未知語の確率モデルと単語の出現頻度の期待値に基づくテキストからの語彙獲得

永田昌明†

本論文では，未知語の確率モデルと単語の出現頻度の期待値に基づいて日本語テキストから未知語を収集する方法を提案する．本手法の特徴は，単語を構成する文字の種類ごとに異なる未知語モデルを使用することによりひらがな語や複数の字種から構成される単語を収集できること，および，単語の出現頻度の期待値を文字列の単語らしさの尺度とすることにより出現頻度が低い単語を収集できることである．人手により単語分割されたEDRコーパスから無作為に選択した10万文（246万語）を用いて語彙数11,521の統計的言語モデルを学習し，EDRコーパスの残りの部分から無作為に選択した10万文（247万語，未知語率7.72%）をプレーンテキストと見なして語彙獲得実験を行ったところ，本手法による語彙獲得の精度は再現率61.5%適合率67.2%であった．

Lexical Acquisition from Japanese Text Based on Statistical Unknown Word Model and Expected Word Frequency

MASAAKI NAGATA†

We present a novel lexical acquisition method from Japanese texts based on a probabilistic model for unknown words and expected word frequency. The benefit of the proposed method is that it can collect *hiragana* words and words which consist of more than one character types by using a different unknown word model for the character type configuration of a word, and that it can collect low frequency words by using the expected word frequency as the likelihood measure of a word hypothesis. We trained a statistical language model with 11,521 vocabulary from 100 thousand manually word segmented sentences (2.46 million words) which were randomly selected from the EDR corpus, and extracted new words from another 100 thousand unsegmented sentences (2.47 million words) which were randomly selected from the rest of the EDR corpus, and whose out-of-vocabulary rate was 2.1%. The lexical acquisition accuracy of the proposed method was 61.5% recall and 67.2% precision.

1. はじめに

コンピュータによる日本語の単語分割は，英語のヒアリングに似たところがある．「makeやgetなど基本単語3,000語の用法をマスターすれば日常英会話は大丈夫!」というような甘言は真っ赤な嘘であることを多くの人は経験的に知っているに違いない．確かに出現頻度順で上位3,000語を覚えていれば，日常会話に出現する異なり単語の大部分はカバーできるだろう．しかし，たった1つの単語を知らないために文の意味が皆目分からなくなることは多い．知らない単語が1つあると文のセグメンテーションに失敗し，前後の単語も聞きとれなくなるからである．

近年，日本語や中国語のような単語を分かち書きしない言語について統計的言語モデルを用いた単語分割法の研究が進み，広範囲のテキストに対して95%程度の高い精度を持つ単語分割プログラムを実現できるようになった[11),18)]．これと同時に，残された単語分割誤りの多くは未知語（辞書未登録語）に起因するため，未知語処理が単語分割の重要課題の1つという認識が広まった．

英語は単語を分かち書きするので未知語は容易に同定できる．そのため英語の形態素解析における未知語処理の中心課題は品詞の推定であり，語頭が大文字であるか否かや接頭辞・接尾辞の情報から未知語の品詞を推定する確率モデルや規則推論法が提案されている[8),19)]．

これに対して日本語や中国語は単語を分かち書きしないので，そもそも入力文中の未知語を同定することが非常に難しい．日本語は単語の種類に応じて字種を使い分ける習慣があるので，従来は「ひらがなから漢

† NTTサイバースペース研究所
NTT Cyber Space Laboratories

字に変化する部分は単語境界である可能性が高い」といった字種に関する発見的規則を未知語の同定に利用する方法[20]が一般的だった．しかし，字種の変化を利用する方法は，ひらがな語や複数の字種から構成される単語に関する精度が低いという欠点がある．

そこで近年では，大規模なテキストコーパスが利用可能になったことから，単語分割における未知語問題の解決策の1つとして，対象分野のテキストに出現する未知語を自動的に収集して単語辞書のカバー率を上げる方法，すなわち，語彙獲得法[3)~5),9),10),12)]や，単語辞書を使わない単語分割法[13),15),17)]がさかんに研究されている．これらの研究は，文字列に関する統計量を使用する方法と単語の確率モデルを使用する方法に大別できる．

前者の方法は，文字 ngram 頻度や相互情報量などテキストから得られる統計量だけを使用して，統計的に有意な水準で（偶然よりも明らかに高い頻度で）固定的に共起する文字 ngram を単語として収集する[4),10),13),15),17)]．この方法は任意のテキストに対して適用できるという利点があるが，出現頻度が低い単語は収集できないという欠点がある．

これに対して後者の方法は，単語を構成する文字列および単語が出現する文脈に関する確率モデルを使用して，文字列の単語らしさを判定する[3),5),9),12)]．この方法は「単語」という概念をモデルの中に持っているので一般に前者の方法より精度は高い．しかし，モデルの学習に単語辞書や単語分割済みコーパスを必要とし，これらは必ずしも容易に入手できないという問題点がある．

本論文では，未知語の確率モデルと単語の出現頻度の期待値に基づく日本語テキストからの語彙獲得法を提案する．本手法は，日本語の正書法に基づいて単語を構成する文字の種類ごとに異なる未知語モデルを使用することにより，ひらがな語や複数の字種から構成される単語を収集できる．また，単語分割に多義が存在する場合の単語の出現頻度の自然な拡張として単語の出現頻度の期待値を定義し，これを文字列の単語らしさの尺度として用いることにより，出現頻度が低い単語を収集できる．

以下では，まず日本語の単語分割を確率論の枠組みで定式化し，未知語の確率モデルについて説明する．次に単語の出現頻度の期待値の計算法を説明する．そして語彙獲得の評価方法と実験結果について報告し，最後に考察と今後の課題を述べる．

2. 未知語の確率モデル

2.1 日本語の単語分割の数学的定義

文字列 $C = c_1 \ldots c_m$ から構成される入力文が単語列 $W = w_1 \ldots w_n$ に分割されるとする☆．数学的には，日本語の単語分割は与えられた文字列 C に対して単語列の条件付き確率 $P(W|C)$ を最大化する単語列 $\hat{W}$ を求める問題と定義できる．ここで文字列 C はすべての単語分割に共通なので $P(W)$ を最大化する単語列を求めればよい．

$$\hat{W} = \arg\max_W P(W|C) = \arg\max_W P(W) \quad (1)$$

本論文では，日本語の単語分割において単語列の同時確率 $P(W)$ を計算するための統計的言語モデルを単語分割モデル（word segmentation model）と呼ぶ．また，任意のテキストに対して単語列の同時確率 $P(W)$ を求めるためには，未知語に確率を割り当てる必要がある．本論文では，未知語に割り当てる確率を求めるための統計的言語モデルを単語モデル（word model）と呼ぶ．

以下では，まず基本的な単語分割モデルを説明し，次に単語モデルを説明する．そして単語モデルを組み込んだ単語分割モデルについて説明する．

2.2 単語分割モデル

単語分割モデルには，単語 ngram モデル（マルコフモデル）や品詞 ngram モデル（隠れマルコフモデル）など，音声認識や品詞タグ付けに使われる統計的言語モデル[6),7)]と同じものを使用できる．一般に，単語 ngram モデルや品詞 ngram モデルは高次のモデルほど言語モデルとしての性能が高いが，より多くの訓練テキストを必要とする．また同じ次数のモデルならば単語 ngram モデルの方が品詞 ngram モデルよりも性能が高い[12)]．

日本語の単語分割モデルの学習には人手により単語分割されたテキストが必要であり，現状では，人手により単語分割されたテキストは，せいぜい数百万語程度しか利用可能ではない．そこで，本論文では単語分割モデルとして次式に示す単語 bigram モデルを使用した☆☆．

☆ 本論文では，単語は表記・読み・品詞の3つ組から構成されると考える．2つの単語はそれぞれの表記・読み・品詞がすべて一致するときに限り等しい．したがって，同形語（表記が同じ）や同音語（読みが同じ）は別々の単語と見なす．

☆☆ 単語を分かち書きする英語では訓練テキストを容易に入手できるので，英語の音声認識では数億語のテキストから単語 trigram モデルを作成することが多い．一方，大規模な単語分割済みコーパスが存在しない中国語の単語分割の研究では，単語 unigram モデルを用いるのが一般的である[18)]．

$$P(W) \approx P(w_1|\texttt{<bos>}) \prod_{i=2}^{n} P(w_i|w_{i-1}) P(\texttt{<eos>}|w_n) \quad (2)$$

ここで <bos> および <eos> は文頭および文末を表す特殊記号である．

2.3 単語モデル

本論文では，単語を構成する文字の種類に応じて複数の単語モデルを使用する．以下では説明を簡単にするために，まず基本的な単語モデルを説明し，次に複数の単語モデルが必要な理由，および，各単語モデルの詳細な説明を行う．

ある単語 w_i が未知語であるとき，その表記が長さ k の文字列 $c_1 \ldots c_k$ である確率 $P(c_1 \ldots c_k|\texttt{<UNK>})$ を計算するためのモデルを単語モデルと定義する．ここで <UNK> は未知語を表す記号である．

$$P(w_i|\texttt{<UNK>}) = P(c_1 \ldots c_k|\texttt{<UNK>}) \quad (3)$$

単語モデルは，一般性を失うことなく，未知語の文字長の分布を表す単語長確率と，ある文字長の未知語の表記の出現確率を表す単語表記確率の積に分割できる．

$$P(c_1 \ldots c_k|\texttt{<UNK>}) = P(k|\texttt{<UNK>}) P(c_1 \ldots c_k|k, \texttt{<UNK>}) \quad (4)$$

Brown ら[2]は，単語長確率 $P(k|\texttt{<UNK>})$ を平均単語長 λ をパラメータとするポワソン分布で近似し，単語表記確率 $P(c_1 \ldots c_k|k, \texttt{<UNK>})$ を文字 zerogram 確率の積で近似した．文字 zerogram とは，すべての文字が等確率で出現すると仮定するモデルである．

$$P(c_1 \ldots c_k|\texttt{<UNK>}) \approx \frac{\lambda^k}{k!} e^{-\lambda} p^k \quad (5)$$

ここで $1/p$ は文字集合の大きさである．日本語の文字集合として JIS-X-0208 を仮定すると $p = 1/6879$ となる．

ポワソン分布は，下に有界な最も簡単な（パラメータが 1 つの）確率分布という理由で，単語長分布の第 1 次近似に用いられている．しかし，Brown のモデルには，長さ 0 の単語に一定の確率を割り当てるという問題，および，文字の出現分布を考慮しないという問題がある．そこで本論文では，単語長確率 $P(k|\texttt{<UNK>})$ は，平均単語長 λ をパラメータとするポワソン分布に従うと仮定するが，下界を 0 から 1 に移動する．

$$P(k|\texttt{<UNK>}) \approx \frac{(\lambda-1)^{k-1}}{(k-1)!} e^{-(\lambda-1)} \quad (6)$$

また，単語表記確率 $P(c_1 \ldots c_k|k, \texttt{<UNK>})$ は，単語内文字 bigram モデルから求めた文字列 $c_1 \ldots c_k$ の出現確率 $P_b(c_1 \ldots c_k)$ と，単語内文字 bigram モデルにおいて長さ k の文字列が出現する確率 $\sum_{|x|=k} P_b(x)$ の比で近似する．

$$P(c_1 \ldots c_k|k, \texttt{<UNK>}) \approx \frac{P_b(c_1 \ldots c_k)}{\sum_{|x|=k} P_b(x)} \quad (7)$$

$$P_b(c_1 \ldots c_k) = P(c_1|\texttt{<bow>}) \prod_{i=2}^{k} P(c_i|c_{i-1}) P(\texttt{<eow>}|c_k) \quad (8)$$

ここで <bow> および <eow> は語頭および語末を表す特殊記号である．

単語内文字 bigram は単語分割されたコーパスから求める．単語内文字 bigram 確率は，単語の先頭（接頭辞）・中間・末尾（接尾辞）に現れる文字 bigram では大きく，単語境界をはさむ文字 bigram では小さくなる．したがって，$P_b(c_1 \ldots c_k)$ は単語を構成する文字列では大きく，そうでない文字列では小さくなり，文字の出現分布のよい近似を与える．

しかし，$P_b(c_1 \ldots c_k)$ がすべての長さの文字列の中で $c_1 \ldots c_k$ が出現する確率であるのに対して $P(c_1 \ldots c_k|k, \texttt{<UNK>})$ は長さ k の文字列の中で $c_1 \ldots c_k$ が出現する確率である．したがって，もし前者で後者を近似すると長い文字列の確率が不適切に小さくなる．これを補正するために，式 (7) のように，長さ k の文字列の出現確率の和 $\sum_{|x|=k} P_b(x)$ で $P_b(c_1 \ldots c_k)$ を割る．

ところが，文字 bigram モデルにより長さ k の文字列が生成される確率 $\sum_{|x|=k} P_b(x)$ を正確に求めるのは難しい．そこでこれを，文字 unigram モデルにより長さ k の文字列が生成される確率，すなわち，語末記号 <eow> 以外の文字が $k-1$ 個続いた後に語末記号が出現する事象の確率で近似する．

$$\sum_{|x|=k} P_b(x) \approx (1 - P(\texttt{<eow>}))^{k-1} P(\texttt{<eow>}) \quad (9)$$

ここで $P(\texttt{<eow>})$ は語末記号が出現する確率である．

2.4 単語を構成する文字の種類による単語モデル

図 1 に後述する EDR コーパスにおける出現頻度 1 の単語の単語長分布，および，出現頻度 1 の単語の平均単語長（$\lambda = 4.8$）から式 (6) のポワソン分布により推定した単語長分布を示す☆．両者は比較的よく一致しているが，2〜4 文字の単語ではポワソン分布による推定値の方が小さく，5〜7 文字の単語では推定値の方が大きい．この推定誤差は，すべての未知語を 1 つの単語モデルで表現することに原因がある．

☆ 一般に，コーパスに 1 度しか出現しない単語（hapax legomenon）の性質は，未知語の性質に近いといわれている[1]．

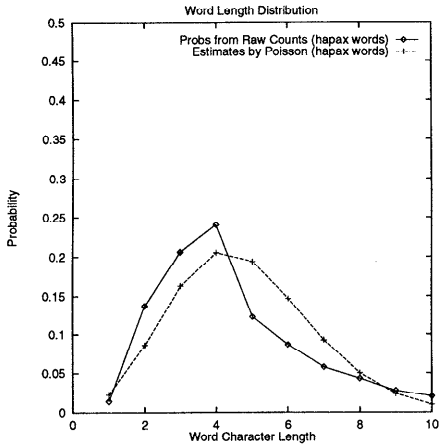


図1 未知語の単語長の分布とポワソン分布による推定値
Fig. 1 Word length distribution of unknown words and its estimate by Poisson distribution.

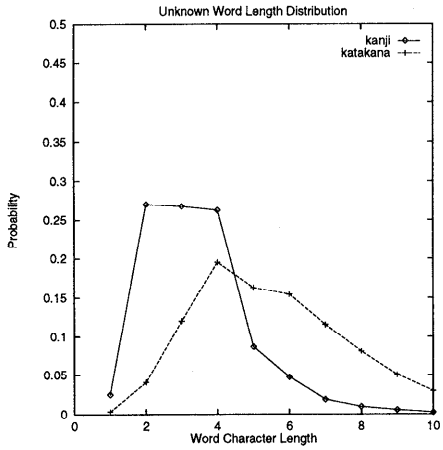


図2 漢字語とカタカナ語の単語長の分布
Fig. 2 Word length distribution of *kanji* words and *katakana* words.

出現頻度1の単語のうち，漢字列から構成される単語の単語長分布とカタカナ列から構成される単語の単語長分布を**図2**に示す．図2から分かるように，漢字語は2〜4文字の単語が多く，カタカナ語は4〜6文字の単語が多い．図1の単語長分布は，この両者の単語長分布を重ね合わせたものとほぼ一致する．

現代の日本語の正書法では，句読点などの記号以外に少なくとも5つの文字の種類（漢字，ひらがな，カタカナ，アルファベット，アラビア数字）がある．漢字は中国系の外来語（漢語），および，中国語と意味的に等しい日本語の表記に（送り仮名をともなって）使用される．ひらがなは助詞や活用語尾などの機能語の表記に使用され，カタカナは西欧系の外来語の発音表記に使用される．アルファベットは西欧系の単語や頭字語の表記に使用され，アラビア数字は数の表記に使用される．

表1 未知語を構成する文字の種類
Table 1 Character type configuration of unknown words.

語を構成する文字種	割合	例
漢字	45.1%	温泉街
カタカナ	11.4%	エスプレッソ
カタカナ–漢字	6.5%	ベル研究所
漢字–ひらがな	5.6%	玉ねぎ，極ま
ひらがな	3.7%	なれそめ
漢字–カタカナ	3.4%	交通ルール
カタカナ–記号–カタカナ	3.0%	ビル・ゲーツ
数字	2.6%	００７
漢字–ひらがな–漢字	2.4%	飲み会，競り合
アルファベット	2.0%	ＶＳＯＰ
漢字–ひらがな–漢字–ひらがな	1.7%	思い違い
ひらがな–漢字	1.3%	えい児，おけ屋

表1にEDRコーパスにおける出現頻度1の単語について，単語を構成する文字の種類を調べたものを示す．表1によれば，1つの字種で構成される単語（漢字，カタカナ，ひらがな，数字，アルファベット）が全体の約65%を占めている．2つ以上の字種で構成される単語のうち，漢字–ひらがな，または，ひらがな–漢字という2つのパターンだけが，形態素，すなわち，これ以上分割すると意味を持たなくなる単語と見なせる．前者は「極ま（る）」のような漢字と送り仮名の組合せ☆に対応し，後者は「えい（嬰）児」のような難しい漢字をひらがなで表記した単語に対応する．その他の字種の組合せから構成される単語，たとえば，「交通ルール」（漢字–カタカナ）や，「思い違い」（漢字–ひらなが–漢字–ひらがな）は，EDRコーパスでは1つの単語として認定されているが，厳密にいえば，形態素ではなく複合語である．

そこで本論文では，標準的な日本語の表記法において形態素を構成する文字の種類のパターンを網羅するように，単語を構成する文字の種類によって日本語の未知語を9種類の未知語タイプに分類する．**図3**に未知語タイプをバッカス記法（Backus Naur Form, BNF）で定義したものを示す☆☆．

<sym>，<num>，<alpha>，<hira>，<kata>，<kan>は，それぞれ記号列，数字列，アルファベット列，ひらがな列，カタカナ列，漢字列という1つ

☆ 本論文で使用した単語辞書にはすでに活用語尾が登録されているので，「極まる」「競り合う」のような活用語では，「極ま」「競り合」などの語幹が未知語収集の対象となる．

☆☆ [···] は文字集合中の任意の1文字と照合することを表す．2つの文字の間に-を書くことで文字範囲を表す．文字コードにはJIS-X-0208を仮定している．* は0回以上の繰り返し，+は1回以上の繰り返しを表す．

```
<sym>::=[、-〇]+
<num>::=[0-9〇一二三四五六七八九十百千万億兆]
        [,. 0-9〇一二三四五六七八九十百千万億兆]*
<alpha>::=[A-z A-ω A-я]+
<hira>::=[あ-ん]+
<kata>::=[ア-ヶ][ア-ヶ ー]*
<kan>::=[亜-瑤][ヽ-〇亜-瑤]*
<kan-hira>::=<kan><hira>
<hira-kan>::=<hira><kan>
<misc>::=上記にいずれにもあてはまらない文字列
```

図3 未知語タイプの定義
Fig. 3 Definition of unknown word type.

の字種から構成される文字列を表す☆. <kan-hira>, <hira-kan> は，それぞれ漢字列–ひらがな列，ひらがな列–漢字列という2つの字種から構成される文字列を表す．そして，これら以外の複数の字種から構成される文字列はすべて <misc> とする．

そして単語モデルでは，式(6)の代わりに次式に示す未知語タイプ別の単語長確率を使用する．

$$P(k|\text{<T>}) \approx \frac{(\lambda_{\text{<T>}}-1)^{k-1}}{(k-1)!} e^{-(\lambda_{\text{<T>}}-1)} \quad (10)$$

ここで <T> は未知語タイプ，$\lambda_{\text{<T>}}$ は未知語タイプ <T> の平均単語長を表す．なお単語表記確率は，未知語タイプ別に文字 bigram 確率を求めるとデータ不足の問題が生じるので，未知語タイプに関係なく式(7)から求める．

2.5 未知語を考慮した単語分割モデル

次に単語モデルを使って，未知語を考慮した単語分割モデルを定義する．未知語の表記が未知語タイプのみに依存し，直前の単語には依存しないと仮定すれば☆☆，未知語 w_i を含む単語 bigram 確率は，w_i を構成する文字列から決まる未知語タイプ <T> を含む単語 bigram 確率と単語モデルが w_i に与える確率の積で近似できる．

$$\begin{aligned} P(w_i|w_{i-1}) &= P(\text{<T>}|w_{i-1})P(w_i|\text{<T>}, w_{i-1}) \\ &\approx P(\text{<T>}|w_{i-1})P(w_i|\text{<T>}) \end{aligned} \quad (11)$$

未知語タイプ <T> を含む単語 bigram 確率 $P(\text{<T>}|w_{i-1})$ は，すべての未知語を対応する未知語タイプ <T> に置き換えた訓練コーパスから推定する．

未知語タイプを含む単語 bigram の例を**表2**に示す．未知語タイプを他の語彙と同等に扱うことにより，特定の未知語タイプが「が」「を」「に」「の」などの助詞の直前に出現する確率や，「い（形容詞）」「る（一段動詞）」「し（サ変動詞）」などの活用語尾の直前に出現する確率が単語 bigram モデルの中に表現される．特に，ひらがな列 <hira> とひらがな表記される助詞や活用語尾との単語 bigram 頻度，および，ひらがな列の未知語が2つ連続する頻度の情報は，ひらがなで表記された未知語を同定する際に非常に重要な役割を果たす．

表2 未知語タイプを含む単語 bigram の例
Table 2 Example of word bigram including unknown word types.

単語 bigram		頻度
<alpha>	の/ノ/助詞	144
<hira-kan>	に/ニ/助詞	60
<hira>	<hira>	7
<hira>	い/イ/語尾	89
<kan-hira>	る/ル/語尾	136
<kan>	<kan>	364
<kan>	が/ガ/助詞	1455
<kan>	さん/サン/接尾語	447
<kan>	し/シ/語尾	624
<kata>	を/ヲ/助詞	442
は/ハ/助詞	<hira>	148
約/ヤク/接頭語	<num>	429

3. 単語出現頻度の期待値

3.1 単語列の N-best 探索

一般に上位 N 個の最適解を求めることを N-best 探索という．本論文では，式(1)の解，すなわち，同時確率 $P(W)$ を最大化する単語列 $\hat{W}$ の N-best 探索に，前向き DP 後向き A^* アルゴリズム[11]を用いる．このアルゴリズムは，文頭から1文字ずつ文末へ進む動的計画法を用いて文頭から任意の単語までの部分解析の確率を求める前向き探索と，文末から文頭方向へ進む A^* アルゴリズムを用いて確率が大きい順に1つずつ任意個の形態素解析候補を求める後向き探索から構成される．

図4に「ペンシルバニア大学は ENIAC の 50 周年を祝う。」という文の文字位置4における前向き探索の様子を示す．前向き探索では，この文字位置で終わるすべての部分解析と，この文字位置から始まるすべての単語候補の組合せの確率を計算し，新しい部分解析を作成する．これを文頭から順番に各文字位置で行えば，文頭から任意の単語までの部分解析の確率が求まる．

図5に同じ例文の後向き探索の様子を示す．前向き探索により文頭から任意の単語までの最適経路が分かっているので，後向き探索では，文末から文頭へ進

☆ 便宜上，数字列には漢数字と小数点のピリオドと位取りのカンマを含め，アルファベットにはギリシャ文字とロシア文字を含め，カタカナには長音記号を含め，漢字には「ヽ」「々」などの記号を含めた．ただし，位取りのカンマや長音記号が語頭の文字にはならないような配慮はしている．

☆☆ これは，シンボルの出現確率は現在の状態のみに依存するという，隠れマルコフモデルの出力独立の仮定と同じである．

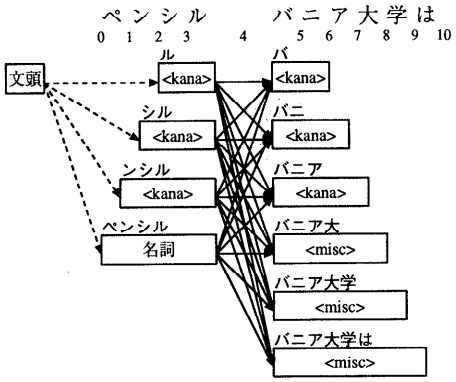


図4 前向き探索の例
Fig. 4 An example of the forward search.

む A^* アルゴリズムを用いて確率が大きい順に任意の数の形態素解析候補を求めることができる．図5では,「ペンシルバニア」と「ENIAC」が未知語であり,「ペンシルバニア」の周辺の単語分割に関して3通りの多義が提示されている．

3.2 字句解析

未知語を同定するため，前向き探索において，入力文中のすべての部分文字列について，もしその文字列が辞書に登録されていなければ，文字列を構成する文字の種類に応じた未知語タイプ <T> を持つ単語候補を生成する．たとえば，図4において,「ペンシル」は辞書登録語であり,「ンシル」(<kana>)「バニア」(<kana>)「バニア大」(<misc>) などは未知語候補である．

前向き探索の計算量は，ある文字位置で生成される単語候補の数の2乗に比例するので，その文字位置から始まるすべての部分文字列を未知語候補とすると，計算量が非常に大きくなる．そこで簡単な字句解析により以下のような制約を加え，未知語候補の数を削減する．

- 括弧や句読点などの自然な単語境界を含む未知語候補は生成しない．
- 数字列とアルファベット列は最も長い未知語候補のみを生成する．
- 長さがある閾値を超える未知語候補は生成しない(数字列，アルファベット列，カタカナ列を除く)．
- 字種の変化の回数がある閾値を超える未知語候補は生成しない．

最後の制約は，図3と同じバッカス記法で表現すれば，以下に示す <token> と <misc> を受理するような字句解析により実現される．

```
<token>::=<sym>|<num>|<alpha>|<hira>|<kata>
          |<kan>|<kan-hira>|<hira-kan>
<misc>::=<token><token>
```

これによって複数の字種で構成される文字列のうち,「漢字–ひらがな」「ひらがな–漢字」に加えて，表1において頻度が大きかった「カタカナ–漢字」「漢字–カタカナ」「漢字–ひらがな–漢字」「漢字–ひらがな–漢字–ひらがな」などのパターンも未知語候補として生成される．

どのような文字列を未知語候補として生成すべきかは，日本語の単語をどう定義するかに依存する．たとえば，上記の字句解析規則では,「C言語コンパイラ」や「過マンガン酸カリウム」は未知語候補として生成されない．しかし，このような複合語を単語分割プログラムが1つの単語として同定すべきかは議論の余地がある☆．ここでは，単語分割の計算量と語彙獲得の再現率のトレードオフを考慮して，上記のような字句解析規則を使用することにした．

3.3 単語の出現頻度の期待値

前向きDP後向き A^* アルゴリズム[11]を用いれば，原理的には，入力文に対するすべての単語分割候補を求めることができる．テキストの第 i 番目の文の第 j 番目の単語分割候補を O_j^i，その確率を $P(O_j^i)$ とするとき，第 i 文における単語 w_α の出現頻度の期待値 $C^i(w_\alpha)$ を以下のように定義する．

$$C^i(w_\alpha) = \sum_j \left(\frac{P(O_j^i)}{\sum_k P(O_k^i)} \times n_j^i(w_\alpha) \right) \quad (12)$$

ここで $n_j^i(w_\alpha)$ は単語 w_α が第 i 文の第 j 候補に出現した回数を表す．

すなわち，ある文における単語 w_α の出現頻度の期待値は，単語 w_α が出現する単語分割候補の相対確率とその単語分割候補における単語 w_α の出現回数の積を，すべての単語分割候補について加算したものである．この定義は単語分割に多義がない場合の単語の出現頻度の定義の自然な拡張になっている．

現実的には，下位候補の確率は急激に小さくなるので，式(12)の分母にあるすべての単語分割候補の確率の総和を上位 N 候補の確率の和で近似する．

$$\sum_k P(O_k^i) \approx \sum_{k=1}^{N} P(O_k^i) \quad (13)$$

テキスト中の単語出現頻度の期待値 $C(w_\alpha)$ は，すべての文に関する単語出現頻度の期待値の総和として以下のように定義する．

☆ 単語分割プログラムの出力は形態素とし，複合語はXtract[16]のような共起表現(collocation)収集ツールで収集すべきだという考え方もある．

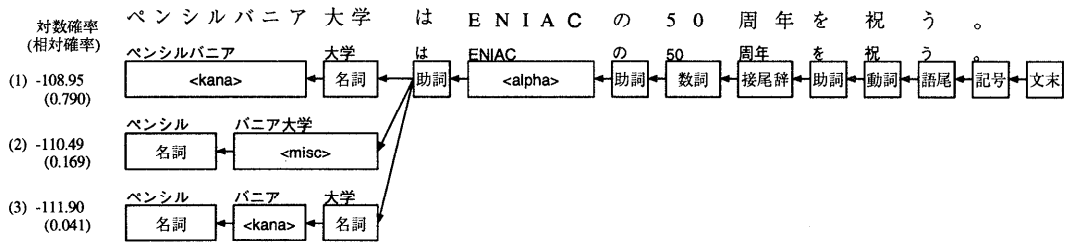


図5 後向き探索の例
Fig. 5 An example of the backward search.

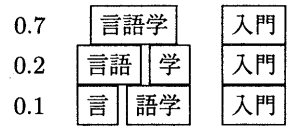


図6 単語頻度の期待値計算の例
Fig. 6 An example for expected word frequency calculation.

$$C(w_\alpha) = \sum_i C^i(w_\alpha) \tag{14}$$

3.4 テキストからの未知語の収集

人間が未知語を同定する場合，その単語候補の綴り（表記）が単語としてもっともらしいかどうか，および，その単語候補がその文脈の中で文法的に許容可能かどうかを判断基準にしていると思われる．これと同様に，本論文の言語モデルでは，未知語の確率モデルが表記のもっともらしさを与え，未知語タイプを含む単語 bigram モデルが文脈中でのもっともらしさを与える．したがって，ある単語候補の出現頻度の期待値は，表記と文脈の両方を考慮したうえで，その単語候補が未知語として同定される回数の期待値に相当し，この期待値が大きいほど単語候補が本当に未知語である可能性が高いと考えられる．

そこで本論文では，式 (14) に示すテキスト中の単語出現頻度の期待値を，テキスト中の任意の部分文字列の「単語らしさ」の尺度として用いる．そして単語出現頻度の期待値の閾値を θ とするとき，辞書に登録されていない単語候補 w_α について，その出現頻度の期待値が θ より大きければ，w_α を未知語として収集する．

$$C(w_\alpha) \geq \theta \tag{15}$$

たとえば，コーパスの第 i 文が文字列 “言語学入門” であり，その上位 3 個の単語分割候補が**図 6** であるとする．図 6 の左端は単語分割候補の相対確率であり，式 (12) の $P(O_j^i)/\sum_k P(O_k^i)$ に相当する．この文における各単語候補の出現頻度の期待値は以下のようになる．

$$C^i(\text{入門}) = 0.7 + 0.2 + 0.1 = 1.0$$
$$C^i(\text{言語学}) = 0.7$$
$$C^i(\text{言語}) = C^i(\text{学}) = 0.2$$
$$C^i(\text{言}) = C^i(\text{語学}) = 0.1$$

もし，どの単語候補も辞書に登録されておらず，閾値 θ が 0.15 ならば，「入門」「言語学」「言語」「学」を単語と認定し，「言」「語学」は無視する．

次に単語頻度の期待値のコーパスにおける総和を求める例を示す．まず「ペンシルバニア大学は ENIAC の 50 周年を祝う。」という文がコーパスに含まれ，その上位 3 個の単語分割候補を図 5 とすれば，「ペンシルバニア」「バニア大学」「バニア」の単語出現頻度の期待値はそれぞれ 0.790，0.169，0.041 となる．さらに「ホワイトハウスはペンシルバニア通りにある.」という文もコーパスに含まれ，「ペンシルバニア」「バニア通り」「バニア」の単語出現頻度の期待値がそれぞれ 0.825，0.127，0.048 とする．コーパス全体における単語出現頻度の期待値は以下のようになる．

$$C(\text{ペンシルバニア}) = 0.790 + 0.825 = 1.615$$
$$C(\text{バニア大学}) = 0.169$$
$$C(\text{バニア通り}) = 0.127$$
$$C(\text{バニア}) = 0.041 + 0.048 = 0.089$$

「ペンシルバニア」は出現頻度の期待値が増加し，未知語と同定される可能性が高くなる．一般に，ある単語候補がコーパス中により多く出現するほど，たとえ文の単語分割に曖昧性があったとしても，その単語候補は未知語として収集される可能性が高くなる．

4. 実　　験

4.1 言語データ

本論文では言語データとして「EDR 日本語コーパス Version 1.0」[14] を用いた．EDR コーパスは，新聞・雑誌・辞書・百科辞典・教科書などから収集され，形態論・統語論・意味論レベルの様々な注釈が人手で付与された約 500 万語（約 20 万文）のコーパスである．この実験では単語区切り・読み・品詞の情報を用いた．

表3 訓練テキストと試験テキストの量
Table 3 The amount of training and test texts.

	訓練テキスト	試験テキスト
文	100,000	100,000
単語（延べ）	2,460,188	2,465,441
単語（異なり）	85,966	85,967
文字（延べ）	3,897,718	3,906,260
文字（異なり）	3,298	3,278

表4 言語モデルに含まれる単語 ngram と文字 ngram の数
Table 4 The number of word and character ngrams in each language model.

	単語		文字	
言語モデル	1-gram	2-gram	1-gram	2-gram
LM-1-1	43,975	172,519	2,902	70,779
LM-10-1	11,531	136,994	2,902	70,779
LM-50-1	3,462	80,335	2,902	70,779

本実験では，人手により単語分割されたコーパスを2つに分割して，一方を言語モデルの学習に使用し，他方を語彙獲得のテストに使用する．この方法は，正解が分かっているので厳密な評価が可能という利点があるが，学習データとテストデータの性質が似ているという欠点がある．現実的な応用の場面では，未知語を収集する対象となるテキストと同じ分野の学習データは必ずしも容易に入手できない．そこで本実験では，言語モデルに含まれる語彙数を少なくすることにより，未知語率☆が高いという意味で，学習データと異なる分野のテキストから語彙を獲得する場合に近い条件を作り出す．

まず EDR コーパスから 10 万文ずつ無作為に文を選択して2つの部分集合を作成し，一方は統計的言語モデルの学習に使用し，他方はプレーンテキストとして語彙獲得のテストに使用する．**表3** に訓練テキストと試験テキストにおける文・単語・文字の数を示す．

語彙数の異なる言語モデルを作成するために，訓練テキストにおける単語の出現頻度（単語 unigram 頻度）に基づき，頻度のカットオフ値を 1，10，50 とする3種類の単語リストを作成した．たとえば，カットオフ値が 10 の単語リストには訓練テキストにおいて出現頻度 11 以上の単語のみが含まれる．そして，各単語リストに基づいて，単語リストにない単語を未知語タイプ記号に置き換えた訓練テキストから単語 bigram 頻度を求め，出現頻度 1 の単語 bigram を捨てて単語 bigram モデルを作成した．以下では，この3つの単語 bigram モデルを単語分割モデルとする言語モデルを，単語 unigram と単語 bigram のカットオフ値に基づいて LM-1-1，LM-10-1，LM-50-1 と呼ぶ．

単語モデルは，すべての言語モデルで同じものを使用した．単語内文字 bigram モデルでは，訓練テキスト中の異なり文字数 3,298 個のうち，頻度 2 以上の 2,900 個を文字リストに登録した．そして文字リストにない文字を未知文字記号に置き換えた訓練テキストから単語内文字 bigram を求めて単語内文字 bigram モデルを作成した☆☆．平均単語長は，訓練テキスト中の出現頻度 1 の単語から未知語タイプ別に求めた．たとえば，単語 unigram 頻度のカットオフ値が 1 の場合，漢字列の平均単語長 $\lambda_{<kan>} = 3.3$，カタカナ列の平均単語長 $\lambda_{<kata>} = 5.6$ であった．

表4 に各言語モデルに含まれる単語 ngram および文字 ngram の数を示す☆☆☆LM-1-1（語彙約 4 万語）は，大規模な辞書と豊富な用例を学習した言語モデルであるのに対して，LM-50-1（語彙約 3 千語）は，わずかな基本単語とその用例のみを学習した言語モデルに相当する．LM-10-1（語彙約 1 万語）は両者の中間的な性質を持つ言語モデルある．なおすべての ngram 確率は削除補間法により平滑化した[6]．

4.2 単語モデルのパープレキシティによる評価

まず，本論文で提案する単語モデルの優位性を示すために，ポワソン分布と文字 zerogram モデルを組み合わせた式 (5) の Brown の単語モデル（Poisson+zerogram），式 (6) のポワソン分布と式 (7) の文字 bigram モデルを組み合わせた単語モデル（Poisson+bigram），式 (10) の未知語タイプ別のポワソン分布と式 (7) の文字 bigram モデルを組み合わせた単語モデル（WT-Poisson+bigram），および，式 (7) の文字 bigram モデルだけを使用する単語モデル（bigram）の4種類について，試験テキスト中で LM-1-1 の単語辞書にない単語（56,121 語）に対する単語あたりのクロスエントロピーと文字あたりのテストセットパープレキシティを求めた結果を**表5** に示す．

一般に，クロスエントロピーやテストセットパープ

☆ 対象テキストの総単語数に対する辞書未登録語の割合．

☆☆ 日本語は文字集合が大きいので，訓練テキストに出現しない文字が試験テキストに出現する可能性が高い．そこで，本論文では，単語 ngram モデルに未知語記号を導入したのと同様に，文字 ngram モデルに未知文字記号を導入する．未知文字記号の出現確率は頻度 1 の文字の出現確率の総和から求め，各未知文字は等確率で出現すると考える．

☆☆☆ 単語 ngram には文頭記号と未知語タイプ記号（9個）と文末記号が含まれ，文字 ngram には語頭記号と未知文字記号と語末記号が含まれている．ただし，文頭記号や語頭記号は文脈を表す記号であって語彙の一部とは見なさない方が理論上都合が良い．そのため，たとえば，文字 unigram の数は，文字リストの大きさ（2,900）に特殊記号数（2）を加えた 2,902 となる．

表5 単語あたりのクロスエントロピーと文字あたりのテストセットパープレキシティ
Table 5 Cross entropy per word and test set perplexity per character.

単語モデル	cross entropy	perplexity
Poisson+zerogram	59.4	2,032
Poisson+bigram	34.4	82
WT-Poisson+bigram	34.0	78
bigram	34.6	85

レキシティが小さいほど，良い単語モデルであるといえる[6),7)]．特に，文字あたりのパープレキシティは，JIS-X-0208の文字集合の大きさ（6,879文字）と比較することにより，単語モデルが文字候補を絞り込む能力を直感的に把握するのに役立つ．

表5から，単語表記のモデルを文字zerogramモデルから文字bigramモデルに置き換えると文字パープレキシティが大幅に減少することが分かる（2,032 → 82）．したがって，Brownのモデルよりも本論文の単語モデルの方が明らかに優れている．また，文字bigramモデルだけの単語モデルよりも，長さの分布としてポワソン分布を導入する方がパープレキシティが減少し（85 → 82），未知語タイプ別のポワソン分布を使用することにより，さらにパープレキシティが減少する（82 → 78）．単純な文字bigramモデルを単語モデルとする場合に比べれば，本論文の単語モデルは長さの分布を考慮することによりパープレキシティが約8%小さくなる．

4.3 比較対象とする語彙獲得法

次に，未知語の確率モデルを用いる語彙獲得法の優位性を示すために，単語分割に関するヒューリスティクスの代表例である最長一致法および字種切り法を用いた以下のような語彙獲得法を実装した．

最長一致法にはいくつか種類があるが，ここでは次のような貪欲なアルゴリズムを用いた．文頭から始め，ある位置で辞書と一致する最長の単語を探し，その終わりの次の文字を開始点として同様の手続きを文末まで繰り返す．もしある位置で辞書と一致する単語がなかった場合は辞書と一致する単語がある文字まで進む．そして辞書と一致する単語がなかった文字列を1つの単語（未知語）と見なす．

字種切り法にもいくつか種類があるが，ここでは記号・数字・アルファベット・ひらがな・カタカナ・漢字の5つの字種を考え，すべての字種の変化点を単語境界と見なした．ただし，この単語分割により得られた単語集合の中で，ひらがな列は助詞・助動詞などの機能語列である可能性が高いので単語集合から取り除く．この単語集合と既知単語集合との差分が未知語集合である．

また，出現頻度の期待値を単語らしさの尺度とする語彙獲得法の優位性を示すために，統計的言語モデルによる単語分割の第1候補，最長一致法による単語分割，および，字種切り法による単語分割から求めた出現頻度を単語らしさの尺度とする語彙獲得法を実装した．

4.4 評価尺度

英語と違って日本語は単語を分かち書きしないので，正書法の中に単語という単位が存在しない．また日本語は膠着語なので，単語境界を一貫性を保ちながら決定するのは日本人にとっても難しい．このため従来は単語分割や語彙獲得の精度を客観的に評価する方法が存在しなかった．本論文では，人手により単語分割されたコーパスの単語境界を唯一の正解と見なし，この正解データに対する再現率・適合率・F尺度で精度を評価する．

単語分割では，正解データ中の単語数を Std，システム出力中の単語数を Sys，一致した単語数を M とすると，再現率は正解データの単語列の中でシステムが正しく同定した単語の割合 M/Std を表し，適合率はシステムが出力した単語列の中で正しく同定された単語の割合 M/Sys を表す．

語彙獲得では，正解データ中の未知語数を Std，システムが未知語と同定した単語数を Sys，両者が一致した数を M とすると，再現率は正解データの未知語の中でシステムが未知語と同定したものの割合 M/Std を表し，適合率はシステムが未知語と同定した単語の中で正解データの未知語と一致するものの割合 M/Sys を表す．

また，単語分割や語彙獲得の精度を1つの指標で表現する場合には，情報検索で用いられるF尺度を用いる．

$$F = \frac{(\beta^2 + 1.0) \times P \times R}{\beta^2 \times P + R} \tag{16}$$

ここで P は再現率，R は適合率，β は適合率に対する再現率の相対的重要度を表す．本実験では $\beta = 1.0$ とし，再現率と適合率の重みを等しくした．

4.5 単語分割と語彙獲得の精度

まず最初に，各言語モデルについて，試験テキストに対する単語分割の精度（再現率・適合率・F尺度），テストセットパープレキシティ（Test Set Perplexity），および，未知語率（Out-Of-Vocabulary rate）を**表6**に示す．LM-1-1，LM-10-1，LM-50-1の順に語彙数が少なくなるほど単語分割の精度が低下している．しかし，LM-1-1とLM-50-1を比較すると，未知語率が

表6 試験テキストに対する単語分割精度・テストセットパープレキシティ・未知語率
Table 6 Word segmentation accuracy, test set perplexity, and out-of-vocabulary rate.

言語モデル	再現率	適合率	F尺度	TSP	未知語率
LM-1-1	94.6%	93.2%	93.9	174	3.59%
LM-10-1	94.0%	92.2%	93.1	212	7.72%
LM-50-1	93.4%	90.8%	92.1	289	14.48%

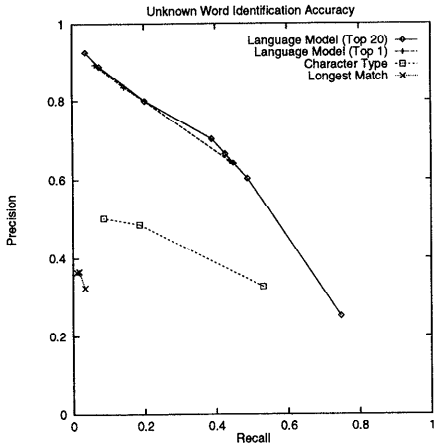


図7 言語モデル LM-1-1 の語彙獲得精度
Fig. 7 Word identification accuracy of LM-1-1.

10%以上増加しているのに対して，再現率および適合率は2%程度しか低下していない．これより統計的言語モデルによる単語分割法は，約15%の未知語を含むテキストでも頑健に動作することが分かる．

試験テキストに対して，3種類の言語モデル LM-1-1，LM-10-1，LM-50-1 を用い，単語出現頻度の期待値の閾値を $\theta = 0$，0.2，0.4，0.6，0.8，1.0，2.0，3.0 とした場合の語彙獲得の再現率と適合率をそれぞれ**図7**，**図8**，**図9** に示す☆．また比較のために，統計的言語モデル，最長一致法，および，字種切り法による単語分割から求めた出現頻度の閾値を $\theta = 1$，2，3 とした場合の語彙獲得の再現率と適合率を各図に示す．また図8に示した語彙獲得精度の具体的な数値を**表7**に示す．

一般に，図7，図8，図9では，再現率と適合率の組を表す点がグラフの右上にプロットされる（再現率と適合率がともに高い）ほど，語彙獲得の精度が高いことを表す．いずれの場合も統計的言語モデルを用いる語彙獲得法は最長一致法や字種切り法より明らかに精度が良い．たとえば，表7（および図8）に示す言

☆ ここでは単語出現頻度の期待値は上位20個の単語分割候補から求め，数字列・アルファベット列・カタカナ列以外の未知語候補の長さは最大8文字までに制限した．

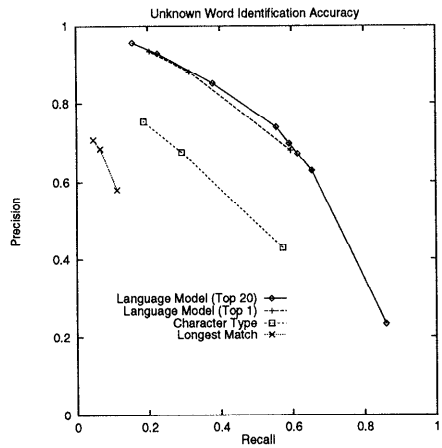


図8 言語モデル LM-10-1 の語彙獲得精度
Fig. 8 Word identification accuracy of LM-10-1.

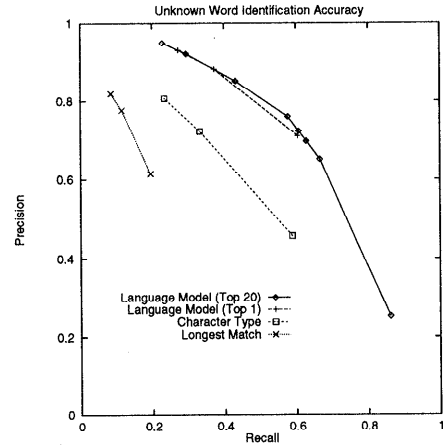


図9 言語モデル LM-50-1 の語彙獲得精度
Fig. 9 Word identification accuracy of LM-50-1.

語モデル LM-10-1 において閾値 $\theta = 1$ のときを比較すると，字種切り法は再現率57.3%適合率42.9%，最長一致法は再現率11.2%適合率58.1%であるのに対して，統計的言語モデルを用いる方法（第1候補のみ）は再現率59.6%適合率68.1%である．

次に単語らしさの尺度として出現頻度を用いる方法と出現頻度の期待値を用いる方法を比較する．まず，再現率と適合率の重みを同等として F 尺度で比較すると，前者の最適値は θ=1 で 63.6，後者の最適値は θ=0.4 で 64.1 であり，後者の方が少し優れている．しかし，出現頻度の期待値を用いる方法の最大の利点は，閾値 θ を変えることにより語彙獲得の再現率と適合率を連続的に自由に設定できることである．一般に出現頻度の期待値の閾値 θ を小さくすると，語彙獲得の再現率は高くなり，適合率は低くなる．LM-10-1 では閾値 θ を0とすれば最大86.0%の再現率を得られる．

表7 言語モデル LM-10-1 の語彙獲得精度
Table 7 Word identification accuracy of LM-10-1.

	言語モデル（上位20候補）			言語モデル（第1候補）			字種切り法			最長一致法		
閾値 θ	再現率	適合率	F尺度	再現率	適合率	F尺度	再現率	適合率	F尺度	再現率	適合率	F尺度
≥0.0	86.0%	23.6%	37.0									
≥0.2	65.5%	62.9%	64.1									
≥0.4	61.5%	67.2%	64.2									
≥0.6	59.6%	69.8%	64.1									
≥0.8	55.6%	74.0%	63.5									
≥1.0	38.1%	85.0%	52.6	59.6%	68.1%	63.6	57.3%	42.9%	49.1	11.2%	58.1%	18.8
≥2.0	22.7%	92.8%	36.5	31.5%	88.0%	46.4	29.3%	67.6%	40.9	6.6%	68.6%	12.0
≥3.0	15.7%	95.5%	27.0	20.5%	93.4%	33.6	18.7%	75.3%	29.9	4.7%	70.8%	8.8

表8 未知語タイプ別の未知語収集精度（LM-10-1, $\theta = 0.4$）
Table 8 Word identification accuracy with respect to unknown word types (LM-10-1, $\theta = 0.4$).

未知語タイプ	再現率	適合率	F尺度
<sym>	100.0	70.6	82.8
<num>	97.4	97.8	97.6
<alpha>	91.4	81.6	86.3
<hira>	53.6	58.0	55.7
<kata>	86.5	80.1	83.2
<kan>	68.1	65.7	66.9
<kan-hira>	50.9	69.8	58.9
<hira-kan>	25.2	79.6	38.3
<misc>	29.6	50.6	37.4

高い再現率を得られることは，実際の未知語収集作業では非常に有利である．現状の技術では語彙獲得の適合率が100%になることはありえないので，テキストから未知語を収集する作業では人手によるチェックが必要である．KWIC などを併用すれば，収集された単語候補が本当に単語であるかどうかを人間は素早く判断できるので，適合率が低いことはあまり問題ではない．これに対して，再現率が高ければより網羅的に単語を収集できる．さらに，収集した単語を辞書に登録し，元のテキストを再び単語分割するという手順を繰り返すと，効率良く単語辞書を構築できる．

表8 に言語モデル LM-10-1 を使用し，閾値 $\theta = 0.4$ とした際の未知語タイプ別の再現率と適合率を示す．一般に，数字列，アルファベット列，カタカナ列から構成される単語は語彙獲得精度が高い．その次に精度が高いのは漢字列から構成される単語である．ひらがな列，漢字–ひらがな列，ひらがな–漢字列から構成される単語は語彙獲得精度が低い．しかし，ひらがな列および漢字–ひらがな列の再現率は50%を超えており，従来，収集が難しいとされていたひらがな語や複数の字種から構成される単語も高い精度で収集できることが分かる．

また言語モデル LM-10-1 を使用したとき，出現頻度1，2，3の単語の語彙獲得の再現率は，それぞれ49.4%，70.6%，79.0%であった☆．出現頻度1の単語の約半数は収集され，出現頻度が大きくなるほど収集精度が高くなる．これより本手法は従来，収集が難しいとされていた低頻度語も高い精度で収集できることが分かる．

5. 考　　察

日本語テキストからの語彙獲得法の性能評価の問題点は，多くの人が合意できる唯一の正解が存在しないことである．一般に，本論文のように唯一の正解（EDR コーパスの単語分割）との完全一致に基づく評価は，（ある被験者の）許容可能性に基づく評価よりも過小評価になる傾向がある．

図10 に，言語モデル LM-10-1 を使用し，単語出現頻度の期待値の閾値 θ が 0.4 の場合に，収集に成功した未知語（`matched`），誤って収集した単語候補（`sys-matched`），収集に失敗した未知語（`std-matched`）の例を示す．（少なくとも筆者にとっては）システムが収集した未知語の多くは許容可能であるし，システムが収集に失敗した未知語の多くはすでに辞書に登録されている単語の組合せに分割できる．

たとえば，「受注価格」は EDR コーパスでは1つの単語として扱われているが，システム出力では「受注」と「価格」（両方とも辞書に登録されている）に分割されたため，未知語収集誤りとなる．システム出力で未知語とされている「ユークリッド」や「ムガール」は，それぞれ「ユークリッド距離」や「ムガール帝国」が EDR コーパスで1つの単語とされているために誤りとなる．大部分の未知語収集誤りは，このような許容可能なパターンである．ちなみに，システム出力から無作為に1000語を選択し，人手により（筆者が）許容可能性を調べたところ，適合率は86.5%で

☆ 試験テキストには76001語の未知語があり，そのうち出現頻度1，2，3の単語がそれぞれ45091語，11970語，5852語ある．

matched=46747（収集に成功した未知語）
田園風景 ガレ場 本務 釈然 総務部 明々 縞 おど 摩耗性
漢方 陣頭指揮 づ 渋谷駅 泣き寝入り はしご 名作
大学院生 書庫 ヒンズークシ山脈 黒猫 胸囲 せっけん
メリヤス 一般競争入札 炎天下 フッカー ５４．３ ふち
他紙 ナメ ...

sys-matched=22802（誤って収集した単語候補）
地元大 割り増しつき 中小商店 一通 ろ過パッド 縁先
地方交付税法 日本印刷 ユークリッド ゾーン符号 箕面
ＩＣ供給 学級講座 開発目的 市交通局 ゆうゆ 高速性能
１００万べん 間高校 汚染井戸水 運転の 情報大学 美夫
物乞 さぐっ 予定額 村田敬 人民議会議長 ムガール
７５チャッド ...

std-matched=29254（収集に失敗した未知語）
売上税反対 症候群 サークル株式会社 受注価格 結合関係
断続的に 移動型 自習教材 スパイ防止法案 工業用酸素
サラウンド機能 ミニフロッピーディスク 規制品目
スリランカ政府 ベッドシーン ランドフォンテン金鉱 農務
平均株価 趙紫陽首相 取材者 シャイ ちりめん友禅
フェスティバルホール 全仏オープン・テニス
電子メール・ゲートウエー 浦項製鉄所 食べ歩き仲間
筋立て 低視程 スペクトル分析 ...

std=76001, sys=69549, matched=46747
rec=61.5(46747/76001), prec=67.2(46747/69549)

図10 未知語収集の例（LM-10-1, $\theta = 0.4$）
Fig. 10 Example of extracted words (LM-10-1, $\theta = 0.4$).

あった．

もちろん明らかな誤りもある．最も顕著なのは，固有名詞と活用語に関する誤りである．固有名詞は表記が多様なので，単語分割を誤ることが多い．たとえば，図10では，「大間高校」（オオマコウコウ）が「大」と「間高校」に，「冨美夫」（トミオ）が「冨」と「美夫」に，「村田敬次郎」（ムラタケイジロウ）が「村田敬」と「次郎」に誤って分割された結果，「間高校」「美夫」「村田敬」が未知語として抽出されている．

これらは中国語系の漢字語（漢語）と和語系の漢字語を単語モデルの中で区別していないことに原因がある．人名・地名などの固有名詞は和語の漢字表記が多く，特に人名の場合，読みが重要で漢字表記は当て字に近いことが多い．現在の単語モデルは単語の表記しか考慮していないが，今後は単語の読みを考慮することにより漢語と和語の違いを単語モデルに反映させる必要がある．

また，後続するひらがな列を活用語尾と解釈し，活用語ではない文字列を活用語と同定する誤りが多い．図10では，「ゆうゆう」（形容動詞）が「ゆうゆ」と「う」（動詞語尾）に，「運転のしやすさ」が「運転の」「し」（動詞語尾）「やす」（形容詞）「さ」（接尾語）に，「物乞い」が「物乞」と「い」（形容詞語尾）に分割された結果，「ゆうゆ」「運転の」「物乞」が未知語として抽出されている☆．

これは本論文の単語モデルが未知語の品詞を考慮していないことに原因がある．未知語の多くは名詞であり，動詞や形容詞などの活用語であることは比較的少ない．また名詞が出現する文脈と動詞や形容詞が出現する文脈は異なる．これらの性質が現在の単語モデルには反映されていない．少なくとも品詞別の未知語の出現頻度を考慮すれば，このタイプの誤りを大幅に削減できる．

6. 関連研究

中国語では，Fung ら[4]は，テキストコーパス（香港立法評議会議事録，広東語，約60万文字）から統計的に有意な文字 ngram を取り出すことにより，地域や分野に特有の単語を収集する方法を提案した．収集された文字列が正しい単語であるかを4人の被験者で評価し，2文字単語，3文字単語，4文字単語の適合率をそれぞれ78.13%，31.3%，36.75%と報告している．

Chang ら[3]は，単語分割されていない大きなコーパス（約31万文）と単語分割された小さなコーパス（1000文）を用いて中国語の辞書を自動的に作成する方法を提案した．彼らの手法は，単語 unigram を用いたビタビ再推定，および，文字 ngram を特徴量として文字列が単語かどうかを判定する2クラス分類器（Two-Class Classifier）を用いる．システム出力を2つのオンライン中国語辞書（約2万語）と比較し，2文字単語では再現率56.88%，適合率77.37%，3文字単語では再現率6.12%，適合率85.97%と報告している．

日本語では，Nagao ら[10]は，suffix array 法を用いて任意の長さの文字 ngram を求める方法を提案し，文字 ngram の出現頻度が語彙獲得に役立つことを指摘しているが，精度に関する具体的な評価はない．

Mori ら[9]は，形態素解析済みのコーパスからある品詞に属する単語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各品詞に属する確率を推定する方法を提案した．EDR コーパスから語彙を収集し，日本語形態素解析システム JUMAN の辞書に登録されていないものを未知語と見なした評価では，再現率57.1%，適合率69.1%と報告されている．ただし，ひらがな語は収集対象外とされている．

伊東ら[5]は，単語を構成する字種により290個にタ

☆ ただしこれらは下位の単語分割候補に出現したものであり，第1候補の単語分割は正しい．

イプ分けし，単語タイプの出現確率と文字クラス別文字 unigram 確率の積から未知語の出現確率を求めるモデルを提案した．新聞やパソコン通信の電子会議室中の文章を対象とした実験では，未知語抽出の再現率40～50%，適合率80～85%と報告されている．彼らのモデルは単語タイプの出現確率を利用している点で，単語タイプの情報が単語長分布にしか反映されない本手法より優れている．しかし，単語表記確率のモデルとして文字クラス別文字 unigram モデル（彼らの手法）と文字 bigram モデル（本手法）のどちらが優れているかは議論の余地がある．

本手法は，単語分割に未知語の確率モデルを使用し，単語らしさの評価尺度には単語出現頻度の期待値を用いる．単語分割されたコーパスを正解と見なし，未知語率 7.72%のテキスト（10 万文）に対して再現率61.5%，適合率 67.2%を得ている．従来手法[3)~5),9)]と本手法を比較することは，言語（中国語と日本語），訓練に用いた（単語分割された/されていない）コーパスの大きさ，既知の単語リストの大きさ，試験に用いたコーパスの大きさと未知語率，正解データの種類（オンライン辞書と単語分割されたコーパス）などの実験条件が異なるので不可能である．しかし，人手により単語分割されたコーパスとの完全一致による評価は過小評価になりやすいことを考慮すると，本手法は従来手法と同等，あるいは，それ以上の語彙獲得精度を持つと思われる☆．

7. おわりに

本論文では，未知語の確率モデルと単語出現頻度の期待値に基づいて日本語テキストから未知語を収集する方法を述べた．本論文で提案した語彙獲得法の特徴は，ひらがな語，複数の文字から構成される単語，出現頻度が低い単語を高い精度で収集できることである．

本手法の今後の課題は，単語の読みと品詞を考慮した未知語の確率モデルを考案することである．これにより固有名詞と活用語に関する誤りの多くを解決できると予想している．

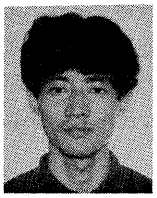
参考文献

1) Baayen, H. and Sproat, R.: Estimating Lexical Priors for Low-Frequency Morphologically Ambiguous Forms, *Computational Linguistics*, Vol.22, No.2, pp.155–166 (1996).
2) Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., Lai, J.C. and Mercer, R.L.: An Estimate of an Upper Bound for the Entropy of English, *Computational Linguistics*, Vol.18, No.1, pp.31–40 (1992).
3) Chang, J.-S., Lin, Y.-C. and Su, K.-Y.: Automatic Construction of a Chinese Electronic Dictionary, *Proc. 3rd Workshop on Very Large Corpora*, pp.107–120 (1995).
4) Fung, P. and Wu, D.: Statistical Augmentation of a Chinese Machine-Readable Dictionary, *Proc. 2nd Workshop on Very Large Corpora*, pp.69–85 (1994).
5) 伊東伸泰，西村雅史：N-gram を用いた日本語テキストの単語単位への分割，情報処理学会研究報告，96-NL-122, pp.57–62 (1997).
6) Jelinek, F.: Self-Organized Language Modeling for Speech recognition, Technical Report, IBM T.J. Watson Research Center (1985).
7) 北 研二，中村 哲，永田昌明：音声言語処理—コーパスに基づくアプローチ，森北出版 (1996).
8) Mikheev, A.: Automatic Rule Induction for Unknown-Word Guessing, *Computational Linguistics*, Vol.23, No.3, pp.405–423 (1997).
9) Mori, S. and Nagao, M.: Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis, *Proc. 16th International Conference on Computational Linguistics*, pp.1119–1122 (1996).
10) Nagao, M. and Mori, S.: A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, *Proc. 15th International Conference on Computational Linguistics*, pp.611–615 (1994).
11) Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A^* N-Best Search Algorithm, *Proc. 15th International Conference on Computational Linguistics*, pp.201–207 (1994).
12) Nagata, M.: Automatic Extraction of New Words from Japanese Texts using Generalized Forward-Backward Search, *Proc. Conference on Empirical Methods in Natural Language Processing*, pp.48–59 (1996).
13) 中渡瀬秀一：正規化頻度による形態素境界の推定，情報処理学会研究報告，96-NL-113, pp.13–18 (1996).
14) 日本電子化辞書研究所：EDR 電子化辞書仕様説明書（第 1 版）(1995).
15) Nobesawa, S., Tsutsumi, J., Jiang, S.D., Sano, T., Sato, K. and Nakanishi, M.: Segment-

☆ ただし，本論文の実験では未知語率が高い状態を人工的に作り出しているので，より厳密な評価のためには，特許文や法律文のような EDR コーパスに収録されていない分野のテキストを EDR コーパスと同じ基準で単語分割し，これを正解データとして評価実験を行う必要がある．これについては今後の課題とする．

ing Sentences into Linky Strings Using D-bigram Statistics, *Proc.16th International Conference on Computational Linguistics*, pp.586–591 (1996).
16) Smadja, F.: Retrieving Collocations from Text: Xtract, *Computational Linguistics*, Vol.19, No.1, pp.143–177 (1993).
17) Sproat, R. and Shih, C.: A Statistical Method for Finding Word Boundaries in Chinese Text, *Computer Processing of Chinese and Oriental Languages*, Vol.4, pp.336–351 (1990).
18) Sproat, R., Shih, C., Gale, W. and Chang, N.: A Stochastic Finite-State Word-Segmentation Algorithm for Chinese, *Computational Linguistics*, Vol.22, No.3, pp.377–404 (1996).
19) Weischedel, R., Meteer, M., Schwartz, R., Ramshaw, L. and Palmucci, J.: Coping with Ambiguity and Unknown Words through Probabilistic Models, *Computational Linguistics*, Vol.19, No.2, pp.359–382 (1993).
20) 吉村賢治，武内美津乃，津田健蔵，首藤公昭：未登録語を含む日本語文の形態素解析，情報処理学会論文誌，Vol.30, No.3, pp.294–301 (1989).

(平成10年8月7日受付)
(平成11年7月1日採録)



永田 昌明（正会員）

1985年京都大学工学部情報工学科卒業．1987年同大学院工学研究科修士課程修了．同年，日本電信電話（株）入社．1989年ATR自動翻訳電話研究所へ出向．1993年日本電信電話（株）へ復帰．現在，サイバースペース研究所勤務．音声翻訳，統計的自然言語処理の研究に従事．工学博士．電子情報通信学会，人工知能学会，言語処理学会，ACL各会員．