日本語形態素解析処理における和語部分の解析性能

山内 佐敏* 南木一典** 米山正秀**
* (株)リコー 研究開発本部 ** 東洋大学工学部情報工学科

1 はじめに

日本語は元々から持っている大和言葉（以下、和語と称する）に中国大陸を初め海外から渡ってくる語を巧みに流用して発展してきている。基本的には和語は平仮名（用言には一文字漢字を使うことが多いが)、外来語として、大きくは漢語を漢字でその他の外来語はカタカナ表記をしており、英語の略記 (Acronym) や記号などはそのまま英文字を利用している。このように、表記により言葉の素性を表わしているのが日本語の書き言葉の一つの特長と言える。

表記の違いを利用して日本語の単語の区切りを見つけたり [1][2]、更に積極的に利用して形態素解析だけでなく係り受け解析まで成果を挙げた研究が既にある [3] が、この目的は辞書をほとんど無くし、軽くて高速な処理を目指したものである。

日本語の形態素解析アルゴリズムとしては最長一致法をスタートとして2文節最長一致法、文節数最小法、最尤評価法、コスト最小法など、もう基本的な考え方が出尽くした感がある。しかしながら、これらは日本語全体の振る舞いを一つの枠組みで対応しているために、個別の振る舞いに付いてはまだこれからであり、対応していかなければならない課題が数多くある。

そこで我々は逆に個別の振る舞いに注目したアルゴリズムを、純粋の和語に準ずる処理から、漢語、カタカナ語、英数字表記の順に実施することにした。幸いにも奈良先端科学技術大学院大学からシミュレータとして日本語形態素解析システム「茶筌」[4] が開放されているのでそれを流用させて頂き、和語部分の特長を分析し、付属語連語（格助詞相当語、接続相当語など）や付属語品詞連糸の導入による性能向上の程度を確認した。

2 益岡・田窪文法と「茶筌」

「茶筌」が採用している文法は基本的には益岡・田窪文法 [5] を下敷きにし、一部分拡張している。当然、我々もこれに従った。しかし、文法の著者も注意書きしている統語的機能に基づく分類にはなじまない「指示詞」が他の品詞と同列で入っているので、この部分に付いては1レベル下げて、名詞の下部の指示代名詞、連体詞の下部の指示連体詞、副詞の下部の指示副詞として分配した。

3 前処理としての文字識別処理

和語解析をする前に対象となる文字列の文字一つ一つの文字種を識別して大局的に和語系か、それ以外かを区分する。そこで和語系以外の文字列を“《”～“》”の記号によって以下のように囲む（識別記号を1 Byte系のコードで表示するため、入力文字列はすべて2 Byte系に変換してから処理を開始する)。

3.1 書式

《αnn| 文字コード列》

《：識別記号：和語以外の始まり（2Byte系）
》：識別記号：和語以外の始まり（2Byte系）
α：語系識別記号：(C,K,S)（1Byte系）
nn：文字数2桁表示（1Byte系）
文字コード列：入力文字列のうち2文字以上の連続する同種文字列（2Byte系）

3.2 文字種区分と識別規則

- 和語系：記号 (、。,. ヽヾ〃仝)、平仮名、カタカナ、漢字、英数字
  - カタカナ、漢字、英数字については、平仮名文字列中に一文字のもののみ
  - 識別記号、語系識別記号などなし(識別処理しない)
- 漢語系：記号（々)、漢字第一水準・第二水準
  - 二文字以上の連続する漢字列（漢字記号を含む)
  - 識別記号：C

- カタカナ表記：記号（ー・ヽゞ）、カタカナ
  - 二文字以上の連続するカタカナ列（長音記号などを含む）
  - 識別記号：K
- 英数字・記号類：記号（，．？！ー（）、その他の記号全て）、英字、算用数字、ギリシャ文字、ロシア文字
  - 二つ以上の記号、二文字以上の連続する英数字
  - 識別記号：S

4 「茶筌」辞書の分割と変更

各表記毎に個別のアルゴリズムを適用するために「茶筌」の辞書を次のように分類した。

4.1 辞書の分割

1. 平仮名表記漢語系辞書
   - 漢語系辞書、カタカナ語系辞書に属するものが見出し語として登録されていて、かつ、平仮名表記で見出し語が登録されているものの平仮名表記語のみを抽出し登録（例：(梅雨期、ばいうき) の内、"ばいうき"のみ登録)
2. 和語系辞書A
   - ひらがなで表記されている単語。(但し、同一見出しに漢字２文字以上 (全て漢字) だけが存在する場合は平仮名表記漢語系辞書Aに入ってしまう。)（例：おでこ、かつぶし、こんにちわ、など)
   - 平仮名と一文字の連続していない漢字やカタカナ、英数字・記号の表記の混ぜ書き単語 (ただし、漢字、カタカナ、英数字・記号は同時に存在しない)、当該漢字があっても平仮名書きする場合の単語（例：(恐れおおい、恐多い) の内、"恐れおおい"のみ登録)
3. 和語系辞書B
   - 二文字以上の連続した漢字を含んだ単語の内、平仮名が存在する場合（例：(恐れ多い、恐多い) の内、"恐多い"のみ登録)
   - 表記の混ぜ書き単語がある場合で、和語系辞書Aに属する和語があった場合と、和語系辞書Bの上記の条件に当てはまるものがあった場合には、二文字以上の連続した漢字 (全て漢字) でも登録（例：(引き受け、引受け、引受、ひきうけ) の内、"引受"を登録し、"引受け"も上記の条件に当てはまるのでやはり登録)
4. 漢語系辞書
   - 二文字以上の連続した漢字を含んだ単語の内、和語系辞書Bに属さないもの（例：(梅雨期、ばいうき) の内、"梅雨期"のみ登録)
5. カタカナ語系辞書
   - 二文字以上のカタカナが存在し、漢字や英数字・記号を含んでいない単語（例：アクティブだ、アルカリ、など)
6. 英数字・記号辞書
   - 二文字以上の英文字、数字、ギリシャ文字、ロシア文字、記号のみで書かれた単語
7. 混在辞書A
   - 単語の中に漢字、カタカナ語の両方が存在し、どちらかが連続している単語 (例: アドレス部、アナログ計算機、メッセージ通信、など)
8. 混在辞書B
   - 単語の中に漢字、英数字・記号の両方が存在し、どちらかが連続している単語 (例：ＣＧＳ単位、ＦＭ放送、など)
9. 混在辞書C
   - 単語の中にカタカナ語、英数字・記号の両方が存在し、どちらかが連続している単語（例：ＪＩＳマーク、¥マーク、など)
10. 混在辞書D
    - 単語の中に連続しない漢字、カタカナ、英数字・記号のどれか２つ以上が同時に存在する単語（例：(株)、ソ連、数ヵ月、鬼ヶ島、など)
    - 日本語には略語や、カタカナの"ヵ"や"ヶ"のように語と語の間に使う語などが多いので、そのための処置
11. 条件以外の辞書
    - １～１１までの条件に入らない単語

4.2 各辞書の使用区分

各辞書は以下の様に各解析時に分けて使用する。

- 和語解析時に用いる辞書
  - 平仮名表記漢語系辞書
  - 和語系辞書A
  - 混在辞書D
  - 付属語連語
  - 付属語連語
- 漢語解析時に用いる辞書
  - 漢語系辞書
  - 和語系辞書B
- カタカナ語解析時に用いる辞書
  - カタカナ語系辞書
- 英数字・記号類解析時に用いる辞書
  - 英数字・記号類辞書
- 統合処理に用いる辞書
  - 混在辞書A
  - 混在辞書B
  - 混在辞書C
  - 条件以外の辞書

4.3 付属語連語の追加

1. 格助詞相当語（74語）
   という, とかいう, とする, として, とはいえ, とともに, でもって, にあたって, に当たって, に当って, において, に於いて, に於て, にかけて, にかんして, に関して, に際して, にしたがって, に従って, 等
2. 接続助詞相当語（80語）
   や否や、やいなや、が早いか、がはやいか、そばから、とたんに、かとみると、かとみれば、上で、うえで、上に、うえに、あげくに、ままで、なりに、に従い、にしたがい、につれ、につれて、かぎりは、等
3. 助詞連語（4語）
   では、には、とは、のは

5 実験

シミュレータの本体は「茶筌」であるが、解析結果の正解の認定は正解文 (RWC[6] が作成したタグ付きコーパス) と自動的に照合できるようにツールを作成した。しかし、RWC の品詞タグの体系と「茶筌」の品詞体系とは若干異なるので対応表で対処したが、どうしても対応付けられない部分はエディタで修正できるようにしている。

5.1 正解文

本研究で使用した日本語文が収められている RWC テキストデータベースは、新情報処理開発機構 (RWCP : Real World Computing Partnership) より公開されたものであり、研究・評価用のテキストデータベースとして作成されたものである。

品詞体系作成の基本方針として、先に述べたように汎用性を考えているため、利用者が自分の研究目的に合わせて取捨選択あるいは変更して利用できることを念頭において設定されている。

尚、本研究では RWC テキストデータベースの中から、チューニング用とテスト用の文として、

1 . 通産省報告書形態素解析データ
通商白書平成 4 年版 (以下、通商白書と略記)

2 . 日本電子工業振興協会報告書形態素解析データ
自然言語処理の動向に関する調査報告書 (以下、電子協報告書と略記)
の、文をそれぞれ 150 文づつに分けて使用した。

5.2 実験結果

表にある A はチューニング用のテキスト、B がテスト用のテキスト、を示している。

1 . 付属語連語導入前の解析結果 (和語部分のみ)

テキスト		全形態素数	正解形態素数	解析率 (%)
通商	A	4226	4188	99.1
白書	B	3922	3875	98.8
電子協	A	1473	1442	98.0
報告書	B	1646	1607	97.6

2 . 付属語連語導入後の解析結果 (和語部分のみ)

テキスト		全形態素数	正解形態素数	解析率 (%)
通商	A	4226	4198	99.3
白書	B	3922	3881	99.0
電子協	A	1473	1447	98.2
報告書	B	1646	1617	98.2

5.3 実験結果と考察

通商白書の一文当たりの形態素数は電子協報告書のより 2 倍を大きく超えているが、正解率は高い。通商白書のほうが「茶筌」の文法と合っていると言える。しかし、付属語連語導入効果は電子協報告書の方が高い。

今回行った改良は副作用がほとんど無いと確信できる範囲の付属語連語に止めているので、格助詞「と」と並立助詞「と」との取り違え、判定詞連体形の「の」と接続助詞の「の」との取り違えなどがまだ残っている。データで示していないが、その他には接続詞、副詞、連体詞などの自立語の誤りが目立つ。

参考に和語以外を含めた文全体の性能を次表に示す。漢字２文字以上の単語やカタカナ語などを括弧で括っているのでその境界の単語の品詞同定は誤っている場合が多い。全形態素数も括弧で括った部分は１形態素として扱っているので少なくなっている（あくまでも参考データである）。

１．付属語連語導入後の解析結果 (文全体)

テキスト		全形態素数	正解形態素数	解析率 (%)
通商	A	6246	5988	95.9
白書	B	6070	5813	95.8
電子協	A	2211	2128	96.3
報告書	B	2389	2250	94.2

さらに、これもあくまでも参考であるが、お借りした「茶筌」の性能を次表に示す。使用マニュアルに従って操作し、最初に出てきた解を取り上げた結果である。

２．「茶筌」の解析結果 (チューニング用のみ)

テキスト		全形態素数	正解形態素数	解析率 (%)
通商白書	A	6332	6240	98.7
電子協報告書	A	2322	2285	98.4

6　おわりに

日本語形態素解析における和語部分の解析性能を確認することができた。特に付属語部分の特徴が把握できた。今後は自立語部分の品詞同定の向上に視点を移したい。更には漢字表記、カタカナ表記、英数字記号を取り込み、和語との境界を含めて解析し、日本語全体の形態素解析アルゴリズムの完成に向かうつもりである。

最後に日本語形態素解析システム「茶筌」を利用させて頂いた奈良先端科学技術大学院大学の松本裕治教授、ならびに、松本研究室の皆様方に感謝します。

参考文献

[1] 長尾真, 辻井潤一, 建部周二,(1978).「国語辞書の記憶と日本語文の自動分割」. 情報処理.Vol.19 No.6,pp.514-521.

[2] 大樫仁司, 小池和弘, 金枝上敦史.(1992).「字種切り法による形態素解析の一改良」. 情報処理学会 第43回 全国大会.

[3] 亀田雅之.(1993).「簡易日本語解析系 Q_JP」. 情報処理学会研究報告 NL-94-4.

[4] 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明.(1997).「日本語形態素解析システム「茶筌」Ver.1.5使用説明書」. 奈良先端大・松本研究室.

[5] 益岡隆志, 田窪行則.(1992).「基礎日本語文法—改訂版—」. くろしお出版.

[6] 井佐原均, 荻野紫穂, 桑畑和佳子, 徳永健伸, 橋本三奈子, 元吉文男.(1996).「RWCテキストデータベース報告書」. 技術研究組合 新情報処理開発機構.