(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2019-16181
(P2019-16181A)
(43)公開日　平成31年1月31日(2019.1.31)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 16/00	(2019.01)	G06F	17/30	220A	5B091
G06F 17/22	(2006.01)	G06F	17/22	664	5B109
G06F 17/27	(2006.01)	G06F	17/27	665	
G06F 16/30	(2019.01)	G06F	17/30	170A	

審査請求 未請求　請求項の数 5　ＯＬ　（全 11 頁）

(21)出願番号　特願2017-133421 (P2017-133421)
(22)出願日　平成29年7月7日 (2017.7.7)

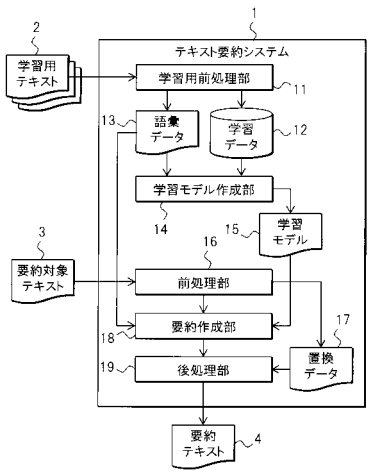
(71)出願人　000155469
株式会社野村総合研究所
東京都千代田区大手町一丁目９番２号
(74)代理人　110002066
特許業務法人筒井国際特許事務所
(72)発明者　北原　真由美
東京都千代田区大手町一丁目９番２号　株式会社野村総合研究所内
(72)発明者　外園　康智
東京都千代田区大手町一丁目９番２号　株式会社野村総合研究所内

Ｆターム(参考)　5B091 AA15　AB01　CA02　CC01　EA01
5B109 ME22　QB14

(54)【発明の名称】テキスト要約システム

(57)【要約】
【課題】テキスト文章の自動要約の精度をより向上させる。
【解決手段】要約対象テキスト３から要約テキスト４を作成するテキスト要約システム１であって、複数の学習用テキスト２に対して、所定の前処理を行って学習データ１２を作成する学習用前処理部１１と、学習データ１２に基づいて機械学習により要約に係る学習モデル１５を作成する学習モデル作成部１４と、要約対象テキスト３に対して所定の前処理を行う前処理部１６と、前処理がなされた要約対象テキスト３に対して、学習モデル１５に基づいて要約テキスト４を作成する要約作成部１８と、要約テキスト４に対して所定の後処理を行って出力する後処理部１９とを有し、前処理では学習用テキスト２および要約対象テキスト３に含まれる語句に所定の加工を行い、後処理では要約テキスト４に対して前処理部１６により行われた加工の内容を復元する。
【選択図】図１

図 1



【特許請求の範囲】
【請求項１】
テキスト文章から要約を作成するテキスト要約システムであって、
原文とその要約を含む複数の学習用テキストに対して、所定の前処理を行って学習データを作成する学習用前処理部と、
前記学習データに基づいて機械学習により要約に係る学習モデルを作成する学習モデル作成部と、
要約対象テキストに対して所定の前処理を行う前処理部と、
前記前処理部により前処理がなされた前記要約対象テキストに対して、前記学習モデルに基づいて要約テキストを作成する要約作成部と、
前記要約テキストに対して所定の後処理を行って出力する後処理部と、を有し、
前記学習用前処理部および前記前処理部における前記所定の前処理では、前記学習用テキストおよび前記要約対象テキストに含まれる語句に所定の加工を行い、
前記後処理部における前記所定の後処理では、前記要約テキストに対して前記前処理部により行われた前記所定の加工の内容を復元する、テキスト要約システム。
【請求項２】
請求項１に記載のテキスト要約システムにおいて、
前記所定の加工は、前記学習用テキストおよび前記要約対象テキストに含まれる数値を所定の記号に置換するものであり、
前記前処理部では、前記置換の際に、前記置換の内容に係る情報を置換データとして記録し、
前記後処理部では、前記置換データの内容に基づいて、前記要約テキストにおける前記所定の記号を対応する前記数値に置換する、テキスト要約システム。
【請求項３】
請求項１に記載のテキスト要約システムにおいて、
前記所定の加工は、前記学習用テキストおよび前記要約対象テキストに含まれる各語句に対して、それぞれ、当該語句の重要度を示す情報を付加するものであり、
前記後処理部では、前記要約テキストにおける前記重要度を示す情報を削除する、テキスト要約システム。
【請求項４】
請求項１に記載のテキスト要約システムにおいて、
前記所定の加工は、前記学習用テキストおよび前記要約対象テキストに含まれる所定の品詞の語句を所定の文字列に置換するものであり、
前記前処理部では、前記置換の際に、前記置換の内容に係る情報を置換データとして記録し、
前記後処理部では、前記置換データの内容に基づいて、前記要約テキストにおける前記所定の文字列を対応する前記所定の品詞の語句に置換する、テキスト要約システム。
【請求項５】
請求項１に記載のテキスト要約システムにおいて、
前記学習用前処理部は、前記学習用テキストにおける要約に含まれる語句のうち、所定の割合以上の語句が、前記学習用テキストにおける原文に含まれているもののみを前記学習データを作成する対象とする、テキスト要約システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、テキスト文章の要約技術に関し、特に、機械学習により生成された学習モデルに基づいて要約を作成するテキスト要約システムに適用するものである。
【背景技術】
【０００２】
機械学習を用いてテキスト文章の要約を自動的に生成する技術が検討されている。要約

を生成する手法には、大別して、抽出型と生成型とがある。抽出型では、例えば、要約の対象となる文章に含まれる重要度の高い単語や文等を抽出し、これらを組み合わせることで要約を作成する。一方、生成型では、例えば、文章を所定の中間表現に変換し、これに基づいて自然言語生成の技術を用いて要約を作成する。現在では、一般的には抽出型の手法が広く研究されており、精度を向上させるための各種の手法が提案されている。

【０００３】

例えば、特開２０１６－１８６７７２号公報（特許文献１）には、要約（短縮文）を作成する対象の文章に含まれる構成要素間を文法的または概念的関係に基づいて連結したツリー構造で表現するとともに、構成要素間のそれぞれの連結に、短縮文に残存する度合いとして得られた結合度を付加し、結合度に基づいて短縮文に含める構成要素を抽出して短縮文を作成することで、自然な短縮文を生成する旨が記載されている。

【０００４】

抽出型の手法では、要約に含まれる単語や文は、原則として原文に含まれる単語や文に制約される。したがって、文章を短縮した場合に不自然な表現や文法的に適切でない要約となってしまう場合がある。これに対し、生成型の手法については、自然な表現を用いることが可能であるが、精度の高い自然言語生成の技術が必要となる等の課題を有している。しかし近年では、例えば、非特許文献１に記載されているような研究もなされており、プログラムやライブラリ等も一般に利用可能となっている。

【先行技術文献】

【特許文献】

【０００５】

【特許文献１】特開２０１６－１８６７７２号公報

【非特許文献】

【０００６】

【非特許文献１】“Research Blog: Text summarization with TensorFlow（登録商標）”、［online］、２０１６年８月２４日、Ｇｏｏｇｌｅ（登録商標、以下同様）、［平成２９年６月２７日検索］、インターネット＜URL：https://research.googleblog.com/2016/08/text-summarization-with-tensorflow.html＞

【発明の概要】

【発明が解決しようとする課題】

【０００７】

抽出型・生成型に関わらず、従来技術のテキスト文章の自動要約技術では、機械学習により生成された学習モデルを用いることで、ある程度の精度の要約を自動的に生成することができる。しかし、精度として十分ではない場合も多く、例えば、原文に含まれる単語等のうち、要約にも現れてほしい重要なものがあるにも関わらず、要約に現れてこないという場合がある等、精度についての改善の余地は多く存在する。

【０００８】

そこで本発明の目的は、テキスト文章の自動要約の精度をより向上させることを可能とするテキスト要約システムを提供することにある。

【０００９】

本発明の前記ならびにその他の目的と新規な特徴は、本明細書の記述および添付図面から明らかになるであろう。

【課題を解決するための手段】

【００１０】

本願において開示される発明のうち、代表的なものの概要を簡単に説明すれば、以下のとおりである。

【００１１】

本発明の代表的な実施の形態によるテキスト要約システムは、テキスト文章から要約を作成するテキスト要約システムであって、原文とその要約を含む複数の学習用テキストに対して、所定の前処理を行って学習データを作成する学習用前処理部と、前記学習データ

に基づいて機械学習により要約に係る学習モデルを作成する学習モデル作成部と、要約対象テキストに対して所定の前処理を行う前処理部と、前記前処理部により前処理がなされた前記要約対象テキストに対して、前記学習モデルに基づいて要約テキストを作成する要約作成部と、前記要約テキストに対して所定の後処理を行って出力する後処理部と、を有する。

【００１２】

そして、前記学習用前処理部および前記前処理部における前記所定の前処理では、前記学習用テキストおよび前記要約対象テキストに含まれる語句に所定の加工を行い、前記後処理部における前記所定の後処理では、前記要約テキストに対して前記前処理部により行われた前記所定の加工の内容を復元する。

【発明の効果】

【００１３】

本願において開示される発明のうち、代表的なものによって得られる効果を簡単に説明すれば以下のとおりである。

【００１４】

すなわち、本発明の代表的な実施の形態によれば、テキスト文章の自動要約の精度をより向上させることが可能となる。

【図面の簡単な説明】

【００１５】

【図１】本発明の一実施の形態であるテキスト要約システムの構成例について概要を示した図である。

【図２】本発明の一実施の形態における事前の学習処理の流れの例について概要を示したフローチャートである。

【図３】本発明の一実施の形態における要約作成処理の流れの例について概要を示したフローチャートである。

【図４】本発明の一実施の形態における前処理および後処理の例について概要を示した図である。

【図５】本発明の一実施の形態における前処理および後処理の他の例について概要を示した図である。

【図６】本発明の一実施の形態における前処理および後処理の他の例について概要を示した図である。

【発明を実施するための形態】

【００１６】

以下、本発明の実施の形態を図面に基づいて詳細に説明する。なお、実施の形態を説明するための全図において、同一部には原則として同一の符号を付し、その繰り返しの説明は省略する。一方で、ある図において符号を付して説明した部位について、他の図の説明の際に再度の図示はしないが同一の符号を付して言及する場合がある。

【００１７】

＜システム構成＞

図１は、本発明の一実施の形態であるテキスト要約システムの構成例について概要を示した図である。本実施の形態のテキスト要約システム１は、機械学習により生成した学習モデルに基づいてニュース記事等のテキスト文章の要約を自動的に生成して出力する機能を有するサーバシステムである。

【００１８】

テキスト要約システム１は、例えば、例えば、サーバ機器やクラウドコンピューティングサービス上に構築された仮想サーバ等により構成される。そして、図示しないＣＰＵ（Central Processing Unit）により、ＨＤＤ（Hard Disk Drive）等の記録装置からメモリ上に展開したＯＳ（Operating System）やＤＢＭＳ（DataBase Management System）、Ｗｅｂサーバプログラム等のミドルウェアや、その上で稼働するソフトウェアを実行することで、自動要約に係る後述する各種機能を実現する。

【００１９】
テキスト要約システム１は、例えば、ソフトウェアとして実装された学習用前処理部１１、学習モデル作成部１４、前処理部１６、要約作成部１８、および後処理部１９等の各部を有する。また、データベースやファイル等として実装された学習データ１２、語彙データ１３、学習モデル１５、および置換データ１７等の各データを有する。学習用前処理部１１、および学習モデル作成部１４は、機械学習における教師データとなる学習用テキスト２に基づいて、機械学習により事前に学習モデル１５を生成する機能を有する。また、前処理部１６、要約作成部１８、および後処理部１９は、要約作成の対象となる要約対象テキスト３について、学習モデル１５に基づいて要約テキスト４を生成して出力する機能を有する。
【００２０】
なお、教師データとなる学習用テキスト２には、要約対象の原文と、正解である要約結果とが含まれている。このような文章としては、例えば、ニュース記事がある。この場合、要約対象となるニュース原文に対して、当該ニュースのタイトルや見出し等を正解である要約結果として用いることができる。本実施の形態では、学習用テキスト２や要約対象テキスト３としてニュース記事を対象に説明するが、これに限られるものではなく、各種の文章に適用することができる。
【００２１】
学習用前処理部１１は、入力となる学習用テキスト２に対して、各種の前処理を施して、機械学習エンジンに入力するために正規化された学習データ１２（および語彙データ１３）を準備する機能を有する。前処理の具体的な内容については後述する。学習モデル作成部１４は、学習用前処理部１１により作成された学習データ１２、および語彙データ１３を入力として、自然言語処理の所定のアルゴリズムを用いて機械学習を行い、要約生成のための学習モデル１５を生成する機能を有する。
【００２２】
機械学習エンジンや自然言語処理のアルゴリズムについては、公知の技術を適宜使用することができる。本実施の形態では、上述の非特許文献１に記載された技術を参照し、例えば、機械学習エンジンとして、Ｇｏｏｇｌｅ社が提供するオープンソースの機械学習ライブラリであるＴｅｎｓｏｒｆｌｏｗ（登録商標、以下同様）を用いる。また、この上で用いる自然言語処理（文章自動要約）のアルゴリズムとして、オープンソースとして提供されているＴｅｘｔｓｕｍのプログラムを用いる。これにより、例えば、ディープラーニングや、ＲＮＮ（Recurrent Neural Network）、ＬＳＴＭ（Long Short-Term Memory）、Sequence to Sequenceモデル、Sequence to Sequence with attentionモデル等の技術を自動要約の際に適用することが可能となる。
【００２３】
前処理部１６は、要約作成の対象となる要約対象テキスト３に対して、上記の学習用前処理部１１の一部と同様の各種前処理を施して、学習モデル１５を適用するために正規化する機能を有する。この前処理の具体的な内容については後述するが、このとき、所定の単語等については、所定の語句や記号等への置換処理が行われ、その結果や内容に係る情報が置換データ１７として記録される。
【００２４】
要約作成部１８は、前処理部１６により正規化された要約対象テキスト３に対して機械学習エンジンにより学習モデル１５を適用して要約を作成する機能を有する。このとき、学習用前処理部１１により作成された語彙データ１３も利用する。機械学習エンジンには、上記と同様に、例えば、Ｔｅｎｓｏｒｆｌｏｗを用いる。なお、ここでの要約は、前処理部１６により行われた置換結果に係る語句や記号等を含んだ状態で作成される。
【００２５】
後処理部１９は、前処理部１６において記録された置換データ１７に基づいて、要約作成部１８により作成された要約における置換結果に係る語句や記号等を元の単語等に置換・復元するとともに、必要に応じて文章の外観を成形して、要約テキスト４として出力す

る機能を有する。なお、後処理部１９では、前処理部１６による要約対象テキスト３に対する置換結果を元の単語等に置換・復元して要約テキスト４を出力しているが、この要約結果を検証するために、同様の後処理を上述の学習モデル作成部１４においても行って、学習用テキスト２に対する学習用前処理部１１による置換結果を元の単語等に置換・復元するようにしてもよい。

【００２６】

＜処理の流れ（学習処理）＞

図２は、本実施の形態における事前の学習処理の流れの例について概要を示したフローチャートである。学習処理では、まず、学習用前処理部１１により、ニュース記事等の学習用テキスト２を読み込み、全ての記事について形態素解析を行って品詞分解を行う（Ｓ０１）。形態素解析は、例えば、ＣｈａＳｅｎ（茶筌）やＭｅＣａｂ（和布蕪）等の一般に入手可能なものも含む各種のプログラムやライブラリを適宜用いて行うことができる。品詞分解により分割した単語や語句の情報の保持方法については特に限定されないが、例えば、学習用テキスト２における対象の単語や語句の区切りの部分に空白を挿入する「分かち書き」により学習用テキスト２に反映させるようにしてもよい。

【００２７】

次に、単語等に分割された状態の学習用テキスト２に対して、学習モデル１５の精度を向上させるための各種の置換処理等の前処理を行う（Ｓ０２）。前処理の内容については後述するが、例えば、学習用テキスト２中に含まれる数値の記載を、桁数も考慮して「＃」等の記号に置換する。また、図示しない辞書データベース等を用いて同義語の表記を統一するように置換してもよい。また、英文の大文字小文字や全角半角を変換して統一するように置換してもよい。このような表記の統一による正規化により、学習用テキスト２において同内容の単語等を集約し、サンプルとしての精度を向上させることができる。単語等の置換に限らず、各単語等に対して重要度等の情報を示すラベルの付加等を行うようにしてもよい。

【００２８】

その後、正規化された学習用テキスト２に含まれる各文章について、所定の選別基準に基づいてノイズ等の不適切な文章を除外するデータクレンジング処理を行う（Ｓ０３）。例えば、学習用テキスト２がニュース記事である場合、要約対象の文章である記事本文に対して、記事のタイトルは正解の要約テキストに相当するが、このタイトルに含まれる名詞が記事本文に含まれていないものばかりである場合は、記事本文に対してタイトルが適切ではないと判断することができる。そこで、例えば、タイトルに含まれる名詞の一定割合（例えば８０％）以上が記事本文にも含まれている文章のみを選別して、これを学習用テキスト２として用いるようにしてもよい。データクレンジングにより残った学習用テキスト２については、これを学習データ１２として記録する（Ｓ０４）。

【００２９】

このとき、学習データ１２に基づいて、これに含まれる各単語等に係るメタデータ等の各種情報を保持する語彙データ１３を併せて生成する（Ｓ０５）。機械学習エンジンとしてＴｅｎｓｏｒｆｌｏｗを用い、文章自動要約のアルゴリズムとしてＴｅｘｔｓｕｍのプログラムを用いる場合、この語彙データ１３は、「ｖｏｃａｂ」ファイルとして作成する。

【００３０】

そして、ステップＳ０４、Ｓ０５で得られた学習データ１２および語彙データ１３を入力として、学習モデル作成部１４により所定の設定条件に基づいて機械学習エンジンによる機械学習を行い、学習モデル１５を作成して（Ｓ０７）、学習処理を終了する。所定の設定条件としては、例えば、学習データ１２の各文章のうち、文頭の２～３文のみを要約の対象とする等の条件を設定することができる。

【００３１】

＜処理の流れ（要約作成処理）＞

図３は、本実施の形態における要約作成処理の流れの例について概要を示したフローチ

ャートである。要約作成処理では、まず、前処理部１６により、ニュース記事等の要約対象テキスト３を読み込み、形態素解析を行って品詞分解を行う（Ｓ１１）。この処理は、図２の学習処理におけるステップＳ０１の処理と同様である。次に、単語等に分割された状態の要約対象テキスト３に対して、図２の学習処理におけるステップＳ０２の処理と同様の前処理を行う（Ｓ１２）。このとき、要約対象テキスト３に含まれる単語等に対して置換を行った場合、置換された単語等の内容や、文章内での出現順序、位置、桁数、単位等の情報を置換データ１７として記録しておく。

【００３２】

その後、前処理による正規化が行われた要約対象テキスト３を入力として、要約作成部１８により、図２の学習処理により作成された学習モデル１５を適用して、機械学習エンジンにより要約を作成する（Ｓ１３）。このとき、上述の図２のステップＳ０５において生成された語彙データ１３も利用する。学習処理と同様に、機械学習エンジンとしてＴｅｎｓｏｒｆｌｏｗを用いることができる。その後、作成された要約の中における置換やラベルの付加等の正規化の内容を、後処理部１９により、置換データ１７を参照して元の単語等に置換・復元し、必要に応じて文章の外観を成形する後処理を行って、要約テキスト４として出力し（Ｓ１４）、要約作成処理を終了する。

【００３３】

＜前処理／後処理の例＞

図４は、本実施の形態における前処理および後処理の例について概要を示した図である。ここでは、学習用テキスト２や要約対象テキスト３に各種の数値が含まれている場合の例について示している。ニュース記事等の文章には、値も桁も異なる多くの種類の数値が含まれている場合があるが、これらの語句を全て異なる語句として取り扱うと、学習用テキスト２においてサンプルが発散して膨大な数となり、語彙データ１３の件数も増えるため、学習モデル１５の精度が低下するとともに学習処理の負荷も増大してしまう。

【００３４】

そこで、本実施の形態では、学習時および要約作成時の前処理（図２のステップＳ０２、図３のステップＳ１２）において数値を「＃」等の記号に置換する。例えば、図４の最上段の文章（分かち書きされた学習用テキスト２および要約対象テキスト３）には、「１２月」と「０．３％」という数値を含む語句が含まれている。この数値部分をそれぞれ、図４の上から２段目の文章のように、「＃＃月」と「＃．＃％」のように「＃」により置換する。

【００３５】

学習モデル作成（図２のステップＳ０７）、および要約作成（図３のステップＳ１３）の処理では、それぞれ、「＃」により置換された状態の文章に対して処理を行い、学習モデル１５の作成、および要約の作成を行う。このとき、要約の文章には、図４の上から３段目の文章のように、「＃＃月」および「＃．＃％」の語句が残存することになる。

【００３６】

本実施の形態では、これらの語句を要約作成時の後処理（図３のステップＳ１４）において元の単語等に戻す。そのために、要約作成時の前処理（図３のステップＳ１２）において数値を「＃」等の記号に置換した際に、置換した数値や文章内の出現位置、桁数、単位等の復元のための情報を置換データ１７に記録しておき、これを後処理時に参照する。これにより図４の最下段の文章のように元の数値を含む要約テキスト４を作成することができる。

【００３７】

図５は、本実施の形態における前処理および後処理の他の例について概要を示した図である。ここでは、学習用テキスト２や要約対象テキスト３に含まれる単語等に対して重要度の情報をラベルとして付加する場合の例について示している。ニュース記事等に含まれる各単語等は、それぞれ、要約作成という観点での重要度が異なる。そこで、本実施の形態では、要約作成において各単語等の重要度を考慮することができるよう、学習時および要約作成時の前処理（図２のステップＳ０２、図３のステップＳ１２）において重要度の

情報を示すラベルを各単語等に付加して一体の単語等とし、これを対象に機械学習の処理を行うものとする。

【００３８】

例えば、図５の最上段の文章（上述の数値置換が行われた結果の学習用テキスト２および要約対象テキスト３）に対して、図５の上から２段目の文章のように、重要度を示すラベルとして「＿Ｘ」（Ｘ＝０～２）の記号を各単語等の末尾にそれぞれ付加して変換する。図５の例では、例えば、「欧州中央銀行」という単語に重要度Ｘ＝０の「＿０」のラベルが付加されて「欧州中央銀行＿０」という語句に変換されている。同様に、例えば、「ユーロ」という単語に重要度Ｘ＝２の「＿２」のラベルが付加されて「ユーロ＿２」に変換されている。なお、図５の例では、重要度を示すＸを、０（重要ではない）＜１（通常）＜２（重要）のように３種類に区分して設定しているが、これに限られず、他の値や区分方法であってもよい。

【００３９】

各単語等の重要度は、文章中の単語の重要度を示す指標として一般的に用いられているＴＦ－ＩＤＦ（Term Frequency - Inverse Document Frequency）値を算出して用いることができる。例えば、算出したＴＦ－ＩＤＦ値を所定の範囲毎に区分して、上記の０～２の重要度を設定する。

【００４０】

学習モデル作成（図２のステップＳ０７）、および要約作成（図３のステップＳ１３）の処理では、それぞれ、単語等の重要度を示すラベルが付加された状態の文章に対して処理を行う、すなわち、各単語等の重要度を考慮した形で学習モデル１５の作成、および要約の作成を行う。このとき、要約の文章には、図５の上から３段目の文章のように、「＿０」や「＿２」等のラベルが付された語句が残存することになる。

【００４１】

本実施の形態では、これらの語句を要約作成時の後処理（図３のステップＳ１４）において元の単語等に戻す。すなわち、文章中の各単語等から、「＿０」や「＿２」等の重要度を示すラベル部分を全て削除する。これにより図５の最下段の文章のように元の単語等による要約テキスト４を作成することができる。

【００４２】

図６は、本実施の形態における前処理および後処理の他の例について概要を示した図である。ここでは、学習用テキスト２や要約対象テキスト３に各種の固有名詞を含む名詞が含まれている場合の例について示している。ニュース記事等の文章には、異なる単語であるが同一の品詞であるものが複数存在する場合がある。例えば、「野村さんは、アメリカよりフランスに住みたい。」という文章には、「野村」、「アメリカ」、「フランス」という固有名詞が含まれている。ここで、「野村」は人名であるが、「アメリカ」と「フランス」はともに国名である。これらの語句を全て異なる語句として取り扱うと、文章中に数値を含む場合と同様に、学習用テキスト２においてサンプルが発散して膨大な数となり、学習モデル１５の精度が低下するとともに学習処理の負荷も増大してしまう。

【００４３】

そこで、本実施の形態では、学習時および要約作成時の前処理（図２のステップＳ０２、図３のステップＳ１２）において、名詞（特に固有名詞）を品詞情報を示す語句に置換する。すなわち、上記の例では、例えば、「野村」を「＿固有名詞人名姓＊１」、「アメリカ」を「＿固有名詞地域国＊１」、「フランス」を「＿固有名詞地域国＊２」のようにそれぞれ置換する。これにより、例えば、「アメリカ」と「フランス」は異なる単語ではあるが、品詞としては同一のもの（「固有名詞地域国」）として取り扱うことができる。

【００４４】

なお、本実施の形態では、「アメリカ」や「フランス」等の「国」が異なる場合でも、「国」レベルの同じ固有名詞として取り扱うものとしているが、これに限られず、同じ取り扱いをする単位・レベルは適宜設定することができる。例えば、「地域」レベルや「都市」レベル等で同じ固有名詞として取り扱うようにしてもよいが、細分化が過剰となると

要約の精度が低下するため、細分化は適当なレベルに止めるのが望ましい。

【００４５】

図６の例では、例えば、最上段の文章（上述の数値置換が行われた結果の学習用テキスト２および要約対象テキスト３）には、「財務省」や「貿易統計」、「貿易黒字」等の固有名詞の語句が含まれている。この固有名詞部分をそれぞれ、図６の上から２段目の文章のように、「＿固有名詞組織＊１」や「＿固有名詞一般＊１」、「＿固有名詞一般＊２」等の品詞情報を示す語句に置換する。ここでは、「固有名詞組織」や「固有名詞一般」等の品詞情報に加えて、「＊１」や「＊２」等の出現順序・位置の情報についても含んでいる。

【００４６】

学習モデル作成（図２のステップＳ０７）、および要約作成（図３のステップＳ１３）の処理では、それぞれ、品詞情報を示す語句により置換された状態の文章に対して処理を行い、学習モデル１５の作成、および要約の作成を行う。このとき、要約の文章には、図６の上から３段目の文章のように、「＿固有名詞組織＊１」や「＿固有名詞一般＊１」等の語句が残存することになる。

【００４７】

本実施の形態では、これらの語句を要約作成時の後処理（図３のステップＳ１４）において元の単語等に戻す。そのために、要約作成時の前処理（図３のステップＳ１２）において固有名詞等を「＿固有名詞組織＊１」等の記号に置換した際に、置換した固有名詞の内容や文章内の出現位置等の復元のための情報を置換データ１７に記録しておき、これを後処理時に参照する。これにより図６の最下段の文章のように元の固有名詞等を含む要約テキスト４を作成することができる。

【００４８】

以上に説明したように、本発明の一実施の形態であるテキスト要約システム１によれば、機械学習を行う前の学習用テキスト２や要約対象テキスト３に対して、前処理の際に、数値を所定の記号に置換したり、各単語等に重要度を示すラベルを付加したり、固有名詞等を所定の語句に置換したり等の所定の加工を行い、加工された文章に対して学習モデル１５の作成や要約の作成の処理を行う。そして、作成された要約に含まれる加工内容（置換・付加された記号や語句）を元の単語等に戻すことで、最終的な要約テキスト４を得る。これにより、学習データ１２のサンプルとしての発散を回避し、単語の重要度を加味して学習モデル１５を生成することで要約作成における精度を向上させることができる。

【００４９】

本発明者らは、学習用テキスト２や要約対象テキスト３としてニュース記事を対象とし、要約テキスト４を実際に作成した上で、新聞記事等の自動要約に対する評価指標として広く用いられているＲｏｕｇｅ（Recall-Oriented Understudy for Gisting Evaluation）－１の値を算出して比較を行った。これによると、前処理において上記のような正規化を行わない場合のＲｏｕｇｅ－１の平均値が０．３～０．４程度であるのに対し、上記のような正規化を個別に、もしくは１つ以上組み合わせて適用することで、Ｒｏｕｇｅ－１の平均値を最大で０．６～０．７程度に向上させることが可能であるという結果が得られた。

【００５０】

以上、本発明者によってなされた発明を実施の形態に基づき具体的に説明したが、本発明は上記の実施の形態に限定されるものではなく、その要旨を逸脱しない範囲で種々変更可能であることはいうまでもない。例えば、上記の実施の形態は本発明を分かりやすく説明するために詳細に説明したものであり、必ずしも説明した全ての構成を備えるものに限定されるものではない。また、上記の実施の形態の構成の一部について、他の構成の追加・削除・置換をすることが可能である。

【００５１】

また、上記の各構成、機能、処理部、処理手段等は、それらの一部または全部を、例えば、集積回路で設計する等によりハードウェアで実現してもよい。また、上記の各構成、

機能等は、プロセッサがそれぞれの機能を実現するプログラムを解釈し、実行することによりソフトウェアで実現してもよい。各機能を実現するプログラム、テーブル、ファイル等の情報は、メモリやハードディスク、ＳＳＤ（Solid State Drive）等の記録装置、またはＩＣカード、ＳＤカード、ＤＶＤ等の記録媒体に置くことができる。

【００５２】

また、上記の各図において、制御線や情報線は説明上必要と考えられるものを示しており、必ずしも実装上の全ての制御線や情報線を示しているとは限らない。実際にはほとんど全ての構成が相互に接続されていると考えてもよい。

【産業上の利用可能性】

【００５３】

本発明は、機械学習により生成された学習モデルに基づいて要約を作成するテキスト要約システムに利用可能である。

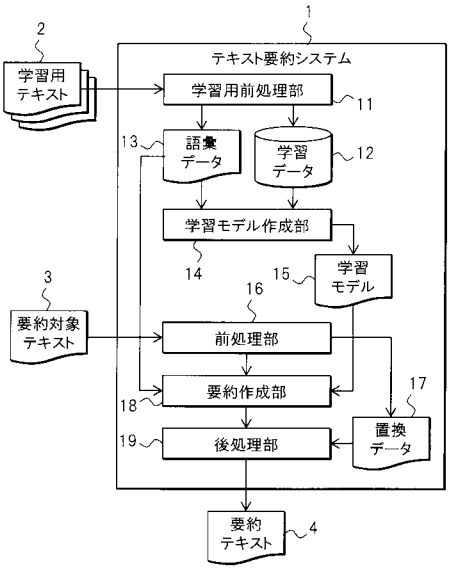
【符号の説明】

【００５４】

１…テキスト要約システム、２…学習用テキスト、３…要約対象テキスト、４…要約テキスト、
１１…学習用前処理部、１２…学習データ、１３…語彙データ、１４…学習モデル作成部、１５…学習モデル、１６…前処理部、１７…置換データ、１８…要約作成部、１９…後処理部

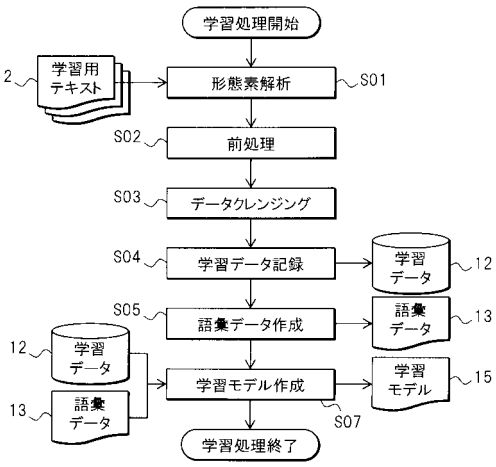
【図１】

図 1



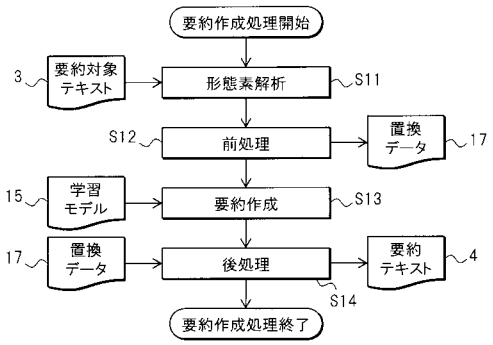
【図２】

図 2



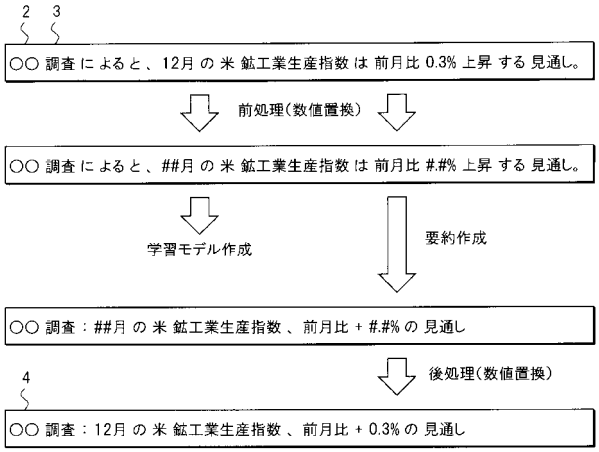
【図３】

図 3



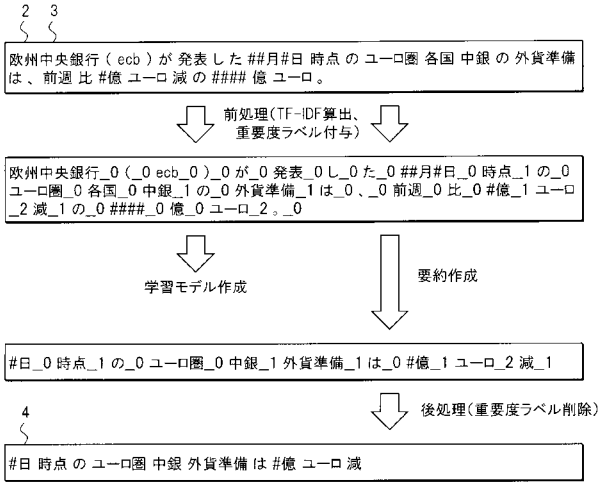
【図４】

図 4



【図５】

図 5



【図６】

図 6