自然言語処理 137－2
（2000．6．1）

機械翻訳システムの今後について

田中康仁
兵庫大学
E－mail:yasuhito@humans－kc. hyogo－dai. ac. jp

概　　要：　機械翻訳システムの本格的開発が行われ始めて約２０年が過ぎた。この間に多くの機械翻訳システムが作られている。ここでは今後の機械翻訳システムの今後の方向について述べる。

キーワード：　機械翻訳、自然言語処理、評価

Future Direction of Machine Translation System Development

Yasuhito Tanaka
Hyogo University
E－mail : yasuhito@humans－kc. hyogo－dai. ac. jp

Abstruct :　*Twenty years have passed since full-fledged machine translation (MT) system development began, and numerous MT systems have been developed in the intervening years. This paper focuses on the future, and discusses likely future directions of MT system development.*

Keyword :　*Machine Translation System, Natural Language Processing Evaluation*

〔０〕はじめに

機械翻訳システムの本格的開発が行われ始めて約２０年が過ぎた。この間に多くの機械翻訳システムが作られている。ここでは今後の機械翻訳システムの今後の方向について述べてみたい。

〔１〕機械翻訳システムの現状

機械翻訳システムを開発している企業は約２０数社ある。またその製品数は約５０個ほどである。そのほとんどのシステムが英語→日本語　日本語→英語である。
日本語→韓国語　韓国語→日本語や日本語→中国語
中国語→日本語も数社で製品化されている。英語、韓国語、中国語以外の言語に対応しているものは数が少ない。それも英語を仲介とした機械翻訳システムである。開発期間の長い英語→日本語への機械翻訳システムでも約７０％程度の翻訳精度である。

機械翻訳のシステムの評価

日本電子化辞書研究所で開発したＥＤＲコーパス（英文）を用いて（２単語～７単語のデータ）評価した。
評価結果は次の通りである。

表１　評価結果　Ａ社の英日翻訳システム

	5	4	3	2	1	合　計	平均値
2単語文	15	1	1	1	1	19	4.47
3単語文	336	97	16	11	0	460	4.65
4単語文	1,369	336	157	25	2	1,889	4.61
5単語文	3,655	808	510	53	4	5,030	4.60
6単語文	4,742	1,379	595	81	1	6,798	5.58
7単語文	4,471	2,331	870	118	1	7,791	4.43
合　計	14,588	4,952	2,149	289	9	21,987	4.53
%	66.35	22.52	9.77	1.31	0.04	100%	

評点の高いものが良い結果を示す。

このような状況を概観して次の２つの大きな問題点がある。

（ｉ）既に開発されている機械翻訳システムの精度を向上させるにはどのようにすればよいか?

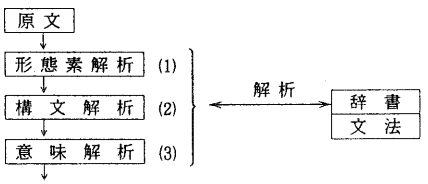
（ii）まだ研究されていない言語についてどのように対応してゆくか?

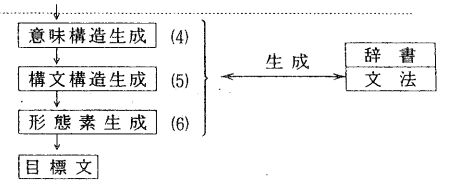
これら２つの問題点について考えてみる。

〔２〕機械翻訳システムの問題点と精度向上の意義

(1) 機械翻訳の構造上の問題点

・機械翻訳システムは次のような構成要素から成り立っている。





おおまかに6つの構成要素から成り立っている。これらの6つの要素の各信頼性をP_1 , P_2 , P_3 , P_4 , P_5 , P_6 とする。

原文が目標文まで正確に変換できる信頼性は

$P=P_1 \cdot P_2 \cdot P_3 \cdot P_4 \cdot P_5 \cdot P_6$　　である。

ここで〔１〕で評価した結果をあてはめると

$0.70=P_1 \cdot P_2 \cdot P_3 \cdot P_4 \cdot P_5 \cdot P_6$　　となる。

P_1 ,P_2 ,P_3 ,P_4 ,P_5 ,P_6 の各確率が個別に測定できないので6つの要素が同じ信頼性と仮定する。

$P_1 = P_2 = P_3 = P_4 = P_5 = P_6$

$0.70 = P_1^{\,6}$

$P_1 = \sqrt[6]{0.70} \fallingdotseq 0.942286$

つまり個々の信頼性は平均0.94、％で表示すると94％程度である。これは個々の構成の信頼性はそのプログラムと使用されている文法、辞書によるものである。全体の信頼性を上げるには個々の要素の信頼性を上げなければならない。

実際には個々の確率は異なっている。開発者はどの確率が低いかよく知っている。

低いところを上げる努力をすれば効果的に全体が良くなるのである。しかも、開発費用や期間のかからないものが、改良の最優先順位になる。

・各構成要素について

各構成要素は語で構成されている。

日本語や中国語、タイ語等の場合は語で構成されているが、膠着語であるため単語分割という作業が入る。
各単語の持つ信頼性をP_{ij} で表す。iは各構成要素とし、jはその構成要素でのj番目の語の要素とする。

各要素が信頼性0.94とすると

$0.94=\prod_{j=1} P_{ij}$ となる。

各単語も信頼性を向上させねばならない。

(2)　機械翻訳システムの問題点と精度向上について具体的に幾つかの内容を示す。

どのようにして信頼性を向上させるか?

(2－1)　単語レベルで考えられること。

（ｉ）未知語を減らして辞書を充実する。

このためには膨大なデータ、コーパスを分析し、単語の収集を行うことが必要である。新聞、www のホームページ、各業界の雑誌、コーパス等を分析する。

（ii）複合語と専門用語を集める。

複合語と専門用語は構成している語の結合で訳語生成できると考えられているが、必ずしもそうではない。

限界利益	marginal gain
限界容量	limiting capacity
価格限界	price limit

対象限界　　bound of object
能力限界　　limit of ability
規格限界　　bound of standard

このため複合語や専門用語を辞書に積極的に追加すべきである。

日本語の場合、多くの国語学者は短単位で語を区切るが、機械可読辞書の場合には、辞書の容量の制限を気にすることなく長単位語をもっと多く取り入れるべきである。このようにすることが曖昧さを減らす方法である。

長単位語を採用することは文を構成している語数を減らすことでもある。

(2－2) 句を集める。

英文では、句は動詞句、副詞句、前置詞句、名詞句等がある。句は一つの意味の集合体である。この句を集めることは文の構成要素を減らすものでもある。

例として［前］at home　家で、家に
　　［名］hope of succeeding　成功の見込み

しかし、長単位語を集めればよいといえない場合もある。このことに注意しなければならない。日本語の場合、長単位語が名詞として使われたり、サ変動詞として使われる場合もある。英語の場合の有名な例外を示す。

① It rains cats and dogs.
② I have cats and dogs.
"cats and dogs"は"rain"と共起する場合と"have"と共起する場合では意味が異なる。このような使用上の差を辞書に印を付けておかねばならない。

(2－3) 並列句を集める。

andやorで結ばれた並列語を集める。英語圏でのandやorで結ばれた慣用表現を日本語に翻訳した場合、訳語の語順が異なる場合がある。

A and B　→　BとA、AとB
A or B　→　B又はA、A又はB

例えば　ladyies gentlemen　→　紳士淑女
又　A and B Cの場合　→ AC and BCの場合と(A) and (B C)の場合がある。
並列処理は規則で処理する方法を採用している機械翻訳システムがほとんどあるが、今後は辞書に登録する方法を取る方が曖昧さが減る。

(2－4) 文法について

(ⅰ) 文法解析にあたって

文を単文、複文、重文にわける。さらに平称文、疑問文、感嘆文に区別して、英語の構文木が日本語の構文木にどのように対応しているか調べる。これを調べるにあたっては、文を構成している個々の動詞ごとに調べることが重要である。構文木といっても単語レベルの構文木が良いのか、句レベルの構文木が良いかも調べなければならない。このためには二言語間の構文上の曖昧性を除去したトリー・バンクが必要である。しかも、多量に必要である。

二つの言語の大量の構文木のデータ・バンクを作り、それを文の種類と動詞により仕分けし、言語間の変換のための方法を作り出す。この一連の作業を簡単にすることを見つけ出したという研究はほとんどない。今後は、二言語間の構文木データを蓄積し、言語変換の文法を実現すべきである。

(ⅱ) 結合価文法データの収集

動詞ごとの結合価文法を多量に集めるべきである。これは意味解析の一つである。

例　［A］provide［B］for（to）［C］
→［A］が［C］に［B］を与える。
［A］:［C］は人間　［B］:は物

(ⅲ) 動詞の重要度別のランク付け

日本語、英語とも動詞の数は大変多い。
日本語も英語もある特定の動詞と名詞が結合し、動詞化するものもある。

日本語：結婚㊞する→結婚する
英　語：make a sign→サインする

工業英語文と普通の英語文では少し動詞の使われ方に差があるが、動詞を重要度に応じてA,B,C,にランクを付けるべきである。ランクA,Bの動詞については色々な辞書を調べ、用法、結合価文法、その他を調べるべきである。又、慣用表現辞書等も調べる必要がある。又、英語、日本語の対訳付きコーパスからも用法を調べるべきであろう。
数人程度で系統的に調べれば数年程度で調べられる。

ランクCの動詞についても同じようにすべきであるが、これは費用と労力に依存する。

(2－5) 文について

機械翻訳システムでは文と訳文をそのまま持ち、文の照合により訳文を出力するシステムが実用化されようとしている。これは複雑な解析と生成を省くものである。このため次のことを提案する。

(ⅰ) マルチリンガル・コーパスの作成

膨大なマルチリンガル又はバイリンガルのコーパスを用意すべきである。

分野も色々な分野を含むものでなければならない。

例 good morning→おはようございます
How do you do?→ごきげんいかがですか

このような文は解析してもあまり意味がない。英文ではa, theを省いて10単語までの文と慣用表現文(ことわざ等)を集めるべきである。

(ⅱ) 定形文、又は、定形句の抽出

ある特定分野で特に使われる定形的な表現形式がある。これを集めうまく使うと質の良い訳文が簡単に生成することができる。

ニュース文や気象の文等で実用化されている。

離散形ｎグラムという簡単な方法で定形文を抽出することができる。

数字部分を〔Num〕と抽象化するのも一つの方法である。

例　There are 2 books.
↓
There are 〔Num〕books.

このほかには
地名　→［Loc］

個人名→［PN］
又は
［PN－1］、［PN－2］……
同一文に二つ以上の個人名がある場合
［PN－m］、［PN－f］
個人名が男性、女性を区別する場合
時間　→［Time］
日付　→［YMD］
等をもうけ、定形文を変形し、より一般的にする。

（ⅲ）文法や用語の解析用のマルチリンガル・コーパス

マルチリンガル・コーパスが望ましいが、バイリンガル・コーパスを大量に収集すべきである。これらは形態素解析、統語解析が済み、曖昧さが除去されたものが望ましい。普通の文、工業文もあれば最も良い。

（ⅳ）シソーラスを用いて文を変形する。

文の構成している一部分の語を類似している他の語と置き換える方法がある。

類似文を見つけだし、その語の一部を変えるものである。

He is a［ teacher ］.→彼は［ 先生 ］である。
↓
He is a student.→　彼は学生である。

このためにも多量のバイリンガル・コーパスを集める必要がある。そしてバイリンガル・コーパスを使用することを考えなければならない。

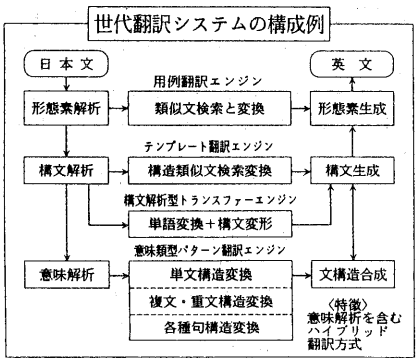
これまでに述べた各種のことを総まとめすると、次のようになる。

（ⅰ）知識データの効率的抽出
（ⅱ）辞書の充実
（ⅲ）文法の大系化
（ⅳ）各種大量のマルチリンガル・コーパスの充実

これらの問題は今後も引続き大きな課題である。

3）機械翻訳システムの方式の全面的な再考察

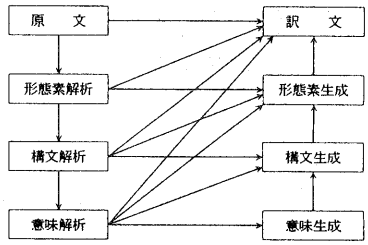
今までの機械翻訳は構文解析を中心としたものにコーパス・ベースを追加したものであるが、我々の頭の中ではこれらのものが一体化されてできているようである。鳥取大学の池原教授は次のようなハイブリート型を提案している。



この方式も一つの方式であるが、私はもっと色々なことが考えられる。例えば日本文から直接英文を生成してもよい。日本文をkeyとして直接一致する英文引き出し、英文を出力してもよい。又日本語の意味解析まで行っていきなり英文の形態素生成にいき英文を生成してもよい。

このように色々な可能性を考えて実際のプログラムの作成、テーブル、ファイルを作成を考えていけばよい。我々の頭の中では池原の提案しているような新しいシステムよりもっと柔軟な処理が行われているはずである。

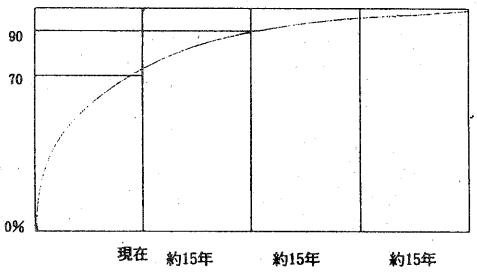
筆者の推薦する翻訳システムの構成例



この原稿が出来あがり、ある国際会議に投稿し1999年11月1日～3日に中国の北京市で行われた全国五届計算語言学联合学術会議（JSCL－99）に出席し同じ図が書かれている論文を見つけた。この内容は発表はされなかった。

この論文には以前に独語→中国語，ロシア語→中国語機械翻訳の研究者達もこのような考えに達しているようであると書かれている。

（4）機械翻訳システムは次のような曲線を描きながら徐々に向上する。



今後１０～１５年間開発努力が行われ、機械翻訳の知識データが充実すれば９０％ぐらいに達するであろう。さらに１０～１５年間かければ９５～９５％ぐらいにまで達するであろう。ある特定分野に限ればもっと早く実現されるかもしれない。

最近本の電子出版化が急速に進んでいる。電子化された内容を再編集するとか、辞書に応用して機械翻訳用の専門用語辞書が簡単に作られる等といった動きが起こっている。個別に開発された本やＣＤ－ＲＯＭが複数個集まり、さらに集約された機械可読の知識へと発展するであろう。

機械可読の知識は、コンピュータ上で容易に統合編集

され、新しい集合体となった知識へと発展するであろう。そのため、今、予測する以上の早さで精度の向上がなされるかもしれない。人々は、便利だと感じると増々使い、さらに便利なものとして使いはじめる。

今後パーソナル・コンピュータはもっと早くなるであろう。そうなれば処理時間は短縮され、もっと大量のデータや知識データを容易に取り扱うことが可能になる。こうなれば機械翻訳システムの精度向上にもつながるであろう。

テスト・データを一挙に２０倍に拡大すれば、約１５年としている期間はもっと縮まるかもしれない。同時に辞書データも１年に改良する量の２０倍程度を投入すれば翻訳精度が向上する。

機械翻訳システムの精度も年々良くなっていくであろうから、誤りデータを一定量得るためには、今までよりも多くテスト・データも増さねばならないし、不足しているかを調べるのが困難にもなっていく。

今、機械翻訳システムは７０％程度の精度であると言われている。そして人手による翻訳の３倍程度の能率が上がる。それでは機械翻訳の精度が上がると、どのようになるのであろうか？

７０％の精度ということは３０％については手作業で修正作業をほどこさなければならないということである。人手による翻訳の３倍の能率という意味が分かる。
同じように考えると

ｑを翻訳精度とする。０≦ｑ≦１
ｑ／（１－ｑ）値が人手による作業より何倍の能率が上がるかを示す値である。

精度　７０％→人手の３倍の能率
８０％→人手の４倍の能率
９０％→人手の10倍の能率
９５％→人手の20倍の能率

これは単純にこのように表示したが、各精度に達するまでには色々な便利な道具や辞書が用意されるであろうから、もっと早い時期に能率が上がるかも知れない。
早く９５％程度の精度にまで向上させたいものである。

〔３〕既に開発された機械翻訳システム以外の言語をどう取り扱うか。

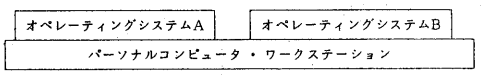
既に開発された機械翻訳システム、英語、中国語、韓国語以外の言語、例えばアラビア語、ポーランド語、モンゴル語等をどのように取り扱えば良いか考えてみよう。

どの言語を考えるかは、その言語がどの程度日本と関係があるかという社会的意義を考えなければならない。しかし、現実には留学生がいるとか、身近にその言語をサポートしてくれる人がいるということが現実的な動機になっている。

まず考えなければならない項目をあげてみる。

（ｉ）ＯＳについて

我々はパーソナル・コンピュータやワーク・ステイションを毎日使っている。しかし、国が異なるとこのＯＳも異なっている。同じWindows'98でも日本では日本用の局所的最適化が行われている。



Ａ国語を使う人々のWindows'98はそれに対応する様に変更されている。それ故、日本語を使えるWindows'98ではＡ国語が取り扱えなければならない。

（ⅱ）入力方式、キーボード、文字フォントについて

ＯＳをどのような環境で使うかということが大きな問題である。次の問題としては、Ａ国語の文字入力をどのようにして行うかということである。

1）　Ａ国語の入力方式はＡ国で標準的に定められているか？　それは日本語の入力と共存できるか？

2）　Ａ国語の入力のためのキーボードは標準的なものとして定められているか、Ａ国語としても使え、日本語のキーボードとしても使えるか？

3）　Ａ国語の文字の表示はうまく表示されるか？

（ⅲ）エディターについて

日本語の表示、Ａ国語の表示がスクリーン上でうまく行えるか日本語とＡ国語を標準的に取り扱うエディターがあるか？

という問題がある。

これは標準化されているか等がある。文字は左から右へ書くことが多いが、アラビア語は右から左へ書く。このように言語が変わるとエディターも大変機能が変わってくる。

（ⅳ）辞書の作成

Ａ国語を機械処理しようとすると、まず機械可読の辞書を作らなければならない。これもＡ国語の辞書と日本語とＡ国語の対になったものが必要である。辞書には一般用語の辞書と分野別の専門用語が必要である。

（ⅴ）コーパスの作成

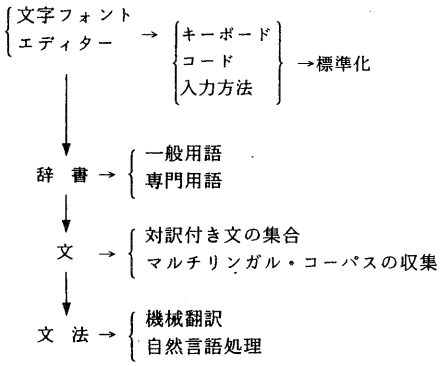
辞書と同じように、文の集合体を集め作成しなければならない。語は単独に使われるのではなく、文の中で文の構成要素として使われるものである。それ故、どのように使われているか、働いているかという実例が必要なのである。ただ単純に１つの言語だけでなく対訳付きの文の集合体が必要である。

（ⅵ）文法

Ａ国語についての自然言語処理を行うとすると、Ａ国語の文法が記号化され、体系化されていなければならない。また文法が正しいという検証もなされなければならない。これもまた大変な作業である。

このようにして道具がそろい機械による言語が取扱い可能になるのである。

以上をまとめると次の図のようなプロセスになる。



〔４〕今後、我々の進む方向について

我々は日本語、英語を中心とした機械翻訳の精度を向上させるために研究を進める方向と、異なった言語を処理するための基礎的な研究、実験、モデル作りの方向がある。この両方を進めてゆかなければならない。世界には色々な言語があり、民族がある。一つの道具日本語と英語だけでは世界に対応することはできない。

早稲田大学にはICMTP（International Conference on Multilingual Text Processing）という学会がある。
URL:http://www.mling.waseda.ac.jp/icmtp/
Email:icmtp@mling.waseda.ac.jp
を参照されたい。

多様な発展方向をもつことが、今後日本が発展し、国際社会の中で生きていく方向ではないだろうか。このようにすることが、国際平和を実現する一つである。

〔５〕おわりに

我々はＥ－mailの発達により世界は縮まってきた。しかし、通信回線上を流れる言語は色々である。これらの情報をうまく変換し、我々に便利な道具の開発も重要なテーマである。ここではその道具の開発について考えてみた。

〔６〕参考文献

(1) ㈱ 日本電子工業振興協会
「自然言語処理システムの動向に関する調査報告書」
平成9年4月
(2) 牧野武則　評価技術
「機械翻訳」Bit別冊　共立出版　1988年9月
(3) 長尾　真　「機械翻訳はどこまで可能か」
岩波出版　1986年6月
(4) Language and Machines:Computers in Tlanslation and Linguistics, National Academy of Sciences National Research Council (1966)
(5) Annual Report of Ikehara Laboratory 1998
Natural Language Processing, Vol.3
Tottori University, Faculty of Engineering
Division of Information and Knowledge Engineering
(Ikehara Laboratory)
(6) 伝愛平　英漢機械翻訳的転換生成策略
計算語言学文集（JSCL－99　Proceeding）
PP341～346　清華大学出版社　1999年10月
(7) 邸海燕　柴佩琪　許玉祥　基于双語語料庫和規則庫的徳漢復合句的転換生成，計算語言学発展与応用，清華大学出版社，1995　264－270
(8) 李向東　周清波等　智能型俄漢機器翻訳系統的句法規則庫的設計原則，1999，中文信息学扱，Vol. 13, №1　16－19

〔７〕データについて

英文データは日本電子化辞書㈱の英文コーパスを利用した。

この論文の翻訳基準について

理解容易性の詳細な内容については参考文献(3)P55を参照されたい。

この原文は　Language and Machines: Computer in Translation and Linguistics, National Academy of Sciences-National Research Council (1966)．(ALPACレポート）の中にある。