論文 / 著書情報
Article / Book Information

題目(和文)	単一文書要約の高度化に関する研究
Title(English)	
著者(和文)	菊池悠太
Author(English)	Yuta Kikuchi
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10377号, 授与年月日:2016年12月31日, 学位の種別:課程博士, 審査員:高村 大也,新田 克己,寺野 隆雄,奥村 学,小野 功
Citation(English)	Degree:, Conferring organization: Tokyo Institute of Technology, Report number:甲第10377号, Conferred date:2016/12/31, 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

専攻： Department of	知能システム科学　専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of　（　工学　）
学生氏名： Student's Name	菊池悠太	指導教員（主）： Academic Advisor(main)	高村大也
		指導教員（副）： Academic Advisor(sub)	

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

文書要約は，入力として単一の文書あるいは複数の関連した文書集合を受け取り，その内容を短くまとめた要約を出力する課題である．長い歴史をもつ文書要約研究のうち，本研究が具体的に取り組む手法は，単一文書を対象とした抽出に基づく要約手法（抽出型要約）である．抽出型要約の抽出単位としては，文からはじまり，近年ではより高い圧縮率を実現するために文圧縮などの技術で文の一部のみを要約に組み込む手法が盛んに研究されている．また，古典的な要約の生成手段は，抽出単位の重要度の高いものから選択していくものであったが，今日においては要約を整数計画問題（ILP）として定式化し大域最適解を求めることで情報の網羅性を高める枠組みが支配的な手段となっている．また，ILP の登場以前から蓄積されていた知見やアイディアを，ILP という新しい枠組みに適用する試みも近年行われるようになっている．また抽出単位の重要度を決定する方法として，機械学習を利用する試みは古くからあるものの，大量の訓練データの不足という問題が常に存在してきた．

本論文は，単一文書を対象とする要約課題において，これまでの研究の歴史を踏まえた上で残された重要な課題のうち， 大きく二つに焦点を当てる．本論文は，それぞれの問題に有効な貢献をすることで単一文書要約を高度化することを目的としており，「単一文書要約の高度化に関する研究」と題し，全 5 章より構成されている．

第 1 章（序論）では，本研究の背景および目的を述べる．これまでの文書要約研究の歴史と残されている課題について，単一文書要約と複数文書要約の違いを中心に振り返る．それらを踏まえ，本研究で取り組む二つの問題について説明する．

第 2 章（関連研究）では，本研究に関連する従来の研究について紹介する．従来提案されてきた ILP による定式化の例や文圧縮の方法，談話構造の利用などいくつかの技術のほか，文書要約における機械学習の利用事例を中心に説明する．

第 3 章（文間の依存関係を考慮した文抽出と文圧縮の同時最適化手法）では，本研究で取り組んだ一つ目の課題について述べる．従来高い精度が確認されている文抽出と文圧縮の同時最適化モデルに，新たな構造的制約として修辞構造に基づく文と文の間の依存関係を組み込む新たなモデルを提案する．提案手法を従来の同時最適化モデルや木制約付きナップサック問題による要約手法と比較評価したところ，文書要約の自動評価指標である ROUGE において最高精度が得られることを確認した．また，高い圧縮率が要求される要約設定へのさらなる柔軟性を獲得するため，文圧縮の制約を緩和するよう拡張を加え，その有効性を示す．

第 4 章（要約器の効果的な訓練のための大規模要約資源の活用手法）では，本研究で取り組んだ文書要約研究おける二つ目の課題について述べる．ターゲットとなる少数の整備された訓練データに加え，大規模であるが整備のされていない要約資源である New York Times Annotated Corpus（NYTAC）がある状況で，後者を有効に要約器の訓練に活用するための手法を提案する．ドメイン適応の分野から標準的な 5 つの手法を取り上げ， さらに訓練する要約器の特性に併せた事例のフィルタリングをオプションとして用意することで，それらの組み合わせが訓練に与える影響を確かめる．実験の結果，要約器を NYTAC で一度訓練したあと，そのパラメータを初期値として所望のターゲットデータで追加的に訓練する手法が有効に働くことが分かった．加えて，使用する要約器の特性に合わせた事例選択を事前に行うことで，一部の評価データセットでは更に精度が大きく向上することを確認した．

第 5 章（結論と今後の課題）では，本研究において単一文書要約の高度化のために取り組んだ二つの課題について達成した内容をまとめ，今後の展望や残された課題について述べる．ここでは，今回取り組んだ課題にかぎらず，文書要約研究全体を通して今後取り組むべき問題についてまとめている．

本論文では，単一文書要約の高度化に関して主に二つの課題に焦点を当てた．要約研究全体でみると，インターネットの発達に伴い 2000 年代以降は複数文書要約が主な研究対象となってきたが．単一文書要約には複数文書要約とは異なる需要や技術的課題が存在している．また，近年の技術の蓄積や大規模なデータの出現など，単一文書要約をより高度化させるための材料が揃いつつあり，一部の研究者の焦点が単一文書要約へ再び集まり始めているという事実もある．そのため今後は単一文書要約も再び盛り上がりを見せていくことが期待される．そのような状況において，本論文は，単一文書要約における二つの重要な課題に取り組み，その有効性を確認しており，今後の単一文書要約研究の発展につながると期待できる．

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	知能システム科学　専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of　（　工学　）
学生氏名： Student's Name	菊池悠太	指導教員（主）： Academic Advisor(main)	高村大也
		指導教員（副）： Academic Advisor(sub)	

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

Text summarization has a long history in natural language processing. The most well-studied approach to text summarization is the extractive approach, which selects a subset of the linguistic units (e.g., sentences, clauses, and words) consisting the input document(s). Formulating extractive summarization as a combinational optimization problem greatly improves the quality of summaries. There has recently been an increasing attention to approaches that jointly optimize sentence extraction and sentence compression.

In this study, we focus on two problems and develop new methods in different layers of problems that are remained in text summarization study, especially in single document summarization. First, although the dependency between words has been used in a number of existing methods based on the joint optimization of sentence extraction and compression, the dependency between sentences has not been exploited in such joint methods. Second, one big problem in text summarization is the lack of training data.

For the first problem, we make use of discourse structure of a document, i.e., rhetorical structure, to obtain the dependency relation between sentences. We use both dependency between words and dependency between sentences by constructing a nested tree, in which nodes in the document tree representing dependency between sentences were replaced by a sentence tree representing dependency between words. We formulated a summarization task as a combinatorial optimization problem, in which the nested tree was trimmed without losing important content in the source document.

For the second problem, we propose methods that attempt to use a large amount of summaries contained in the New York Times Annotated Corpus (NYTAC). We introduce five methods inspired by domain adaptation techniques in other research areas to train our supervised summarization system. We also propose an instance selection method according to the faithfulness of the extractive oracle summary to the reference summary.

We empirically verify the effectiveness of the methods addressing both of these problems. We argue that both of these methods contribute to improve and accelerate the future of single document summarization study.

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).