4B−8

計算機で日本語を判読するための正書法

中牟田 純、水野 聡、島田 静雄、近藤 邦雄、佐藤 尚
埼玉大学工学部知識工学講座

1. 始めに

本研究の目的は、計算機で日本語の文を扱う際の標準的な規則を作ることである。現在計算機で扱う分野としては、論文の構成や検索、自動索引など論理的な文を扱うことが多いため、本研究では、論理的な文に対象を限定している。日本語処理の研究で問題になるのは、形態素レベルでは解釈の曖昧さが残ってしまうことである。従来の研究では、構文解析、意味解析などさらに高いレベルでの解析を行なうことによって、解釈の曖昧さを解決することに重点を置いている[1][3]。これに対し、本研究では、文そのものの問題に着目した。つまり、解釈の曖昧さがある文を解釈することを研究するのではなく、文に解釈の曖昧さが発生しないように書く方法、正書法の提案である。なお、本研究中で扱っている日本語は、漢字仮名交じり文のテキストデータである。

2. 正書法の考え方

自然言語で書かれた文には、次の曖昧さがある[2]。

構文的な曖昧さ
単語の意味の多様性による曖昧さ
文としての意味の曖昧さ
文脈的な曖昧さ

これらの曖昧さをすべて正しく解釈するためには文法だけではなく、文脈を解釈する能力、一般常識の知識などを持たなくてはならない。つまり構文解析、意味解析を行なわずに正確な文の解釈をおこなうのは難しい。そこで、文を書く時に、計算機に読みやすい文を意識的に書く。また、処理をするプログラムも、「計算機に読みやすい規則」に従って書かれている文を対象にして、処理を軽減する。以下では「計算機に読みやすい規則」を正書法と呼ぶ。例えば文を解析する際に使用する辞書を考えてみる。一般に辞書が大きくなると、解析中にある未知語が少なくなるが、逆に解析結果の選択肢が増えてしまい、最適な選択肢を求める作業が多くなる。したがって辞書に登録されている単語を数千語にしておき、辞書に登録されていない単語はユーザー定義の単語として、解析の時に文とともに解析するプログラムに渡すようにする。これにより、最適な選択肢を求める作業が少なくなり、また未知語の処理も不要になる。

3. 正書法の形態素解析への適応

例文（表1）に対して形態素解析を行ない、文の解釈の多様さを調べた。すべての形態素解析の選択肢の数を表2に示す。この選択肢には、表3のような通常現れない選択肢も含まれている。例文6の「話しかける日本語の文章が」の文頭「話しかける」は、形態素解析の段階では「話し（動詞）」＋「書ける（動詞駆けるなど）」か、「話し（名詞）」＋「かける（動詞）」なのか、「話しかける（動詞）」なのか、確定することは不可能である。人は、1つの文だけではなく、前後の文を読むことによってこれらの確定が可能であるが、計算機上で実現するには、構文解析や意味解析を行なわなくてはならない。現在、形態素解析の段階だけでこの問題を解決するために、頻度情報を用いて確定する手法が取られている。しかし、例外処理を頻度によって処理している以上、必ずある確率で、不適切な候補を選んでしまう。表3の中で、通常現れない組合せは以下のものが多い。

Correct Japanese writing methodology for computer processing
NAKAMUTA Jun, MIZUNO Satoshi, SHIMADA Shizuo, KONDO Kunio, SATO Hisashi
Saitama Univ.

他の品詞を人名、固有名詞と認識する。
複合名詞と認識する。
動詞や形容詞の連続と認識する。

日本語は、文節単位で自由に位置を変えることができる。また、名詞、動詞、形容詞は1つで（語幹＋語尾で）1つの文節を作ることができる。したがって解析する時にも、「名詞」＋「名詞」の組合せなどが多くなる。そこで、次の3つの正書法にそって、文を作る。

正書法

人名、固有名詞は定義して使う。（人名、固有名詞として使うのではなく、定義された名詞として使う）
複合名詞は使わない。
動詞、形容詞を連続して使わない。

この正書法に従って書いた文は、人名・固有名詞、名詞の連続、動詞および形容詞の連続は現れない。したがって、形態素解析の選択肢が少なくなる。

4. 正書法導入の効果

正書法を適応した後の形態素解析の結果を表2に示す。正書法1,2,3を適応するに従って、選択肢の数が少なくなっていく。また、候補の内容を調べたところ、期待する解釈は失われていなかった。よって正書法の導入によって、形態素解析の選択肢の確定の作業が容易になる。しかし、正書法を導入したにも関わらず、例文No.4のように数が変わらないものもある。これについては今後の課題である。

5. まとめ

本研究では、従来の方法とは異なり、正書法を導入して、計算機による日本語の文の処理を容易にする方法を提案した。本文は形態素解析に有効な3つの正書法を提案したが、正書法を導入し、日本語の書き方そのものについて研究を重ねれば、計算機にとって読みやすい日本語を書く技術が確立し、また日本語処理をする際にも形態素解析など低レベルの処理だけで正しい解釈が可能になると考える。

1　これを自然言語という
2　このためには
3　計算機にどのように話しかければ
4　誤りなく理解できるか
5　第一に
6　話しかける日本語の文章が
7　論理的に明確でなければならない

表1：例文

No.	正書法導入前	正書法1	正書法1,2,3
1	46	64	8
2	10	8	8
3	4	2	2
4	48	48	48
5	1	1	1
6	312	90	16
7	119	63	36

表2：選択肢の数

話し　かけ　る
名詞　動詞の語幹　活用語尾

話　し　かけ　る
動詞の語幹　活用語尾　動詞の語幹　活用語尾

話　しか　け　る
名詞　係助詞　動詞の語幹　活用語尾

話しかけ　る
動詞の語幹　活用語尾

表3：選択肢の例

参考文献

[1] 長尾 真 監修
「日本語情報処理」
社団法人 電子情報通信学会(1984)

[2] 長尾 真 著
「画像と言語の認識工学」
コロナ社(1989)

[3] 野口 正一 監修 牧野 武則 著
「自然言語処理」
オーム社(1991)

[4] 「TRIE構造辞書とその形態素分類体系の概要」
（財）新世代コンピュータ技術開発機構(1992)