自然言語処理 135－7
（2000．1．27）

文字列が単語になる確率を用いた未知語抽出

池谷昌紀　　新納浩幸

茨城大学 工学部 システム工学科

本論文では2文字あるいは3文字の漢字から構成される未知語をテキストから自動抽出する手法を提案する。漢字から構成される未知語を形態素解析すると、一般にその文字列に対して過分割が生じる。その際、その過分割された各単語に対応する文字列は1単語としてあらわれる確率が低い。この性質を利用して、2文字あるいは3文字の漢字から構成される漢字列が未知語となる尺度を導入する。また漢字から構成される未知語を形態素解析によって過分割した場合、ある分割パターンが認められる。このパターンを利用して未知語候補を集め、次に、上記尺度を適用することで未知語を抽出できる。実験の結果、2文字漢字からなる未知語の抽出はF-値で0.684、3文字漢字からなる未知語の抽出はF-値で0.182であった。本手法は文字列αが未知語かどうかを判定するために、αのトレーニングコーパス中の頻度を必要としない。このため低頻度の未知語も抽出可能という長所を持つ。

Extraction of unknown words by the probability to accept the kanji character sequence as one word

Masanori Ikeya and Hiroyuki Shinnou

Ibaraki University. Dept. of Systems Engineering
{ikeya,shinnou}@nlp.dse.ibaraki.ac.jp

In this paper, we propose a method to extract unknown words, which are composed of two or three Kanji characters, from Japanese text. Generally the unknown word composed of Kanji characters are segmented into some words by the morophological analysis. Moreover, the appearance probability of each segmented word is small. By this characteristic, we can define the measure to accpet two or three kanji characters sequence as an unknown word. On the other hand, we can find some patterns for word segmantation of unkown words. By applying the above measure to Kanji character sequences with these patterns, we can extract unknow words. In the experiment, the F-measure for extraction of unknown words which are composed of two Kanji characters was 0.684 and the F-measure for extraction of unknown words which are composed of three Kanji characters was 0.182. Our method does not need the frequency of the character sequence α in the training corpus to judge whether α is the unknown word or not. Therefore, Our method has the advantage that the low frequent unknown word are extracted.

1 はじめに

本稿では 2 文字あるいは 3 文字漢字からなる未知語をテキストから自動抽出する手法を提案する。

形態素解析の 1 つの課題として未知語処理がある。日本語の場合、未知語によって形態素解析の単語分割が誤り、形態素解析の精度悪化の一因となっている。そのためテキスト中の未知語を検出することは、形態素解析の精度向上に貢献する。また近年、情報抽出の前処理として固有表現抽出の重要性が指摘されている。固有表現抽出を行なう場合、人名、地名、会社名などは未知語としてあらわれやすいため、そこでも未知語を検出する技術が望まれている。

テキスト中のある文字を含むどのような部分文字列をテキストから取り出しても、その文字列が辞書に記載されていない場合、その文字を含む文字列が未知語になっていることがわかる。この種の未知語は記号やカタカナ文字で構成されることが多く、字種の変わり目を確認することで、ある程度、抽出は可能である。未知語抽出の困難な点は既知語から構成される文字列が未知語となる場合である。この種の未知語を検出するには、その単語列が単語列として正しいのか、全体として未知語となっているかの判定が鍵となる。

この判定を行なう場合に、複合語と未知語の区別にまず留意すべきである。例えば、「自然言語」という文字列を考えてみる。これは、「自然」と「言語」という 2 単語から構成される句だと考えれば、複合語である。一方、「自然言語」は 1 つの単語と考え、この単語が辞書に記載されていなければ、未知語となる。1 単語とみなせる複合語を取り出すことも未知語抽出と捉えられるが、ここではそのように捉えないことにする。1 単語として扱える複合語は熟語や連語として捉え、未知語とは考えない。

この立場をとると、既知語から構成される未知語の文字列長は比較的短く、特に、漢字から構成される未知語の長さは 2 あるいは 3 であると考えることができる。本稿では漢字 2 文字あるいは 3 文字から構成される未知語を抽出の対象とする。当然、それ以外の字種から構成される未知語の検出も重要であるが、前述したように、カタカナや記号で構成される未知語は字種の変わり目からある程度判断できるので、ここでは対象としない。また平仮名を含む未知語の抽出も重要な課題ではあるが、新聞記事のような文書では、平仮名で表記される名詞の割合は低いと考え、ここでは対象としない。

ここで提案する手法は、まず形態素解析を行ない、ある単語列が設定したパターンになった場合に、その単語列が未知語の可能性があると考えて、未知語候補として取り出す。次にその単語列を構成する各単語の文字列が 1 単語としてあらわれる確率を定義する。今、単語 w の文字列が α のとき、$w = |\alpha|$ と書くことにする。次に、文字列 α が単語としてあらわれる確率を $P(\alpha)$ と書くことにする。そして単語列 $|\alpha||\beta|$ が未知語となる尺度 $m(\alpha\beta)$ を以下によって定義する。

$$m(\alpha\beta) = 1 - P(\alpha)P(\beta)$$

この値がある閾値を越えた場合に、未知語として検出する。$P(\alpha)$ の計算は、コーパス（トレーニングデータ）から、以下の式で求めておく。

$$P(\alpha) = \frac{c(|\alpha|)}{c(\alpha)}$$

ここで、$c(|\alpha|)$ は単語 $|\alpha|$ のコーパス中の頻度であり、$c(\alpha)$ は文字列 α のコーパス中の頻度である。

本手法は文字列 $\alpha\beta$ が未知語かどうかの判断に、コーパス中の $\alpha\beta$ の頻度を利用していない。このため、コーパス中で低頻度の未知語も検出可能という長所がある。

なお、本稿で用いた形態素解析は「茶筌 (version 2.0b) 」であることを注記しておく。

2 未知語における形態素解析の誤りパターン

本稿では 2 文字あるいは 3 文字の漢字からなる未知語を対象としている。まずはじめに本稿では以下の仮定をおく。

仮定 漢字列 α が未知語であった場合、α を含む文を形態素解析すると α の前と後に単語分割の切れ目が生じる

この仮定に反する例は存在するが、ほとんどの場合この仮定は成立している。

上記の仮定が成立しているとすると、漢字文字 a と b の 2 文字列 ab が未知語であった場合、形態素解析による ab の単語分割は以下のようになる。

パターン 0 ： $|ab| \rightarrow |a|b|$

また漢字文字 a , b , c からなる 漢字列 abc が未知語であった場合、形態素解析による単語分割の結果は以下の 3 つの場合のいずれかになる。

パターン 1 ： $|abc| \rightarrow |a|b|c|$

パターン 2 ： $|abc| \rightarrow |a|bc|$

パターン 3 ： $|abc| \rightarrow |ab|c|$

これらのことから、まずはじめに形態素解析を行ない、単語分割の結果が上記のパターンにあてはまった漢字列を未知語の候補として取り出すことができる。

3 未知語の抽出

3.1 未知語を判定する尺度

文字列 α が 1 単語になる確率を次のように定義する。

$$P(\alpha) = \frac{c(|\alpha|)}{c(\alpha)}$$

ただし、$c(\alpha)$ は コーパス中に文字列 α が出現した回数であり、$c(|\alpha|)$ は文字列 α が 1 単語として出現した回数である。これは、文字列 α があらわれたときに α が 1 単語として解析される確率を表している。

辞書に未登録である α が 1 単語になりやすいということは α が未知語になりやすいということである。このことから、文字列 α が未知語かどうかは $P(\alpha)$ の大きさから判断することができる。しかし、α が未知語である場合、文字列 α が 1 単語として出現する回数 $c(|\alpha|)$ が 0 となるため、α は未知語でないと判断されてしまう。

そこで、本稿では未知語判定の方法として、確率 $P(\alpha)$ の代わりに、文字列 α が未知語になる尺度を用いる。

文字列 α が 形態素解析により $|\beta|\gamma|$ と分割されたとする。文字列が単語となる確率が前後の文字列に依存しないと考えれば文字列 α が 文字列 $|\beta|\gamma|$ に分割される確率は $P(\beta)P(\gamma)$ と近似することができる。

このことから、単語列 α が 1 単語となる尺度 $m(\alpha)$ を以下のように決めた。

$$m(\alpha) = 1 - P(\beta)P(\gamma)$$

この尺度の値が大きいほど、その文字列 α が分割されない可能性が大きい、つまり未知語である可能性が大きい。

またパターン 3 の場合については、m を以下のように定める。

$$m(abc) = \max\{(1-P(a))*m(bc), (1-P(c))*m(ab)\}$$

また 、2 文字列 ab が単語になる確率 $P(ab)$ は

$$P(ab) = \begin{cases} 1 & (c(ab) = 0 \text{ のとき}) \\ \dfrac{c(|ab|)}{c(ab)} & (c(ab) \neq 0 \text{ のとき}) \end{cases}$$

とする。

3.2 適用する順序

前述したパターンについて尺度 m を適用することで、未知語を抽出できる。しかし、尺度の適用には曖昧性が存在する。例えば、漢字列 $abcdefg$ の形態素解析による単語分割の結果が $|ab|c|d|e|fg|$ だとしよう。この場合、$|ab|c|$、$|c|d|e|$、$|c|d|$、$|d|e|$、$|e|fg|$ が未知語の候補となる。ここでもし $|c|d|e|$ が未知語だと判断されれば、その他の $|ab|c|$、$|c|d|$、$|d|e|$、$|e|fg|$ は未知語になることはない。

このような曖昧性に対して、ここでは、パターンによる尺度の適用順序を設けた。具体的には以下の順序を設定した。

パターン 1 --> パターン 0 -->
パターン 2 --> パターン 3

上記の例でいえば、まずパターン 1 の $|c|d|e|$ が未知語かどうかを試す。未知語と判断できれば、単語分割は $|ab|cde|fg|$ に変更する。そして、パターン 0 、パターン 2 、パターン 3 を順次試す。

また、同じパターンでも適用できる場所が複数存在する場合も考えられる。上の例では $|c|d|$ と $|d|e|$ がそれにあたる。このような場合は、尺度 m の大きい方を選択する。

4 実験

毎日新聞 '94 年度の新聞記事 1 年分を形態素解析にかけ単語分割を行なう。これをもとに 2 文字あ

るいは 3 文字の漢字列 α が単語になる確率 $P(\alpha)$ を算出しておく。

4.1 新聞記事からの未知語抽出

毎日新聞 '95 年度の最初の 1,000 文をテスト文として未知語の抽出実験を行なった。最初に手作業により 2 文字列 20 種類、 3 文字列 3 種類の未知語をあらかじめ取り出しておいた。それらを表 1 に示す。

表 1: 未知語

有煙、和魂、硬材、和慶、彌弐、作陶
廣場、星島、彩挺、艶麗、羅湖、老子
祥造、学燈、画期、靖英、普選、深銘
仕梅、女衆、三連星、杉乃井、伊那谷

また、あらかじめ実験により適当な閾値を調べておいた。その結果ここで決定した閾値は、パターン 1 が 0.90、パターン 0、2、3 が 0.98 になった。

表 1 に挙げられている未知語で、本手法で抽出できたものを表 2 に示す。

表 2: 本手法で抽出できた未知語

和魂、硬材、和慶、作陶
星島、彩挺、老子、祥造
画期、普選、深銘、仕梅
女衆、杉乃井

2 文字の漢字からなる未知語において、本手法により抽出した種類数 18 種類のうち、有効な抽出は 13 種類であったので、正解率は $\frac{13}{18} = 0.722$、再現率は $\frac{13}{20} = 0.650$、F 値は $\frac{2 \times 0.722 \times 0.650}{0.722 + 0.650} = 0.684$ となった。また 3 文字の漢字からなる未知語では、本手法により抽出した未知語 8 種類のうち、有効な抽出は 1 種類であったので、正解率は $\frac{1}{8} = 0.125$ 、再現率は $\frac{1}{3} = 0.333$ 、F 値は $\frac{2 \times 0.125 \times 0.333}{0.125 + 0.333} = 0.182$ となった。

4.2 未知語の作成による未知語の抽出

前述した実験のように既存の文書から未知語を抽出する場合、3 文字の漢字からなる未知語の出現確率が低いために、再現率を測ることが困難になる。そこで本稿では 1 つの工夫として、 3 文字の漢字からなる未知語を強制的に作成し、その未知語が抽出できるかどうかを調べる実験を行なった。

まずトレーニングコーパスである毎日新聞 '94 年度の 1 年間分の新聞記事から 3 文字漢字の単語のうち、頻度が 1 であるものを収集し、それらからランダムに 100 種類を選んだ。同時に、選ばれた単語を含む文も取り出しておいた。

次に上記の 100 単語を「茶筌」の辞書から取り除いて、「茶筌」を再構築した。再構築された「茶筌」では、これらの 100 単語は未知語となる。この再構築された「茶筌」と本システムを用いて、取り出しておいた文から、上記 100 個の未知語が抽出できるかどうかを調べた。パターン別に分けた結果を表 3 に示す。ただし、表中パターンの欄で「その他」 は、 3 章で述べた仮定に反しているものを示している。

パターン 1 は比較的良い値だが、パターン 2 と 3 に問題があることがわかる。

4.3 低頻度の未知語の抽出

本手法の長所として低頻度の未知語を抽出できる点がある。そこで、どの程度、低頻度の未知語を抽出できるか検証するため、 4.1 章 で抽出できた未知語の トレーニングコーパス中での頻度を調べた。その結果を表 4 に示す。

表 4: 未知語のコーパス中での頻度

硬材：0、彩挺：0、深銘：0
仕梅：0、祥造：1、女衆：2
杉乃井：2、老子：3、普選：5
和魂：7、和慶：14、星島：15
作陶：17、画期：245

このことから、本手法でコーパス中の出現頻度が低い未知語が抽出可能であることが確認できた。

表 3: 未知語の作成による未知語の抽出

パターン	正解の個数	抽出した個数	正しく抽出できた個数	正解率	再現率	F 値
パターン 1	21	14	11	0.786	0.523	0.628
パターン 2	30	12	7	0.583	0.233	0.333
パターン 3	45	12	10	0.833	0.222	0.351
その他	13	—	—	—	—	—

5 考察

5.1 3 文字の漢字からなる未知語

本稿では 2 文字あるいは 3 文字の漢字からなる未知語を抽出の対象とした。パターン 0 とパターン 1 については比較的良い結果が得られたが、パターン 2 とパターン 3 については良い値を得られなかった。これらの精度が悪かった理由として、パターン 2 と 3 で使われている、2 文字の漢字列が単語になる確率の扱いや、尺度の決め方に問題があると思われる。本手法では 2 文字 の漢字列 ab が単語になる確率を $P(ab) = \frac{c(|ab|)}{c(ab)}$ として与えているが、$\frac{c(|ab|)}{c(ab)} \simeq 1$ となっている場合が多い。例えば実験 4.2 の結果では

/放浪/癖/ : P(放浪) = 70/71 = 0.986
/下/山田/ : P(山田) = 871/957 = 0.910
/計算/尺/ : P(計算) = 1082/1114 = 0.971
/大/番組/ : P(番組) = 2669/2681 = 0.996

などがこれにあたる。

これは、ab が出現しているとき、それらがほとんど 1 単語として出現していることを示している。これを $|ab|c|$ の例で尺度に適用した場合、$P(ab)$ が 1 に近い値なので

$$m(abc) \simeq 1 - P(c)$$

となる。これは c が 1 単語にならない確率のことであり、ab のことは考慮せずに未知語かどうかを判断している。

しかし、上の例を見ればわかるように、本手法で未知語と判定しなかったパターン 2 とパターン 3 の漢字列は複合語と考えることができるものが多い。つまり、パターン 2 とパターン 3 において、本手法の未知語ではないという判定は妥当だと思われる。

5.2 関連研究

テキストから未知語を抽出しようとする研究は、文字列の処理から語彙を自動獲得することで行える。この方向の研究の最初のアイデアは Nagao によって示された。そこでは suffix array 法を用いて大規模コーパスから n-gram を作成し、前後の文字のばらつき具合いから 1 単語と見なせる文字列を抽出している [3]。また Fung らは英語の連語抽出の研究を応用することで、統計的に有意な n-gram を取り出し、中国語の語彙を獲得している [1]。これらの研究は、目的が語彙獲得であり未知語抽出とは異なるが、未知語抽出システムと見る場合、対象としている文字列がトレーニングコーパス中にある頻度以上出現しなくてはならないという欠点を持っている。また Mori は、トレーニングコーパスからある品詞の前後に位置する文字列の分布を求め、任意の文字列の単語らしさとその品詞を推定する手法を提案した [2]。ただしこの手法でも対象としている文字列がトレーニングコーパス中にある頻度以上出現する必要がある。

永田は未知語の検出を未知語を扱える形態素解析の一部として実現した [6]。そこでは、まずある文字列が未知語となる確率を求める。次に未知語を含めて単語分割候補を求め、その候補を利用することである文字列が単語としてあらわれる頻度の期待値を定義した。この期待値が閾値以上のものを単語とみなすことで未知語を抽出できる。ただしこの研究では文字列が未知語となる確率を単語長確率と単語表記確率の積で求めている。さらに単語表記確率は単語内文字 bigram モデルから求めているので、本研究で対象としたような文字列長 2 の低頻度の未知語を抽出することはできない。

形態素解析結果の後修正という観点からも、未知語を検出できる。久光は誤り主導型教師付き学習により後修正のための書換え規則を学習し、その規則を適用することで形態素解析結果を修正している [7]。ただし未知語検出の観点からの評価は述べられていない。また Shinnou は文字ベースの HMM を利用して、複合語の形態素解析による単語分割誤りを修正した [4]。これによって未知語の抽出も可能だが、HMM における状態間を移る際のシンボル出力確率の算出に、トレーニングコーパス内での文字の頻度を利用している。このため低頻度の文字を持つ未知語に関しては、その検出の保証ができない。また未知語部分は形態素解析では誤った分割となるという点から、形態素解析結果を修正することで未知語を認識できる。内山は形態素解析により α と β に分割された単語に対して、形態素 α と β の区切れにくさの尺度を導入することにより α と β の過分割を解消する [5]。この研究は本研究の基本となるアイデアを提供しているが、形態素 α と β の区切れにくさの尺度に文字列 $\alpha\beta$ の頻度を利用している点が本手法と大きく異なる。本手法では α と β の区切れにくさの尺度に文字列 $\alpha\beta$ の頻度を必要としていない。このため、ある未知語に対応する文字列がトレーニングコーパス中で低頻度であっても、その文字列を未知語として抽出可能である。

6 おわりに

本稿では 2 文字あるいは 3 文字の漢字から構成される未知語をテキストから自動抽出する手法を提案した。

まずある文字列が 1 単語となる確率を導入した。次に、未知語が形態素解析によって過分割されることと、過分割された個々の単語に対応する文字列が単語となる確率が低いというヒューリスティクスを用いることにより、ある漢字文字列が未知語となる尺度を設定した。この尺度によって未知語を抽出できる。

実験の結果、2 文字の漢字からなる未知語において、F 値 0.684 を達成できた。また 3 文字の漢字からなる未知語において、パターン 1 では F 値 0.628 が得られた。残りのパターン 2、3 では F 値が 約 0.35 となった。数値的にはパターン 2、3 の未知語に対して F 値が低いが、パターン 2、3 で本手法が未知語として抽出しなかった単語の多くは複合語との区別が微妙である。これらのことから本手法は未知語抽出に有効であったといえる。

本手法は頻度の低い未知語を抽出できるという長所がある。また問題点として 3 文字列に対して精度が悪いという点が挙げられる。3 文字列の未知語抽出の精度向上を今後の課題とする。

謝辞

本研究で利用したコーパスおよび評価文は、毎日新聞 CD-ROM '94 および '95 版から得ています。利用を許可していただいた毎日新聞社に深く感謝します。

参考文献

[1] Fung, P. and Wu, D. : "Statistical Augmentation of a Chinese Machine-Readable Dictionary", Proceedings of the Second Workshop on Very Large Corpora, pp.69–85 (1994).

[2] Mori, S. and Nagao, M. : "Word extration from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis", Proceedings of the 16th International Conference on Computational Linguistics, pp.1119–1122 (1996).

[3] Nagao, M. and Mori, S. : "A New Method of N-gram Statistics for Large Number of n and Automatic Extration of Words and Phrases from Large Text Data of Japanese", Proceedings of the 15th International Conference on Computational Linguistics, pp.611–615 (1994).

[4] Shinnou, H. and Ikeya, M. : "Correction of Word Segmentation Errors through Character-based HMM", PACLING-99, pp.131–136 (1999).

[5] 内山 将夫 : "形態素解析結果から過分割を検出統計的尺度", 言語処理学会, Vol.6, No7, pp.3–28 (1999).

[6] 永田 昌明 : "確率モデルによる日本語処理に関する研究", 京都大学博士論文 (1999).

[7] 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", 情報処理学会自然言語処理研究会, 98-NL-126 (1998).