自然言語処理 124－7
（1998．3．13）

枝分かれモデルによるbi-gramからのtri-gram推定を用いたHMM形態素解析システム

木村 健　　藤本 浩司　　乾 伸雄　　小谷 善行

平成10年3月13日

東京農工大学工学部電子情報工学科
〒184　東京都小金井市中町2-24-16
{kimura,kfujim,nobu,kotani}@cc.tuat.ac.jp

要約

品詞遷移確率を用いた形態素解析システムにおいて，コーパスから獲得するデータの精度は大変重要である。特にn-gramの次元が上がると，データがスパースなってしまい，十分な解析精度が望めない。

本論文では，枝分かれモデル[藤本 98]を用いて品詞bi-gramから品詞tri-gramの頻度を推定し，既存のtri-gramがあればこれをχ^2検定して，HMM形態素解析システムにおいて利用する方法について述べる。

d-bigramを含む品詞bi-gramについて，二次の交互作用がないものと仮定して三次方程式を立て，カルダノの法より実数の解を得る。形態素解析システム側からは，tri-gramを求める部分だけが変わるだけであり，解析精度の向上が見込まれる。

HMM Morphological Analyzer Using Tri-gram Estimation From Bi-gram Nest Structured Transition Probability Model

Takeshi KIMURA, Koji FUJIMOTO, Nobuo INUI, Yoshiyuki KOTANI

Friday 13, March, 1998.

Department of Computer Science
Tokyo University of Agriculture and Technology
2-24-16, Nakamachi, Koganei City, Tokyo 184, Japan.
{kimura,kfujim,nobu,kotani}@cc.tuat.ac.jp

Abstract

It is very important for Morphological Analyzer based on the HMM model by using statistical data from large corpuses to hold high accuracy of statistical data. This paper proposes a more efficient morphological analysis which takes the ordinary HMM model as well as bi-gram nest structured transition probability model. It enables the HMM model to receive two data, one from the corpus, and the other from the nest structured model, and determines to select more appropriate one to calculate the probability. For Part-Of-Speech bi-gram containing d-bigram, by using the cubic equation by Cardano method, it is possible to make the analysis more accurate.

1 はじめに

タグ付きコーパスから品詞（あるいは単語）n-gramを獲得する際，n-gram頻度データ精度は共起する単語の長さが増えるほど低下する。これは，得られるn-gramの頻度の和がほとんど変らないのに対し，出現可能なn-gramの種類が，nについて指数関数的に増加するためである。この問題については，線型補間法などによるn-gramのスムージング[Charniak 93]，エントロピー最大法によるbigramからのn-gram推定[江原 96]などの研究がある。

枝分れモデル[藤本 98]は，異なるクラス間での共起関係をn次元格子によって表現するモデルである。HMMとは違う遷移を仮定しているので，単語どうしの連接確率$P(W_j|W_i)$を使用することで，HMM形態素解析システムとは別のシステムを構築することができる。ところが，両システム間でのデータの交換は，単語出力確率という概念が枝分かれモデルに存在しないため，難しい。

本論文では，まず枝分かれモデルを用いて頻度が高く，より信頼のおけるbi-gramから，交互作用の特徴を用いて作られる三次方程式により，tri-gramを推定する。さらに推定されたtri-gramをχ^2検定し，得られたtri-gramと，(仮に存在すれば）もとからあったtri-gramのうち，どちらが統計的に信頼がおけるか調べ，HMM形態素解析システムにおいて連接確率として使用するtri-gramを決定し，HMM形態素解析システムを大幅に改良することなく解析精度を上げる。

本方法によりHMM形態素解析は，推定することで求められた確率値を，解析系から矛盾のないように利用することができる。

2 枝分かれモデル

枝分かれモデルは，単語を遷移の単位とし，次のような語のクラス階層を前提にしたモデルである。

品詞大分類（クラス1）
→ 品詞小分類（クラス2）
… → 単語（異なり語，クラスn）

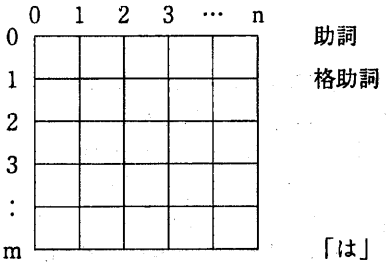
図 1: 単語のクラス

例：

名詞 → 固有名詞 → 人名 → 太郎

クラスの重なりはないものとする。つまりすべての品詞（単語）が，最も詳細な品詞（あるいは単語の表記そのもの）を葉とし，重複のないクラスのノードからなる木の，すべてのクラス階層をただ一回だけ通過する，一意な道(path)によって定義される。

品詞がクラスによって細分化され，さらに最下層に単語の表記を持つn-gramモデルを考える。このモデルには，クラスの違いにより様々なn-gram頻度データが存在する。枝分かれモデルで記述すると，次のように表わすことができる。



名詞，固有名詞，人名，「太郎」

表 1: 枝分かれモデル

表1のすべての交叉は頻度を持つ。本論文では，クラス階層の深さがそれぞれm,nの任意のbi-gramを，(m,n)と表わすことにする。ほとんどは異なるクラス間での品詞bi-gramである。最下層の場合は，単語bi-gramになる。

なお，例外として$(l,0)=(0,l)=(l)$とする。lはクラス階層の深さがlの任意のuni-gramデータである。

ここで，語が共に出現する事象の独立であるか否かを示す「交互作用」という数量を導入し，様々なクラス間でのn-gramを表現する。本節では，一般的な説明に入る前にbi-gramに対象を絞り，後にこれをtri-gramに拡張する。

2.1 交互作用

与えられたbi-gramに前後関係や共起の関係があると，そのbi-gramはuni-gramから直接推定

することができない。逆にuni-gramと同等の情報しか持たず，それぞれが独立な事象であれば，そのbi-gramはuni-gramから求められる。

交互作用$I(i,j)$は，(i,j)に含まれる特定のbi-gramについて，bi-gramが，uni-gramの積と等しいかどうかを示す量である。

$$I(i,j) = \frac{(i,j)(j-1,j-1)}{(i,j-1)(i-1,j)} \tag{1}$$

独立であれば，bi-gramの交互作用は$I(i,j)=1$である。

交互作用を用いて(i,j)を表わすことができる。

$$(i,j) = \frac{(i,j-1)(i-1,j)}{(j-1,j-1)} I(i,j) \tag{2}$$

さらに頻度の項を展開することで，同一の頻度に対し，様々な表現が可能である。

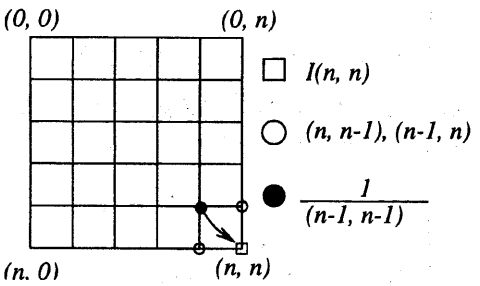


図 2: (n,n)を$I(n,n)$を用いて求める

仮に一次交互作用が1であれば，頻度は他の頻度データから与えられることになる。bi-gramの交互作用の式を図示したものを図3に示す。

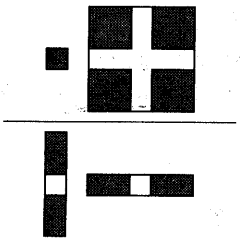


図 3: $I(i,j)$

2.2 二次交互作用

前記の式1の$I(i,j)$は，交互作用の式が一次式であるため多少省略をしてある[1]。tri-gramの場合，二次交互作用も考える必要がある。tri-gramの場合，交互作用$I(i,j,k)$は，次の式となる。

$$I(i,j,k) = \frac{t_1\ t_2\ t_3\ t_4}{t_5\ t_6\ t_7\ t_8} \tag{3}$$

$$\begin{cases} t_1 = (i,j,k) \\ t_2 = (i-1,j-1,k) - t_5 - t_6 - t_1 \\ t_3 = (i-1,j,k-1) - t_5 - t_7 - t_1 \\ t_4 = (i,j-1,k-1) - t_6 - t_7 - t_1 \\ t_5 = (i-1,j,k) - t_1 \\ t_6 = (i,j-1,k) - t_1 \\ t_7 = (i,j,k-1) - t_1 \\ t_8 = (i-1,j-1,k-1) \\ \qquad -t_2 - t_3 - t_4 - t_5 - t_6 - t_7 - t_1 \end{cases}$$

tri-gramの交互作用の式を図示したものを図4に示す。図4の八つの立体の位置は，さきほどの式3の項の位置と対応している。

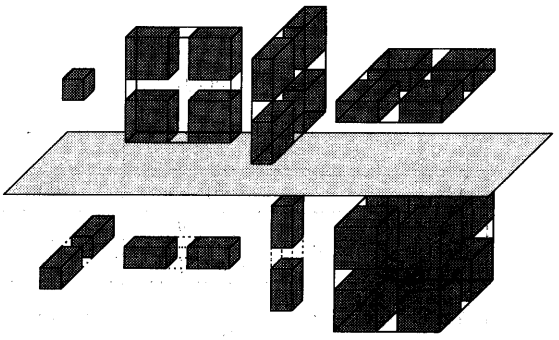


図 4: $I(i,j,k)$

このようにして，n-gramについても同様に交互作用を定義できる。

ここまで述べてきた式は，一つ上のクラスからtri-gramの交互作用を求めた。実は，一つ上のクラスだけでなく，さらにいくつも上のクラスにつ

[1]正しくは，図3のように，つまり

$$I(i,j) = \frac{(i,j)\{(i-1,j-1)-(i,j-1)-(i-1,j)+(i,j)\}}{\{(i,j-1)-(i,j)\}\{(i-1,j)-(i,j)\}}$$

になる。

いても同様に交互作用を求めることができる。なぜなら交互作用は,「自分のクラスと，それ以外のすべてのクラス」についての式なので,「自分以外のすべてのクラス」が一つ上でも，一番上でもよいからである。

次節に述べる bi-gram から tri-gram を推定する方法は，一番上のクラスから最も詳細な品詞のクラスの tri-gram について，交互作用を求めている。

3 bi-gram からの tri-gram の推定

d-bigram を含む uni-gram から tri-gram を推定する方法を述べる。この推定については，[江原 96] などの研究がある。

HMM 形態素解析システムでの連接確率を求めるために，最も詳細な品詞のクラスに属する tri-gram (abc) の頻度を推定する。まず，tri-gram のそれぞれ三つの語について，その語のクラスと，それ以外のすべてのクラスとの二つ分かれた，三次元の分割表を作成する（表 2)。

表 2: 分割表

	C	$\overline{\mathrm{C}}$	
AB	x	$(ab)-x$	(ab)
$\mathrm{A}\overline{\mathrm{B}}$	$(ac)-x$	$a-(ab)$ $-(ac)+x$	$a-(ab)$
$\overline{\mathrm{A}}\mathrm{B}$	$(bc)-x$	$b-(ab)$ $-(bc)+x$	$b-(ab)$
$\overline{\mathrm{A}}\,\overline{\mathrm{B}}$	$c-(ac)$ $-(bc)+x$	$N-a-b$ $-c+(ab)$ $+(bc)+(ac)$ $-x$	$N-a$ $-b+(ab)$
	c	$N-c$	N

未知の tri-gram (abc) を x とおき，他の七つの項を x を含む形の式で表現する。これらの項は，交互作用 $I(a,b,c)$ の項一つ一つに対応する。

次に，交互作用 $I(a,b,c)$ が 1 であるとして，次のような方程式を立てる。

$$\frac{(\mathrm{ABC})(\overline{\mathrm{A}}\mathrm{B}\overline{\mathrm{C}})(\mathrm{A}\overline{\mathrm{B}}\,\overline{\mathrm{C}})(\overline{\mathrm{A}}\,\overline{\mathrm{B}}\mathrm{C})}{(\overline{\mathrm{A}}\mathrm{BC})(\mathrm{A}\overline{\mathrm{B}}\mathrm{C})(\mathrm{AB}\overline{\mathrm{C}})(\overline{\mathrm{A}}\,\overline{\mathrm{B}}\,\overline{\mathrm{C}})} = 1 \qquad (4)$$

$(ABC)\ldots(\overline{ABC})$ には，表 2の対応する箇所の式が入る。これを展開すると，四次の項が消え，三次方程式になる。

$$\begin{aligned} &x(e_1+x)(e_2+x)(e_3+x) \\ &\quad -(h_1-x)(h_2-x)(h_3-x)(h_4-x) \\ &\quad =0 \end{aligned} \qquad (5)$$

ここで，

$$\begin{cases} e_1 = a-(ab)-(ac) \\ e_2 = b-(ab)-(bc) \\ e_3 = c-(ac)-(bc) \\ \\ h_1 = ab \\ h_2 = bc \\ h_3 = ac \\ h_4 = N-a-b-c+(ab)+(bc)+(ac) \end{cases}$$

である。式 5は，次の四つの係数を持つ三次方程式 $a_0x^3+a_1x^2+a_2x+a3=0$ である。

$$\begin{cases} a_0 = e_1+(e_2+e_3)+(h_1+h_2) \\ \quad +(h_3+h_4) \\ a_1 = e_1(e_2+e_3)+e_2e_3-h_1h_2 \\ \quad -(h_1+h_2)(h_3+h_4)-h_3h_4 \\ a_2 = e_1e_2e_3+h_1h_2(h_3+h_4) \\ \quad +h_3h_4(h_1+h_2) \\ a_3 = -h_1h_2h_3h_4 \end{cases}$$

この方程式をカルダノの法を用いて解くことにより，tri-gram 頻度は，未知数 x として得られる。カルダノの法の具体的な式等については，紙面の都合上割愛する。

tri-gram 以上の n-gram について同様の推定を行なう場合は，数値解法によって解を求める必要がある。

4 χ^2検定

前節で述べた tri-gram の推定によって得られたデータは，常によい結果をもたらすとは限らない。むしろ，tri-gram 頻度データとして既にコーパス

から得られた頻度の方がかえって質が高い可能性もある。

いずれかのうち一つを選択するため，χ^2検定を行う。期待値を表2の八つの値として，χ^2 値を求める。

$$\chi^2 = \sum_{i=0}^{7} \frac{\{X_i - Y_i\}^2}{Y_i} \tag{6}$$

$X_i(i = 0 \ldots 7)$ は x を推定した値とした分割表の値，$Y_i(i = 0 \ldots 7)$ は x をコーパスから得られた頻度とした分割表の値である。

低い危険率[2]で検定し，仮説（交互作用が，1.0）が棄却されるかどうかを見て，棄却されればコーパスから得られた頻度を使用する。

5 HMM 形態素解析システムへの応用

これまで述べてきた方法を利用することで，枝分かれモデルによって推定された tri-gram 頻度を，HMM 形態素解析システムで使用できる。一般的な HMM 形態素解析システムの構成を図5に示す。簡単のため，HMM 形態素解析システムは，単語出力確率と，品詞間遷移確率の二つだけを使用しているものとした。

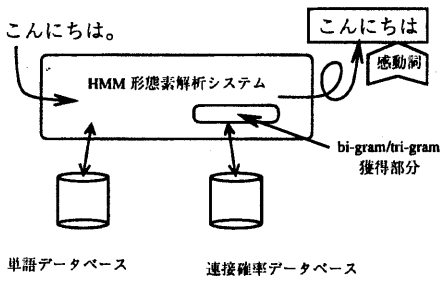


図 5: HMM 形態素解析システム

このうち，システムの bi-gram / tri-gram 獲得ルーチンを，次の図6のように変更する。

このようにして，HMM 形態素解析システムの bi-gram / tri-gram 獲得ルーチン以外のすべての部分を変更することなく，枝分かれモデルから推定して求めた tri-gram 頻度データを使用することができる。

[2][藤本 98] では 0.5%。

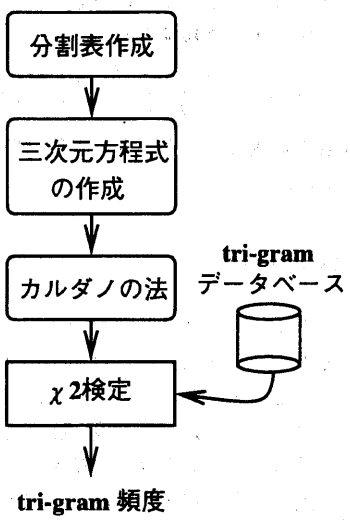


図 6: χ^2検定による tri-gram の選択

6 枝分かれモデルの次元拡張

bi-gram から tri-gram を推定する場合のように一番上のクラスから品詞を求めるのでなく，あらかじめ得られた tri-gram のうち，上のクラスの物を残しておき，そこから下のクラスの値を推定することもできる。

なお，仮に HMM 形態素解析システムの連接コストの計算部分について変更が可能であれば，単語を単位とする遷移確率より求められるコストを，単語出力コストと品詞遷移コストの和のかわりに置き換えることができる。χ^2 検定によって格子上の最適なパスが選択できるので，HMM よりも柔軟なクラス間の遷移関係が記述できる。遷移確率の式は，HMM の場合，

$$P(t_i|t_{i-1}) = \frac{(t_{i-1}t_i)}{t_{i-1}}$$

単語出力確率の式は

$$P(w_i|t_i) = \frac{(w_i t_i)}{t_i}$$

の，それぞれ一つであるのに対し，枝分れモデルの場合，単語間の遷移確率を

$$P(w_i|w_{i-1}) = \frac{(w_{i-1}t'_i)}{w_{i-1}} \frac{(t'_{i-1}w_i)}{(t'_{i-1}t_i)} \frac{(t_{i-1}t_i)}{(t_{i-1}t'_i)}$$

というふうに取ることができる。他の取りかたも可能である。

7 おわりに

本論文では，枝分かれモデルを用いて bi-gram より推定した tri-gram を，HMM 形態素解析プログラムに応用する方法を述べた。比較的大きな HMM 用のコーパスデータが既に存在し，使用する形態素解析システムが品詞を基本的な遷移の単位とする連接確率を使用する場合，本方法は有効である。

参考文献

[江原 96]
江原暉将．最大エントロピー法を用いてバイグラム確率からnグラム確率を求める，情報処理学会研究会報告，NL 113-5，1996.

[白井 96]
白井清昭，乾健太郎，徳永健伸，田中穂積．最大エントロピー法を用いた単語 bi-gram の推定，情報処理学会研究会報告，NL 116-4，1996.

[藤本 98]
藤本浩司，乾伸雄，小谷善行．枝分かれ構造をもつ連接確率モデルによる形態素解析，情報処理学会研究会報告，NL 123-1，1998.

[Charniak 93]
Eugene Charniak, Curtis Hendrickson, Neil Jacobson and Mike Perkowitz. Equations for Part-of-Speech Tagging, In the Proceedings of the Eleventh National Conference on Artificial Intelligence, 1993.

[Nagata 94]
Nagata,M. A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm. In Proceedings of COLING '94, pp. 201–207, 1994.