論文 / 著書情報
Article / Book Information

題目(和文)	ニューラル文書要約の高度化に関する研究
Title(English)	
著者(和文)	牧野拓哉
Author(English)	Takuya Makino
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11676号, 授与年月日:2020年12月31日, 学位の種別:課程博士, 審査員:奥村 学,船越 孝太郎,熊澤 逸夫,高村 大也,篠崎 隆宏
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11676号, Conferred date:2020/12/31, 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 情報通信	系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of	（ 工学 ）
学生氏名： Student's Name	牧野拓哉		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 ）

文書要約技術は与えられた文書を簡潔にした短い文書や文を自動作成する技術であり，たとえば，テレビ局や新聞社における Web ページや電光掲示板といった各媒体向けの新聞記事の要約作成での応用が期待されている．これら要約作成は，媒体ごとに定義された要約長制約や表現に従った人手による作業であり，負荷軽減が求められている．

本論文は「ニューラル文書要約の高度化に関する研究」と題し，全 6 章より構成されている．

第 1 章「序論」では，文書要約手法が抽出型手法，生成型手法に分類されることおよび，それぞれの特徴について説明する．また生成型要約の中でも，ニューラルネットワークに基づく文書要約である生成型ニューラル要約モデルのうち，特に，原文書からの単語のコピーおよび生成用語彙（あらかじめ構築した単語の集合）からの単語の生成によって要約を生成可能な生成型ニューラル要約モデルの発展により高い要約精度が得られるようになってきたことを述べる．一方で，既存の生成型ニューラル要約モデルは，実際の要約作成では要約長制約があるのにもかかわらず要約長制約を超える要約の生成が少なくない点および，生成用語彙の既存構築方法によって要約特有な単語の生成割合が低下し，要約精度に影響しうる点を指摘し，本研究で解くべき課題を述べる．

第 2 章「関連研究」では，抽出型手法と生成型手法の先行研究，評価方法について説明する．抽出型手法では文分類に基づく手法，組み合わせ最適化手法，系列ラベリングに関する手法についてまとめる．生成型手法では統計的要約手法，生成型ニューラル要約モデルについてまとめている．生成型ニューラル要約モデルに関する先行研究については，本論文で提案する手法との関係について述べる．

第 3 章「生成型ニューラル要約モデルの概要」では，本論文で用いる生成型ニューラル要約モデルの概要について述べる．生成型ニューラル要約モデルは原文書をベクトルにエンコードしたのち，逐次的に単語の生成確率を計算し，確率が最大となる単語を生成用語彙から出力することで要約を作成していることを説明する．

第 4 章「要約長制約下における生成型ニューラル要約モデルの学習」では，既存手法が実際の要約作成で設定される要約長制約を超える要約の生成が少なくないという課題に対して，文書要約の自動評価尺度である ROUGE 値が改善するように要約モデルを学習する minimum risk training に，要約長制約を超える要約の生成に罰則を追加する手法を提案する．提案手法は，従来手法と異なり，要約長制約内で ROUGE 値が高い要約の生成確率を向上させ，ROUGE 値が低い要約および，ROUGE 値が高くても要約長制約を超える要約の生成確率を低下させることを述べる．また，提案手法を英語・日本語のデータで評価し，従来手法よりも要約長制約内での要約生成の割合が増加するとともに，ROUGE 値が改善することを示す．さらには，人手による自動生成要約の後編集評価実験から，要約長制約内で要約を生成することが人手による後編集時間の削減に寄与することを示す．

第 5 章「原文書と要約対の差分に基づく生成用語彙構築」では，既存の生成用語彙構築手法は，要約側に出現する言い換えのような，要約特有な表現が生成用語彙から生成されにくく，要約精度に影響しうるという課題に対して，要約側にだけ出現する単語に着目した生成用語彙構築手法を提案する．提案手法は要約に高頻度で出現する単語を対象とした既存の生成用語彙構築手法と異なり，原文書からのコピーによって要約に含めることのできる単語を除きつつ要約特有の単語を生成用語彙に含めるために，原文書と要約対の差分をとることで，要約にのみ出現する高頻度語を用いて生成用語彙を構築する方法であることを説明する．提案手法を 2 つの英語のベンチマークデータで評価し，従来の生成用語彙構築手法と比較し，高い ROUGE 値となりつつ，要約特有の表現を多く生成できることを示す．

第 6 章「結論と今後の課題」では，本研究の結論と今後の課題を説明している．結論では本論文で提案した 2 つの手法によって，生成型ニューラル要約モデルの課題を改善していることを述べている．今後の課題では，本論文では取り組んでいないものの，残された重要な課題を述べる．

（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 情報通信	系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of	（ 工学 ）
学生氏名： Student's Name	牧野拓哉		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 ）

Automatic document summarization aims at generating coherent and short summary for given document(s). Automatic document summarization technique can be used in automatic creation of summaries for news articles. Because of the cost of manual creation of summaries in companies such as newspaper company and television company, developing improved summarization methods are expected to support human laborers.

In this study, we develop two methods for remained problems especially in abstractive summarization methods for single document summarization task. First, although summary-length controllable neural summarization models have been studied, there are still some over-length summaries against the length constraint. Second, the existing output vocabulary construction method may degrade summarization accuracy because some summary-specific words are not included in the output vocabulary while words that often occur in source documents are included.

For the first problem, we propose a training method for generating summaries that are in-length against the length constraint while obtaining high summarization accuracy. The proposed method is based on minimum risk training that trains a model for an arbitrary evaluation metric. Unlike the previous method that rewards summaries that have high similarity with reference summaries, the proposed method rewards summaries that has high similarity with reference summaries while penalizing over-length summaries.

For the second problem, we propose an output vocabulary construction method to include summary-specific words. Unlike the previous method that uses summaries for constructing the output vocabulary, the proposed method constructs the output vocabulary by using difference between a source document and reference summary. The proposed method can generate more summary-specific words than the previous method because the proposed method constructs the output vocabulary by collecting words that occur in a summary and not in a source document.

We empirically verify the effectiveness of the proposed methods. We argue that these methods contribute to improve summarization accuracy of summarization 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).