構で利用して、比較したり補完し合ったりという柔軟な運用を可能にする。

3.トランスファー（言語変換）部の構成

直接、開発を担当したトランスファー（言語変換）部の構成について説明する。必要なソフトウェア資源を列挙し、個々のソフトウェア資源の利用法という観点で説明していく。

3.1.概要

トランスファー部は、以下のソフトウェア資源を前提に構成される。

3.1.1.学習（翻訳準備）時

必要な外部ソフトウェア

(1)日本語形態素解析器

必要なデータ

(2)日中対訳コーパス

(3)日中対訳辞書

(4)汎化品詞表

(5)単語頻度表

3.1.2.ランタイム（翻訳実行時）

必要な外部ソフトウェア

(1)ストリームエディタ

必要なデータ

(2)日中／中日文字対応表

3.2.学習（翻訳準備）時のソフトウェア資源

3.2.1.日本語形態素解析器

翻訳知識作成では形態素の再分割はしないので、細かく分割するタイプのものが適している。(ChaSen ver2.2.8 を利用した。http://chasen.aist-nara.ac.jp/ を参照)

3.2.2.日中対訳コーパス

中国語文と日本語文を「 <--> 」という区切り文字列（空白も含む）で区切って、1 文ずつの対を、1 行 1 対として並べてある。一つの日本語に複数の中国語が対応すれば日本語をコピーし、一つの中国語に複数の日本語が対応すれば中国語をコピーして、各行は必ず中国語、日本語 1 文ずつの対として構成する。ここでいう 1 文とは言語翻訳における処理の単位である。文字コードは、中国語 EUC コードには 0x91 を、日本語 EUC コードには 0x92 をそれぞれ付加する EMACS-MULE コードである。同じコーパスから日中方向の翻

訳パターンと中日方向の翻訳パターンとを生成する。

3.2.3.日中対訳辞書

日本語単語と中国語訳語を空白文字(スペースかタブ)で区切って、1語ずつの対を、1行1対として並べてある。一つの日本語に複数の中国語が対応すれば日本語をコピーし、一つの中国語に複数の日本語が対応すれば中国語をコピーして、各行は必ず日本語、中国語1語ずつの対として構成する。文字コードは、日本語、中国語ともEUCコードである。同じ辞書から日中方向の翻訳パターンと中日方向の翻訳パターンとを生成する。

3.2.4.汎化品詞表

日本語形態素解析器が出力する品詞と、その品詞についての出現頻度の閾値とを空白文字で区切って、1行1対として並べてある。記述された品詞の語については、次項の単語頻度表を参照して得られる出現頻度が、記述された出現頻度の閾値以下の場合に、汎化する。出現頻度の閾値が記述されていなければ1が記述されたものとみなす。記述されていない品詞の語は汎化しない。文字コードは、EMACS-MULEコードである。

3.2.5.単語頻度表

日本語形態素解析器が出力する単語と、その単語についての出現頻度とを空白文字で区切って、1行1対として並べてある。一般名詞は汎化しても述語動詞は汎化しないようにする場合に、例えば、「部屋の予約をする」という場合の「予約」のように意味的には述語であるが形態的には名詞であるものは、そのままでは汎化されてしまう。そこで、「予約」に付与される品詞についての前項の汎化品詞表の閾値を上回る出現頻度と対にして、この「予約」という単語を記述すれば、汎化されなくなる。文字コードは、EMACS-MULEコードである。この表は無くともよい。

3.3.ランタイム（翻訳実行時）のソフトウェア資源

3.3.1.ストリームエディタ

ストリームエディタとは、人間が表示装置を見ながら編集操作を行う対話的なエディタとは異なり、あらかじめ与えられた編集操作を入力単位毎に適用し直ちに出力していくことを入力がなくなるまで続けていく、文字列の変換器である。初期のUNIX（Version 7 AT&T UNIX）から備えられているSEDが代表的であるが、日中間の言語変換を実現するには機能不足であった。代わりにPerl、GNU Awk、Ruby、Tclといったプログラミング言語であれば容易に実現できる。(特にPerl言語のインタプリタはSEDのエミュレーションモードを持っているので、主に利用してきた。 http://www.perl.com/ を参照)

3.3.2.日中／中日文字対応表

日中文字対応表は、中国語の 2 バイト文字コードを対応する日本語文字コードの順に配列したものであり、中日文字対応表は、日本語の 2 バイト文字コードを対応する中国語文字コードの順に配列したものである。例えば、日本語漢字「機」の文字コードから「機」が日本語漢字の中で何番目であるかを算出し、日中文字対応表のその位置の 2 バイトを読み取れば対応する中国語文字「机」の2バイト文字コードが得られるように配列してある。

4.「ＡＴＲ研究発表会 2003」システム

「ＡＴＲ研究発表会 2003」の音声翻訳システム総合展示において、日中の双方向翻訳を構成したシステム(以下、言語翻訳部)について述べる。

4.1.概要

中日言語翻訳部は、

(1) 中日言語変換部
(2) 中日文字変換部
(3) 日本語自動換言部

からなり、日中言語翻訳部は、

(1) 日本語入力自動換言部
(2) 日中言語変換部
(3) 日中文字変換部
(4) 中国語自動換言部

からなる。

研究開発段階では、各部は独立したプロセスであり、それぞれ標準入力から入力し標準出力へ出力するフィルタとして構成してあったが、この構成ではフィルタ間のデータ受け渡しがオペレーティングシステムによりバファリングされ、何文か分のデータがまとめて処理されるため、1 文ずつの最終出力まで処理するのには不都合であった。そこで、音声翻訳システムとして統合するため、各部を Perl 言語によるモジュールとし、言語翻訳部で単一のプロセスとなるように改変した。

4.2.中日言語翻訳部

4.2.1.中日言語変換部

あらかじめ日中対訳コーパスと日中対訳辞書とから自動的に作成した中日翻訳パターンを、Perl 言語によるモジュールに改変して構成されており、中国語音声認識部の出力である中国語文を日本語文に変換する。日中対訳辞書で一つの中国語に複数の日本語が対応付けられていた場合は、その中国語の日本語訳に複数の日本語を羅列して出力する。言語変換部はソフトウェア自体が、日中対訳コーパスと日中対訳辞書とから自動的に変換されており、中国語音声認識部の出力である中国語文を入力するだけで、他のデータを何も参照しない。14 万弱の変換操作の集合体であり、Pentium4 2.8GHz の CPU を搭載したマシンでは 1 文に対して約 1/7 秒ですべての変換操作の適用を試み、適用可能な場合は変換操作を行なった。対訳コーパスには BTEC1 の他に Study、Business の SLT コーパス第一期分の

日中対訳約 13 万対を、対訳辞書には一般的な用語約 4 万 5 千対を利用した。

4.2.2.中日文字変換部

中日言語変換部の出力に翻訳されずに残った中国語漢字を、日本語漢字に変換する。中日対照の異体字表を参照するため、中国語漢字 6768 字のうち 4886 字を日本語漢字に変換できる。

4.2.3.日本語自動換言部

中日言語変換部の出力に残った曖昧さ（例えば中国語「房間」に対する日本語「部屋」「室」「ルーム」等）を解消して、訳語の羅列部分を前後から最適とみなせる一語に変換する。そのため日本語コーパスからあらかじめ 2-gram、3-gram の出現度数を得ておき、各訳語にスコア付けを行うことで最適の訳語を選択する。
さらに、あらかじめ日本語コーパスから自動的に作成した日本語表現データベースを参照して、中日言語変換部の出力を「尤もらしい」日本語表現に自動的に変換する。日本語コーパスには、BTEC1 の他に SLDB、LDB の対話収録部分の合せて約 18 万文と、それらを待遇表現等について換言した約 28 万文との合計約 40 万文を利用した。

4.3.日中言語翻訳部

4.3.1.日本語入力自動換言部

あらかじめ日中対訳コーパスの日本語テキストから自動的に作成した日本語表現データベースを参照して、日本語音声認識部の出力である日本語文を「尤もらしい」日本語表現に自動的に変換する。参照する日本語表現を日中対訳コーパスの日本語テキストに限定しているため、この対訳コーパスを利用する日中翻訳にとって「翻訳しやすい」日本語に言換えていることになる。日本語コーパスには、BTEC1 の他に Study、Business の SLT コーパス第一期分の日本語テキスト約 18 万文を利用した。

4.3.2.日中言語変換部

あらかじめ日中対訳コーパスと日中対訳辞書とから自動的に作成した日中翻訳パターンを、Perl 言語によるモジュールに改変して構成されており、日本語音声認識部の出力である日本語文を中国語文に変換する。日中対訳辞書で一つの日本語に複数の中国語が対応付けられていた場合は、その日本語の中国語訳に複数の中国語を羅列して出力する。言語変換部はソフトウェア自体が、日中対訳コーパスと日中対訳辞書とから自動的に変換されており、日本語音声認識部の出力である日本語文を入力するだけで、他のデータを何も参照しない。15 万強の変換操作の集合体であり、Pentium4 2.8GHz の CPU を搭載したマシンでは 1 文に対して約 1/6 秒ですべての変換操作の適用を試み、適用可能な場合は変換操作を

行なった。対訳コーパスにはBTEC1の他にStudy、BusinessのSLTコーパス第一期分の日中対訳約13万対を、対訳辞書には一般的な用語約4万5千対を利用した。

4.3.3.日中文字変換部

日中言語変換部の出力に翻訳されずに残った日本語漢字を、中国語漢字に変換する。日中対照の異体字表を参照するため、日本語漢字6398字のうち5701字を中国語漢字に変換できる。

4.3.4.中国語自動換言部

日中言語変換部の出力に残った曖昧さ（例えば日本語「バス」に対する中国語「西式浴室」「浴室」と「公共汽車」「巴士」等）を解消して、訳語の羅列部分を前後から最適とみなせる一語に変換する。そのためコーパスBTEC1の中国語テキスト約10万文からあらかじめ2-gram、3-gramの出現度数を得ておき、各訳語にスコア付けを行うことで最適の訳語を選択した。

4.4.評価

本システムは、2004年度に予定している実証的な評価のためのテストベッドに向けた予備結合という意味合いが強く、音声翻訳システムとして音声認識部や音声合成部と統合された点に意義がある。そのため、定量的な性能評価は実施していないが、参考のため本システムの特徴による効果等について述べる。

4.4.1.翻訳速度

本システムでは、言語翻訳部での処理時間を1文当り1秒未満とする前提で、利用するコーパス等の選択を行なった。実際には、Pentium4 2.8GHzのCPUを搭載したマシンで日中、中日とも概ね1/2秒程度で翻訳を実行した。日本語自動換言部では、日中翻訳での日本語入力換言がコーパスを限定して参照する日本語表現データを抑えたためやや高速と感得されるものの、文の違いによる変動が大きかった。
言語変換部は、どの入力文であっても、すべての変換操作の適用を1回ずつ試みるため比較的安定して中日1/7秒程度、日中1/6秒程度の処理時間である。その他の文字変換部と、訳語の曖昧さ解消だけの中国語自動換言部は、言語変換部と比べても短い処理時間であり、言語翻訳としてみれば無視できる程度である。

4.4.2.日本語自動換言の効果

中国語音声認識では、見せるという意味の「看看」kan1kan1が、話をするという意味の「談談」tan1tan1に認識される誤りが比較的目立った。が、運転免許証という意味の目的語が日本語コーパスでは、見せるという意味の動詞と一緒にしか用いられていないため、

認識結果「談談」によって翻訳された中日言語変換部の出力「運転免許証と話をさせて」を、「運転免許証を拝見します」のように換言していた。このため、最終的な翻訳出力は「看看」と正しく認識されていれば「運転免許証を見せて頂けますか」、「談談」と誤認識していても「運転免許証を拝見します」という程度の差でしかなく、認識結果の文字列を注視していなければ誤認識に全く気付かないという、日本語自動換言の狙い通りの効果が見られた。

4.4.3. 日本語入力自動換言部

日本語音声認識では、総合展示の実演中には特に目立った誤りがなかった。難点としては話者が発話中にわずかなポーズをおいても、そこで発話が区切られてしまうことがあり、この場合も、「運転免許証を。」のようなものは日本語コーパスで見せるという意味の動詞と一緒にしか用いられていないため、見せるという意味の述部を補完するように換言して日中言語変換部に出力していた。中国語訳にも当然見せるという意味が訳出されこの部分だけを見れば効果的であったが、こうした場合は、次に区切られた残りの「見せて頂けますか。」という部分も翻訳されてしまうので、発話全体の訳文2文を通して見ると、やはりかなり不自然なものになってしまっていた。

5.翻訳準備の詳細

「3.トランスファー（言語変換）部の構成」節で説明したソフトウェア資源をどのように利用していくかについて説明する。

5.1.翻訳パターン生成

翻訳パターンは、「3.2.2.日中対訳コーパス」項の日本語文を、「3.2.1.日本語形態素解析器」項の形態素解析器により、解析、分割された日本語形態素の列と、「3.2.2.日中対訳コーパス」項の中国語文とを対照させて生成する。

5.1.1.日本語形態素複合語処理

当初、日本語形態素解析器にはTDMT(Transfer-Driven Machine Translation) 用のものを利用してきたが、解析速度、利便性の点からChaSen ver2.2.8を利用することにした。TDMT 用の日本語形態素解析器は、長期にわたり音声対話翻訳の研究に利用されてきており、多くの複合語を単語扱いで辞書に登録している。他の形態素解析器に変更する際に、こうした複合語の登録が容易であるかも重視した。ChaSenにはユーザ辞書の機能があり、TDMT 用の日本語形態素解析器の辞書から複合語を抽出して、この機能によりそれらの複合語をChaSenに登録することで、ほぼ同等の解析結果が得られた。

しかし、形態素解析器の辞書に複合語を単語扱いで登録してしまうと、例えば「計画に手を入れる（修正する等の意味）」という表現を考慮して「手を入れる」という複合語を単語扱いで登録してしまうと、「ポケットに手を入れる」という表現の解析に不都合である、等といった問題があった。機械翻訳システムでは、通常、こうした問題を構文解析の一部として、「計画」「ポケット」等の語の意味分類を検査して複合語にまとめるか否かを判定する等している。

本処理は、日本語と中国語の対応をとるためのものであることから、あらかじめ日中対訳辞書の日本語見出しをChaSenで解析して細分割される見出しを中国語訳語とChaSenの解析結果とを併せて複合語として別の辞書に登録しておき、ChaSenによる日本語文の形態素解析結果と複合語を登録した辞書とを照合していき、複合語として日本語が照合でき、かつ一緒に登録されている中国語訳語が中国語文に含まれている場合にのみ、当該部分を複合語として単語扱いにまとめてしまうこととした。

例えば「手を入れる」という複合語の場合、ChaSenで解析して「手」「を」「入れる」という解析結果と、中国語訳語として「修正」を意味する語や「改善」を意味する語等とが併せて登録されているものとする。対訳コーパスの日本語文の形態素解析結果が「手」「を」「入れる」と一致する部分を含んでいる場合、対訳の中国語文を走査して「修正」を意味する中国語や「改善」を意味する中国語を含んでいるか検査し、含んでいる場合にの

み「手」「を」「入れる」と分割された形態素解析結果を「手を入れる」という複合語で置換する。「修正」や「改善」を意味する中国語が「手」「を」「入れる」以外の部分の訳語であった場合には、誤った複合語扱いをしてしまう可能性は残るが、ユーザ辞書登録してしまうよりは、はるかに安全であると考えられる。

5.1.2.日本語形態素解析後処理

「5.1.1.日本語形態素複合語処理」項で述べたように、当初、形態素解析器には TDMT 用のものを利用してきたが、解析速度、利便性の点から ChaSen を利用することにした。どちらも文字コードは EUC である。

そこで、 ChaSen の解析結果に対する「5.1.1.日本語形態素複合語処理」項の複合語処理が済んだ後で、解析結果の形式を TDMT 用のものに合わせて変換する。これにより、形態素解析器は TDMT 用、 ChaSen のいずれを用いていようと、以降の処理は共通である。さらに、以降の処理で行単位の処理ができるように、形態素毎に 1 行である形式を日本語文毎に 1 行である形式に変換する。

次に、後処理として、日本語漢数字（「一」「二」「三」…「十」「百」「千」等）の品詞を “num” とし、対応する中国語漢数字を情報として持たせておいている。また、当初「エイ」「ビー」等と表記されコーパスに現れないはずであった、英数字「A」「B」や「1」「2」がコーパスに現れるようになったので、これらも品詞を “num” とし、中国語文字（英数字は日本語と中国語で共通なので同じ文字コード）を情報として持たせておいている。

5.1.3.汎化語句の候補列挙

後処理の済んだ日本語形態素の列と中国語文とを、「3.2.3.日中対訳辞書」項の対訳辞書を参照しながら日本語形態素の訳語が中国語文に含まれているか検査し、訳語が含まれていれば、その日本語形態素と中国語文の部分文字列と対にし、汎化語句の候補として列挙していく。

5.1.4.汎化語句の特定

汎化語句の候補のリストを検査して、一対一に対応していないものについては、一対一に対応するように特定する。一対一に対応していないとは、一対多、多対一、多対多に対応している場合であって、例えば日本語が「XX のレンタルは YY でレンタル費が必要です」に対して中国語が「XX の賃貸は YY で費用が必要だ」という構造である場合、日本語の「レンタル」が 2 語とも汎化語句の候補になっている。また、中国語側は形態素解析しないため、日本語の汎化語句の候補「AB」「BC」に対する、中国語の汎化語句の候補「ab」「bc」が中国語文では「…abc…」のように重複して出現している場合もあり得る。

こうした対応の曖昧さに対して、例えば「XX のレンタルは YY でレンタル費が必要です」という例では、前の「レンタル」の前後に位置する汎化語句の候補（ここでは「XX」「YY」

が汎化語句の候補だとする）の対応する中国語文でのそれぞれの位置（先頭からの文字数）の平均値を算出して推定位置とし、後ろの「レンタル」も同様に（ここでは「YY」「費」が汎化語句の候補だとする）推定位置を求める。そして中国語文での「賃貸」にあたる語の実際の位置に、最も近い推定位置を得られた「レンタル」を、「賃貸」にあたる語と一対一に対応させ、他の候補をリストから無効にすることで、汎化語句を特定する。

例とは逆に、日本語が一つに対して中国語が複数候補として対応している場合も、同様にその日本語形態素の中国語での推定位置を求め、最も近い位置にある中国語候補に特定する。

汎化語句のリストとして出力する前に、「3.2.4.汎化品詞表」項の品詞表と、設定されていれば「3.2.5.単語頻度表」項の頻度表を参照して、品詞表に記述された品詞の汎化語句については、単語頻度表から得られる出現頻度が、記述された出現頻度の閾値以下の場合にのみ有効とし、そうでない場合はリストとして出力しない。

5.1.5.整形処理

日本語文、中国語文のそれぞれの文中の汎化語句を、”<x>” や “<y>” 等のタグに置換して、翻訳パターンの形式に変換する。パターンの左辺に原言語、右辺に目的言語を配置する。

5.1.6.可変部連続の解消

”<x>” や “<y>” 等のタグで表記される可変部が連続して現れていた場合、分割してマッチングさせることは非常な処理負荷に対して得るところが小さいので、連続するタグはどちらか一方を残して他を消去する。

例えば「我想<x><y> <--> <y>を<x>たいのですが」というパターンに「我想預訂房間」を正しくマッチングさせるのは実行時の処理負荷が大きくなる。そこで、「我想<x> <--> <x>たいのですが」という書き換えを行なって可変部連続を解消しておく。これによって翻訳できなくなるわけではなく、「預訂<x> <--> <x>を予約する」というパターンがあれば、「我想預訂房間」は「「房間を予約する」たいのですが」となり、ほぼ問題のない訳文ができるのである。

5.2.ストリームエディタ用スクリプト化

評価した時点では、ストリームエディタとしては、機能、速度ともにPerlが優れていた。評価のため他のスクリプトにするのに、 Perlのスクリプトから通常のエディタで変換した。Pentium II 450MHz CPUを搭載したパソコンで、12,000の翻訳パターン（対訳辞書から生成したパターン6,200を含む）を中国語文1,000文に適用した場合の処理時間を以下に示す。

```
Perl    42.830u 0.080s   (31.5 文/秒)
Gawk   112.080u 0.900s   (12.0 文/秒)
Ruby   179.850u 0.340s    (7.5 文/秒)
Tcl    774.040u 0.930s    (1.7 文/秒)
```

上記の他に Python でも簡単に作動させられるが、 Tcl と比較してすら何倍も遅い。ストリームエディタの先駆である SED は指定可能な変換操作に制限がある（初期の SED は 400 操作）ためか、評価時点では正しく変換されなかった。

6.発明提案

上記の研究開発の成果として、以下の３件の発明を提案する。

(1) 翻訳辞書への未登録語自動登録装置

(2) 不等価な対訳辞書登録を許容する機械翻訳装置

(3) 機械翻訳向け換言装置

6.1.翻訳辞書への未登録語自動登録装置

6.1.1.発明の名称

翻訳辞書への未登録語自動登録装置

6.1.2.従来技術

(2-1)従来技術文献

(文献 1)

【公開番号】特開平４－２５６１７１

【公開日】平成４年（１９９２）９月１０日

【出願番号】特願平３－１７２８７

【出願日】平成３年（１９９１）２月８日

【発明の名称】未登録語処理方式

【出願人】富士通株式会社

【発明者】徐 国偉

(文献 2)

【特許番号】第２９９５７８３号

【発行日】平成１１年（１９９９）１２月２７日

【出願番号】特願平２－４１８７７

【出願日】平成２年（１９９０）２月２１日

【発明の名称】カタカナ語の訳語推定装置

【特許権者】日本電気株式会社

【発明者】田邊 裕子

(文献 3)

【公開番号】特開２００３－６１９３

【公開日】平成１５年１月１０日（２００３．１．１０）

【出願番号】特願２００１－１８５８００（Ｐ２００１－１８５８００）

【出願日】平成１３年６月２０日（２００１．６．２０）

【発明の名称】機械翻訳装置および方法
【出願人】株式会社エイ・ティ・アール音声言語通信研究所
【発明者】隅田 英一郎

(2-2)従来技術文献の要旨および問題点

従来の未登録語の訳語推定、自動登録など（以下、自動登録という）は日本語と中国語間の固有名詞（文献 1）や、カタカナ語と英語原語（文献 2）のように対象を限定することにより、翻訳辞書に原語と訳語の対として未登録であるものの一部について登録を自動化するものである。

文献 1 は日本語形態素辞書には日本語側の語が登録されているが日中翻訳辞書には未登録の場合に、文献 2 は英語辞書には英語側の語が登録されているが日英翻訳辞書には未登録の場合に、作動する構成である。

しかし、技術の進歩等により新造語が必要になるのはもちろん、必要でなくとも「新鮮み」を打ち出そう等として次々に新しい言葉が造られていく傾向は、テレビや携帯電話等の情報伝達メディアの一般化、日常化とともに、強まる一方である。

そのため、新造語は意図的に普通の辞書に登録されるような語を避けて、案出、合成、省略、転用されてきており、いわゆる和製英語や多言語からの合成等による新造語を濫用とみる向きもあるものの、これら新造語を避けて現実の会話や文章を成立させることは非常に困難である。

一方で技術進歩は、こうした言語を処理して人間の活動を支援しようとする、いわゆる自然言語処理技術にも大きく影響している。例えば、処理に必要な辞書を装置として構成するための記憶装置は、日本で機械翻訳が商品化され始めた 1980 年代半ばと比較すれば、容量比で十万分の一以下の売価となっている。つまり、当時、記憶装置の百万円相当の容量を用いて構成されていた辞書であれば、現在では記憶装置の 10 円相当の容量を用いて構成可能である。

こうしたことは、辞書を構成する記憶装置の容量を抑えるため、辞書に登録する語を限定するといった、当時の「常識」を完全に陳腐化させており、日本語形態素辞書には登録されているが日中翻訳辞書には未登録である、あるいは英語辞書には登録されているが日英翻訳辞書には未登録である等は、それによって得るコスト低減効果が利用者の受ける不便さに全く見合わないものになってしまっている。

このため、現在のような状況下では、従来のように、完全な未登録語ではなく一部の辞書にだけ未登録といった、登録語の不揃いを解消するような未登録語の自動登録では、機械翻訳等の自然言語処理の有用性に対してはごく限定的な効果しか持ち得ないという問題があった。

6.1.3.発明の要旨

上記の問題点を解決するために、本発明では処理に利用するすべての辞書に未登録であ

る完全な未登録語を、すでに翻訳された対訳文から自動的に、語句として切り出し、訳語および意味的情報等を付加して翻訳辞書に登録する機構を設けた。

6.1.4.発明の効果

本発明の機構により、本装置を組み込んだ自動翻訳システムの操作者にとって、未登録語のピックアップ、訳語の決定、意味分類の付与といった一連の辞書登録作業が自動化され、作業効率が改善される。

従来であれば、例えば、リライトの済んだ訳文と原文とを突き合わせながら、リライトされた部分の原文を突きとめて、辞書に登録可能な語句の形にし、訳語、品詞、意味分類等の情報を付与して、やっと辞書登録の操作が可能であった。このような作業をきちんと続けていかなければ、自動翻訳システムはまた同様の部分で翻訳に失敗して、再度リライトが必要ということになる。そのため、翻訳作業自体が自動化されても、リライトや辞書登録作業を含めた作業全体では、それほど大きな効率改善効果が見られない場合があった。

本発明により、辞書登録作業が大幅に効率化されるので、自動翻訳システムを用いた翻訳作業全体に対しても、はっきりとした効率改善効果をもたらすことができる。

6.1.5.実施例

説明のため、日英翻訳を例にとる。原言語は日本語、目的言語は英語である。

書換えパターン抽出部は、文献 3 に示すような書換えパターンを、本発明を実施する翻訳システムから抽出する。

書換えパターンは少なくとも、原文パターンと訳文パターンとの２つのパターンの対からなり、それぞれのパターンは少なくとも、文字列の情報を持つ固定部と、固定部との位置情報および原文パターンの可変部と訳文パターンの可変部との対応情報の２情報を持つ可変部とからなる。

例えば「私は…に～を送ったＩ sent ～ to … 」という文字列であっても、「私は…に～を送った」という原文パターンと、「 I sent ～ to … 」という訳文パターンとからなっている。原文パターンには「私は」「に」「を送った」という文字列の情報を持つ固定部と、「…」「～」が挿入されていることで固定部との位置情報を持ち、かつ訳文パターンにも同じ「…」「～」が挿入されていることで訳文パターンの可変部との対応情報を持つ可変部とからなっている。訳文パターンも原文パターンと同様であり、書換えパターンになっている。

あるいは、構造化して、例えば

原文固定部数：3
原文固定部１：私は　　原文固定部２：に　　原文固定部３：を送った
訳文固定部数：2
訳文固定部１：I sent　　訳文固定部２：to

可変部数：2
可変部1位置：原文固定部1原文固定部2間　訳文固定部1訳文固定部2間
可変部2位置：原文固定部2原文固定部3間　訳文固定部2訳文末尾間

のように構成しても、同様に書換えパターンになっている。

抽出された書換えパターンは、パターン記憶部に登録され保持される。

ここでは説明のため、「最近～という新技術が話題になっている」という原文パターンを持つ書換えパターンが抽出されたものとする。

対訳用例文照合部は、書換えパターンを順次、パターン記憶部から入力し、対訳用例文群それぞれの原文に原文パターン、訳文に訳文パターンを照合するか検査して、原文パターンと訳文パターンのそれぞれの固定部が共に照合する対訳用例文を探す。固定部が照合するとは、原文パターンの固定部の文字列と同じ文字列が、原文中に固定部の順序でかつ重なりがなく存在することであり、訳文パターンにおいても同様である。
対訳用例文照合部は、原文パターンと訳文パターンが共に照合する対訳用例文において、原文パターンと訳文パターンのそれぞれの可変部を特定する。可変部を特定するとは、原文パターンの可変部の位置情報に合致する範囲の、原文の部分文字列を求めることであり、訳文パターンにおいても同様である。

ここでは、「最近ユビキタスという新技術が話題になっている」という原文の対訳用例文の原文、訳文に照合し、「ユビキタス」を原文の可変部として特定したものとする。同様に「ubiquitous computing」を訳文の可変部として特定したものとする。また、同様に「最近インターネットという新技術が話題になっている」という原文の対訳用例文にも照合し、「インターネット」を原文の可変部として特定したものとする。

当該書換えパターンは、特定された可変部と組にして可変部解析部に送られる。

可変部解析部は、特定された可変部と書換えパターンを入力すると、当該書換えパターンの可変部情報をパターン記憶部から入力し、特定された可変部それぞれについて、以下の処理を行なう。特定された可変部が既に可変部情報として登録された語句であれば、当該可変部情報の内の出現数の情報を 1 増加させる。可変部情報として登録された語句でなければ、登録語辞書検索部に特定された可変部を出力して、検索した結果が戻るのに待機する。

ここでは、「ユビキタス」は可変部情報として登録された語句でなかったものとする。

登録語辞書検索部は、本装置を組み込んだ自動翻訳システムで管理されるすべての辞書データベースを検索し、入力した可変部の全部または一部を見出しとする登録語の情報を得て、可変部解析部に出力する。

ここでは、「ユビキタス」では登録語の情報を何も得られず、「インターネット」では辞書が登録されており品詞分類「普通名詞」や意味分類「技術用語」等の情報が得られたものとする。

可変部解析部は、登録語辞書検索部から検索した結果が戻ると、当該書換えパターンの可変部情報に、特定された可変部と登録語辞書検索部から戻った検索した結果とを追加する。登録語辞書検索部から検索した結果が未登録であることを示していれば、特定された可変部と未登録であることを示す情報とを追加する。特定された可変部すべてについて、処理が終了すると、当該書換えパターンの可変部情報をパターン記憶部に出力して、パターン記憶部に登録され保持される情報を更新する。そして、次の書換えパターンの入力に待機する。

未登録語登録情報生成部は、書換えパターンがすべて照合するかの検査を終え、照合した場合の処理を終えるのに待機しておく。前記処理が終了した後で、パターン記憶部を走査し、未登録であることを示す情報を可変部情報に付加された書換えパターンを順次、パターン記憶部から入力する。

ここでは、１つの可変部を含む「最近～という新技術が話題になっている」という原文パターンを持つ書換えパターンが入力され、その原文パターンの可変部情報には「ユビキタス」が未登録であることを示す情報が付加されており、また「インターネット」が品詞分類や意味分類等の情報を付加されていたものとする。

未登録語登録情報生成部は、入力した書換えパターンの可変部情報を利用して、未登録であることを示す情報を付加された可変部の訳語や品詞分類や意味分類等の辞書に登録する情報を推定する。訳語は、入力した書換えパターンの訳文パターンの可変部により決定する。品詞分類や意味分類等の情報は、それらの情報を追加された他の可変部情報の集合の和をとることで推定する。可変部情報が多い場合には、可変部情報の内の出現数がある閾値以上のものだけの和をとる等する。入力した書換えパターンに複数の可変部が存在する場合は、前記の処理を可変部毎に行なう。

ここでは、「ユビキタス」の訳語は「ubiquitous computing」とする。品詞分類や意味分類等の情報は、「インターネット」の品詞分類「普通名詞」や意味分類「技術用語」と推定するものとする。

未登録語登録情報生成部は、当該書換えパターンにおいて推定した情報を、すでに推定した情報と集合の和をとるようにして追加して保持する。また、可変部情報が多い場合には、可変部情報の内の出現数がある閾値以上のものだけの和をとる等して、保持する情報を限定する。以上の処理を、ここでは「ユビキタス」を可変部情報として含むすべての書換えパターンについて行なう。

未登録語登録情報生成部は、すべての書換えパターンを処理し終えると、未登録であることを示す情報が付加された語毎に保持された情報をリストにし、未登録語登録情報として、辞書登録部に出力する。

辞書登録部は、未登録語登録情報生成部から未登録語登録情報のリストを入力すると、本装置を組み込んだ自動翻訳システムで管理される辞書データベースに登録する。例えば、登録の際に、本装置を組み込んだ自動翻訳システムの操作者に未登録語登録情報のリストを提示して、確認を得た未登録語登録情報のみを辞書データベースに登録する等、登録の具体的手順はあらかじめ定めておき実行する。

以上、説明したように、本装置は機械翻訳に用いる書換えパターンを用いて、対訳用例文群から、未登録語の訳語や品詞分類や意味分類等の情報を自動的に抽出する。

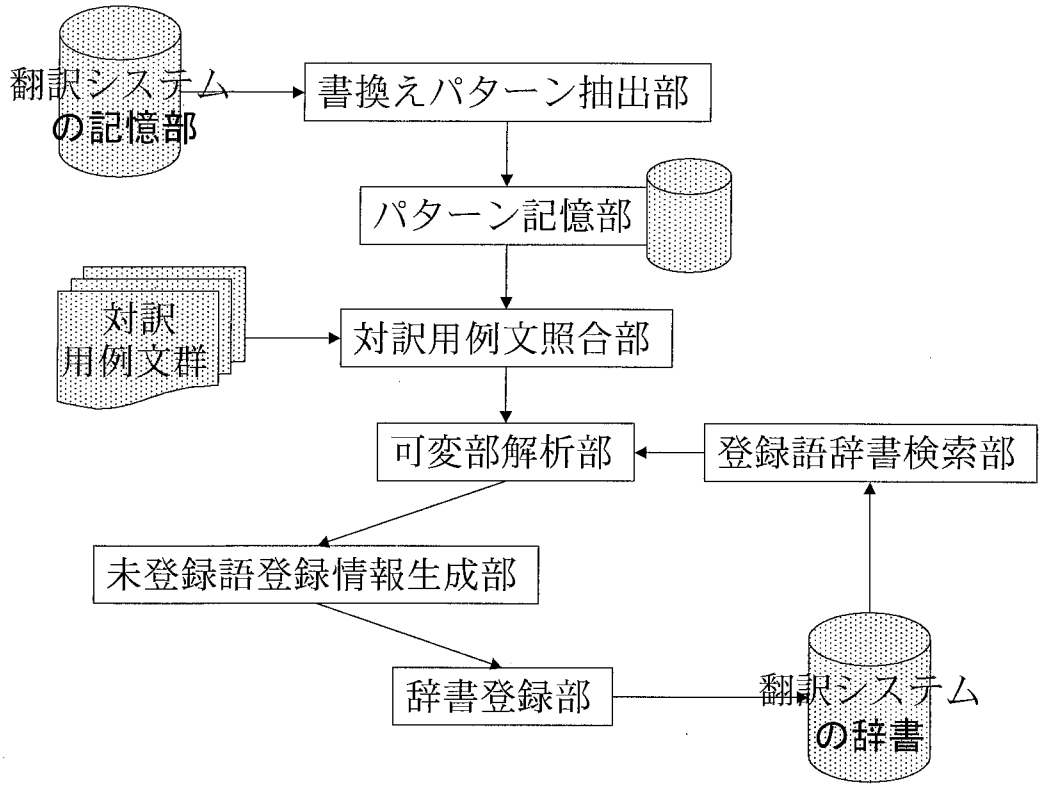
翻訳システム
の記憶部
書換えパターン抽出部
パターン記憶部
対訳
用例文群
対訳用例文照合部
可変部解析部
登録語辞書検索部
未登録語登録情報生成部
辞書登録部
翻訳システム
の辞書

図６．１－１ 装置構成

6.2.不等価な対訳辞書登録を許容する機械翻訳装置

6.2.1.発明の名称

不等価な対訳辞書登録を許容する機械翻訳装置

6.2.2.従来技術

(2-1)従来技術文献

【特許番号】第２６０９１７３号
【発行日】平成９年（１９９７）５月１４日
【出願番号】特願平２－７７４７１
【出願日】平成２年（１９９０）３月２６日
【発明の名称】用例主導型機械翻訳方法
【特許権者】株式会社エイ・ティ・アール自動翻訳電話研究所
【発明者】隅田 英一郎
【発明者】飯田 仁
【発明者】幸山 秀雄

【公開番号】特開２００３－６１９３
【公開日】平成１５年１月１０日（２００３．１．１０）
【出願番号】特願２００１－１８５８００（Ｐ２００１－１８５８００）
【出願日】平成１３年６月２０日（２００１．６．２０）
【発明の名称】機械翻訳装置および方法
【出願人】株式会社エイ・ティ・アール音声言語通信研究所
【発明者】隅田 英一郎

(2-2)従来技術文献の要旨および問題点

従来、原言語による文字列を入力とし、機械処理によって、目的言語による文字列に翻訳して出力する機械翻訳の方式として、原言語文字列の形態素解析・構文解析・意味解析など、原言語から目的言語への構文変換・意味変換など、目的言語文字列への構文生成・形態素生成などの様々な処理の仕方を定めた翻訳規則を蓄積しておき、それらの翻訳規則に基づいて機械処理を行う、規則ベース翻訳方式が知られている。

ただ、翻訳規則は、適用可能範囲が翻訳規則の適用条件によって限定されているため、入力される表現に網羅的に対応するには、入力される表現を網羅する膨大な翻訳規則が必要であった。

そこで、機械翻訳の方式として、原言語表現と、あらかじめ人手などによって翻訳された当該原言語表現の対訳となる目的言語表現と、の対を用例として蓄積し、入力に完全に

一致する用例が存在する場合には、対応する訳を出力し、入力に完全一致する用例が存在しない場合でも、入力に類似する用例を特定し、その特定された用例を参照して、機械処理を行う、用例ベース翻訳方式が提案されている。

しかしながら、規則ベース翻訳方式も用例ベース翻訳方式も、文の構造や「言い回し」等と呼ばれる、複数の語句が関連する部分を対象とする翻訳の方式であり、対象となる部分の構成要素である、個々の語句に関しては原言語の見出しと目的言語での訳語が、意味的にも構文上の機能面においても同等であることが前提であった。これは比較的きちんとした文章においてはあまり問題にならないと考えられていたからであるが、人間の音声を認識して翻訳する音声翻訳、特に特定の人との対話翻訳においては、しばしば問題となる。

例えば、「スーパー」は音声翻訳に限らず新聞記事等でも「スーパーマーケット」の意味で定着してしまっているので、日本語の「スーパー」を正しく翻訳するには「super market」を訳語とせざるを得ないが、そのままでは「スーパーマーケット」ときちんと書くなり話すなりした場合には「super market market」と訳出されてしまうことになる。そこで、超大型タンカーを意味する「スーパータンカー」等、本来、造語成分である「スーパー」を含む語は全体を単語として辞書に登録していなければ誤訳されてしまうという状況になってしまっている。

あるいは「三役」は相撲に関してだけでなく「○○党三役」といった使い方も定着している。仮に「三役」の訳語を「○○党三役」に合わせて辞書に登録した場合は、「三役力士」等のそうでない意味の複合語は、全体を単語として辞書に登録していなければ誤訳されてしまうことになる。

そのため、不用意な辞書登録によって、それまで正しく翻訳されていた多くの語句が誤訳されるという事態を招きかねず、辞書の登録には、作業者に大きな配慮が求められ、作業者にとって大きな負担となるという問題があった。あるいは翻訳システムの辞書登録機能に制限をかけ、固有名詞等の比較的安全な語句のみの登録しか認めなくする等の制限をかけなければならないという問題があった。

6.2.3.発明の要旨

上記の問題点を解決するために、本発明では、語句の意味の範囲や構文上の機能等において同等の範囲や機能を有する訳語を登録していなくとも、訳語を適用する段階で訳語の前後関係から適切な形態に自動的に変形、書換えを行ない、あたかもその訳文に合わせて辞書登録がなされていたかのように動作する機構を設けた。

6.2.4.発明の効果

本発明の機構により、本発明を実施する機械翻訳システムの操作者にとって、原言語と目的言語との対応がとり難い語句でも簡単に辞書に登録しておけるので、辞書登録の作業

負担が軽減する。また、不等価な辞書登録がもたらす重複訳出等の問題が解消し、機械翻訳システムの訳出文がより読み易いものになる。そのため、機械翻訳システムを用いた翻訳作業全体に、効率改善効果をもたらすことができる。

6.2.5.実施例

説明のため、日英翻訳を例にとる。原言語は日本語、目的言語は英語である。

辞書登録部は、本発明を実施する機械翻訳システムの操作者が、原言語の語句を目的言語の語句に変換するための辞書を、当該機械翻訳システムに登録する手段を提供する。その際に、既存の辞書登録手段に加えて、当該操作者が登録しようとする原言語の語句と目的言語の語句とが、構文上や意味上での語句の働きに違いを持つものである場合に、目的言語の語句に不等価を示す標識を付与するように構成する。

ここでは、原言語の語句「ケーキ入刀」に、目的言語の語句「wedding cake」「knife」を羅列して登録し、不等価を示す標識を付与したものとする。

言語変換処理部は、従来知られている方法で原言語を目的言語に変換していく。

ここでは、翻訳の原文として「ウェディングケーキへのケーキ入刀はどうやればいいか教えてください」を入力したものとする。

言語変換処理部は、原言語を目的言語に変換する際に、不等価を示す標識付の目的言語の語句を辞書検索部から入力すると、当該語句を一時記憶に保持して、処理を続行する。当該語句を処理しなければそれ以上処理を進められなくなると、当該語句を関連語拡張処理部へ出力して、拡張処理した結果が返るのに待機する。

ここでは、原文の「ウェディングケーキ」「どうやればいいか教えてください」の部分はそれぞれ「wedding cake」「please show me how to」と部分的に翻訳処理を進めたものとする。

関連語拡張処理部は、語句を入力すると、あらかじめ目的言語のシソーラスや大規模コーパス等の言語資源から作成しておいた関連語データベースを検索し、入力した語句全体との関連性の高い単語を得て、それらの単語と入力した語句を構成する単語とをリストにする。当該リストには、入力した語句全体との関連性を示す数値を、各単語に与える。また、リストの単語数は以降の処理に支障を来たさないように、あらかじめ定めた方法・基準により限定しておく。当該リストを、言語変換処理部へ処理結果として返す。

ここでは、「wedding cake」「knife」に加え「bridal」「cut」「server」「serve」「chocolate」「candle」等の単語を含むリストが返されたものとする。入力した語句全体との関連性を示す数値は、「bridal」が最も高く、以下「cut」「server」その他はリストの順に高いものとする。

言語変換処理部は、既に部分的に翻訳処理を進めた部分のうち、当該語句の近傍の語句に当該語句との近さの情報を付与して近傍語句リストとし、この近傍語句リストと、関連語拡張処理部から返されたリストとを局所生成部に出力して、局所生成部からの処理結果が返るのに待機する。

ここでは、近傍語句リストとして「wedding cake」「please show me how to」の2つの語句からなるリストが、当該語句の拡張処理結果として「wedding cake」「knife」「bridal」「cut」「server」「serve」「chocolate」「candle」等の単語を含むリストが、それぞれ局所生成部に出力されたものとする。

局所生成部は、目的言語の大規模な用例データベースを備え、入力したリストに含まれる語句を多く含む用例文を、入力した語句群に意味的に類似した文であるとみなして用例データベースから検索する（以下、類似文検索という）。この際に、入力した近傍語句リストに付与された、当該語句との近さの情報を類似文検索における重み付けに用い、当該語句とより近い語句を含む類似文をより類似したものであるとみなすようにして、類似文を検索する。また、入力した当該語句の拡張処理結果に与えられた関連性を示す数値を類似文検索における重み付けに用い、関連性のより高い単語を含む類似文をより類似したものであるとみなすようにして、類似文を検索する。辞書登録部により目的言語の語句として登録されていた語句と、近傍語句と、拡張処理結果のリストに含まれる単語と、の3種の語句間でも重み付けをし、拡張処理結果のリストに含まれる単語を含む類似文を比較的類似性が低いものであるとみなすようにして、類似文を検索する。

ここでは、近傍語句「wedding cake」「please show me how to」はどちらも直近（直前または直後）であるため、これらを含む類似文は非常に類似したものであるとみなされる。目的言語の語句として登録されていた「wedding cake」「knife」を含む類似文は非常に類似したものであるとみなされる。拡張処理結果のリストに含まれる単語のうち、入力した語句全体との関連性を示す数値の高い「bridal」「cut」を含む類似文は比較的類似したものであるとみなされる。ここでは、以上から、最も類似した文として「please show me how to cut a wedding cake」が得られたものとする。

局所生成部は、用例データベースを検索して得られた類似文から、近傍語句からの近さが一定範囲内にある、目的言語における自立語、内容語（付属語や文法上の機能語ではない語）を走査して、拡張処理結果のリストに含まれる単語のうち、当該語句の原言語の語句の訳語となっているものを特定する。さらに、前記原言語の語句の訳語となっているものの近傍の付属語や文法上の機能語を含めて、当該語句の原言語の語句の訳語を生成し、言語変換処理部に返す。

ここでは、原言語の語句「ケーキ入刀」に、目的言語の語句「cut」が訳語として生成されたものとする。

言語変換処理部は、局所生成部から処理結果として、当該語句の原言語の語句の訳語を入力すると、一時記憶に保持しておいた当該語句の訳語に関する情報を、局所生成部から入力した訳語の情報で置換し不等価を示す標識を消去して、当該語句を含めて処理を続行する。

ここでは、原言語の語句「ケーキ入刀」に、目的言語の語句「cut」が辞書登録の段階で不等価ではない訳語として与えられていた場合と同様の状態となって、処理が続行されることになる。

以上、説明したように、本発明を実施する機械翻訳システムは、語句の意味の範囲や構文上の機能等において同等の範囲や機能を有する訳語を登録していなくとも、本発明の機構により、訳語を適用する段階で訳語の前後関係から適切な形態に自動的に変形、書換えを行ない、あたかもその訳文に合わせて辞書登録がなされていたかのように動作する。

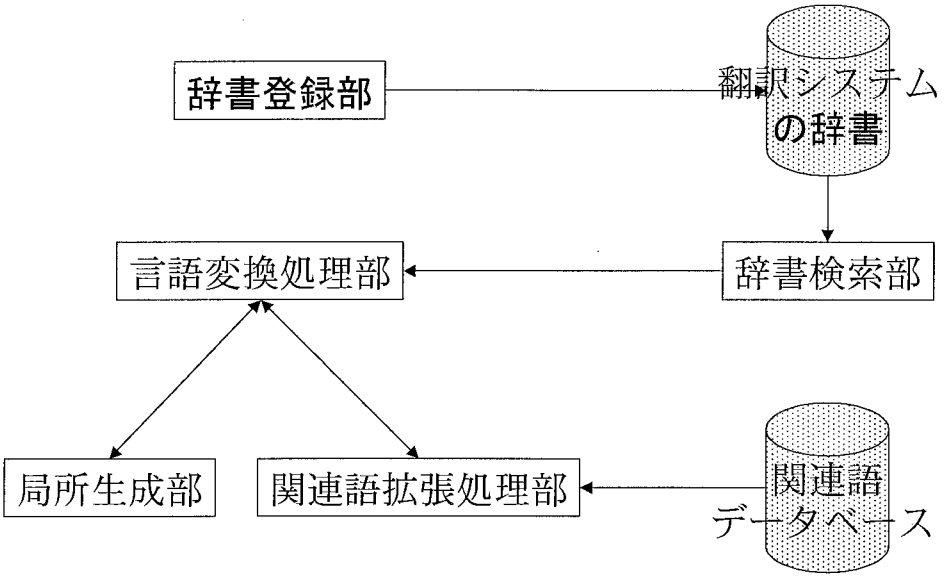

図６．２－１ 装置構成

6.3.機械翻訳向け換言装置

6.3.1.発明の名称

機械翻訳向け換言装置

6.3.2.従来技術

(2-1)従来技術文献

- テキスト自動前編集装置　特開平５－２２５２３２
- テキスト自動前編集装置　特開平６－１３９２７４
- テキスト自動前編集装置及び方法並びにこれに利用される記憶媒体　特開 2000－２６８０３４
- 機械翻訳における後編集装置　特開平６－１７６０５９

- 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自動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)
- 吉見毅彦, 佐田いち子, 福持陽士. 頑健な英日機械翻訳システム実現のための原文自動前編集. 自然言語処理, Vol.7, No.4, pp.99-117 (2000)

(2-2)従来技術文献の要旨および問題点

従来の自動編集、自動書換えなど（以下、換言という）は機構的には既知の機械翻訳機構と同等またはその一部であり、換言の結果が複数個可能であれば、それらの結果の中から人間の操作者が一つを選択することが必要であった。こうした換言機構を備えただけの翻訳装置では、操作者は原言語、目的言語（例えば日英翻訳においては原言語は日本語、目的言語は英語）の知識を持ち、換言の結果を選択する能力を備える必要があったため、当該翻訳装置の利用者が限定されるという問題があった。

システムによっては、例えば、「パスポートを拝見できますか。」を一旦英語に翻訳してから日本語に直訳し直したような、「貴方のパスポートを私に見せて下さい。」と換言しなければならなかった。これでは、その翻訳装置を使えるのは英語に翻訳する能力のある人に限定されてしまうことになる。あるいは、ごく自然な「パスポートを見せて頂けますか」という表現から「パスポートを見せて下さい」や「パスポートを見せろ」まで様々な表現が可能な中から、目の前の翻訳装置ではどの表現ならうまく翻訳できるのか知っていなければならなかった。これでは、その翻訳装置を使えるのはその翻訳装置について熟知した人に限定されてしまうことになるのである。

あるいは、換言の結果が複数個可能にならないように、換言操作の根拠となる知識（以下、換言知識という）の条件をあらかじめ詳細に設定しておかなければならなかった。こうした換言機構を備えただけの翻訳装置では、換言機構以外の狭義の翻訳（以下、言語変

換という）機構における言語変換の根拠となる知識の開発・準備の負荷の一部を換言知識の開発・準備の負荷に転嫁したものであり、当該翻訳装置の性能は開発者の知識・能力に大きく左右される点に変わりなかった。このため、機械翻訳装置の開発には大きな資源が必要であるという問題があった。

6.3.3.発明の要旨

上記の問題点を解決するために、本発明では複数個可能になった換言の結果の中から自動的に一つを選択する機構を設けた。

6.3.4.発明の効果

本発明により、機械翻訳装置の操作者に原言語、目的言語の知識を要求せずに済むため、機械翻訳装置の用途が広がる。例えば、音声認識装置や音声合成装置と組み合わせて音声翻訳（通訳）装置とし、一般の人が海外旅行に持って出るということが可能になる。

あるいは、機械翻訳装置の開発に必要な資源を低減できる。これにより、例えば日英間以外の言語間での機械翻訳装置の開発が容易になる。また、日英間の翻訳装置でも、現在は専門用語辞書の入れ替えといった極一部の変更で幾つかの分野・用途に対応しているのに対し、分野・用途に特化した機械翻訳装置の開発が容易になる。

6.3.5.実施例

説明のため、日英翻訳を例にとる。原言語は日本語、目的言語は英語である。

本発明を実施する機械翻訳システムは、翻訳準備と翻訳実行との 2 つのフェーズからなる。翻訳実行のフェーズにおいて、原言語を目的言語に変換する直接の操作を言語変換ということにする。

翻訳準備のフェーズでは、表現素片収集部は、原言語の大規模なコーパスを走査して原言語表現の類似性を判断するためのデータを収集する。日本語の場合は、例えば、あらかじめ漢字の並びは 2 文字、カタカナの並びは小さな文字「ッ」「ャ」「ュ」「ョ」や「ー」を含めず 2 文字、文字種の変わり目を含む場合は 3 文字、等等の単純な基準を定めておき、その基準の文字数の部分文字列について出現度数を求め、あらかじめ定めた閾値以上の度数を持つ文字列を表現素片データベースに登録し保持しておく。

この例では、対訳コーパスの原言語側の用例文に「パスポートを見せて頂けますか。」があったものとする。他に「パスポート」や「～を見せて」や「～て頂けますか。」を含む用例文も相当数あったものとすると、この文からは「パス」「スポー」「ポート」「を見せ」「見

せて」「せて」「て頂け」「頂けま」「けま」「ます」「すか」が閾値以上の度数を持つ文字列として表現素片データベースに登録され保持されることになる。

換言文生成部は、言語変換部が利用している対訳コーパスの原言語側の用例文を、翻訳する原文に対するのと同様の手段で解析する。さらに換言文生成部は、既知の手段により、文解析結果を換言知識の換言操作に従って換言文に変換し、出力する。既知の手段とは、例えば既存の機械翻訳装置に利用されている、解析手段、変換手段、生成手段等の手段であり、原言語と目的言語が同一言語である点を除けば、換言文生成部は既存の機械翻訳装置の機構と同等またはその一部として構成される。

この例では、対訳コーパスの原言語側の用例文に「パスポートを見せて頂けますか。」があったものとする。この文に対する換言文は、「パスポートを拝見いたします。」「パスポートを見せて下さい。」「パスポートをお願いします。」等となる。

換言文生成部は、当該換言文の生成に使用した変換手段や生成手段等の換言手段を、換言手段データベースに登録し、併せて当該換言文と組にして換言文データベースに登録していく。この時、換言文生成部は、生成した換言文を、表現素片収集部と同じ基準の文字数の部分文字列（以下、表現素片という）に分解し、それらの表現素片を検索キーとして当該換言文が検索されるように換言文データベースに登録していく。

翻訳実行のフェーズでは、本発明を実施する機械翻訳システムは、翻訳すべき原文を入力すると、その原文を原文換言部に出力して、換言文が返されるのに待機する。

この例では、「パスポートを拝見できますか。」という原文が入力され、原文換言部に出力されたものとする。

原文換言部は、原文を入力すると、類似文検索部に入力した原文を出力して、当該原文に類似した換言文が換言文データベースから検索されて戻されるのに待機する。

類似文検索部は、原文換言部から入力した原文を表現素片収集部と同じ基準の文字数の部分文字列（以下、表現素片という）に分解し、それらの表現素片を検索キーとして換言文データベースを検索し、入力した原文と表現素片を共有する換言文をすべて得る。それらの換言文と共有する表現素片を検査し、表現素片データベースに登録された表現素片のうち、より度数の高い表現素片を共有し、より多種の表現素片を共有する換言文をより類似している換言文であるとみなすように構成する、等して、類似した換言文を検索する。

類似文検索部は、類似した換言文と、当該換言文と組にして登録された換言手段とをリ

ストにして原文換言部に出力する。この時、類似文検索部は、入力した原文と類似した換言文との表現素片の差異を、換言手段と同形式で表現し、当該換言文と組にして登録された換言手段とのリストに含めて、原文換言部に出力する。

この例では、「パスポートを拝見いたします。」と「パスカードを拝見できますか。」という２つの換言文が原文換言部に出力されたものとする。

原文換言部は、類似文検索部から当該原文に類似した換言文のリストを入力すると、原文と当該原文に類似した換言文のリストとを表現検証部に出力し、それぞれの換言文が換言文として妥当であるか検証されて戻されるのに待機する。

この例では、「パスポートを拝見いたします。」と「パスカードを拝見できますか。」という２つの換言文が表現検証部に出力されることになる。

表現検証部は、原文と当該原文に類似した換言文のリストとを入力すると、入力した換言文と組にして登録された換言手段を、換言手段データベースから検索し、換言文毎に、換言前の用例文から当該換言文に到る換言操作の妥当性を評価する。ここでいう換言文は、換言文生成部により言語変換部が利用している対訳コーパスの原言語側の用例文から換言されたものであり、換言前の文は用例文にあたる。妥当性を評価するには、換言手段データベースに登録されている換言手段の適用頻度が低い換言手段を用例文／換言文の長さに比して多く含む換言操作の妥当性が低くなるように、換言手段の適用頻度、換言操作に含まれる換言手段の数、用例文／換言文の長さに重み付けを行なう、等することで、妥当性のスコアを算出する。

次に表現検証部は、原文からそれぞれの換言文に到る換言操作の妥当性を評価する。そのために、表現検証部は、類似文検索部が換言手段と同形式で表現した、原文と類似した換言文との表現素片の差異を、換言手段データベースに登録されている換言手段の適用頻度が低い換言手段を原文／換言文の長さに比して多く含む換言操作の妥当性が低くなるように、換言手段の適用頻度、換言操作に含まれる換言手段の数、原文／換言文の長さに重み付けを行なう、等することで、妥当性のスコアを算出する。

この例では、「パスポートを見せて頂けますか。」という用例文から「パスポートを拝見いたします。」という換言文に到る換言操作が含む換言手段も、「パスポートを拝見できますか。」という原文から「パスポートを拝見いたします。」という換言文に到る換言操作が含む換言手段も、換言手段データベースで相当な適用頻度があったものとする。よって妥当性スコアはかなり高く算出されることになる。

これに対し、「パスポートを拝見できますか。」という原文から「パスカードを拝見でき

ますか。」という換言文に到る換言操作が含む換言手段（「ポート」→「カード」）は、換言手段データベースで全く適用頻度が登録されていないものとする。よって妥当性スコアは極めて低く算出されることになる。

表現検証部は、入力した換言文と算出した 2 種類の妥当性のスコアとを組にしたリストを、原文換言部に出力する。

原文換言部は、表現検証部から妥当性のスコアと組にされた換言文のリストを入力すると、入力したリストの最も妥当性の高い換言文と組にされて登録された換言手段を逆方向に適用して、入力した原文を換言する。逆方向とは、換言後の文を換言前の文に逆戻りするように適用することである。ここでいう換言文は、換言文生成部により言語変換部が利用している対訳コーパスの原言語側の用例文から換言されたものであり、換言前の文は用例文にあたる。よって原文は用例文のいずれかに近い文に換言される。

この例では、「パスポートを拝見いたします。」という換言文が入力した原文に類似しているものとして検索され、かつ最も妥当性の高い換言文として評価されることになる。その換言文と組にされて登録された換言手段を逆方向に適用して、「パスポートを見せて頂けますか。」という対訳コーパスの原言語側の用例文に換言する。

本発明を実施する機械翻訳システムは、原文換言部から換言文が返されると、その換言文で原文を置換し、言語変換部に出力して言語変換処理をさせ、言語変換された目的言語文を所定の形式に整形する等の翻訳処理を進める。

以上、説明したように、本発明を実施する機械翻訳システムは、操作者がどのような換言を行えば効果的であるかの知識を有していなくとも、本発明の機構により、当該機械翻訳システムが言語変換処理に利用する知識の獲得源である用例文に最も近い形態に、自動的に換言を行なうことで、あたかも、どのような換言を行えば効果的であるかを熟知した操作者が当該機械翻訳システムに合わせて換言を行なったかのように動作する。

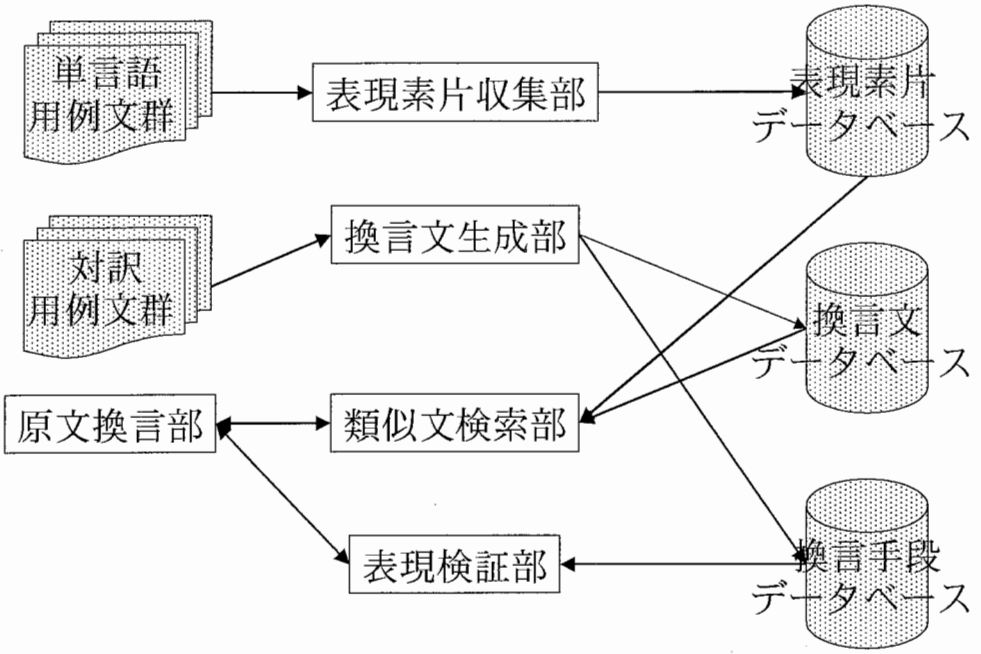
単言語
用例文群
表現素片収集部
表現素片
データベース
対訳
用例文群
換言文生成部
換言文
データベース
原文換言部
類似文検索部
表現検証部
換言手段
データベース

図６．３－１ 装置構成