TR-SLT-0021

bigram 統計量に基づく 中国語形態素解析器の試作

Chinese Morphological Analyzer
based on Bigram Statistics

山本 和英
Kazuhide Yamamoto

2002 年 9 月 30 日

Abstract

研究所内において自作した統計的手法による中国語形態素解析器について、使用法、解析手法などを述べる。また、この解析器において使用されている中国語辞書についても記述情報などの概要を紹介する。

(株) 国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒 619-0288 「けいはんな学研都市」光台二丁目 2 番地 2 TEL:0774-95-1301

Advanced Telecommunications Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288, Japan
Telephone:+81-774-95-1301
Fax:+81-774-95-1308

目次

第 1 章

はじめに

自然言語処理を進める上で、形態素解析器をはじめとする言語解析器は、コーパスなどの言語資源と同様に最も重要な道具である。近年では、この重要性は研究者間でほぼ認識されており、英語や日本語に対する形態素解析器と構文解析器はいずれも複数のものが作成、そして公開または市販され、我々研究者はその恩恵に預かっている。

ところが、中国語に関しては以上の状況は同じではない。我々の知る限り、日本国内はもちろん、中国においても誰もが手軽に使える中国語解析器が研究者の間で広範に知られている、という状況にはなく、まだ十分にツールが整備されているとは言えない。

この背景の一つは、中国語解析の困難性であると考える。中国語は英語のように概ね単語ごとに分かち書きされてはおらず、単語分割が必要である。また、文字種が単語分割のための大きな情報を持つ日本語とは異なり、ほぼ単一文字種 (漢字) である。さらに、複数品詞を持つ語が多いため品詞付与も容易ではない。例えば、中国語の介詞 (前置詞) のほとんどは動詞からの転成であるため日本語や英語にはほとんど存在しない内容語と機能語との間で品詞付与の曖昧性が生じる。また日本語における「ーする」 (動詞) 「ーい」 (形容詞) などの明確な文法標識を持たないため、内容語間の曖昧性も比較的多い。例えば中国語の「担心」は日本語の「心配 (名詞)/ 心配する (動詞)/ 心配だ (形容詞)」のすべてに相当する。

我々は現在、中日翻訳、並びに中国語換言処理の研究を行っている [Zha01, Zha02]。これらの処理は中国語が入力であるため、表層処理を行わない限り中国語解析器が必要である。このため ATR 音声言語コミュニケーション研究所 (以下 ATR) において使用可能な言語資源を用いて作成した。この際、中国語構文木コーパスとして Penn Chinese Treebank を使用して、n-gram 統計による形態素解析器を作成した。本報告は、この解析器で行なっている解析手法の概要を記録すると共に、ツールとして使用する際の方法や辞書等の拡張の方法を述べる。

第 2 章

解析手法

現在の形態素解析手法で最も一般的と考えられる統計的手法を採用した。すなわち、単語分割され、品詞情報が付与されたコーパス (以下、タグ付きコーパス) から n-gram 統計量を求め、これにより解析の曖昧性を解消する。

2.1 概要と全体の流れ

連続する 2 単語 (品詞付き) の連接に対して、出現確率から算出したあるコストを持ち、その値が入力文全体で最も高いものを最終的に選択する。すなわち、単語自身にはコストを持たず、単語分割の曖昧性、品詞付与の曖昧性を共に連接確率のみで解消する。

解析は逐次的 (left-to-right) で決定的 (deterministic) に行なう。すなわち、探索は入力の先頭から行ない、入力中のある位置までに確率の高かった上位 n 候補[1]の解析結果のみを残す絞り込みを、各入力位置で行なうことで、探索範囲を抑えている。

2.2 連接コストの算出

本解析器においては、コストは以下のようにして算出する。以下の説明で、連接する 2 単語を順に w_1, w_2、それらの品詞を順に c_1, c_2、ある 2 要素 i, j が連接していることを $< i, j >$ と記述する。また、 w_i には品詞の情報を含むため、同じ表記であっても異品詞の場合は区別される。

1. まず、 $< w_1, w_2 >$ の出現頻度 $conn(w_1, w_2)$ を調査する。もし $conn(w_1, w_2) > 0$ であれば以下の p_1 をコストとする。

$$p_1 = log(\frac{conn(w_1, w_2)}{\sum_i conn(w_1, i)}) \tag{2.1}$$

2. 次に、 $conn(c_1, c_2)$ を調査し、 $conn(c_1, c_2) > 0$ であれば、以下の p_2 をコストとする。

$$p_2 = log(\frac{conn(c_1, c_2) \sum_k conn(k, w_2)}{\sum_i conn(c_1, i) \sum_j conn(j, c_2)}) \tag{2.2}$$

ただし、この式において、 $\sum_k conn(k, w_2) = 0$ の場合、すなわち w_2 が未知語である場合、形式的に $\sum_k conn(k, w_2) = 0.1$ とする。

[1]標準版では $n = 10$ としているが、適宜変更できる。設定は `ctagger.pm` の冒頭で行なう。

3. 以上の統計量がどちらも得られない場合、一律に -1000 をコストとして与える。さらに、品詞 c_1 または c_2 が "X"(不明、未知) である場合、これにそれぞれ 1000 を減じる。つまり、両品詞とも "X" であれば -3000 をコストとする。

2.3 数字とアルファベットへの対応

探索は、基本的には辞書に記載の語を対象に行なう。しかし、数字 (354) やアルファベット列 (NTT) などをすべて辞書に記述することは不可能であるため、これらに対しては別個の処理を行なっている。すなわち、辞書検索時に、検索対象の文字列がすべて数字からなる場合は 1 語とみなして "CD"(数詞) の品詞を付与することで、辞書に記載されている他の語と同様の効果を与える。同様に、検索対象の語がすべてアルファベット (A-Z, a-z) から構成される場合も 1 語とみなして "NR"(固有名詞) の品詞を付与する。

2.4 未知語への対応

しかし、未知語の可能性を考慮した以下の処理を同時に行なっている。各文字位置までに対する探索時に、その文字 1 字を擬似的な単語とみなし、品詞 "X"(不明、未知) を付与して探索範囲に追加する。この処理によって、どのような入力であっても候補が全くなくなることはなく、解析結果を出力させることが可能となる。前述した通り、複数候補が存在しても連接確率で尤らしい候補を選択するため、辞書記載の語よりも本処理で追加した擬似的な語が優先されることはほぼありえないため、本処理による弊害もほとんどない。

2.5 解析例

以上の処理を、例文 "現在住在飯店" によって説明する。この様子を図 2.1に示す。なお、以下の説明では、候補の足切り値を 10 として説明する。

状態 0 以後の説明のため、便宜上初期状態を状態 0 と呼ぶ。

状態 1 まず第 1 文字目まで、すなわち "現" を解析する。初期状態なので、単に "現" を辞書引きする。この結果、辞書に存在し、品詞 "AD" を持つことがわかる。これとは別に、品詞 "X" を持つ架空の単語 "現" も追加し、 2 候補を解析結果とする。

状態 2 次に第 2 文字目まで、すなわち "現在" を解析する。この場合、 (1) 状態 0 + "現在" (2) 状態 1 + "在"、の両者を検討する必要がある。そこで、それぞれの単語を辞書引きすると、"現在" は 3 品詞が、 "在" も 3 品詞あることがわかる。その結果 (1) は 3 候補 (=1 状態× 3 品詞) が、 (2) は 8 候補 (=2 状態×架空品詞を含め 4 品詞) が生成され、両者の合計で 11 候補となる。これは 10 候補を越えるので、このうちコストの高い上位 10 候補のみを採用する。

状態 3 同様に、 (1) 状態 0 + "現在住" (2) 状態 1 + "在住" (3) 状態 2 + "住" について可能性を検討し、 (1)0 候補 (=1 状態× 0 品詞)(2)0 候補 (=2 状態× 0 品詞)(3)20 候補 (=10 状態×架空語を含め 2 品詞) となり、合計 20 候補となるが、この中からコストの高い上位 10 候補を採用する。

```
(状態 1)
 -8.67  現 /AD
-2000.00  現 /X

(状態 2)
 -5.96  現在 /NT
-10.18  現在 /AD
-11.10  現在 /NN
-12.43  現 /AD 在 /P
-15.74  現 /AD 在 /VV
-16.68  現 /AD 在 /AD
-2008.67  現 /AD 在 /X
-4000.00  現 /X 在 /AD
-4000.00  現 /X 在 /P
-4000.00  現 /X 在 /VV
-5000.00  現 /X 在 /X

(状態 3)
-13.22  現在 /NT 住 /VV
-16.06  現在 /AD 住 /VV
-18.57  現在 /NN 住 /VV
-20.38  現 /AD 在 /P 住 /VV
-22.56  現 /AD 在 /AD 住 /VV
-23.07  現 /AD 在 /VV 住 /VV
-2005.96  現在 /NT 住 /X
-2010.18  現在 /AD 住 /X
-2011.10  現在 /NN 住 /X
-2012.43  現 /AD 在 /P 住 /X
-2015.74  現 /AD 在 /VV 住 /X
-2016.68  現 /AD 在 /AD 住 /X
-4005.88  現 /X 在 /AD 住 /VV
-4007.33  現 /X 在 /VV 住 /VV
-4007.95  現 /X 在 /P 住 /VV
-4008.67  現 /AD 在 /X 住 /VV
-5008.67  現 /AD 在 /X 住 /X
-6000.00  現 /X 在 /P 住 /X
-6000.00  現 /X 在 /AD 住 /X
-6000.00  現 /X 在 /VV 住 /X

(状態 6: 最終状態)
-21.04  現在 /NT 住 /VV 在 /P 飯店 /NN
-23.88  現在 /AD 住 /VV 在 /P 飯店 /NN
-26.39  現在 /NN 住 /VV 在 /P 飯店 /NN
-28.20  現 /AD 在 /P 住 /VV 在 /P 飯店 /NN
-28.63  現在 /NT 住 /VV 在 /P 飯店 /NR
-29.16  現在 /NT 住 /VV 在 /VV 飯店 /NN
-30.38  現 /AD 在 /AD 住 /VV 在 /P 飯店 /NN
-30.89  現 /AD 在 /VV 住 /VV 在 /P 飯店 /NN
-31.47  現在 /AD 住 /VV 在 /P 飯店 /NR
-31.99  現在 /AD 住 /VV 在 /VV 飯店 /NN
...
```

図 2.1: 解析途中の候補とコスト

最終状態 このようにして解析途中結果の絞り込みを行ないながら逐次決定的に解析を行ない、最終的に最もコストの高い候補 “現在 /NT 住 /VV 在 /P 饭店 /NN” を出力する。

第 3 章

使用法

本形態素解析器のプログラムはすべて perl で作成した。本節では使用法を簡単に説明する。

3.1 概要

形態素解析を実行するために必要なファイルは以下の通りである。

- `ctagger.pm`: 形態素解析モジュール (解析器本体)
- `Chinese.dic`: 中国語単語辞書
- `freq`: n-gram 統計情報

このうち、辞書の語彙は追加や削除が容易に可能であり、また n-gram 統計情報は新たにコーパスから再計算させることも可能である。その際は、n-gram 統計情報を得るためのプログラムやコーパスが必要となるが、実行時には必要ない。

なお、以下では、現状のプログラムと言語情報 (単語辞書、n-gram) で行なう形態素解析器を便宜上「標準版 (の解析器)」と呼ぶ。

3.2 ツールとしての実行方法

解析器はモジュール (ライブラリ) として作成した。そのため Unix/Linux[1] 上でツールとして使用する場合には、これを呼び出すプログラムを perl で作成する必要がある。以下に、これを実現する最も簡単なプログラムを示す。

[1] perl で記述されているため、Windows 等の他の OS でも基本的な説明は同じであるが、以下ではすべて Linux 上での説明を行なう。

```
#! /usr/bin/perl

use lib "...";
use ctagger;

use FileHandle;
autoflush STDOUT 1;

while(<>){
  chomp;
  $input = $_;
  my($output) = &morphological_analysis($input, 0, (''));
  print "$output\n";
}
```

プログラム中で、 ··· とあるのは ctagger.pm が存在するパスである。プログラムは標準入力から入力、標準出力に出力する。すなわち、上記のプログラム名を仮に sample.pl とすると、以下のような形で実行させることができる。

```
% sample.pl < input_file > output_file
% echo "我做事总是四平八稳的。" | sample.pl > output_file
```

解析器は GB2312 文字コード体系の中国語を入力とし、1行ごとに形態素解析を行なう。原理上はどのような長い行であっても解析を行なうことができるが、探索空間が広くなるため、可能であれば文単位や句読点単位など、予め入力の切断が明らかな場合はその箇所で入力を分割してから解析を行なったほうが実行時間が早い。

入力の際、行末や行中に句読点は必ずしも必要としない。このため、書き言葉と話し言葉のどちらの入力にも対応する。ただし、標準版は書き言葉のコーパスである Penn Chinese Treebank から n-gram 統計量を取得しているため、解析精度は書き言葉を入力した場合のほうが性能が良いと期待される。また、同じ理由から、句読点があったほうが高性能であろう。

なお、入力は原則として中国語文のみを期待するが、中国語の記号 (=漢字以外の2バイト文字) や1バイト文字も入力として受け付けることが可能である。すなわちどのような書式の中国語ファイルも入力可能であるが、この場合は解析精度は劣化することが予想される。入力中の1バイト空白文字はすべて削除される。

出力は単語単位で出力され、単語間は1バイトの空白が挿入される。各単語には記号 "/" と共に品詞が付与される。ただし、1バイト記号の連続は「単語」とは見なされず、品詞は何ら付与されない。

品詞体系は Penn Chinese Treebank に従っている。ただし、記号などの品詞不明な単語には "X" が付与される点が異なる。

実際に旅行会話ファイル (/DB/LDB/JC/CTEXT/TAS12008.CTEXT) を入力させた場合の出力例を以下に示す。

```
pxs014[20:25]~/tagger)./sample.pl < /DB/LDB/JC/CTEXT/TAS12008.CTEXT
翻译/VV ： /X 您好/IJ ， /PU 这里/PN 是/VC 半岛酒店/NR 。/PU
申请/VV 人/NN ： /X 我/PN 叫/VV 铃木/NR 直子/NR ， /PU 明天/NT 我/PN 将/AD 住/VV 在/P
贵/VA 酒店/NN 。/PU
翻译/VV ： /X 铃木/NR 小姐/NN ， /PU 您好/IJ 。/PU 请/VV 讲/VV 。/PU
申请/VV 人/NN ： /X 估计/VV 到达/VV 酒店/NN 会/VV 比较/AD 迟/VA ， /PU 所以/AD 请/VV
为/P 我/PN 保留/VV 房间/NN 。/PU
翻译/VV ： /X 好的/IJ 。/PU 您/PN 估计/VV 大约/AD 几点/NT 可以/VV 办理住宿登记/VV ？
/PU
申请/VV 人/NN ： /X 估计/VV 可能/VV 得/DER 到/VV 晚上/NN 八/CD 点/M 左右/LC 。/PU
翻译/VV ： /X 好的/IJ 。/PU 不过/AD 从/P 为/P 您/PN 保留/VV 房间/NN 时/NN 起/LC ，
/PU 您/PN 就/AD 需要/VV 付款/VV 。/PU
申请/VV 人/NN ： /X ［/X 哎/IJ ］/X ， /PU 好/VA ， /PU 我/PN 知道/VV 了/AS 。/PU
翻译/VV ： /X 那么/IJ ， /PU 请/VV 告诉/VV 我/PN 您/PN 的/DEG 信用卡/NN 号码/NN 。/PU
申请/VV 人/NN ： /X 是/VC 万事/NN 达卡/NN 。/PU
号码/NN 是/VC 五/CD 二/CD 七/CD 九/CD 、/PU 三/CD 九/CD 二/CD 零/CD 、/PU 二/CD 四
/CD 六/CD 九/CD 、/PU 零/CD 零/CD 九/CD 八/CD 。/PU
翻译/VV ： /X 请/VV 让/VV 我/PN 确认/VV 一下/NN 。/PU 是/VC 万事/NN 达卡/NN ， /PU
号码/NN 是/VC 五/CD 二/CD 七/CD 九/CD 、/PU 三/CD 九/CD 二/CD 零/CD 、/PU 二/CD 四
/CD 六/CD 九/CD 、/PU 零/CD 零/CD 九/CD 八/CD 对/M 吗/SP ？ /PU
申请/VV 人/NN ： /X 对/VA 。/PU
那么/CS ， /PU 对不起/IJ ， /PU ［/X 那个/IJ ——/PU ］/X ， /PU 能/VV 告诉/VV 我/PN 你
的/JJ 名字/NN 吗/SP ？ /PU
翻译/VV ： /X 我/PN 叫/VV 乔/X ·/X 菲利普斯/NR 。/PU 如果/CS 您/PN 有/VE 什么/DT 问
题/NN 的话/SP ， /PU 请/VV 不/AD 要/VV 客气/VA ， /PU 欢迎/VV 来/VV 电话/NN 。/PU
申请/VV 人/NN ： /X 谢谢/IJ 你/PN 。/PU
翻译/VV ： /X 不用/AD 谢/VV 。/PU 那么/CS ， /PU 我们/PN 恭候/VV 您/PN 的/DEG 光临/NN
。/PU
pxs014[20:26]~/tagger)
```

第 4 章

中国語辞書

本節では、形態素解析に使用している中国語辞書について記述する。

4.1 概要

この中国語辞書は、2000年度末においてATRで使用可能な6種類の中国語辞書を統合したものに、手作業で若干の語彙の追加と修正を行なったものである。以下では最初に統合を行なった辞書を辞書A～辞書Fと呼ぶ。

辞書A Penn Chinese Treebankに含まれるすべての語を直接抽出し、これを辞書Aとした。ただし、品詞が"CD/OD/NT/FW"である語、すなわち、数詞やアルファベット表記語はすべて除外した。その結果、規模は10446語となった。

辞書B 日中TDMTにおいて使用されていた語を辞書Bとした。まず、形態素体系の変換を自動的に行なった。ここで、中国語TDMT体系では「形容詞」となっているものはPenn体系ではVAとJJの両者に該当するが、この自動的な識別は困難であるので、一律にVAとして追加した。他の品詞については、ほぼ一対一対応がとれている。次に、辞書Aに含まれている語は削除した。この結果、辞書Bとして新たに7567語を追加した。

辞書C 次に、中国語のピンイン漢字変換ツールcWnnに含まれている全単語を抽出し、これを辞書Cとした。まず、辞書Bと同様に品詞体系が異なるのでまず品詞変換を行なった。この際、「動名詞」という品詞がcWnnにはあるがこれに相当する品詞はPenn体系にはないので、一律にNNとして追加した。次に、辞書Aと辞書Bのどちらかに含まれる語は削除し、重複を取り除いた。この結果、辞書Cとして26100語を追加した。

辞書D／E／F これに、中英翻訳辞書cedictを辞書Dとして、英中翻訳辞書san-ecを辞書Eとして、香港大学の英中翻訳辞書(漢字入力用対応表)を辞書Fとして追加した。これら3辞書には元々品詞が記述されていないため、全単語について品詞名を"X"として追加した。これら3辞書についても、辞書A／B／Cと重複を取り除いた結果、追加した語彙は3辞書合計で34340語となった。

なお、辞書D／E／Fについては、品詞が付与されていない、文字変換や訳語が多いため狭義の単語辞書ではない、などの理由で、本報告で述べる形態素解析の辞書には含まれていない。この結果、形態素解析を行なうための辞書中には品詞"X"は全く含まれていない。

4.2 書式

辞書は GB2312 文字コード体系で記述した。1行に1語を表記し、同表記であっても品詞が異なる場合は別の行に記述する。各行の書式は以下の通りである。

ID	表記	(読み)	品詞	出典	(備考)

ID ID はそれぞれの表記に対して一意に付与した。すなわち、品詞が異なっていても表記が同じ場合は同一の ID とした。現在の ID は2文字または4文字の1バイト英小文字で表記する。概ね、頻出語や重要語は2文字 (aa ～ zz)、その他は4文字 (aaaa-zzzz) で表している。

表記 中国語の単語表記を記述する。中国語は日本語などと異なり、活用を全くしないため、「標準形」などの概念が必要ない。

読み 発音をピンイン (拼音、英字による中国語の発音表記) で表記する。ピンインは四声 (1-4) と共に表記し、軽音の場合は表記しない。前述した6種類の辞書はピンインの表記がそれぞれ若干異なっているので、できる限りその表記を統一した。どの辞書からも読みの情報を得られない場合は空欄とした。

品詞 品詞を Penn Chinese Treebank の品詞体系で記述した。付録に品詞一覧を示す。品詞が不明な場合でも空欄とはせず、"X" と表記した。また、前述した理由から、辞書D／E／Fから追加した単語はすべて品詞 "X" となっている。

出典 これらの表記、読み、品詞の情報をどの辞書から得たかを記述する。原則として、英大文字1文字で表記する。ただし、品詞と読みが別の辞書から得られる場合がある。その場合は、コンマによって併記した。例えば、"A" とあれば辞書Aからすべての情報が得られたことを意味し、"B,C2" とあれば辞書Bから表記と品詞を獲得し、第2項目 (＝読み) は辞書Cから抽出したことを意味する。この欄も空白とはせず、必ず記述した。なお、辞書からの抽出ではなく手作業で追加した語は便宜的に "Z" と記述した。

備考 その他の参考情報を記す。プログラム等では無視される。現在は手作業による追加、修正を行なった行に対し、"lyao:1.3.14" などとメモしている。

第 5 章

問題点

現在分かっている解析器の欠点と問題点を列挙する。

話し言葉への対応 コーパスが Penn Chinese Treebank であり、これは新聞記事等の書き言葉を対象として統計量の学習を行なっている。また、使用される語彙にも大きな差異があり、例えば感動詞、間投詞などはほとんど新聞記事では見受けられないため、これらの語を含む表現に対して形態素解析を行なう際には正しく解析できない可能性が高まる。また、話し言葉特有の倒置表現などの入力表現に対して、うまく対応できない可能性がある。

統計量 前項と同じ理由で、書き言葉としても十分な精度が得られない可能性がある。中国語タグ付きコーパスとして見た場合、Penn Chinese Treebank は小規模と見なされ、容易に入手できる同種のコーパスである Sinica Corpus や人民日報コーパスと比較しても 10 倍以上の規模の違いがある。

第 6 章

まとめ

本報告では、自作した中国語形態素解析器の概要を説明した。まだ問題も多くあるが、形態素解析自体を研究対象にしない場合は、手軽に使用できる。また perl ですべて記述されており、辞書とプログラムが分離されているため、今後品詞体系の変更や新コーパスの入手があった場合に変更も容易と考える。

参考文献

[Zha01] ZHANG, Y., YAMAMOTO, K., and SAKAMOTO, M.: Paraphrasing Utterances by Reordering Words Using Semi-Automatically Acquired Patterns, In *Proc. of NLPRS2001*, pp. 195–202 (2001).

[Zha02] ZHANG, Y. and YAMAMOTO, K.: Paraphrasing of Chinese Utterances, In *Proc. of COLING2002* (2002).

付録 A

Penn Chinese Treebank の品詞タグセット

AD	adverbs
AS	aspect marker
BA	把 in ba-const
CC	coordinating conj
CD	cardinal numbers
CS	subordinating conj
DEC	的 for relative-clause etc.
DEG	associative 的
DER	得 in V-de const. and V-de-R
DEV	地 as the head of DVP
DT	determiner
ETC	tags for 等 and 等等 in coordination phrases
FW	foreign words
IJ	interjection
JJ	noun-modifier other than nouns
LB	被 in long bei-construction
LC	localizer
M	measure word (including classifiers)
MSP	some particles
NN	common nouns
NR	proper nouns
NT	temporal nouns
OD	ordinal numbers
ON	onomatopoeia
P	prepositions (excluding 把 and 被)
PN	pronouns
PU	punctuation
SB	被 in chort bei-construction
SP	sentence-final particle
VA	predicative adjective
VC	copula 是
VE	有 as the main verb

VV other verbs