日本語文章における Katz K Mixture Model の分析

服部 祥大 [1]　吉田 光男 [2]　梅村 恭司 [1]

[1,2] 豊橋技術科学大学　情報・知能工学系
[1]{hattori.shota.mu,umemura}@tut.jp
[2]yoshida@cs.tut.ac.jp

1 はじめに

文書中における語の繰り返し出現を表現するモデルとして，Katz K Mixture Model が存在する．このモデルは，文章中に語がちょうど k 回出現する確率を推定することができる [1].

文章中における語の繰り返しは，語の持つ意味と深く関わっていることが知られている．ある文章に語が 1 回以上出現した場合に，その文章に語が 2 回以上出現する条件付き確率は反復度と呼ばれ，固有名詞や専門用語，キーワードなどにおいて高くなる [2]．更に，日本語のような語の境界の情報を持たない言語においても，部分文字列の反復度を用いることで語の境界を特定し，キーワード抽出を行うことが可能である [3]．また，Katz K Mixture model における語の再出現の条件付き確率が一定であるという仮定をもとにして，条件付き確率が，繰り返しの回数に応じて単調増加しないものをテンプレートに使用される語句として抽出する試みも行われている [4].

これらの手法を利用するためには文書頻度を求める必要がある．この文書頻度を効率的に計算するアルゴリズムが存在し，コーパス中のすべての部分文字列に対して文書頻度を求めることができる [5]. このアルゴリズムを用いることでコーパス中の語を網羅的に調査することが可能となる．

Xu らは Katz K Mixture Model の仮定によって発生する問題に対応することで，推定の精度が向上することを報告している [6]．この報告は英字新聞コーパスにおいて実験・観察を行った結果に基づいており，日本語で同様の現象が生じるかは確認されていない．そのため，本論文では英字新聞コーパスにおいて確認された振る舞いが日本語コーパスにおいても生じているかを確認し，日本語においても同様の改善が可能であるかを調査した結果を報告する．

日本語では，英語と言語の構造が異なり，特に代名詞が使われることが少ないことから，語の繰り返しの出現はより顕著であると考えられる．また，今回の実験では，コーパスに百科事典である Wikipedia を用いており，百科事典は特定の項目に対する説明の集合である．そのため，主題となりうる語の繰り返しは，新聞よりも顕著になると考えられる．

この 2 点から，Wikipedia 日本語版における実験では，繰り返しが多いときに問題となる Katz K Mixture Model における仮定について，より考慮するべきであると考えられる．本論文では，英字新聞の場合と同じようにこの仮定が問題となることを確認した．

2 Katz K Mixture Model

Katz K Mixture Model は 3-way classification をもとに考案されている [1]．3-way classification とはある語に対して文章を

unrelated　語が出現せず，語と文章が関係ない
non-topical　語が出現し，語が文章の主題ではない
topical　語が出現し，語が文章の主題である

の 3 つに分類するもので，これらの文章の比率が Katz K Mixture Model のパラメータとなる．ここで，non-topical な文章と topical な文章の分類は，主題となる語は著者によって繰り返し用いられるという仮定をもとに，語が 1 回のみ出現する文章が non-topical に分類され，複数回出現する文章は topical に分類される．また，topical な文章における語の再出現の条件付き確率は一定であると仮定され，モデル化される．このとき，non-topical の文書を特別扱いしない，2 パラメータモデルも提案されている．

Xu らは繰り返しの条件付き確率が一定値として扱われることで，non-topical な文章において語が偶然に繰り返す場合が見落とされていると考え，再出現の条件付き確率について調査を行った．調査によ

り，形容詞や代名詞，接続詞などの機能語では上記の仮定が成り立つが，固有名詞などの内容語では，語の繰り返しが多いほど再出現の条件付き確率は高くなり，大きな出現数で安定する傾向にあることが分かった．また，この結果を元にして語の出現回数が少ない場合に対応する減衰係数と，分布の末尾に対応する減衰係数を組み合わせることで頻度推定の精度が高まることが示された．

本論文で使用する記号の定義を以下に示す．これらの記号は Xu らの論文に従っている．

D	文書の集合
N	コーパス中の文書数
$cf(w)$	コーパス中における単語 w の出現回数
$df(k;w)$	単語 w が k 回以上出現する文書数（文書頻度）
$tf(d;w)$	文書 d における単語 w の出現回数
$ddf(k;w)$	単語 w がちょうど k 回出現する文書数 $ddf(k+1;w) = df(k+1;w) - df(k;w)$
$cdf(k;w)$	$df(k;w)$ の累積和 $cdf(k;w) = \sum_{i \geq k} df(i;w)$
$P(k+1\|k;w)$	$P(k+1\|k;w) \equiv P(\frac{tf(X;w) \geq k+1}{tf(X;w) \geq k})$, X は文章の確率変数 ある文章において単語 w が k 回出現したときにもう一度単語 w が出現する条件付き確率． 文書頻度をもとに $P(k+1\|k;w) = df(k+1;w)/df(k;w)$ と推定される．

以降は w を省略して cf や $df(k)$ のように表記をする．

Xu らの論文において Katz の 3 パラメータモデルは次の式で表されている．

$$
\begin{aligned}
&P(tf(D) = k) \\
=&(1-\alpha)\delta_{k,0} + \alpha \times (1-\gamma) \times \delta_{k,1} \\
&+ \frac{\alpha \times \gamma}{\beta + 1} \left(\frac{\beta}{\beta + 1} \right)^{k-2} \times (1 - \delta_{k,0} - \delta_{k,1}) \\
&\delta_{i,j} = \begin{cases} 1 & \text{iff } i = j \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

パラメータは次の式で推定される．

$$
\begin{aligned}
\hat{\alpha} &= \frac{df(1)}{N} \\
\hat{\gamma} &= \frac{df(2)}{df(1)} \\
\hat{\beta} &= \frac{cf - df(1) - df(2)}{df(2)} = \frac{cdf(3)}{df(2)}
\end{aligned}
$$

ここで

$$
\frac{\beta}{\beta + 1} = \frac{cdf(3)}{cdf(2)}
$$

が topical な文章における再出現の条件付き確率となり，この減衰係数に従って繰り返し回数が多くなるほど語が丁度 k 回出現する確率が減少していく．

2 パラメータモデルでは，その減衰係数は次の式で表される．

$$
\frac{\beta}{\beta + 1} = \frac{cdf(2)}{cf}
$$

3 実験

本論文の実験ではコーパスとして日本語 Wikipedia のダンプデータ（総記事数 1,225,965 件）を使用し，本文の抽出には WikiExtractor.py を用いた．また，長い文章が多い Wikipedia の本文をそのまま使用すると語が偶然に繰り返すことが多く発生し，条件付き確率の分布の様子を確認することが難しくなってしまったため，本文の最初から 4 段落目までを対象とすることで文章長の制限した．

Xu らと同じように条件付き確率を $P(k+1|k), 1 \geq k \geq 8$ において観察するため，分析の対象とする語は $df(9) \geq 2$ かつ $df(k+1) \neq df(k)$ を満たすものとした．

3.1 機能語と内容語の分布の違い

機能語と内容語の間では，再出現の条件付き確率の分布の様子が異なると，Xu らは報告している．日本語コーパスにおいても同様の特徴があるか確認するため，いくつかの機能語と内容語について条件付き確率の分布を求めた．それぞれの分布を図 1 と図 2 にプロットする．

Katz は，語が一回だけ使用されている non-topical な状態から，語が繰り返し使用される topical な状態へと変化する確率が $P(2|1)$ であり，$P(k+1|k), k \geq 2$ は topical な状態のときにどの程度の確率で，語が再度使われるかの確率であるとしている．図 1 を確認すると，機能語では，分布の末尾部分が不安定になっていることを無視すれば，殆どの語の条件付き

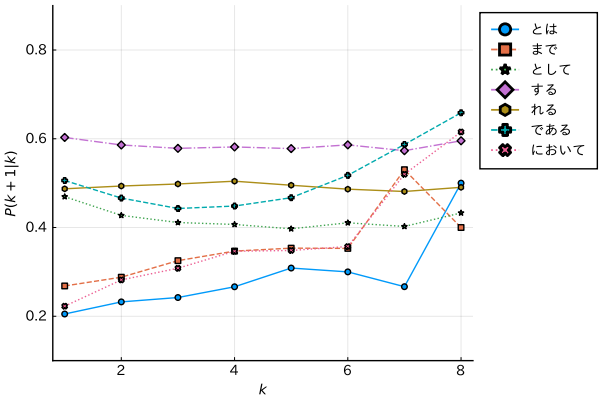


図 1: 機能語の条件付き確率の分布

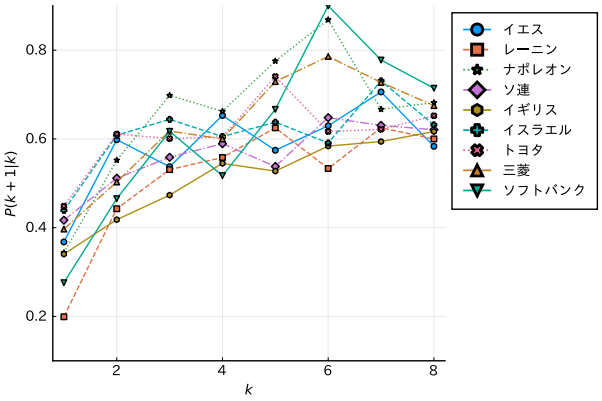


図 2: 内容語の条件付き確率の分布

確率がおおよそ x 軸に水平であり，Katz K Mixture Model の仮定に沿っているといえる．

図 2 の内容語では，$P(2|1)$ が $P(3|2)$ と比べて明らかに小さいことが確認できる．また，$P(3|2)$ と $P(4|3)$ を比較すると $P(4|3)$ の方が大きな値をとっており，出現回数 k が増えるとともに，$P(k+1|k)$ の値が徐々に大きくなり，安定していくことが分かる．これは，繰り返し出現の回数が大きくなるほど，偶発的な再出現が減り，主題となる再出現の影響が大きくなるためである．この条件付き確率の振る舞いは Katz K Mixture Model における条件付き確率が一定値であるという仮定とは一致していない．

3.2 減衰係数と条件付き確率の比較

より詳しく内容語の条件付き確率の分布について調査する．142 個の内容語においてそれぞれのモデルの減衰係数と条件付き確率の平均値，中央値，第 1・第 3 四分位数を比較したグラフを図 3 に示す．すべての統計値において 3.1 で確認された結果と同様に，$P(2|1)$ の確率が $P(k+1|k), k \geq 2$ と比較してかなり小さいことが分かる．また，条件付き確率は繰り返し回数に応じて大きくなっていき一定の値で落ち着く傾向にあることが確認できる．これは，語の繰り返し回数が少ない場合には偶然に語が繰り返すことの影響が残っており，繰り返し回数が多くなるほどに主題に関連した繰り返しの影響が大きくなっていくことが理由であると考えられる．ここで，2 パラメータモデルの減衰係数である $cdf(2)/cf$ を条件付き確率と比較すると $P(k+1|k), k \geq 2$ はこのモデルにおいて実際の値よりも小さく見積もられていることがわかる．同様に 3 パラメータモデルの減衰係数である $cdf(3)/cdf(2)$ においては $P(3|2)$ が過大に見積もられる一方で，$P(k+1|k), k \geq 3$ が実際の値よりも小さく見積もられている．

それぞれの語に対する減衰係数と $P(2|1)$,$P(3|2)$ の比較をを図 4 にプロットする．殆どの語において $df(3)/df(2)$ が，$cdf(2)/cf$ と $cdf(3)/cdf(2)$ の間に存在することが確認でき，日本語においても Katz の 3 パラメータモデルは繰り返しの確率 $P(3|2)$ を高く見積もり過ぎているということが示された．

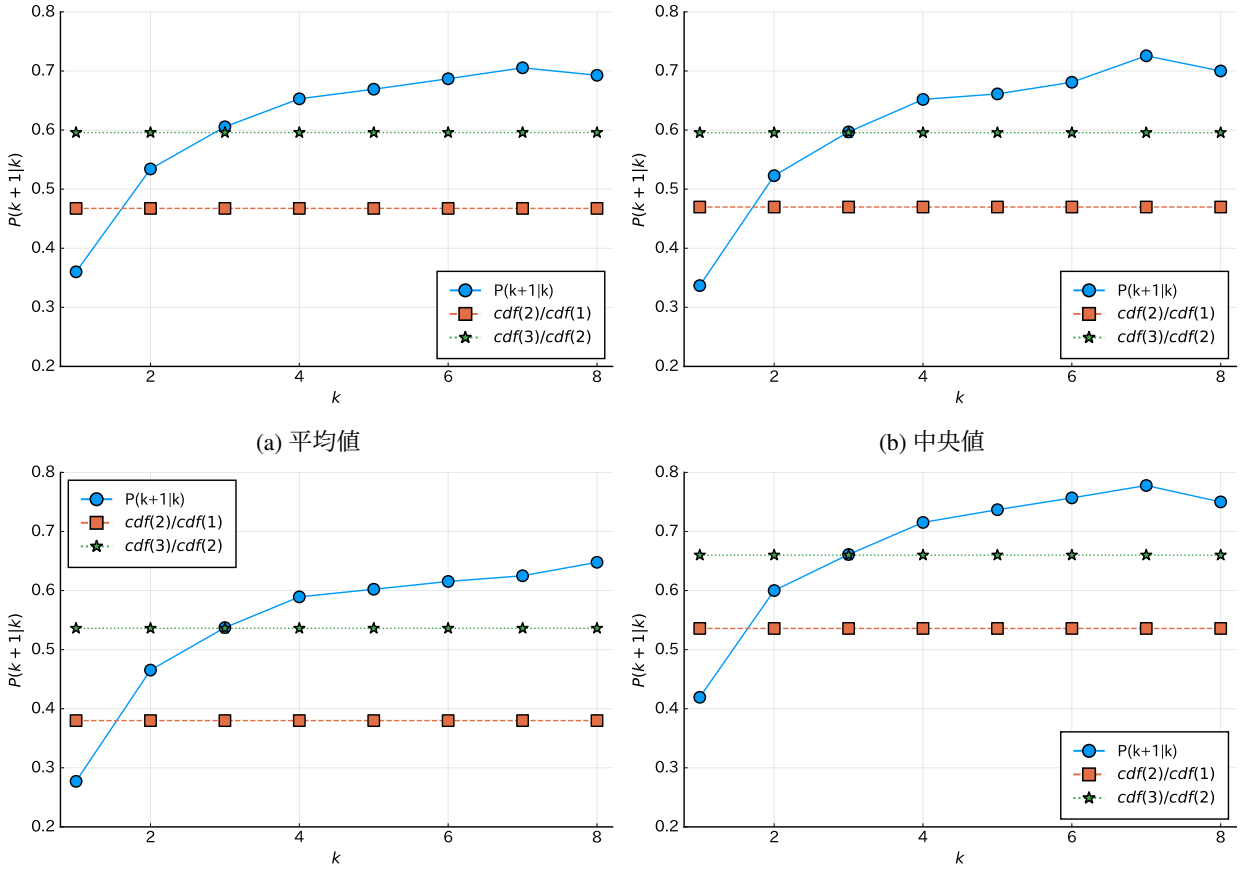
Xu らは，特に $P(3|2)$ について，un-topical な状態と topical な状態の双方の影響を受けているとして，k に応じて減衰係数が変化するモデルを構築することでより実際の分布に近いモデルを作成した [6].

4 おわりに

本論文では，日本語において Xu らが提案したモデルを Katz K Mixture Model の代替として使用することが適当であるか検討した．

結果から，日本語においても語の再出現確率が実際の値よりも高く見積もられる問題が存在することが確認され，Xu らが提案したモデルによって，推定精度の改善が見込めることがわかった．

これから，日本語において Xu らのモデルを用いることで精度の改善が実際に行えるかなどを調査したい．



(a) 平均値

(b) 中央値

(c) 第 1 四分位数

(d) 第 3 四分位数

図 3: 条件付き確率 $P(k+1|k)$ の統計値と減衰率の比較

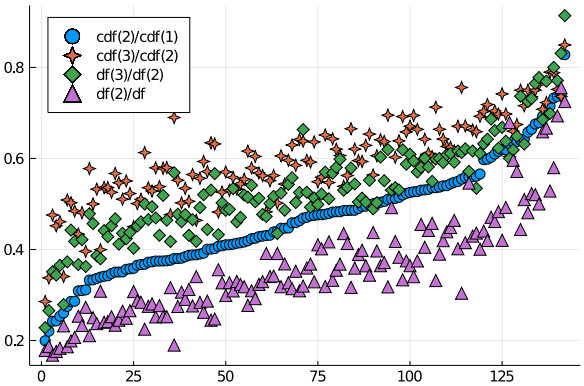


図 4: 語ごとの条件付き確率と減衰率の比較

参考文献

[1] SLAVA M. KATZ. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, Vol. 2, No. 1, p. 15–59, 1996.

[2] Kenneth W. Church. Empirical estimates of adaptation: The chance of two noriegas is closer to $p/2$ than p^2. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*, 2000.

[3] 武田善行, 梅村恭司. キーワード抽出を実現する文書頻度分析. 情報処理学会研究報告. NL, 自然言語処理研究会報告, Vol. 146, pp. 27–32, 2001.

[4] 藤原大輔, 高瀬暁央, 梅村恭司. テンプレートを構成する名詞の Katz モデルによる抽出の試み. 情報処理学会研究報告, 2007-NL-180, No. 25, pp. 145–149.

[5] Kyoji Umemura and Kenneth Church. Substring Statistics. In *Computational Linguistics and Intelligent Text Processing*, pp. 53–71, 2009.

[6] Yinghui Xu and Kyoji Umemura. Improvements of katz k mixture model. *Journal of Information Processing*, Vol. 12, pp. 131–155, 2005.