ISS-A-069

ギブスサンプリングによる代表的文書の確率的推定

佐藤　謙仕　　三浦　孝夫
法政大学理工学部創生科学科

1. 前書き

本研究ではギブスサンプリングによって確率的に文書を生成することで代表的文書を推定する。文書中に多く出現する単語が含まれており、共起的に単語がサンプリングされている文書を代表的文書と定義する。

2. ギブスサンプリングによる文書生成

出現頻度を重要度としたベクトル空間モデルにより文書を表現する。訓練文書から 1 単語ずつギブスサンプリングにより単語の入れ替えを収束するまで行う。条件付き確率分布として、文書集合中の訓練文書との最大類似文書の確率分布を近似する。類似比較には余弦類似度を用いる。

3. 実験

(1) 準備

本研究で用いるコーパスは、毎日新聞 2 週間分のデータで、名詞だけを扱う。

(2) 手順

収束条件は書集合中の訓練文書との最大類似文書が 200 回の総単語の入れ替えで遷移しないこと。文書サイズ、不要語、訓練文書の記憶域の条件を変化させて実験を行い、それぞれの影響をみる。代表的文書は最大生起確率が 0.2 以上であることと最大類似文書が収束していることで評価する。

(3) 結果

表 1 サイズの影響

最大類似文書	最大類似度	見出し語	延べ語
740	1	1	2

表 2 不要語・記憶域の影響

不要語	最大生起確率	記憶域	延べ語数	最大類似文書推移数
有	0.11764	20	17	125
有	0.11111	20	18	217
有	0.14286	20	21	97
無	0.57143	40	25	37
無	0.40909	40	22	40
無	0.30769	40	39	52

表 3 代表的文書の内容

コラム	エッセイ	スポーツ記事	ニュース
4	3	2	1

4. 考察

延べ語の小さな文書に収束した。

不要語を取り除くことによって生起確率が 0.2 以上の特徴語がある文書を抽出できた。

記憶域が長いほどサイズの大きな文書を反映できるが、最大類似文書が遷移しない。

5. 結論

小サイズ文書・不要語を除くことで代表的文書を推定できた。毎日新聞の代表的文書はエッセイ・コラムである。

参考文献

http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt