自 然 言 語 処 理 100−11
（1994. 3. 18）

ユーザー主導型機械翻訳システム TOPTRAN

仁井正治
凸版印刷株式会社　生産技術研究本部

現在、数多くのユーザーで、機械翻訳システムが導入されて使用されている。しかしながら、導入ユーザーの内、コストに見合っただけの機械化による作業の効率化が計られているユーザーは非常に少なく、我々を始めとして、大多数のユーザーは、現状の機械翻訳システムに不満を持っているものと思われる。その原因はいくつか上げられるが、大きな原因の一つは、機械翻訳システムがメーカーの論理に基づいて開発されているためであると思われる。我々は、「ユーザーのユーザーによるユーザーのための」英日・日英機械翻訳システムの研究開発中である。本稿では、我々ユーザー自身が研究開発中の機械翻訳システムＴＯＰＴＲＡＮ（TOPPAN Translation Database System）の概要について報告する。

User Initiative Machine Translation System TOPTRAN

Seiji NII
Corporate Manufacturing, Technology & Research Division
TOPPAN PRINTING CO., LTD.
1, Kanda Izumi-cho, Chiyoda-ku, Tokyo, 101 Japan

Recently, many users are equipped with machine translation systems. However, the enhancement in efficiency brought by the mechanization of translation does not always deserve the introduction and running cost. Many users, including us, seem to be dissatisfied with current machine translation systems. We suspect that one of the biggest reasons would be the developing philosophy based on manufacturer's logic. Conversely, we are in course of R&D enterprise of a MT system "of the user, by the user, for the user." This paper overlooks our system called TOPTRAN (TOPPAN Translation Database System).

１．はじめに

現在の機械翻訳システムのほとんどは、メーカーが企画設計し、開発したものである。したがって、その商品コンセプトは、ユーザーの声を反映したシステムであるとは言うものの、メーカーの論理に基づいて作られている。

その一番の例が、翻訳能力の本質的な向上はメーカーの研究開発者でないと行えないと言うことである。我々ユーザー側でできるのは辞書整備だけであり、本格的に翻訳能力を高めようとすれば、メーカーに改良を依頼するか、メーカーと共同開発を行うしか方法がない。

それでは、メーカーに依頼すれば直ちに翻訳能力が向上するのか。答えは、否である。現在商品化されているほとんどの機械翻訳システムの翻訳能力の善し悪しは、文法規則や構文規則などのようにシステムス内部に保持している規則に大部分が依存する。したがって、翻訳能力を向上させるためには、これらの規則を増やしていく必要があり、メーカーの研究開発者はその努力をする。ところが、これらの規則を増やして行くと、規則同士が干渉しあって、かえって翻訳の質が落ちることが多く、翻訳品質の改良が次第に困難になってくる。

実際に機械翻訳システムを使用して翻訳を行おうとすると、人手翻訳作業には無い別の作業を強いられることになる。例えば前編集や後編集と言われているものである。前編集は、システムが翻訳しやすいように、前もって品詞指定、書換えや分割などを行うことであり、使用者がシステムに合わせなければならない。

前編集を行えば確かに翻訳品質は向上するが、前編集に要する労力に比べて得る結果ははなはだ小さいのが普通である。また後編集で翻訳結果に赤を入れるのだが、もう一度同じ文の翻訳をやらせると、また編集前と同じ訳が出力されて後編集の労力が無駄になる。

以上のようなことから、実際の翻訳業務では、第一線の翻訳者は、費やす労力に比べて得るものが少ないので、マニュアルなどの翻訳以外は、機械翻訳システムを使用することに消極的である。それでは、どうすれば第一線の翻訳者が使いたくなるような作業効率の高い機械翻訳システムが用意できるのか。本稿では、これらの疑問に対する我々の試みについて述べる。

２．ユーザー主導型機械翻訳システム

メーカーの論理に基づく現在の機械翻訳システムを、メーカー主導型機械翻訳システム（Manufacturer Initiative Machine Translation System:ＭＩＭＴＳ）と呼び、ユーザーの論理に基づく機械翻訳システムを、ユーザー主導型機械翻訳システム（User Initiative Machine Translation System:ＵＩＭＴＳ）と呼ぶことにする。

第一線の翻訳者などの意見から、ＵＩＭＴＳに求められる機能は以下のものであると思われる。

(1) 仮に完全翻訳が不可能でも、下訳として使用可能な翻訳品質であること。
(2) 機械翻訳システムの内部を知らない通常の翻訳者が、システムの翻訳能力を向上させられること。
(3) 一度ポストエディットを行うと、それを学習して、次からは同一文に関しては同一の訳文を出力することも可能であること。
(4) プレエディット等、人手翻訳で行わない作業は、行わなくても済む事。
(5) システム操作性はワープロのレベルであること。（現在、ほとんどの翻訳者はワープロを使用して翻訳作業を行っている。）

これらの要求を満たす機械翻訳システムが、我々の欲しいＵＩＭＴＳとなる。

３．ＥＢＭＴ方式

ＥＢＭＴ（Example-Based Machine Translation）[2]方式の機械翻訳システムは、長尾の論文[1]に端を発する、「抽象化された規則に頼るのではなく、豊富な実例／用例を積極的に利用する」[3]方式である。

この方式の特徴は、次のものである。

(1) ある文の翻訳を行うとき、それと良く似た文の翻訳例を見つけ、それを模倣することによって行う。
(2) 新しい翻訳例を追加していくだけで、翻訳能力を向上させることができる。

(3) 人間の訳した翻訳例を真似ることによって、こなれた訳を出力できる。

これらの特徴は、ＵＩＭＴＳの要件の内、システム操作性を除く要件をほぼ満たすものである。

４．TOPTRAN

我々は、ＵＩＭＴＳとしてＥＢＭＴ方式の機械翻訳システムが、現段階では理想のシステムであるとの結論に達したが、この方式の機械翻訳システムは、まだ世の中に存在していないし、それを実現させる技術も確立されていない。そこで、我々独自にＥＢＭＴ方式の機械翻訳システムを研究開発することとし、そのシステムをＴＯＰＴＲＡＮ（TOPPAN Translation Database System）と名付けた。以下にＴＯＰＴＲＡＮの概要について述べる。

４．１　TOPTRANの目指す方向

ＴＯＰＴＲＡＮの目指す方向は、

(1) 翻訳能力を向上させるには対訳データを学習させればよい。
(2) 翻訳作業時プレエディットは行わなくてよい。
(3) ポストエディットした結果は、次の翻訳時に反映される。
(4) 翻訳の質は第一線の翻訳者の下訳として使えるものであれば良く、必ずしも完全翻訳を目指さない。
(5) 操作はワープロ並。
(6) 英日・日英両方向の翻訳を可能とする。

である。

４．２　TOPTRANの特徴

ＴＯＰＴＲＡＮは、対訳登録、対訳学習及び翻訳の機能を持つ。以下にそれぞれの機能の概要を述べる。

(1) 対訳登録

・オペレータが、英文と邦文のペアを用意して（必ずしも同一ファイル内にある必要はない）システムに登録指示を出すだけで、対訳データベースに対訳データが登録される。（「図４－１　同一ファイル内の対訳登録例」及び「図４－２　異なるファイルの登録例」参照。）

・対訳データベース・キーは、英語と日本語それぞれの形態素解析結果の、文字列、品詞情報その他、及びパーツ（(2) 対訳学習：参照）解析結果を使用して作成され、自動的に付加される。

```
                    TOPTRAN訳文登録&学習
h ヘルプ
--------------------------------------------------------------------------------
第31文

英文

Selectively plated with minimum of 150 micro inches of silver on the bonding are
a, measured as in section 6.3.1.


邦文

６．３．１の項の計測方法により、ボンディングされる部分に銀メッキが最低１５０マイ
クロインチの厚さで部分メッキが施されていること。

--------------------------------------------------------------------------------
コマンド>
英語未登録語:
日本語未登録語:
```

図４－１　同一ファイル内の対訳登録例

```
                    TOPTRAN訳文データベース登録
H ヘルプ        英語文番号=31 邦文文番号=31                    訳文登録中
--------------------------------------------------------------------------------
 31 Selectively plated with minimum    ６．３．１の項の計測方法により、ボンデ
    of 150 micro inches of silver on   ィングされる部分に銀メッキが最低１５０
    the bonding area, measured as in   マイクロインチの厚さで部分メッキが施さ
    section 6.3.1.                     れていること。

 32 Plating shall be uniform in        メッキは、見た目に均一で、滑らかで、光
    appearance, smooth, bright in      沢があり、ブリスター、剥離、こぶ、変色
    color, free from blisters,         、不純物、ムラがあったりしてはならない
    peeling, nodules, discoloration,   。
    contaminants and burn marks.
                                       またベースの金属が露出してはならない。
 33 There shall be no exposed base
    metal.                             メッキの下の銅の地金は、見た目に均一で
                                       、滑らかで、かつブリスターや、剥離、変
 34 Copper strike under plating        色、不純物があってはならない。
    shall be uniform in appearance,
--------------------------------------------------------------------------------
対訳登録しますか？ (y/n/q/e/k/r/p/g/t/b/h/l/d/c)
```

図４－２　異なるファイルの登録例

(2) 対訳学習

・登録された対訳の英文、邦文それぞれを、単語レベル、句レベルまたは文レベルに、互いに対応が可能な大きさに分解する。この分解したものを、対訳パーツ、と呼ぶ。

・英語の対訳パーツと日本語の対訳パーツで対応可能なものを、システムが今までの学習結果とパーツ解析結果を基に推論して、自動的に対応付けを行う。これを、自動パーツ・リンク、と呼ぶ。（「図４－３　対

訳学習例」参照。）

・オペレータは、表示された自動パーツ・リンクの結果を見て正しいかどうかを確認し、正しければ、学習指示を出すだけで対訳学習を行う事ができる。もし、誤っていれば、修正操作を行って、正しいパーツ・リンクの対訳学習を行う事ができる。

・システムは、対訳パーツ情報及び対訳パーツ間のパーツ・リンク情報を、学習結果として対訳データベースに蓄積する。この情報は、以後の自動パーツ・リンクに使用される。

(3)翻訳

・プレエディットの必要はなく、原文をそのままキー入力、またはファイル入力する。

・入力原文にたいして対訳登録と同じ方法でキーが生成される。

・生成されたキーを基に対訳データベースが検索され、類似度計算結果で一番近い対訳データ情報が引き出される。（類似度計算には、いわゆるシソーラスは使用せず、ＴＯＰＴＲＡＮ独自の軽い分類を用いている。）

・この引き出された情報が使用されて翻訳文が生成されるが、必要とされる対訳パーツが全部揃わないときは、次に近い対訳データ情報が使用され、それでも揃わない時は更に次に近い対訳データ情報が使用され、対訳パーツが全部揃うまでこれが繰り返される。（「図４－４　日英翻訳例」参照。）

```
                    ＴＯＰＴＲＡＮ対訳学習
  第3646文の学習中
--------------------------------------------------------------------------------
3646 Selectively plated with minimum    ６．３．１の項の計測方法により、ボンデ
     of 150 micro inches of silver on   ィングされる部分に銀メッキが最低１５０
     the bonding area, measured as in   マイクロインチの厚さで部分メッキが施さ
     section 6.3.1.                     れていること。

パーツ・リンク情報
<<1:Selectively 2:plated 18:部分 19:メッキが 20:施さ 21:れ て 22:いる 23:こと >
>
<<3:with #####>>
<<4:minimum 5:of 12:最低 >>
<<8:150 13:１５０ >>
<<7:micro 8:inches 14:マイクロ 15:インチの >>
<<9:of silver 10:銀 >>
<<10:on 11:the bonding area 7:ボンディングさ 8:れる 9:部分に >>
<<13:measured #####>>
<<14:as #####>>
<<15:in section 2:項の >>
<<16:6.3.1 1:６．３．１の >>

リンクＯＫ？  (y/n) n
パーツ・リンク・リストを出力しますか？(y/n) n

英語パーツ情報
1:Selectively 2:plated 3:with 4:minimum 5:of 6:150 7:micro 8:inches 9:of silver
10:on 11:the bonding area 12:, 13:measured 14:as 15:in section 16:6.3.1

日本語パーツ情報
1:６．３．１ の 2:項 の 3:計測 4:方法 5:に より 6:、 7:ボンディングさ 8:れる 9:部
分 に 10:銀 11:メッキ が 12:最低 13:１５０ 14:マイクロ 15:インチ の 16:厚 さ 17:
で 18:部分 19:メッキ が 20:施さ 21:れ て 22:いる 23:こと 24:。

英語リンク：3
日本語リンク：16 17
英語リンク：13 14
日本語リンク：3 4 5
英語リンク：
```

図４－３　対訳学習例

<原文>

ボンディングされる部分に最低１００マイクロインチの厚さで銀による部分メッキが施されていること。

<類似度計算の結果一番似ている文--->３６４６文>

Selectively plated with minimum of 150 micro inches of silver on the bonding area, measured as in section 6.3.1.

６．３．１の項の計測方法により、ボンディングされる部分に銀メッキが最低１５０マイクロインチの厚さで部分メッキが施されていること。

<３６４６文のパーツ・リンク情報>
Selectively plated <-----> 部分 メッキが 施さ れ て いる こと
with <-----> 厚さ で
minimum of <-----> 最低
150 <-----> １５０
micro inches <-----> マイクロ インチの
of silver <-----> 銀
on the bonding area <-----> ボンディングさ れる 部分に
measured as <-----> 計測 方法 により
in section <-----> 項の
6.3.1 <-----> ６．３．１の

<使えるパーツ・リンク>
Selectively plated <-----> 部分 メッキが 施さ れ て いる こと
with <-----> 厚さ で
minimum of <-----> 最低
150 <****> １５０
micro inches <-----> マイクロ インチの
of silver <-----> 銀
on the bonding area <-----> ボンディングさ れる 部分に
measured as <****> 計測 方法 により
in section <****> 項の
6.3.1 <****> ６．３．１の

<訳文>
Selectively plated with minimum of 100 micro inches of silver on the bonding area.

図４－４　日英翻訳例

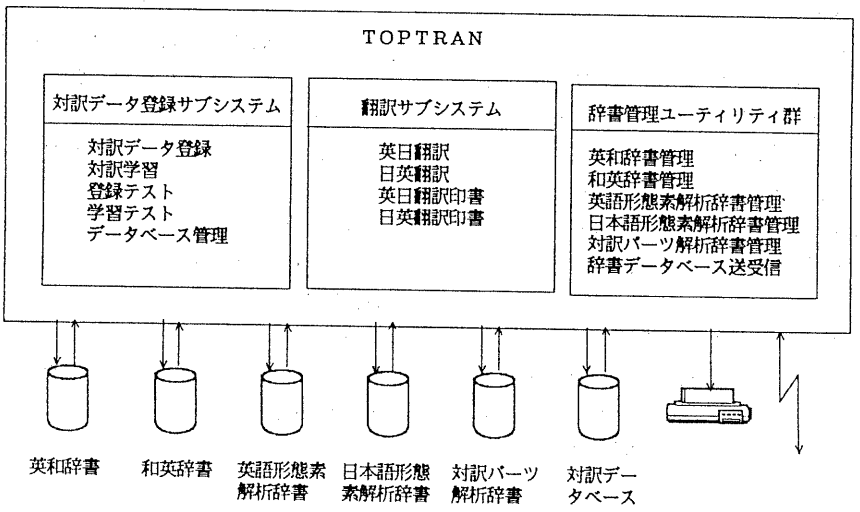

図４－５　ＴＯＰＴＲＡＮシステム構成図

(4) その他

・ポストエディットの結果は、そのまま対訳登録、対訳学習させることができる。

4．3 TOPTRANのシステム構成

TOPTRANは、対訳データ登録サブシステム、翻訳サブシステム及び辞書管理ユーティリティ群により構成される。以下にそれぞれの概要について述べる。(「図4－5 TOPTRANシステム構成図」参照。)

(1) 対訳データ登録サブシステム

対訳データを作成登録処理するシステムであり、以下の機能を持つ。

・対訳データ登録
・対訳学習
・登録テスト
・学習テスト
・データベース管理

(2) 翻訳サブシステム

英日・日英の翻訳を行うシステムであり、以下の機能を持つ。

・英日一括翻訳
・英日一文翻訳
・日英一括翻訳
・日英一文翻訳
・英日翻訳印書
・日英翻訳印書

(3) 辞書管理ユーティリティ群

各種の辞書を管理するユーティリティ群であり、以下のユーティリティよりなる。

・英和辞書管理ユーティリティ
・和英辞書管理ユーティリティ
・英語形態素解析辞書管理ユーティリティ
・日本語形態素解析辞書管理ユーティリティ
・対訳パーツ解析辞書管理ユーティリティ
・辞書データベース送受信ユーティリティ
(ワークステーション間で辞書及び対訳データベースを送受信するもの)

5．実用実験

5．1 実験対象

凸版印刷は、高度リードフレームの生産分野では世界的シェアを持っており、米国を始めとして世界各国に輸出している。リードフレームを受注生産する場合、クライアントと当社の間で、仕様書、解説書、その他の技術文書がやり取りされる。この時、翻訳部門でそれらの技術文書を英日・日英翻訳しているが、専門用語及び独特の表現が多いのと、クライアントの機密保護のため、社内の特定の翻訳チームが翻訳作業を行っている。したがって、リードフレームの受注がラッシュしたときは、このチームに負荷が掛かりすぎる場合があり、機械化による(即ち機械翻訳システムを使う事であるが)効率化が必要とされている。そこで我々は、このリードフレーム関連の技術文書の翻訳を、実用実験として行う事とした。

5．2 実験環境

当社翻訳部門の機械翻訳グループでは、市販の機械翻訳システムを中心にした、ワークステーション及びパソコンによるネットワーク型のドキュメント処理システムを使用して、翻訳作業を行っている。我々は、このネットワークのNSSUN-SP2上にTOPTRANをインプリメントして、既存の翻訳作業の流れの中で実用実験を行った。

登録・学習に使用する英語と日本語の対訳データの内、いわゆる一般科学技術に関する領域の対訳については、人手翻訳の部隊から、リードフレームに関する対訳については、前述のリードフレーム翻訳チームから、それぞれフロッピィでもらった。

翻訳実験時、紙の技術文書からの英文データ入力は、通常の翻訳作業の流れと同じく、市販の機械翻訳システムのOCR入力よるものと、人手によるワープロ入力を利用した。また、日本語データは、ワープロ入力されている翻訳データをそのまま利用した。

５．３　実験結果

５．３．１ 実験条件

翻訳実験は以下の条件下で行った。

(1) 実験時の対訳データベース

・学習対訳データ　　４，３５９対

内、英語の単語数　４４，５３３語

・学習対訳パーツ　　１２，２６８個

(2) 実験１

学習済みデータと同種のリードフレーム関連技術資料で、クライアントが異なる人手翻訳済みのデータを、無作為に５０文、対訳で抽出。その対訳から、英語文と日本語文をそれぞれ別ファイルに格納し、実験データとして使用。

(3) 実験２

学習済みデータとは全く異なる種類のリードフレーム関連技術資料で、人手翻訳済みのデータから、無作為に５０文、対訳で抽出。その対訳から、英語文と日本語文をそれぞれ別ファイルに格納し、実験データとして使用。

(4) 下訳使用可能判定基準

以下を下訳使用可能判定基準とした。

①英日翻訳

・英単語及びフレーズに対応する日本語訳の表記が、文中で日本語として自然なこと。意味が通じるだけでは不可。

・日本語訳の単語または文節間の語順は、狂ってもよい。例えば、「図５－１」の下線部分の単語「銀」は、「図５－２」の下線で示す位置が自然であるが、「図５－１」位置でも良いものとする。

◇ボンディングさ れる 部分 に 銀 最低
１００ マイクロ インチ の 厚 さ で 部
分 メッキ が 施さ れ て いる こと

図５－１　語順例１

◇ボンディングさ れる 部分 に 最低
１００ マイクロ インチ の 厚 さ で 銀
部分 メッキ が 施さ れ て いる こと

図５－２　語順例２

・日本語訳の格助詞その他の助詞は、抜けたり間違ったりしてもよい。例えば、「図５－２」の下線部分の単語「銀」は、「図５－３」の下線で示すように、格助詞「の」が続くのも、自然な日本語文の一つであると言う事ができるが、このような場合、「の」が抜けていても良いものとする。

◇ボンディングさ れる 部分 に 最低
１００ マイクロ インチ の 厚 さ で 銀の
部分 メッキ が 施さ れ て いる こと

図５－３　「の」がつく場合

②日英翻訳

・訳出された英単語及びフレーズの表記が、文中で英語として自然なこと。意味が通じるだけでは不可。

・前置詞at、on、inは、抜けてもよい。例えば、「図５－４」の下線部分のようにinが抜けていても良いものとする。

---

29 加熱された試料は、６・４に書かれている
要領で目視検査する。

◇Heated strip will be visualed
as outlined _ 6 . 4

図５－４　前置詞の抜けた例

・接続詞and、orは、抜けてもよい。例えば、「図５－５」の下線部分のようにandが抜けていても良いものとする。

---

23 釣りリードの傾斜の下限Ａに軸を動かし、
再度焦点を合わせる。

◇Move to the tie bar at the
base of the downset A _ refocus

図５－５ andの抜けた例

５．３．２ 実験結果

それぞれの実験で、下訳として使用可能な訳文の数の、全体に対する比率（以後、下訳率と呼ぶ）は、「表５－１ 実験結果」に示すものであった。

	英日翻訳	日英翻訳
実験１	７４％	７６％
実験２	１８％	１２％

表５－１ 実験結果

ＥＢＭＴ方式の機械翻訳システムの特徴と、この実験結果から、ＴＯＰＴＲＡＮに、翻訳対象分野の同種の対訳データを学習させて行けば、７０％を超える下訳率の翻訳が可能なことが判明した。このことは、ＴＯＰＴＲＡＮがＵＩＭＴＳとして、実用に耐え得る可能性のある事を示すものであると言える。

６．今後の課題

ＴＯＰＴＲＡＮが、現在翻訳対象としているリードフレームの分野だけでも、第一線の翻訳者が下訳として使用可能な品質の翻訳を行えるようになるには、翻訳品質を決める重要な要素である、対訳データベース中の学習対訳データと学習対訳パーツが、またまだ不足している。我々は現在、ＴＯＰＴＲＡＮに既に登録学習済みの対訳データ以外に、約１万対の対訳データを翻訳部隊から既に入手している。これらをいかにロー・コストで短期間に、蓄積するかが大きな課題である。

対訳データの蓄積以外に、翻訳アルゴリズム及び対訳データベースの構造その他で、多大な改良余地が残っているし、実用システムとしては、会話文とテキスト・データの違いはあるものの、ＡＴＲのように翻訳結果の８０％以上[5]が使用可能なようにする必要がある。

また、現在のＴＯＰＴＲＡＮでは問題となっていない翻訳速度についても、対訳データベースが巨大になれば、佐藤の提案[7]にあるように、並列処理計算機等、現在インプリメントしているNSSUN-SP2以外のシステムを考慮する必要があるかも知れない。

謝辞

本研究開発において、貴重な助言を頂いた、京都大学工学部・長尾教授、北陸先端科学技術大学院大学・情報科学研究科・佐藤助教授、大阪大学言語文化部・成田助教授、及び、多量の対訳データを用意して研究開発を支えて下さった、当社ドキュメント・エンジニアリング部の翻訳スタッフ各位に感謝いたします。

参考文献

[1] Nagao,M., A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, in ARTIFICIAL AND HUMAN INTELLIG-ENCE(Elithorn & Banerji,Eds.),Elsevier Science Publishers,pp173-180,1984

[2] Sato,S., Example-Based Machine Translation, Doctorial Thesis,Kyoto University,1991

[3] 佐藤理史：実例に基づく翻訳のアプローチ，日本学術振興会・文字言語音声言語の知能的処理第１５２委員会資料，1992

[4] 佐藤理史：実例に基づく翻訳，情報処理(VOL.33 NO.6 June 1992)

[5] Iida,H., Beyond Analysis-Centerd Large Scale MT Systems, IWST'93

[6] 古瀬、隅田、飯田：変換主導型機械翻訳の実現手法，情報処理学会自然言語研究報告 80-8 (1990)

[7] 佐藤理史：超並列計算機を用いた実例型翻訳の実現，人工知能学会全国大会論文，1993