様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１７　日現在

機関番号：　１２６０８
研究種目：　基盤研究(B)（一般）
研究期間：　2014～2016
課題番号：　２６２８００８０

研究課題名（和文）大規模データに基づいた機械学習による抽出的および非抽出的文書要約手法の開発

研究課題名（英文）Development of methods for extractive and abstractive text summarization based on machine learning with large-scale data

研究代表者

高村　大也（Takamura, Hiroya）

東京工業大学・科学技術創成研究院・准教授

研究者番号：８０３６１７７３

交付決定額（研究期間全体）：（直接経費）　　12,500,000円

研究成果の概要（和文）：要約技術の開発に必要となる大規模要約データを自動構築する技術、またそれを効果的に利用する技術を開発した。また、入力文書に対し、文分割、文圧縮、文融合などの演算を施した上で要約を生成する技術、およびウェブページの推薦システムにおいて、ユーザにカスタマイズしたスニペットを生成する技術を開発した。また、野球のイニング速報を自動的に生成する技術を開発した。さらに、ニューラルネットワークに基づく文要約手法において、出力長を制御する技術を開発した。また、日本語の文圧縮のための大量のデータを自動的に抽出する手法を開発し、実際にこの手法を用いて大規模データを構築し、文圧縮モデルの学習を行った。

研究成果の概要（英文）：We developed a method for automatically creating a large-scale training dataset for text summarization as well as a technique to make use of the dataset. We also developed a method for incorporating sentence division, sentence merge, and sentence compression as components for single-document text summarization. We also developed a method for generating personalized snippets. We also developed a method for creating inning summaries for baseball matches. In the field of neural network-based summarization, we developed a method for controlling the output length for sequence-to-sequence summarization model, and a method for automatically creating a large-scale dataset for Japanese sentence compression.

研究分野：　自然言語処理

キーワード：　文書要約　機械学習　大規模データ　ニューラルネットワーク

１．研究開始当初の背景

電子化テキストの増加を背景にして，自動文書要約の重要性が叫ばれて久しいが，世の中で必要とされているレベルの要約技術の開発には成功していない．

自然言語処理分野における要約研究に大きな影響を与えているのは，アメリカ合衆国の NIST が主催している Document Understanding Conference (DUC)及びその後継の Text Analysis Conference (TAC)である．DUC/TAC は，文書要約の実験データを用意し，いわゆる共通課題形式で世界中の研究チームに要約精度を競わせた．単文書要約課題は 2001,2002 年のみ実施され，それ以降は複数文書要約課題，クエリ指向型要約課題，update 要約や guided 要約など次々と新たな要約課題を始めた．しかし，これらの課題に対して，十分に有効な手法が開発されてきたわけではなく，基本的な単文書要約課題でも，先頭の数文を選ぶだけのリード手法というベースラインとほとんど同等な性能しか実現できていない．複数文書要約課題においても，表層的な情報だけを用いての抽出的要約生成では，これ以上の大幅な性能向上は望めない状況にある．

一方，文書のような構造的出力が可能である機械学習手法が発展してきている．機械翻訳のような言語生成を必要とする研究課題においても，大規模なデータから翻訳知識を抽出し機械学習の枠組みに乗せることで成果が上がっている．また，計算機の能力も向上してきており，従来では活用しきれなかった量のデータをうまく活用できるようになってきている．

このような状況にも関わらず，文書要約では，表層的な手がかりのみに依存し機械学習手法を活かすことができていない．その主な原因は要約データ(元文書と要約文書のペア)の不足である．DUC などの共通要約課題での元文書(集合)と要約文書のペア数はせいぜい 50 程度である．このデータは学習用というより評価用であり，このサイズでは要約文書を生成するための詳細な仕組みの学習は非常に困難である．例えば，DUC2004 の要約文書中の文数は約 4,000 であるのに対し，機械翻訳データである Hansard コーパス中の文ペア数は約 130 万であることからも，データ不足の深刻さが覗える．また，データ不足が原因となり，要約生成という特殊な問題に利用できる機械学習手法は発展していない．既存研究の中にも，各文を要約に入れるか否かの二値分類問題と捉えたり，構造学習の枠組みで文集合を選択する問題と捉えたりして機械学習を導入した研究はあるが，元文書中の各文の重要度を，元文書中でのその文の位置や，その文に含まれる単語の文書全体での頻度を考慮することにより学習するといった試みであり，要約生成の際に各表現がどのように書き換えられるかなどについては学習できていない．

しかし，ここ数年でウェブコンテンツがますます充実してきており，大規模な要約データを構築するための基盤が整ってきた．最も特筆すべきは，一部の新聞記事に人手で要約を付けて配布するサービスが多く出てきていることである．また，Wikipedia のデータも大きくなってきており，そこからも要約データを抽出できる可能性がある．また，新聞記事データ中にも要約技術の開発につながるデータが含まれていることを我々は見出し，大規模要約データに基づく非抽出的要約の要素技術を開発し始めている．

２．研究の目的

本研究課題では，大量の要約データ(元文書-要約文書ペア)に基づく機械学習手法を用いた，抽出的および非抽出的文書要約手法を開発する．そのため，まずは大規模な要約データを構築し，アラインメント技術を開発することで要約研究へ利用できる形にする．その大規模要約データに基づいて機械学習手法を用いて，同じ話題について述べている部分を集めた上で，単語の削除や言い換えを行うことで，より柔軟性が高く圧縮率の高い要約生成を実現する．さらに，複数文書要約へ適用できるように開発手法への拡張を行う．開発した手法は，文書要約ツールという形で，一般に公開する予定である．

３．研究の方法

大量の要約データの入手が重要であるので，３つのルートを用意して確実に準備する．その上で，以下の技術の開発を行う：要約データのアラインメント技術，文書の分割技術，元文書と要約文書及びそれらを横断した共参照解析技術，話題語を中心とした文圧縮と文書要約の技術，元文書にない表現の生成を取り入れたより柔軟な非抽出的要約生成の技術，文書全体の話題語の記述を集めて要約することによるより圧縮率の高い要約生成の技術，複数文書要約への適応方法を探る．

４．研究成果

・New York Times Annotated Corpus は、要約付きの約 65 万の記事を含む。しかし、この要約には様々な種類があり、要約器の学習データとしては適切でない。そこで、ここから学習データとして使用可能なデータを抽出することで要約データを構築した。これにより、約 15 万の元文書-要約文書ペアを獲得できた。またこれらからさらに抽出型の要約アプローチで作成可能な要約文書を取り出すことで, 約 1 万の元文書-抽出型要約文書ペアを獲得した。さらに、要約モデルの学習に適切な学習事例を選択する方法、またドメイン

アダプテーションの手法により大量のデータで学習した要約モデルを、ターゲットとなるドメインにチューニングする方法を開発した。この手法を用いて、実際に文書要約モデルを学習し、高い精度で要約を生成することに成功した。
・単一文書要約のために必要となる、文に対する様々な編集操作を開発した。特に、文を分割する方法、複数の文を融合する方法を開発し、それらを文書要約手法に組み込んだ。より具体的には、分割すべき文を選択したうえで分割し、さらに融合すべき文を選択したうえで融合し、続けてそれらの出力と元々の文から不要な単語を削除することで文を圧縮した。こうすることで多くの候補文を生成し、これらの候補文の中から適切なものを選んで整列させることで要約を生成する手法になっている。最終段階の生成では、最大被覆要約モデルを用いる。これにより、要約における各文の長さを適切に制御することができるようになった。
・野球の打者成績を簡潔に文書で伝えるために、打者成績からイニング速報を自動生成する手法を開発した。打者成績は、すべてのイベントがイニング速報として記述されるわけでなく、重要なイベントが選ばれて記述される。また、複数のイベントがまとめられて一つの文で記述されることもある。このような性質を持つ要約生成問題であり、これを条件付き確率場によりモデル化した。これは、日本語の形態素解析で使われている技術を要約生成に応用したものである。
・製品レビューを対象とした単一文書要約において、どのような箇所が重要であるかを自動的に推定する方法を開発した。推定には、文を重要度にしたがってランキングするランク学習を用いている。
・ニューラルネットワークに基づく文要約手法において、出力長を制御する手法を開発した。これにより、入力文の内容を保ちつつ、自然な文を出力できるようになった。この技術のプログラムは、インターネット上で一般公開している。
・日本語の文圧縮のための大量のデータを自動的に抽出する手法を開発し、実際にこの手法を用いて大規模データを構築し、文圧縮モデルの学習を行った。
・ウェブページの推薦システムにおいて、ユーザにカスタマイズしたスニペットを生成する手法を開発し、実際にこの手法が生成したスニペットが、ウェブページの要約として有用であることを示した。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1　件）

① 渡邉亮彦, 笹野遼平, 高村大也, 奥村学, 「Web ページ推薦システムにおけるユーザ指向型スニペット生成」 人工知能学会論文誌, Vol.31, No.5, p. C-G41_1-12, 2016.

〔学会発表〕（計　10　件）

① Shun Hasegawa, Yuta Kikuchi, Hiroya Takamura and Manabu Okumura. ``Japanese Sentence Compression with a Large Training Dataset''. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL2017), 2017.

②長谷川駿，菊池悠太，高村大也，奥村学．「大規模データを用いた日本語文圧縮」言語処理学会第 23 回年次大会，pp. 4-7，2017.

③ Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura and Manabu Okumura. “Controlling Output Length in Neural Encoder-Decoders”. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), pp. 1328-1338, 2016.

④ Miho Matsunagi, Ryohei Sasano, Hiroya Takamura and Manabu Okumura. “Acquiring Activities of People Engaged in Certain Occupations”. In Proceedings of the 14th Pacific Rim International Conferences on Artificial Intelligence (PRICAI 2016), pp. 327-339, 2016.

⑤ Yuta Kikuchi, Akihiko Watanabe, Ryohei Sasano, Hiroya Takamura and Manabu Okumura. “Learning from Numerous Untailored Summaries”. In Proceedings of the 14th Pacific Rim International Conferences on Artificial Intelligence (PRICAI 2016), pp. 206-219, 2016.

⑥ 菊池悠太，Graham Neubig，笹野遼平，高村大也，奥村学．「Encoder-Decoder モデルにおける出力長制御」 情報処理学会自然言語処理研究会，2016-NL-227(5)，pp. 1-9，July 2016.

⑦ 村上聡一朗，笹野遼平，高村大也，奥村学．「打者成績からのイニング速報の自動生成」言語処理学会第 22 回年次大会，pp. 338-341，2016.

⑧ 田中駿，笹野遼平，高村大也，奥村学．「要約長,文長,文数制約付きニュース記事要約」言語処理学会 第 22 回年次大会，pp.342-345，2016.

⑨ 小池将郎，笹野遼平，高村大也，奥村学．「レビューを対象とした単一文書要約」言語処理学会 第 22 回年次大会，pp.346-349，2016.

⑩ 菊池悠太，渡邉亮彦，高村大也，奥村学．「重要箇所同定用コーパスの構築―New York Times Annotated Corpus の文書要約資源化に向けて― 」言語処理学会第 21 回年次大会，pp.593-596，2015．

〔図書〕（計　0　件）

〔産業財産権〕

○出願状況（計　2　件）

名称：文書要約装置
発明者：田中駿、笹野遼平、高村大也、奥村学
権利者：国立大学法人東京工業大学
種類：特許
番号：特願 2016-035558
出願年月日：2016 年 2 月 26 日
国内外の別：国内

名称：野球のイニング速報の自動生成器
発明者：村上聡一朗、笹野遼平、高村大也、奥村学
権利者：国立大学法人東京工業大学
種類：特許
番号：特願 2016-035560
出願年月日：2016 年 2 月 26 日
国内外の別：国内

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
https://github.com/kiyukuta/lencon

６．研究組織
(1)研究代表者
高村 大也　（TAKAMURA, Hiroya）
東京工業大学・科学技術創成研究院・准教授
研究者番号：80361773

(2)研究分担者
笹野 遼平（SASANO, Ryohei）
東京工業大学・科学技術創成研究院・助教
研究者番号： 70603918

(3)連携研究者
なし

(4)研究協力者
なし