1C－2

単語と品詞の混合 n-gram を用いた形態素解析

山本 和英　河井 淳†　隅田 英一郎　古瀬 蔵
ATR 音声翻訳通信研究所
E-mail: {yamamoto,kawai,sumita,furuse}@itl.atr.co.jp

1. はじめに

本稿では、コーパスより作成した n-gram を用いた日本語形態素解析について述べる。英語の形態素解析ではすでに n-gram を用いた手法が行なわれてきており (例えば [Cha93])、その日本語への適用もいくつか行なわれている (例えば [Nag94])。また混合 n-gram モデルは [Kaw95] や [Mas96] などにより提案され、音声認識に適用されている。本稿ではこのうち [Kaw95] の混合 n-gram モデルを日本語形態素解析に適用し、その性能評価実験を行なった。

以下 2 節で、本稿で使用する混合 n-gram を用いた形態素解析手法の概要、3 節で同手法の評価実験の結果を報告し、4 節で本稿のまとめを行なう。

2. 混合 n-gram による日本語形態素解析

形態素解析の手法として、我々は統計的手法を用いる。コーパスから単語、及び品詞の n-gram 出現頻度を学習し、その連接確率を用いて形態素解析を行なう。

2.1 品詞要素と単語要素の設定

統計モデルによる形態素解析では、周辺の語 (あるいは品詞) との共起によって単語分割および品詞付与の尤度を推定している。ここで、実際に作られた n-gram を観察すると、前後の語の連接状況がほぼ同一であると見做すことのできる語群と、同一品詞であっても接続関係が個別に異なると考えられる語群に大別できることに気付く。例えば、ある二つの一般的な普通名詞の近傍の語の連接状況は多くの場合似通っており、その一方で助動詞に対する接続はそれぞれの語によって大きく異なっている。また以上の観察結果は、我々の直観とも一致する。

以上の考察から、我々は単語群をその品詞によって 2 種類に分類する。一つは品詞単位で連接状況を記述することのできる、つまり品詞に抽象化できる語群でこれらを品詞要素群と呼び、もう一つは単語単位で連接を考慮する必要がある助詞などの語群で、これらを単語要素群と呼ぶ。単語要素群、品詞要素群に記述する品詞には、活用形も考慮する。

どの品詞を品詞要素、あるいは単語要素にするかは品詞体系に依存するが、本稿で行なう以下の実験では品詞要素と単語要素を表 1に示すように分類した。

表 1: 品詞要素と単語要素に該当する品詞

品詞要素	普通名詞、サ変名詞、固有名詞、代名詞 数詞、形容名詞、連体詞、副詞、感動詞 本動詞、形容詞、間投詞、記号
単語要素	格助詞、終助詞、並立助詞、接続助詞 係助詞、副助詞、準体助詞、接続詞 接頭辞、接尾辞、助動詞、補助動詞

Morphological analysis utilizing n-gram of mixed category.
Kazuhide YAMAMOTO, Jun KAWAI, Eiichiro SUMITA and Osamu FURUSE.
ATR Interpreting Telecommunications Research Laboratories.

†現在、(株) 東洋情報システム。

2.2 混合 n-gram の定式化

入力文の単語列 $W = W_1, W_2, \cdots, W_k = W_1^k$、品詞列 $T = T_1, T_2, \cdots, T_k = T_1^k$ としたときの n-gram における単語列と品詞列の同時出現確率 $P(W,T,n)$ を以下の式によって定義する。

$$P(W,T,n) = \prod_{i=n}^{k} \left\{ P(E_i | E_{i-n+1}^{i-1}) \times \frac{P(W_i)}{P(E_i)} \right\} \quad (1)$$

ここで要素 E_i は以下のように定義する。

$$E_i = \begin{cases} W_i & : E_i \text{が単語要素の時} \\ T_i & : E_i \text{が品詞要素の時} \end{cases} \quad (2)$$

ただし、W_i: 該当単語、T_i: 単語 W_i と同一品詞の単語である。

以上のようにして定式化した混合 n-gram の一例 (2-gram の一部) を表 2に示す。

表 2: 混合 2-gram の例 (⟨···⟩ 内は品詞 / 活用形)

E_i	E_{i+1}	頻度
⟨連体詞⟩	⟨普通名詞⟩	680
⟨本動詞 / 連用⟩	ます⟨助動詞 / 終止⟩	1348
お⟨接頭辞⟩	⟨本動詞 / 連用⟩	1966
でしょ⟨助動詞 / 未然⟩	う⟨助動詞 / 終止⟩	907
に⟨格助詞⟩	⟨本動詞 / 連用⟩	1122
···	···	···

2.3 接続表の導入

データのスパース性に対処するためにこれまで種々の平滑化手法が提案されているが、本手法では品詞 2-gram の出現情報からなる接続表を用意する。接続表は混合 n-gram において存在しない場合に参照され、その接続が接続表にある場合は小さい定数をその n-gram 確率として代用する。

3. 性能評価実験

3.1 実験条件

以上の本手法の性能を確認するため性能評価実験を行なった。n-gram モデルとして 2-gram を用い、2-gram の訓練と評価のテスト用のタグ付コーパスとして ATR 対話データベース [Eha92] の国際会議ドメインのデータを採用した。これより無作為抽出した 10000 文を学習文集合とし、この中からさらに無作為抽出した 1000 文を既知のテスト文集合、学習文集合とは別に抽出した 1000 文を未知のテスト文集合とした[1]。また、このコーパスで使われている品詞の異なり数は 25 であり、活用形も含んだ異なり数は 52 である。

3.2 実験結果と考察

実験の結果を表 3に示す。表 3で、再現率 (recall)、適合率 (precision) は最尤候補の出力が正解と形態素単位でどの程度一致したかを示し、文正解率は入力文と形

[1]ただし、未知語の影響を取り除くため、単語辞書は学習文集合と未知のテスト文集合の両者から作成した。

表 3: 実験結果

			再現率	適合率	文正解率
単語分割精度	混合 2-gram	既知文	99.15%	99.48%	93.2%
		未知文	99.13%	99.38%	94.7%
単語分割及び品詞付与精度	混合 2-gram	既知文	98.32%	98.65%	83.1%
		未知文	97.70%	97.95%	81.3%
	品詞 2-gram	未知文	88.01%	88.45%	44.0%
	単語 2-gram	未知文	96.28%	95.96%	69.3%

表 4: 混合 2-gram と単語 2-gram、品詞 2-gram における主な誤り

正解	出力結果	単語	混合	品詞
なっ〈本動詞 / 連用〉	なっ〈補助動詞 / 連用〉	18	18	0
です〈助動詞 / 終止〉	です〈助動詞 / 連体〉	18	1	2
の〈格助詞〉	の〈終助詞〉	14	0	0
いらっしゃい〈本動詞 / 連用〉	いらっしゃい〈補助動詞 / 連用〉	9	0	0
また〈副詞〉	また〈接続詞〉	9	4	3
と〈格助詞〉	と〈並立助詞〉	4	11	0
で〈助動詞 / 連用〉	で〈格助詞〉	2	8	14
事務〈普通名詞〉局〈普通名詞〉	事務局〈固有名詞〉	2	8	6
知ら〈本動詞 / 未然〉せ〈助動詞 / 連用〉	知らせ〈本動詞 / 連用〉	0	5	10
お〈接頭辞〉	お〈普通名詞〉	0	0	50
ます〈助動詞 / 連体〉	ます〈助動詞 / 終止〉	1	0	48
私〈代名詞〉	私〈普通名詞〉	0	0	46
ください〈補助動詞 / 命令〉	ください〈補助動詞 / 連用〉	1	0	35
と〈並立助詞〉	と〈格助詞〉	6	1	18

態素解析出力が完全に一致した文の割合を示す。また参考のため、本手法に関しては単語の分割位置のみを比較したときの精度も示した。

実験の結果、未知文集合に対する実験の単語分割及び品詞 (活用形を含めた) 付与精度がそれぞれ再現率 97.70%、適合率 97.95% となった。この結果は、品詞 3-gram を用いて、同一のコーパス、ほぼ同一の規模で実験を行なった [Nag94] での結果と比較して、再現率で 2.6 ポイント、適合率で 3.4 ポイントだけ高い結果となった。今回の実験では未知語モデルを組み込んでいないため正確な比較はできないが、以上の結果から、本モデルは品詞 3-gram とほぼ同等、もしくはそれ以上の性能であることが期待できる。

また比較実験として単語 2-gram と品詞 2-gram も同一の学習文集合、テスト文集合 (未知文) と接続表で実験した。その結果、表 3に示すように品詞 2-gram は再現率及び適合率が本手法よりも 10 ポイント程度低い結果となった。単語 2-gram は混合 2-gram に近い再現率、適合率を示したが、文正解率が大きく異なった。

単語 2-gram、混合 2-gram、品詞 2-gram のそれぞれにおいて高頻度誤り 5 語を比較して表 4にまとめた。この表より、単語 2-gram で誤った主な語は (「なっ」を除いて) 混合 2-gram で減少していることがわかり (上段)、品詞 2-gram で誤った主な語も混合 2-gram ではほとんど誤っていないことがわかる (下段)。以上の比較実験により、混合 2-gram による形態素解析は単語 2-gram 及び品詞 2-gram よりも高性能であることが確認できた。

4. まとめと今後の課題

本稿では単語と品詞の混合 n-gram のモデル [Kaw95] の日本語形態素解析への適用について報告した。混合 2-gram による実験を行ない、単語 2-gram や品詞 2-gram よりも優れていることを確認した。またこのモデルは、文献 [Nag94] での品詞 3-gram と比較して、ほぼ同等もしくはそれ以上の性能であることが期待できる。

今後は、同手法のさらなる高精度化及び多言語化に取り組んでいく予定である。

参考文献

[Cha93] CHARNIAK, E.: *Stastical Language Learning*, Chapter 3, pp. 39–52, The MIT Press (1993).

[Eha92] 江原暉将, 小倉健太郎, 篠崎直子, 森元逞, 樽松明：電話またはキーボードを介した対話に基づく対話データベース ADD の構築, 情報処理学会論文誌, Vol. 33, No. 4, pp. 448–456 (1992).

[Kaw95] KAWAI, J., WAKITA, Y., and IIDA, H.: Stochastic language model using semantic category and mixed category of words and parts-of-speech for speech understanding, In *Proc. of NLPRS'95*, pp. 107–111 (1995).

[Mas96] MASATAKI, H. and SAGISAKA, Y.: Variable-Order N-gram Generation by Word-Class Splitting and Consective Word Grouping, In *Proc. of ICCASP 96*, pp. 188–191 (1996).

[Nag94] NAGATA, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, In *Proc. of Coling 94*, pp. 201–207 (1994).