

深層学習に基づく学際的主題における 頻出項の発見

2班（情報・人間科学） 趙 秋涵 Zhao Qiuhan
 北京郵電大学 情報工学研究科 修士指導教員：楊 文川 教授

1. はじめに

近年、情報技術の急速な進歩、新しい語彙と知識が絶えず生まれ出されている。科学研究の分野では、さまざまな分野の移行、応用、統合により、多くの新しい学際的な分野が徐々に導き出されてきた。これまでのデータベースまたは検索ツールの分類方法では、学際的な主題に頻出項を発見することは困難である。

この論文では、深層学習に基づいて科学技術文献記述子(descriptor)を生成し、適応パラメータクラスタリング(Adaptive Parameters Clustering, APC)と組み合わせて、頻繁に出現する異分野であるが、研究の方向性が似たトピックが組み合わされた集合、すなわちバッチを記述および発見する方法の研究を行った。

2. 研究の方法と流れ

① データ取得と前処理

この論文では、既存のソフトウェア Hadoop に基づく分散クローラーを使用し、中国最大の文献検索データベースである CNKI¹から、10 のクラスで合計 168 のサブクラスを収集した。各サブクラスは 500、合計 84,000 の文献である。

次に、上記データを、単語のセグメンテーション(Word Segmentation)など前処理し、統計的特性に従ってデータのエンコード(Encode)を完了した。そして、後続のモデルで使用するために、データ拡張(Data Augmentation)と単語ベクトル(Word Embedding)などの事前トレーニング

(Pre-training) をした。

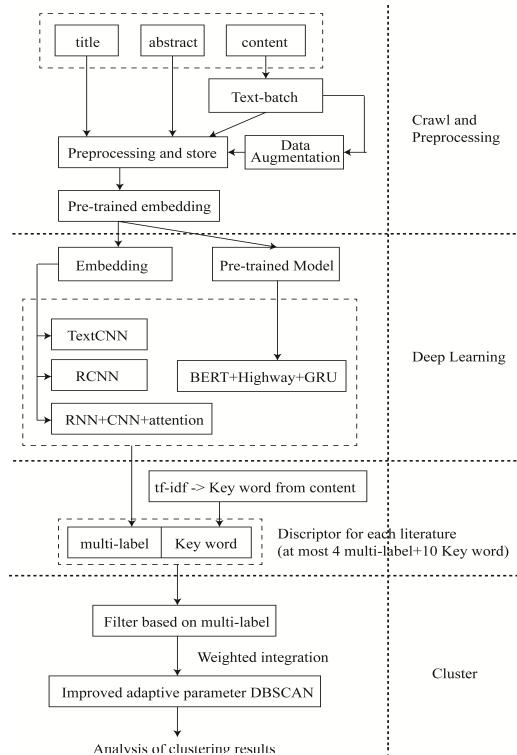


図1 全体的なフレームワーク

② 深層学習

この部分では、さまざまな深層学習モデルを使用して、文献のマルチラベル(multi-label)を生成した。各ラベルの単語ベクトルの次元は、テスト用に 200, 300, および 400 の次元である。そして、統計に基づく tf-idf 方法を使用して、別のキーワードを作った。「マルチラベル+キーワード」は、各文献の記述子として使用される。これには、最大 10 語、つまり $10 * (200 \sim 400)$ 次元のベクトルが含まれる。

¹ CNKI(China National Knowledge Infrastructure): www.cnki.net

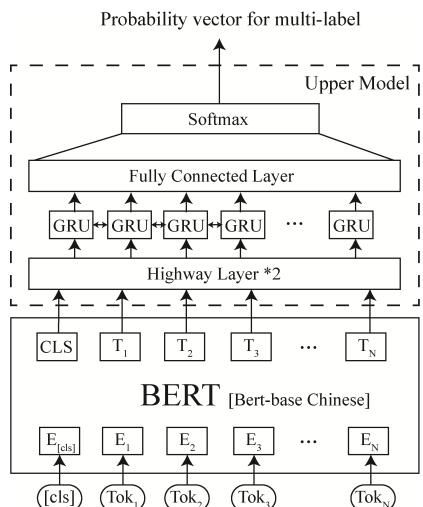


図2 深層学習の例—Bert+Highway+GRU

③ 適応パラメータクラスタリング(APC)

APC を設計し、文献記述子のバッチ($n * 10 * (200\sim 400)$ 次元)をクラスタリングし、それらの中で頻出項を発見する。アルゴリズムの詳しい流れはこの web ページ²に詳述する。

合計 4000 個の 2 次元数から成るランダムポイントを人為的に生成し、設計した APC を使用してその正当性を検証した。結果を図 3 に示す。

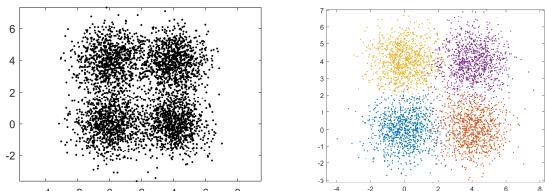


図3 クラスタリング効果の検証

3. 実験結果

事前トレーニングの単語ベクトルとデータ拡張を使用して、候補となる 4 つの深層学習方法の効果を比較する。効果を表 2 に示す。

効果の指標として P(Precision)、R(Recall)、F1 を使用する。P は、予測の結果に基づいて予測が正であるサンプルのうち、真に正であるサンプルの数を表す。R は、元の正のサンプルで正しく予測された正のサンプルの数を表す。F1

は P と R を包括的に考慮した指標である。

表2 4つの深層学習方法の効果

アイテム	P / R / F1
TextCNN	0.8102 / 0.8024 / 0.8063
RCNN	0.8213 / 0.8160 / 0.8186
CNN+RNN+attention	0.8281 / 0.8377 / 0.8329
BERT (FT+TM)	0.8650 / 0.8555 / 0.8602

モデルのパラメータを調整することにより、モデルの効果をさらに最適化する。これらのパラメータには、Batch Size、学習率、トレーニング方法が含まれる。

実験後、Batch Size を 256 に選択すると、学習率が自動的に最適化され、FT + TM(Fine tune + Training Model)方法で、F1 を 0.8740 に到達させた。さらに、収束性(convergence)を進めてモデルの訓練を迅速化することもできる。

4. 結論

クラスタリング効果の検証は、この論文で提案された適応パラメータクラスタリングアルゴリズム(APC)の正しさを証明した。同時に、パラメーター最適化後の深層学習モデルも高い精度を持っている。

この論文で提案された方法は、文献の新しいトピックのバッチを効果的にマイニングでき、優れた実用性を備えている。論文の詳しい内容については、HP³に掲示する。

5. 今後の研究計画

博士課程は東京大学の工学研究科で、技術経営戦略学(TMI)についての研究を進めていく。

具体的には、中米貿易戦を背景に、自然言語処理と深層学習を組み合わせて、科学文献と特許データを通じて、中国とアメリカの革新的な産業(AI、Bigdata、IoT など)の動向と発展を分析することである。

² (pp. 19) www.zhaoliuhuan.cn/files/slide/vsbs.pdf

³ HP: www.zhaoliuhuan.cn/