マルコフ性を用いたギブスサンプリングによる代表文書の生成

大野　奈那子　　三浦　孝夫
法政大学理工学部創生科学科

1. 前書き

直接サンプリングが難しい確率分布の代わりにそれを近似するサンプル列を生成する方法で MCMC の手法の 1 つにギブスサンプリングがある。本研究では、このギブスサンプリングを用いて、新聞記事の文書データをもとに代表文書の確率的生成を行う。

2. 代表文書

文書集合が全体として共通の特性を有すると仮定する。次の 4 つの性質を考える。

(1) 内容がよくみられるトピックである。
(2) 多く出現する単語が含まれている。
(3) 常識的な語のつながりを持った語列である。
(4) 確率的に代表文書を生成できる。

この 4 つの条件を満たしているものを”代表文書”とする。

3. 提案手法と評価方法

本研究で対象となるコーパスは、「CD-毎日新聞 2017」の 1 月 1 日～1 月 14 日の記事(記事数:2302 件、総単語数:97774 語)を文書ベクトル化したものである。まず、ランダムに単語を出現させた文書ベクトルを初期値とする。そして、モンテカルロ法を用いた語生成で 1 語ずつ語の置き換えを行い、収束するまで繰り返す。ここまでが通常のギブスサンプリングである。しかしこの方法では、1 語ずつ置き替えるため語同士のマルコフ性を興梠していないという問題点がある。本生成方法では、「ドナルド・トランプ」のような 2gram を考慮していない。そこで、「ドナルド」に続いて生起する確率を持った「トランプ」を確率的に生成させることで、マルコフ性を持つ語列ができる。本研究では、予めマルコフ遷移表を作成する。ギブスサンプリングとマルコフ遷移表の双方を用いて単語を生成することで、マルコフ性を持った代表文書の生成が期待できる。

そして代表文書の評価については、まず代表文書の記事内容がよくみられるトピックかどうか評価する。そして類似文書の移り変わりを見て確率的に代表文書を生成ができているか評価する。最後に代表文書の 2gram の割合でマルコフ性を評価する。

3. 実験結果

マルコフ性を用いたギブスサンプリングで生成された代表文書は高校ラグビーについての記事、通常のギブスサンプリングで生成された代表文書は自動車販売台数についての記事になる。類似文書の移り変わりを見ると、マルコフ性を用いたギブスサンプリングのほうがテーマの一貫性がある。代表文書における 2gram の割合はマルコフ性を用いた場合は 20 件中 7 件(35%)、通常の場合は 0%という結果になる。

表 1 類似文書

文書番号	テーマ	類似度
文書705	相撲	0.25715143
文書1092	相撲	0.81589244
文書705	相撲	0.80632948
文書1092	相撲	0.77644595
文書295	駅伝	0.82337261
文書1643	サッカー	0.57483383
文書1575	サッカー	0.60791876
文書2046	サッカー	0.65950764
文書83	ゴシップ	0.6092718
文書2343	ゴシップ	0.55205245
文書1114	募集	0.60764362
文書1670	芸術	0.71356011
文書815	文化	0.51178741
文書651	自動車	0.64089084
文書2013	自動車	0.61046528
文書651	自動車	0.62545195
文書1761	自動車	0.70444848
文書651	自動車	0.84481947

表 2 類似文書(マルコフ性)

文書番号	テーマ	類似度
文書705	相撲	0.257151
文書988	相撲	0.367359
文書705	相撲	0.725845
文書1092	相撲	0.721401
文書295	駅伝	0.798529
文書1253	ラグビー	0.83411

4. 考察

マルコフ性を用いたギブスサンプリングで生成された代表文書の内容がスポーツ記事であることからよくみられるトピックであるといえる。また表 1、表 2 より、テーマの一貫性から確率的に代表文書を生成できたといえる。また代表文書の 2gram の割合(35%)より、マルコフ性を持った代表文書が生成できたといえる。したがって、意味の妥当性と収束の合理性の両面で提案手法の有用性が確立される。

5. 結論

本研究において、新聞記事データをもとにマルコフ性を用いたギブスサンプリングによる代表文書の生成ができた。初期値依存性を考慮してバーンイン期間を設けることでより精度が高い確率的生成が可能になると思われる。

参考文献

[1] 須山敦志、杉山将(2017)
「ベイズ推論による機械学習入門」