# 品詞推定を利用した文切り処理

4H−3

宮平知博、神山淑朗

日本アイ・ビー・エム株式会社

## 1. はじめに

計算機ハードウェアの能力・容量の飛躍的な向上を背景として、自然言語処理において膨大な言語データ(コーパス)から言語モデルを構築して応用する、確率・統計的なアプローチが盛んになってきている。

機械翻訳システムにおける確率的品詞推定の応用を別稿[1]で報告したが、さらに、機械翻訳システムの最初の処理として重要な文切りに応用したので報告する。主動詞を探索することで、従来うまく文切りできなかった場合を、精度良く文切りできるようになった。

## 2. 文切り処理

### 2.1 文切り処理の問題点

機械翻訳の最初の処理として文切り処理と呼ばれる１文を切り出す処理がある。英文の場合には文末にはピリオドがあるので、文の切り出しは比較的容易な処理と考えられがちだが、"U.S."などようにピリオドを含む語（以下、ピリオド語と呼ぶ）が存在するため、ピリオドで文が切れるかどうかの判断は必ずしも容易ではない。

さらに、機械翻訳では翻訳すべき文が長くなると処理量が爆発的に増大するため、２文に文切りすべきなのに長い１文として繋げてしまうと、解析処理に時間がかかったり、解析失敗になったりする。したがって、２文に切るべき場合は確実に切ることが重要である。また、文切りの段階で間違えると、後の解析処理ではどのようにしても正しい結果は得られないので、１文とすべきものを２文に切ってしまうこともできるだけ避けたい。現状の機械翻訳技術では文の意味を掴んだ上で文切りの実行・修正を行なうことはできないので、最初にできるだけ正解に近い文切りを行なうことが重要となる。

### 2.2 従来の文切り処理

筆者らが開発するパターンベース機械翻訳システム[2, 3, 4]では、文を短くすることを基本としている。

ピリオド語を含まない単純な英文の場合には、単独のピリオドの次の単語が大文字で始まるかどうかで単純に文切りを行なっている。例えば、

I have a pen. You have a book.

の場合には、"pen."というようなピリオド語は存在しないので、pen の後のピリオドは単独のピリオドだと認識でき、その次の語 "You" が大文字で始まっているので、このピリオドで文切りを行なう。

一方、ピリオド語の中には、"Mr." のように、文末に現われることがない単語と、"U.S." のように、文末にも文の途中にも現われ、そこで文が切れるかどうか判断できない単語が存在する。そのため、"Mr." のような文末に現われることがない単語を辞書中にデータとして格納してあり、そのような単語の後では文を切らない。

さらに、"U.S." のような文が切れるかどうか判断できない単語の場合には、その単語を含む語、たとえば、"U.S. President" を辞書に登録し、登録されている単語の連続の場合には文を切らないようにしている。しかし、この場合には、常にそこでは文を切らないので、例えば、

Japanese Prime Minister Junichiro Koizumi went to U.S. President Bush welcomed him.

のような本来文を切るべき場合にも、文を繋げてしまう。また、ピリオド語を含む複合語をあらかじめすべて登録することは不可能であり、登録されていない場合には何も判断材料が無いため、そこで文を切ってしまう。

## 3. 品詞推定の文切り処理への応用

一般の文では、名詞句がそのまま１文になっている場合もあり、文の構造はさまざまである。しかし、名詞句がそのまま１文になるのは表題やリスト項目の場合が多く、普通のパラグラフ中の文では極めて

---

Sentence Segmentation Processing Using Part-of-Speech Estimation

Tomohiro Miyahira, Yoshiroh Kamiyama
IBM Japan, Ltd. (e-mail: miyahira@jp.ibm.com)
1623-14, Shimo-tsuruma, Yamato-shi, Kanagawa-ken

少ない。特に、ピリオド語で文切りするかどうかの判断に迷う場合には、その一方が名詞句の１文であることはほとんど考えられない。そのため、２文に分割すべき場合は、前後のそれぞれの文は主動詞を持っていると考えることができる。

したがって、２文に分割するかどうかの判断は、その位置の前方と後方に主動詞となるべきものがあるかで決定できることになる。一般に英単語は多品詞語が多いが、品詞推定処理によって各単語の品詞が推定できれば、これを利用することで主動詞を推定できる。

しかしながら、従属接続詞や関係代名詞によって複文が構成されることがあり、単に動詞を捜すだけではそれが主動詞であるかどうかは不明である。例えば、

... it would have been much worse if the U.S. Postal Service had been allowed to ...

のような場合に、"U.S. Postal Service" は、if 以下の従属節中にあるが、単純に前後に動詞を捜しただけでは、前方には have があり、後方には had があるために、間違って文切りしてしまうことになる。

このような複文を含めて、いろいろと試行錯誤した結果、ピリオド語から前後に動詞を探索する際に、以下のアルゴリズムに従って処理を行なうと精度良い判定ができることがわかった。

1. ピリオド語の直後が冠詞または代名詞の場合は、前後の動詞に関わらず文切りする。
2. ピリオド語の前方に従属接続詞か関係代名詞がある場合には、そこまでで動詞の探索を打ち切る。
3. ピリオド語の後方に従属接続詞か関係代名詞がある場合にはネストレベルを +1 し、動詞が出てきたらネストレベルを -1 する。ネストレベルが０の時の動詞を主動詞とみなす。
4. and の直後の動詞は主動詞とは見なさない。

上記１から４の処理を行ないながら、ピリオド語の前後にそれぞれ主動詞が見つかった場合に、文切りを行なう。

なお、現在進行、現在完了、受動態などの場合は、be や have を主動詞として扱っており、動詞の ing/ed 形は主動詞とはしていない。

## 4. 評価

ピリオド語の前後で文を切るべきかどうか従来方法では判断に迷う例文 179 例（２文に切るべき例：30、１文に繋げるべき例：149）をインターネットなどから集め、上記アルゴリズムを実装したプログラムを適用してみた。正しく文切りを判断できた文の数を表１に示す。

表１．正しく文切りできた文の数

| | 処理１ | 処理２ | 従来 |
|---|---|---|---|
| 切る (30 文) | 30 | 27 | 30 |
| 切らない (149 文) | 136 | 127 | 0 |
| 合計 (179 文) | 169 | 154 | 30 |

処理１は上記アルゴリズムによる結果を示し、処理２は、上記アルゴリズムによらず、単純にピリオド語の前後で動詞を探索した場合を示す。また、従来の方法では、切るべきかどうか判断できない場合には常に文切りを実行してしまうため、切るべきでない場合は常に間違ってしまい、正解の数はゼロとなっている。

評価した文の数がまだ充分ではないが、この結果から、推定した品詞を利用して主動詞のチェックを行なうことで、文切りの判断を高精度で行なうことができることがわかる。特に、上記の処理１のアルゴリズムを用いると、処理２では間違って文を繋げてしまう場合にも、うまく文切りの判定ができており、繋げて１文にすべき判断の精度も上がっている。

## 5. まとめ

推定した品詞を用いてピリオド語の前後で主動詞をチェックすることにより、従来は判断できなかった文切りの判定を高精度で行なうことができることがわかった。

今後は、さらに多量の文による評価と、それによるアルゴリズムの改良、および製品への組み込みが課題である。

［参考文献］

[1] 神山, "機械翻訳システムにおける確率的品詞推定とその応用", 情報処理学会第 63 回全国大会, 2001

[2] Takeda, K., "Pattern-Based Machine Translation", Proc. of 16th Coling, Vol. 2, pp. 1155-1158, 1996

[3] 渡辺, 武田, "パターンベース翻訳システム：PalmTree", 情報処理学会第 55 回全国大会, 1997

[4] 宮平, 渡辺, 田添, 神山, 武田, "インターネット機械翻訳の世界", 毎日コミュニケーションズ, 2000