7B-04

医学生物学文献からの数値情報抽出における教師データ削減の検討

大瀧 洋子† 木戸 邦彦† 加藤 千昭† 久保田 一石‡ 安松 勲‡
（株）日立製作所† 第一三共 RD ノバーレ株式会社‡

1. はじめに

医薬品となる化合物の探索では，機械学習を用いたタンパク質と化合物の活性予測が検討されている[1]。この活性予測では，医学生物学論文から数値情報を抽出できれば，データの拡充につながり活性予測の精度向上が期待できる。このようにテキストに記載される数値情報は定量的かつ客観的な情報であるため情報としての価値が高く，正確で網羅的な数値情報の抽出技術が求められている。

数値情報は，属性を表すキーワードと属性値を表す数値の対としてテキストに記載される。属性と属性値の対を抽出する手法については，報告[2]において提案された機械学習を用いた属性と属性値対同定手法が適用できると考えられる。この手法における課題は教師データの作成コストが高いことにあった。属性と属性値の対を抽出する問題に対応するその都度，膨大な教師データを作成することは現実的ではない。

そこで本稿では，数値情報抽出，特に属性と属性値対同定における教師データ削減を目的とし，能動学習を用いることを検討する。

2. 抽出対象の数値情報

本稿では医学生物学論文から抽出する数値情報を ATP 濃度の記述とする。ATP 濃度は，図１に示すように実線枠のような属性候補と，破線枠のような属性値として表される。属性候補と属性値候補間の実線は，属性と対応する属性値であることを示し，破線は属性と対応する属性値ではないことを示している。

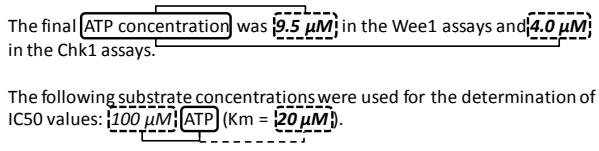

図１：数値情報の記載事例

医学生物学論文における ATP 濃度の記載は次のような傾向がある。

- 属性候補や属性値候補は，一定のルールに基づき記載されている。
- 属性値候補は必ずしも ATP 濃度の値ではなく，属性と属性値の対応付けが必要だが文が自由に記載されルール化が困難。
  一方，属性・属性値の間や前後に出現する単語・文字は類似している。
- 同一文中に属性と属性値の記載がある。文をまたぐ事例はない。また，属性と属性値間の距離は近い。
- １つの属性に対して，複数の属性値が対応する事例がある。

これらの傾向から属性候補と属性値候補の抽出には，正規表現を含むルールで行い，属性と属性値の対の同定には機械学習ベースの手法[2]を用いることが適切であると考えられる。

3. 今回検討した教師データ作成手法

機械学習において性能が高いモデルを作成するには，本稿で検討している属性と属性値の対の同定においても同様に膨大な教師データが必要であり，課題となっている。そこで，属性と属性値の対の同定に能動学習のアプローチを取り入れ，教師データを作成する対象データを効果的に選択し，教師データの作成コストを削減することを検討する。

能動学習では，識別境界の決定に寄与しそうな少数のサンプルのみにラベルを付与することで，効率よくモデルを学習することを目的としている。能動学習のアプローチを取り入れた学習の流れを図２に示す。まず，ラベルが付与された少数の教師データで (a)機械学習アルゴリズムを用いて学習を行い，(b)属性-属性値対同定モデルを生成する。次に，(b)で得られた識別境界に対して，(c)決定に寄与しそうなサンプルをラベル無しサンプルから選択する。(d)サンプルに対してラベルを付与し，再学習を行う。

この能動学習で重要となるのはどのサンプルにラベルを付与するかというサンプル選択である。本稿では 3 つのサンプル選択手法を実装し，教師データ削減コストの比較を行う。

Training Data Reduction for Attribute - Numerical Value Pair Extraction from Biomedical Literature
† Hitachi, Ltd.
‡ DAIICHI SANKYO RD NOVARE CO., LTD

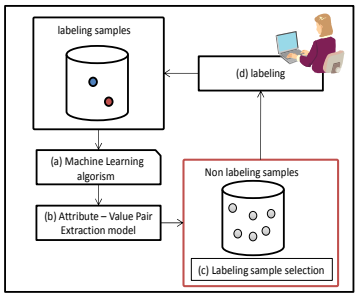

図２：能動学習のフロー

（1）Margin Sampling

サンプル選択手法として Margin Sampling と呼ばれる手法[3]を取り上げる。サンプルの１番目に確率の高いクラス確率と 2 番目に確率の高いクラス確率の差を指標とし，差が最も小さいサンプルをクラス付与の対象とする手法である。

（2）Positive Candidate Margin Sampling

サンプルの選択手法として，Margin Sampling の改良手法として Positive Candidate Margin Sampling を試す。負例に対して正例が少数となる傾向にあるため，正例に対する学習の促進を目的とし，正例候補でかつ最も判別境界に近いサンプルを選択し，ラベルを付与し教師データに追加する。

（3）Random Sampling

サンプルの選択手法として，Random Sampling を試す。この手法は，他の手法のベースラインとして用いる。プールから 1 件のデータをランダムに選択し，ラベルを付与し教師データに追加する手法である。

これらの 3 種類のサンプル選択手法の教師データ削減コストについて次章で比較検討する。

4. 評価実験

4.1. 実験条件

実験データとしては，医学生物学論文 718 件を対象とし，ATP 濃度の抽出を行った。ラベル付与する対象サンプルをサンプル選択手法で選択し，目標精度 F 値 0.8 に達するのに必要な教師データ数を計測した。本実験では，従来法として大量の教師データを使用した場合との比較も行いたいため，全属性候補と属性値候補のペアに対してラベルを付与した。データ数は，5,683 ペアであり，そのうち属性と属性値間に関係性があったのは，995 ペアであった。用いた特徴量数は 15,586 となっている。

4.2. 実験結果

データ追加に伴う F 値の推移を図３に示す。この図３は Margin Sampling, Positive Candidate Margin Sampling, Random Sampling の結果をまとめた図となっている。図中の□で囲んだ数値は，各手法で目標精度 F 値 0.8 を達成したときの教師データ数を表している。

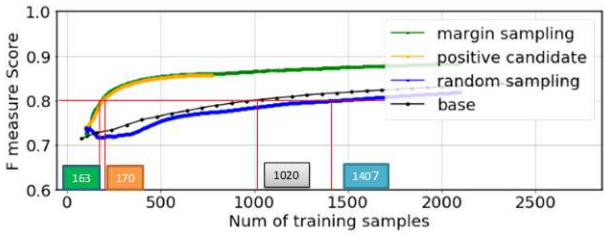

図３：３手法による F 値の推移比較

従来法(base)では，1,020 件のデータで F 値 0.8 に達し，Margin Sampling では 163 件，Positive Candidate Margin Sampling では 170 件，Random Sampling1,407 件で F 値 0.8 を達成する結果となっており，Margin Sampling の識別境界に最も近いデータをサンプリングする手法の教師データ作成コストが最も低く，検討した中では最良の手法となった。識別境界に近いデータを追加することにより，識別境界が微調整されたことで精度向上の速度が Random Sampling より早かったと考えられる。

予め大量の教師データを準備し学習を行う従来法では，1,020 件の教師データにより目標精度を達成できるが，Margin Sampling では 163 件のラベル付与を行えばよいということを考えると，従来と比べ教師データ作成コストを 84.0%（=(1,020-163)/1,020×100）削減可能となった。

5. まとめ

本稿では，テキストに記載されている数値情報を高精度に抽出する技術の開発において，課題であった教師データ作成コストに着目した。

教師データを作成する対象データを効果的に選択する手法について検討し，識別境界に，最も近いデータを選択し，教師データに追加するサンプル選択手法の教師データ作成コストが最も低く，従来と比べ教師データ作成コストを 84.0%削減可能なことを確認した。

[1]浜中 雅俊 他：深層学習に基づくタンパク質と化合物の相互作用予測，情報処理学会第 77 回全国大会，4B-07，2015.
[2]飯田 龍 他：意見抽出を目的とした機械学習による属性-属性値対同定，情報処理学会研究報告，1,21-28,2005.
[3]Simon Tong: Support Vector Machine Active Learning with Application to Text Classification, Journal of Machine Learning Research (2001) 45-66