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

令和　２ 年　 ６ 月　 ５ 日現在

機関番号：１３９０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００４６０

研究課題名（和文）法令のあらましの自動生成

研究課題名（英文）Automatic generation of the Outlines of Japanese Statutes

研究代表者
小川　泰弘（OGAWA, Yasuhiro）

名古屋大学・情報基盤センター・准教授

研究者番号：７０３３２７０７

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究では，法情報の広範な発信を目的に，法令の自動要約に取り組んだ．当初は機械翻訳の手法を利用したが期待した成果が得られなかったため，新たにランダムフォレストを用いて重要文を抽出する手法を開発し，これにより従来よりも高い性能を示した．
研究の後半では，この成果を政治情報である議会会議録の要約にも適用し，ランダムフォレストに基づく新たな手法を開発した．この手法は，評価型ワークショップNTCIR-14 QA Lab-PoliInfoにおいて，参加7チーム14システム中，人手評価で2位，自動評価で1位の成績を収めた．

研究成果の学術的意義や社会的意義
本研究成果における学術的な意義は，学習データが比較的少ない場合におけるランダムフォレストの有効性を示した点にある．現在有効とされるニューラルネットに基づく手法は大量の学習データを必要とするが，本研究の対象では大量のデータが用意できないため，そうした場合にはランダムフォレストの方が効果的であることを示した．また，機械学習における不均衡データに問題に対しても，漸進的アンサンブルランダムフォレストという手法を開発することにより，新たな解決策を示した．
また，これまで取り組まれて来なかった法令や議会会議録の自動要約について成果を上げた点に，本研究の社会的意義がある．

研究成果の概要（英文）：In this research, we worked on automatic summarization of statutes for the purpose of widely disseminating legal information. Initially, we used the machine translation method, but not worked well. Therefore, we used a random forest method to extract important sentences and archived higher performance than traditional methods.
In the latter half of the research, we applied this method to summarization of assembly member speeches. We developed a new method based on the random forest and it got the second place in the manual evaluation and the first place in the automatic evaluation at the NTCIR-14 QA Lab-PoliInfo workshop, in which seven teams and 14 systems participated.

研究分野：自然言語処理，法情報処理

キーワード：自動要約　機械学習　ランダムフォレスト　法令のあらまし　統計的機械翻訳

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

法令は社会の設計図であり，法令を理解することにより，権利や義務などの国民の生活に関する多くの情報を得ることができる．さらに近年では，社会のグローバル化に伴う国際経済活動の活性化や法制度の透明化のため，わが国の法情報を国際的に広く発信し，諸外国の間で法情報を幅広く共有することが，国内外から強く求められている．しかし，法令は，1 文が長かったり，構文が複雑であったりするため，一般に読みにくく，理解しにくい文書であるとされている．また細部を規定する条文などもあるため，文書全体の分量が多い場合があり，その内容の把握は容易ではない．

そこで我々は，官報に掲載される法令の要約文書である「法令のあらまし」に着目してきた．日本の法令情報を海外に発信するために，「法令のあらまし」の翻訳に取り組んできたが，その一方で，あらましの生成は研究対象としてこなかった．しかし，分かりやすい法情報発信のためには，法令の要約である，あらましの生成も求められる．

また近年では，機械学習に基づく自動要約の研究が始められているが，日本語の学習コーパス（要約元と要約後の文書のセット）は充分な量が存在しなかった．しかし法令とそのあらましを要約コーパスとみなすことにより，学習データとして利用することが期待できる．

２．研究の目的

本研究の目的は，分かりやすい日本法の情報発信のために，法令の要約を自動的に生成することである．「法令のあらまし」は，国民の法令理解の促進を目的とし，昭和 48 年 4 月から官報に掲載されている．あらまし作成の対象となるのは，法令のうち，法律・政令・条約であるが，本研究ではもっぱら法律を対象とする．

ここで，政府があらましを作成するのであれば，法律の要約は不要と思われるかもしれないが，そうではない．まず，あらましは昭和 48 年 4 月以前に公布された法律には存在しない．平成 30 年 10 月 1 日時点で有効な法律は 2,261 件あり，これは前述のあらましが存在する新規制定法律 1,156 件の約 2 倍である．しかも，法律はしばしば改正される．実際，上記の 1,156 件中，840 件以上が改正されており，そうした法律のあらましは，現在の法律の要約になっていない可能性がある．よって，現在有効な法律の概要を知るために，自動要約は必要である．

３．研究の方法

本研究においては，最初は法令の自動要約に取り組んだが，研究後半において，同じ政治情報である議会文書の要約にも取り組んだため，それについても記述する．

３．１．重要文抽出による法律の要約

研究当初は，要約元の文書から要約を作成する過程を翻訳とみなし，統計的機械翻訳の技術を利用した．しかし成果が芳しくなかったため，機械学習に基づく分類器を用いる手法など，様々な手法を試みた．

機械学習のためには，そもそも正解とラベル付けされた学習データ（正解データ）が大量に必要となる．一般に要約とは，重要な部分を抽出し，必要に応じて短縮し，場合によっては複数の文を融合するなどの手順で作成される．しかし「法令のあらまし」の内容を分析したところ，約 50%の文が要約元の文とほぼ同一であった．すなわち法令の要約においては，重要文抽出が最初と課題となる．今回は，どの法令文が抽出されたかを示す正解データを構築した．「法令のあらまし」の場合，どの条項に対応する要約かが明記されており，その情報に基づいて正解データを構築した．

機械学習には様々な手法が提案されているが，要約元に含まれる各文を，重要か否かに分類する分類問題と考え，機械学習に基づく分類器を構築する手法を試みた．従来手法を含めて各種手法を試したが，その中でもランダムフォレストと呼ばれる手法が良い性能を示した．

３．２．議会会議録の要約

本研究で扱う法令は広い意味での政治情報の一つである．本研究の期間中に，評価型ワークショップ NTCIR-14 が開催され，その中のタスクの一つである QA Lab-PoliInfo において，議会会議録の要約というテーマが設定されていた．これは本研究の当初の目的には含まれていないが，本研究で開発した手法の他分野での応用可能性を検証するために，このタスクに参加した．

法令の要約と同じ，ランダムフォレストを用いた手法を適用したが，そのままでは性能が低かった．これは，要約元の文書からどの程度の文を抽出して要約を作成するかという要約率が異なることが原因であった．「法令のあらまし」における要約率は 70%であったが，議会会議録の要約においては 8.3%であった．この場合，抽出すべきデータを正例，そうでないデータを負例と呼ぶが，今回のように正例と負例の比率が極端に異なると，何も抽出しない場合でも正解率が 90%以上になるなど，正しく学習できない問題が生じる．そうした場合，学習データにおける正例・負例の比率を変更するアンダーサンプリング，オーバーサンプリングと呼ばれる手法が提案されているが，どのような比率が良いかはデータごとに異なるため，適切な比率の決定は容易ではない．

それに対して本研究では，正例・負例の比率が異なるデータから学習した複数のランダムフォレスト分類器を組み合わせる漸進的アンサンブルランダムフォレストという新たな要約手法を

開発した.

４．研究成果

４．１．法令要約の研究成果

提案手法の有効性を検証するために，提案手法で用いるランダムフォレストと，それ以外の分類器，さらに学習データを利用しない重要文抽出手法との比較実験を行った.

実験では，昭和 48 年 4 月から平成 31 年 1 月までに新規制定された 1,156 法律のうち，1,111 件を元に，正例 8,793 文，負例 15,507 文からなる学習データを構築した．なお，このデータは先行研究である TSC-1 における重要文抽出タスクで使用されたデータセット 4,521 文の 5 倍以上のサイズである．テストデータは，学習データ構築に利用しなかった法律 33 件に対して人手で正解ラベルを付与して構築した.

実験結果を表 1 に示す．ここでは抽出性能を再現率・精度・F 値で，要約性能を ROUGE で評価している．提案手法であるランダムフォレストを用いた方法が，精度以外の指標で最高の性能を示した．精度で SVM に劣った原因を調べたところ，SVM はランダムフォレストよりも抽出する文数が少ないことが分かった．逆に言えば，その点が SVM の再現率の低さの原因である．要約においては再現率が重視されることから，SVM より 7%多く文を抽出しているにも関わらず，SVM に近い精度を示すランダムフォレストが，もっとも有用であると言える．ROUGE の結果も，この点を示唆している．以上により提案手法の有効性を示した.

システム名	再現率	精度	F 値	ROUGE1
提案手法（ランダムフォレスト）	**84.7**	81.9	**81.0**	**88.4**
TextRank	79.2	80.0	77.4	88.0
線形 SVM	77.7	**82.5**	77.4	85.1
多項式 SVM	77.4	82.1	77.7	84.5
リード法(条単位)	76.7	78.0	74.8	81.9
決定木	73.4	80.3	74.7	80.1
TextTeaser	73.0	74.2	71.3	68.4
リード法(オリジナル)	70.9	72.9	69.7	77.0

表 1　法令からの重要文抽出実験の評価結果

４．2. 会議録要約の研究成果

NTCIR-14 QA Lab-PoliInfo には 7 チーム 14 システムの参加があった．評価においては，人手による評価と，ROUGE という自動評価の指標を用いた．提案手法は，人手評価で 2 位，自動評価で 1 位の成績を収めた．表 2 に ROUGE による評価結果の一部を示す．これにより提案手法の有効性を示した.

システム名	N1	N2	N3	N4	L	SU4	W1.2
KitAi-01	0.285	0.145	0.090	0.050	0.278	0.154	0.180
KitAi-02	0.254	0.126	0.083	**0.053**	0.247	0.131	0.156
TTECH-01	0.088	0.028	0.015	0.007	0.082	0.033	0.050
提案手法	**0.326**	**0.164**	**0.094**	0.046	**0.315**	**0.168**	**0.201**
akbl-01	0.256	0.113	0.065	0.034	0.247	0.124	0.148
akbl-02	0.200	0.094	0.051	0.032	0.189	0.095	0.109
KSU-01	0.048	0.001	0.000	0.000	0.047	0.007	0.032
KSU-02	0.069	0.014	0.000	0.000	0.067	0.019	0.043
KSU-03	0.041	0.002	0.000	0.000	0.041	0.007	0.027
KSU-04	0.050	0.002	0.000	0.000	0.048	0.008	0.031
KSU-05	0.067	0.002	0.000	0.000	0.062	0.013	0.041
KSU-06	0.053	0.003	0.000	0.000	0.051	0.008	0.034
LisLb-01	0.171	0.083	0.044	0.026	0.160	0.088	0.106
TO-01	0.116	0.055	0.035	0.012	0.111	0.056	0.070

表 2　PoliInfo の要約タスクにおける自動評価の結果（内容語：ROUGE (recall)）

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Ogawa Yasuhiro、Satou Michiaki、Komamizu Takahiro、Toyama Katsuhiko	４．巻 LNCS 11966
２．論文標題 nagoy Team ' s Summarization System at the NTCIR-14 QA Lab-PoliInfo	５．発行年 2019年
３．雑誌名 NII Testbeds and Community for Information Access Research - 14th International Conference	６．最初と最後の頁 110～121
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1007/978-3-030-36805-0_9	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Yasuhiro Ogawa, Michiaki Satou, Takahiro Komamizu, Katsuhiko Toyama
２．発表標題 nagoy Team ' s Summarization System at the NTCIR-14 QA Lab-PoliInfo
３．学会等名 Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies（国際学会）
４．発表年 2019年

１．発表者名 小川泰弘，佐藤充晃，駒水孝裕，外山勝彦
２．発表標題 法律の要約のためのランダムフォレストを用いた重要文抽出
３．学会等名 人工知能学会全国大会(第33回)論文集
４．発表年 2019年

１．発表者名 佐藤充晃，小川泰弘，駒水孝裕，外山勝彦
２．発表標題 分類器を用いた法令要約に利用する法令文の自動抽出
３．学会等名 平成30年度電気・電子・情報関係学会東海支部連合大会
４．発表年 2018年

１．発表者名 佐藤充晃，小川泰弘，大野誠寛，中村誠，外山勝彦
２．発表標題 統計的機械翻訳の利用による法令のあらましの自動生成
３．学会等名 平成29年度電気・電子・情報関係学会東海支部連合大会
４．発表年 2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考