タグなしコーパスからの確率情報の段階的学習と形態素解析への応用

鈴木由理　山口昌也　乾伸雄　小谷善行　西村恕彦
東京農工大学工学部電子情報工学科
コンピュータサイエンスコース

1．はじめに

近年、自然言語処理の分野では、統計的学習を行う手法が注目されており、形態素解析においても、例外ではない。自然言語に確率モデルを適用して、種々の確率値を求めることによって、形態素解析候補の評価が行われている。

タグ付きのコーパスを用いて確率情報を獲得するのは簡単である。すでに、分かち書きをする必要のない英語などの言語では、この方法を用いて獲得した確率情報を用いて形態素解析を行い、かなり良い精度を得ている。しかし、日本語のタグ付きコーパスの作成には膨大な手間がかかる。そこで、タグなしコーパスからの確率情報の獲得について、種々の研究が行われている。

本研究でも、タグなしコーパスから確率情報を獲得することを考えるが、その際に、間違った方向に学習が傾かないように、補正処理を行うことを考えた。

また、形態素解析に確率情報を用いる場合、一般的に品詞の接続情報を用いる。しかし、これ以外にも形態素解析の手がかりとなるものに、形態素の表記情報がある。この表記情報も確率情報に取り入れられないかと考えた。

このような、形態素解析システムを設計、実現し、確率情報の獲得を行った。さらにその確率情報を利用した形態素解析システムを実現した。

2．解析手法

本研究では、形態素解析結果の確からしさを、出現確率と接続確率で表す。考えられるすべての形態素解析候補を求め、その中で、出現しやすい形態素の組合せで構成しており、形態素が接続しやすい順序に並んでいるものを、確からしい解とする。

一文ごとの形態素解析結果の確からしさを、信頼性パラメータと呼ぶ。信頼性パラメータの値は、出現確率と接続確率の積で表わす。これらの確率値は段階的学習によって獲得する。

形態素解析において、統計的学習によって確率モデルのパラメータを求める場合、大量のコーパスから学習を行えば、接続確率だけで、十分な精度の解析結果が得られる。しかし、二つの形態素の接続頻度を調べなければならない接続確率は、使われる頻度の少ない形態素に関して、精度が悪くなりがちである。その点、同じ量のコーパスを用いても、単独の形態素の頻度だけを調べればよい出現確率は、出現頻度の低い形態素に関しても、比較的精度が良い。入力文に使われる頻度が少ない形態素が含まれている場合、この出現確率を併用することによって、接続確率だけを用いて、解析候補の選好を行うよりも、うまく対応できる。

ここで、二つの確率について、さらに詳しく述べる。

2．1　出現確率

出現確率は、すべての形態素の中で、ある特定の形態素が現れる割合を表す。形態素は語構成により品詞情報と表記情報で分類し、そのカテゴリごとの確率を考える。表記情報としては、文字種と文字数を用いる。分類については後述する。品詞：H、文字種文字数：M である形態素カテゴリを w_{HM} とすると、出現確率：T（w_{HM}）は次のように表される。

$$T(w_{HM}) = \frac{w_{HM} \text{ の語の出現回数}}{\text{テキストの語数}} \quad \text{（式1）}$$

2．2　接続確率

接続確率は、ある特定の形態素カテゴリの後に、別の（あるいは同じ）形態素カテゴリが連接する割合を表す。接続確率も出現確率と同様に、形態素を品詞情報と表記情報で分類し、そのカテゴリごとの確率を考える。品詞：H、文字種文字数：M である形態素カテゴリを w_{HM} とすると、接続確率：$K(w_{H'M'} \mid w_{HM})$ は次のように表される。

$$K(w_{H'M'} \mid w_{HM}) = \frac{w_{HM}\text{ の語と }w_{H'M'}\text{ の語の接続回数}}{w_{HM}\text{ の語の出現回数}}$$

（式２）

２．３　信頼性パラメータ

解析候補の選好をするために、候補の信頼性を表す信頼性パラメータを計算する。パラメータは文単位で、前述した出現確率、接続確率の総積によって計算される。入力文 $W = \{w_1, w_2, \cdots, w_n\}$ に対して、信頼性パラメータ：Q（W）は、次のように表される。

$$Q(W) = \prod_{i=1}^{n} T(w_{HiMi}) * \prod_{i=1}^{n} K(w_{Hi-1Mi-1} \mid w_{HiMi})$$

（式３）

３．段階的学習

前述した出現確率や接続確率は、大量のタグ付きコーパスがあれば、簡単に求めることができる。しかし、現実問題として、タグ付きコーパスを作成するには、莫大な時間と人手が必要である。そこで本研究では、タグなしのコーパスから確率情報を獲得することを考える。

タグなしコーパスからでは、じかに確率情報を獲得することはできない。まず、形態素解析を行って、タグ付きのコーパスに直さなければならない。しかし、それには確率情報が必要になる。そこで、はじめは確率情報ではなく、適当なスコアを与えて形態素解析を行い、その結果から確率情報を獲得することを考える。

そしてさらに、その確率情報を利用して形態素解析を行い、その結果からまた、確率情報を獲得する。これを相互に繰り返すことによって、徐々に実際の値に近い確率情報を獲得する。

スコアを用いた形態素解析を第一段階、確率情報を用いた形態素解析を第二段階とする。

４．補正処理

確率情報は学習によって獲得するので、そこには必ず誤りが含まれている。学習が間違った方向に傾いた場合、それを修正するための補正処理について述べる。

確率情報に誤りが含まれていると、特定の形態素が不当に選ばれやすい（あるいは選ばれにくい）ということが起こる。そこで、不当に選ばれやすいもの（あるいは選ばれにくいもの）に対して、パラメータの補正を行う。補正処理は次の二つに分けられる。

補正処理１：
　不当に選ばれやすいものを選ばれにくくする
　　（信頼性パラメータにある値をかける）
補正処理２：
　不当に選ばれにくいものを選ばれやすくする
　　（信頼性パラメータをある値で割る）

４．１　補正処理の対象と大きさ

補正を行う対象と、補正の大きさは、次のような手順で決定する。

補正処理１の手順
手順１：任意の Ｎ 文をシステムで解析し、間違っている形態素を取り出す。
手順２：システムで解析した結果全体に含まれる形態素を品詞、文字種、文字数で分類し、出現回数を求める。
手順３：分類ごとに間違った形態素の割合を求める。
手順４：間違える割合が Ｘ１ を越え、出現回数が Ｙ１ を越えるものを、補正の対象とする。
　形態素カテゴリ Ｂ に対する補正の大きさ $R1_B$ は、次の式で与えられる。

$R1(w_B)$ ＝
（システム出力中の w_B が正解であった割合）^{3}

（式４）

補正処理２の手順
手順１：任意の Ｎ 文を解析し、間違った形態素を調べ、その部分を正しく解析した場合の形態素を取り出す。
手順２：対象とした文を正しく解析した場合に、その文に含まれる形態素を品詞、文字種、文字数で分類し、出現回数を求める。
手順３：システムが間違った部分に相当する正しい形態素が、全体の出現回数中に占める割合を、分類ごとに求める。
手順４：間違えられた割合が Ｘ２ を越え、出現回数が Ｙ２ を越えるものを補正の対象とする。形態素カテゴリ Ｂ に対する、補正の大きさ $R2_B$ は、次の式で与えられる。

$$R2(w_B) = (\text{Bの正解率})^3 \qquad (式5)$$

例）入力文　「新しい法律だ」

正しい解析結果

新し　＋　い　＋　法律　＋　だ

形容詞語幹　形容詞語尾　名詞　助動詞

システムの解析結果

新　＋　しい　＋　法律　＋　だ

接頭語　名詞　名詞　助動詞

↓

間違われた形態素

「新し（形容詞語幹）」、「い（形容詞語尾）」

間違った形態素

「新（接頭語）」、「しい（名詞）」

処理対象の出現回数にしきい値を設けるのは、「たまたま数回出てきて、たまたま間違って解析された」、という形態素に対して、しなくて良い補正処理を行わないようにするためである。

4．2　補正の方法

補正処理を行った信頼性パラメータ値：Q'（W）は、次のように表される。

$$Q'(W) = Q(W) * \frac{\prod_i R1(w_{B_i})}{\prod_i R2(w_{B_i})} \qquad (式6)$$

5．形態素の分類

すべての単語について、別々に確率情報を獲得するのは困難である。そこで、形態素を語構成により品詞情報と表記情報で分類し、その分類ごとの確率情報を考える。ただし、第一段階だけはスコアを手作業で与えるため、分類を大まかにする。

5．1　品詞情報での分類

品詞情報での分類は、ＩＣＯＴのＴＲＩＥ辞書で使われている形態素分類コードを元にして行った。ただし、ＩＣＯＴの分類カテゴリは、本研究の形態素カテゴリとは別である。分類数は、第一段階が 26 で、第二段階が 75 である。

5．2　表記情報での分類

表記情報としては、形態素を構成する文字種と文字数を用いて分類した。日本語の文章で使われる文字種には、次のようなものがある。

・漢字　・ひらがな　・カタカナ
・英字　・数字　・記号

漢字だけ、ひらがなだけの形態素は、文字数でさらに五種類に分類した。漢字かな混じりの形態素は、組合せを考えると、分類数がかなり増加する。そこでこれに関しては文字数は考えず、文字種も「漢字かな混じり」という一つの分類にまとめた。記号は、句読点や括弧など、特殊なものは分けて分類した。

それ以外の文字種で構成される形態素は、文字種切りするので、他の文字種と混ざる可能性はなく、文字数もあまり影響ないと思われるので、それぞれ単独に扱う。分類数は第一段階で 16 、第二段階で 20 である。

5．3　形態素の分類数

品詞情報と表記情報の二種類を用いて分類すると、全体としては品詞分類数と、文字種、文字数分類数の積の数だけ分類が存在する。形態素の分類数は、第一段階が 416、第二段階が 1500 である。

6．実験

形態素解析システムの評価を行うために、次のような実験を行った。また、出現確率を使わず、接続確率だけを用いて学習、解析をした場合との比較も行った。

6．1　実験方法

まず、第一段階のシステムで形態素解析を行った。その結果から、確率情報を獲得した。さらに学習を二回行い、それぞれ、確率情報を獲得した。

一回目の学習では、信頼性パラメータがスコアのため、補正処理は行わなかった。また、二回目の学習（第二段階システムでの一回目の学習）では、補正処理を行った場合と、行わない場合の二種類の学習を行った。三回目の学習（第二段階システムでの二回目の学習）では、補正処理を行った場合だけで学習を行った。各学習の解析対象は、朝日新聞の社説から任意に抽出した 7000 文である。

各学習で獲得された確率情報を用いて、形態素解析結果の精度を調査した。解析対象は各々、朝日新聞の社説から任意に抽出した 300 文を使用した。ここで用いた文は、学習に用いたコーパスには含まれていない。解析対象は、一文の最大文字数 123、最

小文字数 6 で、平均文字数は 43 である。

さらに、同じ実験環境で、接続確率だけを用いた場合（いわゆる bigram ）との、比較を行った。

6．2　実験結果

正解率の変化を表１に示す。これを見ると、出現確率と接続確率を両方とも使った場合、一回目の学習後は、学習前に比べて正解率が若干良くなっている。しかし、補正を行わずに二回目の学習を行うと、一回目の学習後よりも正解率が悪くなっている。補正を行った場合は、かなり良い結果が得られている。三回目の学習後は、さらに正解率が良くなっている。

接続確率だけを用いた場合も、大体同じような動きをしているが、正解率が若干悪い。また、三回目の学習後には、正解率が大きく下がっている。

表１　正解率変化の比較

		正解率X(%)	正解率Y(%)
学習前		90.7	90.7
一回学習後		92.4	91.8
二回学習後	a	91.2	90.9
	b	94.4	93.0
三回学習後	b	95.1	91.7

a：補正処理なし　　b：補正処理あり
X：出現、接続確率で解析
Y：接続確率だけで解析

7．考察

二回目の学習は、補正処理を行う場合と行わない場合の、二通り行った。この二つを比べると、補正処理を行った方が正解率が高くなっている。これより、補正処理の効果が示せた。

出現確率を用いるか否かで、正解率に差がでた。これは、接続確率だけを用いた方は、三回の学習では、確率値が固定化していないことに起因する。出現確率を用いた方は確率値が収束している。収束の度合いを調べるために、各学習段階でカテゴリごとの確率値の大きさで順位を付け、学習前後の順位の相関係数を調べた。その結果、三回目の学習前後の順位の相関係数は、出現確率が 0.99 接続確率が 0.80 であった。接続確率だけを用いた方は、相関係数が 0.25 であった。

このことから、出現確率を用いることによって、確率情報の収束を早めることができることが分かった。しかし、接続確率だけを用いた場合でも、さらに学習を繰り返せば、確率値が収束して、よい結果を得られる可能性がある。

8．おわりに

解析候補の選好を行うための確率情報の獲得と、それを用いた形態素解析を考えた。形態素解析には、次の方法を用いた。

- 学習した確率情報を用いて形態素解析の候補の選好を行った。
- 確率情報はタグなしコーパスから段階的学習により近似値的に学習した。
- 確率情報には、品詞のほかに表記情報（文字種と文字数）を取り入れた。
- 確率情報の獲得の際に補正処理を行った。

これらを行うための形態素解析システムを設計、実現し、確率情報の獲得を行った。さらに獲得した確率情報を利用した形態素解析システムを実現した。実現したシステムを用いて形態素解析を行った結果、三回の学習によって正解率を 90.7 %から 95.1 %に向上させることができた。

参考文献

[１] 中川聖一，日本語及び英語の確率言語モデルに関する検討，「自然言語処理における学習」シンポジウム論文集，pp.57-64，1994.

[２] 永田昌明，自然言語の確率モデルと統計的学習，「自然言語処理における学習」シンポジウム論文集，pp.17-24，1994.

[３] 小松英二，コスト最小法形態素解析のコストの学習方法，言語処理学会第１回年次大会予稿集，pp.89-92，1995.

[４] ＩＣＯＴ，ＴＲＩＥ構造辞書リリースのためのドキュメント，1992.

[５] 竹内孔一 松本裕治，ＨＭＭによる日本語形態素解析システムのパラメータ学習，情報処理学会研究報告 95-NL-108，pp.13-19，1995.

[６] 朴哲済，統計モデルによる日本語の形態素解析手法，情報処理学会研究報告 95-NL-109，pp.19-26，1995.

[７] Masaaki Nagata, A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, COLING-94, pp.201-207, 1994.