北海道大学工学部研究報告
第138号（昭和63年）

Bulletin of the Faculty of Engineering,
Hokkaido University. No. 138 (1988)

語の接続関係を用いた機械翻訳の一手法

鈴木　康広　宮永　喜一　栃内　香次

（昭和62年9月30日受理）

A Method for Machine Translation Using Conjunctive Relation of Words

Yasuhiro SUZUKI, Yosikazu MIYANAGA, Koji TOCHINAI

(Received September 30, 1987)

Abstract

This paper proposes 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. An experimental translation system was constructed, and experiments were performed using 116 titles extracted from the transactions of IPSJ. As a result of the experiments in English-Japanese translation, the rate of exactly translated titles was 75%, and in Japanese-English translations, the rate was 66%.

1．はじめに

現在，一般的な機械翻訳の手法は原言語の形態素解析や構文解析，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[1]。本報告では，これとは別の手法として語の接続関係を用いた機械翻訳手法について述べる[2]。この手法は形態素解析や構文解析，意味解析を行わないので辞書中に品詞，活用などの文法的な情報を登録する必要がなく，単語情報の辞書への登録が容易であるという特長を持っている。

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]。たとえば，「処理，に関する，言語，研究，自然」という単語群が存在する場合，これらの語を並べて作ることができる文章は，「自然言語処理に関する研究」に限られる。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。そして，対象文章の分野を限定することによってこの性質はさらに著しくなる。したがって，ある特定分野の多数の学術文献から，文章中で接続しているそれぞれの単語の組を予め抽出して辞書に登録しておくことによって，文章を構成する単語群が与えられた場合，上記の情報をもとに文章を生成することができる。本報告では，このような語の接続関係を利用した機械翻訳の手法について述べる。

2．語の接続関係を利用した機械翻訳

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。このことを利用した機械翻訳のアルゴリズムについて以下に述べる。

前述の，「処理，に関する，言語，研究，自然」という語群を考えてみる。この語群から生成することのできる文章は，「自然言語処理に関する研究」という一文に決まる。これは，それぞれの

電子工学科　電子機器工学講座

単語が「T-自然，自然-言語，言語-処理，処理-に関する，に関する-研究，研究-E（Tは文頭，Eは文末を意味する）」という接続関係を持ってい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，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作成することができる。

例） Study on Natural Language Processing （英文）
↓ ↓ ↓ ↓ ↓
研究 に関する 自然 言語 処理 （単語単位の翻訳）
T-自然，自然-言語，言語-処理 （接続情報）
処理-に関する，に関する-研究，研究-E
自然 言語 処理 に関する 研究 （訳文）

なお，接続情報は予め大量の文献から抽出し，接続情報辞書に蓄えておく，また，このアルゴリズムは日英翻訳についても同様に適用可能である。また，多義語における訳語の選択も接続情報を用いることによって行うことができる。

3．実験システム

上述の機械翻訳アルゴリズムによる実験システムを作成した。以下，このシステムの概要と翻訳手順について述べる。実験システムは PL/I で書かれており，北海道大学大型計算機センタの HI TAC M-680H 上に作成されている。翻訳の手順は基本的には英日と日英翻訳で共通である。なお，本システムは小規模な実験システムであり，翻訳対象文章は情報処理関係の論文，研究会報告等の表題に限定している。また，単語辞書および接続情報辞書には予め情報処理関係の論文表題など 1500 例から抽出した情報を登録してある。登録された単語数は 1463 語，日本語接続情報数は 5317 組，英語接続情報は 5747 組である。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を述べる。

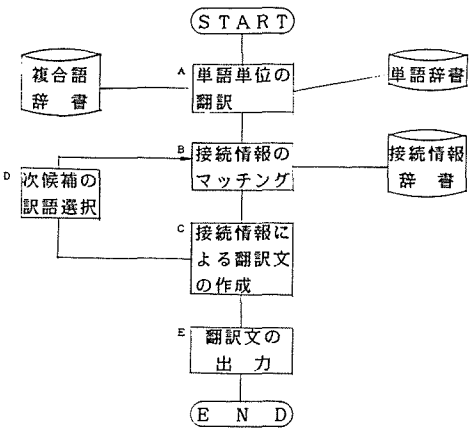


図 1 翻訳処理の流れ

3．1 英日翻訳処理

A．単語単位の翻訳

単語辞書を用いて翻訳対象文章の単語単位の翻訳を行う。単語辞書には，英単語とそれに対する日本語の訳語が 3 種類まで記録されている。単語辞書の構造を図 2 に示す。前述のように，単語辞書中には単語，訳語およびその頻度情報が記録されているだけで，品詞，活用など文法的情報は記録されていない。この辞書から最も頻度の大きい訳語を選択して単語単位の翻訳を行う。以下に "A Study on Natural Language Processing" の単語単位の翻訳例を示す。

例） A Study on Natural Language Processing
↓ ↓ ↓ ↓ ↓
研究 に関する 自然 言語 処理

単語単位の翻訳を行う際，ある単語が複数の訳語を持ち，かつほぼ等頻度で出現している場合は頻度情報による訳語の選択は困難である。このような場合，当該の語がその前後の語と複合語になっていれば，それを利用して訳語を選択することができる。すなわち，単語辞書とは別に複合語辞書を設け当該の語が複合語辞書に登録されている場合は無条件で複合語辞書中の訳語を選択する。辞書に登録されている複合語の一部を表1に示す。

表1　複合語辞書の内容

英語複合語	日本語複合語
Production Rule	プロダクション　ルール
Expert System	エキスパート　システム
Machine Translation	機械　翻訳
Personal Computer	パーソナル　コンピュータ
Kana-Kanji Translation	かな漢字　変換
Natural Language	自然　言語
Production System	プロダクション　システム
Image Processing	画像　処理
Computer Graphic	コンピュータ　グラフィック
User Interface	ユーザ　インターフェース
Vector Processor	ベクトル　プロセッサ
Word Processor	ワード　プロセッサ
Parallel Computer	並列　計算機
Poisson Process	ポアソン　過程
Error Correction	誤り　訂正
Error detection	誤り　検出
Color Graphic	カラー　グラフィック
Graphic Display	グラフィック　ディスプレイ
Integrated Circuit	集積　回路

B．接続情報のマッチング

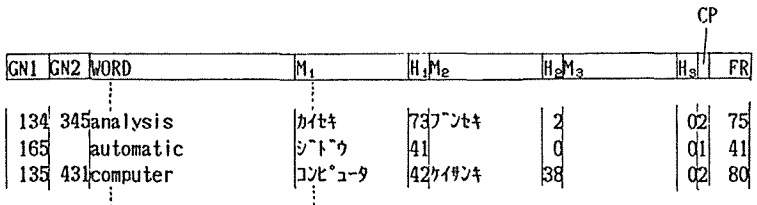
単語単位の翻訳によって得られたそれぞれの訳語について，日本語接続情報辞書を用いて接続可能な訳語の組を捜す。以下に示す例は“Study on Natural Language Processing”を単語単位に翻訳した結果から得られる接続情報である。

例）　T-自然，T-言語，T-処理，研究-E，
処理-E
に関する-研究，自然-言語，言語-処理
言語-E，処理-言語，処理-に関する

日本語接続情報辞書の構造を図3に示す。図に示すようにこの辞書には訳語の組，すなわち接続情報とその組の出現頻度および訳語間の助詞が3種類までその出現頻度と共に記録されている。

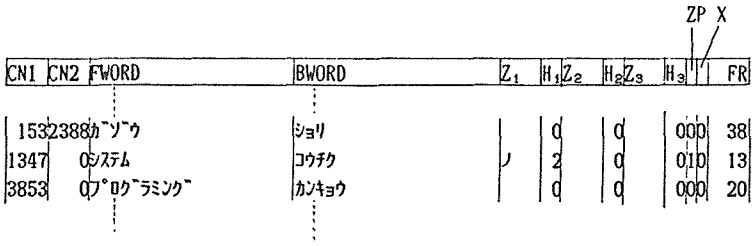
C．接続情報による翻訳文の作成

ステップBで求めた訳語の組を用いて文章の組み立てを行い，すべての訳語が含まれ，かつ文頭から文末まで接続している訳語列を訳文とする。以下に，"A Study on Natural Language



GN1 ：語番号
GN2 ：この語の類似語番号（単数，複数形）
WORD：英単語
M_1 ：訳語
H_1 ：訳語の頻度
CP ：登録されている訳語の数
FR ：英単語の出現頻度

図2　単語辞書の構造



CN1 ：接続情報番号　　CN2 ：類似接続情報番号
FWORD：前の単語　　BWORD：後の単語
Zi ：単語間の助詞　　H_i ：助詞の出現頻度
ZP ：登録されている助詞の数　　X ：未使用
FR ：接続情報の出現頻度
*) FWORD,BWORD中の t, e は文頭，文末を表す

図3　日本語接続情報辞書の構造

Processing" に対する接続可能な訳語列を示す。

例）
- T
  - 自然 - 言語
    - 処理
      - E
      - に関する - 研究 - E　○
    - E
  - 言語
    - 処理
      - E
      - に関する - 研究 - E
    - E
  - 処理
    - 言語 - E
    - E
    - に関する - 研究 - E

上の例では，○印の付いている文章が訳文となる。

翻訳文の生成過程で辞書中に接続情報が存在しないために訳文が生成されない場合がある。このような場合は，以下のような強制接続ルールを用いて訳文を生成する。

〈ルール〉 1ヶ所または2ヶ所で接続情報が辞書中に存在しない場合は，その部分を強制的に接続し訳文とする。

例1）「接続-関係」が辞書にない。

T-語-の-接続⇦⇨関係-を用いた-機械-翻訳-E

例）「接続-関係」，「を用いた-機械」が辞書にない。

T-語-の-接続⇦⇨関係-を用いた⇦⇨機械-翻訳-E

上の例で，例1は一つの接続情報「接続-関係」が接続情報辞書中にない場合で，例2は二つの接続情報「接続-関係」，「を用いた-機械」が接続情報辞書にない場合であり，'⇦⇨' は強制的に接続した部分である。なお，上述のルールは日英翻訳時にも同様に適用される。また，3ヶ所以上で接続情報が辞書中に存在しない場合は，次候補の訳語選択を行う。

D．次候補の訳語選択

ステップCで訳文が生成されなかった場合は，以下の手順で訳語の次候補を選択する。

①．今，単語単位の翻訳の結果 A_1，B_1，C_1，D_1，E_1，F_1 という訳語が得られたとする。このう

ち次候補を持つものを A_1, E_1 とし，これらの次候補を A_2, E_2 とすると以下のように次候補選択が行われる。

イ）次候補を持つ訳語 A_1, E_1 についてこれらを含む接続情報をステップBで得られた接続情報から捜す。

ロ）A_1, E_1 を含む接続情報の出現頻度の総和

$$S_A=\{\Sigma(A_1-X)+\Sigma(X-A_1)\}$$

$$S_E=\{\Sigma(E_1-X)+\Sigma(X-E_1)\}$$

を求める。ここで，(A_1-X) $(X-A_1)$ は訳語 A_1 を含む任意の接続情報の出現頻度であり，$\Sigma(A_1-X)$，$\Sigma(X-A_1)$ は接続情報の出現頻度の総和である。

ハ）S_A, S_E のうち最小のものについて次候補を選択する。

$S_A<S_E$→次候補 A_2 を選択

$S_A>S_E$→次候補 E_2 を選択

ただし，複数の訳語を持つ語が一語しか存在しない場合はその訳語の次候補を選択する。

②．①で取り出した次候補を用いて再び接続情報のマッチングを行う。

③．訳文が生成されない場合は，候補がなくなるまで①，②を繰り返す。

④．最終的に訳文が生成されなかった場合は，解析不能のメッセージを表示する。

E．翻訳文の出力

生成された文章が一文の場合はその文章を翻訳文として出力する。生成された文章が複数の場合は以下の手順で翻訳文を選択する。

①．それぞれの接続情報には，頻度情報が付加されているので文頭から文末までの各接続情報の頻度の総和を求める。

②．①で求めた接続情報の頻度の総和が最大の文章を翻訳文として選択する。

3．2　日英翻訳処理

A．単語単位の翻訳

日英翻訳の場合は単語辞書を逆引きして最も頻度の大きい単語を選択し，単語単位の翻訳を行う。以下に「自然言語処理に関する研究」の単語単位の翻訳例を示す。

例）

自然	言語	処理	に関する	研究
↓	↓	↓	↓	↓
Natural	Language	Processing	on	Study

単語単位の翻訳を行う際，英日翻訳と同様に複合語辞書中の訳語を優先させる。

B．接続情報のマッチング

英語接続情報辞書を用いて接続可能な訳語の組を捜す。以下に示す例は「自然言語処理に関する研究」を単語単位に翻訳した結果から得られた接続情報である。

例）T-Natural, T-Language, T-Processing
T-Study, Natural-Language, Study-E
Language-Processing, Language-E
Processing-Natural, Processing-Language
Processing-E, On-Natural, On-Processing

英語接続情報辞書の構造を図4に示す。図に示すようにこの辞書には接続情報とその出現頻度および後の訳語に付随する冠詞が3種類までその出現頻度とともに記録されている。

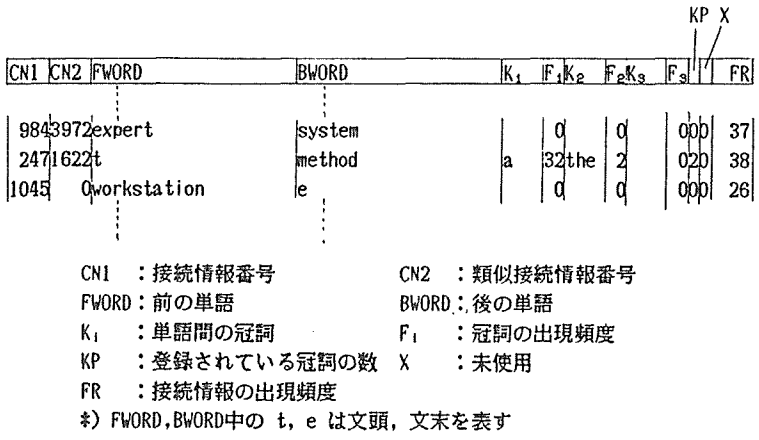


図4 英語接続情報辞書の構造

C．接続情報による翻訳文の作成

ステップBで得られた接続情報を用いて翻訳文を生成する。以下に「自然言語処理に関する研究」の訳文生成過程を示す。

例） T ┌ Natural - Language ┌ Precessing - E
　　 │　　　　　　　　　　└ E
　　 ├ Language ┌ Processing - Natural
　　 │　　　　 └ E
　　 ├ Processing ┌ Natural - Language - E
　　 │　　　　　 └ E
　　 └ Study - on ┌ Natural - Language ┌ Processing - E ○
　　 　　　　　　 │　　　　　　　　　 └ E
　　 　　　　　　 └ Processing ┌ Natural - Language - E ○
　　 　　　　　　 　　　　　　 └ Language - E

この例では○印の2文が訳文として生成される。

D．次候補の訳語選択

英日翻訳の場合と同様に次候補の訳語を選択する。

E．翻訳文の出力

ステップCで生成された訳文を以下に示す。

例）訳文

① A Study on Natural Language Processing (124)

② A Study on Processing Natural Language (116)

上の例で（ ）内は接続情報の頻度の総和である。上の例では①の訳文が最終的な翻訳結果として選択される。

3．3 日英翻訳における冠詞の処理

日英翻訳時には冠詞の処理問題が生じる。本システムでは冠詞の処理を統計的に行っている。以下に，冠詞の処理手順を示す。

①. 単語単位の翻訳結果から英語接続情報辞書によって得られた訳語の組に対して，訳語の組の出現頻度，および冠詞とその出現頻度を辞書から取り出す。

②. ①で取り出した訳語の組の出現頻度をM，冠詞の出現頻度（複数の冠詞が付属する場合はそれぞれの冠詞の出現頻度の総和）をNとし

$$(N/M)\times 100 \quad [\%]$$

の値が50%を越える場合，最も出現頻度の大きい冠詞を付ける。

③. 50%以下の場合は冠詞を付けない。

上述の50%という値は以下に示すように実験的に定めたものである。すなわち，接続情報の出現頻度に対する冠詞の出現頻度の割合をパラメータとして冠詞の当てはめ実験を行ったところ表2に示すような結果が得られた。表2は冠詞がつく可能性のある267ヶ所について冠詞の当てはめを行った結果であり，PARAは冠詞を付けるか付けないかの判断の基準となる前述のパラメータである。正解率は267ヶ所中原文と一致した割合を示す。

表2　冠詞の当てはめ実験の結果

PARA	正解率	a	b	c
10 %	57.3 %	8.8 %	86.9 %	5.2 %
30 %	67.0 %	44.3 %	53.4 %	2.3 %
50 %	72.7 %	94.7 %	4.0 %	1.3 %
70 %	70.8 %	100 %	0 %	0 %
90 %	64.8 %	100 %	0 %	0 %

表2でA，B，Cは誤りの種類で，それぞれ原文に冠詞が付いているのに付けなかった誤り(a)，原文に冠詞が付いていないのに付けたあやまり(b)，誤った冠詞を付けた誤り(c)である。表2をみるとPARA＝50%のときが最良の値となっている。

3．4　英日翻訳における助詞の処理

英日翻訳時には，「は」，「が」などの助詞の処理の問題が生じる。助詞の処理についても冠詞の場合と同様の処理を行っている。PARAの値は，接続情報の延べ出現数に対して助詞の付随する割合が0.9%と小さく，実験的に決定することができなかったので，冠詞の場合を参考にして50%としている。

4．翻訳実験と実験結果

実験システムの性能を評価するために，情報処理関係の文献の表題116例について英日，日英の翻訳実験を行ったのでその結果を表3に示す。表3は英日，日英それぞれの正翻訳率を示すものである。①，②，③はそれぞれ前述の強制接続ルールが用いられなかった場合，1ヶ所のみ強制接続を行った場合および2ヶ所について強制接続を行った場合の実験結果である。また，Iは翻訳結果が一文であった場合，IIは複数の翻訳結果が得られた場合の実験結果である。

表3　翻訳実験結果

英日翻訳実験結果				
	文数	正翻訳	誤翻訳	正翻訳率
①	60	53	7	88.3 %
I	29	28	1	96.9 %
II	31	25	6	80.6 %
②	21	18	3	85.7 %
I	17	15	2	88.2 %
II	4	3	1	75.0 %
③	19	16	3	84.2 %
I	12	12	0	100.0 %
II	7	4	3	57.1 %
総合	100	87	13	87.0 %
I	58	55	3	94.8 %
II	42	32	10	76.2 %
日英翻訳実験結果				
	文数	正翻訳	誤翻訳	正翻訳率
①	58	38	20	65.5 %
I	20	17	3	85.0 %
II	38	21	17	55.3 %
②	25	23	2	92.0 %
I	17	15	2	88.2 %
II	8	8	0	100.0 %
③	15	15	0	100.0 %
I	11	11	0	100.0 %
II	4	4	0	100.0 %
総合	98	76	22	77.6 %
I	48	43	5	89.6 %
II	50	33	17	66.0 %

原則として，原文と一致した訳文が得られた場合を正しく翻訳されたと見なした。ただし，以下に示すような冠詞の違いや単数，複数の違いがあった場合でも正しく翻訳されたと見なしている。

例) A Method～ ＝Method～
～Expert Systems～ ＝～Expert System～
～を用いた～ ＝～を利用した～

翻訳結果をみると，正翻訳率は日英翻訳の方が約10%低い値となっている。これは，辞書構造の関係で英語と日本語単語の多義性に違いがあるためである。英単語の場合は，辞書の構造から最大3種類の訳語までしか持つことができないのに対して，日本語単語の場合は，辞書を逆引きするので制限がない。このことによって日英翻訳の場合，その接続情報にばらつきが生じ，正しい翻訳ができず正翻訳率が低くなっていると考えられる。今回の実験では単語辞書を英日，日英翻訳で共用しているが日英翻訳のための単語辞書を別に作成して訳語の個数に制限を設ければ英日翻訳と同程度の正翻訳率が得られると予想される。

また，一般的に複数の訳文が出力されたときの正翻訳率が低くなっている。これは，複数の訳文が作成された場合，接続情報の頻度の総和が最大のものを適訳として選択するアルゴリズムを用いているためで，接続情報の頻度の総和が最大のもの以外でも正解であった例がいくつか存在した。

最終的な正翻訳率は，英日翻訳で116文中87文が正しく翻訳され76%であった。また，日英翻訳では116文中76文が正しく翻訳され66%であった。

5．おわりに

本報告では，語の接続関係を利用した機械翻訳手法について述べてきた。本手法は，文法的な情報を用いず接続情報のマッチングで訳文を生成するので以下のような利点がある。

1．辞書への単語情報，接続情報の登録が容易である。
2．システムでの実現が容易である。

実際に，実験システムを作成し情報処理関係の文献の題目を対象とした実験を行った結果，英日翻訳で75%，日英翻訳で66%の最終的な正翻訳率が得られた。

今後の課題としては，複数の訳文が出力された場合の適訳の選択アルゴリズムの改良があげられる。また，今回は論文表題について翻訳実験を行っているが，同様のアルゴリズムが一般文にも適用可能であるかどうかも今後の課題となる。

謝　辞

本研究の遂行にあたり，種々有益な御討論をいただいた研究室の皆様に感謝致します。

参考文献

1）情報処理　Vol.26　No.10（1985）「機械翻訳特集」
2）鈴木康広，宮永喜一，栃内香次：語の接続関係を用いた機械翻訳　情報処理学会第35回全国大会，3S-3（1987）
3）鈴木康広：日本語情報処理における語の接続関係とその応用　北海道大学工学部修士論文（1985）