自 然 言 語 処 理 102－8
（1994．7．22）

TOPTRANにおける事例学習

仁井正治
凸版印刷株式会社　生産技術研究本部

現在の機械翻訳システムは、言語感覚や翻訳レベルで、開発した研究開発者のそれら以上には、決してなり得ない。超一流の翻訳者といえども、勉強して日々研鑽に励んでいるのに、機械翻訳システムは、購入した時点の言語レベルのままである。もはや、研究開発者（私も含む）が、システムの言語レベルで、主導権を取る時代ではなく、真に実用に耐え得る機械翻訳システムの言語レベルは、ユーザーに任せる時期にきている。我々は、EBMT方式で、ユーザーが手作りに近い感覚で、言語レベルを向上させる得る、機械翻訳システムTOPTRANを開発した。本稿では、TOPTRANにおける学習機能の捉え方と事例の学習方法について述べる。

TOPTRAN Example Learning

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

Current machine translation systems cannot surpass the language snese and translation ability of the researchers who develop them. Although even the best translators pursue their studies daily, machine translation systems remain at exactly the same level of language ability as when they were purchased. Now it is no longer an age when research and developers (myself included) take the lead in determining system language ability; it has become, rather, an age when this is to be left to the users of machine translation systems that can stand up to actual use. We have developed TOPTRAN whose language ability is increased by users in an almost a do-it-yourself fashion. This paper will explain TOPTRAN's learning functions and methods of example learning.

１．はじめに

近年、国際化の進展とビジネス構造の変化につれて、あらゆる業種で翻訳を伴う業務が増大している。凸版印刷は、総合印刷業として、専門部門を設けてこの翻訳業務の増大に対処してきた。

工場などで作業の効率化を図るためには、機械化を行う。当社も、印刷工程その他に、自動搬送ロボットを含む多数の機械を導入して、効率化を図っている。翻訳作業は知的作業のため、数年前までは人手がすべてであり、機械化は不可能であった。現在では機械翻訳システムが市販されていて、機械化が可能となり、我々も市販の機械翻訳システムを導入して、翻訳作業の効率化を目指している。

ところが、工場などに導入した機械と違って、市販の機械翻訳システムは、導入することによって、かえって作業効率が悪くなる場合が多く、機械を導入した効果が得られていない。

その大きな理由の一つは、工場の機械が、ユーザーの作業の何を効率化するかを、考えて作られたものであるのに対して、既存の機械翻訳システムのほとんどが、「初めに理論ありき」で、学問的、あるいは、研究開発者の立場で、開発されているからであると思われる。基本的翻訳能力は、システム内部の、言語理論に基づく文法規則その他の規則に依存しているばかりでなく、それらの規則に合わない部分をも「例外処理」と称して更に規則化し、すべてを規則に依存しているのが現状である。システムによっては、「例外処理が７割以上もあり、残りがいわゆる規則」と言われているものもある。

その結果、システムの基本的翻訳能力を向上させることができるのは、研究開発者だけとなり、なおかつ、その研究開発者さえも、規則の余りの多さに、次第に対処できなくなって、翻訳能力の向上を行うことができなくなっている。

そこでここ数年、この技術のボトルネックを打ち破る新しい方法として、「事例に基づく翻訳方式（ＥＢＭＴ）」[4] [5] [6] [7] [8] [9]を初めとする、用例／実例／コーパスを用いて翻訳を行う方式が注目され、多くの研究開発が行われている。

しかしながら、このＥＢＭＴ方式の機械翻訳システムを初めとして、現在発表されている技術のほとんどは、既存のシステムとの融合であったり、規則を作るために事例を用いていたり、システム全体ではなく一部の理論の検証のための小さな実験システムであったりして、学問的、あるいは、研究開発者の立場で、研究開発されようとしている。したがって、仮に現在研究開発されている機械翻訳システムが世の中に出たとしても、我々ユーザーにとっては、相変わらず使いづらいシステムになる可能性があると思われる。

以上の状況から、我々は、機械翻訳システムのユーザーの立場に立って、「初めに理論ありき」でない、「初めにユーザーありき」の、ＥＢＭＴ方式で英日・日英翻訳可能な「ユーザー主導型機械翻訳システムＴＯＰＴＲＡＮ」[1]を開発した。本稿では、ＴＯＰＴＲＡＮにおける学習機能に対する考え方を述べ、ＥＢＭＴ方式の利点と問題点を洗い出して、我々の研究開発の目指す方向、及び、事例の学習方法について述べる。

２．機械翻訳システムに於ける学習機能

２．１ 翻訳能力の向上

第一線の翻訳者といえども、最初から高レベルの翻訳ができたわけではなく、翻訳実務を行いながら経験を積んで、レベルが上がって行った結果、現在のレベルに達している。更に、既に最高のレベルにある人でさえ、第一線で翻訳作業を行っている限り、たとえ８０歳を超えようとも、勉強をしてレベルを上げようと努力し続けている。

ところが、既存の機械翻訳システムは、翻訳能力で人間に遙かに劣るにもかかわらず、翻訳能力の向上を考慮した設計になっていない。前述のように、向上させることができるのは、メーカーの研究開発者だけである。そこで、われわれユーザーは、システムを購入した後、辞書整備を必死で行い、できる限り翻訳レベルを上げようと、努力する。更に、辞書整備だけでなく、うまく訳せないような原文に対する訳文を、名詞句として登録

して、少しでもいい訳がでるように、努力して使用している。

ユーザーが、何故、辞書整備や名詞句登録を行うかというと、機械翻訳システムの翻訳レベルは、内部の規則を書いた研究開発者のレベル以上には、決してならないからである。また、たとえ言語レベルの最高の研究開発者が、規則を書いたとしても、言語は生き物で、日々新しい言葉や概念が発生しており、すぐに対応できなくなるからである。いくら優秀な研究開発者といえども、すべての分野で最高の翻訳ができるとは限らない。前述のように、最高レベルの翻訳者でさえ、日々研鑽に励んでいるのに、機械翻訳システムは、システムを購入した時点でのレベルに固定されるか、せいぜいバージョン・アップによる、若干の向上が図られるに過ぎない。

２．２ 学習機能の位置付

以上のことから、研究開発者（私も含む）が、システムの言語レベルまでやるのは、間違いであると考える。研究開発者が、なんだかんだと、主導権を取って行う時代はもう過ぎており、実用に耐え得る機械翻訳システムの翻訳レベルは、ユーザーに任せるべきであると考える。これをやらない限り、良い翻訳ができないし、システムの持つ言語感覚も、実際の翻訳とはかけ離れた、研究開発者の域を出ないもののままである、と思われる。

機械翻訳システムは、道具としての位置づけを、鮮明にし、ユーザーが、自由に翻訳レベルを向上させられることを、基本機能とすべきである。この基本機能こそが、機械翻訳に於ける学習機能であると考える。ユーザーの手作りに近い感覚で、機能アップできるシステムであれば、優秀な翻訳者がシステムに教えれば、実用に耐え得る優秀なシステムになる、と思われる。

２．３ 学習機能とＥＢＭＴ方式

ＥＢＭＴ（Example-Based Machine Translation）[4]方式の機械翻訳システムは、長尾の論文[3]に端を発する、「抽象化された規則に頼るのでなく、豊富な実例／用例を積極的に利用する」[5]方式である。

ＥＢＭＴ方式の機械翻訳システムでは、翻訳能力を向上させるには、対訳例を追加すれば良い。ということは、システムの研究開発者だけでなく、システムの内部を知らない通常の翻訳者でも、対訳例を追加することで、翻訳能力を向上させることができることを、意味している。対訳例を追加するということは、人間的に言い換えれば、ユーザーから見て、機械翻訳システムに、翻訳の仕方を教えることであり、システム側からみると、翻訳の仕方を学習する、ことになる。

それでは、ＥＢＭＴ方式を取り入れるだけで、ユーザーが自由自在に翻訳レベルを向上させることができるようになるのであろうか。現実は、前述したように、現在発表されているＥＢＭＴ方式の技術のすべてが、システムを研究開発する側に都合の良いものばかりで、ユーザーのことを考慮したシステムは皆無である。また、これら発表されているシステムでは、機械翻訳に対する特別の知識無しに、ユーザーが対訳例を教えるだけで、翻訳能力を向上させることは不可能であると思われる。

３．ＥＢＭＴ方式の利点と問題点

ＥＢＭＴ方式で、ユーザーが自由自在に翻訳レベルを向上させることのできるシステムは、どのようなシステムにすれば良いのであろうか。我々は、ＥＢＭＴ方式を取り入れるに当たって、ＥＢＭＴ方式の利点及び問題点について、ユーザーの立場から洗い出してみた。

３．１ ＥＢＭＴ方式の利点

ＥＢＭＴ方式の利点を、研究開発者及びユーザーの両方から捉えてみると、以下のようになると思われる。

(1)研究開発者から見た利点

佐藤が[9]で指摘しているように、

・システムの構築・改良が容易である。

・知識のポータビリティと安定性に優れている。

・こなれた訳を出力できる可能性がある。
・システムの安定性に優れている。
・その訳がどれくらい信頼できるかの指標を求めることができる。
がある。

(2) ユーザーから見た利点

ユーザーから見て考えられる利点は、
・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。
・前編集は行う必要がない。
・後編集の結果を、次回の翻訳に生かすことができる。
・専門の機械翻訳システム・オペレーターは必要なく、通常の翻訳者が容易に操作できる。
である。

3．2 ＥＢＭＴ方式の問題点

ＥＢＭＴ方式の問題点を、研究開発者及びユーザーの両方から捉えてみると、以下のようになると思われる。

(1) 研究開発者から見た問題点

これも、佐藤が[7]で指摘しているように、
・計算量の問題
・翻訳ユニットの大きさの問題
・シソーラスの問題
・翻訳ユニットの応用的な利用の問題
がある。

(2) ユーザーからみた問題点

ユーザーから見た問題点は、
・多量の事例を集められるのか。
・事例はどうやって学習させるのか。
・シソーラスは誰が作るのか。
・類似度計算にシソーラスは役に立つのか。
・ユーザー側で何ができるか。
がある。

(3) ユーザーから見た技術的疑問

研究発表されている技術に対しての、ユーザーの素朴な疑問、
・構文解析ですべて解析できるのか。
世の中では、英語、日本語共に、構文が滅茶苦茶なのが堂々と通用している。
・構文変換方式との融合はどうなるのか。
・例をいちいち集めるのは効率が悪く、ある程度を規則化すると効率がよい。その通りだが、規則化するには、例が必要では。それなら、システムに例を学習させるのと、大差無いのでは。
・メーカー側しか翻訳能力を向上させることができないのでは、ＥＢＭＴの意味が半減するのでは。
がある。

4．ユーザーの素朴な願い

4．1 素朴な願いの言いたい放題

ＥＢＭＴ方式の利点及び問題点について、ユーザーの立場から洗い出してみた結果、我々は、技術的問題を棚に上げて、以下のような機械翻訳システムを開発できないか、素朴に考えてみた。
・最小限の規則でシステムが開発できないか。
・文法は言葉の動きをまとめただけ、初めに文法があった訳ではない。したがって、大量の対訳データーがシステム内部に集まってから、規則にできるものだけ規則化できないか。
・技術が完成してから作るのではなく、使いながら改良すれば良い。つまり、いわゆる技術の学習。そのための仕組みを作れないか。
・シソーラスを使用しないシステムが開発できないか。
・構文解析を行わないシステムを開発できないか。
・翻訳を行いながら、後編集後の対訳データを集めればよい。対訳データを集めてからシステムを作るのではなく、システムを先に作ってデータを集めれば良い。
・能力アップは常に必要。したがって、システムを実際に使用するユーザーしか、能力向上を行えないのでは。であれば、ユーザーが主役の、能力向上の為のオペレーションができないか。

・システムが間違えることを前提にしたシステムを作れないか。システムが誤訳をしたら人が直せば良い。人手翻訳でも、チェッカーが赤を入れているのだ。誤訳しないようにすることからくる、技術の大きな壁。そちらの方が問題なのでは。

4．2 素朴な願いのまとめ

素朴な願いをまとめてみると、以下のようになる。

・人間同様、例文を学習することにより、日々賢くなる機械翻訳システム。
・システムを開発するために例文を集めるのではなく、実作業で使用しながら例文を集める。
・最初は、必要最小限の規則で稼働する。
・システム内部のことを全く知らない普通の翻訳者が、容易に例文を追加できる。
・いわゆるシソーラスは使用しない。
・構文解析は行わない。
・システムが間違えることを前提にアルゴリズムを作る。

5．研究開発の目指す方向

以上の「3．EBMT方式の利点と問題点」及び「4．ユーザーの素朴な願い」から、ユーザーの自由自在になる機械翻訳システムとはなにか、色々考えた結果が、ユーザー主導型機械翻訳システムTOPTRANであり、その目指す方向は以下のものである。

(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良い。
(2) 翻訳作業時、前編集は行わなくても良い。
(3) 後編集した結果は、次の翻訳時に反映される。
(4) 翻訳品質は、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) システムの操作性は、ワープロのレベルであれば良い。
(6) 英日、日英の双方向の翻訳を可能とする。

6． TOPTRANにおける事例学習

「5．研究開発の目指す方向」に記述されている方向のシステムを開発するに当たって、その中心となる機能は、事例学習機能である。一口に翻訳事例を学習させるといっても、ユーザーにとって使い易い学習方法とは、とのようにすれば良いのか、明白ではない。また、各ユーザーがバラバラに、翻訳事例をデータベースに蓄積するのであれば、個々のシステムに蓄積できる数も限られてくる。

そこで我々は、これらの問題を、対訳データの学習の問題、及び、知識データの学習の問題、として捉え、以下に述べる方法で解決を試みた。

6．1 対訳データの学習

対訳を利用して翻訳を行うには二つの方法がある。それは、英語または日本語の翻訳原文が、対訳データベースに蓄積されている対訳データの英語または日本語と、全く一致する場合の翻訳と、部分的に一致する場合の翻訳である。TOPTRANでは、対訳登録を行うことにより、前者が可能となり、対訳学習を行うことにより、後者が可能となる。

6．1．1 対訳登録

対訳登録とは、英文、邦文のペアを対訳データベースに登録することをいう。勿論、対訳登録は、文だけでなく、単語でも良い。

対訳登録を行うことにより、英語または日本語の翻訳原文が、対訳データベースに蓄積されている対訳データの英語または日本語と、全く一致する場合の翻訳が可能となる。

本稿では、紙面の関係で詳細を記述しないので、参考文献[2]を参照されたい。

6．1．2 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、また文レベルに互いに対応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、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もとに推論して、自動的に対応させて表示する。オペレータは、その表示を見て正しければ学習指示を行い、正しくなければ修正操作を行う。「図1 対訳学習例」を使用して、以下に対訳学習の方法を説明する。

(1)自動パーツ・リンク

システムに対して登録命令を出すと、システムは内部で自動的に固有番号を対訳データに付与する。例えばこれが 4923 番とすると、対訳を学習する段階で、この 4923 という番号を入力する。 4923 が入力されると、システムは、4923 文の、対訳データとパーツ・リンク情報を表示し、「リンクＯＫ？ (y/n)」と表示する。これは、システムが過去の学習結果に基づいて推論して、The plastic と「このプラスチックの」という部分が対応するのか、と聞いてきているのである。更に、box に対して「箱は」が、will be に対しては、間違って「密封できる」が対応するか、と聞いてきているのである。以下 with desiccant まで同様である。＊印は、システムが分からない、と答えているのである。

(2)パーツの意味

ＴＯＰＴＲＡＮでは、「パーツ」という言葉を使用していて、「トランスレーション・ユニット」という言葉を使用していないし、構文、ノードという言葉も使用していない。なぜ使用していないかというと、この英語と日本語の対応を、文法的に無茶苦茶にやっているからである。The plastic に対して「このプラスチックの」、box に対しては「箱は」を対応させている。したがって、いわゆる「トランスレーション・ユニット」とは言えないし、構文、あるいは、ノードという言葉とは、意味が異なるので、使わないのである。

(3)パーツとオペレータの言語感覚

オペレータが、「この英語のパーツとこの日本語のパーツは、このように対応する。」と決めてやると、その通りになる。それがこのシステムの大きな特徴である。なぜこういうことをや

```
                    ＴＯＰＴＲＡＮ対訳学習
  第4923文の学習中
--------------------------------------------------------------------------------
4923 The plastic box will be inside     このプラスチックの箱は、乾燥剤と一緒に
     an airtight plastic bag with       密封できるプラスチックの袋に入れる。
     desiccant.

パーツ・リンク情報
<<1:The plastic 1:この 2:プラスチック の >>
<<2:box 3:箱 は >>
<<3:will be 8:密封 できる >>
<<4:inside #####>>
<<5:an airtight #####>>
<<6:plastic 9:プラスチック の >>
<<7:bag 10:袋 に >>
<<8:with 9:desiccant 6:乾燥 剤 と 7:一緒 に >>

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:The plastic 2:box 3:will be 4:inside 5:an airtight 6:plastic 7:bag 8:with 9:de
siccant
日本語パーツ情報
1:この 2:プラスチック の 3:箱 は 5:、 6:乾燥 剤 と 7:一緒 に 8:密封 できる 9:プ
ラスチック の 10:袋 に 11:入れる 12:。
英語リンク：3 4
日本語リンク：11
英語リンク：5
日本語リンク：8
英語リンク：
```

図1 対訳学習例

るかというと、直訳調なら良いのだが、意訳調になると、構文解析など、文法に忠実な解析では、実際の翻訳をやるとうまく訳が出てこない。ここで The plastic box を「このプラスチックの箱は」と対応させないで、なぜ The plastic を「このプラスチックの」に対応させ 、box を「箱は」に対応させるのかということは、理屈ではないのである。要するに、この英語を日本語に訳したときに、あるいは日本語を英語に訳したときに、英語のどこが日本語のどこと、どのように対応するか、オペレータが勝手に決めることができる。それが重要なのである。この機能により、意訳調の翻訳が可能となり、オペレータの持つ言語感覚を、システムに取り入れることができるようになる。

(4)パーツ・リンクの修正

システムは、「リンクOK？ (y/n)」と聞いてきているが、この例では、will be のリンクを間違えているのと、inside 、an airtight については、パーツ・リンクが行われていないので、「n」と答えている。「n」が答えられると、システムは、英語と日本語それぞれのパーツ情報を、番号と共に表示する。パーツは、どのようにして作られているかということは、ここでは説明を省略するが、過去にシステムが学習した結果と、ある規則に則って作られている。オペレータはこれを見て、この例では、3番の will be を間違えてリンクしているので、英語パーツの 3、4番 will be inside に対して、日本語パーツの 11番「入れる」を対応させ、英語パーツ 5番 an airtight は、なにも対応していないので、日本語の 8番「密封できる」を対応させるように指示する。これらの指示が終わると、システムは、今修正指示された結果を取り入れて、再度パーツ・リンク情報を表示する。オペレータは、自分の望むパーツ・リンクが表示されるまで、この操作を繰り返すことができる。

修正指示が終了すると、システムは、リンク情報すべてを、対訳データベースに書き込んで行く。オペレータのやることはこれだけである。オペレータは、システムの内部のことを知らなくても良いし、何ヵ月も費やしてオペレーションを覚える必要もない。また、専任のオペレータを置く必要もない。

６．２ 知識データの学習

前述したように、ＥＢＭＴ方式の問題点の中に、多量の対訳データを集められるか（「３．２ ＥＢＭＴ方式の問題点」参照）、という問題がある。ＴＯＰＴＲＡＮでは、分野別専門システムとＴＯＰＴＲＡＮネットワークにより、この問題を解決している。

６．２．１ 分野別専門システム

ＴＯＰＴＲＡＮは、オペレータが実作業で使用しながら、例文を集める仕組みになっている。したがって、翻訳者一人一人に一台づつＴＯＰＴＲＡＮの稼働するシステムを渡して、それぞれの翻訳者が、自分の得意な分野の対訳データを収集し、ＴＯＰＴＲＡＮに学習させることが、容易にできる。その結果、それぞれのシステムを、それぞれの分野の翻訳に強い、専門のシステムにすることができる。即ち、Ａシステムは化学に強く、Ｂシステムは情報処理に強い、等のように、システムを専門化させることができる。

６．２．２ ＴＯＰＴＲＡＮネットワーク

ＴＯＰＴＲＡＮは、辞書及び対訳データを再利用可能な形で蓄積したもの（以後、知識データと呼ぶ）を、他のＴＯＴＲＡＮに送信することができると共に、受信側のＴＯＰＴＲＡＮで、その受信した知識データを必要なものだけ取り込んで、知識データの学習を行うことができる。（「図２ 知識データの送受信」参照）

この機能により、ＴＯＰＴＲＡＮでは、複数のシステムで独立に、異なる分野の対訳データの登録・学習を行い、それぞれのシステムの知識データを、一台の知識データ・サーバー的なシステムに集約し、一つの巨大な知識データとすることが

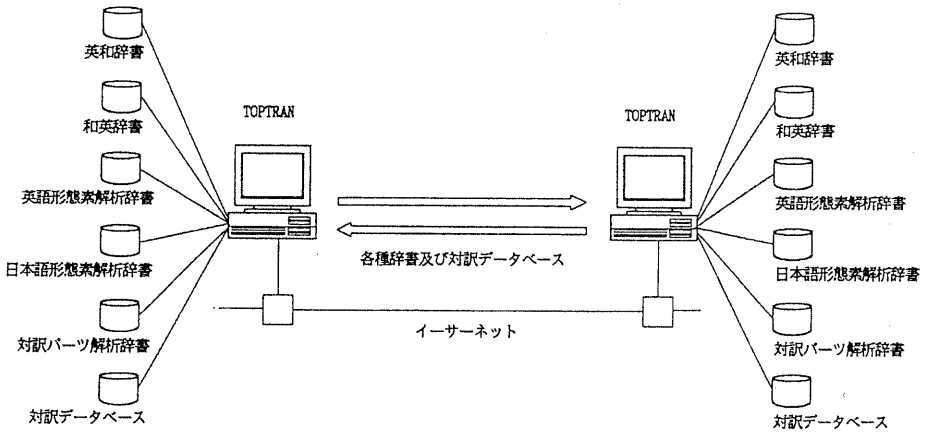

図2 知識データの送受信

できる。

7．今後の課題

対訳データをシステムに学習させる過程で、一番コストがかかるのは、パーツ・リンクの修正作業である。この部分は、過去の学習結果から推論して、自動的にパーツ・リンクを行うようになっているが、まだまだ不十分で、システムがよく間違えてリンクする等、改良の余地が残っている。自動パーツ・リンクのアルゴリズムを改良して、もっとオペレータの負荷を減らすことが、今後の課題である。

参考文献

[1] 仁井正治：ユーザー主導型機械翻訳システムＴＯＰＴＲＡＮ、情報処理学会自然言語処理研究報告 100-11 (1994)

[2]仁井正治：ユーザー主導の機械翻訳、ＡＡＭＴ Journal、No.7 June 1994

[3] 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

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

[5] 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

[6] 佐藤理史：ＭＢＴ１：実例に基づく訳語選択、人工知能学会誌、Vol.6 No.4 1991

[7] 佐藤理史：ＭＢＴ２：実例に基づく翻訳における複数翻訳例の組み合わせ利用、人工知能学会誌、Vol.6 No.6 1991

[8] 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

[9] 佐藤理史：実例に基づく翻訳、情報処理(Vol.33 No.6 June 1992)