参考文献書誌情報抽出における能動サンプリングのための有効な確信度の検討

Examination of effective confidence measures for active sampling in bibliography extraction from reference strings

浪越 大貴
Daiki Namikoshi

岡山大学 太田研究室
Ohta Laboratory, Okayama University

概要 電子図書館の運用には書誌情報データベースの整備が必須である．特に，学術論文の参考文献欄には，著者名やタイトルなどの有用な書誌情報が集約されている．しかし，CRF により書誌情報を高精度に抽出するには一定量の学習データが必要で，その生成コストが問題だった．そこで川上らは，能動サンプリングを利用して学習データを削減する方法を提案した．本研究では，学習データの選別に用いる確信度としてどのようなものが有効であるか実験により評価した．

1 はじめに

多数の学術論文を蓄積する電子図書館を快適に利用するには，検索やソート，リンク等の機能が必須である．しかし，人手でこれらのための書誌情報をデータベースに登録するコストは膨大である．また，Conditional Random Field（CRF）[1]により，書誌情報を高精度に抽出するには雑誌ごとに一定量の学習データが必要であり，その生成コストは無視できない．そこで本研究では川上ら[2]と同様に，学習データが少ない場合に，能動サンプリングを利用して抽出精度を改善する方法について検討する．本稿では，学習データを選別する能動サンプリングにおいて，どのような確信度が有効であるか実験により評価した．

2 能動サンプリングを用いた書誌情報抽出

2.1 確信度

能動サンプリングは，ある時点での学習モデルで書誌情報抽出が困難な参考文献文字列を，優先的に選択して次回の学習データとし，逐次学習モデルを更新する．そのため，[3]を参考に書誌情報抽出の困難さを表す尺度として，以下の確信度を定義する．

- Token Margin (TM)
  この確信度は各トークンに付与された，上位二つのラベルの周辺確率を用いる．まず参考文献文字列中の入力系列 $\mathbf{x}$ の各トークンに対する上位二つのラベルの確率の差を求める．ここで，入力系列 $\mathbf{x}$ に対して，Y_i を参考文献文字列中のi 番目のトークンx_i に対して付与されうるラベルを表す確率変数とする．トークン列中の最小の差をその参考文献文字列の書誌情報抽出結果の確信度とする．具体的には以下の式で定義する．
  $$C_{TM}(\mathbf{x}) = \min_{1 \le i \le |\mathbf{x}|} P(Y_i = l_1 | \mathbf{x}) - P(Y_i = l_2 | \mathbf{x}) \quad (1)$$
  ここで，l_1，l_2 はそれぞれ周辺確率が 1，2番目に大きいラベルである．

- Nonnormalized Likelihood (NNLH)
  この確信度は，川上ら[2]が定義した確信度 Normalized Likelihood（NLH）において，入力系列の長さによる正規化を行わないものである．CRF は入力系列に対する条件付き確率が最大になるような出力ラベル系列$\mathbf{y}^*$を導出する．$P(\mathbf{y}^*|\mathbf{x})$ の値が小さければ，ラベル付与は困難であると見なし，この値を確信度として利用する．入力系列が長い場合，そもそもラベル付与が困難[3]なため，この確信度は長さによる正規化を行わない．具体的には以下の式で定義する．
  $$C_{NNLH}(\mathbf{x}) = \log\left(P(\mathbf{y}^* \mid \mathbf{x})\right) \quad (2)$$

次に川上らが[2]で定義した確信度を説明する．

- Normalized Likelihood (NLH)
  この確信度は，上記で定義した NNLH を入力系列$\mathbf{x}$ の長さで正規化した確信度である．具体的には以下の式で定義される．
  $$C_{NLH}(\mathbf{x}) = \frac{\log(P(\mathbf{y}^*|\mathbf{x}))}{|\mathbf{x}|} \quad (3)$$

- Minimum Probability of Token Assignment (MP)
  この確信度は，参考文献文字列中の各トークンに付与されたラベルの周辺確率そのものを利用している．L を付与できるラベルの集合とする．確率 $\max_{l \in L} P(Y_i = l \mid \mathbf{x})$は$i$ 番目のトークンに着目したラベル付与の確信度と見なし，参考文献文字列中の各トークンに対するラベル付与の確信度の中で最小のものを，その参考文献文字列の書誌情報抽出の確信度とする．具体的には以下の式で定義される．
  $$C_{MP}(\mathbf{x}) = \min_{1 \le i \le |\mathbf{x}|} \max_{l \in L} P(Y_i = l|\mathbf{x}) \quad (4)$$

- Average Token Entropy (ATE)
  この確信度は，全ラベル候補の周辺確率のエントロピーに基づいて定められる．エントロピーの値が大きいほど，より多くの書誌要素ラベルに確率が分散しているためラベル付与が困難であると判断する．具体的には以下の式で定義される．
  $$C_{ATE}(\mathbf{x}) = -\frac{\sum_{1 \le i \le |\mathbf{x}|} \sum_{l \in L} -P(Y=l|\mathbf{x}) \log P(Y=l|\mathbf{x})}{|\mathbf{x}|} \quad (5)$$

2.2 能動サンプリング

2.1 節で示した確信度を用いて，本研究では，川上ら[2]と同様に，以下の手順で能動サンプリングを行う．まず，ラベル未付与の参考文献文字列S を大量に収集する．次に，少量の参考文献文字列$S_0 \subset S$ を選出して，ラベルを付与しこれを第 1 回目の学習データとして CRF M_0を学習する．その後，以下の手順を繰り返す． CRF M_{t-1} を用いて，参考文献文字列$S - \bigcup_{i=0}^{t-1} S_i$の確信度をそれぞれ算出し，各確信

度の昇順にランキングする．その後，上位n件の参考文献文字列$S_t \subset S - \bigcup_{i=0}^{t-1} S_i$を学習データとして人手でラベルを付与し，ラベル付与した参考文献文字列$\bigcup_{i=0}^{t} S_i$を用いてCRF M_tを学習する． これは書誌情報抽出が困難なサンプルは学習に有効であるという考え方に基づく．

3　評価実験

3.1　実験概要

能動サンプリングにおいてどのような確信度が有効であるか実験により評価する．実験データとして，以下の参考文献文字列コーパスを利用する．

IEICE-J　2000 年の電子情報通信学会和文論文誌に含まれる参考文献文字列 4,787 件（内，和文 2193 件）

IEICE-E　2000 年の電子情報通信学会英文論文誌に含まれる参考文献文字列 4,787 件（内，和文 0 件）

IPSJ　2000 年の情報処理学会論文誌に含まれる参考文献文字列 4,574 件（内，和文 1,537 件）

また，書誌要素ラベルは，Author，Title 等の 18 種類のラベルを定義したが，評価の際には荒内ら[4]の研究に倣って Title や Booktitle 等の似ている種類のラベルを 9 種類に再分類したものを用いる．再分類の際には，同じ書誌要素は正解判定において区別しない．評価指標として，参考文献文字列を構成する全てのトークンに正しく書誌要素ラベルを付与できた参考文献文字列数を全参考文献文字列数で割ったものを用いる．また，5 分割交差検定で書誌情報抽出精度を算出する．

3.2　能動サンプリングの効果

能動サンプリングの効果を図 1，2，3 に示す．ここで，2 初期学習データは学習用データSから無作為に 10 件選出し，その後，確信度が低い参考文献文字列を 10 件ずつ学習データに加える．また，ラベルを付与する参考文献文字列を無作為に選出した場合を RAND と記し，ベースラインとする．図 1 より，IEICE-J では，MP が有効であるとわかる．図 2 より，IEICE-E では，比較的 ATE が有効であることがわかる．また，30 件の学習データ件数においては，NNLH も有効であることがわかる．図 3 より，IPSJ では，TM が有効であることがわかる．また，学習データ件数が 80 件を超えた後は，NNLH が高い書誌情報抽出精度を実現している．また，全ての学術論文誌において，NNLH は NLH に比べ高い抽出精度が得られた．特に，図 3 の IPSJ では，その効果が顕著であった．

4　まとめ

本稿では，能動サンプリングにおける有効な確信度を検討した．実験の結果，IEICE-J では MP，IEICE-E では ATE，NNLH，IPSJ では TM，NNLH が比較的有効だとわかった．今後の課題としては，複数の確信度を組み合わせ，有効な組み合わせを見つけることがあげられる．

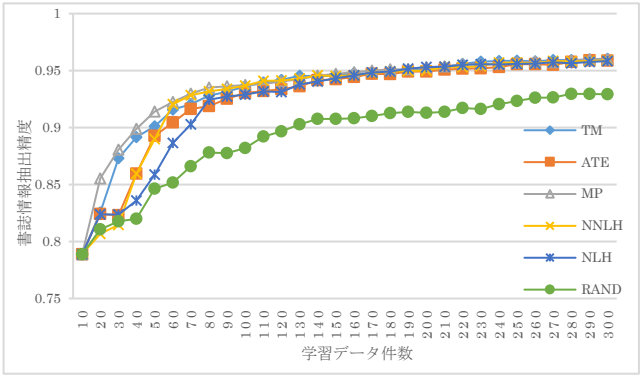

図 1：能動サンプリングにおける書誌情報抽出精度（IEICE-J）

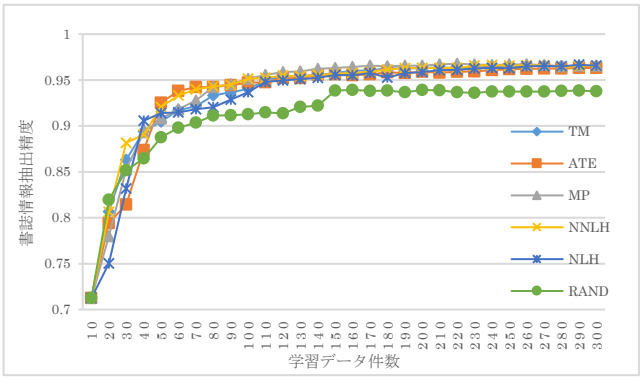

図 2：能動サンプリングにおける書誌情報抽出精度（IEICE-E）

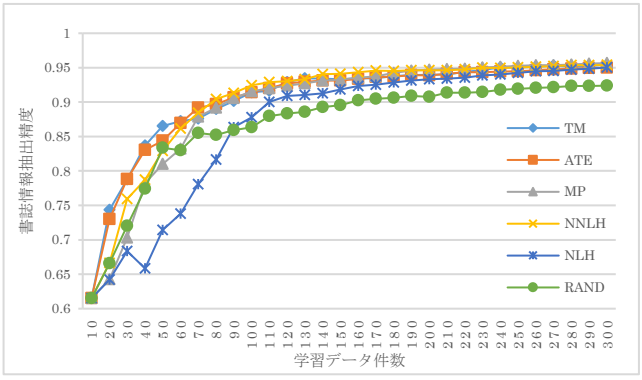

図 3：能動サンプリングにおける書誌情報抽出精度（IPSJ）

参考文献

[1] Lafferty，J.，McCallum，A. and Pereira，F.: Conditional Random Fields ： Probabilistic Models for Segmenting and Labeling Sequence Data，In Proc．of 18th International Conference on Machine Leaning，pp. 282-289(2001).

[2] 川上尚慶，太田学，高須淳宏，安達淳：少量学習データによる参考文献書誌情報抽出精度の向上，情報処理学会論文誌データベース，Vol. 8，No. 2，pp. 1–12 (2015).

[3] Settles，B. and Craven，M .: An Analysis of Active Leaning Strategies for Sequence Labeling Tasks，In Proc. Of the Conference on Empirical Methods in Natural Language Processing (EMNLP)，pp. 1070-1079，ACL Press(2008).

[4] 荒内大貴，太田学，高須淳宏，安達淳：CRF による和英文の参考文献文字列からの自動書誌要素抽出，情報処理学会研究報告，Vol.2012-DBS-156，No.1，pp.1‐8(2012).