(別紙様式博5)

学位論文要旨

学位授与申請者
大内　智仁

題目：不要文削除及び重要文追加によるデータ拡張に関する研究

本研究は、ディープラーニングを用いた自動要約システムにおいて、学習データのデータ拡張手法について検討し、トピックモデルで評価した重要度により、最も重要でない一文を削除したデータを拡張データとして用いる手法が、比較検討した手法の中で最も効果的であることを明らかにした。

第1章　序論

近年、ネット上の情報量は日々、指数関数的に増加している。このような状況の中で、情報を取捨選択する必要性は高まってきている。そのため、これからの時代、できるだけ効率よく情報の内容を理解するために、自動要約という技術の需要は高まってくるものと思われる。自動要約システムにおいて、自然な要約を作るためには、生成型要約システムというのが必須となってくる。しかし、生成型要約システムを構築するには、大量の学習データが必要となる。そして、大量の学習データは、学習記事に対して、人手で要約をつけていかなければならず、大量のコストがかかってしまう。そのために、自動要約システムにおけるデータ拡張という技術によって、少ないデータ量でも自動要約の精度を高めることが必要となってくる。

第2章　関連研究

自動要約システムには、抽出型要約と呼ばれるものと、生成型要約と呼ばれるものがある。抽出型要約には、大きく分けて、3つの種類があり、それぞれGraph Base手法と、Feature Base手法と、Topic Base手法がある。生成型要約は、最初、Encode-Decoderモデルを基にした手法が提案された。しかし、記事特有の固有名詞などが生成できないなどの問題があった。これに対し、Copy Mechanismというものが提案されて、解消されたが、未だ、要約が同じフレーズが繰り返されるという問題があった。そこでCoverage Mechanismというものが提案された。本研究では、Copy MechanismとCoverage Mechanismを組み合わせたPointer-Generatorモデルというものを用いて実験を行う。

第3章　実験条件

本研究では、要約の精度を測るのにROUGEという評価指標を用いた。これは、人間の作った要約と、モデルが生成した要約の単語の一致率を測るというものである。また、実験に用いたデータセットには、CNN/DailyMailデータセットを用いた。このデータセット

には、訓練データが 287、226 記事、検証データが、13、368 時、テストデータが、11、490 記事存在する。

第 4 章　自動要約システムにおけるデータ拡張の有効性

第 4 章では、Encoder-Decoder モデルに Attention Mechanism を加えたモデルを使って、データ拡張の有効性について検証した。データ拡張の方法として、入力記事に対して、トピックモデルを用いて各文の重要度を算出し、最も重要度の低い一文を抜いたものを拡張データとする。元データに加えて拡張データを加えて学習したものを提案手法と呼ぶ。実験の結果、提案手法の有効性が確認された。

第 5 章　Pointer-Generator モデルにおけるデータ拡張の有効性

第 5 章では、第 4 章のモデルの代わりに、Pointer-Generator モデルを用いて、提案手法の有効性を検証した。実験した結果は、提案手法の有効性が確認された。

第 6 章　不要文削除によるデータ拡張に関する研究

第 6 章では、比較手法として、EDA、　LexRank、Luhn という手法を用いたデータ拡張手法を用いた。EDA 手法は、ある単語を同義語に置き換える手法と、ある単語の同義語を記事のランダムな位置に挿入する手法と、ランダムな二つの単語を入れ替える手法と、ランダムな単語を削除する手法と、それら 4 つの手法を組み合わせた手法の 5 つの手法からなる。LexRank は、検索ランキングアルゴリズムの PageRank 手法を応用した方法で、この手法で文の重要度を測定し、最も重要度が低かった一文を抜く手法を比較手法とした。Luhn は、単語の位置と頻度を基に、文の重要度を測る手法で、この手法で最も重要度の低かった一文を抜く手法を比較手法とした。結果は、提案手法が最も精度が高かった。

第 7 章　Pointer-Generator モデルにおける不要文削除と重要文追加によるデータ拡張手法の比較

第 7 章では、提案手法として、重要文を追加する手法を提案する。ここで、今まで提案手法としてきた手法を不要文削除手法（“remove”手法）と名付け、新たに提案した手法を重要文追加手法（“add”手法）と名付ける。“add”手法は、最重要文を記事の最初に付けるか最後に付けるかによって、二つに分けられる。それぞれ、“add-s”手法、“add-e”手法と呼ぶ。実験の結果、“remove”手法が最も良い結果となった。

第 8 章　不要文削除と重要文追加を組み合わせたデータ拡張手法

第 8 章では、提案手法として、不要文削除と重要文追加を組み合わせた手法（“hybrid”手法）を提案する。この手法も二つに分けられて、最重要文を最初につけて不要文を削除する手法を“hybrid-s”手法、最重要文を最後につけて不要文を削除する手法を“hybrid-e”手

法と呼ぶことにする。実験の結果、“remove”が最も良い結果となった。

第 9 章　記事数と拡張の効果の関係性

第 9 章では、記事数と拡張の効果の関係性について実験した。実験の結果、記事数が少なければ少ないほど拡張の効果が高いことが示された。

第 10 章　生成した要約例

第 10 章では、第 6 章で生成した拡張なし手法と提案手法の要約例を示す。

第 11 章　結言

本研究では、自動要約システムにおけるデータ拡張について実験を行った。自動要約システムには、Pointer-Generator モデルを使用し、そのモデルにおいてデータ拡張することにおける ROUGE 値の増加が確認できた。特に、効果が高かった手法として、不要文を一文削除するというものであった。また、不要文を選択する手法も、Luhn や LexRank によって決めるよりもトピックモデルを用いて選択する手法が最も効果が高かったことが確認された。また、EDA を用いたデータ拡張を自動要約システムに適応したのも本研究が初めてである。特に、SR 手法において、拡張なし手法より効果があることが確認された。が、提案手法には及ばなかった。また、第 7 章以降で提案した重要文追加手法や不要文削除手法と組み合わせたハイブリッド手法なども試し、拡張なし手法より効果があることが確認されたが、提案手法には及ばなかった。

以上