4AE-01 文字 n-gram データからの k-NN 法に基づく統計的形態素推定 *

永松健司　田中英彦†
東京大学大学院 工学系研究科

1 はじめに

日本語コーパスの整備に伴い，統計的情報を基にした形態素解析の研究が日本語に対しても多く行なわれている．しかし，分かち書きされないという日本語の性質のため，辞書引きによる形態素分割が前処理に入り，速度的に接続ルールとコストに基づく形態素解析手法を置き換えるには至っていない．

一方の接続ルールとコストに基づく形態素解析手法も，コストの調整や複数の接続ルール間での適用の順番などとの兼ね合いもあり，保守の点のオーバーヘッドも無視できなくなってきている．

これに対して，本稿では文字単位の n-gram データをコーパスから抽出し，そのデータから k-NN 法を用いて形態素分割および品詞属性の付加を行なう統計的形態素解析手法を提案する．評価の結果，不要なデータの刈り込み等を行なわない時点でも 93% 近くの精度と，従来の形態素解析プログラムの 1.5 倍，約 20000 文字 / 秒という処理速度が実現された．

2 利用する文字 n-gram データ

文字 n-gram データを用いると前処理としての形態素分割を行なう必要がなくなる．この考えに基づいて，[3] は文字 n-gram データをマルコフモデルによる統計的形態素解析に適用した．しかし，この方法では，処理時間が n の冪乗に比例するため実用的ではない．

そこで本手法ではより単純な k-NN 法を利用するために，n-gram データもマルコフモデルで用いられる n-gram データとは異なる形式のものを使用する（図 1）．ここで特徴的なのは，n-gram データを構成する各々の切れ目位置と各文字に確率情報を付与している点である．

本研究では，この n-gram データ ($n = 1 \sim 4$) を EDR コーパス全文 207,802 文 [1]（表 1）と評価用の 1000 文を除いた 206,802 文からそれぞれ抽出して次節で述べる形態素推定を行なっている．

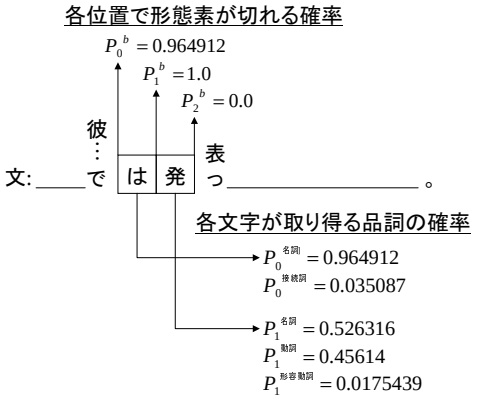

図 1: 本研究で利用する n-gram データの形式

n-gram	1-gram	2-gram	3-gram	4-gram
n-gram 総数	9,448,229	8,799,431	8,172,812	7,724,156
n-gram 種類数	5710	345,108	1,572,361	3,319,892
種類当りの頻度	1654.68	25.498	5.1978	2.3266

表 1: EDR コーパス全文から抽出された n-gram

3 k-NN 法による形態素推定

本研究で提案する k-NN 法による形態素推定は以下の手順で行なわれる．

1. 入力文中の各位置 i を含む全部分文字列を n-gram データから検索し，一致文字数を近さと見なして最も近いデータ (nearest neighbors) を選択する．
2. 選択された n-gram データ中で位置 i に対応する位置で形態素が切れる確率を基に，入力文の位置 i で形態素が切れる / 切れないを決定し，それらの多数決により，その位置で形態素が切れるかどうかを決定する．
3. 2. と同様の多数決により，各文字に与える品詞属性を決定する．
4. 切れると決定された位置の間の文字列を形態素として取り出し，その中の文字に与えられた品詞属性中，最大確率を持つ属性をその形態素の品詞属性とする．

以上の手順は k-NN 法の $k = 1$ の時に相当する．また，データ間の近さを上述のように定めることで，通常の k-NN 法で問題となる nearest neighbors の検索を辞書検索と同様のアルゴリズムで検索できるようになる．

---

*A Stochastic Morphological Analysis for Japanese employing Character n-Gram and k-NN Method

†Kenji Nagamatsu　Hidehiko Tanaka
{naga, tanaka}@mtl.t.u-tokyo.ac.jp
Faculty of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo, 113, Japan

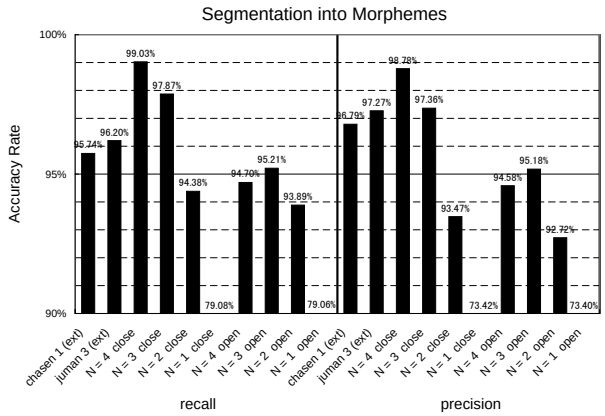

図 2: 形態素文字列の一致度

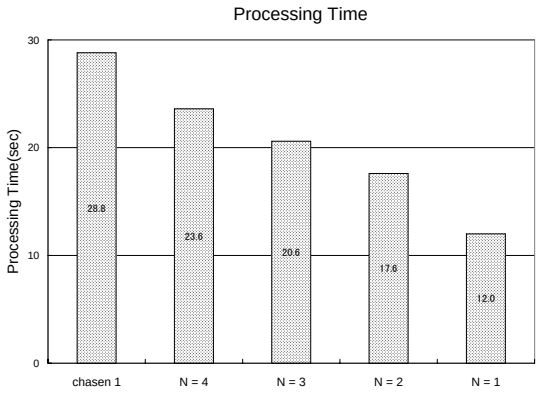

図 4: 10000 文当りの解析時間

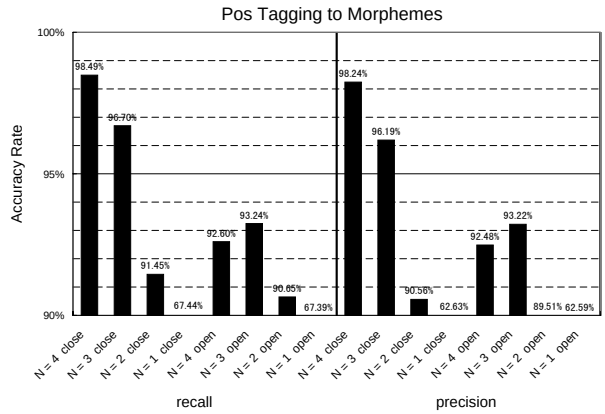

図 3: 品詞属性まで含めた一致度

4 評価

このアルゴリズムによる形態素推定を，EDR コーパスを元にして評価した．使用した n-gram データは前節に示したもので，評価用に残した 1000 文を解析して結果が EDR コーパスの正解と完全に一致するかどうかで，「形態素文字列としての一致度」と「品詞属性まで含めた一致度」を求めた．

また，利用する n-gram データを「4-gram まで」「3-gram まで」「2-gram まで」「1-gram だけ」の 4 通りの状況 ($N = 1,2,3,4$) を設定し，それぞれ評価文が既知となる n-gram データ (close) と，未知となる n-gram データ (open) を用いた場合の計 8 通りに対して評価を行なった．

評価結果から「形態素文字列としての一致度」を図 2に，「品詞属性まで含めた一致度」を図 3に示す．juman3.0beta と chasen1.0[2] の結果 (ext) は直接比較できないため，人手で適切かどうかの判断を行なったものである．

図 4は close 評価用データを用いた場合に EDR コーパスの先頭 10,000 文 (404,994 文字) を解析するのに要した時間 (user time) を求めたものである．従来のプログラムとの比較のため，chasen1.0 に対しても同様の設定の下で評価した．

5 考察

k-NN 法の性質から予想される通り，close 評価の下では利用する n-gram データの種類 (N) が大きくなる程，解析精度も上がることが示されている．

一方，open 評価の下では $N = 3$ の時に最も精度が高くなるが，これは手順 2. において n-gram データの左右端の情報が採用された場合に精度が下がることに依る．4-gram において，それが最も顕著に現われたためだと思われる．

処理速度の点を見ると，open 評価で最も高い精度が得られる $N = 3$ の場合に，約 20000 文字 / 秒の速度で解析でき，これは接続ルールとコストに基づく解析プログラムである chasen 1.0 の約 1.5 倍となる．

k-NN 法の研究では，事例の刈り込みが処理速度と分類精度の向上に寄与することが知られている．本研究の課題としても，現在，コーパスから抽出されたまま使用している n-gram データから不要なものを刈り込むことで，更なる解析精度と処理速度の向上が期待される．

6 おわりに

本稿では，文字 n-gram データから k-NN 法により形態素分割および品詞属性の付加を行なう形態素解析手法を提案した．評価の結果，k-NN 法という単純な方法を用いるだけでも 93% という解析精度と，k-NN 法の単純さによる 20000 文字 / 秒という高速な解析速度が実現された．

参考文献

[1] J. Electronic Dictionary Research Institute Ltd. Edr electronic dictionary technical guide, 1995.

[2] Y. Matsumoto, A. Kitauchi, T. Yamashita, O. Imaichi, and T. Imamura. *Japanese Morphological Analysis System ChaSen Manual.* Nara Institute of Science and Technology, 1997. NAIST Technical Report NAIST-IS-TR97007.

[3] 山本, 増山. 品詞・区切り情報を含む拡張文字の連鎖確率を用いた日 本語形態素解析. 言語処理学会 第 3 回 年次大会 発表論文集, pp. 421–424, mar 1997.