D-5-10 CRFを用いたBlog著者年代推定(D-5. 言語理解とコミュニケーション,一般講演)

泉, 雅貴 / 三浦, 孝夫

(出版者 / Publisher)
一般社団法人電子情報通信学会
(雑誌名 / Journal or Publication Title)
電子情報通信学会総合大会講演論文集
(号 / Number)
1
(開始ページ / Start Page)
50
(終了ページ / End Page)
50
(発行年 / Year)
2007-03-07

D-5-10

CRFを用いたBlog著者年代推定

泉 雅貴[1]　　三浦孝夫[1]

[1] 法政大学工学部電気電子工学科

1 動機と解決法

昨今、Blogの利用者は爆発的に増大している。Blogには、人々の主観的な意見や、考えというのが多く記述されいる。さらに、Blogの記事は日々更新され、その情報はリアルタイム性に優れている特徴を持つ。

これらのことより、このBlogをマーケティングなどの情報源と捕らえ、Blogの記事から現在流行しているものや、商品に対する意見や評判情報などを獲得しようとする研究が近年行われるようになっている。

しかし、これらの情報というのは、人々の年齢や性別、職業など、人々の属性によって大きく傾向が異なる場合が多く、１０代で流行しているものはなにか？という情報などを取り出すためには、予めそのBlog記事をそれらの属性に仕分けしておく必要がある。

そこで、本論文ではこれらの属性のうち、「著者の年代」に対して、Blogを推定する手法を提案する。年代を推定する手法は、CRFを用いてBlogの記事中の単語に"10代","20代","30代","40代","判定不能"のいずれかのラベルを付加し、記事中で一番多く付加したラベルをそのクラスとする。また、実験によりこの推定方法が有効であることを示す。

2 Conditional Random Fields

CRF(Conditional Random Fields)は、観測系列Xが与えられたとき、それに対するラベル系列Yを一意に決定するためのアルゴリズムである。Markovモデルを用いたHMMやMEMMとは異なり、１つの指数分布モデルによって条件付確率$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_i^n \sum_a \lambda_a f_a(X_i, Y_i)\right)$を表現する。$f_a(X_i, Y_i)$は二値の出力をもつ素性関数で、その素性関数に対しパラメータλ_aが付随する。$Z(X)$は正規化項である。これにより、柔軟な素性の定義、ラベル偏り問題が解決できる。入力Xに対数最適なラベル列Yは、Viterbiアルゴリズムを用いて効率よく探索できる。また、パラメータの学習には予め用意した訓練データより、反復スケーリング法などを用いることにより値を決定する。これより、CRFで用いる素性とCRF対して適切な訓練データを与えることによって、各単語にラベル付けができる。

3 CRFによる著者年代推定

CRFを用いた著者年代推定に対する手順を以下に示す。

CRFで使用する素性は、単語、品詞の両方を用いて、前後3gramまでとし、"単語+ラベル","品詞+ラベル","単語+ラベル+1つ前のラベル"を素性として定義した。この素性を用いて、日本語固有名詞抽出、NP Chunking, Base-NP chunkingなどで9割以上の精度を出している。

次に、訓練データの作成について説明する。まず各クラスごとにbigramの単語で出現頻度を計算し、それをもとに特徴語を上位一定数選ぶ。次に、その特徴語を元に、各クラスの特徴を表す文章ををを一つ選択し、それを訓練データとしてパラメータ推定に利用する。

4 実験

本実験では、CRFで定義した素性は固定した上で、訓練データの作成に着目し、訓練データに用いる特徴語の抽出件数を変化させ実験を行った。実験に使用するBlogは、Yahoo!Blogより各クラスごとに1500件ずつ収集する。そのうち各クラスごとに1000件を訓練データ作成用に、残りの500件をテストデータとして割り当てる。また、前処理として形態素解析を行う上で一番の障害となる顔文字を、予め用意した顔文字辞書を使い、特殊文字として認識できるようにする。今回使用するCRFは、CRF++を使用し、形態素解析にはJUMANを使用した。

分類結果より、各年代別に$recall$,$precision$,F値をそれぞれ求め、その値より正しく分類できているかを判断する。

抽出件数	recall	precision	学習時間(s)	F値
100件	0.16	0.41	28	0.23
500件	0.37	0.42	340	0.393
1000件	0.38	0.44	1010	0.40
2500件	0.41	0.49	2654	0.446
5000件	0.45	0.50	4768	0.474
10000件	0.46	0.50	8476	0.479

図 1: 特徴語抽出件数を変化させた時の分類結果

	recall	precision	F値
10代	0.44	0.59	0.50
20代	0.36	0.40	0.38
30代	0.31	0.44	0.36
40代	0.71	0.57	0.63

図 2: 抽出件数5000件時の分類の詳細

5 考察

図より、特徴語抽出件数10000件時が一番良い結果であることを示している。続いて5000件時が良いが、10000件時と比較するとF値はほとんど変わらず、誤差範囲内と考えても良い。また、学習時間を見ると5000件時と比較し、10000件時は二倍近い学習時間がかかることから、特徴語5000件時が一番妥当であると言える。

次に図には、抽出件数5000件時の分類結果の詳細を示す。この結果には判定不能は含まれていない。これより、10代、40代は分類結果が良く、両方ともF値は0.5以上とれている。特に40代に関しては、precisionが6割以上取れており、かつrecallが7割以上取れていることから、システムとして有効であるということが言える。それに対し、20代、30代の分類結果はF値が0.4を下回り、あまり分類精度を上げることができていない。さらに、例えば30代ならその年代の周辺の20代、40代に対しての誤分類が多いことがわかる。このことから、20代、30代の著者のもつ特徴は、その周辺の年代の著者がもつ特徴と類似している部分があり、それにより分類精度を下げてしまっている可能性がある。

6 結び

本論文では、CRFを用いて、訓練データ作成の観点から、Blogにおける著者性別推定の有効性を示した。しかし、20代、30代では特徴をつかむことができず、今後はCRFの素性について考えていかなけらばならない。

参考文献

[1] J. Lafferty, A. McCallum and F.Pereira: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Intn'l Conf. on Machine Learning(ICML), 2001