自然言語処理 33－5
（1982. 10. 22）

Prologを用いた機械翻訳システムにおける意味処理

村木一至　市山俊治
（日本電気(株) C&Cシステム研究所）

1．はじめに

近い将来に実用的翻訳システムが出現するであろうという予測がある。この予測は機械翻訳システムへの強いニーズによって支えられている。こうした状況を受け、日本、欧米では、機械翻訳システム（MTシステム）実現への研究が多角度から進められている。

MT研究は計算機の誕生と同時に始まり、語彙データ主体の翻訳、構文主体の翻訳、意味を用いた翻訳へと歩みを進めてきた。この潮流は、変形・句構造文法や意味文法（格文法）といった言語理論の誕生と符合していると同時に、計算機の効率向上、容量増大といった研究環境の改善に支えられてきた。その結果、言語理解、MTに語彙データ、構文情報、意味情報を扱う統一的計算機モデルが必要であることが広く認識されるようになってきた。

他方、MT研究に於いては翻訳方式としてピボット方式・トランスファ方式が明確に概念化され、最近それらを折衷する方式として融合方式[7]も提案されてきた。

こうした方式とは別に最近例文による翻訳、類似による翻訳を提唱する動きがある。この考え方は、対象分野のテキストから生のデータを取出し、生データの持つ言語的形式をそのまま利用して翻訳を行うものであり、データとしての文パターンとの照合だけによって翻訳文を生成するという単純さの利点を持つものであるが、例文とその変形（語の置換）だけで入力文の全てをパターン化することはかなり困難である。

筆者らは、こうした対象領域のテキストからの生データが持つ情報を翻訳用テンプレートとして用いるのではなく、テキスト中の語と語、句と句の係り関係に関する生データを用いる擬ピボット方式MTシステムを提案する。この係り関係はプラグマティックステーブルと呼ぶ知識として形式化される。プラグマティックステーブルは、テキストに出現する語と語の共起関係を保持し、この内には対象領域で出現しない共起関係は含まれない。

従来MTシステムではこの共起関係を構文的情報や意味情報として人間が抽象化し、解析、(変換)、生成のメカニズムに組込んできた。この際抽象化を進める余り、現実のテキストデータに存在する多様性を切り捨ててしまい、話の統語的・意味的曖昧さをそのままメカニズムと知識の中に抱え込んでしまう傾向があった。その曖昧さを少しでも解消する目的で、多くのヒューリスティックを導入する必要があったが、それはシステムの見通しのよさ、健全性を損う危険性をいつも孕んでいた。

本稿で提案する方式は、Prolog[1]を用いた翻訳システムに於ける意味処理方式を基礎付けるものであり、形態素解析、構文・意味検証、翻訳テンプレートとしての内部表現生成、翻訳文生成の各フェーズにプラグマティックステーブル内の生の語関係情報を適用し、曖昧さの解消、訳語の選択を行うものである。この方式によれば、対象分野の生データとしてのプラグマティックス（意味、文中の文脈情報を含む）を翻訳システムのメカニズムと独立に管理できるとともに、翻訳の質を実例を用いることにより高度化できる。

筆者らは、Prologを用いた英日翻訳システムTRAPを開発してきた。[2][5]

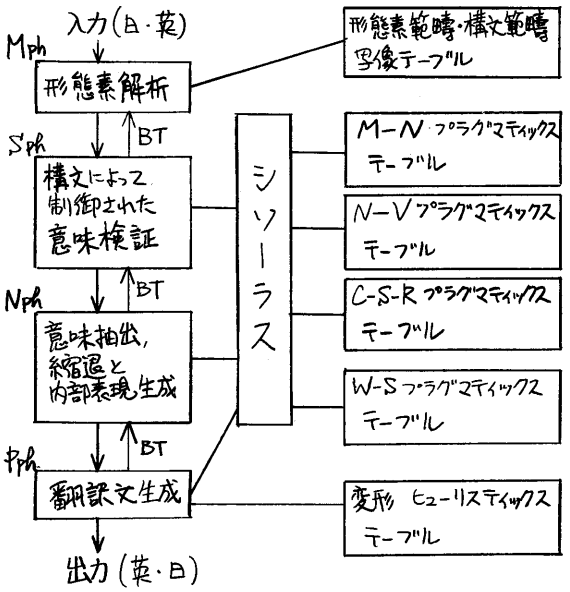


図1. 翻訳処理フローとプラグマティックステーブル

TRAPには4つのフェーズがあり（図1参照）、各々形態素解析、構文によって制御された意味検証、意味抽出・縮退と内部表現生成、翻訳文生成と呼ばれる。これら4つのフェーズは、前フェーズ内の曖昧さから発生する後フェーズでの誤り発生（処理の行詰り）によって自動的に前フェーズへバックトラッキングがかかり、そのフェーズで抱える曖昧さの1つを選択的に後フェーズへ伝える。

以下、翻訳システムにおけるプラグマティックステーブルの意義を明確化しつつ、TRAPシステムの各フェーズにおける処理の概要を説明する。

2. Morphological phase (Mph)

形態素解析フェーズでは、形態素発見とともに、構文範疇の決定を行う。この際、形態素範疇から構文範疇へ写像を行う。この写像では、ある構文範疇に写像されたストリングには、形態素情報の集合が特徴素として付加される。こうすることによって、形態素の持つ結合情報が、後のフェーズでも利用し得るとともに、形態素・構文範疇写像によって構文規則に現れる範疇が少なくなり、構文規則の爆発を回避することができる。又、形態素解析内の範疇集合と構文範疇との対応表を処理の外に置くことによって、形態素解析以後の文法モデルの変更を吸収することができる。表1は、写像テーブル内容の例であり、"スル・コト・ガ・デキ・ル・ダロウ"という形態素列から、"スルコトガデキルダロウ"というサ終（可，推）なる構文範疇が得られることを示す。

表1. 形態素範疇・構文範疇写像テーブル

(スル) (コト) (ガ) (デキ) (ル) (ダロウ) サ連 形1 助1 助動3 助動形5 助動4 終	サ終(可,推)
指示代名詞	連体詞
代名詞	名詞
名詞	名詞

3. Syntactically controlled Semantic check phase (Sph)

構文規則によって処理を制御する意味検証フェーズでは、形態素解析フェーズから得られた辞書内容と、M-N、N-Vプラグマティックテーブルが主要な処理データとして使われる。M-N，N-Vテーブルは高度に抽象化されたデータでなく、対象分野（計算機マニュアル）の実テキストからサンプルされた語と語の共起関係を生に近い形式で保持するものである。故にこれらのテーブル中には、語と語が同一の文中で互いに係り受け関係になっている生のデータが、係り受け関係詳とともに貯えられている。M-N，N-Vテーブルは基本的に名詞格関係と動詞格関係を保持しており、表2、表3のような構成を持っている。

表2に示すN-Vテーブルは、Vに対する必須格関係を保持しており、例

表2．N-V プラグマティックテーブル

Verb	Noun	Case		Pointer
ひく	風邪	O	ヲ	P1
取付ける	車	G	ニ	
ひく	人	A	が	
ひく	人	O	ヲ	

翻訳辞書

V CATCH, A TC, O TC

表3．M－N プラグマティックテーブル

Noun	Modifier	Attr.	Pointer
人	車	POsses	
車	車輪	PArt	
風邪	ひどい	DEGree	

中の"ひく"というVと"風邪"は目的格関係に在り、"取付ける"と"車"は対象格関係を満たすという事実を保持している。この表中でP項は解析の早い段階で訳語が決定でき、更に、入力語の訳語としてかなり特殊な訳が望まれる場合、翻訳辞書を直接（むろん最優先的に訳語選択される）指定するものである。"風邪をひく"という場合には、"ひく"に対して"catch"の選択を指定することを表中に示している。表3のM－Nテーブルは名詞格関係を保持し、その構成はN－Vテーブルとほぼ同一である。

両テーブル中のP項と辞書の関係について述べておく。P項は特殊な訳語選択機構を提供しているが、解析用辞書と翻訳用辞書によって行われる訳語選択機構を補って、適切な訳語選択を行うものである。表2中の"風邪をひく"という記述等は、直接熟語として辞書登録する方法も考えられるが、"ひどい風邪をひく"という記述では、構文的構造を認定した後（他例では格関係認定後）に初めて訳語を選択できるようになるため慣用的あるいは特殊表現を全て熟語登録することはできない。また翻訳用辞書中の訳語選択制約と対に"ひく"に対し"catch"を選ぶような指定情報を持たせることも行われているが、このような慣用的表現については、早い段階（解析途上）で訳語を決定し、余分な処理を排除することが有効であると考える。

さて、次に意味検証プロセスでの辞書内容とプラグマティックステーブル情報の関係と処理概要について述べる。

解析用辞書は、図2に示すような内容を持つ。基本的には、構文範疇、構文・意味属性の集合（図中SF_i、V_i）、意味範疇（図中CAT_i）と格構造から成る。ここで、SF_iには、外部辞書中に記述された統語的情報の他に形態素解析から得られた情報が含まれる。

SY			
	V_1 ひく (SF_1…)	CAT_1	A C_1^1, O C_2^1
	V_1 ひく (SF_1…)	CAT_2	A C_1^2, O C_2^2
	V_2 ひく (SF_2…)	CAT_3	A C_1^3

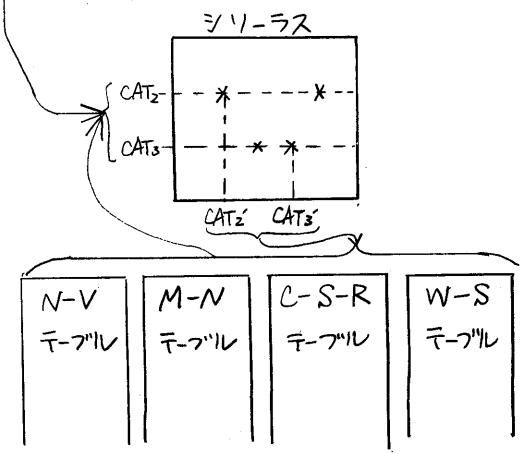


図2．解析用辞書項目と
プラグマティックテーブル

SF_i, V_iは、TRAPシステムで扱った浅い意味処理（実際には統語情報に関する処理をも含む）[2][5]で用いられる情報や付属語情報が記述され、構文規則適用時にいつも参照されるものである。この情報も統語的な語と語の係り関係や、意味的関係の妥当性を構文規則でまとめあげられるコンポーネント間で検証する役割を果す。

この情報の他に、場の意味（対象分野固有の意味、対象の記述目的というプラグマティックスあるいは文中の局所的共起や文内の脈絡）がプラグマティックステーブルに貯えられている。通常SF_i、V_iを用いて抽象化された統語・意味シンボルを基に粗い検証が行われるが、こうしたシンボルだけで、すべての許される係り関係を保持することは非常に困難である。その理由は、個々の異なった係り受けをシンボルの意味の細分化によって区別しようとすれば、それらの間に類似が多くなるため、かえって区別しにくくなり、そこで導入される分類の軸を管理することが"人手では"無理になってくるということである。

プラグマティックステーブルと辞書内容は図2に示すように、シソーラスを介して内部関係をもつ。辞書内容中のCAT_iと格関係記述中の制約C_iは、シソーラスのエントリーに対応し、プラグマティックステーブルの単語もシソーラスのエントリーに対応している。シソーラスはループを含まないネットワーク構造であり、上位ー下位、全体ー部分関係を保持している。

これらのデータを用いて意味検証が進行する。以下にその概略を述べる。このフェーズへは前述したように、構文範疇とその辞書内容が入力として与えられる。構文規則に従って語と語の間の関係が意味的に検証され、句、節へとまとめられてゆく。この段階で、SF_i, V_iの情報で浅い意味検証が行われると同時に、語（あるいはまとめられたコンポーネント）の$\{CAT_i\}$が構文規則で予測される中心語、例えば、名詞句相当コンポーネント中の幹名詞、節の動詞（用言）に伝播され、その中心語の$\{CAT_i\}$と、従属語からの$\{CAT_i\}$によって（シソーラスを介して）N-V, M-Nプラグマティックステーブルが検索される。これによって語句の共起関係の妥当性が検証され　許される関係だけが抽出される。しかし、この意味検証は今だ共起関係の検証を行ったに過ぎず、抽出された内容から格構造を発見することが必要である。

この意味検証フェーズが1つの解析結果を生成したときは、語・句の格関係の候補を全て含んだリスト（木）構造が得られることになる。

4. Normalization phase (Nph)

Nphフェーズでは、省略語の発見、格構造の同定、縮退の後、内部表現としての格構造を生成する。

ここでは前Sphにより生成された格候補情報を元に、解析辞書の格構造パターンへの埋込みを行う。この処理はすべて手続きとして与えられるが、格構造を同定する時点では格の候補には全ての場合について何らかの格関係記述子が付加されている（N-V, M-Nテーブル参照）ため処理上は、格構造パターン中の制約C_iと語の関係をシソーラスをたどって判定する検索操作を行うことになる。この結果必須格を埋めるべき格要素が無いことが発見された段階で文脈に沿って省略語の発見を行う。この省略語処理もN-V, M-Vテーブルを用いて実行される。

さて、格構造が同定されたあとには、C-S-R（Case Structure Reduction）テーブルを用いた縮退が行われる。縮退テーブルは、源言語に固有な構文構

造を言語独立な格構造に縮退するものである。言語独立というのは必ずしもその格構造が理想的なピボット言語としての記述内容を持つことを意味していない。表4にC-S-Rテーブルの内容の一例を示す。

表4. C-S-Rプラグマティックステーブル

行う	O: 開発, G: 装置	開発	G: 装置

一般に日本語では不必要な叙述が多い。特に筆者らの対象であるマニュアル類は本来事実のみが記述されているべきで、その文章は簡潔な報告文であるべきであると考えられる。しかし、表3中の例のように、"装置の開発を行う"という類いの文が非常に多く現れる。英文でこの表現をニュアンスまで含め正しく表現する手法は無いし、事実を伝えるという目的では"装置を開発する"と同一視できると考える。そこでC-S-Rテーブルを用いて格構造の正規化を行う。この処理を施すことにより、格構造が簡潔になりPphフェーズの負荷を減じるばかりでなく、訳文生成時に源言語に依存した処理を不必要にする。

さて、こうした慣用的表現に係わる語用法的問題は、Mph中のM-Sテーブルによっても扱われた。M-Sテーブルでは語の隣接関係だけで決定される叙述表現を扱ったのに対し、Nph中のC-S-Rテーブルでは格関係の変形を伴うという違いがあるため、この正規化は、格関係の同定終了後にNphで行わざるを得ない。

5. Paraphrasing phase (Pph)

Pphでは、Nphの出力結果から翻訳辞書を用いて訳語選択し、最終的に形態素合成を経て訳文を生成する。このフェーズでは、Nphからの正規化された格構造中の語とそれに共起する格要素をキーとして訳語選択が行われる。格構造は、1つの中心語とその格要素から成立しており、その中心語に対応する翻訳辞書エントリーが検索される。そしてそのエントリー中の複数の翻訳パターンの1つを選択し、以下順に部分となっている格構造単位に翻訳パターンを選択してゆく。この翻訳パターンの選択段階では、解析途上に発見されたM-N, N-Vテーブル中のP項で指定される翻訳パターンが優先的に採用される。

さて、全ての翻訳パターンがM-N, N-Vテーブルで指定されていれば理想的な訳出が可能となるが、そうした対応を全て持つことは不可能である。そのために訳語選択に関し、場(特に対象分野)に関する語の組合わせについての例を用いてより適切な訳語選択を行う。このデータは、訳語選択W-Sプラグマティックステーブルと呼ばれる。W-Sテーブルは表5の形式を持つ。基本的に2項関係をもとにした対訳テーブルであり、格による訳し分けを行うものである。翻訳パターンの選択時にW-Sテーブルを引用し、翻訳パターンの曖昧さを解消する。このテーブルを用いる効果は2つある。第1は、こうしたテーブルを用いることによって、翻訳辞書中のパターン選択制限記述を少なくすることができることと、分野での常識を訳出に反映させることができる。第2の点は、訳語選択時の早い段階で曖昧さを解消できることである。W-Sテーブル、翻訳辞書を用いた翻訳パターンの選択の後、構

表5. W-Sプラグマティックステーブル

風邪	ひどい	COLD	BAD

文的変形を行い訳文を生成する。

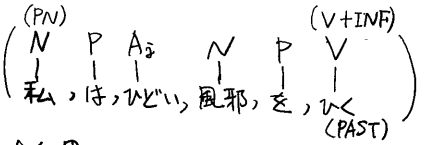
6. 処理例

図3に"私はひどい風邪をひいた"という入力文に対する各フェーズの処理結果を示す。各フェーズでは、表1～表5に示したようなパラグマティックデータを参照しながら処理を行う。

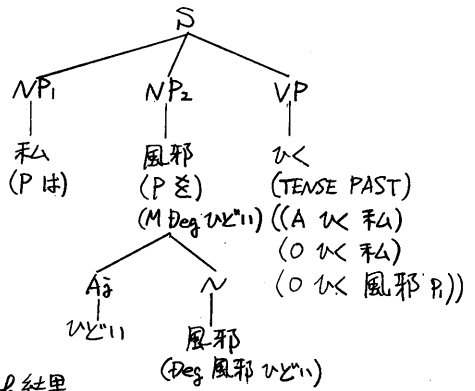
a. 入力文

私はひどい風邪をひいた。

b. Mph結果



c. Sph結果



d. Nph結果

(ひく1 (tense PAST) P1
((A 私 ひく)
(O 風邪 ひく (H 風邪 ひどい))))

e. 出力

I caught a bad cold.

図3. 翻訳処理例

7. おわりに

本稿では、TRAP (Translation by Prolog) におけるプラグマティックステーブルを用いた擬ピボット翻訳方式について述べた。本方式は、具体的なテキストから対象分野固有の意味、文体情報を抽出し、翻訳各フェーズの曖昧さを解消し、高い質の翻訳を可能にする枠組を提供する。

一般に翻訳システムは、大量の非定形な文書を高い精度で翻訳できることが要求されている。高い精度を達成するには分野固有の言語的・非言語的知見をシステムに埋込むことが必要であるが、この種の知識は分野毎に異なり、ある場合には排他的でさえある。そのため、全ての分野に固有な知識を一挙に統一的な知識体系へとまとめあげることは無理がある。

筆者らは上記の問題を解決するため、非言語的性格の強い分野固有の語と語の共起、文体情報、文内容などの知識をプラグマティックステーブルとして形式化し、対象分野に独立な核となる翻訳メカニズムと知識から分離した。こうすることによって、言語内的知識とそのメカニズムだけでは正しい翻訳が困難であった文書の翻訳を、プラグマティックステーブルを活用することで可能にすることができる。更に、このプラグマティックステーブルを分野毎に用意することで、格システムを各々の対象分野に適合させることができる。

今後、TRAPシステムによる、より広汎な実験を通し、本方式に基づく実用的翻訳システム開発を目指す。

[参考文献]

[1] F.C.N. Pereira and D.H. Warren, "Definite Clause Grammar for Language Analysis," Artificial Intelligence, Vol. 13, 1980.
[2] 市山, 村木, "機械翻訳に向けるための一アプローチ," 情処自然言語研資 1980.
[3] 村木, 市山, "日本語質問文解析におけるデータセンティックスの利用," 情処自然言語研資, 1981.
[4] 市山, 村木, Prologを用いた自然言語入力質問応答システム, 情処春全大, 1982.
[5] 市山, 村木, "Prologを用いた翻訳システムにおける意味処理," 情処秋全大, 1982.
[6] 三留, 東野, 村木, "マニュアル文における省略と係り受けの分析" 同上.
[7] 田中元吉, 安川, "意味表現用言語SRLの機械翻訳への応用", 情処自然言語研資 1982.