

英日機械翻訳システムにおける 英文の解析方法について

野上宏康 杉浦裕美子 平川秀樹 天野真家
(株)東芝 総合研究所

我々は既に科学技術文献等を対象にした英日機械翻訳システムを開発した。実際の英文を大規模な辞書と文法を用いて解析する場合には、(1) 多品詞による曖昧さの解消、(2) 構文解析に要する時間の増加への対処、(3) 係り受け関係・並列関係の曖昧さのような構文的曖昧さの解消、が大きな問題となる。

本稿では、これらの問題を解決するために以下の解析方法が有効であることを具体的な現象例を用いて述べる。

- (1)(2)の問題に対しては、局所的な情報だけではなく、先読みにより入力文全体の情報も用いて解析する。
- (2)(3)の問題に対しては、構文的曖昧さを含む文を解析する場合、構文解析ではこの曖昧さを内在する唯一つの解析結果だけを構築し、後の意味解析で曖昧さを解消する。

English Analysis in an English-Japanese Machine Translation System

Hiroyasu Nogami Yumiko Sugiura Hideki Hirakawa Shin-ya Amano

Toshiba Research and Development Center
1, Komukai Toshibacho, Saiwai-Ku, Kawasaki-City, Kanagawa, 210 Japan

We have developed an English-Japanese machine translation system for the treatment of scientific and technical documents. Analysing realistic documents using large-scale dictionaries and grammars requires that the following problems be solved: (1) categorial disambiguation, (2) reduction of parsing time, (3) disambiguation of syntactic ambiguities such as prepositional attachment and coordinations.

This paper describes methods which are used in the solution of such problems, as follows:

- o Parsing method reinforced by use of global information obtained by a look-ahead function.
- o Disambiguation method for syntactic ambiguities, in which the parser produces a unique result incorporating ambiguities which are disambiguated by semantic.

1. はじめに

我々は、既に科学技術文献、マニュアル等を対象とした英日機械翻訳システムを開発した。実用システムで、実際の英文に現われる種々の複雑な言語現象を、大規模な辞書と文法を用いて解析する場合には、多品詞による曖昧さの解消、構文解析に要する時間の増加への対処、係り受け関係・並列関係の曖昧さのような構文的曖昧さの解消が大きな問題となる。

本稿では、これらの問題に対して、2章で、多品詞による曖昧さを解消する方法、3章で、構文解析に要する時間を短縮する解析方法、4章で、構文的曖昧さを解消する方法について、CFGモデルに基づき具体的な現象例を用いて述べる。

2. 多品詞現象

英単語は、複数の品詞になるものが多い。例えば、多くの名詞は動詞にもなり、また、規則動詞の～ed形は過去形にも過去分詞形にもなる。実用の大規模な辞書を用いる場合、1文中には多くの多品詞語が存在し、これらの組み合わせの数は非常に大きくなるので解析の効率が悪くなる。しかし、これらの品詞列の中で確からしい品詞列の順に解析を行えば効率を上げることができる。

ここでは、確からしい品詞を優先または決定するために、特に、大域的に入力文全体を眺めることが有効である例を示す。なお、これらの適用においては、()等で囲まれた挿入の部分は除く必要がある。

- 文中の等位接続詞の後方に動詞があれば、等位接続詞より前方に動詞が存在する。

ex. …V …… and V…

ex. …V …… and ……V ……

- which, who, whoso, how, what, whichever 等の後方には必ず動詞が存在する。

ex. which ……動詞……………

- 同じ単語は同じ品詞を優先する。

- 文頭に従属接続詞と前置詞の両方になりうる単語がある場合、後方に動詞が1つしか存在しない場合には前置詞を優先する。

- [that 冠詞], [従属接続詞 名詞] の後方には動詞が存在する可能性が高い。

ex. that the………V ……

ex. because the ………V ……

- 等位接続詞の後方の規則動詞の～ed形は、前方にhaveかbeと過去分詞形が存在すれば、過去分詞形を優先する。

ex. …have partially exposed the plates or fogged (過去形、過去分詞形) the…

- 過去を表わす副詞があったら、過去形を優先する。

ex. I read (現在形、過去形、過去分詞形) the book yesterday.

- when の後方に動詞が存在しないならば、関係副詞の可能性はない。

- 等位接続詞の直後の動詞は、前方に助動詞が存在すれば、原形を優先する。

ex. …would V ……and cut (原形、過去形、過去分詞形) ……

3. 一般的な構文現象と文法

ここでは、科学技術文献、マニュアル等によく出現する構文現象を先読みによって効率よく解析する方法について述べる。

3-1 句構造現象

英語の構造は、いくつかの要素が集まって1つの句が構成され、これらの句が集まってさらに大きな句が構成される。これらの句を構成するには、不可欠な要素が存在する。例えば、英語の句には名詞句、前置詞句、動詞句、従属節、文などがあり、文には名詞句と動詞句、前置詞句には前置詞と名詞相当語が不可欠な要素である。一般に、要素aを不可欠とする句 α は①のように表現できる。また、aを含まない句 β は②のように表現できる。

① α 句: □□□□ a □□□□

② β 句: □□□□□□□□

③文法規則: $\alpha \rightarrow \dots a \dots$

④文法規則: $\alpha : [a \text{ の存在を先読みチェック}]$
 $\rightarrow \dots a \dots$

①の α 句は、概念的に記述した③の文法規則を用いることによって解析できる。しかし、ここで問題なのは、

②のような a を含まない句に対しても、③の規則を適用し無駄な処理を行うことである。そこで、③の規則の適用を制限し、この規則を適用する際に入力文の後方に a が存在するか先読みによって調査し、存在する場合のみ③の規則を適用するようにすれば、a を含んでいない②の句に対する無駄な処理を回避できることになる。この先読みによって制限した規則を④に示す。

現象例とその文法記述

(1) 句の並列現象

句の並列現象は、句を構成するのに不可欠な要素の存在を、先読みで調べることにより効率良く解析できる。文並列の例を①に、名詞並列の例を②に示す。また、各々の文法規則を③、④に示す。

①NP VP NP and NP pre NP VP NP.

②NP1 VP NP2 and NP3 pre NP4.

③文:[動詞の存在を先読みチェック] →名詞句・動詞句

④名詞句→名詞・前置詞句

②の文を解析する場合を考える。andの左側(NP1 VP NP2)で、名詞句(NP2)、動詞句(VP NP2)、文(NP1 VP NP2)のいずれも成功するので、andの右側には名詞句、動詞句、文の可能性を試行することになる。ここで問題なのは文の可能性を試行する場合に、「NP3 pre NP4 を文の主語として解析する」という無駄な処理を行うことである。

③の規則は、文の規則を適用する際に、入力文の後方に文の不可欠の構成要素である動詞が存在するか先読みして調べ、動詞が存在する場合のみ「文の規則」を適用することを意味している。この適用条件により、②のandの右側に対して、③の「文の規則」は適用されず、「NP3 pre NP4 を文の主語として解析する」という無駄な処理を回避することができる。

(2) 動詞による一般的現象

一般に動詞は複数のパターンをとり、使用頻度の高い動詞ほどその傾向が高くなる。このような動詞を含む文を解析する場合に問題なのは、その動詞がとりうる全てのパターンを試行し無駄な処理を行うことである。この場合も、あるパターンを試行する際に、そのパターンに不可欠な要素が存在するかどうか先読みして調べることで無駄な解析を回避することができる。

目的語にthat節と名詞句の両方をとりうるsupposeを含む文を①に、文法規則を②③に示す。

①An chemist supposes the chemical nature of something …… (V は非存在) …….

②動詞句→ suppose: [動詞の存在を先読み]・that節

③動詞句→ suppose・名詞句

supposeの後方の動詞の存在を先読みすることによって、「the chemical以下を②のthat節の主語として解析する」という無駄な処理を回避でき、直ちに③の名詞句の処理から始めることができる。

(3) 名詞句が連続する現象

名詞句が連続する現象は多品詞語の存在や挿入現象によってよく出現する。他動詞と目的語の間に前置詞句が挿入されて名詞句連続が出現する現象の例を①②に示す。

①NP1 vt pre NP2 NP3 pre NP4…………….

②… instruction indicates to the program the …

この現象を解析する場合に問題なのは、2番目の名詞句NP3以降が1番目の名詞句NP2の関係詞省略の関係代名詞節となる可能性を試行することである。この場合も、関係代名詞節の不可欠な要素である動詞が存在するかどうか先読みして調べるにより無駄な解析を回避することができる。

3-2 相関的な構文現象

英文には、as~as, so~that, not~but, either~or, it~that, the比較級~ the比較級のような相関的な構文現象が出現する。このような相関的な構文現象は、一般に①のように表現できる。また、非相関現象は②のように表現できる。

①相関的現象: □□□■□□□■□□□

a1 a2

②非相関現象: □□□■□□□□□□□□

a1

③α→a1……a2

④α→a1: [a2の存在を先読みチェック] ……a2

①に示す相関的現象は③の文法規則を用いることによって解析できる。しかし、ここで問題なのは、②のような非相関現象に対しても、③の規則を適用し無駄な処理を行うことである。そこで、③の規則の適用を制限し、a1を読み込んだ時点で、入力文のa1より後方にa2が存在するか先読みして調べ、a2が存在する場合のみ③の規則を適用するようにすれば、②のような非相関現象に対する無駄な処理を回避できることになる。この先読みによって制限した規則を④に示す。

現象例とその文法記述

相関的な構文現象の例として it …that構文を取り上げて説明する。①に、itが真主語の構文の例を示す。また、it構文の文法規則を②に示す。

①itが真主語の構文：

It is the important point by way of some……

②it構文の文法規則：

文→it：[thatの存在を先読みチェック] …that…

②の規則は、itを読み込んだ時点で、itより後方にthatが存在するか先読みして調べ、thatが存在する場合のみ②の規則を適用することを意味している。従って、it構文でない①の文に対して②の規則は適用されず、直ちにitが真主語の構文の解析を始めることができ、it構文としての無駄な解析を回避することができる。

3-3 カンマを伴う並列現象

カンマを伴う並列現象は、一般に①のように表現できる。また、区切り・挿入表現などの、カンマを伴う非並列現象は②のように表現できる。

①並列現象：a1, a2 and a3

②非並列現象：……a1, a2 ……

③a→a1・カンマ・a2・and・a3

④a→a1・カンマ:[等位接続詞存在を先読みチェック]
・a2・and・a3

①に示す現象は、③の文法規則を用いることによって解析できる。しかし、ここで問題なのは、②のような非並列現象に対しても③の規則を適用し、並列としての無駄な処理を行うことである。そこで、③の規則の適用を

制限し、カンマを読み込んだ時点で入力文のさらに後方に並列接続詞が存在するか先読みして調べ、並列接続詞が存在する場合のみ③の規則を適用することにする。この先読みによって制限した規則を④に示す。これにより、②のような非並列現象に対する無駄な処理を回避できることになる。

現象例とその文法記述

動詞句の並列でない現象例を①に、動詞句並列の文法規則を②に示す。

①To confirm it, see the text.

②動詞句→動詞句1・カンマ:[等位接続詞の存在を先読みチェック]・動詞句2・and・動詞句3

③カンマ:[動詞、等位接続詞、動詞の存在を先読みチェック]

②の規則は、カンマを読み込んだ時点で、等位接続詞が存在するか先読みして調べ、等位接続詞が存在する場合のみ②の規則を適用することを意味している。従って、②の規則によって「①のconfirm itとsee the textを並列として試行する」という無駄な解析を回避することができる。また、さらに③に示すように適用条件を強くすることも可能である。③の適用条件は、「カンマの間、及び並列接続詞の後方に動詞が存在すること」を要求している。

3-4 文頭修飾現象

文頭に従属節、不定詞句、前置詞句などの修飾要素が、主文との間にカンマを伴って出現する現象は、一般に①のように表現できる。

①□□□□□□, □□□□□□□□

修飾要素 主文

①のような表現を解析する場合に問題なのは、修飾要素と主文との区切りのカンマを別の機能のカンマ(並列、副詞句の区切り、挿入等の機能を持つカンマ)として、先に解析する可能性があることである。そこで、文頭修飾要素の処理中にカンマを読み込んだ場合、先読みによって入力文のさらに後方にカンマが存在するか調査し、カンマが存在しない場合は、この読み込んだカンマを主

and, or, if, when等の直後、前置詞の直後、動詞・助動詞の直後、名詞句の直後、関係代名詞の直後、thatの直後等がある。

(11)については、⑥の規則の適用を先読みによって制限し、挿入の先頭のカンマを読み込んだ時点で、後方にカンマが存在するかどうか先読みして調べ、カンマが存在する場合のみ⑥の規則を適用するようにすれば、非挿入現象に対する無駄な処理を回避できることになる。この先読みによって制限した規則を⑦に示す。但し、この解析は非挿入現象を挿入現象として誤って解析する可能性があるため、最も優先度を低くしておく必要がある。

現象例とその文法記述

主語の直後に名詞句が挿入されている例を①に、文法規則を②③に示す。

①挿入現象：

The op-code extending register, OER, is used as...

②文→主語・挿入・動詞句

③挿入→カンマ:[カンマの存在を先読みチェック]

・名詞句・カンマ

③の規則は、主語の直後のカンマを読み込んだ時点で、後方にカンマが存在するかどうか先読みして調べ、カンマが存在する場合のみ挿入の規則を適用することを意味している。この先読みによる適用条件を使うことによって挿入現象を効率良く解析することが可能となる。

また、本方法で効率良く解析できる他の例を④～⑧に示す。⑦の2番目のカンマは、挿入と、従属節と主節の区切りの両方の機能を、また⑧の2番目のカンマは挿入と、主節と分詞構文の区切りの両方の機能を兼ねている。本方法では同じ枠組みの中で挿入部と非挿入部を解析するので、このような挿入に対しても効率よく解析できる。また、挿入部の係り先の決定に関しては4章で述べる。

④bubbles, entering the, will follow ...

⑤machine was, as the user expected, operated ...

⑥The computer knows, and very wisely, that ...

⑦If.....register, which, the

⑧NP VP NP, which, ING

4. 構文的に曖昧な現象

ここでは、係り受け関係の曖昧な現象、並列関係の曖昧な現象の効率の良い解析方法について述べる。

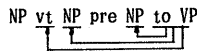
4-1 係り受け関係の曖昧な現象

係り受け関係の曖昧な現象例を以下に示す。

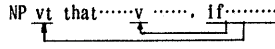
・名詞・動詞と前置詞



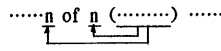
・名詞・動詞と不定詞



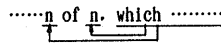
・動詞と従属節



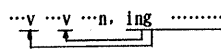
・挿入



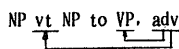
・関係節と名詞



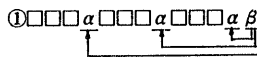
・動詞と分詞構文



・動詞と副詞



上述の係り受け関係の曖昧な現象は一般に①のように表現できる。この現象は②③の規則を用いることによって解析できる。



②αP → α・β

③αP → α

④αP → α : [②の規則失敗の場合]

しかし、この場合には次のような問題が存在する。

・βの解析を係り受け関係の曖昧な回数だけ行ってしまい、βが大きな要素の場合は特に効率が悪くなること

・複数の曖昧な構造を独立して出力するので、複数の候補間に関係がなく、この中から最も確からしい係り受け関係を決定する処理に適さないこと

そこで、構文解析では、この曖昧さを内在する唯一の概念構造だけを構築し、その後、語彙や意味に依存する情報等を使って曖昧な係り受け関係に優先度を付け、最も確からしい係り受け関係を決定することにする。唯一の解析結果だけを構築するためには、例えば、③の規則の適用を制限し②の規則に失敗した場合のみ適用するようにすればよい。この制限された規則を④に示す。この場合は係りうる最も近い要素にかかることになる。以下に、この方法が有効な理由を列挙する。

- ・βの解析を2回以上行なうことを回避でき、効率がよくなる。
- ・係り受け関係の曖昧さを内在した唯一の構文解析結果に対して係り受け関係を決定する処理をするので、優先度を付ける処理に適している。
- ・特に、句構造規則では扱いにくい挿入の係り受け関係等の曖昧さを解消する処理に適している。
- ・構文規則と係り受け関係の優先規則とを明確に区別できる。

現象例とその解析方法

(1) 名詞と関係節の係り受け関係の曖昧な例を①に示す。また、文法規則を②③に示す。

- ①…data in the register which is transferred …
- ②名詞句→冠詞・名詞・関係節
- ③名詞句→冠詞・名詞：[②の規則失敗の場合]

③の規則は、②の規則が失敗した場合に適用することを意味している。従って、②③の規則による構文解析の結果、which以降の関係節は、係り得る最も近い要素で

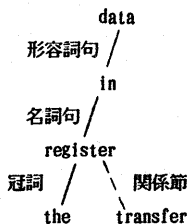


図1 構文解析結果

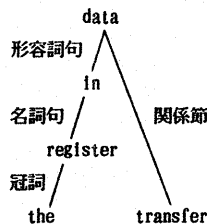


図2 係り受け解析結果

あるregisterにだけ係ることになる。この結果を図1に示す。次に、図1の結果に対して次の規則

- ④関係節は名詞に係る
- ⑤dataは register より transfer の対象になりやすいを適用することによって、関係節が data に係ることが決定でき、図2の結果を得ることができる。ここで、④の規則は、係り受け関係の曖昧さを表現する規則であり、⑤の規則は語彙の意味に依存した係り受け関係の優先度に関する規則である。

(2) 名詞と挿入部の係り受け関係の曖昧な例を①に、係り受け関係に関する規則を②に示す。この場合も(1)と同様の処理によって挿入部はdeviceに係ることが決定できる。

- ①…devices invoked by the operation(a printer or a counter)…
- ②deviceはprinter とcounter の上位概念である

4-2 並列関係の曖昧な現象

並列関係の曖昧な現象は一般に①のように表現できる。この現象は、②③のような規則を用いることによって解析できる。

- ①並列現象： $\square \alpha \square \alpha \square \alpha$ and α
- ② $aP \rightarrow a \cdot \text{and} \cdot a$
- ③ $aP \rightarrow a$
- ④ $aP \rightarrow a$ ：[②の規則失敗の場合]

しかし、この場合には次のような問題が存在する。

・and の右側のαの解析を並列の曖昧な回数だけ行い、αが文や節などの大きな要素の場合は特に効率が悪くなること

・複数の曖昧な構造を独立して出力するので、複数の候補間に関係がなく、この中から最も確からしい並列関係を決定する処理に適さないこと

そこで、構文解析では、この曖昧さを内在する唯一の概念構造だけを構築し、その後、構文パターン・語彙や意味に依存する情報等を使って曖昧な並列関係に優先度を付け、最も確からしい並列関係を決定することにする。唯一の解析結果だけを構築するためには、例えば、③の

唯一の解析結果だけを構築するためには、例えば、③の規則の適用を制限し、②の規則に失敗した場合のみ適用するようにすればよい。この制限された規則を④に示す。この場合は、並列関係になりうる最も近い要素と並列になることになる。以下に、この方法が有効な理由を列挙する。

- ・αの解析を2回以上行うことを回避でき、効率がよくなる。
- ・並列関係の曖昧さを内在した構文解析結果に対して並列関係の決定処理をするので、優先度を付ける処理に適している。
- ・構文規則と並列関係の優先規則とを明確に区別できる。

現象例とその解析方法

名詞の並列関係の曖昧な現象例を①に示す。また、文法規則を②～⑥に示す。

- ① the contents of the register and the output of the detector
 ②親名詞句→名詞句・and・名詞句
 ③親名詞句→名詞句:[②の規則失敗の場合]
 ④名詞句→冠詞・名詞・前置詞句
 ⑤名詞句→冠詞・名詞:[④の規則失敗の場合]
 ⑥前置詞句→前置詞・親名詞句

③の規則は、②の規則が失敗した場合に適用することを意味している。従って、②～⑥の規則による構文解析の結果、outputは並列になりうる最も近い要素であるregisterと並列関係になる。この結果を図3に示す。次に、図3の結果に対して次の規則

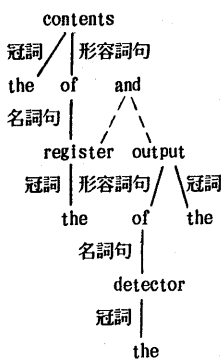


図3 構文解析結果

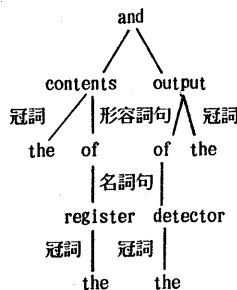


図4 並列関係解析結果

⑦同じ品詞は並列関係になる

⑧ [NP pre1 NP and NP pre2 NP: pre1=pre2] のパターンは [NP pre1 NP] と [NP pre2 NP] が並列関係になりやすい

⑨同じ意味素性を持つ単語同志は並列になりやすいを適用することによってcontentsとoutputが並列関係であることを決定でき、図4の結果を得ることができる。ここで、⑦の規則は並列関係の曖昧さを表現する規則であり、⑧⑨の規則は構文パターン、語彙の意味に依存した並列関係の優先度に関する規則である。

5. おわりに

大規模な辞書と解析文法を用いる実用システムでは、多品詞による曖昧さの解消、構文解析時間の短縮のために、局所的な情報だけでなく先読みによる大域的な入力文全体の情報を使って解析する方法、係り受け関係・並列関係の曖昧な現象に対しては、句構造解析で、この曖昧さを内在する唯一の解析結果だけを構築し、後の意味解析で曖昧さを解消する方法が有効であることを述べた。

[参考文献]

[1] 辻井：文解析方式，情報処理，Vol.27，No.8，pp.924-932 (Aug.1986)
 [2] 吉田，日高：機械翻訳のための構文解析手法，Vol.26，No.10，pp.1157-1164 (Oct.1985)
 [3] A S Hornby: Guide to patterns and Usage in English, Oxford University Press, 1976
 [4] Fernando Pereira: Logic for Natural Language Analysis, SRI International Technical Note 275, January 1983
 [5] 久米，辻井，長尾：MUプロジェクトにおける英日翻訳システムの構成と翻訳結果，情報処理学会自然言語処理研究会，59-6，1987
 [6] 岡島，山野，平井，新田：自然言語処理における中間語表現と多義性解消のしやすさとの関係，自然言語処理技術シンポジウム論文集，1983
 [7] 天野，平川：英日機械翻訳用パーサについて，情報処理学会自然言語処理研究会，32-1，1982
 [8] 中里，堤，平川，天野：日英機械翻訳システムにおける日本語解析について，情報処理学会第32回全国大会予稿 pp.1611-1612
 [9] 野上，杉浦，中里，天野：英日機械翻訳システムにおける構文解析について，情報処理学会第34回全国大会予稿 pp. 1259-1260