PDF 形式の論文データからの章の構造推定に関する研究

松戸　直樹

現在，多くの論文がインターネット上にアップロードされている．このことが原因で，自らが望んだ論文を探すことが困難となっていたり，既に存在する手法や事柄についての論文を再び執筆しかねない事などの様々な問題が起こっている．このような問題を解決するために，学術分野におけるデータマイニングの研究が行われている．その研究分野では主にニューラルネットワークという手法が用いられ，学習データが大量に必要になる．しかし論文のデータは PDF というプログラムで扱いづらい形式で配布されている場合が多く，そのような問題を解決するために様々な情報抽出ツールが存在している．しかし，メタデータに関しての抽出に関しての研究は多くされているが，本文の章の構造の抽出に関しての研究はあまり行われておらず，改善の余地があると考えられる．

PDF の論文データからメタデータを抽出できるツールの一つに CERMINE がある．CERMINE は本文データを抽出した際のデータは本文とそれ以外の構造のみの抽出となるので，表，式や図を別々で抽出しない．

本研究では PDF から章の構造を推定し本文の章の構造を抽出することを目的とする．

次に提案手法について述べていく，提案手法は論文からの章の構造推定を系列ラベリング問題に置き換えて推定することである．まずは pdf2json というツールを用いて y 軸やテキストなどの抽出を行う．次にその抽出されたテキストに対してアノテーションというその文がどの構造に属するかを人が判断してラベルを付与する作業を行う．次にそのデータを用いて長期的な依存関係を学習することのできる LSTM という手法を用いてそのテキストがどのラベルに属するかを単語のみを手掛かりとして推定できるモデルを作成し，それぞれのテキストについて特徴スコアを出力として得る．最後に系列全体で最適なラベルを推定できる CRF という機械学習手法を用いる．

実験設定は train データを 9 本，validation データを 1 本，test データを 1 本使用し，CRF，LSTM，LSTM-CRF の 3 つの手法を用いた結果を比較する．

実験結果は，精度 74%，再現率 68%，F1 値 71%で圧倒的に CRF のみを用いたモデルが良い結果を出した．そのように結果になった考えられる原因は LSTM 層の出力スコアがあまり良くなかったため，CRF 層を用いてもあまり良い結果が出なかったと考えられる．LSTM 層のスコアが良くなかった原因は，使用できた論文数が少なかったことが考えられる．LSTM のような機械学習手法を用いる際には一般的に大量の学習データがあるほど良いスコアが得られるとされているからである．今後の展望に関しては学習データとして使用する論文のデータをどのようにして増やしていくか，あるいは少ないデータでも良いスコアが出るようなモデルを考えていく必要があると考えている．

（指導教員　若林啓）