論文 / 著書情報
Article / Book Information

題目(和文)	高品質教師データが得られないドメインにおける要約手法の研究
Title(English)	
著者(和文)	狩野竜示
Author(English)	Ryuji Kano
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第11758号, 授与年月日:2022年3月26日, 学位の種別:課程博士, 審査員:奥村 学,熊澤 逸夫,中山 実,篠﨑 隆宏,船越 孝太郎
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11758号, Conferred date:2022/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報通信 系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 (工学) Doctor of
学生氏名： Student's Name	狩野 竜示	指導教員（主）： Academic Supervisor(main)	奥村 学
		指導教員（副）： Academic Supervisor(sub)	

要旨（和文 2000 字程度）
Thesis Summary（approx.2000 Japanese Characters ）

ニューラルネットワークの台頭により，自動要約手法は目覚ましい発展を遂げたが，多くの研究は教師データが整備された一部のドメインに集中して行われてきた．その代表例が新聞記事を対象としたものである．これは，新聞記事の見出しを要約とみなすことで本文から要約を生成，あるいは抽出する機械学習モデルを学習させることができるからである．ただし，世の中で要約が望まれている文章は新聞記事に限らない．ソーシャルメディア，メール，チャット，医療文書など様々なドメインの文書が要約対象となりうる． 新聞記事は文章執筆を生業とした記者が不特定多数の読者向けに執筆したものであるため，その質がある程度担保されている．それに対し，メールは個人間のプライベートなやり取りの中で書かれたものであり，ソーシャルメディアは匿名性の高い場で非業務的活動として行われているため，質が低いものが多く含まれている．要約は本文の重要箇所を反映しているべきという前提があるが，メールやソーシャルメディアの件名，タイトルは本文の重要情報が記載されていなかったり，本文に無い情報を含むことがある．

こうした教師データが未整備のドメインの文書を要約する際に，2 つのアプローチが存在する． 1 つ目のアプローチは，教師データを使用しない教師なし手法である． 2 つ目は，品質の低い教師データを使って効率的に学習を行う方法である． 本論文ではこれら 2 つのアプローチにおいて，それぞれ新しい手法を提案する．

第 1 章の序論では，既存の要約研究の流れとその課題「教師データが未整備なドメインに対する要約研究の不足」について論じ，その解決手段としての 2 つのアプローチについて議論する．

第 2 章の関連研究では，要約モデルの発展について述べながら，上記 2 つのアプローチに関して，これまでにどのような関連研究が行われてきたかを論じる．

第 3 章では，1 つ目のアプローチである「教師なし要約手法」について論じる．まず，既存の教師なし抽出型要約手法の課題を述べ，既存手法とは異なる要約の指標と，それに基づいたモデルを提案する．既存の教師なし要約手法の多くは，頻度を元に要約としてのふさわしさを定量化している．ここでは，頻繁に言及される話題は重要であるという仮定が前提にされている．ただし，実際には重要な話題が必ずしも多く言及されるとは限らない．そこで新たな手法として，返信による言及のされやすさを重要度の指標として考慮したモデルを提案する．提案したモデルは，要約対象の本文と，返信の対を用いて学習を行う．モデルは正しい返信とランダムサンプリングにより得られた偽返信であるかを判別する学習を行う．その際，Gumbel Softmax により本文の一部の文を抽出し，正しい返信と偽返信の判別に使用する．モデルは学習の過程で返信によって言及されやすい文を本文から抽出するようになるため，これを要約とみなして評価する．提案手法をメールデータセット及び，ソーシャルメディアデータセットで評価し，従来手法と同等もしくは上回る性能を発揮することを確認した．また，提案手法が従来手法では抽出できない重要文を抽出できていることを定量的定性的両方の分析で確認した．

第 4 章では，2 番目のアプローチである「低品質教師データの効率的学習」として，カリキュラムラーニングを要約データセットに利用した手法を提案する．カリキュラムラーニングは，ある指標に基づき学習するデータの順序を変更することでモデルの性能を上げる手法であり，ノイズを含むデータセットに対する有効性が示されてきた．ただし，既存研究における手法では高品質教師データと，低品質学習データ両方を必要とする課題があった．今回，このカリキュラムラーニングを要約データセットに応用するにあたって，単一のノイズを含むデータセットからノイズを定量化する手法を提案し，既存手法よりも高い性能を発揮することをメールデータセット及びソーシャルメディアデータセットを用いて確認した．また，これまでカリキュラムラーニングに用いられてこなかった抽出率や含意判定確率が要約データセットに対するカリキュラムラーニングに有効であることも合わせて示す．

第 5 章では第 3 章と第 4 章で提案した手法に対する課題と今後の展望について述べる．

備考：論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note：Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報通信 系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 (工学) Doctor of
学生氏名： Student's Name	狩野 竜示	指導教員（主）： Academic Supervisor(main)	奥村 学
		指導教員（副）： Academic Supervisor(sub)	

要旨（英文 300 語程度）
Thesis Summary（approx.300 English Words ）

Rise of neural networks led to the great advancement of automatic summarization. Most of the previous research focused on domains with organized datasets such as news articles. However, there are many other domains of texts that are demanded to be summarized. In this paper, we discuss two approaches to deal with summarization when there is no organized training dataset.

The first approach is unsupervised summarization. Previous methods of unsupervised summarization are based on an assumption that important topics are frequently referred. However, in fact, important topics can only be referred once. To overcome the Achilles heel of the previous methods, we propose ″probability of being referred to by replies″ as an important factor of summarization. Our proposed model extracts sentences that are more likely to be mentioned by replies by distinguishing true replies from false replies. Experimental results on mail and social media datasets show our model outperforms or performs equally as the previous methods of unsupervised summarization.

The second approach is to efficiently train summarization models from noisy datasets. In news article summarization, we leverage headlines as summaries. Some other datasets such as mail datasets and social media datasets also have titles and subjects but the quality is lower compared with news headlines. One way to train models efficiently from noisy datasets is to use curriculum learning. Previous methods quantified noise of translation training data using a model trained by a noisy dataset and a model trained by a clean dataset. However, there are no such datasets in summarization fields. We propose a method to quantify noise from a single corpus. We conduct experiments on mail and social media datasets to verify our method outperforms previous methods.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).