

統計モデルによる日本語の形態素解析手法

朴 哲済⁺⁺ 李 鍾赫^{**} 李 根培^{*}

+早稲田大学情報科学研究中心

#浦項工科大学情報通信研究所

*浦項工科大学電子計算学科

E-mail: cjpark@madonna.postech.ac.kr

{jhlee, gblee}@vision.postech.ac.kr

本稿では、拡張CYK法に単語接続に関する統計モデルを利用した日本語の形態素解析手法を提案する。形態素解析の過程はまず、接続情報を検査し接続が可能な形態素解析結果をすべて得たのち、ヒューリスティックスを利用して優先順位を決める。我々は、接続情報表の値を確率として用いることにより接続の強度を表現し、その強度により形態素候補の優先順位を決めた。このとき用いる確率情報は、形態素解析の対象言語に関する確率モデルとして統計情報抽出機構から得られる。このような形態素解析手法を日本語を対象に約24万形態素のコーパスから接続情報を抽出し実験した結果95.2%の解析成功率を得た。本論文では確率接続表を用いた形態素解析能力と、接続の可不可のみを表した接続表を用いた形態素解析機構を比較し評価を行う。

A Japanese Morphological Analysis Based on Statistical Method

Chul-Jae Park⁺⁺ Jong-Hyeok Lee^{**} GeunBae Lee^{*}

+Centre for Informatics, Waseda University

#Information Research Laboratories, POSTECH

*Computer Science & Engineering, POSTECH

E-mail: cjpark@madonna.postech.ac.kr

{jhlee, gblee}@vision.postech.ac.kr

This paper proposes a method for Japanese morphological analysis based on the CYK algorithm using a statistical model about the conjunctive relationships between words in a sentence. The morphological analysis process is comprised of two main stages: The extraction of the several analysis results each of which are checked for morphological connectivity, and the computation of an analysis score of each morpheme chain for the results of the first stage. Through the application of statistical probabilities in connectivity information between neighbouring morphemes, the most preferable analysis among the several possible ones can be selected. Probability information is extracted from a corpus of about 240,000 morphemes. Based upon these concepts, we developed a Japanese Morphological analyzer, and obtained 95.2% of accuracy for Japanese sample sentences from news. By comparing experimentally with other algorithm which represent connectivity information by bits, we demonstrate that the statistical method is more efficient.

1. はじめに

統計的手法に基づいた自然言語処理システムにおいて、処理する言語に関する統計情報（以下言語情報と呼ぶ）の正確性は、そのシステムの性能に大きな影響を与える。このため、言語情報をより正確なものにするために膨大な量の蓄積を行う必要がある。しかし、この膨大な量の蓄積を自動的に行なうことは困難であり、膨大な時間と努力をかけざるを得ない。これまで研究されてきた統計的方法は、主として単語構造を対象に1次マルコフモデル（Markov model）に従ってコストを設定した形態素解析[4]がある。この方法は最小コストパスの近似解が高速に得られるように、最長一致法を一般化したアルゴリズムを採用している。また、

文字列情報を利用したN-Gramアルゴリズムや文節の構造的特徴を利用した形態素解析[2]等がある。本研究では、文レベルの言語情報の自動抽出および抽出された言語情報を利用した形態素解析手法を提案する。必要な言語情報は統計的方法によってコーパスから自動的に構築できるようにした。統計的方法は経験によってコーパスから、もっと頻繁に使われているものが自然に受け入れられるものだという判断に従って、形態素候補の中でその経験値が一番高いものを選択する方式である。

筆者らは、動的プログラミング技法であるCYK法を拡張した形態素解析を日一韓機械翻訳システムの解析部として構築した。これはまず、すべての解析結果(path)を求めたのち、接続の強さ優先ヒューリスティックスと形態素数最小ヒューリスティックスを用いて解析結果の優先順位を決める。このとき連続に現れた二つの形態素間の接続は、接続情報表を利用して検査する。ある形態素の右に接続可能なものと、左に接続可能なものに関する情報を持っているのが接続情報表である。われわれは、この接続情報表を確率値を使って具現した。左接続情報Lと右接続情報Rが接続可能なら、接続情報表の該当位置に0から1のあいだの確率を接続の強度として与えた。この値が1に近いときに強い接続を、0に近いときに弱い接続を表す。この

値は、動詞の後に名詞が来る確率のような品詞間接続に関する確率である。

本論文の構成は次のようになっている。まず初めに、2章で日本語に関する統計モデルの設定と統計データの収得方法について述べる。3章で、確率接続表を用いた日本語の形態素解析手法を提案する。4章では、本手法による実験結果を提示し、性能評価を行う。最後に、5章で、まとめを行なう。

2 統計モデルの設定

大量の日本語文書を対象に言語を構成する字種と品詞間の接続関係について統計を抽出し、この統計を基に形態素解析に必要な確率接続表を設定した。

2.1 Tri-gramモデル

一般的に品詞は与えられた単語に対して唯一に決められるものではなく、多くの単語が文脈によって異なる品詞を持つ。ある単語が複数の品詞を持つ場合、その単語は品詞的衆意性があるという。

統計情報抽出部で使用する品詞は、一般的に細分化するほど統計の間違いが少なくなる。一方、細分化しすぎると不必要的品詞まで区別され、実際利用の価値がなくなる。本論文で使用する接続情報は左接続情報数 256個、右接続情報数 256個である。日本語の文節は、2または3個以上の形態素に分割できる。そこで処理の用意性と正確性のため衆意性解消の単位を文節としてtri-gramを使った。文節タグ(tag)は単語に対するタグの結合で構成されている。

Tri-gramモデルは、一つの文Sに対して形態素解析の結果を入力として各文節が持っている品詞の集合、またはパスの中で式(1)を満足するタグ列Tを探すことである。以下で用いるすべての式では、ランダム変数(random variable)の表記は省略する。

$$T(w_{1..n}) = \arg \max_{T_{1..n}} \prod_{i=1}^n P(t_i | w_i) P(t_i | t_{i-N+1..i-1}) \quad (1)$$

ここで、 $P(t_i | w_i)$ は語彙的確率(lexical probab

ility), $P(t_i | t_{i-N+1..i-1})$ は文脈的確率(contextual probability)と呼ぶ。これは以下のような頻度によって推定できる。

$$P(t_i | w_i) = \frac{\text{freq}(t_i | w_i)}{\text{freq}(w_i)} \quad (2)$$

$$P(t_i | t_{i-N+1..i-1}) = \frac{\text{freq}(t_{i-N+1..i-1})}{\text{freq}(t_{i-N+1..i-1})} \quad (3)$$

2.2 統計抽出

本研究では、情報理論的概念である相互情報測定値(mutual information measure)を利用した統計的アプローチ方法を提案する。相互情報測定値はコーパスから自動取得が出来るので情報の構築が容易なことと、処理分野の変化に能動的に対応可能である。

2.2.1 相互情報

文脈上特別な意味を持つ特定な単語は、他の特定な単語と同じ文の中で一緒に現れる現状を言語学では連語関係(collocation relation)、または、共記関係(co-occurrence relation)と呼ぶ。単純な例として「目」と「閉じる」、「口」と「つぐむ」は共記関係にある単語として、これを使う文の中ではほとんど一緒に現れる。即ち、共記関係にある単語の集合としての共記パターンは、文章の内部で単語たちが互い交わって現れることを判断する選択制約知識を表現する一つの方法になる。

本稿では、情報理論的概念である相互情報測定値(mutual information measure)を用いて、共記関係にある単語たちの相互関連性(word association)を計る。相互情報測定値はコーパスから比較的簡単に求められる。客観的評価式として、任意の二つの単語 x と y の統計的相関関係 $I(x, y)$ は次のようになる[2]。

$$\begin{aligned} I(x, y) &= \log_2 \frac{P(x, y)}{P(x)P(y)} \approx \log_2 \frac{\frac{f(x, y)}{N}}{\frac{f(x)}{N} \frac{f(y)}{N}} \\ &= \log_2 \frac{Nf(x, y)}{f(x)f(y)} \end{aligned} \quad (4)$$

この式で $P(x)$ と $P(y)$ は単語確率(word probability)として、単語 x と y がそれぞれ独立的に使用される確率を表す。そして $P(x, y)$ は共記確率(joint probability)で、単語 x と y が一緒に使用される確率を表す。この $P(x)$, $P(y)$ および $P(x, y)$ は、それぞれコーパスの中で x と y が現れた頻度 $f(x)$, $f(y)$ および文の中で w 個の文節で単語 x と y が一緒に現れる頻度 $f_w(x, y)$ を計って、コーパスの大きさ N で正規化した近似値として求めた。コーパスから上の式によって求めた相互情報測定値が大きいほど、二つの単語の間の意味関係はもっと緊密である。

2.2.2 統計抽出手法

大量の日本語文を対象としているため、人手で統計をとるのでは、膨大な時間と手間がかかってしまう。よって、既存のシステムを利用することで手間を最小限に押えることを考えた。今回考案・使用した統計取得手順は次の通りである。

(1) 日本語文を形態素解析する。

日本語文を、形態素解析システムJUMANにかけ、形態素に分割された解析結果を得た(解析結果として、形態素文字列、読み、活用の基本形、品詞、活用の種類、活用形などが outputされる)。

(2) 形態素解析の出力結果を修正する。

JUMANで定義されている文法と、我々が定義した文法の異なる部分を修正するツールを作成し、これを実行することで機械的に修正した。

(3) 統計をとる。

(2)の出力結果から、語を構成する字種と、品詞間の接続関係についての統計をとるツールを作成・実行した。

以上の手順ことで形態素解析で用いる初期確率接続表の値を自動的に獲得した。

2.2.3 信頼度評価及び再構成処理

形態素解析結果の分析によって確率接続表の学習効果を図る。接続情報がすでに構成されている場合、前回までの確率値と今回の形態素解析の結果による確率値を計算し、新しい確率接続表を構

成する。再構成処理はこの値を計算し、確率接続表の信頼度を評価して再構成する。今回の形態素解析における確率値と前回までの確率値を次のように計算し、最終確率値を設定する。まず、今までの確率接続表に登録されていない品詞の場合は、今回の確率接続値をそのまま設定する。前回まで設定されている場合は、次のような計算によってその値を変更する。

$$\text{確率値} = P + Q - (P \times Q)$$

P : 今回求められた確率値

Q : 前回までの確率値

3. 統計モデルによる形態素解析手法の提案

本システムは日本語の形態素に関する左右接続情報と形態素間の接続可能性を検査する接続検査表を利用して宣言的に処理する。宣言的処理は、プログラムを簡単にするとともに、接続情報の維持・補修を容易にする。形態素解析にはCYK法を拡張して用いた。形態素解析の過程はまず、CYK法により可能なすべての解析結果を得たのち、ヒューリスティックスを用いて優先順位を決める。このヒューリスティックスは自由に選択できる。また、ヒューリスティックスを変えることによりさまざまな優先順位が得られる。

3.1 形態素解析の方法

形態素解析は文節を単位として形態素の辞書検索および、形態素と形態素間の接続検査によって行う。CYKアルゴリズムは三角テーブルTを用いる。一つのエレメント $T[i, j]$ は、候補文字の組み合わせである全文字列の i 番目文字から j 番目文字までの形態素解析結果を表す。全体の長さ n の文字列に対して、この $T[i, j]$ は一つの形態素、または、 $T[i, k]$ と $T[i+k+1, j]$ ($0 \leq k < n-i$) の結合としてまた現れる。このことを利用すると形態素の結合形態は、 $T[i, n-1]$ のみ考慮してもできる。そして時間計算量は $O(n^2)$ になる。図 1 にアゴリズムを簡略に示す。

```

for (row = 0; row < n; row++) {
    for (col = row; col < n; col++) {
        sub1 = T(col + 1, n);
        if (sub1 != NULL)
            sub2 = Read_From_Dictionary(row, col);
        connect = Connection_Check(sub1, sub2);
        if (connect)
            T(row, n) = sub1 + sub2;
    }
}

```

図 1: 形態素解析のアルゴリズム

3.2 接続情報

接続情報は文節を成す形態素間の接続に制約を与えて、正確な分析を行うための情報である。これは辞書に形態素と一緒に登録する。われわれは入力言語である日本語の各形態素に対して左右接続情報を附加した。本システムで使用する辞書は日一韓対訳辞書である。辞書には日本語表題語、その表題語ごとの左右接続情報、そして対応する韓国語、韓国語に与えられる接続情報と意味情報を登録した。

連続に現れる形態素間の接続は、接続情報表を利用して検査する。ある形態素の右に接続できるものと、左に接続できるものに関する情報を持っているのが接続情報表である。本論文では、この接続情報表を二つの形式に構成して比較評価を行った。一つはビットマップ形式として具現した。もう一つは接続情報の値を確率値として提供する確率接続情報表として具現した。ビット形式の接続情報表は左接続情報と右接続情報が接続可能なら、接続情報表の該当する位置にビットで接続可能性を表した。その値が1なら接続可能、0なら接続が不可能である。反面、確率情報表は統計情報抽出機構から得られた形態素間の接続可能性に関する経験値を基に、左接続情報と右接続情報間の確率的接続可能値を接続情報表の該当位置に与えた。

3.3 ビット接続表でのヒューリスティックス

形態素間の接続情報を用いてCYK法により形態素解析をすると、接続が可能なすべての結果が得られる。形態素選択部は、形態素解析機構の後処理として形態素解析結果が一つ以上になったとき、いろいろなヒューリスティックスを利用して解析結果を優先順位別に整列する。形態素選択部では、他のヒューリスティックスも自由に使えるようにした。そして使うヒューリスティックスによって違う結果を得ることができる。

(1) 強い接続優先ヒューリスティックス

ビット形式の接続表を用いた接続検査では、接続不可能、弱い接続、よく使う強い接続の三つに接続の形態を細分した。「名詞+ピリオド」、「副詞+名詞」、「名詞+動詞」等が弱い接続の例である。強い接続の例としては、「名詞+助詞」、「動詞語幹+語尾」等である。この接続形態によってコストを与えた。コストは形態素解析結果パスの中で、弱い接続が多ければ高いコストを持つ。本システムでは強い接続にはコスト1、弱い接続にはコスト15を与えた。次にその例を示す。

例)「女性より男性が多くなる。」の解析結果

(a) [女性(名詞)][より(名詞)][男性(名詞)]

[が][多][く][なる].[.]

コスト = 35

(b) [女性(名詞)][より(助詞)][男性(名詞)]

[が][多][く][なる].[.]

コスト = 7

になってコストが小さい(b)が選択される。

(2) 最小形態素数ヒューリスティックス

このヒューリスティックスは形態素の個数が少ないパスを選択する方法である。本システムでは辞書に単語を登録する方法として、よく使うものを複合形態として登録した。それで、形態素の数が少ないパスをもっと優先に選択した。この二つのヒューリスティックスを結合する方法として、個々のコストを掛け算をして総合コストを求める。総合コストを得たのち、その順に整列して総合コストが小さい一つのパスを選択する。

3.4 確率接続表でのヒューリスティックス

確率形式の接続表を用いた接続検査の値は、0から1までの品詞間の接続確率である。0は接続不可能、1に近い値ならよく使う強い接続を表す。例えば、「名詞+名詞」、「名詞+ピリオド」、「副詞+名詞」、「名詞+動詞」の品詞構成においてそれぞれ、0.32, 0.14, 0.10, 0.64等の確率が接続の強さとして付加される。この確率を形態素解析結果パスの接続強度として与える。この方式では、接続強度の値が接続の強さをそのまま表す。したがって、接続強度が高い結果が選ばれる。次は拡張CYKアルゴリズムを通して二つの解析結果が出たとき、確率接続表を用いて一つのパスを選択する例を示している。

例)「女性より男性が多くなる。」の解析結果

(a) [女性(名詞)][より(名詞)][男性(名詞)]

[が][多][く][なる].[.]

接続強度 = 0.64

(b) [女性(名詞)][より(助詞)][男性(名詞)]

[が][多][く][なる].[.]

接続強度 = 1.08

になって接続強度が高い(b)が選択される。

4. 実験結果及び考察

4.1 実験方法

本論文で提示した統計情報抽出機構と形態素解析機構を構築し、確率接続表を用いた形態素解析能力を分析した。本稿での実験環境は、SPARCStation 10 (90 MIPS)であり、C言語で実装した。これは鉄鋼関係の特許文書を翻訳するための機械翻訳システムの解析部として開発した。まず、統計情報抽出機構を利用して日本語に関する確率接続表を構成した。形態素解析手法はCYK法を拡張して用いた。CYK法で可能な解析結果をすべて得たのち、ヒューリスティックスを利用して優先順位を決めた。このときビット接続表を用いた形態素解析結果と、確率接続表を用いた形態素解析結果を比較し、それぞれの正解率を計った。実験のための準備事項は次のようなものである。

• 辞書の構成

3万単語の表題語と接続情報で構成されている。辞書アクセスはハッシュ法を利用している。

• 確率接続表の設定

確率接続表での語彙的確率、文脈的確率および共記パターンでの相互情報測定値を求めるためにはタギング(tagging)されたコーパスが必要である。これを人によってタギングするのには膨大な時間と努力が要求されるのでタギングされたコーパスの自動収集が必要である。筆者らは朝日新聞社説6カ月分のコーパス(約24万形態素)から自動タギングを行った。この結果の間違ったところを修正して、各単語間の接続可能確率を求め確率接続表を作成する。

• 実験データ

定量的評価は、「NHKのニュース」1993年6~12月までのデータから月別に抽出した文章に対して、ビット接続表(方法1)と確率接続表(方法2)を用いて実験を行った。

4.2 実験結果

現在稼動している日一韓機械翻訳システムの形態素解析部を修正して実験を行った。評価の目的は、形態素解析部での接続表の構成において確率的接続情報表を用いた形態素解析システムの利点と欠点を洗い出すことに置いた。評価項目は、正解率、文の長さと解析時間、品詞別解析率、単語特徴別解析率を設定した。

実験データの1文当たり平均形態素数は31.08個で、平均文節数は13.20個の長文である。ビット形式の接続表を用いた形態素解析(方法1)では、総1,523個の形態素に対して93.3%の解析成功率を得た。一方、確率形式の接続表を用いた形態素解析(方法2)では、同じデータに対して95.2%の解析成功率を得た。失敗原因の大部分は接続情報表の未整備にあった。表1に、文の長さと解析時間の相関関係を表した。実験結果、1文字当たり平均解析時間は、方法1で、0.0402秒、方法2で、0.0380秒になった。表2、表3に、品詞と解析成功率の相関関係、単語特徴と解析成功率の相関関係を表した。

表1：文の長さと解析時間

文の長さ	総文数	比率 (%)	方法1 (BIT接続表)		方法2 (確率接続表)	
			平均時間 (秒)	合計時間 (秒)	平均時間 (秒)	合計時間 (秒)
- 19	3	6.1	0.213	0.639	0.265	0.795
20 - 29	4	8.2	0.947	3.788	0.893	3.572
30 - 39	11	22.4	1.233	13.563	1.235	13.585
40 - 49	6	12.2	1.242	7.452	1.496	8.976
50 - 59	4	8.2	1.649	6.596	1.622	6.488
60 - 69	7	14.3	1.905	13.335	1.874	13.118
70 - 79	2	4.1	2.540	5.080	2.180	4.360
80 - 89	4	8.2	3.290	13.160	2.587	10.348
90 - 99	3	6.1	3.244	9.732	3.151	9.453
100 -	5	10.2	3.769	18.845	3.294	16.470
合計	49	100	-	92.190	-	87.165

表 2: 品詞別解析成功率

品詞名	形態素数	比率 (%)	方法1(BIT接続表)		方法2(確率接続表)	
			成功個数	成功率 (%)	成功個数	成功率 (%)
名詞	631	41.4	626	99.2	618	97.9
動詞	194	12.7	169	87.1	168	86.6
形容詞	26	1.7	26	100	25	96.2
形容動詞	14	0.9	14	100	14	100
副詞	13	0.9	11	84.6	11	84.6
連体詞	10	0.7	9	90.0	8	80.0
接続詞	3	0.2	3	100	3	100
感動詞	0	0	0	-	0	-
助詞	422	27.7	363	86.0	398	94.3
助動詞	57	3.7	49	86.0	52	91.2
接辞	27	1.8	25	92.6	27	100
特殊	126	8.3	126	100	126	100
合計	1,523	100	1,421	93.3	1,450	95.2

表 3: 単語特徴別解析成功率

区分	形態素数	比率 (%)	方法1(BIT接続表)		方法2(確率接続表)	
			成功個数	成功率 (%)	成功個数	成功率 (%)
略語	6	0.39	6	100	6	100
地名	56	3.68	56	100	46	82.1
人名	15	0.98	15	100	15	100
固有名詞	17	1.12	17	100	17	100
カタカナ表記	30	1.97	30	100	30	100
2字成語漢字	329	21.6	317	96.4	315	95.7
特殊なもの	31	2.04	16	51.6	30	96.8

5. まとめ

本稿では、拡張CYK法に確率接続表を用いた形態素解析手法を提案した。システムの構成は統計情報抽出部、形態素解析のためのCYK部およびコスト計算部に分けられる。このシステムは日一韓機械翻訳システムの日本語形態素解析機構として具現した。日本語形態素解析のため、形態素の左側と右側に来ることが出来る形態素の種類別にコード化した接続情報を付加した。日本語の接続情報は左右個々、256個のコードで表した。日本語文の形態素解析方法は拡張CYK法を用いた。接続情報を検査して接続が可能な形態素解析結果をすべて得たのち、ヒューリスティックスを利用して優先順位を決めた。

我々は接続情報表の値を確率として用いることにより接続の強度を表現し、その強度により形態素候補の優先順位を決めた。このとき用いる確率情報は形態素解析対象言語に関する確率モデルとして統計情報抽出機構から得られる。CYK部とコスト計算部を通す本形態素解析手法をビット形式の接続表と確率形式の接続表に分けて、それぞれNHKニュースの文書を持って実験したところ、93.3%、95.2%の解析成功率を得た。ビット形式の接続表を用いたときより確率接続表を用いた方が1.9%正解率が向上された。また、実行時間においても1文字当たり平均0.0022秒程度短縮された。全体の形態素解析において第一順位の結果では対象言語に関する統計モデルを使用することによって、より自然な解析結果を得ることができ、本確率接続情報を用いた形態素解析方法の有効性を確認した。

本研究で提案した確率接続表を用いた拡張CYK法による形態素解析システムは、鉄鋼関係特許文書(約13万件の特許文書)翻訳のための、日一韓機械翻訳システムの形態素解析機構として構築した。現在は、実用的形態素解析機構としての利用可能性を検証している。

参考文献

- [1] EunJa Kim, Jong-Hyeok Lee : Implementation of Japanese-Korean MT System: Japanese Morphological Analysis Using Heuristics, Proc. of The Korea Information Science Society'93, pp. 797-800, 1993 (in Korean).
- [2] JinHee Yoo, Jong-Hyeok Lee, GeunBae Lee : Post-Processing for Character Recognition Using Morphological Analysis and Linguistic Evaluation, Journal of The Korea Information Science Society, Vol. 22, No. 6, pp. 880-891, 1995 (in Korean).
- [3] 三部 裕史, 大森 健児 : 信頼性の低い文字認識結果に対する言語情報を用いた誤認識文字の訂正, 情報処理学会論文誌, Vol. 34, No. 10, p. 2117-2124, 1993.
- [4] 下村 秀樹, 並木 美太郎, 中川 正樹, 高橋 延匡 : 最小コストパス探索モデルの形態素解析に基づく日本語誤り検出の一方式, 情報処理学会論文誌, Vol. 33, No. 4, pp. 457-464, 1992.
- [5] 鈴木 哲也, 朴 哲済, 中山 康徳, 谷口 清継, 寛 捷彦 : 信頼度評価に基づく活用形の推定, 日本ソフトウェア科学会第11回大会, 1994.