３．日本語のN-gramによるモデル化

- データ量に対するエントロピーおよびカバー率の変化
- 新聞記事
- X線CTレポート
- ATRの対話データ

学習データ量に対するEntropyの変化

	entropy		
Unigram	Σ i	p(wi)	log p(wi)
Bigram	Σ (i, j)	p(wi, wj)	log p(wj | wi)
trigram	Σ (i, j, k, l)	p(wi, wj, wk)	log p(wk | wi, wj)
4-gram	Σ (i, j, k, l)	p(wi, wj, wk, wl)	log p(wl | wi, wj, wk)

カバー率

カバー率Ｘ％の種類の数

入力データのＸ％をカバーするのに必要最小限の
マルコフ連鎖の種類の数

例:unigram　　データ（ａ ｂ ａ ｂ ａ ａ ｃ ｄ ）

カバー率　　100%の種類の数4 (a b c d)
カバー率　　75%の種類の数2 (a b)
カバー率　　50%の種類の数1 (a)

大蔵省はことし四月から新銀行法が施行されるのに伴い、在日外銀の営業活動を日本の銀行同様に扱うとの基本方針を決め、これを盛り込んだ政令を二月中にも公布する。おもな内容は（１）企業向け貸し出しに対する大口融資規制を在日外銀にも適用し、五年間の猶予期間を設けるなどの配慮をする（２）利益準備金の積み立てを義務づけ、外銀に対する信頼を高める（３）邦銀の支店を買収することや現地法人化を認める――など。大蔵省はこれによって在日外銀に関する法的根拠が明確になるほか、在日外銀の国内活動がしやすくなり、欧米諸国の間に出始めているわが国の金融制度に対する不満を和らげるのに役立つとみている。（在日外国銀行は「きょうのことば」参照）

新聞記事

新聞記事　　（74日分）

・コンピュータによる形態素解析
・コンピュータによる品詞および読み（音節）の付加
・文節単位

全データ量約170万文字（漢字かな）

音節
長音・促音を1音節と計算
記号・外国語読み・数詞は削除
種類の和　111音節

漢字仮名
漢字JIS1級、約3000文字

品詞
人名・地名などの意味的なカテゴリを含んで約450種類

頭部CT 単純および造影
１、３月１３日のCTと比較した。
２、スライスのレベルが若干異なっているので正確な比較はできないが、鞍上槽の正中からやや右上方へ向かって進展している増強効果を示す腫瘤の大きさは本質的に変わっていない。ただし前回のCTでこの結節性腫瘤の右前方に見られた嚢胞性の成分については今回は描出されていない。
３、側脳室の大きさ形も前回と同様である。

impression.....
鞍上槽の頭蓋咽頭腫の残存については明らかな変化はないが、右後方に見られた嚢胞性成分が消失しているかもしれない。

X線CT所見作成の例

X線CT所見作成の文章

Total　約25万文字
人手による文節区切り

音節

“mass effect”, “large magna”などの外来語が数多く出現
　音節の種類の数は118音節

漢字かな

　外来語はアルファベットの全角文字に変換
　（例 “mass effect” は “MASS EFFECT”）

単語

　語彙数　約3000
　ただし、出現率が高い　100文節は単語として
　登録　　（例　“脳実質を”）

・[あっ、あえーっと]そちら第１回の通訳電話国際会議の事務局でしょうか。
・はいそうです。
・[えーっとちょっと]その会議のことでねあのー登録のことでお伺いしたいんですが。
・はい。
・どうぞ。
・[えーっと]今手元にあの登録用紙があるんですけれども
　[えーっと]その中でちょっとあの
・クレジットカードをね[あのー]クレジットカードの名前となんかナンバーを書くところ
・があるんですがはいそうです。[えーっと]それをちょっとクレジットカードを
　　持っていない者がいるんですけれどもその場合はどうなんでしょうか。
・はい。

国際会議の対話例

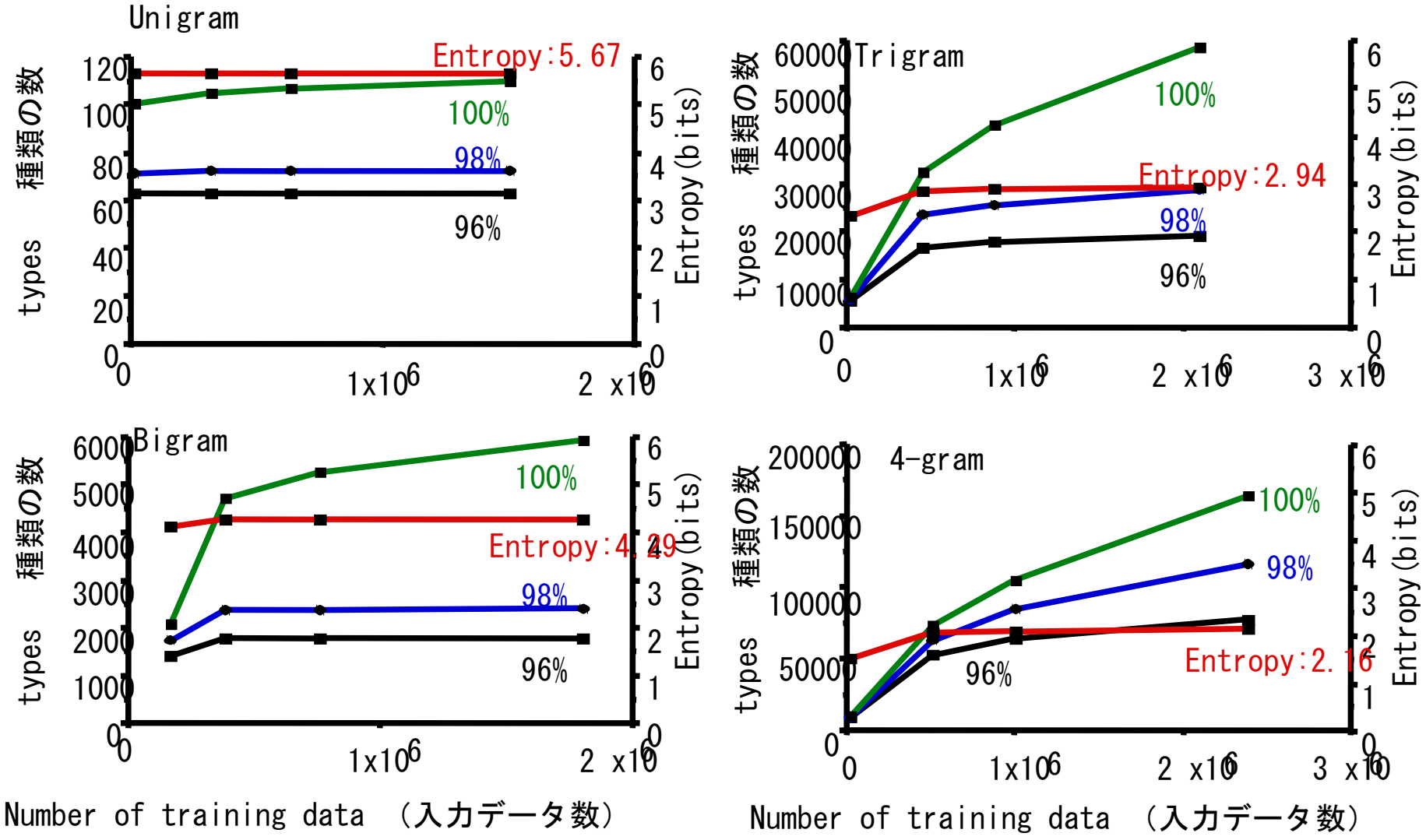
	unigram	bigram	trigram
音節	5.5	3.21	1.57
漢字かな	7.3	2.55	1.61
単語	8.13	3.75	2.61

X線CT所見作成(エントロピ)

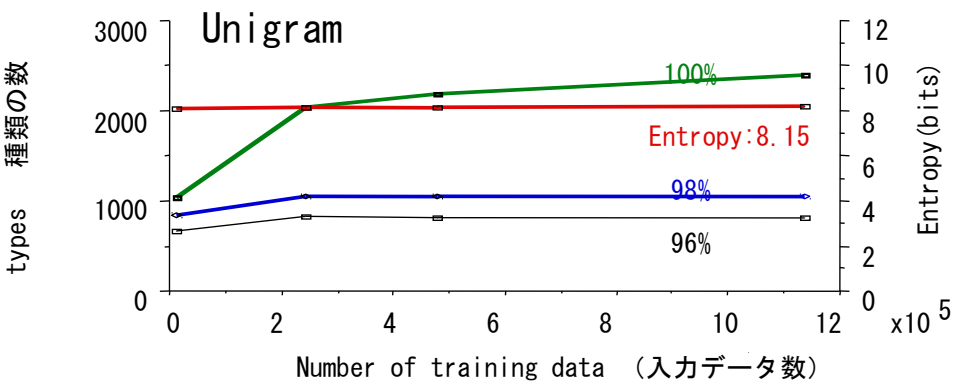
	unigram	bigram	trigram	4-gram
音節	5.67	4.29	2.94	2.16
漢字かな	8.15	4.45	2.87	2.29
品詞	5.57	2.69	2.03	1.63

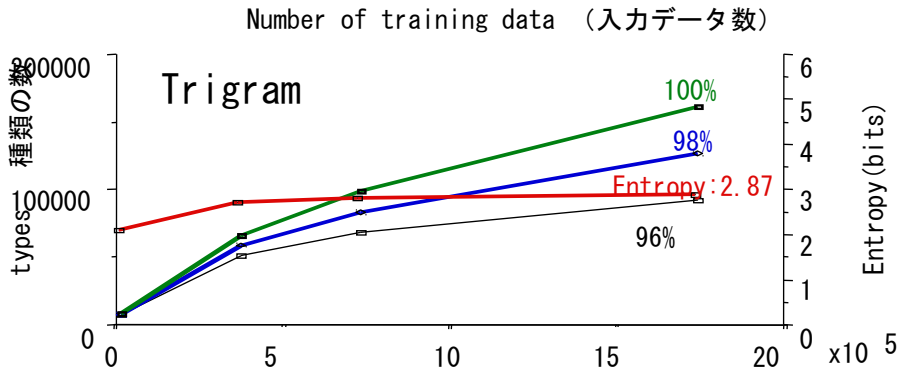
新聞記事(エントロピ)

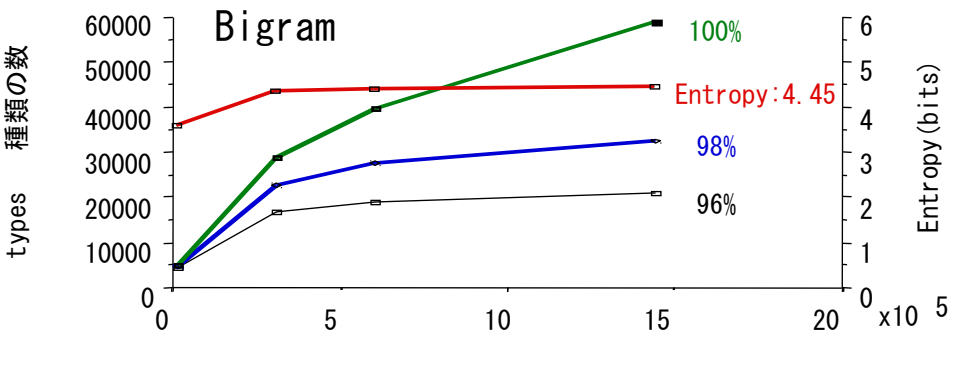
新聞記事とX線CT所見作成の比較
X線CT所見作成の文章は新聞記事と比較して単純

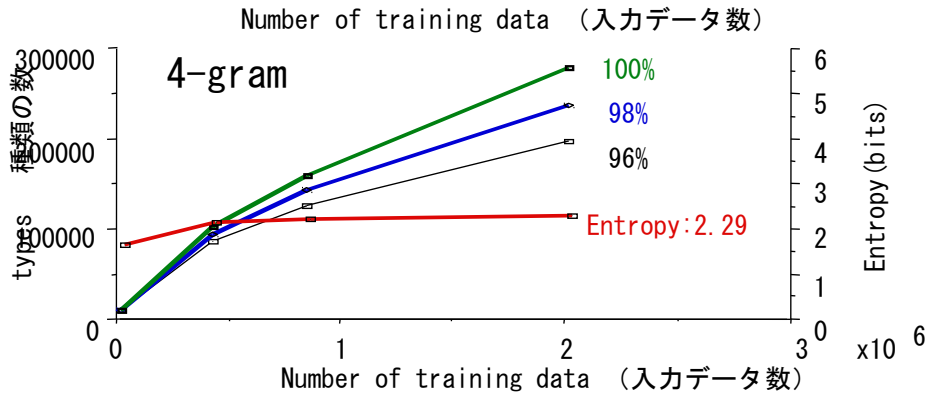


新聞記事における学習データ数にたいする音節の
マルコフ連鎖確率値のカバー率およびエントロピー

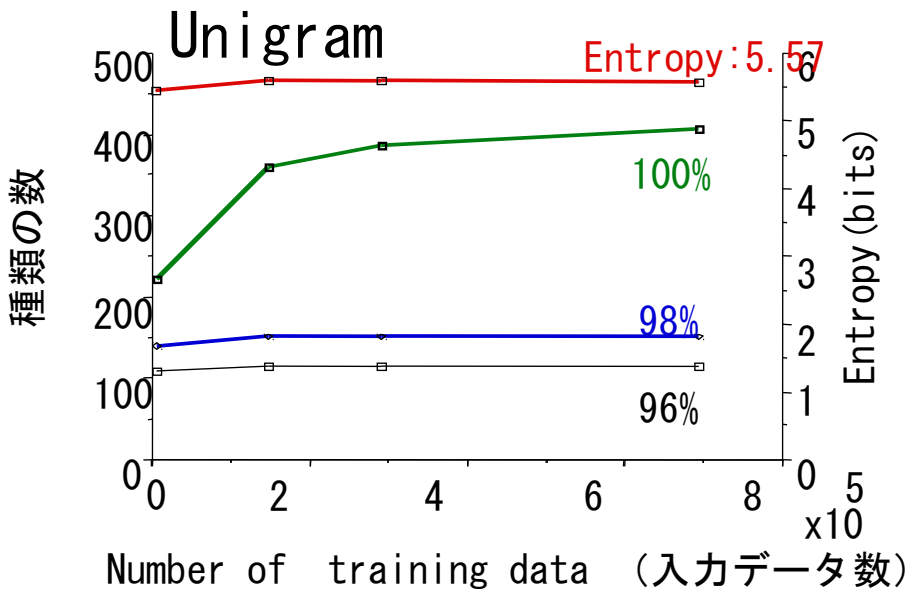


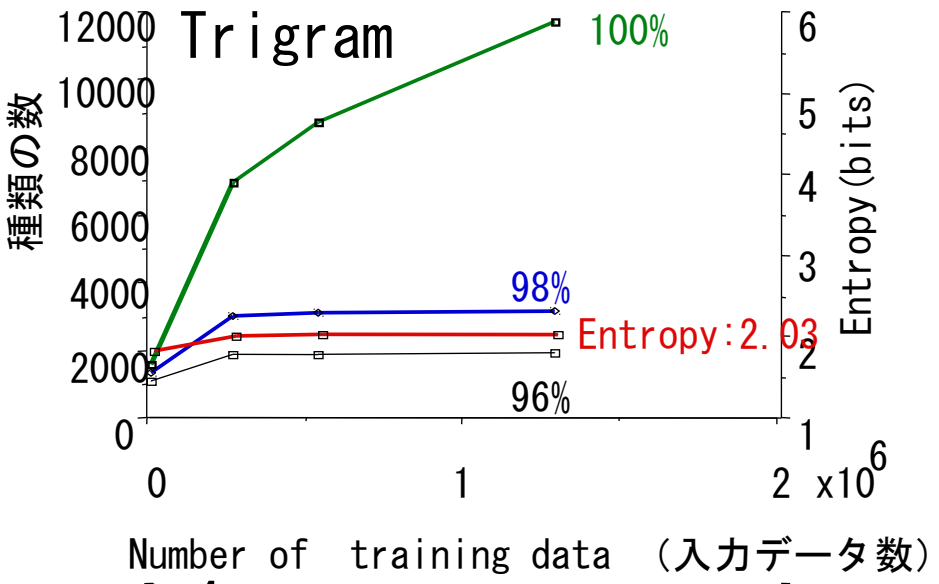


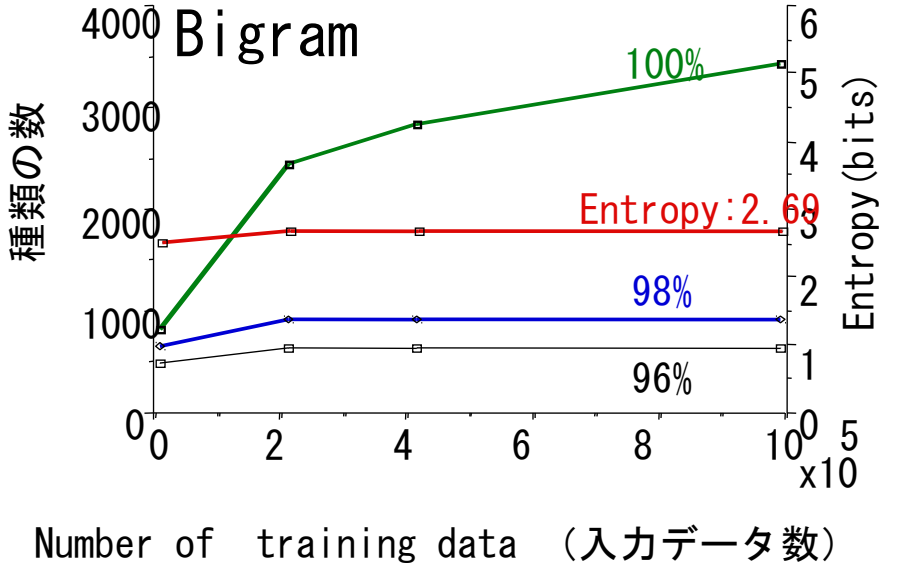


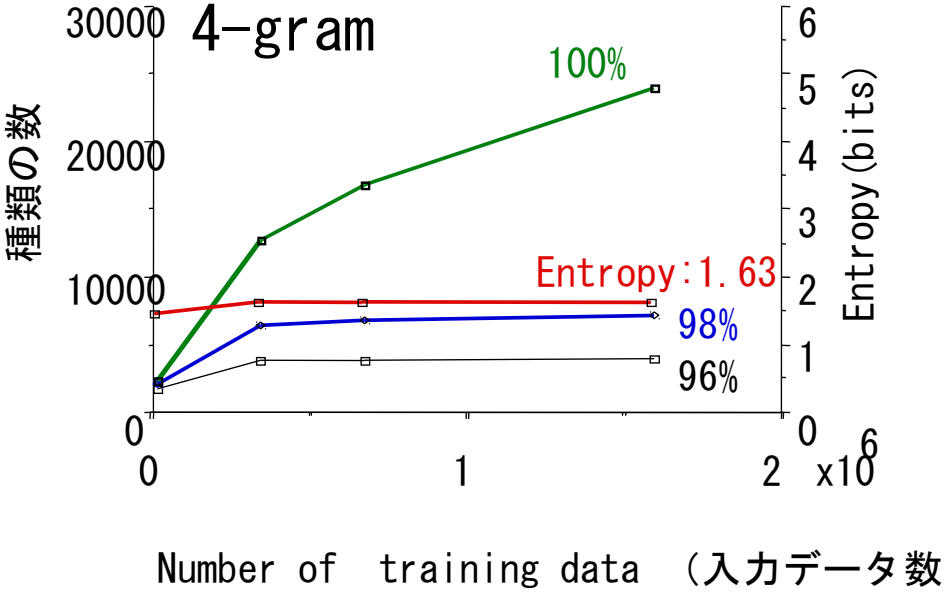


新聞記事における学習データ数に対する漢字仮名の
マルコフ連鎖確率値のカバー率およびエントロピー

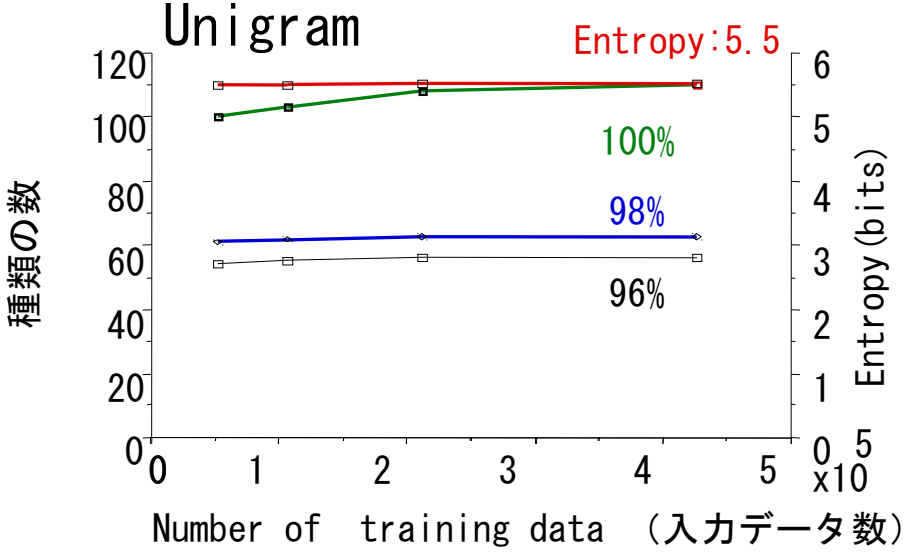


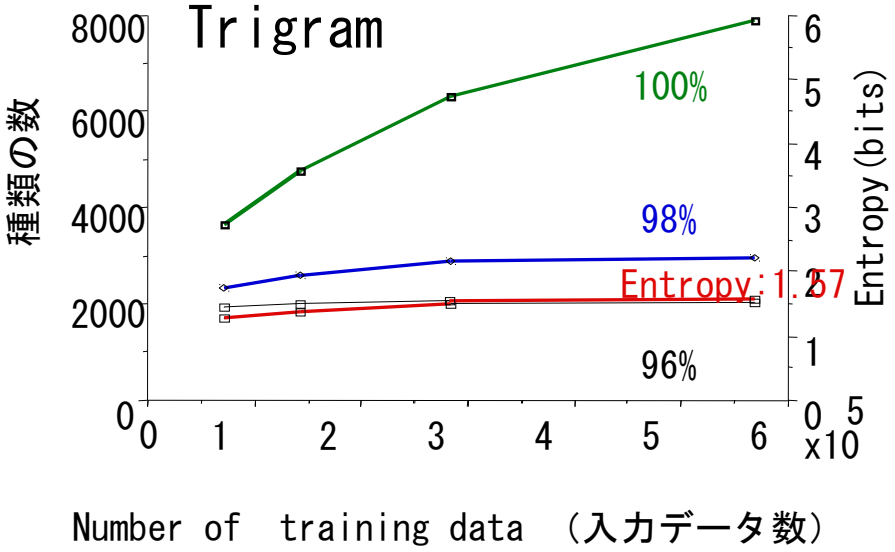


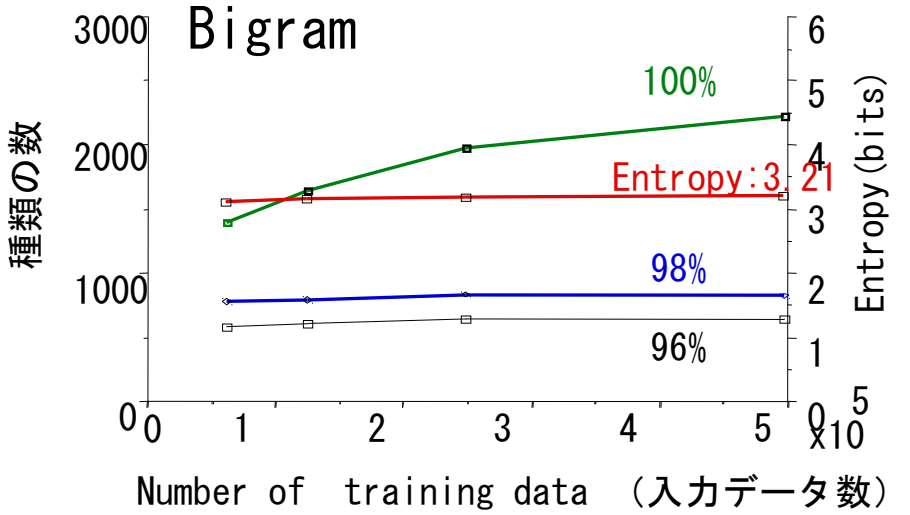




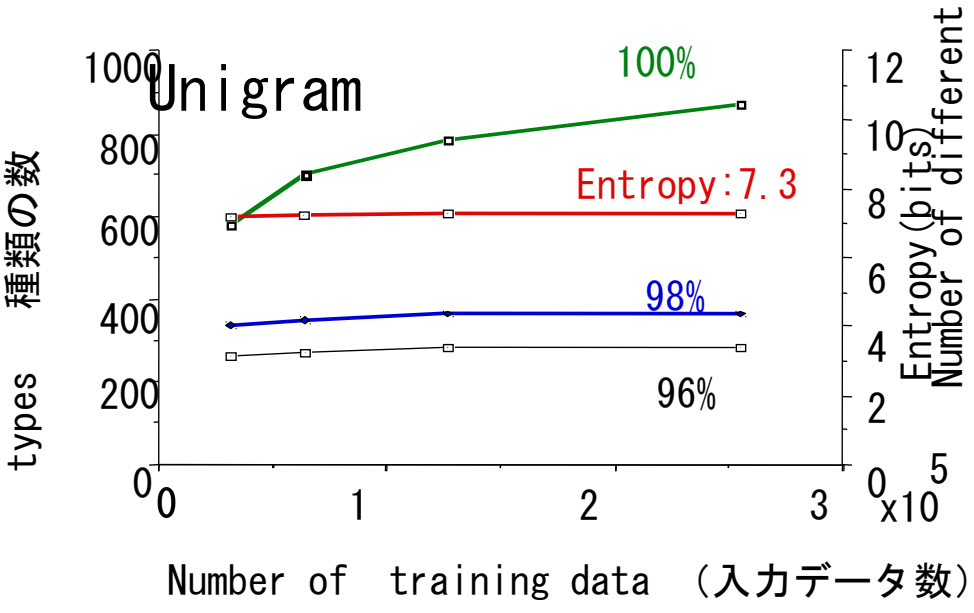
新聞記事における学習データ数に対する品詞の
マルコフ連鎖確率値のカバー率およびエントロピー

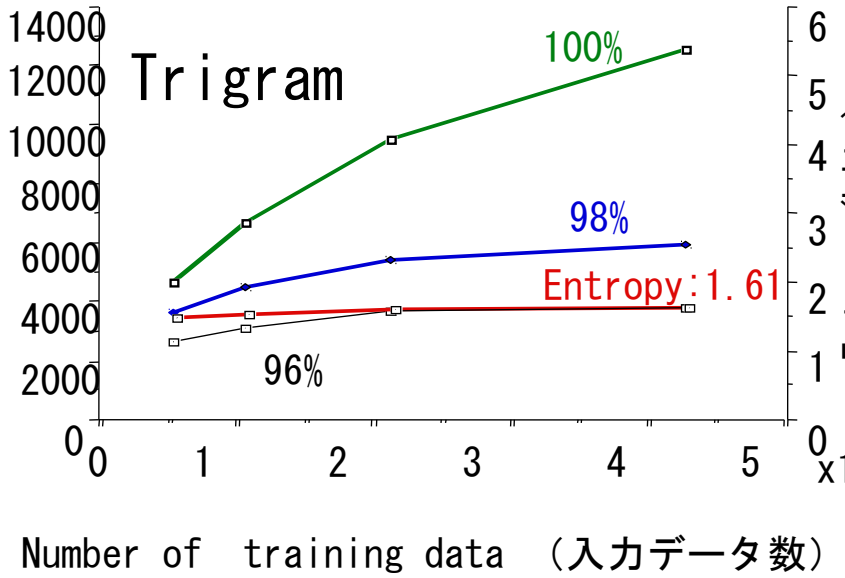


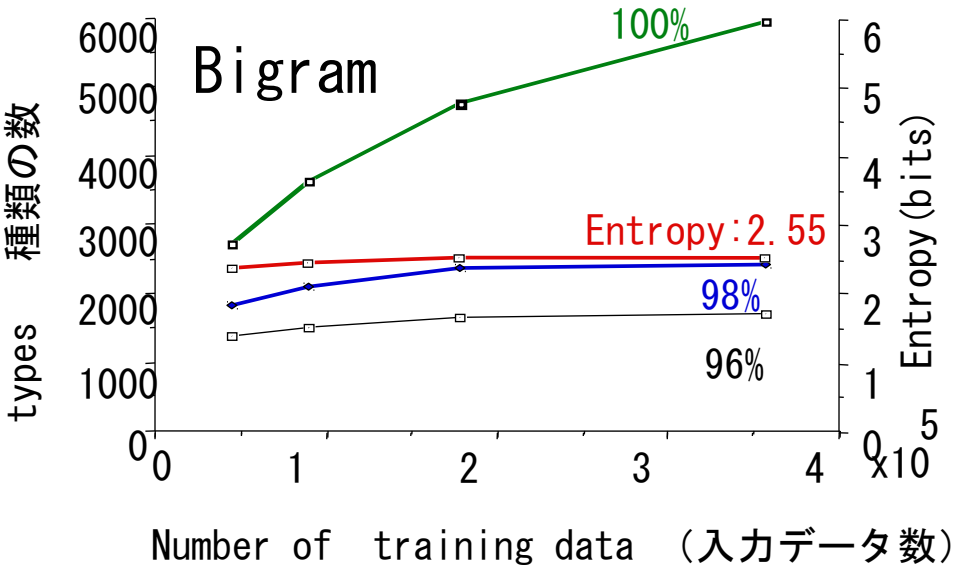




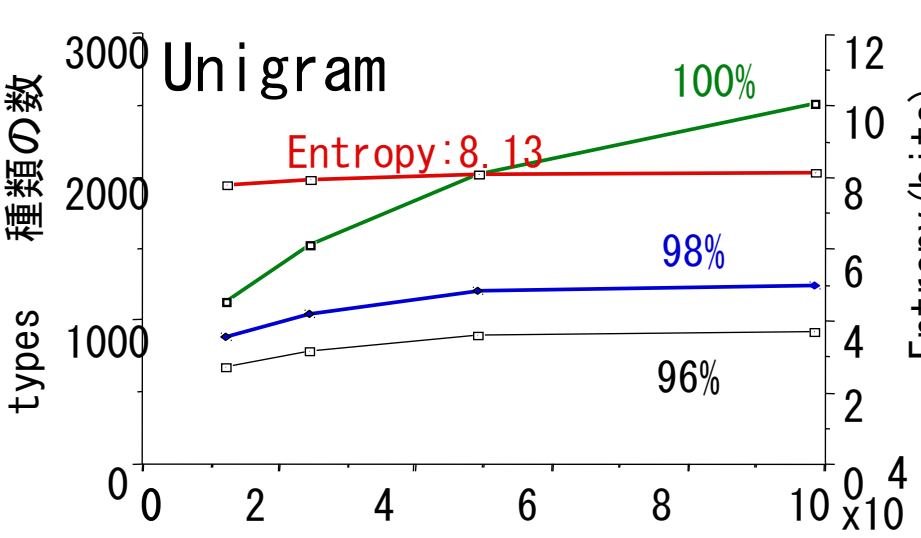
Ｘ線ＣＴ所見における学習データ数に対する音節のマルコフ連鎖確率値のカバー率およびエントロピー

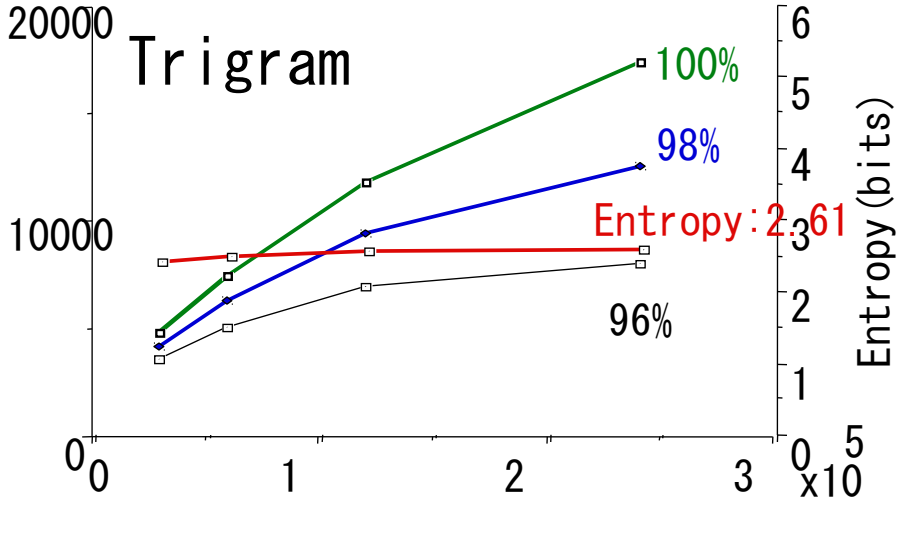


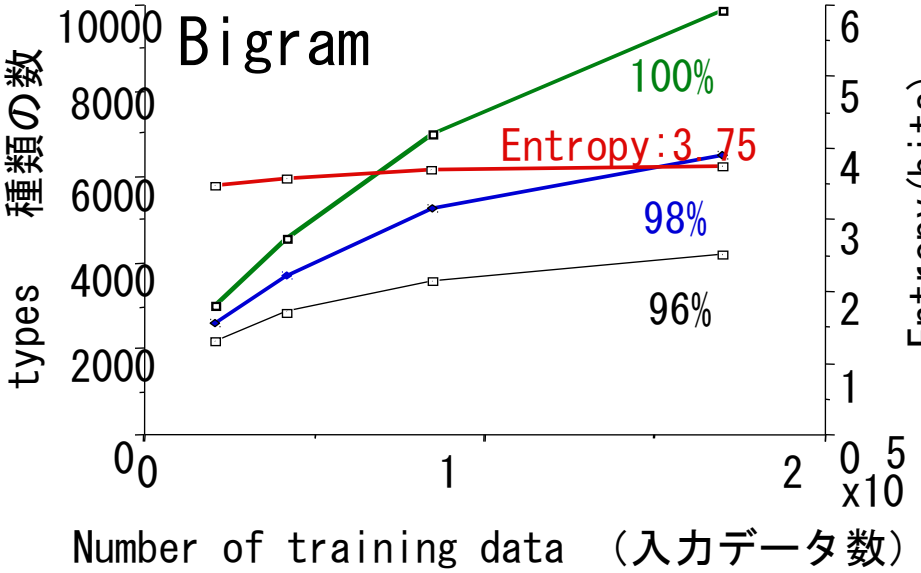




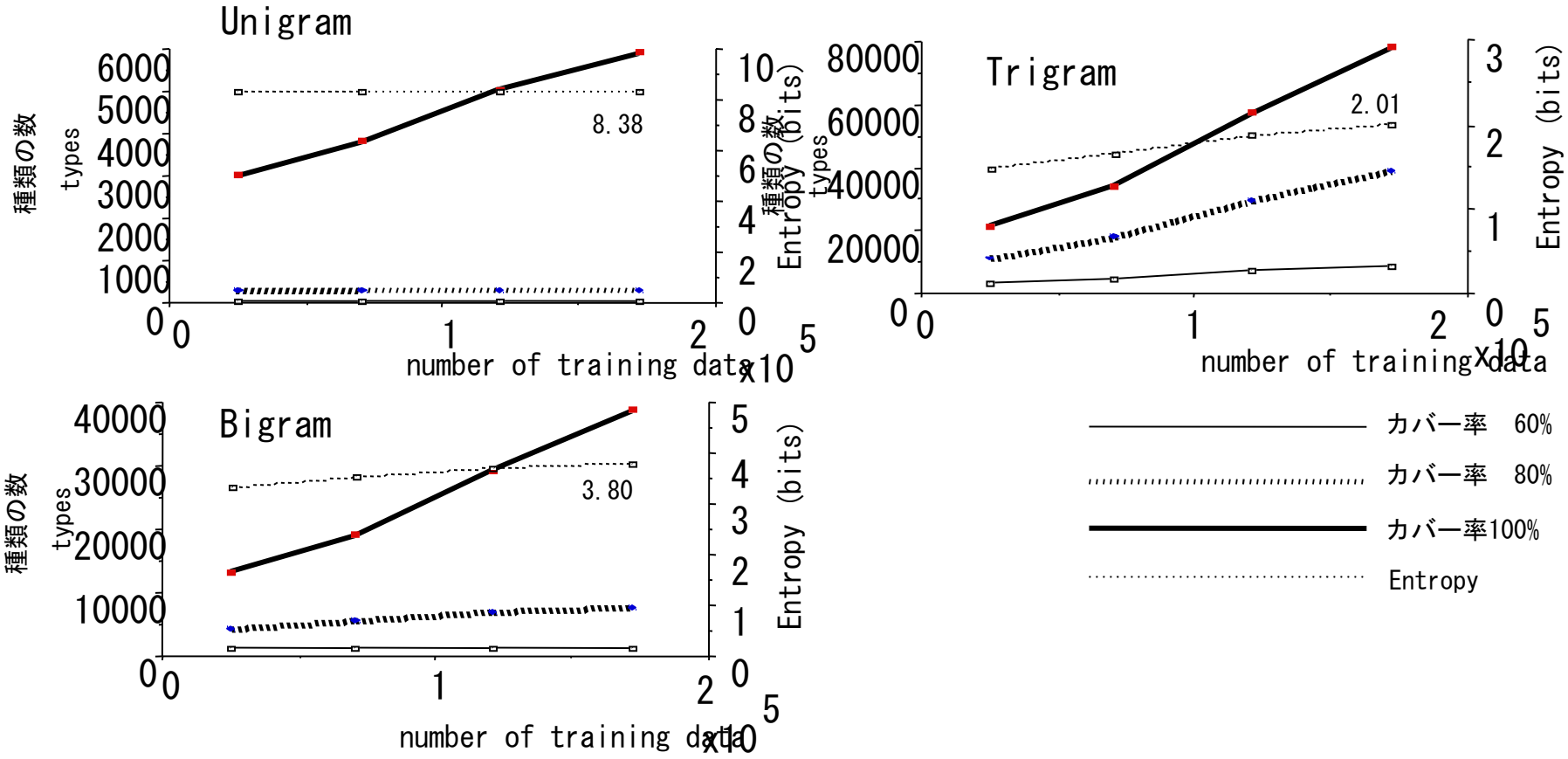
Ｘ線ＣＴ所見における学習データ数に対する漢字仮名のマルコフ連鎖確率値のカバー率およびエントロピー







Ｘ線ＣＴ所見における学習データ数に対する単語のマルコフ連鎖確率値のカバー率およびエントロピー



ＡＴＲの国際会議のデータベースにおける、学習データの
入力データに対するエントロピーおよびカバー率の変化(単語)

入力データ量に対するマルコフ連鎖確率値の変化のまとめ

1　エントロピーとカバー率

安定になるまでの学習データ数

エントロピー　〈カバー率〉

エントロピーだけでなく、カバー率も調査する必要あり

2　カバー率100%と98%

学習データが増加した場合、全体に占める割合は少ないが、
たえず新しい種類の連鎖が出現する

言語モデルとしてのマルコフモデルの妥当性

・滅多に出現しない言語現象は、
あえてモデルに適合させる必要がない

入力データ量に対するマルコフ連鎖確率値の変化のまとめ

3　新聞記事とX線CT所見作成の比較
X線CTの所見作成の文章は新聞記事と比較して文章が単純

漢字かな bigram　2.55 ＜ 4.45
(X線CTの所見作成)　(新聞記事)

4　形態素解析プログラムの精度
形態素解析プログラムの精度　単語認定率で約95%
人手によって文節単位に区切られた時の値との差
品詞のtrigramに有意性が見られない？

5　ATRの国際会議における単語trigramの信頼性
低い(データ量が必要)