コスト最小法と確率モデルの統合による形態素解析

山下 達雄，松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{tatuo-y,matsu}@is.aist-nara.ac.jp

本研究では、コスト最小法に基づく日本語形態素解析システム『茶筌』で用いられる人手により与えられたコストと、品詞 bi-gram モデルに基づき品詞タグ付きコーパスから学習された確率パラメータを統合し新たなコストを作成するという手法を提案する。実験により、品詞タグ付きコーパスが少量の場合、新たに作成されたコストを用いることで他の手法に比べ高い精度が得られることが確かめられた。この手法を用いることで、品詞タグ付きコーパスの作成の際のブートストラップを効率良く行うことができる。

[キーワード] 形態素解析，コスト最小法，品詞タグ付きコーパス，学習

Integration of Minimum Cost Method and Stochastic Model for Morphological Analysis

Yamasita Tatuo, Matsumoto Yuji
Graduate School of Information Science, Nara Institute of Science and Technology

Morphological analysis is a process that assigns a part-of-speech tag to each word in a sentence. It is an inevitable process in natural language processing. So far, two major approaches have been studied. One is the hand-crafted rule-based approach. Human observes various linguistic phenomena, and compiles rules for coping with them. However, it becomes harder and harder for human to maintain the whole rules as more and more exceptional rules are added. The other is the corpus-based probabilistic approach. This approach guesses probabilities of part-of-speech tag sequences estimated from part-of-speech tagged corpus as parameters of morphological analyzer. This approach can easily achieve high accuracy, although it requires a large part-of-speech tagged corpus, which needs tremendous amount of manual labor.
To solve these problems, We propose an approach that integrates the hand-crafted rules and the probabilities estimated from a small-size corpus. We show that this approach effectively achieves.

[keyword] morphological analysis, tagger, minimum cost method, part-of-speech tagged corpus, learning

1 はじめに

形態素解析は、構文解析・意味解析などの高度な自然言語処理の基盤として重要な技術である。

日本語形態素解析において、今までに研究されてきた最適解の選択手法は大きく次の二つの流れに分けれられる。

一つは、人手により作成された制約や優先規則を用いた方法である。これは、人間がさまざまな言語現象をおおまかにとらえて、規則などの形に抽象化したものと言える。記述量や理解の容易さといった面において効率の良いものであるが、記述者の考慮外の現象がとらえきれていないことが多い。そのような例外的な現象を規則として追加・修正することによって、解析精度を向上させていくのが一般的である。しかし、それにつれて、規則は次第に複雑になり、保守・管理が人間の手には負えなくなってしまう。つまり、一つの規則を修正することによって他に与える影響が予測不能になり、全体的な精度をあげていくのが難しくなる。しかし、これら

の規則は今までの経験の蓄積であり、有効な資源と言える。このような方法に基づくシステムとして、Breakfast[1]・『茶筌』[2]・JTAG[3]・JUMAN[4] などが知られている。

もう一つは、品詞タグ付きコーパスから学習された確率パラメータを用いた方法であり、人手による手間・解析精度などの問題をある程度解決できる[5]。しかし、ある程度の精度を出すには大量の品詞タグ付きコーパスが必要となり、それらが整備されていない未開拓な分野には不向きである。

本研究では、未開拓な分野で高い解析精度を得るための手法として、少量の品詞タグ付きコーパスから学習された確率パラメータと有用な言語資源である人手により作成された制約や優先規則の統合による形態素解析を提案する。

2 最適解の選択手法

本研究では、人手により作成された制約や優先規則を用いた方法のうち、現在流布している日本語形態素解析システムの多くが採用している人手により与えられたコストを用いたコスト最小法に着目する。また、品詞タグ付きコーパスから学習された確率パラメータを用いた方法のうち、品詞 bi-gram モデルに基づく統計的学習により得られた確率パラメータを利用する方法に着目する。

どちらの方法も、コスト・確率パラメータへの値の与え方が異なるだけで、基本的には、同じアルゴリズム (ヴィテルビ・アルゴリズム) で、最適解の選択を行うことができる。

ここでは、コスト最小法及び品詞 bi-gram モデル[1] に基づく最適解の選択法について説明し、この二つの手法の関係について述べる。

2.1 コスト最小法

分割された各形態素間の接続と形態素そのものにコストを与えて、その合計が最小の解を優先するという手法である。

分割された各形態素間の接続に与えられるコストを連接コスト、形態素そのものに与えられるコストを形態素コストと呼ぶことにする。

コスト最小法を用いた日本語形態素解析システムでは、これらのコストを人手により与えることが多い。

例えば、図 1のように「名詞と判定詞の連接コストは 30」「名詞『さかな』の形態素コストは 100」というようにコストを設定する。この場合の最適解は、文頭から文末までの連接・形態素コストの合計が最小 (240) である「さかな:名詞、だ:判定詞、よ:助詞」となる。

実際には形態素一個一個にコストを与えるのは大変なので、品詞ごとに設定することが多い。

2.2 品詞 bi-gram モデル

ある品詞の次にある品詞が現れる確率 $p(t_i \mid t_{i-1})$、ある品詞のときにある形態素が現れる確率 $p(w_i \mid t_i)$ を用いて、これらの確率の積が最大になるパスを優先する手法である。確率の積 Pは以下のような式で表される。w_iは形態素、t_iは品詞を表す。

$$P(w_1, \cdots, w_n) \cong \prod_{i=1}^{n} p(w_i \mid t_i) p(t_i \mid t_{i-1})$$

図 2に例を挙げる。名詞の次に判定詞が現れる条件付き確率は p(判定詞 | 名詞) $= 0.03$、名詞というカテゴリー内で『さかな』という語が現れる確率は p(さかな | 名詞) $= 0.05$ となっている。この場合の最適解は、文頭から文末までの確率の積が最大 (4.5×10^{-7}) になる「さかな:名詞、だ:判定詞、よ:助詞」となる。

このような条件付き確率 (確率パラメータ) は、品詞タグ付きコーパスがあれば最尤推定によって簡単に求めることができる。例えば、ある品詞タグ付きコーパス中で、名詞が 100 回現れて、その名詞の直後に判定詞が 30 回現れたとすると

$$p(\text{判定詞} \mid \text{名詞}) = \frac{30}{100} = 0.3$$

と推定でき、また、名詞 100 回のうち「さかな」が 5 回現れたとすると

$$p(\text{さかな} \mid \text{名詞}) = \frac{5}{100} = 0.05$$

と推定できる。

2.3 二つの手法の関係

コスト最小法におけるコストの和は品詞 bi-gram モデルにおける確率の積 P の逆数の対数を取ったものとみなすことができる [6]。

$$\log \frac{1}{P} \quad = \quad -\log\Big(\prod_{i=1}^{n} p(w_i \mid t_i) p(t_i \mid t_{i-1})\Big)$$

[1] 実際には品詞 trigram モデルが用いられることもある (文献 [5] に詳しい) が、人手で trigram ルールを作成することは非現実的であり、本研究の目的である人手によるルールとの統合には適さないので採用しなかった。

連接コスト　名詞 — 判定詞 $= 30$,　名詞 — 助詞 $= 45$

形態素コスト　名詞『さかな』$= 100$,　助詞『よ』$= 10$

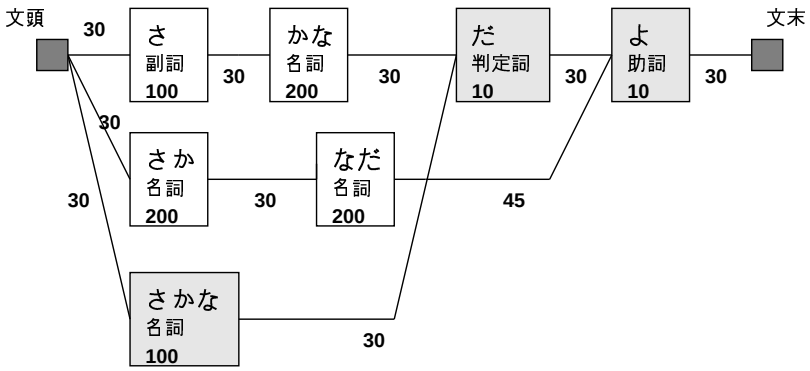

図 1: コスト最小法

p(判定詞 | 名詞) $= 0.3$,　p(助詞 | 名詞) $= 0.02$

p(さかな | 名詞) $= 0.05$,　p(だ | 判定詞) $= 1$

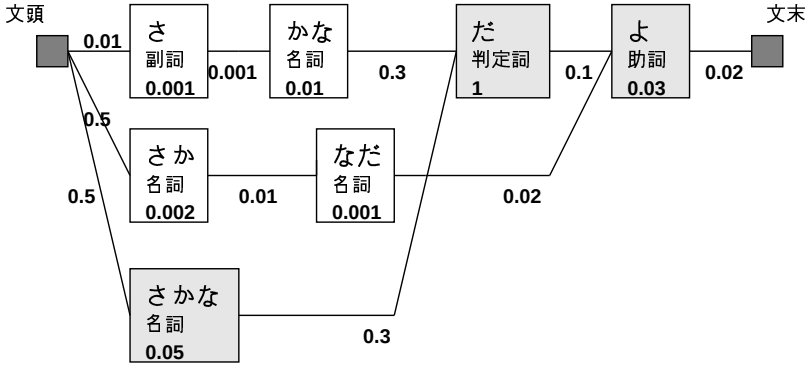

図 2: 品詞 bi-gram モデル

$$= \sum_{i=1}^{n}\Big(-\log p(w_i \mid t_i)\Big) + \sum_{i=1}^{n}\Big(-\log p(t_i \mid t_{i-1})\Big)$$

$-\log p(w_i \mid t_i)$ がコスト最小法における形態素コスト、$-\log p(t_i \mid t_{i-1})$ が連接コストに対応する。

これにより人手によるコスト体系と品詞タグ付きコーパスからの学習による確率パラメータを同じ土俵で扱うことが可能であることが分かる。つまり、確率パラメータをコストに変換するか、コストを確率パラメータに変換すれば良い。

3　コストと確率パラメータの統合

人手によるコスト体系とコーパスから学習された確率パラメータをある比率で混ぜ合わせ新たな値をつくり出し、それを最適解の選択に用いる。

どのような方法で統合すれば最適なのかは不明であるが、ここでは確率値の線形和という一番単純な方法を採用する (3.2節)。そのために両者を確率として扱うことにする。それには人手によるコストを

確率パラメータに変換することを考えなければならない (3.1節)。

処理の流れを図 3に示す。まず人手によるコストを確率パラメータへ変換する。次に二つの確率パラメータをある比率で統合する。そして最後にその新たな値をコストに変換する。

3.1 コストから確率パラメータへの変換

2.3節の議論により、確率パラメータの逆数の対数を取ったものがコスト最小法におけるコストに対応する。そこで、指数関数を用いてコストを確率パラメータに変換することを考える[2]。

コストは人間による恣意的な値であり、修正しやすいように比較的大きい値 (100 など) が用いられていることが多い。このような値を直接確率値に変換しても、極端に小さくなってしまい、確率値として直感に合わない。

そこで、まずコストのスケール変換を行う。最大のコストが最小の確率になるように係数を求めて、それを用いてスケール変換を行う。この係数を**確率化係数**α_pと呼ぶことにする。確率化係数の最適値については実験の章で述べる。

変換は式 (1) を用いて行う。人手によるコストC_hを確率化係数α_Pを用いてスケール変換を行ったコストを指数関数で確率パラメータ P_hに変換する。

$$P_h \quad = \quad \exp(-\alpha_p \times C_h) \qquad (1)$$

3.2 二つの確率パラメータの統合

人手によるコストを変換した確率パラメータ P_h と、コーパスからの学習による確率パラメータ P_c を以下の式を用いて足し合わせ、新たなパラメータ P_{new}を作り出す。λは混ぜ合わせの比率で**統合比率**と呼ぶ。

$$P_{new} = \lambda P_h + (1-\lambda)P_c \qquad (2)$$

図 3では、人手によるコスト 45 (名詞と助詞の連接コスト) を前節の方法で確率値 $0.04(= P_h)$ に変換し、それを品詞タグ付きコーパスから得た確率パラメータ $0.02(= P_c)$ と、1:3 の比率 ($\lambda = 0.25$) で足し合わせている。P_{new}は、0.025 となる。

統合比率の最適値については実験の章で述べる。

[2]コストから変換されたものは、厳密には確率値とは言えないが、ここでは、確率値として扱うことにする。

3.3 コストへの変換

本研究では実装にあたって、コスト最小法に基づく形態素解析システム『茶筌』[2] を使用しているため、確率パラメータをコストに変換する必要がある。確率パラメータの逆数の対数を取り、形態素解析システムに合わせてスケール変換を行う。

『茶筌』ではコストの範囲は、1〜255 であるので、図 3では、それに合わせて $-\log P_{new}$ をスケール変換している。

4 実験

前節で述べたように、本研究では、形態素解析システム『茶筌』により実装を行っている。

統合に用いる人手によるコストは、『茶筌』に付随の定義ファイルに記されているものを利用する[3]。

実験に用いた品詞タグ付きコーパスは、日経新聞 CD-ROM 94 年版の記事 1000 文 (約 30000 形態素) と ATR 経路探索課題コーパス 30 対話 (約 87000 形態素) で、共に当研究室において人手により品詞タグが付与された。これらは「書き言葉」「話し言葉」という異なる分野を代表するデータとみなせる。

日経新聞では 10 fold、対話コーパスでは 30 fold の cross validation を行った。

また、統合手法との比較のために、人手によるコストのみでの解析、品詞タグ付きコーパスからの学習結果のみを用いた解析も行った。

確率化係数及び統合比率は、経験上の最適値を用いた。学習に用いるデータの量によるが、確率化係数は 3〜11 、統合比率は 0.1〜0.001 の範囲内で良い精度が得られる。

図 4に実験結果を示す。図中の “rule” は人手によるコストのみでの解析、“probablistic” は品詞タグ付きコーパスからの学習結果のみでの解析、“integration” は本研究で提案した統合手法による解析を表す。横軸は、日経新聞では学習に用いた文の数 (一文約 30 形態素) を、ATR 経路探索課題コーパスでは学習に用いた対話数 (一対話約 3000 形態素) を表す。縦軸は、解析精度を表す。解析結果が正しいか否かの判断には「形態素の持つ全ての情報が等しい」という基準を採用した。これは、単語分割、読み、品詞情報の全てが一致していなければ、誤りとみなすという、大変厳しい基準である。

[3]連接コストは連接規則ファイルに記されているもの、形態素コストは形態素辞書に記されている形態素ごとの重みとリソースファイル (.chasenrc) に記されている品詞ごとの重みを掛け合わせたものを用いる。実装上の問題点については、文献 [7] を参照のこと。

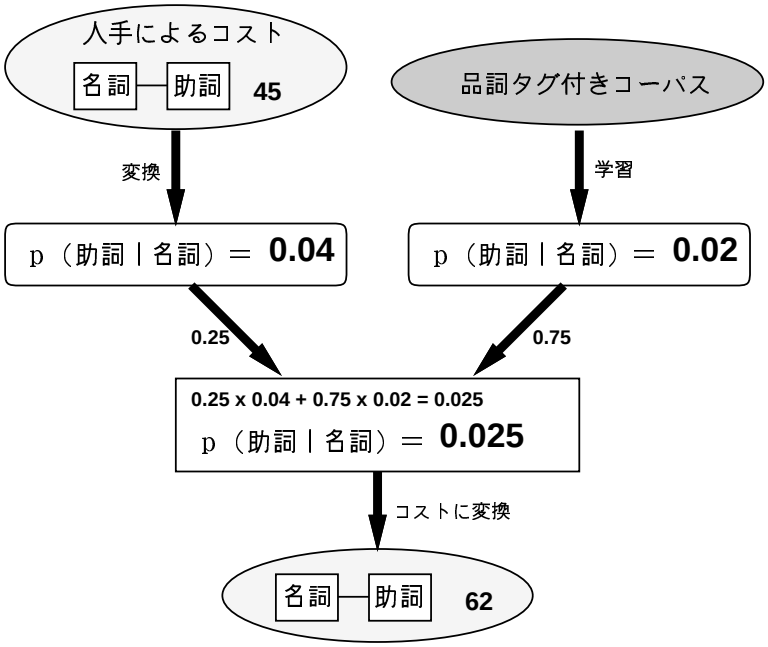

図 3: 統合処理の流れ

この実験により、統合手法は、人手作成コストによる手法や、同量の品詞タグ付きコーパスから学習された確率パラメータのみを利用する手法といった既存の手法と比べ、学習に用いる品詞タグ付きコーパスが小規模の場合において、解析精度が優位であることが示された。小規模な品詞タグ付きコーパスであっても、形態素解析の精度向上に有効に利用できることにより、次章で述べるようなブートストラップを効率良く行うことができる。

5 応用

適用分野を限定し形態素解析を行う場合、その分野に大量の品詞タグ付きコーパスが存在すれば、統計的学習により確率パラメータを獲得し、高い解析精度を得ることができる[4]。

しかし未開拓な分野ではそのような大量のコーパスの入手は不可能であり、自らその分野の品詞タグ付きコーパスを作成しなければならない。形態素解析の精度が低いと品詞タグ付け作業は非常に困難になる。

本研究で提案した手法を用いれば、そのような未開拓な分野においても小規模の品詞タグ付きコーパスを作成しさえすれば既存の手法に比べ高精度の解析が可能である。これにより、その分野の品詞タグ付きコーパスの作成が容易になる。つまり、ある程度の量の品詞タグ付きコーパスが整備される度に、そのコーパスを対象として統合手法を用い新たなコストを作成し、それを用いて形態素解析精度を徐々に向上させていくことができ、作業効率の向上につながる。

現在、このような品詞タグ付きコーパス作成支援環境の構築を文献 [8] で提案した形態素解析結果の視覚化システムを中心にして進めている。

[4]分野が異なると解析精度が低くなることは、新聞・対話の二つのデータで確認した。分野依存に関する最近の研究として文献 [9] を挙げておく。

謝辞

本研究では、ATR 経路探索課題コーパス、及び、日経新聞 CD-ROM 94 年版を利用させて頂きました。新聞記事データの研究利用許諾を頂いた日経新聞社に感謝します。

付記

『茶筌』に関する情報は以下の URL を御参照下さい。本研究で使用した学習プログラムも公開する予定です。

http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

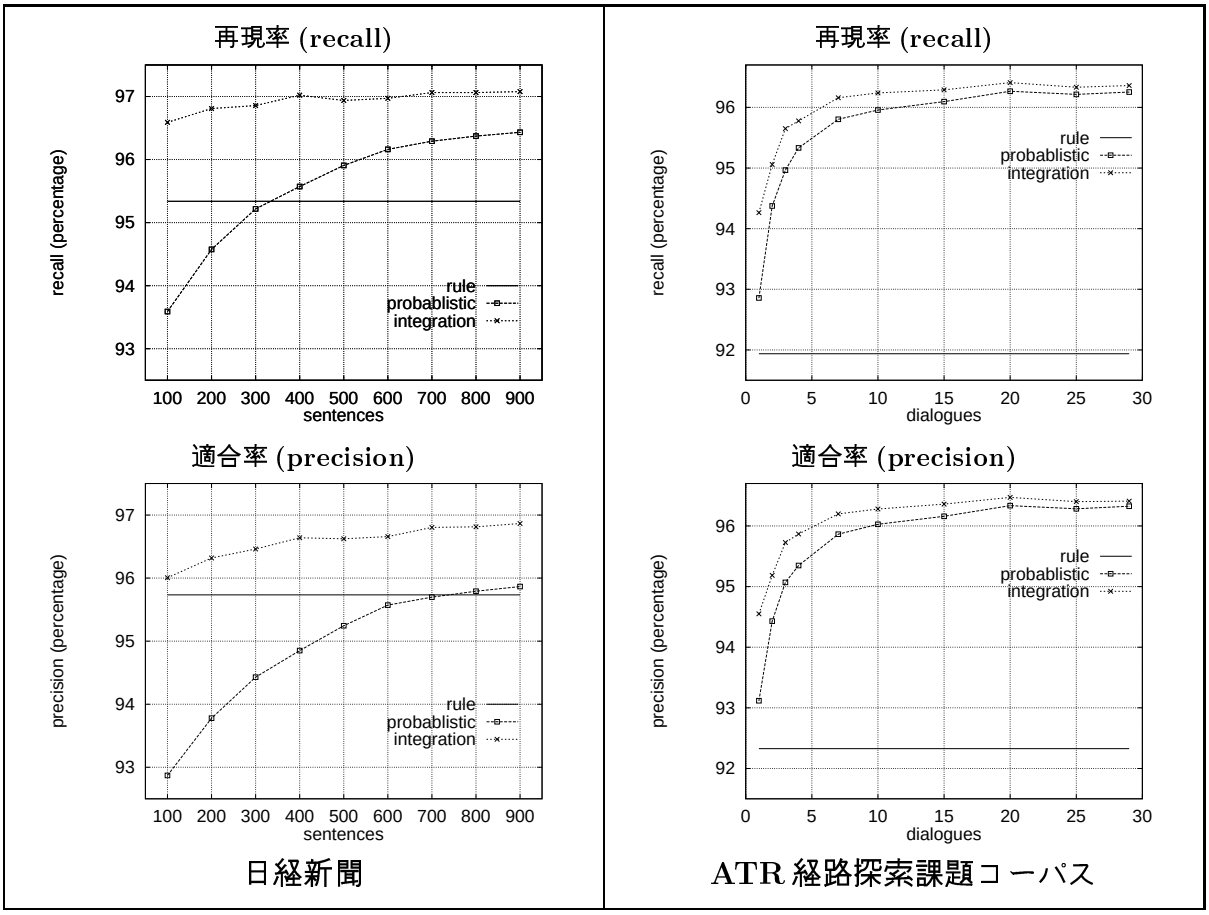

図 4: 実験結果

参考文献

[1] 颯々野学, 斎藤由香梨, 松井くにお. “アプリケーションのための日本語形態素解析システム”, 言語処理学会第 3 回年次大会発表論文集, pp. 441-444, March 1997.

[2] 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明. “日本語形態素解析システム『茶筅』version1.0 使用説明書”, NAIST Technical Report, NAIST-IS-TR97007, February 1997.

[3] 渕武志, 松岡浩司, 高木伸一郎. “保守性を考慮した日本語形態素解析システム”, 情報処理学会研究報告, 97-NL-117, pp. 59-66, January 1997.

[4] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真. “日本語形態素解析システム JUMAN 使用説明書 version 2.0”, NAIST Technical Report, NAIST-IS-TR94025, July 1994.

[5] 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—”, pp. 91-98, 森北出版, 1996.

[6] 長尾真編. “自然言語処理”, pp.117-137, 岩波書店, 1996

[7] 山下達雄. “規則と確率モデルの統合による形態素解析”, 奈良先端科学技術大学院大学修士論文, NAIST-IS-MT9551119, March 1997.

[8] 山下達雄, 松本裕治. “形態素解析結果の視覚化システム ViJUMAN とその学習機能”, 情報処理学会研究会報告, 95-NL-110, pp. 71-78, September 1996.

[9] Satoshi Sekine. “The Domain Dependence of Parsing”, Fifth Conference on Applied Natural Language Processing, pp. 96-102, 1997.