(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2013-97723
(P2013-97723A)
(43)公開日　平成25年5月20日(2013.5.20)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	220A	5B109
G06F 17/21	(2006.01)	G06F	17/30	170A	
		G06F	17/21	550A	
		G06F	17/30	220Z	

審査請求 未請求　請求項の数 5　ＯＬ　（全 13 頁）

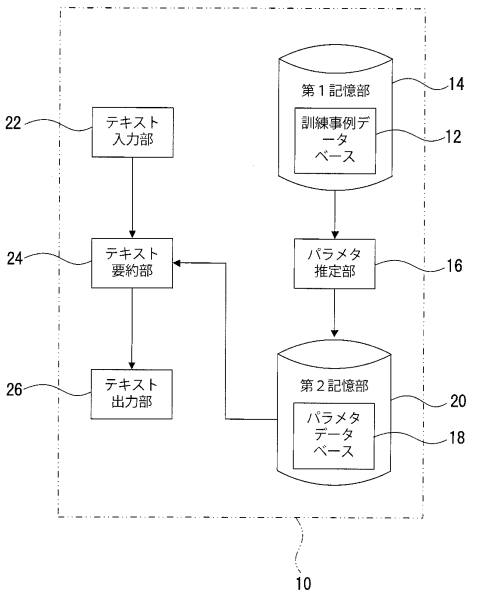
(21)出願番号　特願2011-242529 (P2011-242529)
(22)出願日　平成23年11月4日 (2011.11.4)

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(74)代理人　110001519
特許業務法人太陽国際特許事務所
(72)発明者　西川　仁
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　牧野　俊朗
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　松尾　義博
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

Ｆターム(参考)　5B109 QA05

(54)【発明の名称】テキスト要約装置、方法及びプログラム

(57)【要約】
【課題】ＲＯＵＧＥ等のような要約の品質を評価する評価尺度の値が直接最大化されるようにパラメタを学習する。
【解決手段】第１記憶部14に記憶された訓練事例データベース12には、テキストx_iとテキストx_iから予め生成された要約y_iとを対応付けた訓練事例が格納されており、パラメタ推定部16は、訓練事例のテキストx_iから要約y'を生成し、訓練事例の要約y_iに対する生成した要約y'の誤差を、要約の品質を評価する評価尺度を用いて求め、要約ｙの生成に用いるパラメタを、求めた誤差が大きくなる程更新の幅が大きくなるように更新し、パラメタデータベース18に格納する。パラメタデータベース18に格納されたパラメタは、テキスト要約部24による要約対象のテキストｘからの要約ｙの生成に用いられる。
【選択図】図１



【特許請求の範囲】
【請求項１】
テキストと、当該テキストから予め生成された要約と、を対応付けた訓練事例を格納した訓練事例データベースを記憶する第１記憶部と、
前記訓練事例に含まれ前記テキストと対応付けられた要約の品質に対する、要約対象のテキストからの要約の生成に用いられるパラメタの更新前の値を用いて、前記訓練事例に含まれる前記テキストから生成された要約の品質の誤差を、要約の品質を評価する評価尺度を用いて求め、前記パラメタの更新前の値を、求めた前記誤差が大きくなる程更新の幅が大きくなるように更新することで、前記パラメタを推定するパラメタ推定部と、
を含むテキスト要約装置。
【請求項２】
前記パラメタ推定部は、前記パラメタとしての重みベクトルの更新前の値をw_{old}、前記訓練事例に含まれる前記テキストをx_i、前記訓練事例に含まれる前記要約をy_i、前記パラメタとしての重みベクトルの更新前の値を用いて前記訓練事例に含まれる前記テキストx_iから生成された要約をy'、テキストｘ及び要約ｙの特徴ベクトルをｆ(ｘ,ｙ)、前記評価尺度をＲＯＵＧＥ、前記誤差を$loss(y';y_i)$としたときに、
【数１】
$$w_{new}=w_{old}+\lambda(f(x_i,y_i)-f(x_i,y'))\quad\cdots(1)$$
但し、
$$\lambda=\frac{loss(y';y_i)-w_{old}\cdot f(x_i,y_i)+w_{old}\cdot f(x_i,y')}{|f(x_i,y_i)-f(x_i,y')|^2}\quad\cdots(2)$$
であり、
$$loss(y';y_i)=1-ROUGE(y';y_i)\quad\cdots(3)$$
上記(１)～(３)式に従って前記パラメタとしての重みベクトルの更新後の値w_{new}を演算することで前記パラメタを推定する請求項１記載のテキスト要約装置。
【請求項３】
前記パラメタ推定部によって推定された前記パラメタを格納するパラメタ・データベースを記憶する第２記憶部と、
要約対象のテキストを受け付けるテキスト入力部と、
前記テキスト入力部によって受け付けされた前記要約対象のテキストと、前記パラメタ・データベースに格納された前記パラメタと、に基づいて、前記要約対象のテキストの要約を生成するテキスト要約部と、
前記テキスト要約部によって生成された要約をテキストとして出力するテキスト出力部と、
を更に備えた請求項１又は請求項２記載のテキスト要約装置。
【請求項４】
テキストと、当該テキストから予め生成された要約と、を対応付けた訓練事例を格納した訓練事例データベースが第１記憶部に記憶された状態で、
前記訓練事例に含まれ前記テキストと対応付けられた要約の品質に対する、要約対象のテキストからの要約の生成に用いられるパラメタの更新前の値を用いて、前記訓練事例に含まれる前記テキストから生成された要約の品質の誤差を、要約の品質を評価する評価尺度を用いて求め、前記パラメタの更新前の値を、求めた前記誤差が大きくなる程更新の幅が大きくなるように更新することで、前記パラメタを推定するパラメタ推定ステップ
を含むテキスト要約方法。
【請求項５】
コンピュータを、請求項１～請求項３の何れか１項記載のテキスト要約装置を構成する各手段として機能させるためのテキスト要約プログラム。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明はテキスト要約装置、テキスト要約方法及びテキスト要約プログラムに関する。
【背景技術】
【０００２】
近年、電子化されたテキストが大量に流通するようになってきたことを背景として、それらのテキストに記述されている情報を迅速に把握するために、コンピュータ等の機械によってテキストを要約させる(要約を生成させる)技術に対するニーズが高まっている。
【０００３】
機械によってテキストを要約させる技術では、一般に、要約対象のテキストから、その内容を代表していると思われる文(重要文)、或いは、そのような文の集合(重要文集合)を１つ以上選び出し、それらを連結することで要約が生成される。文、或いは文の集合に対して要約対象のテキストの内容を代表しているか否かを評価する際には、各々の文、或いは文の集合を特徴ベクトルとして表現し、この特徴ベクトルと、予め何らかの方法で推定した重みベクトル(以下、パラメタともいう)と、の内積がスコアとして算出される(例えば非特許文献１も参照)。
【０００４】
また、文、或いは文の集合を表現した特徴ベクトルとの内積を求めるパラメタは、要約対象のテキスト、或いは要約対象のテキストの集合と、それらに対応する要約と、から成るペアの集合(以下、訓練事例という)に基づいて予め学習される。パラメタの学習に際し、パラメタは何らかの誤差関数を最小化するように学習される。
【先行技術文献】
【非特許文献】
【０００５】
【非特許文献１】Wen-tau Yih,Joshua Goodman,Lucy Vanderwende and Hisami Suzuki、“Multi-Document Summarization by Maximizing Informative Content-Words.”、In Proceedings of International Joint Conference on Artificial Intelligence(IJCAI),2007.
【非特許文献２】Koby Crammer,Ofel Dekel,Joseph Keshet,Shai Shalev-Shwartz and Yoram Singer、“Online Passive-Aggressive Algorithms.”、Journal of Machine Learning Research,Vol.7,2006.
【発明の概要】
【発明が解決しようとする課題】
【０００６】
しかしながら、テキストを要約する技術で用いるパラメタの学習に際し、例えば非特許文献１に記載の技術では、要約の品質が直接最大化されるようなパラメタの学習を行っておらず、要約に含まれ易い文が高いスコアになるようにパラメタの学習を行っている。機械によって生成された要約の品質を評価するための尺度としては、例えばＲＯＵＧＥ(非特許文献２を参照)と称される評価尺度が多用されているが、要約に含まれ易い文が高いスコアになるようにパラメタを学習することと、学習したパラメタを用いて生成した要約のＲＯＵＧＥ等の評価尺度が高い値になることには直接関係がなく、パラメタの学習精度に関して改善の余地があった。
【０００７】
本発明は上記事実を考慮して成されたもので、ＲＯＵＧＥ等のような要約の品質を評価する評価尺度の値が直接最大化されるようにパラメタを学習できるテキスト要約装置、テキスト要約方法及びテキスト要約プログラムを得ることが目的である。
【課題を解決するための手段】
【０００８】
本発明に係るテキスト要約装置は、テキストと、当該テキストから予め生成された要約

と、を対応付けた訓練事例を格納した訓練事例データベースを記憶する第１記憶部と、前記訓練事例に含まれ前記テキストと対応付けられた要約の品質に対する、要約対象のテキストからの要約の生成に用いられるパラメタの更新前の値を用いて、前記訓練事例に含まれる前記テキストから生成された要約の品質の誤差を、要約の品質を評価する評価尺度を用いて求め、前記パラメタの更新前の値を、求めた前記誤差が大きくなる程更新の幅が大きくなるように更新することで、前記パラメタを推定するパラメタ推定部と、を含んで構成されている。

【０００９】

また、本発明において、前記パラメタ推定部は、前記パラメタとしての重みベクトルの更新前の値をw_{old}、前記訓練事例に含まれる前記テキストをx_i、前記訓練事例に含まれる前記要約をy_i、前記パラメタとしての重みベクトルの更新前の値を用いて前記訓練事例に含まれる前記テキストx_iから生成された要約をy'、テキストｘ及び要約ｙの特徴ベクトルを$f(x,y)$、前記評価尺度をＲＯＵＧＥ、前記誤差を$loss(y';y_i)$としたときに、

【００１０】

【数１】

$$w_{new}=w_{old}+\lambda(f(x_i,y_i)-f(x_i,y')) \quad \cdots(1)$$

但し、

$$\lambda=\frac{loss(y';y_i)-w_{old}\cdot f(x_i,y_i)+w_{old}\cdot f(x_i,y')}{|f(x_i,y_i)-f(x_i,y')|^2} \quad \cdots(2)$$

【００１１】

であり、

$$loss(y';y_i)=1-ROUGE(y';y_i) \quad \cdots(3)$$

【００１２】

上記(１)～(３)式に従って前記パラメタとしての重みベクトルの更新後の値w_{new}を演算することで前記パラメタを推定することができる。

【００１３】

また、本発明は、前記パラメタ推定部によって推定された前記パラメタを格納するパラメタ・データベースを記憶する第２記憶部と、要約対象のテキストを受け付けるテキスト入力部と、前記テキスト入力部によって受け付けされた前記要約対象のテキストと、前記パラメタ・データベースに格納された前記パラメタと、に基づいて、前記要約対象のテキストの要約を生成するテキスト要約部と、前記テキスト要約部によって生成された要約をテキストとして出力するテキスト出力部と、を更に備えることができる。

【００１４】

本発明に係るテキスト要約方法は、テキストと、当該テキストから予め生成された要約と、を対応付けた訓練事例を格納した訓練事例データベースが第１記憶部に記憶された状態で、前記訓練事例に含まれ前記テキストと対応付けられた要約の品質に対する、要約対象のテキストからの要約の生成に用いられるパラメタの更新前の値を用いて、前記訓練事例に含まれる前記テキストから生成された要約の品質の誤差を、要約の品質を評価する評価尺度を用いて求め、前記パラメタの更新前の値を、求めた前記誤差が大きくなる程更新の幅が大きくなるように更新することで、前記パラメタを推定するパラメタ推定ステップを含んでいる。

【００１５】

本発明に係るテキスト要約プログラムは、コンピュータを、請求項１～請求項３の何れか１項記載のテキスト要約装置を構成する各手段として機能させる。

【発明の効果】

【００１６】

本発明は、ＲＯＵＧＥ等のような要約の品質を評価する評価尺度の値が直接最大化され

るようにパラメタを学習できる、という効果を有する。

【図面の簡単な説明】

【００１７】

【図１】実施形態で説明したテキスト要約装置を示すブロック図である。

【図２】テキスト要約装置として機能するコンピュータの概略ブロック図である。

【図３】テキスト要約処理の概略を示すフローチャートである。

【図４】訓練事例データベースに格納される訓練事例の一例を示す図表である。

【図５】(Ａ)はテキスト及び要約の一例を示す図表、(Ｂ)は(Ａ)に示すテキスト及び要約に対応する特徴ベクトルの一例を示す説明図である。

【図６】パラメタ学習処理のアルゴリズムの一例を示すフローチャートである。

【図７】パラメタ・データベースの一例を示す図表である。

【図８】要約対象のテキストの入力形式の一例を示す図表である。

【図９】要約生成処理のアルゴリズムの一例を示すフローチャートである。

【発明を実施するための形態】

【００１８】

以下、図面を参照して本発明の実施形態の一例を詳細に説明する。図１には本実施形態に係るテキスト要約装置１０が示されている。テキスト要約装置１０は、パラメタ(本実施形態では重みベクトルw)を推定するために必要となる訓練事例を格納した訓練事例データベース１２を記憶する第１記憶部１４と、訓練事例データベース１２に格納された訓練事例を受け付けてパラメタを推定するパラメタ推定部１６と、パラメタ推定部１６で推定されたパラメタを格納したパラメタ・データベース１８を記憶する第２記憶部２０と、要約の対象とするテキストを受け付けるテキスト入力部２２と、テキスト入力部２２で受け付けられた要約対象のテキストとパラメタ・データベース１８に格納されたパラメタを入力として要約を生成するテキスト要約部２４と、テキスト要約部２４によって生成された要約を出力するテキスト出力部２６と、を備えている。

【００１９】

テキスト要約装置１０は、例えば図２に示すコンピュータ３０で実現することができる。コンピュータ３０はＣＰＵ３２、メモリ３４、不揮発性の記憶部３６、キーボード３８、マウス４０、ディスプレイ４２を備え、これらはバス４４を介して互いに接続されている。なお、記憶部３６はＨＤＤ(Hard Disk Drive)やフラッシュメモリ等によって実現できる。記録媒体としての記憶部３６には、コンピュータ３０をテキスト要約装置１０として機能させるためのテキスト要約プログラム４６、訓練事例データベース１２、パラメタ・データベース１８が各々記憶されている。ＣＰＵ３２は、テキスト要約プログラム４６を記憶部３６から読み出してメモリ３４に展開し、テキスト要約プログラム４６が有するプロセスを順次実行する。

【００２０】

テキスト要約プログラム４６は、テキスト入力プロセス５０、パラメタ推定プロセス５２、テキスト要約プロセス５４及びテキスト出力プロセス５６を有する。ＣＰＵ３２は、テキスト入力プロセス５０を実行することで、図１に示すテキスト入力部２２として動作する。またＣＰＵ３２は、パラメタ推定プロセス５２を実行することで、図１に示すパラメタ推定部１６として動作する。またＣＰＵ３２は、テキスト出力プロセス５６を実行することで、図１に示すテキスト出力部２６として動作する。なお、テキスト要約プログラム４６は本発明に係るテキスト要約プログラムの一例である。

【００２１】

テキスト要約装置１０がコンピュータ３０で実現される場合、訓練事例データベース１２を記憶する記憶部３６は第１記憶部１４として用いられ、メモリ３４の一部領域は訓練事例データベース１２として用いられる。また、パラメタ・データベース１８を記憶する記憶部３６は第２記憶部２０として用いられ、メモリ３４の一部領域はパラメタ・データベース１８として用いられる。これにより、テキスト要約プログラム４６を実行したコンピュータ３０が、テキスト要約装置１０として機能することになる。

【００２２】
次に本実施形態の作用として、まず図３を参照し、コンピュータ３０のＣＰＵ３２でテキスト要約プログラム４６が実行されることで実現されるテキスト要約処理の概略を説明する。テキスト要約処理のステップ６０において、パラメタ推定部１６は、訓練事例データベース１２から訓練事例を受け付け、受け付けた訓練事例に基づいてパラメタの学習を行う(パラメタ学習の詳細は後述)。また、次のステップ６２において、パラメタ推定部１６は、学習したパラメタをパラメタ・データベース１８に格納する。
【００２３】
次のステップ６４において、テキスト入力部２２は、入力された要約対象のテキストを受け付ける。なお、要約対象のテキストは、例えばキーボード３８を介して入力されたテキストでもよいし、例えば記憶部３６に予め記憶されたテキストでもよいし、例えば通信回線を介してコンピュータ３０と接続された他の機器から受信したテキストでもよい。
【００２４】
また、ステップ６６において、テキスト要約部２４は、テキスト入力部２２によって受け付けされた要約対象のテキストと、パラメタ・データベース１８に格納されたパラメタと、から要約を生成する(要約生成の詳細は後述)。
【００２５】
そしてステップ６８において、テキスト出力部２６は、テキスト要約部２４によって生成された要約を外部へ出力する。要約の出力は、例えばディスプレイ４２にテキストとして表示してもよいし、要約を読み上げる音声として出力してもよいし、テキストデータとして記録メディアに記録するか他の機器へ送信するようにしてもよい。
【００２６】
また、図３では、パラメタの学習及びパラメタ・データベース１８への格納(ステップ６０,６２)と、要約の生成・出力(ステップ６４～６８)と、を一連の処理として示しているが、パラメタの学習及びパラメタ・データベース１８への格納を行った後、時間を空けて要約の生成・出力を行ってもよい。
【００２７】
次に、テキスト要約処理の詳細を説明するにあたり、まず、訓練事例データベース１２に格納される訓練事例について、図４を参照して説明する。図４には、訓練事例データベース１２に格納される訓練事例の一例が示されている。図４において、訓練事例データベース１２は、各レコードに１つの訓練事例を各々格納している。訓練事例データベース１２の各レコードに格納されているそれぞれの訓練事例は、個々の訓練事例を識別するための事例番号、要約対象のテキスト、及び、当該要約対象のテキストから生成された要約の各情報を含んでいる。なお、訓練事例に含まれる要約は、例えば人手によって作成され、後述するパラメタ(重みベクトルｗ)の学習における基準として用いることが可能な正しい要約である。
【００２８】
続いて、パラメタ(重みベクトルｗ)を用いた要約の生成について説明する。要約対象のテキストをｘ、要約対象のテキストｘから生成された要約をｙとする。要約対象のテキストｘが与えられると、要約対象のテキストｘと、要約対象のテキストｘから生成された要約ｙと、から特徴ベクトルｆ(ｘ,ｙ)が得られる。特徴ベクトルｆ(ｘ,ｙ)と、パラメタ(重みベクトルｗ)の内積ｗ・ｆ(ｘ,ｙ)が要約ｙのスコアである。要約を生成する際には、要約対象のテキストｘから生成される要約ｙのうち、パラメタ(重みベクトルｗ)と特徴ベクトルｆ(ｘ,ｙ)の内積ｗ・ｆ(ｘ,ｙ)が最大となる要約ｙを要約ｙ’として生成する。上記を式で表すと次の(４)式のようになる。
【００２９】
$$y' = \arg_y \max\ w \cdot f(x,y) \quad \cdots(4)$$
要約の生成方法の一例については後述するが、任意の要約生成方法を適用することができる。
【００３０】

要約対象のテキストｘと当該テキストｘから生成された要約ｙから得られる特徴ベクトルｆ(ｘ,ｙ)は、各要素が互いに異なる単語に対応し、一例として、要約ｙが或る単語を含んでいれば対応する要素の値が１、含んでいなければ対応する要素の値が０となるベクトルを適用することができる。図５(Ａ)に示すテキストｘ及び要約ｙから得られる特徴ベクトルｆ(ｘ,ｙ)の一例を図５(Ｂ)に示す。図５(Ａ)に示す要約ｙは「路上」や「強盗」といった単語を含むため、図５(Ｂ)に示す特徴ベクトルｆ(ｘ,ｙ)では、これらの単語に対応する要素の値が１となっている。一方、図５(Ａ)に示す要約ｙは「道路」や「泥棒」といった単語は含んでいないため、図５(Ｂ)に示す特徴ベクトルｆ(ｘ,ｙ)では、これらの単語に対応する要素の値が０となっている。

【００３１】

次に、パラメタ推定部１６によって行われるパラメタの学習を説明する。ｎ個の訓練事例が与えられたと仮定し、そのうちのｉ番目の訓練事例のテキストをx_i、要約をy_iとする。例えば事例番号１のテキストをｘ1、要約をｙ1とする。

【００３２】

パラメタの学習に際しては、ｎ個の訓練事例が１つずつ取り上げられる。その時点でのパラメタｗの下で、テキストx_iから生成された要約ｙ’と、人手によって作成された正しい要約y_iが得られた際に、要約ｙ’の要約としての良さ(品質)は、要約ｙ’を評価するための評価尺度、一例としてＲＯＵＧＥを用い、要約y_iを基準とした要約ｙ’のＲＯＵＧＥの値(ＲＯＵＧＥ(ｙ';y_i))として算出できる。ＲＯＵＧＥの値は０～１の範囲であるので、前出の(３)式のように１からＲＯＵＧＥの値を減算することで、要約y_iを基準(正しい要約)としたときの要約ｙ’の悪さ、すなわち要約y_iに対する要約ｙ’の誤差が得られる。なお、前出の(３)式では誤差として誤差関数loss(ｙ';y_i)を用いている。

【００３３】

上記の誤差を減らすようにパラメタを更新できれば、より良い要約を生成できるパラメタを得ることができる。そこで、以下の(５)式のようにパラメタを更新する。

【００３４】

【数２】

$$w_{new}=\arg_w \min \frac{1}{2}|w-w_{old}|^2 \qquad \cdots(5)$$

但し、$w_{new}\cdot f(x_i,y_i)-w_{new}\cdot f(x_i,y')\geqq \mathrm{loss}(y';y_i)$

【００３５】

(５)式において、w_{old}は現在(更新前)のパラメタであり、w_{new}は更新後のパラメタである。上記の(５)式は、パラメタｗの変化を最小にしつつ、更新後のパラメタw_{new}において、正しい要約y_iのスコアが現在(更新前)のパラメタw_{old}を用いて生成された要約ｙ’のスコアより必ず大きくなるように、パラメタｗを更新している。その際、要約y_iのスコアから要約ｙ’のスコアを減算した残差が誤差関数loss(ｙ';y_i)以上となるようにしている。上記の(５)式を解くと、前出のパラメタ更新式((１),(２)式)が得られる。

【００３６】

前出の更新式((１),(２)式)では、現在(更新前)のパラメタw_{new}を、誤差関数にＲＯＵＧＥを用いて求めた要約y_iに対する要約ｙ’の誤差が大きくなる程更新の幅が大きくなるように更新しているので、パラメタｗの更新に伴ってＲＯＵＧＥの値が直接高くなるようにパラメタｗを更新(学習)することができる。

【００３７】

次に図６を参照し、パラメタ推定部１６によるパラメタ学習処理(図３に示すテキスト要約処理のステップ６０,６２の詳細)の具体的なアルゴリズムの一例を説明する。パラメタ学習処理のステップ７０において、パラメタ推定部１６は、事例番号ｉが付与された訓練事例に含まれるテキストx_i及び要約y_iを訓練事例データベース１２からｎ個(事例番号ｉ＝１～ｎの訓練事例を)取得する。またステップ７２において、パラメタ推定部１

６は、予め設定された繰り返し回数Ｔを取得する。次のステップ７４において、パラメタ推定部１６は、パラメタ(重みベクトルｗ)を、全ての要素の値が０のベクトルへ初期化する。またステップ７６において、パラメタ推定部１６は変数ｔに１を設定し、次のステップ７８において、パラメタ推定部１６は変数ｉに１を設定する。

【００３８】

次のステップ８０において、パラメタ推定部１６は、事例番号ｉの訓練事例について、現在のパラメタ(重みベクトルｗ)と特徴ベクトルｆ(ｘ$_{i}$,ｙ)との内積ｗ・ｆ(ｘ$_{i}$,ｙ)が最大となる要約ｙを、要約ｙ'として生成する(次の(６)式も参照)。

【００３９】

$$y' = \arg_y \max\ w \cdot f(x_i, y) \quad \cdots(6)$$

【００４０】

次のステップ８２において、パラメタ推定部１６は、まず前出の(３)式により要約ｙ$_{i}$に対する要約ｙ’の誤差loss(ｙ';ｙ$_{i}$)を演算した後に、前出のパラメタ更新式((１),(２)式)によりパラメタ(重みベクトルｗ)を更新する。これにより、要約ｙ$_{i}$に対する要約ｙ’の誤差が大きくなる程更新の幅が大きくなるようにパラメタ(重みベクトルｗ)が更新される。

【００４１】

ステップ８４において、パラメタ推定部１６は、変数ｉが先のステップ７０で取得した訓練事例の総数ｎ以上になったか否か判定する。ステップ８４の判定が否定された場合はステップ８６へ移行し、ステップ８６において、パラメタ推定部１６は変数ｉを１だけインクリメントした後にステップ８０に戻る。これにより、ステップ８４の判定が肯定される迄ステップ８０～ステップ８６が繰り返され、ｎ個の訓練事例を１つずつ用いてパラメタ(重みベクトルｗ)の更新(学習)がｎ回繰り返される。

【００４２】

ステップ８４の判定が肯定されるとステップ８８へ移行する。ステップ８８において、パラメタ推定部１６は、変数ｔが先のステップ７２で取得した繰り返し回数Ｔ以上になったか否か判定する。ステップ８８の判定が否定された場合はステップ９０へ移行し、ステップ９０において、パラメタ推定部１６は変数ｔを１だけインクリメントした後にステップ７８に戻る。これにより、ステップ８８の判定が肯定される迄ステップ７８～ステップ９０が繰り返され、ｎ個の訓練事例が各々Ｔ回ずつ用いられてパラメタ(重みベクトルｗ)の更新(学習)が繰り返される。これにより、要約の品質を評価する評価尺度であるＲＯＵＧＥの値が直接最大化されるようにパラメタ(重みベクトルｗ)が更新(学習)される。

【００４３】

ステップ８８の判定が肯定されると、パラメタ(重みベクトルｗ)の更新を終了してステップ９２へ移行し、ステップ９２において、パラメタ推定部１６は、上記処理で更新(学習)されたパラメタ(重みベクトルｗ)をパラメタ・データベース１８に格納し、パラメタ学習処理を終了する。パラメタ・データベース１８の一例を図７に示す。パラメタ・データベース１８は、各レコードに、或る単語を含んでいるか否か等の「特徴」と、当該特徴の「重み」の各情報が設定されている。図７に示すパラメタ・データベース１８における最初のレコードは、単語「道路」を含んでいるという特徴の重みが-0.03であるということを示している。

【００４４】

上述したパラメタ学習処理が行われた後、図３に示すテキスト要約処理のステップ６４において、テキスト入力部２２は、入力された要約対象のテキストｘを受け付ける。テキスト入力部２２によって取得されるテキストｘの入力形式の一例を図８に示す。図８に示すテキストｘの入力形式では、テキストｘが一文毎に分割され、更に形態素解析が行われた結果が入力される。図８において、表の各行は各々単一の形態素に対応しており、個々の形態素毎に「表記」「品詞」「読み」及び「標準形」の各情報が付加されている。

【００４５】

続いて図９を参照し、テキスト要約部２４による要約生成処理(図３に示すテキスト要

約処理のステップ６６の詳細)の具体的なアルゴリズムの一例を説明する。要約生成処理のステップ１００において、テキスト要約部２４は、要約対象のテキストｘ及び要約長Ｋをテキスト入力部２２から取得する。またステップ１０２において、テキスト要約部２４は、パラメタ・データベース１８からパラメタ(重みベクトル)ｗを取得する。また、ステップ１０４において、テキスト要約部２４は、ステップ１００で取得したテキストｘを文の集合Ｕに格納し、次のステップ１０６において、テキスト要約部２４は、要約を表す集合Ｓを空集合φに初期化する。

【００４６】

ステップ１０８において、テキスト要約部２４は、文の集合Ｕが空集合φよりも大きいか(文の集合Ｕに何らかの文が格納されているか)否かを判定する。ステップ１０８の判定が肯定された場合はステップ１１０へ移行する。ステップ１１０において、テキスト要約部２４は、文の集合Ｕに格納されている文の中から、要約の集合Ｓに文s_iを入れた時のスコア(＝特徴ベクトルｆとパラメタｗの内積)と要約の集合Ｓに文s_iを入れていない時のスコアの差を、文s_iの長さで除した値が最大の文s_i(次の(７)式を満たす文s_i)を選択する。

【００４７】

$$s_i = \arg\max_{s_i \in U}((w \cdot f(x,\{S,s_i\}) - w \cdot f(x,S))/\mathrm{length}(s_i)) \quad \cdots(7)$$

なお、(７)式において、length(s_i)は文s_iの長さである。

【００４８】

次のステップ１１２において、テキスト要約部２４は、要約の集合Ｓに既に入っている文の長さに、ステップ１１０で選択した文s_iの長さを加えた長さが要約長Ｋ以下か否か判定する(次の(８)式も参照)。

$$\mathrm{length}(\{S,s_i\}) \leqq K \quad \cdots(8)$$

【００４９】

ステップ１１２の判定が肯定された場合はステップ１１４へ移行する。ステップ１１４において、テキスト要約部２４は、ステップ１１０で選択した文s_iを要約の集合Ｓに加えた後に(次の(９)式も参照)、ステップ１１６へ移行する。

$$S = \{S,s_i\} \quad \cdots(9)$$

【００５０】

また、ステップ１１２の判定が否定された場合は、ステップ１１４をスキップして(文s_iを要約の集合Ｓに加えることなく)ステップ１１６へ移行する。

【００５１】

ステップ１１６において、テキスト要約部２４は、ステップ１１０で選択した文s_iを文の集合Ｕから除去する(Ｕ＝Ｕ－s_i)。ステップ１１６の処理を行うとステップ１０８に戻り、ステップ１０８の判定が肯定される迄ステップ１０８～ステップ１１６を繰り返す。文の集合Ｕが空集合φになると、ステップ１０８の判定が否定されてステップ１１８へ移行する。

【００５２】

ステップ１１８において、テキスト要約部２４は、テキストｘの文の中で、長さが要約長Ｋ以内で、単一の文として最もスコアの高い文s_iを文ｖとして選択する(次の(10)式も参照)。

$$v = \arg\max_{s_i \in x:\mathrm{length}(s_i) \leqq K} w \cdot f(x,s_i) \quad \cdots(10)$$

【００５３】

次のステップ１２０において、テキスト要約部２４は、要約の集合Ｓに入っている文のスコアが、先のステップ１１８で選択した文ｖのスコア以上か否か判定する。ステップ１２０の判定が肯定された場合はステップ１２２へ移行する。ステップ１２２において、テキスト要約部２４は、要約の集合Ｓに入っている文をテキストｘの要約として出力し、要約生成処理を終了する。また、ステップ１２０の判定が否定された場合はステップ１２４へ移行する。ステップ１２４において、テキスト要約部２４は、先のステップ１１８で選

択した文ｖをテキストｘの要約として出力し、要約生成処理を終了する。

【００５４】

上述した要約生成処理のステップ１２２又はステップ１２４において、テキスト要約部２４によって出力されたテキストｘの要約は、図３に示すテキスト要約処理のステップ６８において、テキスト出力部２６により外部へ出力される。

【００５５】

なお、図９に要約生成処理として示した要約の生成方法は一例であり、本発明におけるパラメタの更新方法は、任意の要約生成方法と組み合わせることが可能である。

【００５６】

また、上記では、要約の品質を評価する評価尺度としてＲＯＵＧＥを用いた態様を説明したが、本発明はこれに限定されるものではなく、ＲＯＵＧＥ以外の評価尺度を適用することも可能である。

【００５７】

また、上記では訓練事例データベース１２が、コンピュータ３０に設けられた記憶部３６に記憶されている態様を説明したが、本発明はこれに限定されるものではなく、通信回線を介してコンピュータ３０と接続された別のコンピュータに設けられた記憶部に記憶されていてもよい。この場合、パラメタの学習にあたり、テキストx_i及び要約y_iを各々含むｎ個の訓練事例は、通信回線を介して前記別のコンピュータから受信するように構成することができる。また、要約対象のテキストｘについても、通信回線を介して前記別のコンピュータから受信する構成であってもよい。

【００５８】

更に、上記では、コンピュータ３０がテキスト要約プログラム４６を実行することで、コンピュータ３０がテキスト要約装置１０として機能する態様を説明したが、本発明はこれに限定されるものではなく、図１に示した各機能ブロック(パラメタ推定部１６、テキスト入力部２２、テキスト要約部２４及びテキスト出力部２６)は、それぞれハードウェアで構成することも可能である。

【００５９】

また、上記ではテキスト要約プログラム４６が記憶部３６に予め記憶（インストール）されている態様を説明したが、本発明に係るテキスト要約プログラムは、ＣＤ－ＲＯＭやＤＶＤ－ＲＯＭ等の記録媒体に記録されている形態で提供することも可能である。

【００６０】

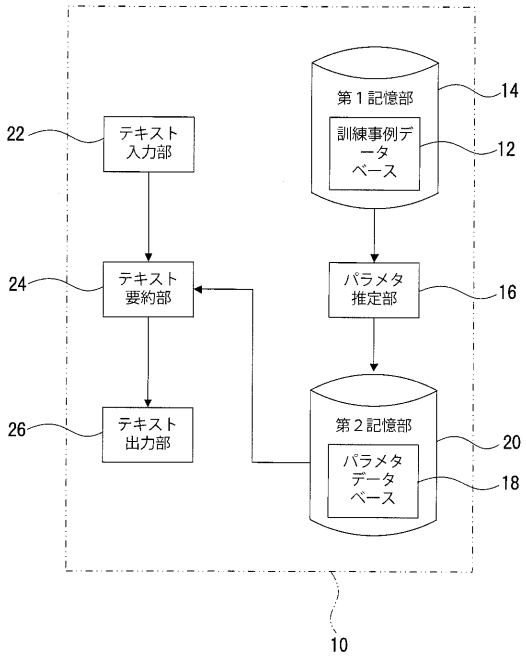
本明細書に記載された全ての文献、特許出願及び技術規格は、個々の文献、特許出願及び技術規格が参照により取り込まれることが具体的かつ個々に記された場合と同程度に、本明細書中に参照により取り込まれる。

【符号の説明】

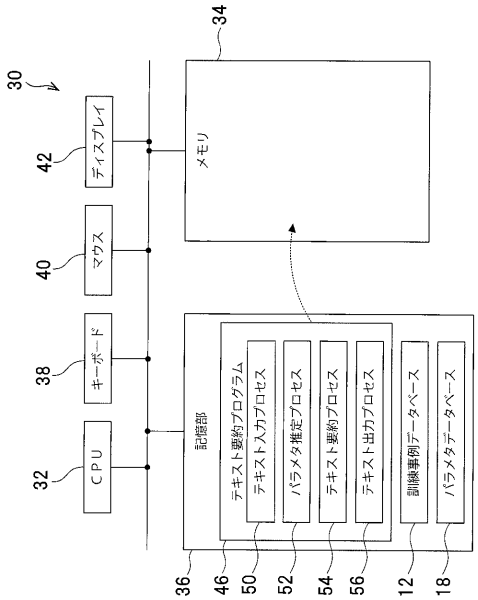
【００６１】

１０　テキスト要約装置
１２　訓練事例データベース
１４　第１記憶部
１６　パラメタ推定部
１８　パラメタ・データベース
２０　第２記憶部
２２　テキスト入力部
２４　テキスト要約部
２６　テキスト出力部
３０　コンピュータ
３４　メモリ
３６　記憶部
４６　テキスト要約プログラム

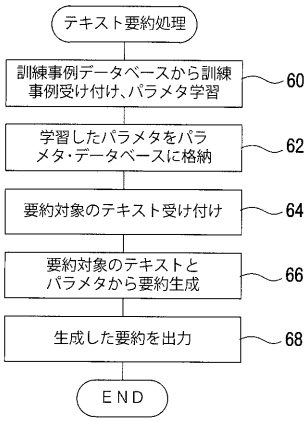
【図１】



【図２】



【図３】



【図４】

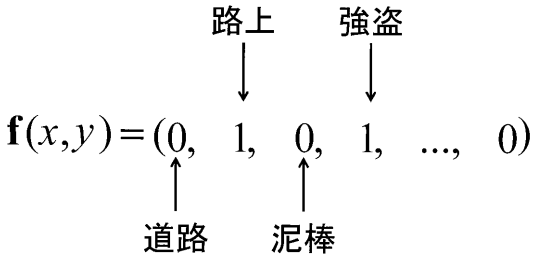
事例番号	テキスト	要約
1	9月26日午前、神奈川県横須賀市の路上で強盗事件が発生した。被害者は市内在住の70歳の男性。犯人は身長170センチメートル程度の中肉中背の男で、サングラスにマスク姿だったと言う。被害者の男性によれば男は原付バイクに乗って被害者に近づいたという。	9月26日午前、神奈川県横須賀市の路上で強盗事件が発生した。
2	9月27日未明、神奈川県横須賀市光の丘で火災が発生した。火元の民家は全焼した。火災当時居住者は外出しており建物内におらず、怪我人はいなかった。警察、消防は放火の可能性が高いとして調査を行っている。	9月27日未明、神奈川県横須賀市光の丘で火災が発生した。
…	…	…

【図５】

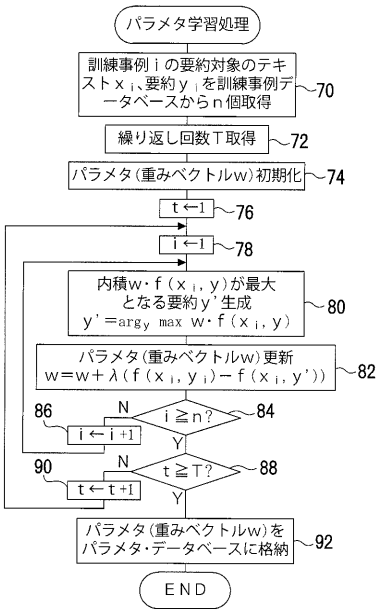
(A)

テキストx	要約y
9月26日午前、神奈川県横須賀市の路上で強盗事件が発生した。被害者は市内在住の70歳の男性。犯人は身長170センチメートル程度の中肉中背の男で、サングラスにマスク姿だったと言う。被害者の男性によれば男は原付バイクに乗って被害者に近づいたという。	9月26日に神奈川県横須賀市の路上で強盗事件が発生した。

(B)



【図６】



【図７】

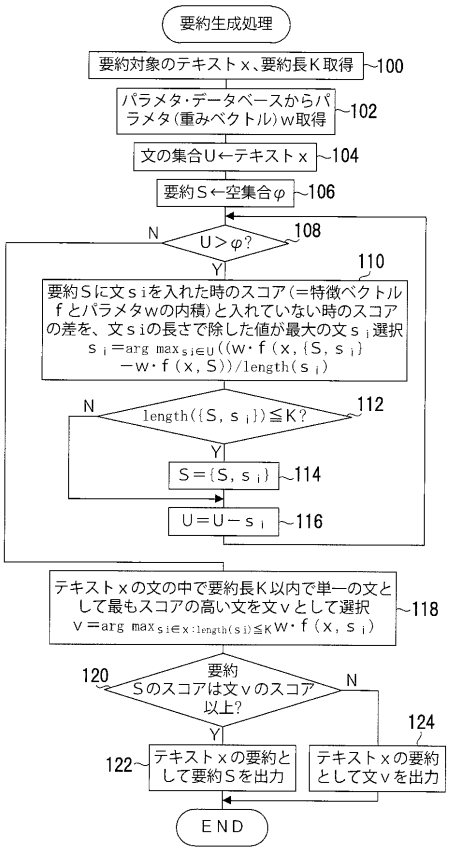
W

特徴	重み
道路	-0.03
路上	0.51
泥棒	-0.49
強盗	0.38
…	…

【図８】

表記	品詞	読み	標準形
9月	名詞:日時:連用	クガツ	9月
26日	名詞:日時:連用	ニジュウロクニチ	26日
午前	名詞:日時:連用	ゴゼン	午前
、	読点		、
神奈川	名詞:固有:地	カナガワ	神奈川
県	名詞接尾辞:名詞	ケン	県
横須賀	名詞:固有:地	ヨコスカ	横須賀
市	名詞接尾辞:名詞	シ	市
の	格助詞:連体	ノ	の
路上	名詞	ロジョウ	路上
で	格助詞:連用	デ	で
強盗	名詞	ゴウトウ	強盗
事件	名詞	ジケン	事件
が	格助詞:連用	ガ	が
発生	名詞:動作	ハッセイ	発生
し	動詞活用語尾	シ	し
…	…	…	…

【図９】


要約生成処理
要約対象のテキストｘ、要約長Ｋ取得
100
パラメタ・データベースからパラメタ（重みベクトル）ｗ取得
102
文の集合Ｕ←テキストｘ
104
要約Ｓ←空集合φ
106
Ｎ
Ｕ＞φ?
108
Ｙ
110
要約Ｓに文ｓｉを入れた時のスコア（＝特徴ベクトルｆとパラメタｗの内積）と入れていない時のスコアの差を、文ｓｉの長さで除した値が最大の文ｓｉ選択
ｓｉ＝arg max ｓｉ∈Ｕ((ｗ・ｆ(ｘ,{Ｓ,ｓｉ})−ｗ・ｆ(ｘ,Ｓ))/length(ｓｉ)
Ｎ
length({Ｓ,ｓｉ})≦Ｋ?
112
Ｙ
Ｓ＝{Ｓ,ｓｉ)
114
Ｕ＝Ｕ−ｓｉ
116
テキストｘの文の中で要約長Ｋ以内で単一の文として最もスコアの高い文を文ｖとして選択
ｖ＝arg max ｓｉ∈ｘ:length(ｓｉ)≦Ｋｗ・ｆ(ｘ,ｓｉ)
118
120
要約Ｓのスコアは文ｖのスコア以上?
Ｎ
Ｙ
124
122
テキストｘの要約として要約Ｓを出力
テキストｘの要約として文ｖを出力
ＥＮＤ