マルコフチェインによるワードスパムの合成実験とその評価について

鴨志田芳典 菊池浩明
東海大学情報理工学部情報メディア学科

1 はじめに

ワードサラダは，マルコフチェインを利用して学習元のコーパスから構文的な間違いのない，人工的な文章を生成するスパムブログ方式である．本研究では，マルコフチェインによる文章合成の効果を明らかにしてその防止対策を検討するために，ワードサラダを実装していくつかの評価実験を行った．ワードサラダによる人工的に合成された文章と人が文脈を判断して作成した文章の違いを，被験者による評価実験から明らかにする．

2 ワードサラダ

2.1 合成方法

ワードサラダは，コーパスから n 階マルコフ連鎖を利用して合成される．コーパスとなる文章に形態素解析を行い，分かち書きされた形態素の n 階マルコフ連鎖モデルを求め，それを基に確率的に形態素を並べていく．n 階マルコフ連鎖において i 番目に出力される語 X は

$$P(X_i) = P(X_i|X_{i-1}, X_{i-2}, \ldots, X_{i-n})$$

の条件付確率に従って生起する．この n を階数と呼ぶ．

図1に単純マルコフ連鎖 ($n = 1$) から作られたワードサラダの出力の例を示す．形態素解析器として MeCab を用いた．

> つまり自分が、怒りに引き揚げても、謂わばいいくらいでしたのぞ》を食べなければ通俗の苦しみ、それは、子供のは爽快《もっ》のこぶしを感じるの腰布（しかし、めしを、もじもじした。

図 1: 単純マルコフ連鎖による文の例

3 実験

3.1 評価データ

同じコーパスから合成した階数 $n = 1, \ldots, 4$ のワードサラダ，文章を行単位で並べ替えたセンテンスサラダ，人がコーパスから文脈を考慮した切り貼り文章を評価する．コーパスには 5000 文字程度の政治・経済に関する記事と最近更新された goo ウェブ検索総合急上昇ランキングの内，最もポイントの高いキーワードを検索クエリとし，Google ニュース検索から抽出したニュース記事 3,5 件の 2 種類を用いた．またコーパス長を L とする．

3.2 実験方法

実験 1. 主観評価 情報系の学生 9 名に対し次の主観評価実験を行う．評価データ計 100 題を提示し，その文章が機械的に合成されたものであるかどうかを判断させ，その正答率と応答時間を計測した．事前実験では鍵括弧の位置が大きな判断基準としてあったため，評価データからは括弧表現を全て除去した．

実験 2. 投稿評価 ブログ名，記事タイトル，投稿日時，ジャンル等の条件を揃えて各々4 件ずつ，計 24 件公開し，アクセス数を観測した．記事を公開するブログサービスとしてココログを用いた．

実験 3. 復元評価 コーパスと同じ文章を復元した回数を復元率とする．ワードサラダを 1000 文ずつ合成し，コーパス長 $L = 2500, \ldots, 10000$ と階数 $n = 1, \ldots, 6$ について，復元率の変化を調べる．

4 実験結果

実験 1 の正答率と応答時間を図 2 と 3 に各々示す．実験 2 による総アクセス数，検索経由アクセス数，コメント数，トラックバック数を表 1 に示す．ただし，n=5 はセンテンスサラダ，n=6 は原文または切り貼りとする．実験 3 の復元率を図 4 に示す．異なるコーパスで検証した結果もほぼ同様の振る舞いになった．

5 考察

実験 1 の結果では，$n = 6$ の正答率が最も低い．単一コーパスから合成したワードサラダは自然な文章と判別が困難であると言える．

実験 2 では，ワードサラダの生成方法によってアクセス数等の明確な差は生じなかった．これは，コメントや

An Experimental Evaluation of Distinguishability with Word-Spams Synthesized by Markov Chain
Yoshifumi Kamoshida and Hiroaki Kikuchi
School of Information Secience and Technology, Tokai University

表 1: 実験 2:アクセス統計量

階数 n	アクセス数	検索経由	コメント	トラック
1	3	2	0	1
2	25	3	1	1
3	7	3	2	7
4	6	2	0	2
5	8	5	1	1
6	8	1	0	1

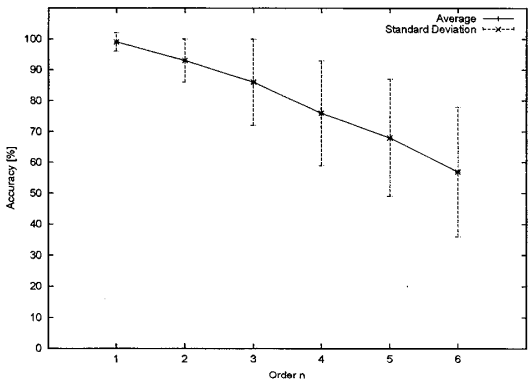

図 2: 実験 1:階数 n についての正当率

トラックバックといった要因に大きく左右されている為であると考えられる. また明らかに自然な文章ではない単純マルコフ連鎖から作られたワードサラダには, アフィリエイトを促す様なコメントがついた. n=2 以上のワードサラダには, 記事内容についてのコメントやトラックバックが付く事もあった.

実験 3 の結果より, $n \geqq 5$ で復元率が 80%以上になった.

5.1 検出方法の検討

[2] は階数 n についてのカルバック・ライブラー情報量の差でワードサラダを自動判別する方法を検討して

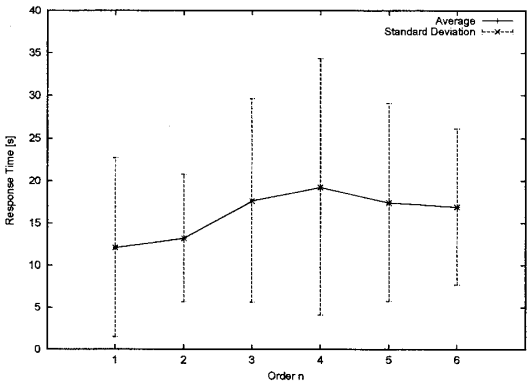

図 3: 実験 1:階数 n についての応答時間

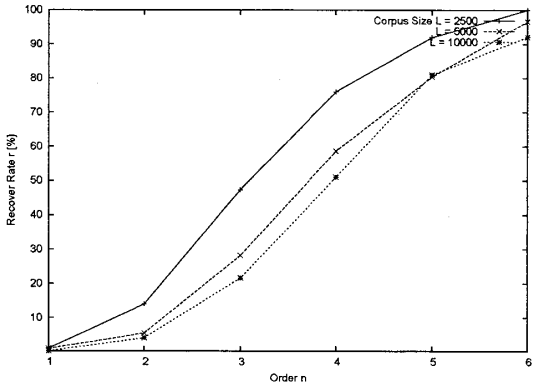

図 4: 実験 3:階数 n についての復元率

いる. しかし, 本実験によると $n \geqq 5$ で通常文と識別不能になり [2] では検出できない. 従って従来手法と, Web からの検索を合わせて使用してフィルタリングする必要がある.

また, マルコフ連鎖による文生成には, データベースの増大に伴い一文の文字数が多くなる事が知られている. 一文をまるまる表示するワードサラダなら, 文字数からコーパスサイズが予想できる可能性がある. 文末までを表示しないワードサラダならば, 文末に相応しくない単語が出現しているはずなので, そこから検出できると考える.

6 おわりに

ワードサラダの検出性能を評価するため, 人を対象とした実験を行った. 大規模なコーパスや, 口語表現を多く用いたコーパスを用いた場合についての実験を今後の課題とする.

参考文献

[1] 森本, 片瀬, 山名：“N-gram と離散型共起表現を用いたワードサラダ型スパム検出手法の提案”, 情報処理学会研究報告, DBS-148, No.24, pp.1-8,2009.

[2] T. Larvergne, et al., ：“Detecting Fack Content with Relatine Entropy Scoring”, CEVR, Vol.377, pp. 27-31,2008.