自然言語処理 134－17
（1999．11．26）

誤り駆動モデルに基づく中国語未登録語の認識

韓東力　古郡廷治
電気通信大学　情報工学科

本稿は、中国語単語分割における未登録語の認識誤りを書き換え規則によって訂正する手法について述べる。書き換え規則は、誤り駆動学習モデルをもとに、コンピュータによる未登録語を含む中国語単語分割の結果と「正解」の差分から、単語の表記や品詞などの情報を使って生成する。その後で、生成した書き換え規則を学習データに適用し、その信頼度を評価する。信頼できない規則は削除する。最後に、未登録語を含む単語分割の結果を用いて未登録語の認識実験を行う。実験の結果、８４％の未登録語の認識に成功した。

キーワード：中国語、単語分割、未登録語、書き換え規則、誤り駆動学習モデル、実験

An Application of Error-Driven Model to Unregistered Chinese Word Identification

Dongli Han and Teiji Furugori
Department of Computer Science
The University of Electro-Communications

A method of identifying unregistered words using transformation rules is proposed for Chinese word segmentation. We use various information of words and parts of speech convey to generate the transformation rules automatically with an error-driven learning model. Each rule is assigned a relative value that shows its reliability. And then we discard unreliable rules. Using our method and a closed corpus, we got a success rate of 84% in an unregistered word identification experiment.

Keywords: Chinese, word segmentation, unregistered word, transformation rules, error-driven learning model, experiment

1 はじめに

日本語のような膠着語には、機械による文法分析や意味解析を行う前に、単語単位への分割の問題がある。中国語も膠着語のため、単語分割は避けて通ることのできない問題である。

本稿では、誤り駆動型モデルをもとにした中国語単語分割における未登録単語の認識手法を提案し、その実験結果を報告する。

2 中国語単語分割での未登録語の認識

近年、自然言語処理が急速な発展をみている中で、中国語の単語分割の研究がますます脚光を浴びるようになってきた。

入力文を単語単位に分割するとき、電子化辞書を用いて単語を認識する。しかし、すべての単語が辞書に登録してあるわけではない。たとえば、中国語の人名や地名など固有名詞は無制限にあるため、辞書には登録されていない方が普通である。

中国語単語分割には、未登録語を検出することと、複数の分割が可能になった場合、そのうちのどれを正解として選択するかの問題がある。本稿では、未登録語、とくに固有名詞の正しい認識の仕方に重点をおいた分析手法を考える。

中国語の人名や地名などの固有名詞は、文中にまったく普通の漢字列として現れる。固有名詞を構成する漢字のほとんどはそれぞれ独立しても使われる。つまり、固有名詞の一部分として使われる漢字も普通の単語か、その一部分になり得るものである。しかも、その多くは二つ以上の品詞範疇をもち、たとえば、ある人の名前は形容詞にも動詞にもなり得る。これは、逆に動詞や形容詞と見なされたものが人名や地名の場合もあるということである。一例をあげよう。

王平原　（姓名）
王　（姓、名詞）
平　（形容詞、動詞）
原　（名詞、形容詞）
平原　（名詞）

姓としてしか使えない漢字の数は決まっている。しかし、名を構成する漢字にはほとんど制限がない。どんな漢字でも名の一部分として使える。したがって、人名を構成する漢字の組み合わせは無制限にある。

中国人の姓名の字数は2から5までである。時代、地域によって、姓名として使われている漢字は異なる。

単語分割をする際、未登録語に遭遇したとき、それを正確に処理しないと、分割が不可能になったり、別の単語として分割されたりすることによって、その後の処理に重大な影響を与える。未登録語を誤って検出した例と正しく検出した例を示そう。

江泽民主席预定下月访问日本

↓

誤った未登録語の検出：
江 | 泽 | 民 | 主席 | 预定 | 下月 | 访问 | 日本

正しい未登録語の検出：
江泽民 | 主席 | 预定 | 下月 | 访问 | 日本

中国語言語処理においては、未登録語の認識問題が大きなボトルネックとなっている。この問題により良い解決策を与えられれば、中国語言語処理はさらに前進しよう。

3 未登録語の認識手法

未登録語の認識には、大別して規則に基づく手法と統計的な手法とがある。

規則に基づく手法は、中国語の諸現象を整理することによって、規則を引出し、それに従って未登録語を認識する手法である。たとえば、Liang-Jyh Wang らは、サブ言語(sub-language)の考えをもとに、タイトル名詞（たとえば、地位や役目などを表す名詞）を情報として用い、人名の認識を行った[1]。Zheng Jiaheng と Liu Kaiying は、中

国語のコーパス（１０万語）に現れたすべての中国人の姓と名を人手で分析したうえで、人名を認識するための規則をまとめた[2]。Sun Maosong と Zhang Weijie は、いくつかの表をつくり、それらを用いることによって、中国語に訳された英語の人名認識をした。これらの表として、常用英語人名漢字表やタイトル名詞表などがある[3]。Shen Dayang らは中国の地名を認識するため、地名コーパスをつくり、それに基づいて地名としてよく使われている漢字や地名の構成規則などを抽出している[4]。

規則に基づく手法は、解選択の基礎となる規則を抽出するために人間の直感や参照用の小規模なテキストデータだけを用いているが、比較的効率のよい結果（規則の集合）を得ている。規則の集合は、可読で、後の修正や追加なども簡単である。

統計的な手法では、大規模テキスト、つまりコーパスを用いて、漢字と漢字の共起情報を求めることによって、特定の漢字列を固有名詞として認識する。 Chang, J.S.らは二重コーパスの手法を提案し、普通のコーパスと人名コーパスを用いて、入力文にあるすべての可能な単語や人名の確率を計算する上で、最優パスを求めている [5]。Hsin-his Chen と Jen-Chang Lee は三種類の固有名詞（人名、地名、組織名）を認識する方法を検討している。彼らはコーパスからある漢字が人名として使われる確率とそれが普通の単語として使われる確率を計算した上で、隣接している漢字の相互情報量を考慮に入れながら、入力文にある漢字が人名なのかほかの単語なのかを決定している[6]。

最近の研究には、統計的な手法を用いたものが多い。しかし、この方法には次のような欠点がある。

（１）スパース問題を避けて高精度なモデルを構成するために、タグ付きの大規模なデータが必要となる。これを作成するには、多くの人手と時間がかかる。

（２）データの巨大化は、解析速度に悪影響を与える。

（３）確率を計算することによって得られた知識は人間には読みにくいので、知識の形式化が難しくなる。

我々は、それぞれの手法の利点と欠点を考慮した上で、未登録語の認識を規則に基づいて行う手法を提案し、実験によりその評価をする。ここでは、人手ではなく、機械学習の結果を用い、恣意性を排し、一般性をもった規則の生成を試みる。

4 誤り駆動モデルによる未登録語の認識

４．１誤り駆動型学習モデル

誤り駆動型モデルは、機械学習モデルの一つとして、自然言語処理にも使われる。図１に示すように、誤り駆動型モデルは未解析のデータをトレーニングデータとして初期状態生成プロセスに渡す。次に、生成された初期状態を比較学習メカニズムの入力とし、ここであらかじめ用意された正解データと比較しながら規則を生成する。最後に、生成された初期規則の信頼度を評価することにより優先順位を付けた最終規則をつくる。

Eric Brill は誤り駆動モデルを用いて英単語品詞の決定を試みた[8]。久光徹と丹羽芳樹は日本語形態素解析における誤りを誤り駆動モデルにより修正した[9]。Julia Hockenmaier と Chris Brew [7]、David D.Palmer[11]は、誤り駆動モデルを中国語単語分割にうまく導入している。同様に、Chang Baobao、Liu Ying と Liu Qun は、誤り駆動モデルを中英機械翻訳上での冠詞選択の問題に使っている[10]。

４．２．誤り駆動モデルによる未登録語の認識

初期状態生成プロセス 我々は、初期状態生成プロセスを初期単語分割と初期品詞付けの二つのステップに分ける。初期単語分割では、トレーニングデータを大ざっぱな単語単位に分割する。この過程で、香港中文大学（The Chinese University of Hong Kong）のオンライン中国語処理システム Jasmine[12]の一部（単語分割）を用いる。トレーニングデータには、未登録語が多数入っている新聞記事からランダムに抽出したテキストを使用す

る。初期単語分割の終了後、分割されたトレーニングデータに対し初期品詞付けを行う。この時、品詞分類を拡張し、一般の中国語文法辞書にある品詞体系より詳細な品詞体系を用いて品詞付けをする。

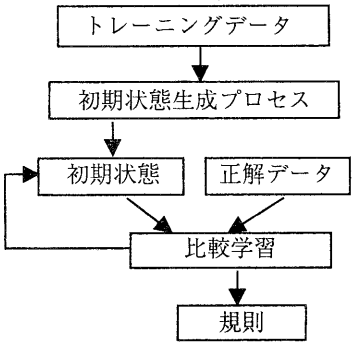

図１　誤り駆動型学習モデル

中国語の拡張品詞体系　中国語の文法辞書に従うと、品詞には次のような体系がある。

名詞
- 普通名詞 ng (common noun)
- 場所名詞 nf (locative noun)
- 時間名詞 nt (temporal noun)
- 人名 nm (personal name)
- 地名 nd (name of a place)
- 組織名 nz (name of an organization)
- その他の固有名詞 no (other names)

動詞
- 普通動詞 vg (common verb)
- 助動詞 vz (auxiliary verb)
- 連結動詞 vx (link verb)

その他
- 前置詞 pp (preposition)
- 形容詞 aa (adjective)
- 副　詞 dd (adverb)
- 助詞 us (genitive)

我々は、中国語未登録語特に固有名詞の特徴について検討した上で、さらに次のような品詞を定義する。

名字１nx (surname・姓としてしか使わない漢字)

名字２nb (surname+other・姓としても他の品詞としても使える漢字)

身分表示名詞 ni (identity or official rank of a person)

接頭辞 np (prefix・よく氏名の前につける漢字)

機構種類名詞 nj (name of the type of an organization)

行政単位名詞 na (name of the administration unit)

非単語漢字 nc (single character that never works as a word or as a surname
単語でない独立漢字)

人間行為動詞 vp (verb that usually follows a personal name or is followed by a personal name)

学習メカニズムと初期規則の生成　まず、初期単語分割と初期品詞付けをしたトレーニングデータに含まれる固有名詞の認識誤りを手動で修正し、正解データを用意する。次に、手動で修正する前のデータ（初期状態）と正解データを比較しながら、初期規則を以下の手順で生成する。ここで、各記号の意味は

W_i ：i 番目の単語　tag_i ：W_i の品詞 $(2 \leq i \leq n)$

A：W_1 の品詞　B：W_{n+1} の品詞

if

正解文 = W_1/A $[W_2 W_3 \cdots W_n]$/固有名詞

W_{n+1}/B

and

初期状態 = $W_1/A \quad W_2/tag_2 \quad W_3/tag_3 \quad \cdots$

$W_n/tag_n \quad W_{n+1}/B$

then 次のような規則を生成する。

If A, tag_2, tag_3, … ,tag_n,B

then W_2W_3 … W_n is a proper name

初期規則の信頼度評価 規則には信頼度を付与する。信頼度の付与は次の形式に従う。

まず上の手順で生成された規則 rule(i)を学習データ全文に適用する。そしてこの規則で未登録語を正しく識別した個所の数を Y(i)、誤った個所の数を R(i)とする。

$$信頼度\ R(i) = Y(i) - N(i)$$

ここで、試行実験の結果から、信頼度 R の閾値 r（下限値）を 2 とする。しかし、場合によっては R が大きくても規則があまり信頼できないことがある。たとえば、Y(i)と N(i)が、それぞれ 10 と 8 である場合、未登録語の認識はうまくいった１０個所に対し、間違った個所が 8 個にも達しているので、あまり信頼できる規則だとはいえない。このような状況を避けるため、信頼度 R に加え、もう一つの評価関数を用意する。

$$補助信頼度\ A(i) = \frac{Y(i) - N(i)}{Y(i) + N(i)}$$

A の閾値 a（下限値）は試行実験の結果から、0.5 とする。

次に、以下の手順で信頼度 R と補助信頼度 A を用いてルールの信頼度を評価し、信頼できないルールを削除する。

```
For i = 1 to number(rules) {
  if (R(i) < r) or (A(i) < a) then delete rule(i)
}
```

最終規則の生成 残っているすべての規則を R に従って順序リストに格納し、未登録語を認識するための最終規則を生成する。

規則の生成例 トレーニングデータから規則の生成に至るまでの機械学習の過程を簡単な例で示す。

トレーニングデータ文：

工业局局长张向东发表关于工业形势的讲话

（日本語の訳文：工業局の局長である張向東が工業発展の状況について発言した）

↓

初期単語分割の結果：

工业|局|局长|张|向|东|发表|关于|工业|形势|的|讲话

↓

中国語拡張品詞体系に基づいて初期品詞を付けた結果：

工业/ng 局/na 局长/ni 张/nb 向/pp 东/nf 发表/vp 关于/pp 工业/ng 形势/ng 的/us 讲话/ng

手動で得られた正解：

工业/ng 局/na 局长/ni 张向东/nm 发表/vp 关于/pp 工业/ng 形势/ng 的/us 讲话/ng

↓

初期品詞を付けた結果と正解の比較：

局长/ni ［张/nb 向/pp 东/nf］ 发表/vp

局长/ni ［张向东/nm］ 发表/vp

↓

比較により生成された規則：

If [W_1/ni W_2/nb W_3/pp W_4/nf W_5/vp]

Then tag $W_2W_3W_4$ as a whole with nm（人名）

↓

最後に、生成された初期ルールを学習データ全文に適用し、信頼度の評価を行う。

5 実験

中国語の新聞記事から固有名詞の入っている200 文をトレーニングデータとしてランダムで抽出し、Jasmine システムの単語分割ツールにより単語分割をした。その結果から書き換え規則の生成及び規則を用いる未登録語の認識実験を行った。行われた実験は閉じた実験（closed test）である（トレーニングデータとテストデータが同じもの）。

次の再現率（Recall）と適合率（Precision）は、実験結果を評価するものである。

$$(1)\ \text{Recall} = \frac{C}{T}$$

$$(2)\ \text{Precision} = \frac{C}{K}$$

ここで、K は抽出した規則の集合を用いて検出した未登録語の数、C はそのうちで正しく認識された未登録語の数である。T は手動で検出した固有名詞の数である。

抽出した規則の数は２３８個あったが、その中に重複したものが４６個あったため、実際に抽出したのは１９２個である。これらの１９２個の規則を用いて、学習データに適用した。その結果、Recall＝84%、 Precision ＝92%を得た。

6 おわりに

本稿では、誤り駆動学習モデルにより生成された書き換え規則を用いて未登録語を認識する手法とその実験結果を述べた。その手法の特徴は次の諸点にある。

- 関連研究の多くが固有名詞の一部（人名や地名）に焦点を当てた分析をしているのに対し、すべての辞書未登録語を対象とした分析をしていること。
- 認識規則を小規模なコーパスから抽出することができるので、手法全体がほかの言語（日本語やタイ語など）に容易に応用可能なこと。
- より自動的、客観的な認識規則の生成が短時間でできること。

しかしながら、実験の精度をさらに上げるためには次のようなことを考慮する必要があろう。

- 対象とする単語の左右一単語までを考慮して、未登録語の認識規則の生成を行った。これを左右二単語まで拡張して、認識規則を生成すること。
- 今回の実験では 200 文の学習データしか使わなかったので、より全面的な規則集合を得るために、学習データ増やすこと。
- 学習データを拡大すると、生成された規則の集合が巨大化する恐れがある。これを避けるため、生成された規則を分類し、統合化すること。

なお、今回の実験ではクローズドテストを試みたものである。当然のことながら、今後はトレーニングデータとテストデータに違いのあるオープンテストを行い、さらにこの手法の有効性を検証する必要もある。

参考文献

1. Liang-Jyh Wang, Wei-Chuan Li, and Chao Huang Chang: Recognizing Unregistered Names for Mandarin Word Identification, Proceeding of Coling-92, Nantes,Aug.23-28, 1992
2. Zheng Jiaheng, Liu Kaiying: Approach of Processing Tactics on the Names and Surnames in Chinese Automatic Segmenting System, <<计算语言学研究与应用>>, 北京语言学院出版社 1993.10
3. Sun Maosong,Zhang Weijie: Translated English Name Identification in Chinese Texts, <<计算语言学研究与应用>>, 北京语

言学院出版社 1993.10

4. Shen Dayang,Sun Maosong : Identifying Chinese Place Names in Unrestricted Text, <<计算语言学进展与应用>> 北京语言学院出版社 1993.10
5. Chang, J.S. et al: Large –Corpus Based Methods for Chinese Personal Name Recognition, Journal of Chinese Information Processing, Vol.6, No.3
6. Hsin-his Chen and Jen-Chang Lee: Identification and Classification of Perpor Nouns in Chinese Texts, Proceedings of 16th International Conference on Computational Linguistics. Copenhagen, Denmark, August 5-9, 1996
7. Julia Hockenmaier and Chris Brew: Error-Driven Learning of Chinese Word Segmentation, Language, Information and Computation (PACLIC12), 18-20 Feb,1998
8. Eric Brill: Transformation-Based Error-Driven Learning and Natural Language Processing。A Case Study in Part-of-Speech Tagging, Computational Linguistics, 1995 Vol.21, No.4
9. 久光徹、丹羽芳樹: 書き換え規則と文脈情報を用いた形態素解析後処理, 自然言語処理 126-8,1998
10. Chang baobao, Liu ying and Liu qun: Research on the processing of articles in Chinese-English machine translation, Journal of Chinese Information Processing, Vol.3, 1998
11. David D.Palmer: A Trainable Rule-based Algorithm for Word Segmentation, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL '97), Madrid, 1997
12. W. S. Wong and A. Qin: Jasmine Automatic Chinese Processing System (http。//peflinux0.ie.cuhk.edu.hk/~access/index.ht ml), December,1995