自然言語処理 86－9
（1991. 11. 15）

後編集結果を利用する機械翻訳システム

伊藤悦雄・武田公人・天野真家

(株)東芝 総合研究所

機械翻訳システムの一次出力に対する後編集は翻訳作業の中で大きな割合を占める。一度行った後編集作業の結果を保存し、再利用することができれば機械翻訳システムの運用効率を大幅に向上できる。本稿では、以前翻訳した文書と同一の文、あるいは類似する文を翻訳する場合、以前の翻訳結果を利用することによって、後編集作業を軽減する機械翻訳システムについて述べる。

MACHINE TRANSLATION SYSTEM UTILIZING EXISTING TRANSLATED DOCUMENTS

Etsuo Itoh　Kimihito Takeda　Shin-ya Amano

TOSHIBA　R & D Center
Information Systems Lab.

1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 210, Japan

Post-editing the primary output of a machine-translation system covers much of translation work. Saving the result of post edition for later use can significantly upgrade the overall efficiency of a MT system. This paper describes a MT system which reduces the burden of post-edition when translating sentences identical on similar to those which have been already translated by utilizing existing translated documents.

1. はじめに

現在開発されている機械翻訳システムは、自動翻訳ではなく、人間の翻訳作業を支援する「翻訳支援システム」である。このため、品質の良い翻訳結果を得るためには、機械が翻訳した結果に人間が介入して修正を加える作業(ポストエディット)が不可欠である。このポストエディットに要する時間は翻訳全体の時間の内で非常に大きな割合を占める。従って、ポストエディットに要する時間の短縮は機械翻訳システムの導入効果を高めるための重要な課題である。

ポストエディットを短縮するためには、編集機能などポストエディット支援機能を強化することも有効である。しかし、本質的な解決には、翻訳品質を向上させ、機械翻訳による一次出力と最終目標文との差を小さくすることが必要である。

翻訳品質を向上させるために、機械翻訳で使用する辞書・文法の改良が行われてる。この改良には多くの時間と労力を要するという問題がある。もう一つの翻訳品質向上のアプローチとして類例利用がある。このアプローチでは「Example Based Machine Translation」の研究が行われている。しかし、まだ研究段階にあり、実用化までには解決すべき課題が多い。

しかし、実用システムにおける類例利用では、比較的簡単な方式を用いても非常に大きな効果を上げる場合がある。例えば、マニュアルのバージョンアップや、非常に類似した表現を多用する場合などである。この様な場合では、同じユーザが行った翻訳を利用することにより、ポストエディットの結果に反映されている知識のフィードバックすることができ、翻訳品質の向上が期待できる。

この様な観点から、我々は、参照する文として同じユーザがすでに翻訳を行った文書を利用することによって、検索・参照した文をそのまま利用する方法を開発した。

この方法には以下の利点がある。

(1) ユーザは意識してデータベースを作る必要がない
(2) 電子化辞書、用例集等の汎用データベースと違い、ユーザ特有の表現・訳語等が利用できる
(3) 翻訳で得られた品詞情報などを文と共に記憶し、検索条件に利用することで柔軟な検索が可能となる

本稿では文書の管理機能を強化することにより、以前の翻訳結果の参照・利用を可能にし、ポストエディットの結果をその後の翻訳結果として出力できる機械翻訳システムについて述べる。

2.機械翻訳への要求

上述の様に文章の翻訳時に従来の翻訳結果を参照することは重要である。しかし、機械翻訳システムが実用化され、翻訳された文書量が大量になっても、これを文書データベースとして利用する方法が確立されていなかったため、従来の翻訳システムでは、入力文書は機械翻訳システムにとってはそれぞれ独立した文書として扱われていた。

従って、マニュアルのバージョン・アップなどのように、ポストエディットまで終了している文書の一部が修正された場合には、以前に行ったポストエディットが修正された文書の翻訳結果に全く生かされなかった。このため修正後の文書の翻訳結果の全体に渡って、再度ポストエディットをすることが必要である。

更に、既に翻訳されている文書とまったく等しい部分が修正された文書に多く含まれていても、その部分も含めて文書全体を翻訳しなおすこととなり、翻訳に多くの時間を要すという問題がある。

以上の様に、部分的に修正された文書の翻訳を行う際、既に存在するポストエディット済みの文書を利用することによって、

(1) 以前の翻訳結果を利用した部分へのポストエディットが不要になり、作業の軽減が図れ

る。

(2) 以前の翻訳結果を利用した部分は、ポストエディットの結果を取り入れているので、オペレータの知識などを反映できる。

(3) 文書全体でなく一部分のみに対し翻訳処理を行うため、翻訳処理時間の短縮が可能である

という利点が生じる。

ところが、従来のシステムでこの機能を実現させるには、

(1) 修正情報を管理し

(2) 修正された部分のみに対し機械翻訳を行い

(3)その結果を以前の翻訳結果と差替える

という作業をすべてユーザが行わなければならない。しかし、細かい修正が広範囲に渡って行なわれた場合や、文書作成者と翻訳者が異なる場合などには修正情報のユーザによる管理が困難である。

そこで、以前の翻訳結果の内、利用可能な部分を抽出し、過去の翻訳結果を再利用する機能を提供する。この際、この機能を有効に動作させるために以下を前提として導入する。

[前提1]

参照する文書は翻訳する文書と大幅には異ならない。

この前提は、新規に翻訳する文書と大幅に異なる文を参照しても利用可能な文が少なく、効率の向上が図れないことに基づく。また、実際の文書データにおいても、プログラムのマイナーチェンジに伴うマニュアルの更新などにおいては全体の1割程度しか修正されないことから、この前提は適当であるといえる。

3.同一文における後編集結果の再利用

新規に翻訳する文と同一な文を以前に翻訳している場合にはその文の翻訳結果を利用できると考え、同一文を他の文書から効率良く検索する方式を考案した。ただし、同一文であっても、文脈によって翻訳結果が異なる場合がある。しかし、前章で述べた前提1によって、参照する文書と翻訳する文書では内容が類似するので同一文の翻訳の仕方があまり変わらないということが仮定できる。

3.1 文単位の再利用

ポストエディットが行われた結果を利用する場合、利用することによる翻訳作業の高効率化を最大にするために、利用できるか否かの判別を文単位で行うこととした。その理由は文の独立性である。現在の翻訳システムでは、文脈情報を翻訳に利用することは困難なため、翻訳は文単位に行われる。このため、翻訳結果は文単位に独立し、文単位の入れ替えが可能である。

また、文単位の比較を行うために、次の前提を導入する。

[前提2]

以前機械翻訳を行った文書は、翻訳過程において文単位に分割され、しかも、文単位で原文-訳文の対応付けが成されている.

従来の機械翻訳システムではこの機能を実現している[1]。

文を単位として既存の文書を再利用して翻訳を行った場合には、その翻訳結果として、既存文書から検索された文を再利用した訳文と新規に機械翻訳された訳文とが混在した文書が得られる。このような翻訳文書を効率良く編集するために、再利用した訳文と新規翻訳した訳文を区別してユーザに明示する機能も不可欠である。通常、新規翻訳した部分に対してのみポストエディットが必要であるため、ユーザはその場所を文章中から検索する必要があるためである。

3.2 利用可能箇所の判別

上記の方針に従い翻訳支援を行うためには、既に翻訳されていて参照される文書(以下参照文書)と新たに翻訳する文書(以下新規翻訳文書)との差、すなわち利用できる箇所を求める機能をいかに効率良く実現するかが重要である。

この機能を実現するため、簡易文抽出方式および文の高速比較方式を考案した。本節では、その詳細について、主に英文を例にとって述べる。

3.2.1 文抽出方式

通常、文書から文を切り出す際に、その精度を上げるためには辞書引き・形態素解析を必要とする。しかし、我々が開発した利用可能な箇所を抽出する機能の実現においては、処理速度の向上のために形態素解析を必要としない簡易文抽出方式を採用する。

簡易文抽出方式は下記の基準に基づいている。

<文の抽出基準>

- 文とは「標題」またはそれ以外の「通常文」である。
- 標題は、「前置詞や冠詞以外の単語の先頭文字が大文字で、かつ、次の行が空行であるか、次の行の先頭単語が大文字で始まる」という性質を持つ。このため、この性質を持つ行が出現した場合、その行は「標題」であり、その行末で文は終了する。
- 通常文の文末には文の終りを示す記号(.!?;:等)が存在し、その次の文字が空白または改行である。

この簡易文抽出を新規翻訳文書に対してのみ行い、参照文書については行う必要がない(前提2)。従って、新規翻訳文書と参照文書では異なった文抽出方式を採用することとなる。このため、文の抽出に差が生じる可能性がある。これに対しては、文同士の比較において、隣接する文を複合して一文としたものを一方の比較対象とすることにより文の抽出時における誤差を解決できる。

3.2.1 高速検索アルゴリズム

文同士の比較を効率よく行うためには、新規翻訳文書中のある文と比較する参照文書中の文をいかに無駄なく決定するかが鍵となる。

そこで、本項では、文書中の文に対して一括して利用できるか否かを判断するための高速アルゴリズムに付いて述べる。

(1) 文の比較

まず、文書の第一文同士、あるいは一致した文の次の文同士(新規翻訳文書の第N文と参照文書の第n文)を比較する。前提1によると、前の文同士が一致した場合は、次の文が一致する可能性が高い。一致した場合は翻訳結果を利用可能である。

一致しない時には以下の場合が考えられる。

a. N文の一部が修正された場合
b. N文が挿入された文である場合
c. 文が削除されている場合
d. 移動された文である場合

このうち、c・dの場合には文書中の他の位置にこの文と一致する文が存在する。ただし、この時、探索空間は文書全体となる。

他の場所に一致する文があるか否かの検索を高速に行うための手順を以下で述べる。

(2) 次の文の比較

新規翻訳文書のN文と参照文書のm文が一致している場合は以下の手順で次の文(第N+1文)との比較を次の手順で行う。

(a) 新規翻訳文書の第N+1文と参照文書の第m+1文と比較する。複数の文に渡って複写/移動がおこなわれた場合や文が削除されている場合には一致する。
(b) この比較で一致しない場合は、新規翻訳文書の第N+1文と参照文書の第n文と比較する。一致した場合は、一文のみに複写/移動がおこなわれた場合である。
(c) 上記の二文との比較において、どちらも一致しない場合は第N+1文に対しても修正が加えられている場合である。この第N+1文に対しては、第N文に行った処理と同様の検索・比較を行う。

一方、新規翻訳文書のN文と同一な文が発見できなかった場合には、次の文の比較を次の手順で行う。

(a')新規翻訳文書の第N+1文と参照文書の第n文と比較する。一致した場合は、N文が挿入された文である場合である。

(b')この比較で一致しない場合は、新規翻訳文書の第N+1文と参照文書の第n+1文と比較する。一致した場合は、N文中の一部が修正された場合である。

(c')上記の二文との比較において、どちらも一致しない場合は、複数の文に渡って修正された場合である。この第N+1文に対しては、第N文に行った処理と同様の検索・比較を行う。

上記の検索方式を採用することにより、文書の比較のための探索空間を大幅に減少することができる。すなわち、文書全体が探索空間となるのは不一致文が出現した時である。その場合も、複数の文に渡り連続して修正が加えられている場合(CとC')以外は、次の文の探索空間は2文と非常に狭い範囲の探索で一致文を発見することができる。

さらに、文書の修正情報を利用することにより一層高効率の検索が可能である。すなわち、参照文書の原文のどの部分にどのような修正を加えたかという情報を利用し、検索範囲を限定するのである。

たとえば、ある文に修正が加えられたという情報がある場合には、一致する文が存在しないことが検索前に解る。このため、この文に対する比較・検索を行う必要がなくなる。また、移動・複写の情報を利用することによりその文と一致する翻訳済文を検索を行うことなく発見できる。

3.2.2 文の検索方式

(1) 文末の単語による一次検索

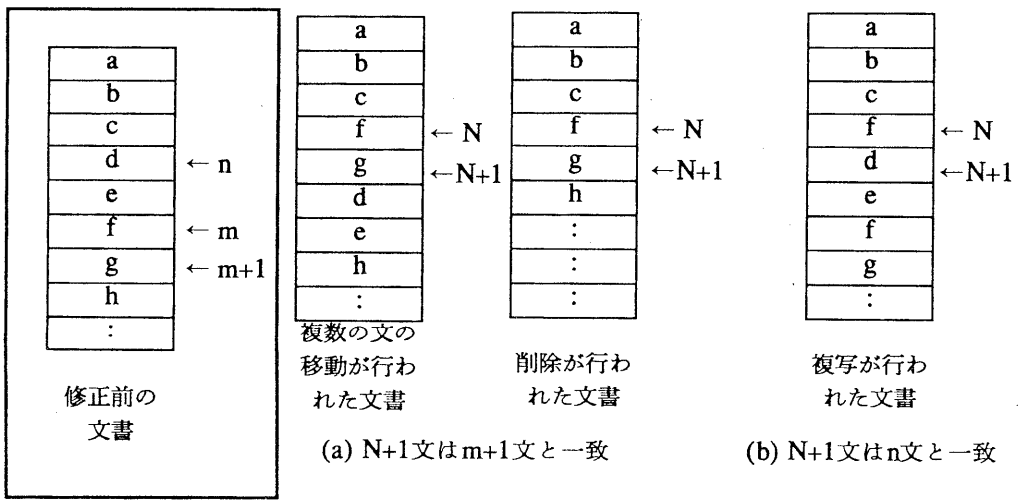
上記の手順を用いても、参照文書中の利用できる文の有無の検索のための探索空間が文書全体となる場合がある。このため、文書サイズが大きくなると検索時間が増加する。この検索を高速に行うために文末の二単語をキーとして一次検索を行う。

この検索方式は、以下の2点を根拠としている。

(a) これらの単語は文中からの抽出が容易である

(b) これらの単語によって文の限定が可能である

文からの抽出の容易さは文頭の単語が第一であり、次いで文末の単語となる。しかし、文



第一図 N文とm文が一致した場合

頭に出現する単語は、副詞・冠詞・代名詞・接続詞等であり、文を限定するための形態的な情報に乏しい。これに対し、文末の単語は一致する場合が少ないという性質がある。この性質を下記の通り検証した。

文書中の任意の文の文末の一単語および二単語を取り出した時、その単語で終了する文が文書中に(その文も含めて)何文存在するかを調査した。その結果を第一表に示す。この表から明らかなように約70%の単語は一文のみに出現する。つまり、文末の1単語によって70%の文は1文に限定できる。また、文末の二単語によって1文に限定できる割り合いは80%以上になる。更に、文末の二単語によって2文以下に限定できる割合は90%以上に上る。

しかし、「同一文書中において同一単語で終了する文は少ない」という性質に反し、文末に出現する頻度の大きい単語もまれに存在する。この単語はその文書のキーワードである。キーワードは文章中への出現頻度が高いため、文末への出現頻度も高くなっている。

たとえば、第一表の文書3において文末の出現頻度がもっとも多い単語は "network"である。この単語は63文の文末に出現している。この文書は、コンピュータネットワークの解説書であり、"network"はこの文書のキーワードである。

従って、キーワードが文末にある場合は探索空間の減少割合が低いが、それでも文末の2単語を用いる事によって文書中の2%(882文中21文)にまで減少させることができる。

(2) 検索テーブルによる検索

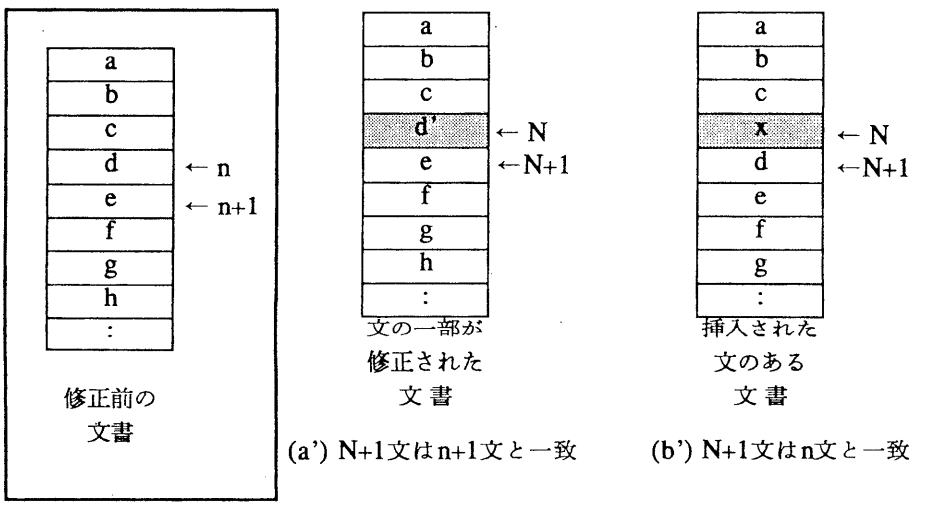
上記の文末の単語による一次検索を高速に行う手段として検索テーブルを使用する。

この検索テーブルは「文末の単語と文の番号」の対応を示すものである。このテーブルは参照文書を検索する際、文にアクセスする毎にその文の情報をテーブルに記録する方法で作成する。しかし、参照文書は既に翻訳処理時にすべての文に対し、文の抽出が行なわれている。このため、翻訳処理時に「文末単語 - 文番号テーブル」を作成することにより、一層高速な検索も可能となる。

このテーブルは検索の高速性を重視し、文末の単語は単語全体を1バイトに圧縮した形式で記録し、これを用いて一次検索を行う。これは、文字列比較は比較的時間がかかる処理であるため、極力それを避けるためである。

4. 修正情報抽出機能の拡張

上記の機能の導入により機械翻訳の性能は大幅に向上する。しかし、実際の文書の翻訳時



第二図 N文と一致する文がない場合

には、文中の一部のみが異なる文も多く出現する。このような類似する文はすべての文を翻訳し後編集するより、代表的な文を翻訳し、後編集した結果を複写・修正した方が効率良い。

このような場合に以前の翻訳結果を利用するためには、上述の方法は使用できない。そこで、文の一部が異なるような文(類似文)を使用できるように機能を拡張する。

4.1 文の類似性

類似文には様々な定義がある[14][15]。しかし、本方式では類似文として検索された文を極力そのまま翻訳に利用することを前提とし、従来とは異なる定義を以下の通り提案する[16]。

[類似文]

以下の条件を満たすものを類似文とする。ただし、(a)から順に類似性が高い文とする。

(a) 訳文の言語において出現しない、あるいは必須でない情報のみ相違する文

例: 冠詞、名詞の単数・複数のみ異なる文
異形語への置換のある文

(b) 翻訳に於ける意味変換に与える影響の少ない部分が相違する文

例: 時制が異なる文
複合語が代表する一語に置換された文
名詞が代名詞に置換された文

(c) 原文の意味の差の少ない文

例: 同一品詞の他の語に置換されている語がある文
修飾語の付加 / 欠落がある文

この定義に従って、検索を行うために以下に示す類似度判定条件を設定する。この条件を品詞毎に設定し、これを組合せることにより、上記の類似文の検索が可能である。

[類似度判定条件]

(A)完全一致 ..... 検索文に出現するすべての単語が出現し、それ以外の単語が出現しない。

(B)包含一致 ..... 検索文に出現するすべての単語が出現する。

(C)部分一致 ..... 検索文に出現する単語のうち一部の単語が出現する。出現の割り合いを%で指定する。

第一表 同一単語の文末での出現頻度

出現文数	文書1 : 996文 自然言語処理の論文		文書2 : 2291文 計算機マニュアル		文書3 :882文 ネットワーク解説書	
	1単語	2単語	1単語	2単語	1単語	2単語
1	444.(77)	764.(88)	340.(59)	1257.(80)	268.(67)	528.(81)
2	69.(11)	66.(8)	80.(14)	169.(11)	63.(16)	86.(13)
3	29.(5)	27.(3)	38.(6)	70.(4)	19.(5)	15.(2)
4	20.(3)	7.(1)	35.(6)	32.(2)	16.(4)	11.(4)
5	13.(3)	1.(0)	23.(4)	17.(1)	11.(3)	5.(0)
6	8.(1)	1.(0)	16.(3)	10.(1)	6.(1)	0.
7	3.(0)	1.(0)	16.(3)	10.(1)	5.(1)	3.(0)
8	4.(0)	0.	8.(1)	4.(0)	2.(0)	4.(0)
9	4.(0)	0.	6.(0)	4.(0)	2.(0)	0.
10	4.(0)	0.	3.(0)	1.(0)	1.(0)	0.
11~	3.(0)	0.	23.(4)	5.(0)	11.(3)	3.(0)
	(最大13文)		(最大33文)	(最大17文)	(最大63文)	(最大21文)

単位：単語数(括弧内は%)

(D)品詞一致 .....検索文に出現する単語の一部が同一品詞の他の語に置換されている。

(E)一致不要 .....一致の必要がない品詞に対して設定する。

また、単語の検索においては、動詞などの活用や名詞の単複変化などに対応して柔軟な検索を可能にするための「同一形」「原形」「その他の活用形」の順で検索を行う。また、必要に応じて語順一致条件を付加できる。

4.2 ユーザの指定による類似文の利用

上述の類似文検索では、柔軟な検索が可能である。しかし、マニュアルなどテクニカルライトされている文書では、類似する文同志の異なる単語を入れ替えることによって、翻訳結果がえられる場合が多い。例えば、「○○ボタンを押してから××ボタンを押す」の○○の部分のみ異なる、あるいは、数字部分のみが異なる場合などである。

この場合は上記の柔軟な類似文検索を行うより、むしろ、文中の単語の内、異なっても文意の変わらない単語を指定し、それ以外の部分が同一の文を類似文として検索したほうが効率が良い。

5. 実験および評価

5.1 同一な文のみを検索する場合

3節で述べた検索方式を用い、利用可能文抽出処理の速度評価テストを行った。このテストでは、簡略化のために文末の1 単語のみをキーとする方法を用いた。約1,000文からなる参照文書とそれに約一割の修正を加えた新規翻訳文書との比較を実施した結果、修正情報の抽出を2秒以内で行うことができた(EWS AS3260使用)。

5.2 類似文を検索する場合

次に類似文検索の実験結果について述べる。この評価実験では約1,600文の翻訳結果を検索対象とした。このデータに対する類似文の検索所用時間は検索条件によって異なるが、いずれも3秒以内である(EWS AS3260使用)。

また、検索条件を種々変更して行った結果、有効な検索を行うために類似度判定条件を以下の設定にする必要があることが判明した。

(1) 動詞は出現割合70%の部分一致より厳しい条件とする。

(2) 前置詞は出現割合50%の部分一致より厳しい条件とする。

(3) 英日翻訳に利用する際は語順を考慮し、日英翻訳に利用する際は語順は考慮しない。

6. おわりに

本論文では、以前の翻訳結果を用いることにより、後編集の結果を新規翻訳結果に反映できる高効率な機械翻訳システムについて述べた。

今後は、構文的な情報を利用する検索方式の検討を行う予定である。

参考文献

[1]安達他、日英相互翻訳システムのエディタ方式、情報処理学会34回全国大会、1988

[2]伊藤、長谷部、武田、天野：「過去の翻訳結果を利用した翻訳支援システム」、情報処理学会第38回全国大会 (1989)

[3]隅田、堤「構文の照合による柔軟なテキスト検索機能を備えた翻訳支援システム」情報処理学会第37回全国大会 (1988)

[4]佐藤、長尾「実例に基づいた翻訳」情報処理学会第38回全国大会 (1989)

[5]中村「用例検索翻訳支援システム」情報処理学会第38回全国大会 (1989)

[6]伊藤、長谷部、武田、天野：「類似文検索機能を備えた翻訳支援システム」、情報処理学会第39回全国大会 (1989)