自然言語処理 88−14
（1992. 3. 13）

単語接続情報を利用した機械翻訳

鈴木 康広　　栃内 香次

北海道工業大学　　北海道大学

近年、パーソナルコンピュータの性能向上に伴い、パーソナルコンピュータ上で稼動する機械翻訳システムが広く研究・開発されている。一般に、これらの手法は原言語の形態素解析や構文解析、さらには意味解析を行った後、その情報に基づいて目標言語に変換するものである。このような手法は、膨大で複雑な文法情報を利用するため、いくつかの問題を抱えている。これに対して、我々は単語と単語のつながり易さ（接続情報）を利用する機械翻訳手法を提案した[1)]。実際に、提案した手法に基づく実験システムをパーソナルコンピュータＰＣ９８０１上に構築し様々な実験を行い、その有効性を示したので報告する。

Machine Translation System Using Conjunctive Relation of Words

Yasuhiro SUZUKI　　Koji TOCHINAI

Hokkaidou Institute of Technology　　Hokkaido University

The recent development of computer hardware technology has stimulated active research in machine translation system on personal computer. We proposed a new method for machine translation. This method uses the conjunctive relation of words, instead of the syntactic and semantic analysis employed in the conventional machine translation system.

In this paper, we constructed experimental translation system on personal computer PC9801. We show some experimental results and validity in this method.

1．はじめに

現在一般に用いられている機械翻訳手法は、原言語の形態素解析、構文解析、さらには意味解析を行った後、その情報に基づいて目標言語に変換する方式である[2)]。我々は、これとは異なる手法として単語の接続情報を機械翻訳に応用する研究を行っている[1)]。

ある限られた分野の文献を対象とする場合、文章を構成するそれぞれの単語の間には特定の接続関係がある[3)]。例えば、「機械、システム、我々、翻訳、開発する」という単語群がある場合、これらの単語を並べ換えて生成することができる文章のうち、最も自然な文章は「我々（は）機械翻訳システム（を）開発する」というように１文に限定することができる。すなわち、ある単語に接続可能な単語は少数に限定されていると考えられる。従って、文章中で接続している２つの単語を接続情報として大量に抽出してあらかじめ辞書に登録しておき、英文を単語単位に翻訳して得られた訳語群から接続情報を用いて一意に翻訳文を生成することができる。

この方式の利点としては、第１に辞書の保守・管理の容易さが挙げられる。この方式では、文法情報をほとんど用いないため、辞書への登録情報数が少なく、その内容も単語に対する訳語と接続する訳語の組を登録するだけでよい。第２にシステム構築の容易さが挙げられる。本方式では、翻訳アルゴリズムが簡単であるため、システム構築を容易に行うことができ、高速な処理速度を得ることができる。第３に同じアルゴリズムで日英の翻訳が可能なことである。英文上での単語の接続情報を抽出し、辞書を構築することにより日英の翻訳が同様のアルゴリズムで可能となる。

本論文では、実際に上述の方式による実験システムを作成し性能評価実験を行った結果について報告する。

2．翻訳アルゴリズム

前述のように、文章中のある単語について、その単語に接続可能な単語は特定の単語に限ることができる。このことを利用して、以下に述べるような機械翻訳アルゴリズムが考えられる。

前述の、「機械、システム、我々、翻訳、開発する」という語群を考えてみる。この語群から生成することができる文章は、「我々は機械翻訳システムを開発する」という一文に定まる。これは、それぞれの単語が「Ｔ－我々、我々（は）－機械、機械－翻訳、翻訳－システム、システム（を）－開発する、開発する－Ｔ（Ｔは文頭または文末を示す）」という接続関係を持ち、文はこのような接続関係の連鎖で表されることを示している。この関係を接続情報と呼ぶことにする。そこで、以下に示すように英文を構成する各単語を単語単位に翻訳して得られた語群から、接続情報を用いて翻訳文を生成することができる。

例）

We develop machine translation system.
↓ ↓ ↓ ↓ ↓（英文）
我々 開発する 機械 翻訳 システム
（単語単位の翻訳）

Ｔ-我々、我々(は)-機械、機械-翻訳、翻訳-システム、システム(を)-開発する、開発する-Ｔ
（接続情報）

我々（は）機械翻訳システム（を）開発する
（日本語文）

ここで、接続情報はあらかじめ同一分野の大量の文献から抽出し、接続情報辞書に蓄えておくものとする。また、このアルゴリズムは日英翻訳についてもそのまま適用可能である。すなわち、英文についても同様に接続情報、「T-we,we-develop,develop-machine,machine-translation,translation-system,system-T」をあらかじめ接続情報辞書に登録しておくことによって、以下に示すように翻訳文が生成される。

例）

我々 機械 翻訳 システム 開発する
↓ ↓ ↓ ↓ ↓(日本語)
We machine translation system develop
（単語単位の翻訳）

T-we,we-develop,develop-machine,machine-translation,translation-system,system-T
（接続情報）

We develop machine translation system（英文）

３．実験システム

我々は、前述のアルゴリズムに基づき単語接続情報を利用した機械翻訳システムを作成した。以下、このシステムの概要と翻訳処理過程について述べる。

本システムはＮＥＣのパーソナルコンピュータＰＣ９８０１上にＣ言語で構築されている。なお、本システムは小規模な実験システムであり、翻訳対象文章は主として情報処理関係の論文、研究会報告等のアブストラクトに限定している。実験システムの処理の流れを図１に示す。以下、図１に従って翻訳処理の概略について述べる。

Ａ．入力文章の単語分割（日英翻訳時のみ）

入力文章が英語の場合は単語がスペースで区切られているが、日本語の場合はかな漢字混じりのべた書き文であり、単語分割の必要性が生じる。以下、日本語文の単語分割処理について述べる。

一般の機械翻訳システムでは、入力文章の単語分割時に入力文章の構文解析を行うのが普通である。これに対して我々のシステムは、入力文章の文法的な解析を行っていないため、非常に簡単な手法を採用している。この手法では、単語辞書中の見出し語が文字列の長い順に登録されている分割辞書を用いて、文字列の長いものから順に単語として切り出される。従って、最後に残った文字列が助詞の「は、が、を」等として認識される。なお、この方式における日本語文章の正分割率は約９５％である。

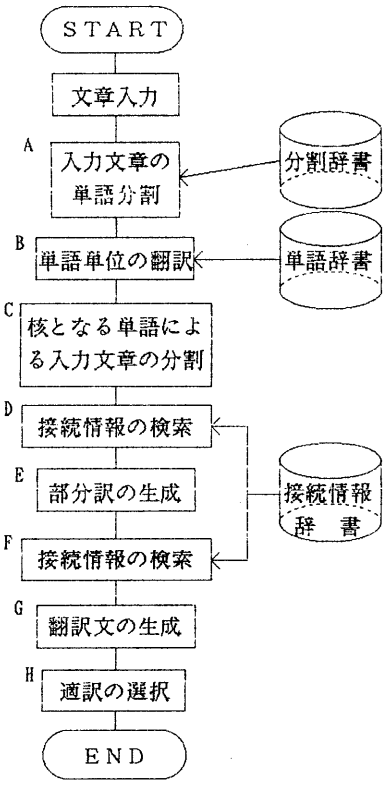


図１　翻訳処理の流れ

Ｂ．単語単位の翻訳

単語単位の翻訳では、単語辞書を用いて翻訳対象文章を単語単位に翻訳する。単語辞書には英単語とその訳語（日英の場合は日本語の単語とその訳語）および頻度情報などが登録されている。複数の訳語が単語辞書に登録されている場合は、全ての訳語を考慮する。

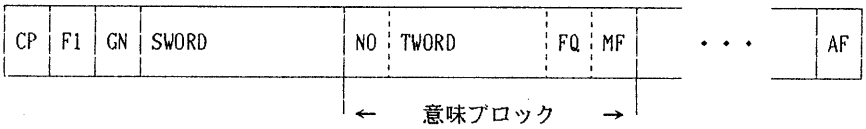
単語辞書の構造を図２に示す。単語辞書は可変長であり、図２の単語SWORDに対する訳語TWORDが１～９種類まで登録することが可能である。訳語には訳語番号NOが付加されており、訳語の接続処理は同類語概念[4]を用いたこの訳語番号で行われる。同類語とは、訳文中の同位置で同様の使われ方をする訳語である。例えば、 "use"の訳語「～（を）用いる」と "introduce"の訳語「～（を）導入する」は同類語であり同じ訳語番号が付加される。同類語概念の採用により、接続情報数の増大を抑えている。

Ｃ．核となる単語による入力文章の分割

本アルゴリズムで長文が入力された場合、接続可能な訳語の組（接続情報）が増大し翻訳速度等に悪影響を及ぼす可能性がある。このため、実験システムでは、入力文における単語の位置情報を翻訳処理過程で利用している。

これは、入力文中においてある単語を核として入力文を分割し、その分割部分ごとに部分訳を生成してから、部分訳同士の接続により翻訳文を生成するものである。これにより、不必要な訳語の接続解析を減少させている。核となる単語の定義は以下に示す通りである。

①入力文章をある単語で分割した場合、分割部分内で部分訳を生成することができ



CP　：意味ブロック数（登録訳語数）　1byte
F1　：予備用フラッグ　1byte
GN　：語番号　4byte
SWORD　：単語　25byte
意味ブロック：訳語ブロック　28byte
　NO　：訳語番号　4byte
　TWORD　：訳語　20byte
　FQ　：訳語の出現頻度　3byte
　MF　：訳語のフラッグ　1byte
AF　：単語の総出現頻度　4byte

図２　単語辞書の構造

②その単語が任意の同一分野の文章に出現した場合も同様に分割区分内で部分訳を生成することができる

場合は、この単語を核となる単語とする。核となる単語には、英日の場合で前置詞や動詞系の単語、日英では「は、が、を」等の助詞や前置詞が挙げられる。核となる単語の一覧とその分割区分を表１に示す。

図３に英日、図４に日英の場合の部分訳生成および訳文生成過程を示す。部分訳生成時には、分割部分外の訳語との接続解析を行わないために、不必要な訳語の接続解析数が減少している。このため、処理速度および翻訳文の質が入力文の分割を行わない場合に比べて大幅に改善されている。日英翻訳の場合も同じことが言える。

表１　核となる単語の一覧と分割区分

	核となる単語	分割区分
英日翻訳	・前置詞 ・only ・such as　など	核となる単語から次の核となる単語の前まで
	・動詞系の訳語を持つもの ・関係代名詞系の訳語を持つもの など	核となる単語の前後で分割
	・助動詞系の訳語を持つもの など	核となる単語から次の動詞系の訳語を持つ単語まで
日英翻訳	・助詞「は、が、を」	その位置で分割
	・動詞系の訳語を持つもの ・前置詞系の訳語を持つもの	核となる単語の前後で分割

＜入力文＞　We propose a new translation algorithm for machine translation .

＜分　割＞　We ／ propose ／ a new translation algorithm ／ for　machine translation.
　　我々　提案する　新しい　翻訳　アルゴリズム　のための 機械　翻訳
　　↓　↓　↓　↓

＜部分訳＞ 我々　提案する　新しい翻訳アルゴリズム　機械翻訳のための

＜翻訳文＞　我々は機械翻訳のための新しい翻訳アルゴリズムを提案する。

図３　単語の位置情報を利用した翻訳処理過程（英日）

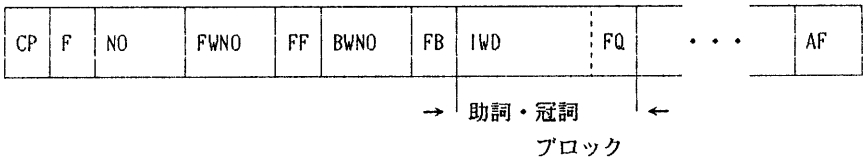
<入力文> 我々は機械翻訳のための新しい翻訳アルゴリズムを提案する

<分　割> 我々 ／ は ／ 機械 翻訳 ／ のための ／ 新しい 翻訳 アルゴリズム ／ を ／ 提案する。

<部分訳> we　machine translation　for　new translation algorithm　propose

<翻訳文> We propose a new machine translation for machine translation.

図４　単語の位置情報を利用した翻訳処理過程（日英）



CP	：助詞・冠詞ブロック数	1byte
F	：予備用フラッグ	1byte
NO	：接続情報番号	4byte
FWNO	：前訳語番号	4byte
FF	：予備用フラッグ	1byte
BWNO	：後訳語番号	4byte
FB	：予備用フラッグ	1byte
助詞・冠詞ブロック	：助詞・冠詞の処理用	8byte
IWD	：助詞・冠詞	6byte
FQ	：助詞・冠詞の出現頻度	2byte
AF	：接続情報の総出現頻度	4byte

図５　接続情報辞書の構造

Ｄ．接続情報の検索－１

核となる単語によって分割された入力文の分割部分内におけるそれぞれの訳語について、接続可能な訳語の組を接続情報辞書より全て抽出する。

接続情報辞書の構造を図４に示す。接続情報辞書も単語辞書と同様に可変長であり、訳文中で接続する可能性のある訳語の組が、単語辞書における訳語番号の組（FWNO, BWNO）としてその出現頻度とともに登録されている。また、日本語（英日翻訳）における接続情報辞書には訳語間に出現する可能性のある助詞「は」、「が」、「を」等がその出現頻度とともに３種類まで登録されている。英語（日英翻訳）の場合は冠詞"a"、"an"、"the"がその頻度情報とともに３種類まで登録されている。本実験システムでは、これらの情報を基に助詞または冠詞の付加処理を統計的に行っている。助詞または冠詞の付加処理は以下に示す様に行われる。

接続情報の出現頻度をＡ、助詞または冠詞の出現頻度の合計をＢとした場合、次式

$$S = (B/A) \times 100\ [\%]$$

で$S > 50$のとき、最大頻度の助詞または冠詞を訳語間に付加する。域値５０は実験的に定められたもので、この場合は約８０％の正答率が得らている。

E．部分訳の生成

抽出された接続情報を用いて、分割部分内で得られた訳語を接続して部分訳を生成する。この時、分割部分内の全ての訳語を用いて部分訳が生成されるため、複数の部分訳が生成される場合もある。以下に、英日および日英翻訳時における部分訳生成の例を示す。

例）英日における部分訳生成

＜分割部分＞　new　machine　translation
＜訳　　語＞　新しい　機械　翻訳
　　マシン
＜部 分 訳＞　新しい－機械－翻訳
　　新しい－翻訳－マシン

例）日英における部分訳生成

＜分割部分＞　新しい　機械　翻訳
＜訳　　語＞　new　machine　translation
　　machines
＜部 分 訳＞　new-machine-translation

生成された部分訳には優先順位が付けられる。それぞれの接続情報には前述したように、過去における出現頻度が付加されている。部分訳における先頭の訳語から末尾の訳語までの接続情報の出現頻度の総和を求め、これを部分訳の優先順位とする。従って、優先順位の高い部分訳における訳語列ほど過去において多く出現しており、部分訳としての確からしさが高いと言える。翻訳文の生成過程では、優先順位の高い部分訳から順に接続解析が行われる。なお、接続情報の未登録等により部分訳の生成が不可能な場合（部分訳が１つも生成されない場合）は、強制接続により訳語を接続し部分訳とする。強制接続は、入力文における語順を考慮して訳語を並べ換えるものである。

F．接続情報の検索－２

作成されたそれぞれの部分訳における先頭の訳語と末尾の訳語を用いて、接続する可能性のある訳語の組を接続情報辞書より全て抽出する。この時、先頭と末尾以外の訳語については接続解析を行わないので、訳文生成時に発生する無駄な解析木の生成が抑えられている。

G．翻訳文の生成

接続情報の検索－２で抽出された接続情報を用いて、文頭から文末まで部分訳を接続し翻訳文を生成する。この時、文頭から文末までそれぞれの部分訳が過不足なく重複せずに接続したものを翻訳文とする。従って、本システムでは複数の翻訳文が生成される場合がある。翻訳文の生成過程を図５に示す。図５で、＿＿の部分が生成された翻訳文である。この場合は翻訳文として２文が生成されている。なお、接続情報の未登録等によって翻訳文の生成が不可能な場合（翻訳文が１文も生成されない場合）は、強制接続によって翻訳文を生成する。これは、部分訳生成時と同様の処理で、入力文における部分訳の位置を考慮して部分訳を強制的に接続し、翻訳文とするものである。

＜入 力 文＞　We propose a new translation algorithm for machine translation.

＜部 分 訳＞　我々　提案する　新しい翻訳アルゴリズム　機械翻訳のため

＜訳文生成＞
T ┬ 我々 ┬ は新しい翻訳アルゴリズム ― を提案する ― T
　　　　├ は新しい翻訳アルゴリズム ― が機械翻訳のため
　　　　└ は機械翻訳のため ― の新しい翻訳アルゴリズム ― を提案する ― T
　　├ 新しい翻訳アルゴリズム ┬ を提案する ― T
　　　　　　　　　　　　　　└ が機械翻訳のため ― に我々
　　└ 機械翻訳のため ┬ の新しい翻訳アルゴリズム ― を提案する ― T
　　　　　　　　　　└ に我々 ― は新しい翻訳アルゴリズム ― を提案する ― T

（Tは文頭・文末を表す）

図５　部分訳の接続による翻訳文生成過程

H．適訳の選択

生成された翻訳文には部分訳の生成時と同様に優先順位が付けられる。これは、生成された訳文の文頭から文末までの接続情報の頻度の総和を求めて、総頻度の高い順番に優先順位を付加するものである。従って、翻訳文における優先順位は部分訳における優先順位と同じく、順位の高い翻訳文ほど妥当な翻訳文であると言える。

4．性能評価実験

実験システムの翻訳性能を評価するために翻訳実験を行った。以下、実験手順について述べる。

①情報処理関係の英文論文誌から任意に抽出した２０００文の論文アブストラクトを人手により翻訳する。

②人手による翻訳結果から単語情報、接続情報を抽出し、単語辞書および接続情報辞書をそれぞれ構築する。

③辞書構築に用いた文章と異なる１００文について実験システムにより翻訳を試みる。

①では、人手によって作成された翻訳文から直接単語辞書と接続情報辞書が生成されるので、単語と訳語の関係が１対１になるように翻訳を行う必要がある。②では、作成された翻訳文中の訳語および隣合う訳語の組が、それぞれ単語辞書と接続情報辞書に頻度情報とともに登録される。この時、さらに冠詞や助詞の付加情報が、それらの出現頻度とともに接続情報辞書に登録される。

なお、今回の実験では日英翻訳用の辞書構築が間に合わなかったため、性能評価実験は英日翻訳の場合についてのみ行った。

5．実験結果および考察

上述の手順により行った翻訳実験の結果を以下に示す。

①試験文１００文中６２文が正しく翻訳された。

②試験文１００文中複数の翻訳文が出力された例が１４例であった。

③上記の１４例のうち、正解文を含んでいたものが８例であった。

なお、単語辞書の容量が約１５００語、接続情報辞書の容量が約４０００組であった。

①で、誤って翻訳された３８文を見ると大きく２つの誤りに分けることができる。１つは、接続情報の未登録により強制接続によって部分訳や翻訳文を生成したものである。以下に翻訳例を示す。

例）We test the behavior of the model again.
→我々はそのモデル再びのテスト動作
※"again"の訳語「再び」との接続情報の未登録による誤翻訳

これについては、より多くの情報を辞書に登録して、辞書を充実させることにより改善が可能である。もう１つは、接続詞を含む文における核となる単語の分割誤りによるものである。以下に翻訳例を示す。

例）A parallel sorting algorithm by A and B includes a merging process.
→AによるとBの並列ソーティングアルゴリズムはマージ処理を含む
※"and"を核となる単語にして入力文を分割し部分訳を生成したことによる誤翻訳

これについては、核となる単語の再検討や接続詞に関する特別規則の設定などが必要である。

なお、③の正解文を含んでいたもの８例に関しては、優先順位５位以内に全ての正解が含まれていた。以下に翻訳例を示す。

例）Some practical systems have been developed for character recognition.
→いくつかの実用的なシステムが文字認識のために開発されている　[293]
→文字認識のためにいくつかの実用的なシステムが開発されている　[267]
※[]内は文頭から文末までの接続情報の出現頻度の総和

また、翻訳速度に関しては応答時間で平均２～３秒であった。以上の結果から、パーソナルコンピュータ上での本システムの有効性を確認できたと言える。

6．おわりに

本論文では、単語の接続情報を利用した機械翻訳手法について述べた。本手法は、複雑な文法的情報をほとんど用いず接続情報のマッチングで翻訳文を生成するため、以下に示すような利点を持っている。

1．翻訳用辞書の保守・管理が容易である。
2．システム構築を容易に行うことができる。
3．辞書を構築するだけで他言語間の翻訳が可能である。

実際に、実験システムを作成し性能評価実験により、その有効性を確認することができた。

今後の課題としては、実験結果で述べたように翻訳用辞書の充実化が挙げられる。また、接続情報辞書を一度構築すると、単語辞書を構築するだけで他言語の翻訳が可能になる利点を持っているので、他言語の辞書構築も今後の課題である。

さらに、本システムは文法的情報をほとんど用いていないため、翻訳文の質に限界がある。従って、文法情報の利用も今後の課題の１つと言える。

謝　辞

本研究を行うにあたり、終始適切な御示唆をいただいた北海道大学工学部電子工学科電子機器工学講座各位に感謝致します。

参考文献

[1]鈴木、栃内:語の接続関係を利用した機械翻訳システム、情報処理学会論文誌Vol.29 No.4 1988
[2]野村、田中:機械翻訳、共立出版bit別冊 1988
[3]鈴木:日本語情報処理における語の接続関係とその応用に関する研究、北海道大学工学部修士論文 1985
[4]鈴木、栃内:単語接続情報を用いた科学技術文献の自動翻訳システム、北海道工業大学研究紀要第19号 1991