形態素と文字の情報を用いた中国語形態素解析

侯 海霞（東京農工大学大学院 情報工学専攻）
古宮 嘉那子（東京農工大学　工学研究院　先端情報科学部門）
柴原 一友（テンソル・コンサルティング株式会社，東京農工大学）
藤本 浩司（テンソル・コンサルティング株式会社，東京農工大学）
小谷 善行（東京農工大学　工学研究院　先端情報科学部門）

Chinese Morphological Analysis Using Morphemes and Characters

Haixia Hou (Graduate School of Engineering, Tokyo University of Agriculture and Technology)
Kanako Komiya (Institution of Engineering, Tokyo University of Agriculture and Technology)
Kazutomo Shibahara (Tensor Consulting Co.Ltd., Tokyo University of Agriculture and Technology)
Koji Fujimoto (Tensor Consulting Co.Ltd., Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Institution of Engineering, Tokyo University of Agriculture and Technology)

１．はじめに

形態素解析は自然言語処理の重要な基本技術の一つである。中国語や日本語などには分かち書きがないため、形態素解析は特に重要である。日本語の形態素解析は多く研究されているが、中国語の形態素解析の研究は比較的少ない。形態素解析において重要な課題となっているのは、未知語（辞書に登録されていない、あるいは訓練コーパスに出現されていない、形態素となり得る文字列）をどのように扱うかということである。

ここれまでの研究の多くは、内元(2001)や竹内(1997)などのような形態素の情報だけを使用していた。しかし、形態素の情報だけでは、未知語の扱いが困難であるという問題があった。これに対して、我々は中国語のほとんどは漢字によって書かれるという事実を踏まえ、文字情報を素性として用いることができると考えた。しかし、小田(1999)では文字の情報だけでは、既知語に対する精度が低いと証明した。そのため、我々は、形態素の情報と文字の情報を素性として使用する形態素解析の手法を形態素解析の手法を提案する。

本研究に最も近い研究は、中川哲治(2004)の形態素と文字の情報を併用した中国語の分割であるが、これは文を形態素に分解する部分だけを行い、形態素解析は行っていない。また、我々の手法では、形態素の情報と文字の情報に対して柔軟な素性設計の可能な最大エントロピー(ME)モデルを使用し、品詞遷移や文字位置などの情報を用いてコーパスから未知語の性質を学習した。人民日報のタグ付きコーパスを使用して実験を行った結果、素性として形態素と文字の両方を使用した手法は形態素だけ使用した手法より高い解析精度が得られた。

２　素性の学習

素性の学習には、Berger (1996)の ME モデルを用いる。このとき、ある形態素 x_i がある品詞 y_i である確率 $p(y_i \mid x_i)$ を計算する。ME モデルにおける $p(y_i \mid x_i)$ の計算は

素性に依存する。素性は形態素解析に役に立つ情報によって定義され、素性関数の引数として利用される。素性関数は下記のように定義する。

$$f_{ijk}(x,y)=\begin{cases}1: & x\text{は情報}\ g_{ij}\text{持つ}\quad\text{かつ}\quad y=f_k\\ 0: & \text{その他}\end{cases}$$ 式２－１

x：着目している形態素

y：着目している形態素の品詞

素性(g_{ijk})：x は情報 g_{ij} が持つ　かつ　$y=f_k$

素性集合と訓練データが与えられた時、エントロピーを最大化にすることによって、モデルが生成される。全ての素性 g_{ijk} ごとにパラメータ λ_{ijk} を用い、Berger（1996）のような条件付き確率として表される。

$$p^*(y\mid x)=\frac{1}{Z(x)}e^{\sum_{ijk}\lambda_{ijk}f_{ijk}(x,y)}$$ 式２－２

$$Z(x)=\sum_{y}e^{\sum_{ijk}\lambda_{ijk}f_{ijk}(x,y)}$$ 式２－３

パラメータ λ_{ijk} を推定する際には、下記式２－４のように、訓練コーパスにおける全ての素性 g_{ijk} に対し、ME モデルから計算される（x,y）の確率が訓練コーパスにおいての（x,y）の出現確率と等しくなるようにする。p は訓練コーパスによって計算される確率である。

$$\sum_{x,y}\overline{p}(x,y)f_{ijk}(x,y)=\sum_{x,y}\overline{p}(x)p^*(y\mid x)f_{ijk}(x,y)$$ 式２－４

下記の式２－５で収束するまでパラメータ λ_{ijk} を更新しながら学習する。（括弧の部分が０になるのが一番よい）

$$\lambda_{ijk}^{(n+1)}=\lambda_{ijk}^{(n)}+c\left[\sum_{x,y}\overline{p}(x,y)f_{ijk}(x,y)-\sum_{x,y}\overline{p}(x)p^*(y\mid x)f_{ijk}(x,y)\right]$$ 式２－５

n：推定回数

c：学習率

３．形態素の素性と文字の素性

本論文では、形態素の情報である「形態素の連接品詞」、「形態素の表層」、文字情報である「文字が形態素においての位置」の３種類の素性を定義する。これ以降、形態素の連接品詞を「前品:バイグラム」、形態素の表層を「単品:表層」、文字位置を「文位:文字が形態素においての位置|文字」と表記する。最大エントロピー法では、ラベル（形態素解析器なので本研究では品詞に相当する）ごとに素性を作成する必要があるため、

最も単純な素性でも 46 種類の素性となる。以下に、素性を列挙する。

A.「前品:バイグラム」

本論文が利用しているコーパスにおいては、品詞は 46 種類あるため、合計で 2162（47 種類×46 種類）個の素性を利用する。

A-1　品詞バイグラム（46 種類×46 種類）

下記に例をしめす。

例：「前品：名詞」

「着目している形態素が名詞で、かつ、その直前の形態素の品詞が名詞である」ときに 1 になる素性。

A-2　文頭-品詞バイグラム（46 種類）

例：「前品：文頭」

「着目している形態素が名詞で、かつ、その直前の形態素の品詞が文頭である」ときに 1 になる素性。

B.「単品:表層」

辞書においては、形態素は 57760 個あるため、合計で 2656960（57760 種類×46 種類）個の素性を利用する。下記に例をしめす。

例：「単品:我」

「着目している形態素が名詞で、なおかつ「我」である」ときに 1 になる素性。

形態素の連接品詞と表層を素性としてのパラメータの格納の仕方を表 3-1 に示す。品詞は本論文の訓練コーパスにおいては４６種類があるが、ここでは、品詞を n:名詞、v:動詞、a:形容詞、d:副詞、w:記号の５種類を例にする。

表 3－1　形態素素性に対してのパラメータ

情報番号	情報名	f_1:n	f_2:v	f_3:a	f_4:d	f_5:w
g_{10}	前品:文頭	λ_{101}	λ_{102}	λ_{103}	λ_{104}	λ_{105}
g_{11}	前品:v	λ_{111}	λ_{112}	λ_{113}	λ_{114}	λ_{115}
g_{12}	前品:n	λ_{121}	λ_{122}	λ_{123}	λ_{124}	λ_{125}
…	…	…	…	…	…	…
g_{21}	単品:我	λ_{211}	λ_{212}	λ_{213}	λ_{214}	λ_{215}
g_{22}	単品:是	λ_{221}	λ_{222}	λ_{223}	λ_{224}	λ_{225}
g_{23}	単品:学生	λ_{231}	λ_{232}	λ_{233}	λ_{234}	λ_{235}
…	…	…	…	…	…	…

「前品: 文頭」:現在着目している文字列が文頭である
「前品: v」:直前前の品詞が v である
「前品:n」:直前前の品詞が n である
「単品:我」: 着目している文字列が「我」である
「単品:是」:着目している文字列が「是」である
「単品:学生」:着目している文字列が「学生」である
「品詞」の種類: n:名詞, v:動詞, a:形容詞, d:副詞, w:記号 という 5 種類を例にしている
「品詞」の種類:全部で 46 種類がある
λ_{101}: 情報 g_{10} を持っている文字列が n という品詞になる重みである
λ_{111}: 情報 g_{11} を持っている文字列が n という品詞になる重みである
λ_{ijk}: 情報 g_{ij} を持っている文字列が f_k 品詞になる重みである
λ_{ijk}: 前品に対して 2162 個になる。単品に対して 2656960 個になる

C.「文位:文字が形態素においての位置|文字」

文字の形態素における位置の表記に[7]の表記方法を用いる。この表記を表 3－2 に示す。

表 3－2　文字が形態素においての位置

表記	意味
S	形態素が 1 文字である文字
B	形態素の先頭にある文字
I	形態素の中間にある文字
E	形態素の末尾にある文字

辞書において、文字は 12977 個あるため、合計で 596942 個の素性を利用する。下記に例をしめす。

例 1:「文位: S｜我」

「着目している形態素が名詞で、なおかつ「我」という 1 つの文字である」ときに 1 になる素性。

例 2:「文位: B｜我」

「着目している形態素が名詞で、なおかつ「我」という文字であり、形態素の先頭になる」ときに 1 になる素性。

例 3:「文位: I｜我」

「着目している形態素が名詞で、なおかつ「我」という文字であり、形態素の中間にある（先頭でも末尾でもない)」ときに 1 になる素性。

例 4:「文位: E｜我」

「着目している形態素が名詞で、なおかつ「我」という文字であり、形態素の末尾になる」ときに 1 になる素性。

文字情報である「文位:文字が形態素においての位置|文字」を素性としてのパラメータの格納の仕方を表３-３に示す。

表３－３　文字素性に対してのパラメータ

情報番号	情報名	f_1:n	f_2:v	f_3:a	f_4:d	f_5:w
g_{31}	文位:S｜我	λ_{311}	λ_{312}	λ_{313}	λ_{314}	λ_{315}
g_{32}	文位:S｜是	λ_{321}	λ_{322}	λ_{323}	λ_{324}	λ_{325}
g_{33}	文位:B｜学	λ_{331}	λ_{332}	λ_{333}	λ_{334}	λ_{335}
g_{34}	文位:E｜生	λ_{341}	λ_{342}	λ_{343}	λ_{344}	λ_{345}
g_{35}	文位:I｜一	λ_{351}	λ_{352}	λ_{353}	λ_{354}	λ_{355}
…	…	…	…	…	…	…

「文位:S｜我」：着目している文字列が「我」という１つの文字である
「文位: B｜学」：着目している文字列の先頭が「学」という文字である
「文位: E｜生」：着目している文字列の末尾が「生」という文字である
「文位:I｜一」：着目している文字列の中間に「一」という文字がある
「品詞」の種類: n:名詞, v:動詞, a:形容詞, d:副詞, w:記号 という５種類を例にしている
「品詞」の種類:全部で 46 種類がある
λ_{311}: 情報 g_{31} を持っている文字列が n という品詞になる重みである
λ_{312}: 情報 g_{31} を持っている文字列が v という品詞になる重みである
λ_{ijk}: 情報 g_{ij} を持っている文字列が f_k 品詞になる重みである
λ_{ijk}:全部で 596942 個になる

４．形態素と文字の情報を用いた中国語形態素解析システム

すべての文字列を形態素としてシステムを実行するには時間かかるため、本研究では態素解析候補にする文字列の長さは、あらかじめ 5 文字までとし、辞書に入ってない５文字以上の文字列を考慮しないようにした。これは、中国語において、５文字以上の未知語の形態素が非常に少ないためである。本論文の訓練コーパスにおいて、６文字以上の形態素は僅か 0.62%であった（**表**４－１）。

また、未知語が入っている文章を解析できるようにするため、辞書にある文字列だけではなく、長さが５文字以内の文字列は全て形態素候補とした。この際、形態素候補には全種類の品詞を付けて展開する。ここで、形態素候補と品詞のセットをノードと呼ぶ。

表４-１　訓練コーパスにおいての形態素の長さ

形態素の長さ	出現確率
1 文字	47.34%
2 文字	44.73%
3 文字	4.46%
4 文字	2.12%
5 文字	0.73%
6 文字以上(6 文字含む)	0.62%
形態素の総数：1083411	

また、高速化のため、句読点などのマークは、周り見ずにそのマークを直接一つの形態素として扱った。

作成した形態素候補をリンクで繋いで、ラティスを作成する。この様子を図４－１に示す。ここでは、ノード全てを図に表示するのが困難であるため、一部のみ表示している。

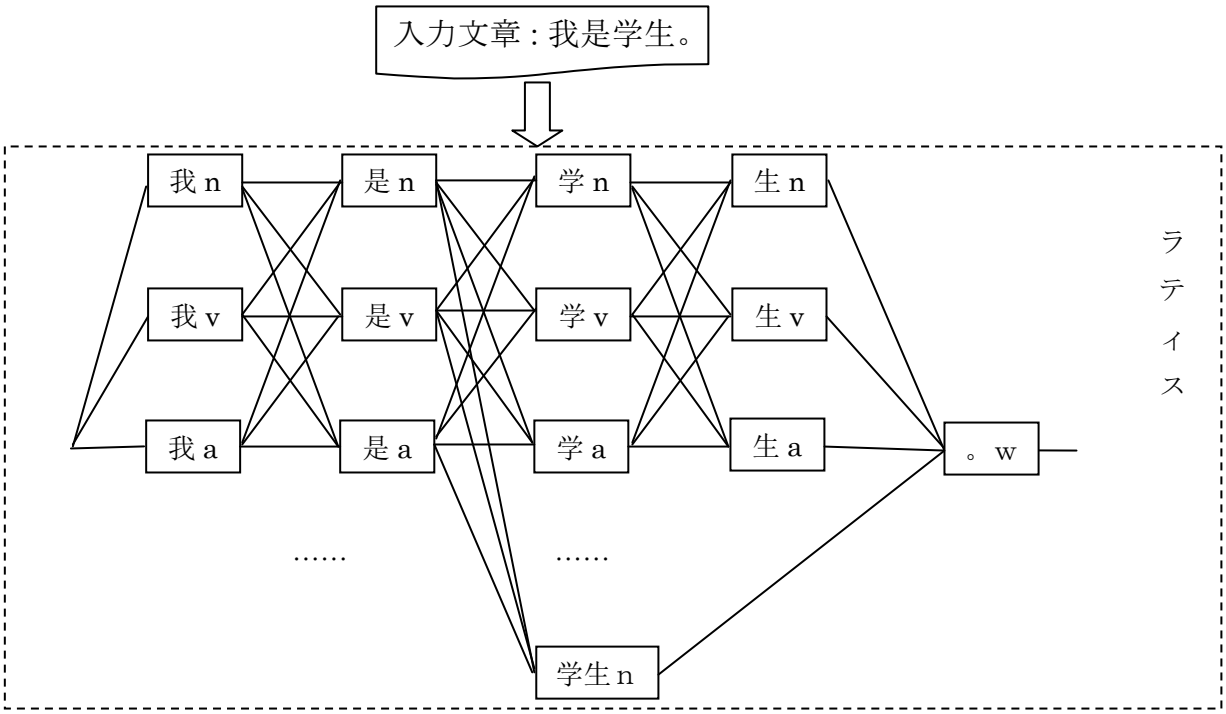

図４－１　ノードセットとラティス

ラティスにあるすべてのルートの確率を ME モデルで計算し、そのうち、確率が最大になっているルートを結果として出力する。

５．実験

形態素と文字の情報を用いた中国語形態素解析を実装し、人民日報、富士通と北京大学によるタグ付きコーパス(人民日報タグ付きコーパス)を使って、形態素と文字の情報を用いた手法と、形態素だけ使用した手法の比較を行った。

人民日報タグ付きコーパスからランダムで取り出した1541文をテストコーパスとして、残りは訓練コーパスとして利用した。人民日報タグ付きコーパスの分類を表５－１にて示す。

表５－１　人民日報タグ付きコーパスの分類

総数	合計	訓練 コーパス	テスト コーパス
文	46, 251	44, 710	1, 541
形態素	1, 083, 411	1, 048, 121	35, 290

解析精度の評価を行うために、Closed テストと Open テストを行った。このテストデータとして closed テストデータと open テストデータをそれぞれ作成する。closed テストデータは訓練コーパスからランダムで取りし、open テストデータは訓練コーパスから分けられたテストコーパスからランダムで取り出す。

上述した形態素と文字の情報を用いた中国語形態素解析の手法で closed テストと open テストをそれぞれ 3 回行ったため、テストデータは毎回 300 文をランダムに 3 回取り出して作成した。表５－２の形態素の数は 3 回のデータの平均値である。

表５－２　Closed test と Open test によるテストデータ

数	Closed テストデータ	Open テストデータ
文	300	300
形態素	7,025	6,870

６．評価

上述した実験の結果を表６-1 にて示す。確率は 3 回実験において得られた確率の平均値である。

表６－１　実験結果

素性	Closed テスト		Open テスト	
	適合率	再現率	適合率	再現率
形態素だけ	96.1%	95.4%	83.71%	89.2%
形態素と文字	**96.1%**	**95.9%**	**90.31%**	**93.2%**
文字情報効果	**0**	**0.5**	**6.6**	**4**

表６－１から、未知語のない closed データでの実験では、文字情報の効果はほとんどないことが分かる。また、同じ表から、未知語のある open データでの実験では、文字情報の効果は明らかである。適合率において 6.6 ポイント、再現率において 4 ポイントを上がっている。以下に具体的な例をとって考察する。

例１、未知語である「细致」

形容詞である「细致」は辞書と訓練コーパスにないが、形容詞である「细心」「细微」「细嫩」「别致」「雅致」はあったため、「文位:B|细」「文位:E|致」との２つの情報を用いて、コーパスから「细致」を学習することができた。

例２、未知語である「圆润」

形容詞である「圆润」は辞書と訓練コーパスにないが、形容詞である「圆满」「圆浑」「滋润」「红润」「湿润」はあったため、「文位:B|圆」「文位:E|润」との２つの情報を用いて、コーパスから「圆润」を学習することができた。

７．おわりに

本論文は、形態素と文字の情報を用いた中国語形態素解析の手法を提案した。実験により、全て漢字により書かれている中国語は形態素解析を行う場合には、形態素の情報も文字の情報も有効であることが分かった。適合率が 6.6 ポイント、再現率が 4 ポイントが上がった。未知語の処理に文字の情報が役に立っていることが分かった。

謝辞

本論文の作成にあたり、人民日報のコーパスを利用させていただきました。人民日報のコーパスを作成した各社に感謝いたします。

文献

Berger,A.L.,Pietra,S.A.D.,and Pietra,V.J.D.,A(1996)「Maximum Entropy Approach to Natural Language Processing」Computational Linguistics,22(1),pp.39-71.

内元清貴、関根聡、井佐原均(2001)「最大エントロピーモデルに基づく形態素解析―未知語の問題の解決策―」,自然言語処理, Vol.8:NO.1,pp.127-142,

竹内孔一、松本裕治(1997)「隠れマルコフモデルによる日本語形態素解析のパラメータ推定」情報処理学会論文誌,Vol.38:NO.3,pp.500-509.

小田裕樹、森信介、北研二(1999)「文字クラスモデルに基づく日本語単語分割」情報処理学会研究報告,99-NL-130,pp.1-8.

中川哲治、松本裕治(2004)「単語レベルと文字レベルの情報を用いた中国語・日本語単語分割」 自然言語処理研究会報告 2004(73),pp.197-204.