ISS-P-058

能動学習を用いた実験情報抽出の一手法

八田谷 翔太† 太田 学††
†岡山大学工学部情報系学科 ††岡山大学大学院自然科学研究科

1. はじめに

近年，学術論文データベースの充実により，膨大な数の論文を手軽に入手できるようになり，文章を効率よく読むことを支援する研究が行われている．

本稿では，学術論文から能動学習を利用して Conditional Random Field(CRF)による実験情報抽出を行い，無作為に選択した学習データで学習した場合と比較し，高い抽出精度が得られることを実験により示す．

2 関連研究

平井ら[1]は，論文中の実験に関連する記述を実験情報，また図や表等を論文構成要素と定義し，これらの自動抽出を行った．彼らは，まず論文構成要素をルールで抽出し，その抽出結果を用いて CRF で実験情報を抽出する 2 段階抽出と，論文構成要素を抽出せずに，直接 CRF で実験情報である論文構成要素を抽出する 1 段階抽出を行っており，2 段階抽出の方が，抽出精度が高かったことが報告されている．

川上ら[2]は，参考文献書誌情報の自動抽出において能動学習を適用することで，必要な学習データ量の削減に成功した． 本研究では，平井らの実験情報抽出において能動学習が有効であることを確認する．

3. CRF による実験情報抽出

本研究では，平井ら[1]と同じ方法で実験情報を抽出する．まず，論文 PDF ファイルを XML ファイルに変換する．2 段階抽出では，論文 XML ファイル中の TEXT にルールにより論文構成要素ラベルを付与し，その結果を用いて TEXT が実験情報か否かを CRF により判定する．1 段階抽出では，論文 XML ファイルから，直接 CRF によって実験情報である論文構成要素を抽出する．

4. 能動学習

機械学習において，少量学習データで高精度な結果を得ようとすると，効率よく学習を進める必要がある．本稿では，ある時点での学習モデルで実験情報抽出が困難なデータを優先的に学習させ，モデルを更新する．この時，実験情報抽出の困難さを表す尺度として式(1)の確信度[2]を定義する．

$$C_{MP}(\mathbf{x}) = \min_{1 \leqq i \leqq |\mathbf{x}|} \max_{l \in L} P(Y_i = l|\mathbf{x}) \qquad (1)$$

ここで $\mathbf{x}$ は入力 TEXT の系列，Y_iはi番目の TEXT に付与されうるラベルを表す確率変数，L は付与できるラベルの集合，$P(Y_i = l|\mathbf{x})$はラベル$l \in L$がx_iに割り当てられる周辺確率である．すると$\max_{l \in L} P(Y_i = l|\mathbf{x})$は$i$番目の TEXT に注目したラベル付与の確信度といえるため，式(1)は学習データに対する実験情報抽出の確信度を表している．本稿では，この確信度の値が低いほど抽出が困難と判定し，逐次データを追加する．

5. 評価実験

能動学習を適用して実験情報抽出実験を行った． 2 段階抽出と， 1 段階抽出の 2 通りを評価する．式(1)の確信度に基づいてデータを選出した場合と無作為にデータを選出した場合を比較する．両者ともに無作為に選出した 10 件を初期学習データとする． なお，この 10 件は同一のデータである．また逐次追加するデータも 10 件とする．

抽出結果を図 1 に示す．図 1 は，2 段階と 1 段階ともにそれぞれ 10 回の抽出結果の平均を示している．2 段階抽出と 1 段階抽出の両者ともに，能動学習を適用した方が抽出精度(F 値)は高かった．1 段階と 2 段階を比較すると，2 段階の方が，抽出精度が高かった．

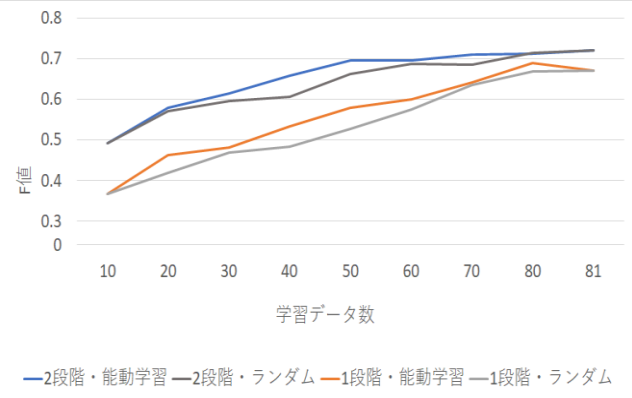

図1. 実験情報の抽出結果

6. まとめ

能動学習を適用した実験情報抽出を行い，抽出精度を評価した．少量学習データで高い抽出精度を得るために，確信度を定義し，確信度に基づいて学習データを選択した．2 段階抽出と 1 段階抽出どちらの場合も能動学習を適用した方が抽出精度は高く，また 2 段階抽出の方が 1 段階抽出より抽出精度が高かった．

参考文献

[1] 平井久貴，新妻弘崇，太田学，高須淳宏 ： 学術論文からの実験情報抽出の一手法，DEIM Forum 2015，F3-1，2015.

[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（TOD），vol. 8，no. 2，pp. 18-29，2015.