

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5747066号
(P5747066)

(45) 発行日 平成27年7月8日(2015.7.8)

(24) 登録日 平成27年5月15日(2015.5.15)

(51) Int.Cl. F I
G06F 17/21 (2006.01) G O 6 F 17/21 6 8 0
G06F 17/27 (2006.01) G O 6 F 17/27

請求項の数 20 (全 21 頁)

(21) 出願番号	特願2013-231893 (P2013-231893)	(73) 特許権者	309027126
(22) 出願日	平成25年11月8日(2013.11.8)		ニュアンス コミュニケーションズ, イン
(62) 分割の表示	特願2011-170125 (P2011-170125) の分割		コーポレイテッド
原出願日	平成16年11月12日(2004.11.12)		アメリカ合衆国 マサチューセッツ州 O
(65) 公開番号	特開2014-59896 (P2014-59896A)		1 8 0 3 パーリントン ワン・ウェイサ
(43) 公開日	平成26年4月3日(2014.4.3)	(74) 代理人	100107766
審査請求日	平成25年11月8日(2013.11.8)		弁理士 伊東 忠重
(31) 優先権主張番号	03104316.9	(74) 代理人	100070150
(32) 優先日	平成15年11月21日(2003.11.21)		弁理士 伊東 忠彦
(33) 優先権主張国	欧州特許庁 (EP)	(74) 代理人	100091214
			弁理士 大貫 進介

最終頁に続く

(54) 【発明の名称】 トピック特異的言語モデルおよびトピック特異的ラベル統計によるユーザー対話を用いたテキストセグメント分割およびラベル付与

(57) 【特許請求の範囲】

【請求項1】

コンピュータによって実行される方法であって：

構造化されていないテキストをテキストセクションにセグメント分割した結果を、セグメント分割およびトピック付与手段から、受領手段によって、受領する段階であって、前記結果は少なくとも一つのテキストセクションについて該少なくとも一つのテキストセクションの内容を示すトピックを含む、段階と；

前記少なくとも一つのテキストセクションおよび該少なくとも一つのテキストセクションについてのセクション見出しを含む第一の構造化されたテキストをユーザーに対して、出力手段によって、出力する段階であって、前記セクション見出しは前記少なくとも一つのテキストセクションに割り当てられた前記トピックに対応し、前記少なくとも一つのテキストセクションに付与された前記トピックは複数のセクション見出しに関連付けられており、前記少なくとも一つのテキストセクションについての前記セクション見出しは前記複数のセクション見出しから選択される、段階と；

前記第一の構造化されたテキストに対する少なくとも一つの修正を指示するユーザー入力を、入力手段によって、受領する段階と；

前記第一の構造化されたテキストを、ユーザーから受領された前記少なくとも一つの修正に従って、修正手段によって修正して第二の構造化されたテキストを生成する段階とを含む、方法。

【請求項 2】

コンピュータによって実行される方法であって：

構造化されていないテキストをテキストセクションにセグメント分割した結果を、セグメント分割およびトピック付与手段から、受領手段によって、受領する段階であって、前記結果は少なくとも一つのテキストセクションについて該少なくとも一つのテキストセクションの内容を示すトピックを含む、段階と；

前記少なくとも一つのテキストセクションおよび該少なくとも一つのテキストセクションについてのセクション見出しを含む第一の構造化されたテキストをユーザーに対して、出力手段によって、出力する段階であって、前記セクション見出しは前記少なくとも一つのテキストセクションに割り当てられた前記トピックに対応する、段階と；

前記第一の構造化されたテキストに対する少なくとも一つの修正を指示するユーザー入力を、入力手段によって、受領する段階と；

前記第一の構造化されたテキストを、ユーザーから受領された前記少なくとも一つの修正に従って、修正手段によって修正して第二の構造化されたテキストを生成する段階とを含み、

前記少なくとも一つのテキストセクションに付与されたトピックは複数のセクション見出しに関連付けられており、前記少なくとも一つのテキストセクションについての前記セクション見出しは、前記複数のセクション見出しから選択され、

前記セクション見出しの選択は、トピックに付与されるテキストセクションの前に特定のセクション見出しがくる頻度を反映するトレーニング・データに基づく計数統計を使うことによって、および/またはテキストセクションの先頭に見出される明示的な言語表現を使うことによって行われる、

方法。

【請求項 3】

前記少なくとも一つのテキストセクションについての前記セクション見出しは、前記複数のセクション見出しのうちで、前記少なくとも一つのテキストセクションに割り当てられたトピックについて最も頻繁に選択されるセクション見出しである、請求項 2 記載の方法。

【請求項 4】

前記少なくとも一つのテキストセクションに付与されたトピックに関連付けられた前記複数のセクション見出しをユーザーに提供する段階をさらに含み、前記少なくとも一つの修正は、前記複数のセクション見出しのうちからの、前記少なくとも一つのテキストセクションについての代替的なセクション見出しのユーザーによる選択を含む、請求項 2 記載の方法。

【請求項 5】

前記少なくとも一つの修正は、前記少なくとも一つのテキストセクションについて挿入されたセクション見出しを置換するためにユーザーによって入力される新しいセクション見出しを含み、前記新しいセクション見出しは、前記少なくとも一つのテキストセクションに付与されたトピックに関連付けられた前記複数のセクション見出しのうちのどのセクション見出しとも異なっている、請求項 2 記載の方法。

【請求項 6】

前記セクション見出しは前記第一の構造化されたテキスト中に、ある第一の位置において挿入され、前記少なくとも一つの修正は、前記セクション見出しを前記第一の位置とは異なる第二の位置に移動させて、前記少なくとも一つのテキストセクションの境界を再定義することを含む、請求項 1 記載の方法。

【請求項 7】

前記複数のテキストセクションが第一の複数のテキストセクションであり、当該方法がさらに：

ユーザーから受け取られた前記少なくとも一つの修正を無効にすることなく、前記第二の構造化されたテキストの少なくとも一部を第二の複数のテキストセクションに再セグメ

10

20

30

40

50

ント化する段階と；

前記第二の複数のテキストセクションおよび該第二の複数のテキストセクションのそれぞれについての対応するセクション見出しを含む第三の構造化されたテキストを生成する段階とを含む、

請求項 1 記載の方法。

【請求項 8】

コンピュータによって実行される方法であって；

構造化されていないテキストをテキストセクションにセグメント分割した結果を、セグメント分割およびトピック付与手段から、受領手段によって、受領する段階であって、前記結果は少なくとも一つのテキストセクションについて該少なくとも一つのテキストセクションの内容を示すトピックを含む、段階と；

10

前記少なくとも一つのテキストセクションおよび該少なくとも一つのテキストセクションについてのセクション見出しを含む第一の構造化されたテキストをユーザーに対して、出力手段によって、出力する段階であって、前記セクション見出しは前記少なくとも一つのテキストセクションに割り当てられた前記トピックに対応する、段階と；

前記第一の構造化されたテキストに対する少なくとも一つの修正を指示するユーザー入力を、入力手段によって、受領する段階と；

前記第一の構造化されたテキストを、ユーザーから受領された前記少なくとも一つの修正に従って、修正手段によって修正して第二の構造化されたテキストを生成する段階とを含む、

20

あるテキスト部分を、前記少なくとも一つのテキストセクションについての前記セクション見出しの完全なまたは部分的な言語表現として識別する段階と；

ユーザーに提供される前記第一の構造化されたテキストから前記テキスト部分を除去する段階とをさらに含む、

方法。

【請求項 9】

セグメント分割の粒度がカスタマイズ可能な粒度パラメータを使ってユーザーによって制御される、請求項 1 記載の方法。

【請求項 10】

コンピュータによって実行される方法であって；

30

構造化されていないテキストをテキストセクションにセグメント分割した結果を、セグメント分割およびトピック付与手段から、受領手段によって、受領する段階であって、前記結果は少なくとも一つのテキストセクションについて該少なくとも一つのテキストセクションの内容を示すトピックを含む、段階と；

前記少なくとも一つのテキストセクションおよび該少なくとも一つのテキストセクションについてのセクション見出しを含む第一の構造化されたテキストをユーザーに対して、出力手段によって、出力する段階であって、前記セクション見出しは前記少なくとも一つのテキストセクションに割り当てられた前記トピックに対応する、段階と；

前記第一の構造化されたテキストに対する少なくとも一つの修正を指示するユーザー入力を、入力手段によって、受領する段階と；

40

前記第一の構造化されたテキストを、ユーザーから受領された前記少なくとも一つの修正に従って、修正手段によって修正して第二の構造化されたテキストを生成する段階とを含む、

前記セグメント分割およびトピック付与手段が、構造化されていないテキストをセグメント分割するおよび/または少なくとも一つのテキストセクションにトピックを付与することにおいて、注釈付けされたトレーニング・データから構築される少なくとも一つの統計モデルを使い、ユーザーから受け取られる前記少なくとも一つの修正が前記少なくとも一つの統計モデルを適応させるためにログに記録され、解析され、

前記少なくとも一つの統計モデルはトピックシーケンス確率、トピック位置確率、セクション長確率および/またはテキスト放出確率を含む、

50

方法。

【請求項 1 1】

コンピュータ・システムを有する装置であって、前記コンピュータ・システムは：
 構造化されていないテキストをテキストセクションにセグメント分割した結果をセグメント分割およびトピック付与手段から受け取る段階であって、前記結果は少なくとも一つのテキストセクションについて該少なくとも一つのテキストセクションの内容を示すトピックを含み、前記トピックは複数のセクション見出しに関連付けられている、段階と；

前記少なくとも一つのテキストセクションおよび該少なくとも一つのテキストセクションについてのセクション見出しを含む第一の構造化されたテキストをユーザーに対して出力する段階であって、前記セクション見出しは前記少なくとも一つのテキストセクションに付与されたトピックに関連付けられた前記複数のセクション見出しから選択される、段階と；

前記第一の構造化されたテキストに対する少なくとも一つの修正を指示するユーザー入力を受け取る段階と；

前記ユーザーから受け取られた前記少なくとも一つの修正に従って前記第一の構造化されたテキストを修正して第二の構造化されたテキストを生成する段階とを実行するよう構成されている、
 装置。

【請求項 1 2】

前記少なくとも一つのテキストセクションに付与されたトピックに関連付けられた前記複数のセクション見出しが、前記第一の構造化されたテキストとともにユーザーに提供され、前記少なくとも一つの修正は、前記複数のセクション見出しのうちからの、前記少なくとも一つのテキストセクションについての代替的なセクション見出しのユーザーによる選択を含む、請求項 1 1 記載の装置。

【請求項 1 3】

前記少なくとも一つの修正は、前記少なくとも一つのテキストセクションについて挿入されたセクション見出しを置換するためにユーザーによって入力される新しいセクション見出しを含み、前記新しいセクション見出しは、前記少なくとも一つのテキストセクションに付与されたトピックに関連付けられた前記複数のセクション見出しのうちどのセクション見出しとも異なっている、請求項 1 1 記載の装置。

【請求項 1 4】

前記セクション見出しは前記第一の構造化されたテキスト中に、ある第一の位置において挿入され、前記少なくとも一つの修正は、前記セクション見出しを前記第一の位置とは異なる第二の位置に移動させて、前記少なくとも一つのテキストセクションの境界を再定義することを含む、請求項 1 1 記載の装置。

【請求項 1 5】

コンピュータ・システムを有する装置であって、前記コンピュータ・システムは：
 構造化されていないテキストをテキストセクションにセグメント分割した結果をセグメント分割およびトピック付与手段から受け取る段階であって、前記結果は少なくとも一つのテキストセクションについて該少なくとも一つのテキストセクションの内容を示すトピックを含み、前記トピックは複数のセクション見出しに関連付けられている、段階と；

前記少なくとも一つのテキストセクションおよび該少なくとも一つのテキストセクションについてのセクション見出しを含む第一の構造化されたテキストをユーザーに対して出力する段階であって、前記セクション見出しは前記少なくとも一つのテキストセクションに付与されたトピックに関連付けられた前記複数のセクション見出しから選択される、段階と；

前記第一の構造化されたテキストに対する少なくとも一つの修正を指示するユーザー入力を受け取る段階と；

前記ユーザーから受け取られた前記少なくとも一つの修正に従って前記第一の構造化されたテキストを修正して第二の構造化されたテキストを生成する段階とを実行するよう構

10

20

30

40

50

成されており、

前記複数のテキストセクションが第一の複数のテキストセクションであり、前記コンピュータ・システムがさらに：

ユーザーから受け取られた前記少なくとも一つの修正を無効にすることなく、前記第二の構造化されたテキストの少なくとも一部を第二の複数のテキストセクションに再セグメント化する段階と；

前記第二の複数のテキストセクションおよび該第二の複数のテキストセクションのそれぞれについての対応するセクション見出しを含む第三の構造化されたテキストを生成する段階とを実行するよう構成されている、

装置。

10

【請求項 16】

コンピュータ・システムを有する装置であって、前記コンピュータ・システムは：

構造化されていないテキストをテキストセクションにセグメント分割した結果をセグメント分割およびトピック付与手段から受け取る段階であって、前記結果は少なくとも一つのテキストセクションについて該少なくとも一つのテキストセクションの内容を示すトピックを含み、前記トピックは複数のセクション見出しに関連付けられている、段階と；

前記少なくとも一つのテキストセクションおよび該少なくとも一つのテキストセクションについてのセクション見出しを含む第一の構造化されたテキストをユーザーに対して出力する段階であって、前記セクション見出しは前記少なくとも一つのテキストセクションに付与されたトピックに関連付けられた前記複数のセクション見出しから選択される、段階と；

20

前記第一の構造化されたテキストに対する少なくとも一つの修正を指示するユーザー入力を受け取る段階と；

前記ユーザーから受け取られた前記少なくとも一つの修正に従って前記第一の構造化されたテキストを修正して第二の構造化されたテキストを生成する段階とを実行するよう構成されており、

前記コンピュータ・システムが前記セグメント分割およびトピック付与手段を有し、前記セグメント分割およびトピック付与手段が、構造化されていないテキストをセグメント分割するおよび/または少なくとも一つのテキストセクションにトピックを付与することにおいて、注釈付けされたトレーニング・データから構築される少なくとも一つの統計モデルを使うよう構成されており、前記コンピュータ・システムがさらに、ユーザーから受け取られる前記少なくとも一つの修正を、前記少なくとも一つの統計モデルを適応させるためにログに記録し、解析するよう構成されており、

30

前記少なくとも一つの統計モデルはトピックシーケンス確率、トピック位置確率、セクション長確率および/またはテキスト放出確率を含む、

装置。

【請求項 17】

実行可能な命令がエンコードされている少なくとも一つのコンピュータ可読記憶デバイスであって、前記命令は、コンピュータ・システムによって実行されたときに：

構造化されていないテキストをテキストセクションにセグメント分割した結果をセグメント分割およびトピック付与手段から受け取る段階であって、前記結果は少なくとも一つのテキストセクションについて該少なくとも一つのテキストセクションの内容を示すトピックを含み、前記トピックは複数のセクション見出しに関連付けられている、段階と；

40

前記少なくとも一つのテキストセクションおよび該少なくとも一つのテキストセクションについてのセクション見出しを含む第一の構造化されたテキストをユーザーに出力する段階であって、前記セクション見出しは前記少なくとも一つのテキストセクションに付与されたトピックに関連付けられた複数のセクション見出しから選択される、段階と；

前記第一の構造化されたテキストに対する少なくとも一つの修正を指示するユーザー入力を受け取る段階と；

前記ユーザーから受け取られた前記少なくとも一つの修正に従って前記第一の構造化さ

50

れたテキストを修正して第二の構造化されたテキストを生成する段階とを含む、
方法を実行する、
コンピュータ可読記憶デバイス。

【請求項 18】

前記少なくとも一つのテキストセクションについての前記セクション見出しは、前記複数のセクション見出しのうちで、前記少なくとも一つのテキストセクションに割り当てられたトピックについて最も頻繁に選択されるセクション見出しである、請求項 17 記載のコンピュータ可読記憶デバイス。

【請求項 19】

前記方法がさらに、前記少なくとも一つのテキストセクションに付与されたトピックに関連付けられた前記複数のセクション見出しをユーザーに提供する段階を含み、前記少なくとも一つの修正は、前記複数のセクション見出しのうちからの、前記少なくとも一つのテキストセクションについての代替的なセクション見出しのユーザーによる選択を含む、請求項 17 記載のコンピュータ可読記憶デバイス。

10

【請求項 20】

前記少なくとも一つの修正は、前記少なくとも一つのテキストセクションについて挿入されたセクション見出しを置換するためにユーザーによって入力される新しいセクション見出しを含み、前記新しいセクション見出しは、前記少なくとも一つのテキストセクションに付与されたトピックに関連付けられた前記複数のセクション見出しのうちどのセクション見出しとも異なっている、請求項 17 記載のコンピュータ可読記憶デバイス。

20

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、非構造化テキストをテキストセクションにセグメント分割して各セクションにセクション見出しとしてラベルを付与することによって、非構造化テキストから構造化文書を生成する分野に関する。テキストのセグメント分割およびテキストセクションへのラベル付与（ラベル付けとも記される）が提供されるユーザーは、セグメント分割およびラベル付け手続きの制御を有する。

【背景技術】

【0002】

音声テキスト化転記プロセスによって生成されたテキスト文書は通例、いかなる構造も与えてくれない。従来式の音声テキスト化転記システムや音声認識システムは録音された発話に対応するテキストに文字として転記するのみだからである。テキストの整形、テキストのハイライト、句読点付けまたはテキスト見出しの明示的に口述されたコマンドが、音声認識システムによって、あるいは音声認識プロセスによって生成されたテキストにその後適用されるテキスト整形手続きによって適正に認識されて処理されなければならない。

30

【0003】

自動音声認識も、典型的にはトレーニング・データおよび/または手作業で設計されたテキスト整形規則に基づく自動テキスト整形システムもいずれも、複雑な整形コマンド、セクション境界、そしてたとえばセクション見出しを表す特定のテキスト部分を適正に識別するのに必要とされる人間の専門知識を欠いているため、必然的に誤りを生じる。したがって、通常音声テキスト化転記プロセスまたはテキスト整形プロセスの結果は人間の校正者を通す必要がある。校正者は、文書全体を通して見ることでその文書の内容についての情報を集め、音声テキスト化転記プロセスの出力が理にかなった結果であるかどうか、テキスト整形が文書内容に関して正しく実行されたかどうかを判断しなければならない。

40

【0004】

文書の構造が明示的に口述されていないとき、すなわち多くの見出しやセクション境界が発話による口述で明示的にエンコードされていないときには校正者の仕事はさらにひど

50

くなる。さらに、文構造すなわち句読点記号さえもがほとんど口述されないときには、そうした句読点記号は校正者が手動で挿入してやる必要がある。

【0005】

特にテキストのセクションへの分割は校正者にとっては負担の大きい仕事である。これは、あるセクション種別の変化を検出するのは校正者が新しいセクションのより長い部分を読んでからでないと判断できないからである。ここで、校正者は、セクション境界および適切な見出しを挿入するためにすでに吟味したテキスト内の何らかの位置にジャンプして戻らなければならない。特に、文書内のさまざまな位置の間で絶えずジャンプすることは、人間の校正者にとって、非常に時間がかかり、骨の折れることである。

【発明の概要】

【発明が解決しようとする課題】

【0006】

本発明は、ユーザーの判断に反応して非構造化テキストのセグメント分割およびラベル付けを実行するための、方法、コンピュータプログラムプロダクト、テキストセグメント分割システムおよびテキストセグメント分割システムのためのユーザーインターフェースを提供することをねらいとしている。

当業者は、本発明の目的はこれに限定されないことを理解するであろう。たとえば、本発明はユーザーの校閲に反応しての構造化テキストの修正の手段を提供することもねらいとしている。

【課題を解決するための手段】

【0007】

本発明は、テキストをテキストセクションにセグメント分割し、各セクションにトピックを付与し、各テキストセクションにセクション見出しの形でラベルを付与する方法を用いるテキスト処理システムのための効率的なユーザーインターフェースを提供する。これらのタスクは注釈付けされたトレーニング・データに基づいてトレーニングされた統計モデルを使って実行される。まず、当該方法はトレーニング・データから抽出された統計モデルを使うことによってテキストのテキストセクションへのセグメント分割を実行する。テキストがテキストセクションにセグメント分割されたのち、各テキストセクションはテキストセクションの内容を示すトピックを付与される。テキストセクションへのトピックの付与にあたってはトレーニング・データから抽出された統計モデルが使用される。テキストセグメント分割およびトピック付与が実行されたのち、当該テキストにラベルをセクション見出しとして挿入することによって構造化されたテキストが生成される。ラベルは当該テキスト中で、そのラベルのすぐあとにそのラベルが言及しているセクションが続くよう、セクション境界に対応する位置に挿入される。この挿入されたラベルは、続くテキストセクションに先立つ見出しとして理解されるべきである。

【0008】

上記の方法で構造化テキストが生成されたら、該構造化テキストは、セグメント分割、トピック付与およびテキストの一般的な構造化の制御を有するユーザーに提供される。当該方法は最後に、ユーザーの校閲に反応してその構造化テキストの修正を実行する。

【0009】

本発明のある好ましい実施形態によれば、セクション見出しとしてのラベルの挿入には、句読点付け、ハイライト、インデントおよび書体の修正といった整形ステップを組み込んでいるテキスト整形手続きが含まれる。

【0010】

本発明のあるさらなる好ましい実施形態によれば、テキストセクションへのトピック付与は、当該テキストセクションへの複数のラベルの集合の付与をも含む。付与されたラベルの集合のうちの一つのラベルが最終的にセクション見出しとしてそのテキストに挿入される。ここではトピック (topic) は、セクションの明確に区別される (distinct) 類 (class) または種別 (type) のやや抽象的な宣言を表している。そのような宣言は、典型的なまたは所定の構造に従ういわゆる組織化文書に特に適用可能である。たとえば、医療報

10

20

30

40

50

告書は人口学的ヘッダ、患者の病歴、健康診断および使用薬のようなトピックの系列を特徴とする。

【0011】

そのような構造化文書の各セクションは抽象的なトピックによって識別できる。抽象的なトピックとは対照的に、ラベルはそのようなセクションの具体的な見出しを示す。たとえば、患者の検査に言及するセクションは「健康診断」「診断」「検査」「外科的検査」といった複数のさまざまな方法でラベル付けできる。テキストのあるセクションにラベル付けする方法がどうであれ、セクションの内容、すなわち今の場合だと検査は、付与されたトピックによって同定される。

【0012】

テキストのテキストセクションへのセグメント分割は、たとえば米国特許第6,052,657号で開示されている方法によって実行できる。これは言語モデルならびにテキストのあるブロックと言語モデルとの間の相関を示すための言語モデルスコアを利用している。テキストのセグメント分割およびトピック付与のためのより精密で信頼できる手続きは、本出願と並行して出願された特許出願“text segmentation and topic annotation for document structuring”において開示されている。この文書は、テキストセグメント分割およびトピック注釈付けのための、トピックシーケンス確率、トピック位置確率、セクション長確率およびテキスト放出(emission)確率を明示的に利用することによる統計モデルを記載している。これらの確率は、根底にある注釈付けされたトレーニング・データが組織化文書であるときには特に助けになる。

【0013】

本発明のあるさらなる好ましい実施形態によれば、テキストセクションへのラベルの集合のうちの一つのラベルの付与および付与されたその一つのラベルを前記テキストセクションのセクション見出しとして当該テキストに挿入することは、トレーニング・データに基づく計数統計および/またはセクション冒頭において見出される明示的なもしくは部分的な言語表現を考慮する。計数統計は何らかのトピックに対応付けられたセクションの前にある特定のラベルがくる観察された頻度を反映するものである。この方法により、最も好適なラベルまたは見出しについてのほかの手がかりが全くテキスト中に見出されない場合に、トピックごとに最も頻繁に付与されるラベルがデフォルト見出しとして選択される。換言すれば、計数統計量によってデフォルトラベルがテキストセクションに付与されるのである。

【0014】

あるいはまた、前記計数統計に基づくラベル付与は、あるセクションの冒頭に、当該セクションに付与されているラベルの集合のうちの一つに厳密に一致する明示的な言語表現が見出された場合には無効にされる(overruled)。さらに、セクションの冒頭で明示的な言語表現に厳密に一致するラベルがない場合、当該セクションの冒頭に見出された何らかの言語表現に部分的にのみ一致するラベルがデフォルトラベルの代わりに挿入されてもよい。一つのラベルのテキストセクションへの付与、すなわち当該テキストセクションに付与されているラベルの集合のうちの一つのラベルの選択は、トレーニング・データに基づく計数統計とセクション冒頭に見出された完全な明示的なまたは部分的な言語表現とを

【0015】

本発明のあるさらなる好ましい実施形態によれば、セクション冒頭で何らかの完全または部分的な言語表現が見出された場合、この言語表現が当該セクションから除去されてもよい。これは、その言語表現が、挿入されるラベルによって置き換えられる明示的に口述された見出しを表している場合に有用である。例を挙げると、「薬患者が使用するのは...」で始まるセクションは「薬」というラベルに対応付けできる。このラベルはそのあとに続くセクションのための見出しの役割をするので、「薬」という用語そのものはセクションのテキストからは除去して、「患者が使用するのは...」で始まる適切なセクション内容が残るようにするべきである。この方針の修正としては、口述された見出しの一部または

10

20

30

40

50

何らかのセクションの冒頭句でありうる何らかの所定のつなぎ語 (filler) を、たとえそのつなぎ語がラベルの一部でなくても除去することが考えられる。たとえば、あるセクションが「薬はX、Y、Z...」で始まっている場合に、これを「薬」という見出しとそれに続く薬の列挙「X、Y、Z...」に変換し、つなぎ語「は」をスキップする。

【0016】

本発明のあるさらなる好ましい実施形態によれば、たとえば明示的な言語表現とラベルとの間の厳密な一致などによる当該テキストへのセクション見出しの挿入は、ユーザーによって無効にされることができる。この場合、当該方法によって挿入が取り消され、もとのテキスト部分が復元される。より具体的には、何らかのセクション冒頭語が付与されたラベルとの一致したために除去された場合、ユーザーがこれらの除去された語に一致しない別のラベルを使うことに決めるときにはこれらの語が再挿入される必要がある。

10

【0017】

本発明のあるさらなる好ましい実施形態によれば、構造化テキストのユーザーへの提供はさらに、各テキストセクションに付与されているラベルの完全な集合を提示することを含む。ラベルの集合の各ラベルはセクション見出しのための代替を表しているので、ユーザーは自動的に挿入されたセクション見出しを代替の見出しと簡単に比較することができる。

【0018】

本発明のあるさらなる好ましい実施形態によれば、構造化テキストのユーザーへの提供はさらに、代替的なセクション境界の指標を提示することを含む。この方法では、本方法によってテキスト中に自動挿入されたセクション境界がユーザーに見えるばかりでなく、より簡単かつ簡便な校正のために、代替のセクション境界がユーザーに提示される。この方法では、当該文書の正しいセクション境界を見出すという校正者の仕事は、自動挿入されたセクション境界および代替的なセクション境界の取得に還元される。

20

【0019】

本発明のあるさらなる好ましい実施形態によれば、ユーザーの校閲に反応しての構造化テキストの修正は、当該テキストのテキストセクションへのセグメント分割および/またはラベルとテキストセクションとの間の対応付けの修正に関わる。さらに、句読点付け、ハイライトなどといった実行された整形の修正も考えられる。

【0020】

本発明のあるさらなる好ましい実施形態によれば、ユーザーの校閲に反応して実行されるテキストセグメント分割の修正およびテキストセクションへのラベルの付与の修正は、ユーザーが、提示されたラベルの一つまたは代替的セクション境界の一つを選択することによって開始される。するとユーザーによって選択されたその修正が本方法によって実行され、セクション見出しを選択されたセクション見出しで置き換えたり、セクション境界の位置を変えたりする。

30

【0021】

ある第一のテキスト修正を達成することは、第二のテキスト修正を実行しなければならないことを意味することがある。たとえば、セクション見出しに番号が付いている場合、あるテキストセクションを除去すれば、後続のテキストセクションまたはセクションラベルの再番号付けが必要になる。したがって、本発明はさらに、ユーザーの校閲に反応して実行される先の修正に起因する修正を動的に実行するよう適応される。

40

【0022】

本発明のあるさらなる好ましい実施形態によれば、テキストセクションへのセクション見出しとしてのラベルの付与の修正は、ユーザーが、テキストセクションに付与されているラベルの与えられた集合のうちの一つのラベルを選択することに反応して、あるいはユーザー定義のラベルを入力してこのユーザー定義ラベルをセクション見出しとして当該テキストセクションに付与することによって実行される。このようにして、ユーザーは迅速かつ効率的に与えられているラベルの集合のうちの一つのラベルを正しいセクション見出しとして同定したり、あるいはまた当該テキストセクションに対してそれまで知られてい

50

なかった見出しを定義したりすることができる。

【0023】

ラベルの集合のうちの一つのラベルの選択も、ラベルの入力も、当該テキスト中でセクション境界として識別された位置に限定されるものではない。それに加えて、ユーザー要求に基づき、当該テキスト中の任意の位置に適切なラベルの集合を与えることができる。このようにしてユーザーはやはり文書の構造化およびラベル付けの完全な制御を有する。

【0024】

本発明のあるさらなる好ましい実施形態によれば、ユーザーの校閲に反応しての修正の処理があると、テキストのテキストセクションへの再セグメント分割およびテキストセクションを指すセクション見出しとしてラベルを挿入することによる構造化テキストの再生成がその後引き起こされる。再セグメント分割も構造化テキストの再生成も、トレーニング・データから抽出された統計モデルを利用し、ユーザーの校閲に反応して処理されたすでに実行済みの修正を参照する。たとえばユーザーがセクション境界の再定義の形で、あるいはセクション見出しの再ラベル付けの形でテキスト中に修正を導入した場合、本発明の方法は、最初に行われたユーザーの修正は変えずに残して当該構造化テキストのその後の再セグメント化および再生成を実行する。このように、ユーザーによって導入された修正が本発明の方法によって無効にされたり再修正されたりすることは決してない。

【0025】

本発明のあるさらなる好ましい実施形態によれば、テキストのセクションへの再セグメント分割も、セクション見出しとしてラベルを挿入することによる構造化テキストの再生成も、校正者またはユーザーによって実行される校閲プロセスの間、動的に行われる。テキストの再セグメント分割も構造化テキストの再生成も、全テキストセクションに適用することもできるし、現在のセクションおよび後続の全セクションに適用することもできるし、ユーザーによって指定されれば単一のセクションに適用されることもできる。たとえば、新しいセクション境界が導入されたとき、あるいはユーザーによって見出しが除去されたときには、さらなる再構造化や見出し更新は現在のセクションのみに限定されることが理にかなっている。このようにして、本方法は、テキストに導入される必要のある小規模な、よって局所的な変更に対してより速く反応できる。

【0026】

本発明のあるさらなる好ましい実施形態によれば、テキストセグメント化の粒度を、粒度パラメータというものをカスタマイズすることによってユーザーが制御できる。このようにして、ユーザーは、テキストの構造化が細かめか粗めかを決定できる。カスタマイズ可能な粒度パラメータを変更すれば、その結果としてテキストセクションの除去または挿入が生じる。

【0027】

本発明のあるさらなる好ましい実施形態によれば、ユーザーの校閲に反応して実行される修正は、統計モデルをさらにトレーニングするために本方法によってログに記録され、解析される。このようにして、本方法全体をユーザーの嗜好に効率的に適応させることができる。たとえば、ユーザーがテキストからある特定のラベルを繰り返し除去している場合、本テキストセグメント分割方法は、将来の適用においては、この特定のセクション見出しを挿入することを控える。ユーザーによる修正が本方法の適応に影響する度合いつまり適応の感度もユーザーが制御できる。これはたとえば、あるラベルの挿入または除去が所定回数生じしてはじめて本方法がこの特定のユーザー導入修正に適応するということを意味する。導入された変更对本方法が適応するまでに手動による変更が何度加えられる必要があるかは、ユーザーが与えてもよい。

【0028】

さらに、ユーザー導入修正への本方法の適応は、現在の文書における後続セクションをもすでに指定していることができる。本方法は、ある文書の始まる部分でユーザーによって導入された修正に適応し、後続のテキストセクション内では対応する修正を自動的に実行する。したがって、この適応は現在の文書ならびに本発明の方法が適用される将来の

10

20

30

40

50

文書にも適用される。

以下で、本発明の好ましい実施形態について図面を参照しつつより詳細に説明する。

【図面の簡単な説明】

【0029】

【図1】本発明のセグメント分割法のフローチャートを示す図である。

【図2】ユーザー導入修正の解析を組み込んだテキストセグメント分割のためのフローチャートを示す図である。

【図3】音声認識プロセス中に本発明を実装するフローチャートを示す図である。

【図4】本発明のユーザーインターフェースのブロック図である。

【図5】セグメント分割システムのブロック図である。

10

【発明を実施するための形態】

【0030】

図1は、テキストセグメント分割およびトピック付与方法のフローチャートを示している。第一のステップ100では、音声テキスト化転記システムなどによって生成された非構造化テキストが入力される。入力されたテキストに基づき、ステップ102では本方法は当該テキストをテキストセクションにセグメント分割して各テキストセクションにトピックを付与することによって、構造化およびトピック付与を実行する。ステップ102におけるテキストセグメント分割およびトピック付与を実行するためには、トレーニング・データから抽出される言語モデルまたは統計モデルがステップ104によってステップ102に提供される。ステップ105は、あるラベルがあるトピックに付与される確率を示すラベル計数統計を提供する。ラベル計数統計は、トレーニング・データに基づいてあるラベルがあるトピックに付与される頻度を反映するものである。

20

【0031】

ステップ106では、ステップ105で提供された計数統計およびステップ102で提供されたセグメント分割テキストを参照することによって、各テキストセクションにラベルがセクション見出しとして付与され、当該テキスト中の適切な位置に挿入される。ステップ106によってラベル付与が実行されたのち、セグメント分割されたテキストおよび挿入されたラベルならびに代替的なラベルがステップ108でユーザーに提示される。さらに、ステップ108では代替的なセクション境界がユーザーに提示される。後続ステップ110において、ユーザーはステップ108の提供されたセグメント分割およびラベル付与が許容できるかどうかを決定する。あるいはまた、ユーザーは、ステップ108によって提示された代替見出しまたは代替セクション境界によって提示された代替セグメント分割を選択することもできる。

30

【0032】

提示された代替のいずれもユーザーの嗜好を満たさない場合、ユーザーはセクション境界やセクション見出しを入力することもできる。ステップ110のユーザーの決定に反応して、ステップ112でユーザーの決定が本方法によって処理される。ユーザーの決定の処理は、挿入されたセクション見出しの置き換え、後続セクション見出しの再ラベル付け、当該文書の後続部分の再構造化または当該文書全体の再構造化および再ラベル付けを含む。さらに、ユーザーが導入した修正の動的処理も考えられる。動的処理とは、ユーザーが修正を導入すると、後続テキストセクションに関係したさらなる修正または当該構造化方法の以後の適用の際に実行されるべき修正を自動的に引き起こすということを意味する。

40

【0033】

ステップ112でユーザー決定が処理されたのち、次のステップ114で結果としての修正が実行される。

【0034】

図2は、ユーザーの導入した修正の解析を組み込んだテキストセグメント分割およびテキスト付与方法のフローチャートを示している。第一のステップ200では、音声テキスト化転記プロセスなどの結果として得られる非構造化テキストがステップ202に提供さ

50

れる。ステップ202では、テキストセクションへのテキストセグメント分割が、ステップ204によって提供される言語モデルまたは統計モデルを使用することによって実行される。さらに、ステップ202では、ステップ204によって提供される言語モデルに保存されている統計情報を使用することによって各テキストセクションにトピックが付与される。

【0035】

ステップ202で当該テキストがテキストセクションにセグメント分割されたのち、そして各テキストセクションがあるトピックに対応付けられたのち、後続ステップ206で、各テキストセクションにセクション見出しとしてラベルが付与され、テキスト中の適切な位置に挿入される。ステップ206で実行されたラベル付与は、ステップ205によってステップ206に提供されるラベル計数統計を明示的に使用する。ラベル計数統計は、トレーニング・データに基づいて、あるラベルがあるトピックに付与される頻度を反映するものである。

10

【0036】

当該テキストをテキストセクションにセグメント分割し、各テキストセクションにトピックを付与し、さらに各テキストセクションにラベルを付与することによってテキストが構造化されたのち、付与された見出しが代替候補とともにステップ208でユーザーに提示される。ユーザーに提示される代替候補とは、代替的なテキストセグメント分割ならびに代替的なセクションラベルのことである。次のステップ210では、ユーザーは実行されたテキストセグメント分割および実行されたセクションラベル付与を受け入れるかどうか、あるいは提示された代替候補の一つを選択するかどうかを決定する。さらに、ユーザーは任意のセグメント分割ならびに任意のセクション見出しを自分の嗜好に従って入力することもできる。ステップ210のユーザー決定ののち、次のステップ214で本方法はユーザーによって何らかの修正が導入されたかどうかを調べる。ステップ214でユーザー導入修正が検出されなければ、本方法はステップ218で終了し、ステップ206で実行されたような構造化およびラベル付けされたテキストを結果として与える。これに対し、ステップ214でユーザー導入修正が検出された場合、本方法はステップ212に進む。ここではユーザー導入修正が処理され、実行される。ユーザーの決定の処理および実行は異なるテキストセグメント分割、テキストラベル付けおよびテキスト整形手続きの複数を組み込む。

20

30

【0037】

ステップ212でユーザー決定が処理され、実行されたのち、本方法はステップ216に進む。ステップ216では、ユーザー導入修正が、本構造化および付与手続きの次の適用のための永続条件として保存される。ステップ216後、前記ユーザー修正の種類がテキスト構造化に関わるかテキストセクションへのラベル付与に関わるに応じて、本方法はステップ202またはステップ206に戻り、新しい構造化または新しいラベル付与が実行される。

【0038】

同様にして、ステップ202およびステップ206によって実行される当該テキストの新しい再構造化および新しい再付与は、ステップ216によって提供されるすでに実行された修正を明示的に考慮する。このようにして、ユーザーが実行した修正がテキスト構造化ステップ202およびラベル付与ステップ206によって決して無効にされないことが保証される。

40

【0039】

図3は、テキストセグメント分割およびトピック付与方法の音声認識システムへの実装を示している。ステップ300で音声が入力される。次のステップ302では、音声の第一の部分 $p=1$ が選択される。ステップ302によって選択された音声の第一の部分はステップ304に提供され、該ステップ304は言語モデル m を使用することによって音声テキスト化転記を実行する。言語モデル m はステップ306によってステップ304に提供される。ステップ304によって音声部分 p がテキスト部分 t に転記されたのち

50

、ステップ308では結果として得られる、音声部分pに対応するテキスト部分tが保存される。次のステップ310では、音声部分の添え字pが最後の音声部分を示す p_{max} と比較される。pが p_{max} より小さければ、pは1インクリメントされ、本方法はステップ304に戻る。ステップ304、308、310は音声部分の添え字pが最後の音声部分 p_{max} に等しくなるまで繰り返し適用される。音声部分の添え字pが最後の音声部分 p_{max} に等しい場合には、音声信号全体がテキストに転記されたことになる。そのとき、結果として得られるテキストは、複数の音声部分pに対応する複数のテキスト部分tからなる。

【0040】

転記されたテキストに基づいて、ステップ312において、当該テキストのテキストセクションへのセグメント分割が実行され、各テキストセクションは各セクションの内容に特異的なトピックに対応付けられる。ステップ312のこのセグメント分割手続きは、ステップ314によってステップ312に提供される、テキストセグメント分割のために設計された統計モデルを使用する。ステップ312で当該テキストがセグメント分割されたトピックに対応付けられたら、次のステップ316では、各テキストセクションに付与されたトピックが該テキストセクションの対応する音声部分pと並んで決定される。この決定に基づいて、次のステップ318ではある特定のセクションを指している音声部分pの第二の音声認識が実行されうる。ステップ306によって、あるテキストセクションに付与されているトピックに応じて、第二の音声認識のためのトピック特異的な言語モデルが提供される。ステップ300から310によって記述された手続きにおいて音声はすでにステップごとに転記されているので、繰り返しの音声認識は、音声部分pに対応する特定のテキスト部分について選択的に実行できる。

【0041】

前記繰り返しの音声認識ステップが当該テキストの各セクションについて実行されたら、ステップ320においてユーザーは当該テキストのセグメント分割に関するさらなる修正を導入することができる。ステップ320のユーザー導入修正により、本方法はテキストセグメント分割ステップ312に戻る。ここで、ユーザーのフィードバックに応じて新しいセグメント分割が生起してもよいし、ならびに/またはセクションをトピックおよびラベルに再対応付けしてもよい。

【0042】

ステップ312の実行されたテキストセグメント分割およびステップ318の繰り返しの音声認識ステップがいずれもユーザーによって了承されたら、本方法はステップ322で終了する。

【0043】

ステップ316において実行されるトピックとセクションとの間の対応付けは、ステップ304によって実行される音声転記とともにやはり、前記特許出願“Text segmentation and topic annotation for document structuring”において、および本出願人によって本願と並行して出願された特許出願“Text specific models for text for matting and speech recognition”によって記載されているようなテキストセグメント分割およびトピック注釈付けの方法を明示的に利用する。

【0044】

このようにして、人間の校正者の専門知識を、普遍的かつ効率的にテキストセグメント分割およびテキストラベル付けに、ならびに対応する音声認識手続きに結び付けることができる。

【0045】

図4は、本発明のユーザーインターフェースのブロック図を示している。ユーザーインターフェース400は好ましくはグラフィカルユーザーインターフェースとして適応される。ユーザーインターフェース400はテキストウィンドウ402および提案ウィンドウ404を有する。テキストセグメント分割およびラベル付与にかけられたテキストはテキストウィンドウ402内に提示される。テキスト内にセクション見出しとして挿入されたラベル406は、テキストウィンドウ402内で提示されるテキストの中でみつけやすい

10

20

30

40

50

ようハイライトされる。ユーザーがたとえばポインタ408を使うとき、そのユーザーはラベル406を選択でき、ラベル406の選択に反応してラベルリスト410がユーザーインターフェース内で提示される。ラベルリスト410は、ラベル406の代わりにテキスト中に挿入されうる代替的なラベルとなることのできるラベル412、414、416の集合全体を提示する。

【0046】

追加的または代替的に、ラベルリスト410は提案ウィンドウ404内でも提示されることができる。ポインタ408によって、ユーザーはラベルリスト410によって提示されるラベル412、414、416のうちの一つを選択して所与のテキスト中のラベル406を置き換えることができる。ラベル406、412、414、416のいずれもユーザーの嗜好に合わないときには、ユーザーはユーザー入力欄418を使ってラベルを入力できる。ひとたび代替ラベルがユーザーによって選択または入力されたら、ラベル406は代替ラベルによって置き換えられる。同じようにして、代替的なテキストセグメント分割を用いたテキストのセグメント分割が代替的なセクション境界の形でユーザーに提示され、ユーザーの選択に基づいて実行されうる。

ユーザーは、当該テキスト中の第一の位置で付与されたラベル(406)を選択し、その付与されたラベルを当該テキスト中の第二の位置に移動することによってセクション境界を再定義してもよい。前記第二の位置がセクション境界を定義し、前記選択されたラベルが前記セクション見出しを定義する。

【0047】

図5は、本発明のセグメント分割システムのブロック図を示している。セグメント分割システム500はグラフィカルユーザーインターフェース520、構造化テキストを保存するための構造化テキストモジュール518、処理ユニット516、統計モデルを保存する統計モデルモジュール514、非構造化テキストを保存する非構造化テキストモジュール512および音声テキスト化転記を実行する音声認識モジュール510を有している。セグメント分割システム500は外部記憶装置508および入力装置504に接続されている。ユーザー506は、セグメント分割システム500の入力装置504およびグラフィカルユーザーインターフェース520を通じてセグメント分割システムと対話できる。

【0048】

セグメント分割システムに入力された音声502は音声認識モジュール510によって処理される。音声認識モジュール510は非構造化テキストモジュール512に接続されており、そこに音声テキスト化転記プロセスの結果として得られる非構造化テキストが保存される。該非構造化テキストモジュール512は処理ユニット516に接続されており、非構造化テキストを該処理ユニット516に与えるようになっている。処理ユニット516は統計モデルモジュール514に双方向的に接続されている。統計モデルモジュール514に保存されている統計モデルによって与えられる統計情報を使うことによって、処理ユニット516は、前記非構造化テキストモジュール512によって与えられた非構造化テキストに基づいて、テキストセグメント分割および当該テキストのセクションへのラベル付与を実行する。音声認識モジュールはさらに、前記統計モデルモジュールによって保存され、提供される言語モデルを使用する。このようにして、統計モデルモジュールはテキストセグメント分割のための言語モデルとともに音声認識のための言語モデルをも提供する。テキストセグメント分割が通例一字接続を使うのに対して音声認識は通例三字接続を使うので、後者は典型的には、テキストセグメント分割のための言語モデルと比べて異なる種類である。

【0049】

処理ユニット516がテキストセグメント分割およびテキストセクションへのセクション見出しとしてのラベルの付与を実行したら、そのようにして生成された構造化テキストは構造化テキストモジュール518に保存される。構造化テキストモジュールはグラフィカルユーザーインターフェース520に接続されており、構造化テキストモジュール518に保存されている構造化テキストをユーザー506にグラフィカルユーザーインターフ

10

20

30

40

50

エース520によって提示するようになっている。ユーザー506は入力装置504を通じてセグメント分割システムと対話できる。したがって、入力装置504はグラフィカルユーザーインターフェース520と処理ユニット516とに接続されている。ユーザー506がテキスト構造化またはラベル付与のどちらかの修正を導入すると、処理ユニット516は構造化テキストモジュール518に保存されている構造化テキストの再構造化および再対応付けを実行する。再構造化および再対応付けされた構造化テキストは、実行された修正がユーザーの嗜好に一致するまで繰り返しユーザーに提示される。それ以上の変更がユーザーによって導入されなくなったら、構造化テキストモジュール518に保存されている構造化テキストは外部記憶装置508に伝送される。

【0050】

さらに、構造化テキストモジュール518に保存されている構造化テキストは、音声認識モジュール510によって実行される改良された音声認識のために活用されることもできる。したがって、構造化テキストモジュール518は音声認識モジュール510に直接接続される。このコンテキスト特異的なフィードバックの利用により、より正確で特定の音声認識手続きが音声認識モジュール510によって実行できるようになる。

【0051】

このように、本発明は、文書を構造化し、テキストセクションにセクション見出しのたらしきをするラベルを付与する方法を提供する。特に、自動音声認識および自動音声転記の分野において、人間の校正者によって実行されるべき校正の仕事がきわめて容易化される。提案されるテキストのセグメント分割については、校正者にとって、何らかの見出しに続くテキストが本当に対応する種別のセクションを表しているかどうかを確認することは、テキストの一部分を吟味し、セクションを判別し、セクションの始まりに戻ってテキストに見出しを挿入しなければならない従来式の校正手続きに対比してずっと簡単である。

【0052】

さらに、本方法は、校正者によって簡単に選択できる代替的なセクション境界および代替的なセクションラベルを供給する。さらに、校正プロセスの間、システムは校正者によって導入された最も頻繁な訂正を学習し、この情報を将来の適用のために利用する。

【0053】

いくつかの態様を記載しておく。

〔態様1〕

テキストをテキストセクションにセグメント分割し、注釈付けされたトレーニング・データに基づいて各テキストセクションにトピックを付与する方法であって、

- ・トレーニング・データから抽出された統計モデルを使うことによって当該テキストをテキストセクションにセグメント分割し、
 - ・前記トレーニング・データから抽出された統計モデルを使うことによって各テキストセクションに該テキストセクションの内容を示すトピックを付与し、
 - ・前記ラベルを前記テキストセクションに付与するために当該テキストにラベルをセクション見出しとして挿入することによって、構造化されたテキストを生成し、
 - ・前記構造化されたテキストをユーザーに提示し、
 - ・ユーザーの校閲に反応して前記構造化されたテキストの修正を処理する、
- ステップを有することを特徴とする方法。

〔態様2〕

テキストセクションに付与された前記トピックがさらにラベルの集合に付与されており、該ラベルの一つが前記テキストセクションに付与されて当該テキストにセクション見出しとして挿入されることを特徴とする、態様1記載の方法。

〔態様3〕

前記構造化されたテキストをユーザーに提示することがさらに、各テキストセクションについて、該テキストセクションに付与された前記トピックに付与された前記ラベルの集合を提示することを含むことを特徴とする、態様1または2記載の方法。

10

20

30

40

50

〔 態様 4 〕

前記テキスト修正が、当該テキストのセクションへのセグメント分割の修正ならびに / またはラベルとテキストセクションとの間の対応付けの修正を含むことを特徴とする、態様 1 ないし 3 のうちいずれか一項記載の方法。

〔 態様 5 〕

態様 3 または 4 記載の方法であって、前記構造化されたテキストの修正が：

- ・あるテキストセクションにラベルを、該テキストセクションに付与されている前記トピックに付与された前記ラベルの集合のうちの一つのラベルを選択することによって付与し、

- ・当該テキスト中の第一の位置で付与されたラベルを選択し、その付与されたラベルを当該テキスト中の第二の位置に移動することによってセクション境界を再定義し、前記第二の位置がセクション境界を定義し、前記選択されたラベルが前記セクション見出しを定義し、

- ・ラベルを入力し、該入力されたラベルを前記テキストセクションに付与する、ことを含むことを特徴とする方法。

10

〔 態様 6 〕

態様 1 ないし 5 のうちいずれか一項記載の方法であって、前記構造化されたテキストの修正の前記処理が、ユーザーの校閲に反応して当該テキスト中の修正を実行し、その後：

- ・前記トレーニング・データから抽出された統計モデルを使い、かつ前記実行された修正を参照することによって、当該テキストをテキストセクションに再セグメント分割し、

- ・前記実行された修正を参照することによって当該テキストにラベルをセクション見出しとして挿入することによって構造化されたテキストを再生成し、前記ラベルを前記テキストセクションに付与し、前記構造化されたテキストを校閲のためにユーザーに提示する、ステップを引き起こすことを含むことを特徴とする方法。

20

〔 態様 7 〕

態様 1 ないし 6 のうちいずれか一項記載の方法であって、前記構造化されたテキストの修正の前記処理が、当該テキスト内であるテキスト部分がセクション見出しを記述する定型として識別されたときに、該テキスト部分をラベルによって置き換えることを含むことを特徴とする方法。

〔 態様 8 〕

態様 1 ないし 7 のうちいずれか一項記載の方法であって、前記テキストセグメント分割の粒度が、カスタマイズ可能な粒度パラメータによってユーザーにより制御されることを特徴とする方法。

30

〔 態様 9 〕

態様 1 ないし 8 のうちいずれか一項記載の方法であって、前記統計モデルを適応させるために、前記構造化されたテキストの修正がログに記録され、解析されることを特徴とする方法。

〔 態様 10 〕

テキストをテキストセクションにセグメント分割し、注釈付けされたトレーニング・データに基づいて各テキストセクションにトピックを付与するテキストセグメント分割システムであって、

- ・トレーニング・データから抽出された統計モデルを使うことによって当該テキストをテキストセクションにセグメント分割する手段と、

- ・前記トレーニング・データから抽出された統計モデルを使うことによって各テキストセクションに該テキストセクションの内容を示すトピックを付与する手段であって該トピックがさらにラベルの集合に付与されている手段と、

- ・前記ラベルを前記テキストセクションに付与するために当該テキストに前記ラベルの集合のうちの一つのラベルをセクション見出しとして挿入することによって、構造化されたテキストを生成する手段と、

- ・前記構造化されたテキストをユーザーに提示する手段と、

40

50

・ユーザーの校閲に反応して前記構造化されたテキストの修正を処理する手段、
とを有することを特徴とするシステム。

〔態様 1 1〕

態様 1 0 記載のテキストセグメント分割システムであって、前記構造化されたテキストの修正を処理する手段が、当該テキストのセクションへのセグメント分割の修正ならびに / またはラベルとテキストセクションとの間の対応付けの修正を実行するよう適応されていることを特徴とするシステム。

〔態様 1 2〕

態様 1 0 または 1 1 記載のテキストセグメント分割システムであって、前記構造化されたテキストの修正を処理する手段がさらに：

・あるテキストセクションにラベルを、該テキストセクションに付与されている前記トピックに付与された前記ラベルの集合のうちの一つのラベルを選択することによって付与し、

・当該テキスト中の第一の位置で付与されたラベルを選択し、その付与されたラベルを当該テキスト中の第二の位置に移動することによってセクション境界を再定義し、前記第二の位置がセクション境界を定義し、前記選択されたラベルが前記セクション見出しを定義し、

・ラベルを入力し、該入力されたラベルを前記テキストセクションに付与する、ことを実行するよう適応されていることを特徴とするシステム。

〔態様 1 3〕

態様 1 0 ないし 1 2 のうちいずれか一項記載のシステムであって、前記構造化されたテキストの修正を処理する前記手段が、ユーザーの校閲に反応して当該テキスト中の修正を実行するよう適応されており、さらに：

・前記トレーニング・データから抽出された統計モデルを使い、かつ前記実行された修正を参照することによって、当該テキストをテキストセクションに再セグメント分割し、

・前記実行された修正を参照することによって当該テキストにラベルをセクション見出しとして挿入することによって構造化されたテキストを再生成し、前記ラベルを前記テキストセクションに付与し、前記構造化されたテキストを校閲のためにユーザーに提示する、ステップをその後引き起こす手段を有することを特徴とするシステム。

〔態様 1 4〕

態様 1 0 ないし 1 3 のうちいずれか一項記載のシステムであって、前記構造化されたテキストの実行された修正をログに記録して解析する手段をさらに有しており、該ログに記録して解析する手段が前記統計モデルを適応させるよう適応されていることを特徴とするシステム。

〔態様 1 5〕

テキストをテキストセクションにセグメント分割し、注釈付けされたトレーニング・データに基づいて各テキストセクションにトピックを付与するためのコンピュータプログラムであって、

・トレーニング・データから抽出された統計モデルを使うことによって当該テキストをテキストセクションにセグメント分割し、

・前記トレーニング・データから抽出された統計モデルを使うことによって各テキストセクションに該テキストセクションの内容を示すトピックを付与し、該トピックはさらにラベルの集合に付与されており、

・前記ラベルを前記テキストセクションに付与するために当該テキストに前記ラベルの集合のうちの一つのラベルをセクション見出しとして挿入することによって、構造化されたテキストを生成し、

・前記構造化されたテキストをユーザーに提示し、

・ユーザーの校閲に反応して前記構造化されたテキストの修正を処理する、プログラム手段を有することを特徴とするプログラム。

〔態様 1 6〕

10

20

30

40

50

態様 15 記載のコンピュータプログラムであって、前記構造化されたテキストの修正を処理するプログラム手段が、当該テキストのセクションへのセグメント分割の修正ならびに / またはラベルとテキストセクションとの間の対応付けの修正を実行するよう適応されており、該ラベルとテキストセクションとの間の対応付けの修正のために前記プログラム手段がさらに：

- ・あるテキストセクションにラベルを、該テキストセクションに付与されている前記トピックに付与された前記ラベルの集合のうちの一つのラベルを選択することによって付与し、
 - ・当該テキスト中の第一の位置で付与されたラベルを選択し、その付与されたラベルを当該テキスト中の第二の位置に移動することによってセクション境界を再定義し、前記第二の位置がセクション境界を定義し、前記選択されたラベルが前記セクション見出しを定義し、
 - ・ラベルを入力し、該入力されたラベルを前記テキストセクションに付与する、
- ステップを実行するよう適応されていることを特徴とするプログラム。

〔態様 17〕

態様 15 または 16 記載のコンピュータプログラムであって、前記構造化されたテキストの修正を処理する前記プログラム手段が、ユーザーの校閲に反応して当該テキスト中の修正を実行するよう適応されており、さらに：

- ・前記トレーニング・データから抽出された統計モデルを使い、かつ前記実行された修正を参照することによって、当該テキストをテキストセクションに再セグメント分割し、
 - ・前記実行された修正を参照することによって当該テキストにラベルをセクション見出しとして挿入することで構造化されたテキストを再生成し、前記ラベルを前記テキストセクションに付与し、前記構造化されたテキストを校閲のためにユーザーに提示する、
- ステップをその後引き起こすためのプログラム手段を有することを特徴とするプログラム。

〔態様 18〕

テキストをテキストセクションにセグメント分割し、注釈付けされたトレーニング・データに基づいて各テキストセクションにトピックを付与するためのユーザーインターフェースであって、

- ・トレーニング・データから抽出された統計モデルを使うことにより構造化されたテキストをユーザーに提示する手段と、
 - ・各テキストセクションに付与された各トピックに付与されているラベルの集合をユーザーに提示する手段と、
 - ・ユーザーの校閲に反応して前記構造化されたテキストの修正を処理する入力手段と、
 - ・統計モデルをトレーニングするために前記構造化されたテキストの処理された修正をログに記録して解析する手段、
- とを有することを特徴とするユーザーインターフェース。

〔態様 19〕

態様 18 記載のユーザーインターフェースであって、前記構造化されたテキストがグラフィカルユーザーインターフェースによってユーザーに提示され、前記入力手段が、ユーザーが前記提示されたラベルの集合のうちの一つのラベルを選択してその選択されたラベルがテキストセクションに付与されるという形で、前記構造化されたテキストの修正を処理するよう適応されていることを特徴とするユーザーインターフェース。

〔態様 20〕

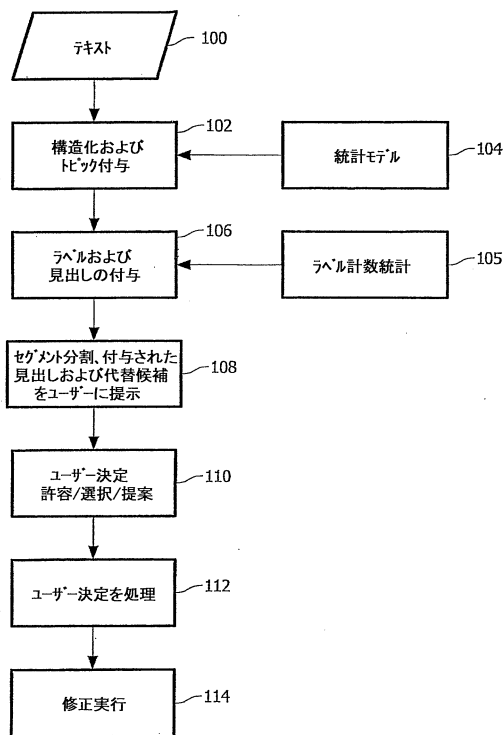
態様 18 または 19 記載のユーザーインターフェースであって、統計モデルを使い、かつ前記処理された修正を参照することによって、ユーザーの校閲に反応して再セグメント分割され、再ラベル付けされたテキストを提供する手段をさらに有することを特徴とするユーザーインターフェース。

【符号の説明】

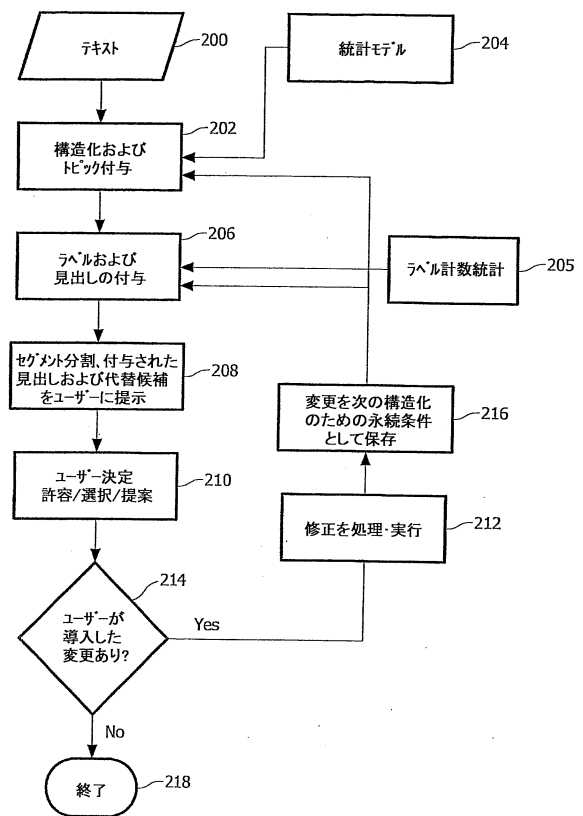
【0054】

- 4 0 0 ユーザーインターフェース
- 4 0 2 テキストウィンドウ
- 4 0 4 提案ウィンドウ
- 4 0 6 ラベル
- 4 0 8 ポインタ
- 4 1 0 ラベルリスト
- 4 1 2 ラベル
- 4 1 4 ラベル
- 4 1 6 ラベル
- 4 1 8 ユーザー入力欄 10
- 5 0 0 セグメント分割システム
- 5 0 2 音声
- 5 0 4 入力装置
- 5 0 6 ユーザー
- 5 0 8 外部記憶装置
- 5 1 0 音声認識モジュール
- 5 1 2 非構造化テキストモジュール
- 5 1 4 統計モデルモジュール
- 5 1 6 処理ユニット
- 5 1 8 構造化テキストモジュール 20
- 5 2 0 グラフィカルユーザーインターフェース

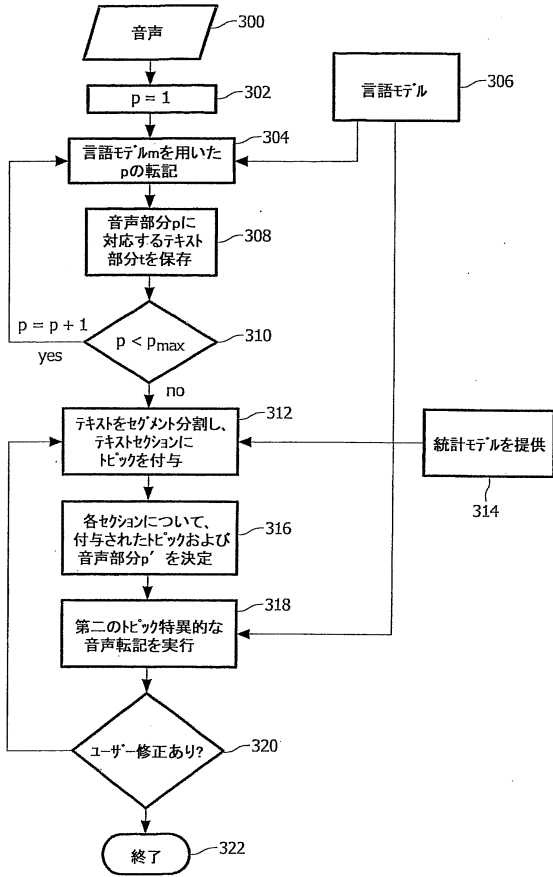
【図 1】



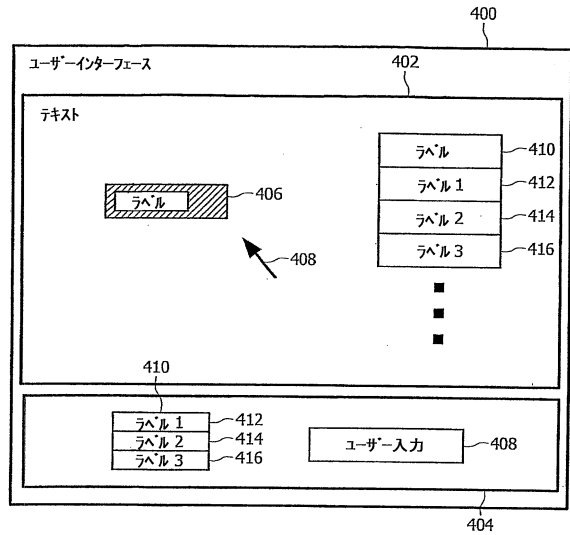
【図 2】



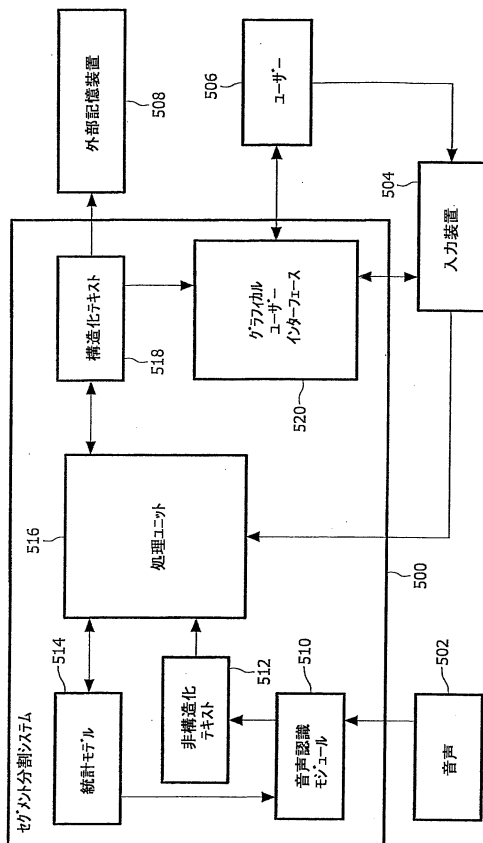
【 図 3 】



【 図 4 】



【 図 5 】



フロントページの続き

- (72)発明者 ペーテルス, ヨーヘン
ドイツ連邦共和国, 5 2 0 6 6 アーヘン, ヴァイスハオスシュトラッセ 2, フィリップス イ
ンテレクチュアル プロパティ アンド スタンダーズ ゲーエムベークーヘン内
- (72)発明者 マツソフ, エフゲニー
ドイツ連邦共和国, 5 2 0 6 6 アーヘン, ヴァイスハオスシュトラッセ 2, フィリップス イ
ンテレクチュアル プロパティ アンド スタンダーズ ゲーエムベークーヘン内
- (72)発明者 マイヤー, カルステン
ドイツ連邦共和国, 5 2 0 6 6 アーヘン, ヴァイスハオスシュトラッセ 2, フィリップス イ
ンテレクチュアル プロパティ アンド スタンダーズ ゲーエムベークーヘン内
- (72)発明者 クラコワ, ディートリッヒ
ドイツ連邦共和国, 5 2 0 6 6 アーヘン, ヴァイスハオスシュトラッセ 2, フィリップス イ
ンテレクチュアル プロパティ アンド スタンダーズ ゲーエムベークーヘン内

審査官 川 崎 博章

- (56)参考文献 特開2003-196296(JP, A)
特開2000-235574(JP, A)
米国特許第05111398(US, A)
成田 えりか, 松本 裕治, 言語処理学会第1回年次大会発表論文集(1995) 論文全文データベース作成のためのSGMLタグ付けの自動化, [online], 日本, 言語処理学会, 1995年 3月31日, [検索日2014.12.9], インターネット, p.329-p.332, URL, http://www.anlp.jp/proceedings/annual_meeting/1995/index.html
八木 玲子, 見えてきた次のMS Office XML使った新たな情報共有が目玉, 日経バイト 第240号, 日本, 日経BP社, 2003年 4月22日, 第240号【ISSN】0289-6508, p.14-p.15

(58)調査した分野(Int.Cl., DB名)

G 0 6 F 1 7 / 2 0 - 1 7 / 2 8