文字ベースのHMMによる複合語単語分割の誤り修正

○池谷昌紀　　新納浩幸

茨城大学 工学部 システム工学科

1 はじめに

本論文では複合語の単語分割を行うために，通常の形態素解析と文字ベースのHMMとを相補的に利用する手法を提案する．

複合語の単語分割は従来の言語処理システムの一要素技術として重要であるだけでなく，全文検索で生じるインデックス作成や検索式の解析などにも必要とされ，その重要性は高い．複合語の単語分割は一般に形態素解析によって行えるが，分割精度の他に未知語の扱いも問題となるため，形態素解析を用いない手法も提案されている．その中の一つとして文字ベースのHMMもある．文字ベースのHMMでは，文字ベースなので未知語の問題は生じない．ただし文字ベースのHMMにおいて，状態 i から状態 j に遷移するときに文字 a が出力されるコスト $B_{ij}(a)$ を uni-gram や bi-gram などから得ると，長い文字列が一単語となる場合に正しく単語分割が行えないことが多い．そのためにプラスアルファの何らかの工夫が必要となる．山本は単純な文字ではなく拡張文字として品詞などの情報も文字に付加している [7]．また Tsuji は対訳辞書から形態素数を得る工夫を行っている [4]．また小田は PPM^* モデルを利用して出力シンボルを可変長 n-gram にしている [8]．

ただし上記の問題は形態素解析では生じない．形態素解析による単語分割では，長い文字列からなる一単語を正しく認識することはむしろ容易である．一方，形態素解析による単語分割の誤りは文字列の局所的な部分であり，あるパターンが存在する．このような文字列の局所的な部分の単語分割に対しては，文字ベースのHMMが有効である．

つまり単語分割に対して文字ベースのHMMと一般の形態素解析は相補的に利用できる．そこで本論文では，形態素解析で生じた誤りを文字ベースのHMMから修正することで単語分割を行う．形態素解析の誤りをどのように判断するかが問題であるが，ここでは形態素解析の誤りパターンに注目し，該当パターン部分を文字ベースのHMMによる単語分割結果と比較することで，誤りかどうかを判断する．

また本論文で利用した形態素解析システムは JUMAN 3.5 であることを注記しておく．

2 文字ベースのHMMによる単語分割

2.1 文字ベースのHMM

HMM $\boldsymbol{M}$ は以下の六つの組で定義される．

$$\boldsymbol{M} = (\boldsymbol{S}, \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}, \boldsymbol{F})$$

$\boldsymbol{S}$: 状態の集合
$\boldsymbol{Y}$: 出力シンボルの集合
$\boldsymbol{A}$: 状態遷移コストの集合
$\boldsymbol{B}$: 出力コストの集合
$\boldsymbol{\pi}$: 初期状態コストの集合
$\boldsymbol{F}$: 最終状態の集合

複合語の単語分割は文字間に単語の境界が存在する (1) か存在しない (0) かのどちらかの記号を割り当てる問題に一般化できる．今，$\boldsymbol{S}$として 1 と 0 の 2 つの状態を用意し，$\boldsymbol{Y}$として文字を考える．また$\boldsymbol{A}$は考慮せず，$\boldsymbol{\pi}$ と $\boldsymbol{F}$ を $\{1\}$ とおく．あとは$\boldsymbol{B}$を設定すれば，文字ベースのHMM $\boldsymbol{M}$ が構築できる（図1参照）．

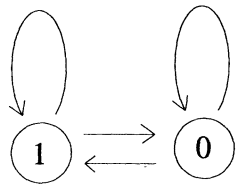

図 1: 構築した HMM

HMM では出力シンボル系列がどの状態をたどってきたかを Viterbi アルゴリズムより推定す

ることができる．すなわち文字間に単語の境界が存在するかしないかが推定でき，単語分割が行える（図2参照）．

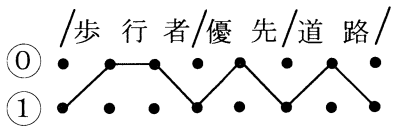

図 2: HMM による単語分割

2.2 文字の出現コストの算出

$B_{ij}(a)$ を状態 i から状態 j へ移るとき文字 a が出現するコストとする．この場合，各文字 a に対する $B_{11}(a)$，$B_{10}(a)$，$B_{01}(a)$，$B_{00}(a)$ を求めることが$\boldsymbol{B}$の構築に対応する．

$B_{ij}(a)$ を求めるために，まず日経新聞 CD-ROM '90 の 1 年分の新聞記事を形態素解析し，複合語を取り出した．取り出した複合語の各文字を見ると，その文字の左右に単語境界があるかどうか，つまり状態がわかる．今，文字 a の左の状態が i，右の状態が j であるような個数を $C(i,j,a)$ とする．例えば，「自然言語」という単語は形態素解析により，/自然/言語/ と分割されるので，$C(1,0,$自$)$,$C(0,1,$然$)$,$C(1,0,$言$)$,$C(0,1,$語$)$ にそれぞれ 1 が加えられる．$B_{ij}(a)$ は以下の形で計算した．

$$B_{ij}(a) = \log_2 \frac{C(i,\bar{j},a)+1}{C(i,j,a)+1}$$

ここで $\bar{j}$ は状態 j ではない状態とする．つまり $j=1$ なら $\bar{j}=0$ であり，$j=0$ なら $\bar{j}=1$ である．

上記の形だけでも$\boldsymbol{B}$は構築できるが，精度をあげるために bi-gram を利用する．まず $C(i,j,ab)$ を考える．これは文字 a の左の状態が i 右の状態が j で更に，文字 a の次の文字が b である個数を示す．この値は先に取り出した複合語から得られる．また $B_{ij}(ab)$ を状態 i から状態 j へ移るときに文字 a が出現し，しかも文字 a の次の文字が b であるときのコストとする．$B_{ij}(ab)$ は以下の式で計算した．

$$B_{ij}(ab) = \log_2 \frac{C(i,\bar{j},ab)+1}{C(i,j,ab)+1}$$

$B_{ij}(a)$ と $B_{ij}(ab)$ から線形和をとり新たな $B'_{ij}(a)$ を以下のように定義した．

$$B'_{ij}(a) = \alpha \cdot B_{ij}(a) + (1-\alpha) \cdot B_{ij}(ab)$$

ここでは $\alpha = 0.3$ とした．

3 文字ベースの HMM による単語分割の修正

3.1 文字ベースの HMM の誤りパターン

文字ベースの HMM では長い文字列が一単語となるような場合に過剰に分割を行い単語分割が誤る．これは文字ベースが基本的に局所的な部分から，単語の境界があるかどうかを判断するために生じている．例えば，JUMAN では「1 次産品共通基金」という文字列が一単語となっている．この文字列が言語的に一単語かどうかは問題ではない．想定しているアプリケーションにおいて一単語として扱いたいと考え，辞書に一単語として登録してあれば，一単語として解析すべきであろう．

さて，この「1 次産品共通基金」を単語分割する場合は一単語とするのが正解である．一方,「1 次産品」を単語分割する場合は，/1 /次/産品/ と，三つの単語列として解析するの正解である．つまり “1” と “次” の間に単語境界が存在するかどうかを判断するには，“1” と “次” だけの局所的な関係では判断できない．さらにこれは “1” と “次産” の関係あるいは “1” と “次産品” の関係などに拡張しても単語境界が存在するかどうかは判断できない．

一方，一般の形態素解析は「1 次産品共通基金」を一単語,「1 次産品」を/1 /次/産品/と解析するのは容易である．

3.2 形態素解析の誤りパターン

一般に形態素解析の誤りは未知語によって生じる．日本語の単語はほとんどの場合，2 文字あるいは 3 文字から構成される．このため未知語部分はほとんど以下のように過分割される．

2 文字の未知語	○○	→	/○/○/
3 文字の未知語	○○○	→	/○/○/○/
			or
			/○/○○/

また未知語ではないが，JUMAN の場合，以下の誤りのパターンも多い．

形態素解析結果	正しい分割
/○/○○～	/○○/○～

上記のケースに共通する特徴として，先頭が1文字で分割されている点がある．つまり形態素解析で1文字単語と認識されている部分は，誤りである可能性がある．

一方，文字ベースの HMM は，このような局所的な単語分割に有効である．

3.3 相補的利用

前述したように，形態素解析と文字ベースの HMM のそれぞれの欠点は，互いに補えることが分かる．

本論文では，形態素解析と文字ベースの HMM を相補的に利用した複合語の単語分割を行う．まず形態素解析による単語分割と文字ベースの HMM による単語分割を並行して行い，両者の結果を比較する．基本的には形態素解析の結果を採用するが，以下のパターン部分は，その部分を HMM の結果に修正する．

表 1: 形態素解析の誤りパターン

形態素解析	HMM
/○/○～/	/○○～/

注意として，/○/○～/ と /○○～/ の文字列の長さは等しく，しかも，単語分割のマークである "/" が一致している部分は，最初と最後の部分の2個所しか存在しない．例えば表2の解析結果の下線部分が上記のパターンに当てはまる．

表 2: パターンの例

形態素解析	HMM
/鈴木/健/四郎/	/鈴木/健四郎/
/永島/帝/二/さん/	/永島/帝二/さん/
/河/口利/加さん/	/河口/利加/さん/
/東/大卒/	/東大/卒/

4 実験

毎日新聞の CD-ROM '94 年度版の最初から 8,543 種類の複合語を取り出した．形態素解析と文字ベースの HMM による単語分割を行った結果，分割結果が一致したものは 7,760 種類 (90.8%)，一致しなかったものは 783 種類 (9.2%) であった．一致しなかったものの中で，修正が生じたものは 212 種類 (2.5%) であった．一部を表 3 に示す．

表 3: 分割の修正

形態素解析	HMM
/延/岡市/	/延岡/市/
/奥谷/喬/司/	/奥谷/喬司/
/追/加点/	/追加/点/
/病/人食/	/病人/食/
/島/内/監督/	/島内/監督/
/若/田光/一/さん/	/若田/光一/さん/
/若/貫兄/弟/	/若貫/兄弟/
…	…
/小泉/順/一郎/郵政相/	/小泉/順一郎/郵政相/
/織田/大/次郎/店長/	/織田/大次郎/店長/

この 212 種類について修正の結果が正解であったかどうかを調べると，178 種類 (84.0%) は正解であったが，34 種類 (16.0%) は不正解であった．ただし，34 種類の中には，形態素解析による単語分割でも誤りである場合や，正解が曖昧なものが 15 種類あった．純粋に形態素解析の方が正しかったものは 34 種類中 19 種類である．つまり修正による悪影響は，修正したもの全体に対して実質 $19/212 = 9.0\%$ であった．

5 考察

5.1 誤りパターンのカバー率

本論文では表1に示した誤りパターンだけに注目している．形態素解析の単語分割の誤りには当然このパターン以外のものも存在する．そのため前述した実験ではどの程度形態素解析の単語分割の精度を向上させることができたのかが示せていない．

ここでは実験で利用した複合語からランダムに 783 種類取り出し，それらに対する形態素解析による単語分割結果を調査した．結果，単語分割の誤りは 28 種類あり，そのうち 26 種類 (92.8%) が本論文で注目したパターンとなっていた．注目したパターンになっている誤りに対して，実験では 84.0% を正しく修正できているので，結果的に誤りの $0.928 * 0.840 = 78.0\%$ を本手法で修正できると考えられる．

5.2 文字ベースの HMM の単独利用

本論文では形態素解析の単語分割結果を優先し，誤りだと予想できるものを文字ベースの HMM

により修正している．この関係を逆転させ，文字ベースの HMM の単語分割結果を優先し，形態素解析により修正することも考えられる．また文字ベースの HMM を単独で利用できるまで改良する方向もある．しかし本論文の立場からは，文字ベースの HMM は形態素解析よりも精度が低く，上記のような方向への発展は難しい．

文字ベースの HMM が形態素解析よりも精度が低いのは，単語の認定の問題があるからだ．複合語の単語分割には本質的に曖昧な部分を持っている．例えば，「修学旅行」を一単語と考えるか /修学/旅行/ と区切るかは曖昧であろう．本論文では単語の認定を形態素解析で利用する辞書に準拠する方針をとった．形態素解析用の辞書で「修学旅行」が一単語で登録されていれば，/修学/旅行/ と区切るのは誤りとしている．この方針であれば形態素解析の方が精度が高くなるだろう．また本論文の学習データは形態素解析結果から作成されているという点もあろう．

5.3 品詞の付与

本手法を使って単語区切りを修正した場合，品詞の情報が欠落するという欠点がある．この問題に対しては部分的ではあるが JUMAN により全ての区切り方を生成することで対処できる．例えば，「東大卒」の JUMAN 解析結果は 4 通りであり，そのうち，/東大/卒/の単語区切りを持つものは以下のものだけである．この結果を流用できる．

東大	（とうだい）	東大	組織名
卒	（そつ）	卒	普通名詞

5.4 関連手法

本論文では未知語に対処するために辞書を使わない手法として文字ベースの HMM を利用した．辞書を使わない手法としては文字間の結合の強さを測る手法もある．結合の強さの測り方として相互情報量 [2] や尤度比検定 [6] を利用することが提案されている．ただしこれら手法も文字ベースの HMM で問題にあげたように長い文字列からなる一単語を過分割する問題がある．

本論文の文字ベースの HMM は形態素解析結果を利用してパラメータを得ているので教師付きの学習に分類できる．この枠組みでは，未知語が大量に出現するような新たな対象領域へ展開することは難しく，教師なしの単語分割の学習が必要になる（[5, 3, 1] など）．この点での拡張が今後の課題である．

6 おわりに

本論文では複合語の単語分割を行うために，通常の形態素解析と文字ベースの HMM とを相補的に利用する手法を提案した．

文字ベースの HMM による単語分割では，長い文字列からなる一単語が過分割されやすい．また形態素解析による単語分割の誤りは，局所的であり，ある単語列のパターンが認められる．このパターンの場合に，文字ベースの HMM による単語分割結果を採用する．

実験では新聞記事から得た 8,543 種類の複合語に対して単語分割を行った．形態素解析による複合語単語分割誤りの 178 種類を正しく修正でき，19 種類は余計な修正であった．対象とした誤りパターンのカバー率から考察すると，誤りのほぼ 4 分の 3（78.0%）を本手法により修正できると推測する．

参考文献

[1] Xiaoqiang Luo and Salim Roukos. An Iterative Algorithm to Build Chinese Language Models. In *The 34th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, pp. 139–143, 1996.

[2] Richard Sproat and Chilin Shih. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Language*, No. 4, pp. 336–351, 1990.

[3] Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. A Stochastic Fnite-State Word-Segmentation Algorithm for Chinese. *Computational Linguistics*, Vol. 22, No. 3, pp. 377–404, 1996.

[4] Keita Tsuji and Kyo Kageura. An HMM-based Method for Segmenting Japanese Terms and Keywords based on Domain-Specific Bilingual Corpora. In *The 4th Natural Language Processing Pacific Rim Symposium*, pp. 557–560, 1997.

[5] 永田昌明. 単語頻度の再推定による自己組織化単語分割. Technical Report NL-121-2, 情報処理学会自然言語処理研究会, 1997.

[6] 影浦峡. 文字単位の bigram 尺度に基づく複合漢字列の単位切り手法. 言語処理学会第 3 回年次大会, pp. 477–480, 1997.

[7] 山本幹雄, 増山正和. 品詞・区切り情報を含む拡張文字の連鎖確率を用いた日本語形態素解析. 言語処理学会第 3 回年次大会, pp. 421–424, 1997.

[8] 小田裕樹, 北研二. PPM^* モデルによる日本語単語分割. Technical Report NL-128-2, 情報処理学会自然言語処理研究会, 1998.