技術文書の機械翻訳における常識と文脈情報の利用†

向 仲 顯††

現在実用化されている機械翻訳システムでは，構文解析と同時に動詞の格フレームと名詞の意味素性およびヒューリスティックなルールを用いた意味解析を行い，単語の意味や掛り受けの曖昧性を解消している．しかし，これだけでは句と句の関係や多義語の意味などを十分解析することができず，誤った解釈が生じる．このため，常識と文脈情報を機械翻訳に利用する試みが行われている．しかし，あらゆるタイプの文書に適合するような，すべての常識および文脈情報を蓄積するシステムは，当面実用化困難である．本稿は，翻訳対象文書を技術文書に限定することによって，蓄積する常識と文脈情報を最も効果があると考えられるものに限定したモデルを作成し，その場合の翻訳品質の改善効果を調査した結果を述べる．一般的にマニュアル，仕様書，技術論文などの技術文書は，システムや装置の構成，機能，操作法などを述べていることが多い．このため，概念辞書に常識として，単語の表す実体の一般的な属性，構成および機能を記憶させておくと同時に，入力テキストから翻訳時に構成，機能を抽出して文脈情報ファイルに蓄積し，構文・意味解析にこれらの情報を利用する．このモデルを，実際の技術文書で評価した結果，概念辞書の利用により動詞の意味不適切が75%，並列句のスコープ，掛り受けなどの構文エラーが40%改善できた．また，文脈情報の利用により構文エラーが13%改善できた．

1. はじめに

機械翻訳システムでは，構文解析と同時に意味解析を行って，単語の意味や掛り受けなどの曖昧性を解消している．現在実用化されているほとんどの機械翻訳システムでは，動詞の格フレームと名詞の意味素性およびヒューリスティックなルールを用いた意味解析を行っている．

しかし，単語の意味や句と句，句と節の関係などは，格フレームと意味素性だけでは捉えられない場合も多い．このため，並列句のスコープ，連体修飾句や連用修飾句の掛り先などに関する構文エラーおよび動詞の意味不適切などのエラーが生じる．

そのため，常識および文脈情報を利用する自然言語理解の技術を，翻訳に適用する様々な試みが行われている．Dahlgren[1]は，常識を単語の表す実体の分類，特徴などの形でデータベース化しておき，その情報を用いることにより，単語や構文解析の曖昧性を解消するシステムとしてKind Types Systemを提案している．石崎，井佐原らは，文脈情報を有する概念辞書により文脈理解を行うシステムCONTRAST[2),3)]について発表している．CONTRASTの概念辞書は，ある概念が分解されうる複数のシーンに関する情報を有する．

しかし，あらゆるタイプの文書に適合するような，すべての常識および文脈情報を蓄積するシステムは，当面は実用化困難である．

本稿は，翻訳対象文書を技術文書に限定することによって，蓄積する常識と文脈情報を最も効果があると考えられるものに限定したモデルを作成し，その場合の翻訳品質の改善効果を調査したものである．

マニュアル，仕様書，技術論文などの技術文書では，最初にシステムの構成や機能を述べ，それに基づいて詳細な動作や操作法を述べていることが多い．したがって，技術文書では，単語の表す実体の属性，構成，機能などで常識を表すのが適切である．また，技術文書の最初に出てくるシステムの構成や機能説明が，後の動作説明や操作法の説明の理解に役立つ．このため，文脈情報として構成，機能情報を蓄積して利用するようにすればよい．本稿で述べるモデルでは，常識として単語の表す実体の属性，構成および機能を蓄積し，文脈情報として構成，機能情報を翻訳時に抽出して蓄積し，機械翻訳における単語の曖昧性の解消，構文の曖昧性の解消に利用する．

2. 常識を用いた曖昧性解消

技術文書の機械翻訳における誤解釈の原因には，多義語の訳し分けの誤り，掛り受けの誤解釈，並列句のスコープの誤解釈などが多い．これらの場合について，構文・意味解析に単語の表す実体の属性，構成，機能などの常識を用いて曖昧性を解消する方法について検討した．

† Utilization of Common Sense and Context Information in the Machine Translation of Technical Documents by KOU MUKAINAKA (EDP Systems Engineering Division, NEC Corporation).

†† 日本電気(株)情報処理システム技術本部

2.1 多義語の訳し分け

多義語の中でも，用言の訳し分けが翻訳品質に重大な影響を与える．従来の用言の訳し分けは，用言の特定の語義に対応する格パターンと体言の意味素性を使って行っていた．しかし，従来の意味素性は体言を一意に分類しているため，用言の特定の語義が要求する体言の特徴を十分に表すことができなかった．

体言の表す実際の物や事は，各種の特徴を持っている．体言の表す物の特徴を，形状，性質，用途など色々の面から見て複数の属性を与えてやるようにすると，その物の特徴をかなりよく表すことができる．例えば，カードに対しては，人工物，紙状，軽い，数えられる，文字が書けるなどの属性を与えることができる．

用言の特定の語義が要求する体言の特徴を，これら属性により表すことができる．例えば，「穴を開ける」という日本語の表現を考えた場合，対象が紙状の物であれば，英語の punch に，板状または塊状の物であれば，drill に訳し分けることができる．

このように，ある用言が複数の語義を持っている場合に，各語義ごとの格フレームが要求する体言の属性に違いが見られる．例えば，**表1**に示すように，日本語の用言「入れる」は，「を格」が「固体」であれば「中に納める」という意味になり，「液体」であれば，「そそぐ」という意味になり，「スイッチ」であれば「機能させる」という意味になる．

このような，格フレームが要求する体言の属性の違いによって，用言の意味を訳し分ける．すなわち，用言の特定の語義が要求する属性を体言が持っているかどうかを調べ，属性の一致する語義を選択する．**図1**

フロッピィディスクを (固体) 箱に (容器) 入れる. (納める)
ビーカに (容器) 湯を (液体) 入れる. (そそぐ)
最初に電源スイッチを (スイッチ) 入れて下さい. (機能させる)

図1 体言の属性値と格フレームを用いた用言の訳し分けの例
Fig. 1 Examples of lexical ambiguity resolution by nominal attributes filling slots of case frames.

の例で，「入れる」という用言の「中に納める」という語義は「を格」に「固体」を，「に格」に「容器」を要求するので，「フロッピィディスク」に「固体」という属性があるかどうか「箱」に「容器」という属性があるかどうかを調べ，見つかればその語義に決定する．

属性に上位／下位関係がある場合は，体言が持っている属性に対して，用言の要求する属性が上位属性であれば条件を満足するものとする．

図2に示すように，概念の表す実体の属性は概念辞書に登録される．概念辞書は日本語，英語など各言語に対して共通に設けられ，概念の表す実体に対する常識を蓄積する．

各言語固有の情報は各々の語彙辞書に記述される．したがって，日本語の用言の各語義に対して各フレームが要求する体言の属性は，日本語の語彙辞書に記述される．各言語の語彙辞書の表層語は，品詞を中継してパスにより概念辞書の概念と結ばれる．語義が違えば，おのおの別の概念と結ばれる．逆に，同義語は同一概念に結ばれることになる．

属性の種類は動詞の訳し分けを目的として選定した．技術文書に使用される動詞約 3,000 語[10),11)]の中

表1 用言の各語義に対して格フレームが要求する体言の属性の例
Table 1 Examples of nominal attributes that fill slots of case frames corresponding to each sense of a verb.

日本語用言	語義	格フレームが要求する体言の属性 (が格)	(を格)	(に格)	…
入れる	中に納める (PUT-IN)	人 (HUMAN)	固体 (SOLID)	容器 (CONTAINER)	
	そそぐ (POUR-IN)	人 (HUMAN)	液体 (LIQUID)	容器 (CONTAINER)	
	加入する (ENTER)	人，組織 (HUMAN, ORGANIZATION)	人 (HUMAN)	集団，組織 (GROUP, ORGANIZATION)	
	機能させる (TURN-ON)	人 (HUMAN)	機械，スイッチ，火 (MACHINE, SWITCH, LIGHT)		
	ロードする (LOAD)	人 (HUMAN)	情報，材料 (INFORMATION, MATERIAL)	機械 (MACHINE)	

から，多義性の大きいもの約180個を選び，これらの動詞の語義を訳し分けるために必要な属性約220個を選定した．その他の動詞に対して必要な一般的な属性約40個を加え，全体で約260個の属性を定めた．各概念に対する属性は，この260個の中から，その概念の特徴をよく表すと考えられる属性だけを付与することにした．

2.2 掛り受けの解析

従来，掛り受けの解析は，格関係および句読点や修飾句と被修飾句の位置関係などを手掛りとして行われてきた．しかし，これらの解析は，手掛りの発生頻度に依存しており，発生頻度の低いケースが表れると誤って解釈される．

このため，従来の手掛りを用いて，掛り受けの候補を複数抽出し，これらの候補について，修飾句と被修飾句が機能的に関係があるかどうか，または構成要素と全体の関係にあるかどうかを調べ，関係のある句と句の関係を優先させるようにすれば，より確実に掛り受けの解析を行うことができる．

概念辞書には，前述のように概念の表す実体の持つ属性が登録されるが，そのほかに**図3**に示すように，ある概念の表す実体の構成，機能，上位／下位／関連概念などの常識を登録しておく．概念辞書を照合することにより，ある句と句が機能的に関係があるかどうか，部分全体関係にあるかどうかを調べることができる．

例えば，**図4**(a)の文を従来の方法で解析すると「フロッピィディスクに」は「送った」に掛り誤った解釈が生じる．しかし，**図5**に示すように，概念辞書に，フロッピィディスクの機能として，「データを記録する」というのがあれば，「フロッピィディスク」は，「記録する」に掛ることを知ることができる．

このように，連体／連用の別，格関係，句読点，修飾句と被修飾句の位置関係などの手掛りにより，複数の掛り先候補を抽出する．これらの候補について，機能上の関係や部分全体関係を評価し，関係のある句と句の関係を優先させる．

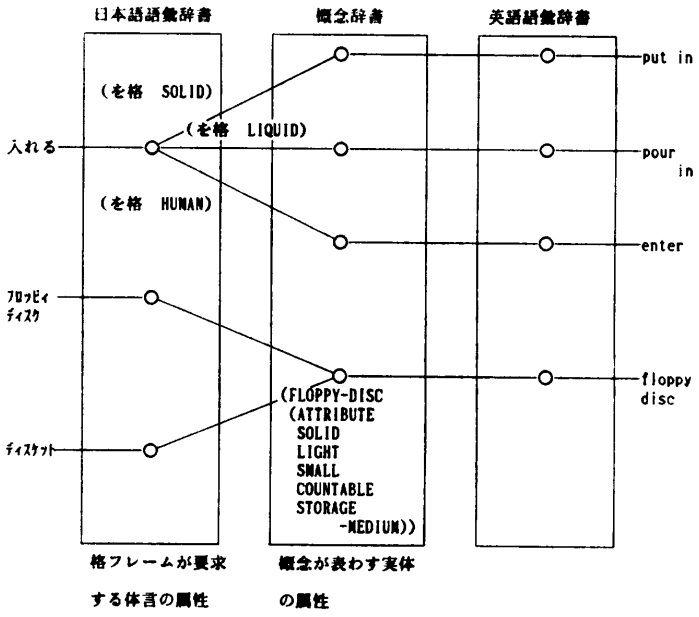

図2 多義語の訳し分けにおける語彙辞書と概念辞書の関係
Fig. 2 Relation between the concept dictionary and each language lexicon for lexical ambiguity resolution.

```
(概 念  (ATTRIBUTE 属性1 属性2 ……)
        (HAS-PART 構成要素1 構成要素2 ……)
        (FUNCTION (概念1 (関係1 概念2)
                         (関係2 概念3)
                         ……)
                  (概念n ……
                  ……))
        (THESAURUS (BT 上位概念)
                   (NT 下位概念1 下位概念2 ……)
                   (RT 関連概念1 関連概念2 ……)))
```

図3 概念辞書の形式
Fig. 3 Format of the concept dictionary.

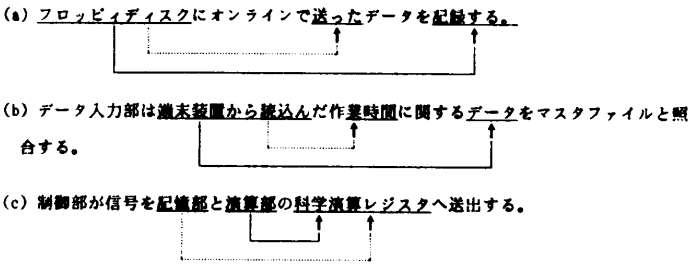

図4 構成・機能情報を用いた掛り受けおよび並列句の解析例
Fig. 4 Examples of structural ambiguity resolution by general compositions and functions of entities that are represented by words.

構成・機能関係だけで決まらない時は，従来の手掛りにもとづき発生頻度による優先順位によって選択するようにする．図4(b)に連体修飾句の掛り受けの解析の例を示す．

機能上の関係を解析する時に，機能を表す概念は必ずしも文中に使われている概念と一致するとはかぎらない．このため，概念辞書のシソーラスを用い，文中に表れる概念の上位語，下位語，関連語が，特定の概念の機能と一致しても機能上の関係があるものとする．

2.3 並列句のスコープ

並列句のスコープの決定は，従来から句読点，助詞，句の長さなどの表層的な手掛りにより解析されてきた．しかし，これらの手掛りも発生頻度に依存しており，発生頻度の低いケースが表れると誤って解釈される．

このため，並列句のスコープの解析についても，掛り受けの解析と同様に，従来の手掛りで複数の候補を抽出し，機能的な関係および部分全体関係を評価し，関係のある候補を選択する．例えば，図4(c)で，科学演算レジスタが演算部の構成要素の1つであることがわかっていれば，「記憶部」と「演算部の科学演算レジスタ」で並列句を構成していることがわかる．従来のヒューリスティックな方法で解析したのでは誤って解釈される．

3. 文脈情報を用いた曖昧性解消

機械翻訳の過程で，構文・意味解析の結果として概念構造が生成される．この概念構造から構成・機能情報を抽出し，文脈情報ファイルに登録する．文脈情報ファイルに蓄積した文脈情報は，構文・意味解析の過程で利用する(図6)．

3.1 文脈情報の抽出と蓄積

構成・機能情報の抽出にあたり，否定，推量，疑問などを表す文は避けなければならない．そのため，断定，可能などを表す文だけ選択し，次の手順により，概念構造中の各概念の表す実体の構成・機能情報を抽出する．

(1) 用言および助動詞による判断の表現を解析し，断定，可能などを表す文だけ選択する．

(2) 概念構造から，物と物との関係が部分全体関係を表している部分を抽出し，全体を表す概念に対する構成情報とする．

(3) 概念構造から，動作動詞の格に道具・手段など機能を持つ物を有する単文を抽出し，道具・手段などを表す概念に対する機能情報とする．

```
(FLOPPY-DISK (ATTRIBUTE SOLID LIGHT SMALL COUNTABLE STORAGE-MEDIUM)
    (HAS-PART MAGNETIC-DISK FD-JACKET)
    (FUNCTION (STORE (OBJ DATA))
    (THESAURUS (BT    MAGNETIC-DISK)
               (NT    MICRO-FLOPPY
                      MINI-FLOPPY)))
```

図5 概念辞書の例
Fig. 5 An example of the concept dictionary.

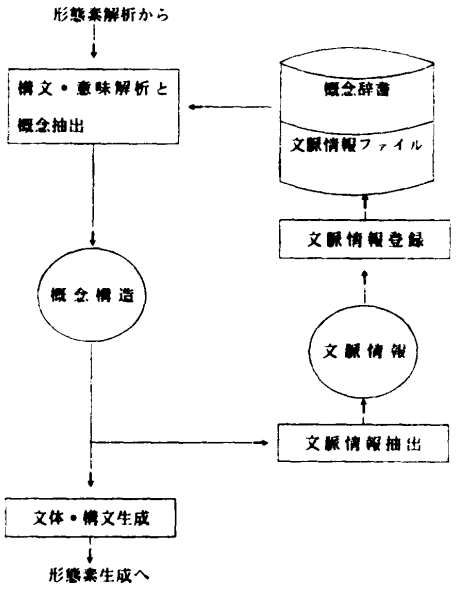

図6 文脈情報の抽出と蓄積
Fig. 6 Extraction and storage of context information.

抽出した概念は，翻訳中の文書に一度出てきた語で，文脈情報ファイルに既に登録されている概念かどうかを，情報ファイルを検索して判別する．また，概念辞書に登録されているかどうかも検索し，次の手順により抽出した構成・機能情報を，文脈情報ファイルに登録する．

(1) 文脈情報ファイルに概念が登録されている場合には，登録されている内容と比較し，新規情報の時だけ追加登録する．

(2) 文脈情報ファイルに概念が登録されていない場合には，概念辞書を検索し，概念辞書にある場合は，文脈情報ファイルに新規登録し，概念辞書にリンクする．

(3) 概念辞書にない場合は，そのまま新規登録す

表 2 モデルの評価結果
Table 2 Evaluation result of the model.

項目		格フレームと頻度だけでは間違う文の比率	名詞の属性情報により改善する比率	構成・機能情報により改善する比率	文脈情報の利用により改善する比率
構文エラー	連体修飾句の掛り先	2% (8)	0	50% (4)	0
	連用修飾句の掛り先	8% (28)	0	29% (8)	14% (4)
	並列句のスコープ	7% (24)	0	50% (12)	17% (4)
	全体	16% (58)	0	40% (23)	13% (8)
動詞の意味の訳し分け		19% (68)	75% (51)	0	0

注) 評価文は全体で357文あり，カッコ内に該当する評価文の文数を示す．

る．

3.2 文脈情報の構文・意味解析への利用

蓄積した文脈情報は，構文・意味解析過程での掛り受けの解析および，並列句のスコープの決定に利用する．利用方法は，概念辞書の構成・機能情報の利用と同様で，修飾句と被修飾句が機能的に関係があるかどうか，または構成要素と全体の関係にあるかどうかを調べ，関係のある句と句の関係を優先させる．

常識と違って，文脈情報は翻訳対象の技術文書にしか書かれていないような，具体的な情報を含んでいる．このため，文脈情報ファイルを先に検索し，関係が見つからない時は概念辞書も調べるようにする．

4. モデルの評価

上述のモデルを，技術論文(NEC 技報から2点)およびマニュアル（ソフトウェアの手引書）に適用して調査した．調査にあたって，従来の格フレームと，発生頻度に基づく解析では問題が生じる文を次のケースについて抽出した．

（1） 連体修飾句の掛り先に問題の出る文

連体修飾句が直後の体言に掛らない場合および連用修飾句と間違う場合

（2） 連用修飾句の掛り先に問題の出る文

連用修飾句の掛り先が格フレームだけでは判別できない場合

（3） 並列句のスコープに問題の出る文

並列句のスコープが表層の手掛りでは間違う場合

（4） 動詞の意味の訳し分けに問題の出る文

動詞の意味が格フレームと意味の使用頻度では訳し分けられない場合

(1)～(3)に対して，概念辞書の構成・機能情報を用いて解析することにより改善できる文および文脈情報ファイルを用いて解析することにより改善できる文を求め，抽出した文との比率を求めた．

(4)に対して，名詞の属性情報を用いて解析することにより改善できる文を求め，抽出した文との比率を求めた．

結果は表2に示すように，動詞の意味の間違いは概念辞書の名詞の属性情報により 75% 改善できる．連体修飾句，連用修飾句の掛り先，並列句のスコープなどの構文エラーについては，概念辞書の構成・機能情報を用いることにより平均40%，文脈情報ファイルの構成・機能情報により平均 13% 改善できる．

5. むすび

技術文書を対象にして，機械翻訳の翻訳品質を向上させるために常識と文脈情報を用いるモデルを作成し評価した．常識として，体言の表す実体の属性・構成・機能を蓄積し，文脈情報として文章中に出てくる体言の構成と機能情報を蓄積しておいて利用する方法をとった．評価結果は，常識の利用については比較的良好な結果が得られたが，文脈情報の利用については必ずしも満足する結果は得られなかった．

本稿で提案したモデルは記憶容量および処理量が従来の2倍以下で実現することを目標にしたので，きわめて限定した常識と文脈情報を用いている．さらに情報を増やした場合の検討，効率的な処理方法，および動詞の訳し分け，掛り受けの解析，並列句の解析以外に適用した場合の検討は，今後の課題である．

謝辞 本稿をまとめる機会を与えていただいた日本電気(株)登家取締役，内田本部長に感謝するとともに，検討を進めるにあたり協力をいただいた石森部長，小関課長および中国日本電気ソフトウェア(株)青木部長に謝意を表す．

参考文献

1) Dahlgren, K.: *Naive Semantics for Natural Language Understanding*, p. 102, Kluwer Academic Publishers, Massachusetts (1988).
2) 石崎，井佐原：文脈情報翻訳システム CONTRAST，情報処理，Vol. 30, No. 10, pp. 1240-1249 (1989).
3) 石崎，井佐原，徳永，田中：文脈と対象世界モデルを利用した機械翻訳に向けて，人工知能学会誌，Vol. 4, No. 6, pp. 660-670 (1989).
4) 稲垣，壁谷，小橋：意味連結パターンを用いた係り受け解析，情報処理学会自然言語処理研究会報告，67-5 (1988).
5) 牧野，小関：統合自動翻訳システム PIVOT, bit 別冊 機械翻訳，pp. 184-190 (1988).
6) 市山，村木：機械翻訳システム PIVOT の中間言語，第38回情報処理学会全国大会論文集，pp. 345-346 (1988).
7) 野村，村木：機械翻訳システム PIVOT の日本語格フレームモデル，第 38 回情報処理学会全国大会論文集，pp. 390-391 (1988).
8) 野村，村木：機械翻訳システム PIVOT における格パターンの処理，第 38 回情報処理学会全国大会論文集，pp. 392-393 (1988).
9) 井上：変形文法と日本語，大修館書店 (1976).
10) インタープレス対訳センタ：技術英文を書くための動詞辞典，アイピーシー (1987).
11) 小泉，船越，本田，仁田，塚本：日本語基本動詞用法辞典，大修館書店 (1989).
12) 長尾，辻井，田中，石川：科学技術論文における並列句とその解析，情報処理学会自然言語処理研究会報告，36-4 (1983).
13) 長尾，辻井：機械翻訳における訳語選択と構造変換過程，情報処理，Vol. 26, No. 11, pp. 1261-1270 (1985).
14) 横尾，林：日本語埋め込み構造の解析，昭和 62 年度人工知能学会全国大会論文集，pp. 369-372 (1987).
15) 長尾，辻井，中村，坂本，鳥海，佐藤：科学技術庁機械翻訳プロジェクトの概要，情報処理，Vol. 26, No. 10, pp. 1203-1213 (1985).

(平成2年1月19日受付)
(平成2年4月17日採録)

向仲　頼（正会員）

1958年九州大学工学部電気工学科卒業．同年日本電気(株)入社．現在，情報処理システム技術本部副技師長．ミニコン，オフコンのオペレーティングシステムの開発，オフコンの標準アプリケーションの開発，機械翻訳システムの開発に従事．人工知能学会会員．