文字ベースの中国語単語分割における出所の異なる辞書の比較

内山 将夫　　井佐原 均

独立行政法人情報通信研究機構 (NICT)

1　はじめに

中国語の単語分割は，単言語情報検索や言語横断検索を含む中国語処理において，最初に必要な処理である．中国語の単語分割は，これまでに，評価型ワークショップを含めて，様々な研究が行なわれている [5]．それらの研究の焦点の１つが未知語の処理である．ただし，未知語とは，解析システムの辞書に含まれない単語であると定義する．

未知語に対処する方法としては，辞書を拡張する方法と，解析アルゴリズムを拡張する方法があるが [12]，本稿では，辞書を拡張する方法について検討する．なお，本稿での辞書とは，単に，単語をリストしたものである．

本稿で比較する辞書は，出所の異なる次の 2 種類の辞書である．

- 訓練データから作られた辞書 (訓練辞書)
- 雑多なデータから作られたなるべく大きな辞書 (雑多辞書)

雑多辞書の利点は，色々な単語を含むため，訓練データ以外のテキスト (テストデータ) に対してシステムが適用されたときであっても，そのテストデータ中に含まれる単語を多く含むことが期待できることである．すなわち，テストデータにおける未知語の割合 (未知語率) が低いことである．一方，雑多辞書の欠点は，辞書中の単語の区切りと訓練データ中の単語の区切りとが一致しないことにより，解析誤りが生じる可能性があることである．また，訓練データとテストデータの分野が同じときには，雑多辞書を利用すると，余計な単語を考慮する必要がでるため，かえって精度が低下する可能性がある．

これと反対に，訓練辞書においては，そこに含まれる単語は，訓練データ中に含まれる単語のみからなるので，辞書中の単語と解析単位としての単語とに区切りの不一致が生じることはない．また，訓練データとテストデータの分野が同じときには，適切な単語のみが辞書に登録されていることが期待できる．しかし，訓練辞書は雑多辞書に比べると小さいので，テストデータにおける未知語率は一般には高くなる．

たとえば，第 2 回 Chinese Word Segmentation Bakeoff [5] における Microsoft Research コーパス (MSR) と Peking University コーパス (PKU) においては，それぞれの 10 万 6 千語と 10 万 4 千語のテストデータに対して，訓練辞書の未知語率は，MSR では 0.021 であり PKU では 0.033 であるが，雑多辞書[1]の未知語率は，MSR では 0.016 であり PKU では 0.013 である．

次に，訓練辞書や雑多辞書を拡張する簡単な方法として，本稿では，テストデータを自動解析した結果に含まれる単語のリスト (テスト辞書) を元の辞書に追加することを考えた[2]．この追加は，訓練辞書に対しても雑多辞書に対しても行なうことができる．そして，それによる未知語率の減少は大きい．たとえば，訓練辞書においては，MSR では 0.021 から 0.011，PKU では 0.033 から 0.016，雑多辞書においては MSR では 0.016 から 0.007，PKU では 0.013 から 0.006 にまで未知語率が減る．

このように，テスト辞書の追加により，未知語率は大きく減少する．しかし，それと同時に，間違った区切りからなる「単語」も追加されてしまう．そのため，未知語率の減少が必ずしも解析精度の向上につながらない可能性がある．

なお，テスト辞書の追加により精度向上が見込めるのは，以下の理由による．まず，未知語の多くは固有名詞であると考えられるが，固有名詞は，

[1] 本稿での雑多辞書とは，MSR と PKU の訓練辞書を統合し，更に，同 Bakeoff における Academia Sincia コーパスと City University of Hong Kong コーパスの訓練辞書を加えたものに，Linguistic Data Consortium (LDC) の中英・英中辞書から中国語単語を抽出したものを加えたものである．

[2] テスト辞書の追加は，詳細は異なるが，文献 [6] で提案されている．

一回出現すると，続けて出現する可能性が高い [4]．こうした場合，この複数回の出現のうちで，一回でも解析に成功した場合には，それがテスト辞書に登録される．そうすれば，そのテスト辞書を利用して再び解析した場合には，今度は，初回に解析に失敗した個所においても，テスト辞書に登録されているおかげにより，解析に成功する可能性がある．また，このようなことが期待できるテストデータのサイズは，1 記事単位程度からであるので，テスト辞書を利用する方法は，簡単でかつ適用の範囲が広い．

以上のように，訓練辞書や雑多辞書やテスト辞書には，利点と欠点があるため，本稿では，それらの利用が，どの程度のトレードオフの関係にあるかを調査した．

このようなトレードオフを調べるためには，様々な解析システムを利用することが理想であるが，本稿では，最大エントロピー法 [1, 9] に基づく文字ベースの中国語単語分割システムのみを利用した．中国語単語分割システムには，単語ベースのもの，文字ベースのもの，両者を混合したものが考えられるが [12]，他のシステムについての調査は今後の課題とする．しかし，本稿での調査結果は，他のシステムに対しても，同様に成立する可能性が高いと考える．

2 単語分割システム

本稿では，最大エントロピー法に基づく文字ベースの中国語単語分割システムを利用した．その大枠は，文献 [7, 10] 等と同様に，各文字に，その文字が単語の先頭なら「B」，単語の中間なら「I」，単語の末尾なら「E」，1 文字で単語を構成するなら「S」をタグとして付与するというものである．なお，解析システム[3]の実装には maxent[4] を利用し，Viterbi 探索により最大確率のタグ列を得た [8]．ただし，探索時には「BS」のような矛盾したタグ列は考慮しないようにした．また，以下で記述する素性については，訓練データにおける頻度が 1 よりも大きいもののみを利用した．更に，過適合を防ぐために，Gaussian Prior を利用した [3]．また，パラメタ推定には L-BFGS-B 法 [11] を利用した．

3 分割のための素性

分割に利用した素性は，文字およびタグに関する素性と辞書に関する素性である．

文字およびタグに関する素性は，文字を C，タグを T とすると以下である．

$$
\begin{array}{l}
C_{-2}, C_{-1}, C_0, C_{+1}, C_{c+2}, \\
C_{-1}C_0, C_0C_{+1}, \\
C_{-2}C_{-1}C_0, C_{-1}C_0C_{+1}, C_0C_{+1}C_{+2}, \\
T_{-2}, T_{-1}, T_{-2}T_{-1}, \\
T_{-2}C_{-2}, T_{-1}C_{-1}, T_{-2}C_{-2}T_{-1}C_{-1}, \\
T_{-1}C_{-1}C_0, T_{-2}C_{-2}T_{-1}C_{-1}C_0
\end{array}
$$

ただし，タグ付けの対象である文字の添字を 0，その前後 i 番目の添字を $-i, +i$ で表現している．なお，T は前述の「BIES」のどれかである．また，文字を素性として利用する場合には，アラビア数字と漢数字は「一」に置き換え，アルファベットは「A」に置き換えた[5]．これらの素性は文献 [7, 10] 等を参考にした．

辞書に関する素性は，各文字に付与する後述する「辞書タグ」D により，以下で定義される．

$$
\begin{array}{l}
D_{-2}, D_{-1}, D_0, D_{+1}, D_{+2}, \\
D_{-1}D_0, D_0D_{+1}, \\
D_{-2}D_{-1}D_0, D_{-1}D_0D_{+1}, D_0D_{+1}D_{+2}, \\
D_{-2}C_{-2}, D_{-1}C_{-1}, D_0C_0, D_{+1}C_{+1}, D_{+2}C_{+2}, \\
D_{-1}C_{-1}D_0C_0, D_0C_0D_{+1}C_{+1}, \\
D_{-2}C_{-2}D_{-1}C_{-1}D_0C_0, D_{-1}C_{-1}D_0C_0D_{+1}C_{+1}, \\
D_0C_0D_{+1}C_{+1}D_{+2}C_{+2}
\end{array}
$$

「辞書タグ」とは，解析対象の文に対して，辞書を利用した分割をしたときに，その結果として付与される可能性のある「BIES」タグの和集合である．たとえば「VWXYZ」という文について，辞書に「VW」「WX」「X」「Z」という単語があるときには,「VW」「WX」「X」「Y」「Z」という分割が得られる可能性がある．すると，V には B，W には E か B，X には E か S，Y には S，Z には S が分割のタグとして割り当てられる可能性がある．したがって，辞書タグとしては，V には「B」，W には「BE」，X には「ES」，Z には「S」がそれぞれ割り当てられる．ただし，辞書にない単語

[3] 第 1 著者から入手可能である．

[4] www2.nict.go.jp/jt/a132/members/mutiyama/software.html

[5] 「はじめに」で述べた各コーパスの未知語率は，これらの文字を置き換えた後の単語についてである．

Y には，特別に NIL を辞書タグとして割り当てる．辞書タグの定義は，文献 [2] を参考にした．

なお，辞書タグを利用するにあたって，訓練データ中に出現する単語を全て辞書に登録すると，訓練データでは，NIL となる文字がなくなるので，テストデータを解析するときとのミスマッチが起きる．それを防ぐために，訓練においては，訓練データ中に一回しか出現しない単語を訓練辞書から除いた．ただし，テスト時には，訓練データに一回しか出現しない単語も辞書に追加した．

また，訓練辞書やテスト辞書を構築するときには，データ中に何回以上出現した単語を加えるかという選択が考えられる．これについては，全て加える場合と，2 回以上出現した場合などを比較したが，その結果，解析精度にはそれほどの差がなかったので，全て加える場合についてのみの実験結果を示す．

4 実験

実験では，簡体字のコーパスである MSR と PKU を利用した．その理由は，我々が中国語単語分割をする必要があるテキストが簡体字のテキストであるので，そのための解析システムを作成するというのが，中国語単語分割における当初の目的であったからである．

表 1 には，単語分割の評価結果を示す[6]．表における「コーパス」の列は，単語分割の対象としたコーパスであり,「辞書」の列は，解析システムが利用した辞書が「訓練辞書」か「雑多辞書」であるかを示している.「なし」の場合には辞書は利用していない.「追加」の列は，テスト辞書を各辞書に追加したかどうかを示す.「再現率」の列は，テストデータにおける正解として与えられた単語分割において，どの程度の割合 R が，実際に，システムによる分割に現れたかを示し,「適合率」の欄は，システムに分割された単語のうちで，どの程度の割合 P が正解の単語であったかを示す.「F 値」は $\frac{2PR}{P+R}$ である.「Roov」は，テストデータに出たが訓練データに出ない単語を未習語と定義する (訓練データに出た単語を「既習語」と定義する) とき，未習語のうちでシステムが正しく再現した割合である.「Riv」は，既習語をシステムが正しく再現した割合である．なお，再現率 R（および適合率 P）が 2 項分布に従うと仮定すると，単語数が n のときに，その標準偏差は $s = \sqrt{\frac{R(1-R)}{n}}$ である．そして，$R = 0.95, n = 10$ 万 のときには，$2s \sim 0.001$ である．したがって，再現率や適合率に 0.002 程度の差がある場合には，それらの間には有意水準 1%程度の有意差がある．参考のために，Bakeoff における MSR と PKU のオープンテスト[7]での最高の F 値は，MSR では 0.972，PKU では 0.969 である．

まず，F 値に注目すると，表 1 から，MSR と PKU の双方において，辞書を利用した方が精度が高いことが分かる．

次に，訓練辞書と雑多辞書を比較すると，MSR では訓練辞書の方が F 値が高く，PKU では雑多辞書の方が F 値が高い．これは，MSR のテストデータでの未習語の割合 (未習語率) が 0.026 であるのに対して，PKU では 0.058 であることから，未習語率の比較的少ない MSR ではノイズの少ない訓練辞書が有効であり，未習語率の高い PKU では多くの単語を含む (未知語率の低い) 雑多辞書が有効であると考える．すなわち，訓練データとテストデータが良く似ているときには，訓練辞書を利用し，そうでないときには，雑多辞書を利用するのが良いと考える．

最後に，テスト辞書の追加のありなしを比較すると，MSR でも PKU でも，テスト辞書の追加により，0.001 だけ F 値が向上している．ここで Roov に注目すると，テスト辞書の追加により，0.01〜0.05 程度向上している．また，Riv については，0.000〜0.003 程度の減少である．したがって，テスト辞書の追加は，既習語の再現率を減少させることなく，未習語の再現率の向上に役立つ．ただし，テストデータにおける未習語の割合は小さいので，全体における F 値の向上は少ない．

[6] Bakeoff で使われた score スクリプトを利用して評価した．なお，文献 [5] にあるように，score スクリプトの利用において，再現率が誤って高く計算される場合があった．これは環境変数 LANG が ja_JP.EUC-JP のときに生じた．環境変数 LANG を C にした場合には正しい再現率が計算された．表 1 にはその正しい結果が示されている．

[7] テストデータの単語分割に際して，訓練データ以外にも任意の言語資源を利用できる形式のテストをオープンテストと呼ぶ．

表 1: 評価結果

コーパス	辞書	追加	再現率	適合率	F 値	Roov	Riv
MSR	なし	なし	0.948	0.947	0.948	0.644	0.956
MSR	訓練	なし	0.973	0.960	0.967	0.490	0.986
MSR	訓練	あり	0.974	0.963	0.968	0.543	0.985
MSR	雑多	なし	0.959	0.959	0.959	0.636	0.967
MSR	雑多	あり	0.959	0.961	0.960	0.682	0.967
PKU	なし	なし	0.927	0.933	0.930	0.758	0.937
PKU	訓練	なし	0.945	0.945	0.945	0.617	0.965
PKU	訓練	あり	0.944	0.949	0.946	0.647	0.962
PKU	雑多	なし	0.951	0.958	0.954	0.807	0.959
PKU	雑多	あり	0.951	0.959	0.955	0.817	0.959

5 おわりに

本稿では，中国語単語分割における未知語に対する対処という観点から，訓練辞書と雑多辞書の有効性を比較した．また，これらに，テストデータにおける単語を追加することにより，訓練データ中に出現した単語の再現率を減少させることなく，訓練データ中に出現しない単語の再現率を向上できることを確かめた．テスト辞書の利用は，簡単で適用範囲が広い方法であるため，本稿で利用した文字ベースの中国語単語分割システムだけでなく，その他の解析システムにおいても有効な方法であると考える．

参考文献

[1] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language modeling. *Computational Linguistics*, Vol. 22, No. 1, 1996.

[2] Andrew Borthwick. *A Maximum Entropy Approach to Named Entity Recognition*. PhD thesis, New York University, 1999.

[3] Stanley F. Chen and Ronald Rosenfeld. A Gaussian prior for smoothing maximum entropy models. Technical Report CMU-CS-99-108, Carnegie Mellon University, 1999.

[4] Kenneth W. Church. Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than p^2. In *COLING-2000*, pp. 180–186, 2000.

[5] Thomas Emerson. The second international Chinese word segmentation bakeoff. In *Fourth SIGHAN Workshop on Chinese Language Processing*, 2005.

[6] Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto. Pruning false unknown words to improve Chinese word segmentation. In *PACLIC-18*, pp. 139–149, 2004.

[7] Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. A maximum entropy approach to Chinese word segmentation. In *Fourth SIGHAN Workshop on Chinese Language Processing*, 2005.

[8] Masaaki Nagata. A stochastic Japanese morphological analyzer using a forward-DP backward-A* n-best search algorithm. In *COLING-94*, pp. 201–207, 1994.

[9] Adwait Ratnaparkhi. *Maximum Entropy Models for Natural Langauge Ambiguity Resolution*. PhD thesis, University of Pennsylvania, 1998.

[10] Nianwen Xue. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, Vol. 8, No. 1, 2003.

[11] C. Zhu, R. H. Byrd, and J. Nocedal. L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, Vol. 23, No. 4, pp. 550–560, 1997.

[12] 中川哲治, 松本裕二. 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. 情報処理学会論文誌, Vol. 46, No. 11, 2005.