係り受け解析を用いた読みにくい文の抽出と対策について

小林悠子（情報科学コース）
伊藤俊一（情報教育講座）

１．序論

言語による情報伝達においては、読みにくい文を前もって特定し、それに対する解決策を考えることが必要となる。同じ文に対する複数の読み手の解釈が異なる場合、その文は曖昧で読みにくい文であると言える。本研究は、複数の読み手が考えた係り受け構造に違いが生じる文を大量に収集することにより、違いが生じる文に見られる共通の特徴を見つけ、それらに対する対策を考察することを目的とする。

複数の読み手としては、機械１種類と人間１名を用いる。現状では機械は人間と比べると書き手の意図や文脈を読み取る能力が低いため、係り受け構造の解析が難しい文であればあるほど、単純にそれに比例して誤って解析してしまう確率も高まると考えられる。そのことを利用して、機械と人間がそれぞれに解析した係り受け構造に違いが生じる文を特定することによって、読みにくい状況が生じている文を見つけ出すという手法を本研究は用いる。機械１種類としては、日本語係り受け解析器 Cabocha を選んだ。Cabocha は、日本語係り受け解析器としては最も精度が高いシステムであると言われている(正答率 89.29%)。

```
(a)東北から----------------D
    関東にかけての-D        |  1
          東日本一帯に---D  2
              甚大な-D|  3
               被害を-D  4
            もたらした。  5
(b)東北から---D
  関東にかけての-D  1
          東日本一帯に---D
              甚大な-D|
               被害を-D
            もたらした。
```

図１：Cabocha (a)と人間(b)の係り受け解析結果例

２．方法

大学一年生 20 人が書いたレポート初稿を分析の対象とする。初めに、すべての文に対し、Cabocha による解析を行う。その後、人間が目を通し、Cabocha の解析で間違っていると思われる部分を正しく解析し直す。Cabocha の解析が間違っていると人間が判断した 285 件の係り元の文節を用いて、係り元、Cabocha の示した係り先、人間が示した係り先それぞれの文節の、最初と最後の品詞を調べる。品詞は形態素解析器 Mecab を使って解析する。同時に、Cabocha の示した係り受けの距離と人間が示した係り受けの距離をそれぞれ測る。距離とは、係り先の文節が係り元の文節よりどれだけ後にあるかを表す。（図１参照。）

３．結果と考察

285 件のデータをまずは係り元の読点の有無、「Cabocha の係り受けの距離－人間の係り受けの距離」の値がプラスであるかマイナスあるかで分類する。以降では、プラスを「飛越」、マイナスを「割込」と呼ぶことにする。さらに、Cabocha と人間の係り先の句点・読点の有無で分類する。

続いて、係り元の文節の最後の品詞を接続助詞・動詞・助動詞・格助詞・係助詞・副助詞・助詞副詞化・副詞・名詞・形容詞・助詞連体化・並立助詞・接続詞に分類する。動詞と助動詞については、その係り先を確認し、述部に係るものと格に係るものとに分類する。それらの品詞を係り受けの形態によって述部 to 述部［接続助詞・動詞（to 述部）・助動詞（to 述部）］、格 to 述部［格助詞・係助詞・副助詞・助詞副詞化・副詞・名詞］、述部 to 格［動詞（to 格）・助動詞（to 格）・形容詞］、格 to 格［助詞連体化・並立助詞］、接続詞の５つに分ける。述部 to 述部は複文構造の誤りを、格 to 述部は格構造の誤りを、述部 to 格は関係節構造の誤りを、格 to 格は名詞句構造の誤りを、接続詞は接続構造の誤りを示すものと考えられる。

これらの分類に基づいて 285 件のデータを集計した結果、以下のことが明らかになった。

割込による係り受けの誤りが全体の 70%を占めた。誤りのほとんどは割込によるものであり、飛越による誤りは多くはないと言える。

また、誤りのほとんどは格構造（62%）・複文構造（22%）・名詞句構造（9%）を解析するときに生じるものであり、これらで全体の９割以上を占める。関係節構造･接続構造の誤りは少なかった。

割込による誤りは、Cabocha と人間のどちらの係り先にも読点がない場合に生じやすかった（係り元に読点がない内の 81%、ある内の 72%）。

割込による誤りの種類としては、係り元に読点がない場合もある場合も格構造の誤りが多かった（それぞれ 68%、47%）。読点がない場合には名詞句構造の誤り（13%）、読点がある場合には複文構造の誤り（47%）もそれぞれ多く認められた。

これらのことから、特に以下のような状況の時に割込による係り受けの誤りが生じやすいと言える。（件数は収集された誤りの件数を示す。）

[M__C__H。]

・読点のない格が文末の述部に係るとき（格構造の誤り）：27 件

「大垣共立銀行が４月１１日に「生体認証ＡＴＭ」を９月下旬に導入することを発表した。」

[M__C__H__]

・読点のない格が読点のない文中の述部に係るとき（格構造の誤り）：44 件

「これはグーグルが日本国内を越えて世界中で使用されているインターネットサービスを用いた検索システムの一つであるからだと思う。」

・読点のない格が読点のない文中の格に係るとき（名詞句構造の誤り）：17 件

「例えば iPad や iphone などのパソコンの代わりになるようなものが近年出てきたからである。」

[M、C__H。]

・読点のある格が文末の述部に係るとき（格構造の誤り）：13 件

「最後に、これを利用したことによって起こる不便さだ。」

・読点のある述部が文末の述部に係るとき（複文構造の誤り）：17 件

「動画や画像などをダウンロードした時、そのダウンロードしたファイルは動画や画像などではなく、個人情報を勝手に流出させるプログラムが仕組まれているウイルスだ。」

[M、C__H__]

・読点のある格が読点のない文中の述部に係るとき（格構造の誤り）：6 件

「そして「正しい情報」とは、今回のように、それを信ずるに足る確固たる証拠を確認してから発表することだろう。」

・読点のある述部が読点のない文中の述部に係るとき（複文構造の誤り）：7 件

「この事件は情報がどれだけ貴重であるかがわかり、コンピュータに関する知識がある社員なら、会社から情報を取得できてしまうということがわかる事件である。」

４．まとめ

最後に、係り受け構造の解析が困難なために文が読みにくくなる状況を回避するには書き手がどのような対策を取るべきかについて提案する。

対策１：

係り元が読点のある述部あるいは読点のある格であり、かつ、係り先に読点がないならば、述部として解析されてしまう可能性のある文節を読点なしで割り込ませるべきではない。

係り元が読点のない格［格助詞・係助詞・副助詞・助詞副詞化・副詞・名詞］であり、かつ、係り先に読点がないならば、述部として解析されてしまう可能性のある文節を読点なしで割り込ませるべきではない。

係り元が読点のない格［助詞連体化・並立助詞］であり、かつ、係り先に読点がないならば、格として解析されてしまう可能性のある文節を読点なしで割り込ませるべきではない。

対策２：

割込が原因で文が読みにくくなる状況を回避するためには、係り元と係り先の距離をできるだけ短く配置する必要がある。

対策３：

割込がどうしても必要な場合には、書き手の意図する係り先に読点を配置することで誤りを低減させる必要がある。

参考文献

横林博・菅沼明・谷口倫一郎（2004）係り受けの複雑さの指標に基づく文の書き換え候補の生成と推敲支援への応用 情報処理学会論文誌，45(5), 1451-1459.

須藤崇志・丸山広・中村太一（2008）文を分かりにくくする要因の分析と改善支援手法の提案 電子情報通信学会技術研究報告．SS，ソフトウェアサイエンス 108(64), 41-46.