英文における非文法的要素の解析

7J-8

青山 千秋[+] 野上 宏康[++] 天野 真家[++] 堤 義直[+]
+ (株)東芝 情報通信システム技術研究所
++ (株)東芝 総合研究所

1. はじめに

機械翻訳において翻訳の対象を科学技術文献などのあいまい性が少ない文とすることで解析率が向上するといわれている。ところが、実文においては科学技術文献に特有の非文法的要素がしばしば現れるために解析率の向上をはばんでいる。そこで通常の文法規則(ATNルール)では扱うのが困難で、かつ、非文でないという非文法的要素について、その形態と出現頻度を英文の科学技術文献、マニュアルについて調べた。

2. 非文法的要素とは

非文法的要素とはATNパーサのもつ解析ルールでは解析が困難な要素のことである。たとえば、挿入やハイフンなどである。

ATNパーサに挿入やハイフンなどの解析も行うようにルールを与えると正しくない解析をしたり、処理時間の増大が生じ、通常の解析に支障が生じる。また、すべての単語の前後に発生しうる要素をATNルールに書くのは全ATNルールのノードに書かなくてはならず大変である。たとえすべてのノードについて適用されるようなルールが書けるように、ATNを拡張しても処理時間は短縮されず現実的ではない。

以上のような問題を生じさせる要素をまとめて非文法的要素と呼ぶことにした。

3. 調査の対象

科学技術関係の9文献から 165項3103文を無作為で抽出し、その中から非文法的要素を捜し出して、分類、カウントした。

項を単位として文を取り出したのは、1文ごとに取り出したのでは非文法的要素の発生がとらえにくいからである。

4. 文とは

機械翻訳での処理単位が文である。文章から文を切り出す基準は、一般の文であれば終止符で切り、タイトルや見出しなどは改行で切っている。文を切る処理ではあまり文を短く切らないようにしている。

5. 非文法的要素の種類

非文法的要素は以下の10種類に分類した。

1. タイトル
2. 見出し
3. ハイフン
4. スラッシュ
5. 挿入
6. 強調
7. 話法
8. 文中文
9. 不要区
10. 並列

5.1 タイトル

タイトルとは文章の始めにある名詞句のことである。これを拡張して名詞句で翻訳させるべき文をタイトルと呼ぶことにした。

通常の文には動詞が必須であるとした。そこで名詞だけの文などを部分訳としてでなく訳させるためにはタイトルという解釈が必要になった。

また、タイトルとあらかじめ識別しておくことは解析するために文のレイアウト情報を利用することになる。

例 D. ~~THE REVERSE EXAMPLE~~ //部がタイトル
File Sorting

5.2 見出し

見出しとは行頭にある数字や記号のことである。文中でも列挙するときには現れる。

見出しも、タイトルと同様にレイアウト情報を利用した解析率を向上させるためのパーサに渡す情報になる。

例 ~~1.~~ SIMPLIFYING CLAUSES //部が見出し
~~(1)~~ Objective judgements

5.3 ハイフン

ハイフンとは単語と単語を-で結んで1つの形容詞として利用することである。

例 English-Like language
be single-valued

The analysis of un-grammatical elements in English sentences
Chiaki Aoyama[+] Hiroyasu Nogami[++] Shinya Amano[++] Yoshinao Tsutsumi[+]
+ TOSHIBA Information and Communication System Laboratory
++ TOSHIBA Research and Development Laboratory

5.4 スラッシュ

スラッシュとは単語と単語を／で結び、／がorの働きをすることである。

例 Input/output
position/associativity

5.5 挿入

挿入とは文の中に構文としての役割を持たずに任意に挿入される文や単語のことである。

例 the precedence (an intger)

このほかにも～，～，～という形の挿入などがあるが、今回は括弧で示される挿入についてカウントした。

5.6 強調

強調とはいくつかの単語の前後を引用符ではさむことで1つの単語として扱い強調することである。一般に動詞は含まれず単語数は少ない。

例 relation "r3"
in the "truth value"

5.7 話法

話法とはいくつかの文の前後を引用符ではさみ一つの単語として扱うことである。強調とは異なり動詞が含まれる。

例 The query "list the ～" is ～
Thus "Is the ～?" should get ～

5.8 文中文

文中文とは文の中に現れる文のうち話法でないものである。挿入中に現れることが多い。

例 文 (文. 文. 文.) 文.

5.9 不要区

不要区とはいくつかの単語が翻訳の対象とならずまとめて1つの単語として扱われることである。挿入とは異なり文の一要素として不要区は解析される。

例 operators =, <, >;
is (EQUAL "NIL" X)

5.10並列

並列とはいくつかの単語や文が並列されたもので、特に意味処理をしないと正しく解析されないものを示す。

今回の調査ではカウントを行っていない。

6．カウント方法

各非文法的要素が発生した数を文についてそれぞれの要素ごとにカウントした。つまり、1文中にハイフンが3箇所あって、スラッシュが4箇所あるとハイフンは1、スラッシュも1とカウントした。

さらに、非文法的要素がまったくない文の数も数えた。これらの文はＡＴＮパーサのルール次第で確実に解析できる文である。つまり、総文数からこの文の数を引いたものがＡＴＮパーサで解析困難な文の数を示す。

7．非文法的要素の出現頻度

	出現回数（回）	頻度（%）
タイトル	196	6．3
見出し	283	9．1
ハイフン	173	5．5
スラッシュ	8	0．2
挿入	321	10．3
強調	226	7．2
話法	47	1．5
文中文	16	0．5
不要区	595	19．2
なし	1550	50．0

この表から非文法的要素を考慮しないパーサは科学技術文献に対して50%程度の解析率が得られるにすぎないことがわかる。

つまり、実文の解析率の向上には非文法的要素の解析ができるシステムの作成が求められるのである。とくにＯＣＲから読み込ませた文章をそのままプレエディットしないで翻訳させるには非文法的要素を避けて通れない。

非文法的要素のうち、タイトル、見出し、ハイフン、スラッシュ、括弧による挿入、強調、話法は識別が機械的にやりやすいので、扱うのが容易である。

これらを特別に処理する部分をパーサに付加することで手堅く解析率の向上が見込まれる。

文中文、不要区、コンマによる挿入、並列はプレエディットを要する場合が多く、機械的な処理は困難である。

プレエディターを使って制御記号を挿入することが現実的な解決方法である。

9．おわりに

本稿で述べた調査結果に基づき英日機械翻訳システムに非文法的要素の解析を行う処理部を組み込んだ。現在は実文を翻訳させて解析率の向上を計っているところである。