(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5379138号
(P5379138)
(45)発行日　平成25年12月25日(2013.12.25)　(24)登録日　平成25年10月4日(2013.10.4)

(51)Int.Cl.　　F I
G06F 17/27 (2006.01)　　G06F 17/27　Z
G06F 17/28 (2006.01)　　G06F 17/28　C

請求項の数 52　(全 45 頁)

(21)出願番号　特願2010-521289 (P2010-521289)
(86)(22)出願日　平成20年8月25日 (2008.8.25)
(65)公表番号　特表2010-537286 (P2010-537286A)
(43)公表日　平成22年12月2日 (2010.12.2)
(86)国際出願番号　PCT/CN2008/072128
(87)国際公開番号　WO2009/026850
(87)国際公開日　平成21年3月5日 (2009.3.5)
審査請求日　平成23年8月25日 (2011.8.25)
(31)優先権主張番号　11/844,067
(32)優先日　平成19年8月23日 (2007.8.23)
(33)優先権主張国　米国 (US)
(31)優先権主張番号　11/844,153
(32)優先日　平成19年8月23日 (2007.8.23)
(33)優先権主張国　米国 (US)

(73)特許権者　507103802
グーグル・インコーポレーテッド
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ヴュー・アンフィシアター・パークウェイ・１６００
(74)代理人　100108453
弁理士　村山　靖彦
(74)代理人　100064908
弁理士　志賀　正武
(74)代理人　100089037
弁理士　渡邊　隆
(74)代理人　100110364
弁理士　実広　信哉

最終頁に続く

(54)【発明の名称】領域辞書の作成

(57)【特許請求の範囲】
【請求項１】
トピック文書コーパスにおける第1のトピック語分布の、文書コーパスにおける第2のトピック語分布に対する比に比例するトピック相違値を算出するステップと、
前記トピック文書コーパスにおける候補トピック語の第1の分布の、前記文書コーパスにおける前記候補トピック語の第2の分布に対する比に比例する該候補トピック語に関する候補トピック語相違値を算出するステップと、
前記候補トピック語相違値および前記トピック相違値に基づいて、前記候補トピック語が前記トピックのための新たなトピック語であるかどうかを判定するステップと、
を備え、
前記トピック文書コーパスは、或るトピックと関係するトピック文書のコーパスであり、
前記文書コーパスは、前記トピック文書および他の文書を含む文書のコーパスであることを特徴とするコンピュータによって実施され、
前記候補トピック語は、前記トピックのためのトピック辞書に存在しない語であるとともに、前記トピックのための新たなトピック語として識別され前記トピック辞書への格納対象の語となるための候補であることを特徴とする方法。
【請求項２】
前記トピックのための前記トピック辞書の中の既存の語を、前記トピック相違値が算出されるトピック語として選択するステップをさらに備えることを特徴とする請求項1に記

載の方法。
【請求項３】
トピック相違値を算出するステップは、
前記トピックのためのトピック語を選択するステップと、
前記トピック文書コーパスにおける各トピック語の第1の分布の、前記文書コーパスにおける各トピック語の第2の分布に対する比に比例する前記トピック語のそれぞれに関するトピック語相違値を算出するステップと、
前記トピック語相違値の中心傾向に基づいて、前記トピック相違値を算出するステップと、
を備えることを特徴とする請求項1に記載の方法。
【請求項４】
前記トピック文書コーパスにおける前記候補トピック語の前記第1の分布は、前記トピック文書コーパスにおける前記候補トピック語の分布の、前記分布の対数に基づく値に対する比に比例することを特徴とする請求項1に記載の方法。
【請求項５】
前記候補トピック語が新たなトピック語であるかどうかを判定するステップは、前記候補トピック語相違値が前記トピック相違値より大きい場合、前記候補トピック語が新たなトピック語であると判定するステップを備えることを特徴とする請求項1に記載の方法。
【請求項６】
前記候補トピック語が、新たなトピック語であると判定された場合、前記候補トピック語を前記トピック辞書の中に格納するステップをさらに備えることを特徴とする請求項1に記載の方法。
【請求項７】
前記文書コーパスの中でトピックと関係する文書を識別するステップと、
前記トピックと関係する文書クラスタを生成するステップと、
前記文書クラスタのそれぞれの中の語を識別するステップと、
前記文書クラスタのそれぞれの中の前記識別された語から候補トピック語を選択するステップと、
をさらに備えることを特徴とする請求項1に記載の方法。
【請求項８】
前記文書コーパスの第1のサブセットを備える訓練コーパスにおける既存の語、およびそれぞれが辞書の中の既存の語である構成要素語の系列によって定義される候補語に関する第1の語頻度を算出するステップと、
前記文書コーパスの第2のサブセットを備える開発コーパスにおける前記構成要素語および前記候補語に関する第2の語頻度を算出するステップと、
前記候補語の前記第2の語頻度、および前記構成要素語および前記候補語の前記第1の語頻度に基づいて、候補語エントロピー測度を算出するステップと、
前記構成要素語の前記第2の語頻度、および前記構成要素語および前記候補語の前記第1の語頻度に基づいて、既存語エントロピー測度を算出するステップと、
前記候補語エントロピー測度が前記既存語エントロピー測度を超えている場合、前記候補語が候補トピック語であると判定するステップと、
をさらに備えることを特徴とする請求項1に記載の方法。
【請求項９】
訓練コーパスにおける既存の語および候補語に関する第1の語頻度を算出するステップは、前記訓練コーパスにおける前記既存の語および前記候補語の確率に関する言語モデルを訓練するステップを備え、
開発コーパスにおける前記構成要素語および前記候補語に関する第2の語頻度を算出するステップは、前記開発コーパスにおける前記構成要素語および前記候補語のそれぞれに関する語カウント値を算出するステップを備えることを特徴とする請求項8に記載の方法。

【請求項１０】
前記候補語の前記第2の語頻度、および前記構成要素語および前記候補語の前記第1の語頻度に基づいて、候補語エントロピー測度を算出するステップは、
前記候補語および前記構成要素語の前記確率に基づいて、第1の対数値を算出するステップと、
前記候補語の前記語カウント値、および前記第1の対数値に基づいて、前記候補語エントロピー測度を算出するステップと、
を備え、
前記構成要素語の前記第2の語頻度、および前記構成要素語および前記候補語の前記第1の語頻度に基づいて、既存語エントロピー測度を算出するステップは、
前記候補語および前記構成要素語の前記確率に基づいて、第2の対数値を算出するステップと、
前記構成要素語の前記語カウント値、および前記第2の対数値に基づいて、前記既存語エントロピー測度を算出するステップと、
を備えることを特徴とする請求項9に記載の方法。
【請求項１１】
前記候補トピック語は、1つまたは複数のHanzi文字を備えることを特徴とする請求項1に記載の方法。
【請求項１２】
或るトピックと関係するトピック語を備えるトピック辞書を選択するステップと、
トピック語、文書コーパス、およびトピック文書コーパスに基づいて、トピック語相違値を算出するステップと、
前記文書コーパスおよび前記トピック文書コーパスに基づいて、候補トピック語に関する候補トピック語相違値を算出するステップと、
前記候補トピック語相違値および前記トピック語相違値に基づいて、前記候補トピック語が前記トピックのための新たなトピック語であるかどうかを判定するステップと、
を備え、
前記トピック語は、前記或るトピックと関係し、
前記文書コーパスは、トピック文書および他の文書を含む文書のコーパスであり、
前記トピック文書コーパスは、前記トピックと関係するトピック文書のコーパスであり、
前記候補トピック語は、前記トピックのためのトピック辞書に存在しない語であるとともに、前記トピックのための新たなトピック語として識別され前記トピック辞書への格納対象の語となるための候補であることを特徴とするコンピュータによって実施される方法。
【請求項１３】
前記候補トピック語が、新たなトピック語であると判定された場合、前記候補トピック語を前記トピック辞書の中に格納するステップをさらに備えることを特徴とする請求項12に記載の方法。
【請求項１４】
トピック語相違値を算出するステップは、
前記トピック辞書の中の既存のトピック語を選択するステップと、
前記文書コーパスおよび前記トピック文書コーパスに基づいて、前記トピック語のそれぞれに関する既存トピック語相違値を算出するステップと、
前記既存トピック語相違値の中心傾向に基づいて、前記トピック語相違値を算出するステップと、
を備えることを特徴とする請求項12に記載の方法。
【請求項１５】
前記文書コーパスおよび前記トピック文書コーパスに基づいて、前記候補トピック語に関する候補トピック語相違値を算出するステップは、

前記トピック文書コーパスにおける前記候補トピック語に関連する第1の確率を算出するステップと、
前記文書コーパスにおける前記候補トピック語に関連する第2の確率を算出するステップと、
前記第1の確率の、前記第2の確率と、前記第1の確率に基づく対数値との積に対する比に基づいて、前記候補トピック語相違値を計算するステップと、
を備えることを特徴とする請求項12に記載の方法。
【請求項１６】
前記候補トピック語は、1つまたは複数のHanzi文字を備えることを特徴とする請求項12に記載の方法。
【請求項１７】
一時的でないコンピュータ可読媒体の中に格納されたソフトウェアを備える装置であって、
前記ソフトウェアは、コンピュータ可読命令を備え、
前記コンピュータ可読命令は、コンピュータ処理デバイスによって実行可能であり、さらにそのような実行時に、前記コンピュータ処理デバイスに、
トピック語、文書コーパス、およびトピック文書コーパスに基づいて、トピック語相違値を算出させ、
前記文書コーパスおよび前記トピック文書コーパスに基づいて、候補トピック語に関する候補トピック語相違値を算出させ、
前記候補トピック語相違値および前記トピック語相違値に基づいて、前記候補トピック語が前記トピックのためのトピック語であるかどうかを判定させ、さらに
前記候補トピック語が、トピック語であると判定された場合、前記候補トピック語をトピック辞書の中に格納させ、
前記トピック語は、前記トピックと関係するトピック辞書の中の語であり、
前記文書コーパスは、前記トピック文書および他の文書を含む文書のコーパスであり、
前記トピック文書コーパスは、トピックと関係するトピック文書のコーパスであり、
前記候補トピック語は、前記トピックのためのトピック辞書に存在しない語であるとともに、前記トピックのための新たなトピック語として識別され前記トピック辞書への格納対象の語となるための候補であることを特徴とする装置。
【請求項１８】
データストアと、
トピック語処理モジュールと、
辞書アップデータモジュールと、
を備え、
前記データストアは、或るトピックと関係するトピック語を備えるトピック辞書を格納し、
前記トピック語処理モジュールは、
或るトピックと関係するトピック辞書の中の語であるトピック語、トピック文書および他の文書を含む文書のコーパスである文書コーパス、および該トピックと関係する該トピック文書のコーパスであるトピック文書コーパスに基づいて、トピック語相違値を算出し、
前記トピック辞書の中のトピック語のための候補として候補トピック語を選択し、
前記文書コーパスおよび前記トピック文書コーパスに基づいて、前記候補トピック語に関する候補トピック語相違値を算出し、さらに
前記候補トピック語相違値および前記トピック語相違値に基づいて、前記候補トピック語が前記トピックのためのトピック語であるかどうかを判定するように構成されており、
前記辞書アップデータモジュールは、前記候補トピック語が、トピック語であると判定された場合、前記候補トピック語を前記トピック辞書の中に格納するように構成されていることを特徴とするシステム。

【請求項１９】
前記トピック語処理モジュールは、
前記トピック文書コーパスにおける前記候補トピック語に関連する第1の確率を算出し、
前記文書コーパスにおける前記候補トピック語に関連する第2の確率を算出し、さらに
前記第1の確率の、前記第2の確率と、前記第1の確率に基づく対数値との積に対する比に基づいて、前記候補トピック語相違値を計算するように構成されていることを特徴とする請求項18に記載のシステム。
【請求項２０】
トピック文書コーパスに関する相違閾値を算出するステップと、
候補語に関する候補語相違値を算出するステップと、
前記候補語相違値が、前記相違閾値を超えている場合、前記候補語が前記トピックに関するトピック語であると判定するステップと、
を備えており、
前記相違閾値は、トピック文書コーパスにおけるトピック語に関する第1のトピック語確率の、前記文書コーパスにおける前記トピック語に関する第2のトピック語確率に対する比に比例し、
前記トピック文書コーパスは、或るトピックと関係するトピック文書のコーパスであり、
前記トピック語は、前記トピックと関係するトピック辞書の中の語であり、
前記文書コーパスは、前記トピック文書および他の文書を含む文書のコーパスであり、
前記候補語相違値は、前記トピック文書コーパスに関連する候補語に関する第1の候補語確率の、前記文書コーパスに関連する前記候補語に関する第2の候補語確率に対する比に比例することを特徴とする方法。
【請求項２１】
トピック相違値を算出するための手段と、
候補トピック語に関する候補トピック語相違値を算出するための手段と、
前記候補トピック語相違値および前記トピック相違値に基づいて、前記候補トピック語が前記トピックのための新たなトピック語であるかどうかを判定するための手段と、
を備え、
前記トピック相違値は、トピック文書コーパスにおける第1のトピック語分布の、文書コーパスにおける第2のトピック語分布に対する比に比例し、
前記トピック文書コーパスは、或るトピックと関係するトピック文書のコーパスであり、
前記文書コーパスは、前記トピック文書および他の文書を含む文書のコーパスであり、
前記候補トピック語相違値は、前記トピック文書コーパスにおける候補トピック語の第1の分布の、前記文書コーパスにおける前記候補トピック語の第2の分布に対する比に比例し、
前記候補トピック語は、前記トピックのためのトピック辞書に存在しない語であるとともに、前記トピックのための新たなトピック語として識別され前記トピック辞書への格納対象の語となるための候補であることを特徴とするシステム。
【請求項２２】
或るトピックと関係するトピック語を備えるトピック辞書を選択するための手段と、
トピック語、文書コーパス、およびトピック文書コーパスに基づいて、トピック語相違値を算出するための手段と、
前記文書コーパスおよび前記トピック文書コーパスに基づいて、候補トピック語に関する候補トピック語相違値を算出するための手段と、
前記候補トピック語相違値および前記トピック語相違値に基づいて、前記候補トピック語が前記トピックのための新たなトピック語であるかどうかを判定するための手段と、
を備え、

前記トピック語は、前記トピック辞書の中の語であり、
前記文書コーパスは、トピック文書および他の文書を含む文書のコーパスであり、
前記トピック文書コーパスは、前記トピックと関係する該トピック文書のコーパスであり、
前記候補トピック語は、前記トピックのためのトピック辞書に存在しない語であるとともに、前記トピックのための新たなトピック語として識別され前記トピック辞書への格納対象の語となるための候補であることを特徴とするシステム。
【請求項２３】
トピック語、文書コーパス、およびトピック文書コーパスに基づいて、トピック語相違値を算出するための手段と、
前記文書コーパスおよび前記トピック文書コーパスに基づいて、候補トピック語に関する候補トピック語相違値を算出するための手段と、
前記候補トピック語相違値および前記トピック語相違値に基づいて、前記候補トピック語がトピック語であるかどうかを判定するための手段と、
前記候補トピック語が、トピック語であると判定された場合、前記候補トピック語を前記トピック辞書の中に格納するための手段と、
を備え、
前記トピック語は、或るトピックと関係するトピック辞書の中の語であり、
前記文書コーパスは、トピック文書および他の文書を含む文書のコーパスであり、
前記トピック文書コーパスは、前記トピックと関係する該トピック文書のコーパスであり、
前記候補トピック語は、前記トピックのためのトピック辞書に存在しない語であるとともに、前記トピックのための新たなトピック語として識別され前記トピック辞書への格納対象の語となるのための候補であることを特徴とするコンピュータ処理デバイス。
【請求項２４】
トピック文書コーパスに関する相違閾値を算出するための手段と、
候補語に関する候補語相違値を算出するための手段と、
前記候補語相違値が前記相違閾値を超えている場合、前記候補語が前記トピックに関するトピック語であると判定するための手段と、
を備え、
前記相違閾値は、トピック語に関する第1のトピック語確率の、文書コーパスにおける前記トピック語に関する第2のトピック語確率に対する比に比例し、
前記トピック文書コーパスは、或るトピックと関係するトピック文書のコーパスであり、
前記トピック語は、前記トピックと関係するトピック辞書の中の語であり、
前記文書コーパスは、前記トピック文書および他の文書を含む文書のコーパスであり、
前記候補語相違値は、前記トピック文書コーパスに関連する候補語に関する第1の候補語確率の、前記文書コーパスに関連する前記候補語に関する第2の候補語確率に対する比に比例することを特徴とするシステム。
【請求項２５】
訓練コーパスにおける既存の語、およびそれぞれが辞書の中の既存の語である構成要素語の系列によって定義される候補語に関する第1の語頻度を算出するステップと、
開発コーパスにおける前記構成要素語および前記候補語に関する第2の語頻度を算出するステップと、
前記候補語の前記第2の語頻度、および前記構成要素語および前記候補語の前記第1の語頻度に基づいて、候補語エントロピー関連測度を算出するステップと、
前記構成要素語の前記第2の語頻度、および前記構成要素語および前記候補語の前記第1の語頻度に基づいて、既存語エントロピー関連測度を算出するステップと、
前記候補語エントロピー関連測度が前記既存語エントロピー関連測度を超えている場合、前記候補語が新たな語であると判定するステップと、

を備えることを特徴とするコンピュータによって実施される方法。
【請求項２６】
前記訓練コーパスおよび前記開発コーパスは、ウェブ文書を備えることを特徴とする請求項25に記載の方法。
【請求項２７】
前記候補語が、新たな語であると判定された場合、前記候補語を既存の語の辞書に追加するステップをさらに備えることを特徴とする請求項25に記載の方法。
【請求項２８】
第1の語頻度を算出するステップは、前記訓練コーパスにおける前記既存の語および前記候補語の確率に関する言語モデルを訓練するステップを備え、
第2の語頻度を算出するステップは、前記開発コーパスにおける前記構成要素語および前記候補語のそれぞれに関する語カウント値を算出するステップを備えることを特徴とする請求項25に記載の方法。
【請求項２９】
候補語エントロピー関連測度を算出するステップは、
前記候補語および前記構成要素語の前記確率に基づいて、第1の対数値を算出するステップと、
前記候補語の前記語カウント値、および前記第1の対数値に基づいて、前記候補語エントロピー関連測度を算出するステップと、
を備え、さらに
既存語エントロピー関連測度を算出するステップは、
前記候補語および前記構成要素語の前記確率に基づいて、第2の対数値を算出するステップと、
前記構成要素語の前記語カウント、および前記第2の対数値に基づいて、前記既存語エントロピー関連測度を算出するステップと、
を備えることを特徴とする請求項25に記載の方法。
【請求項３０】
前記語はそれぞれ、1つまたは複数のHanzi文字を備えることを特徴とする請求項25に記載の方法。
【請求項３１】
前記語はそれぞれ、1つまたは複数の表語文字を備えることを特徴とする請求項25に記載の方法。
【請求項３２】
前記候補語が、新たな語であると判定された場合、前記辞書を前記候補語で更新するステップをさらに備えることを特徴とする請求項25に記載の方法。
【請求項３３】
第1のコーパスにおける既存の語、およびそれぞれが辞書の中の既存の語である構成要素語の系列によって定義される候補語に関する第1の語確率を算出するステップと、
第2のコーパスにおける前記構成要素語および前記候補語に関する第2の語確率を算出するステップと、
前記第2の候補語確率、および前記候補語のおよび前記構成要素語の前記第1の語確率に基づいて、第1のエントロピー関連値を算出するステップと、
前記第2の構成要素語確率、および前記候補語および前記構成要素語の前記第1の語確率に基づいて、第2のエントロピー関連値を算出するステップと、
前記第1のエントロピー関連値が前記第2のエントロピー関連値を超えている場合、前記候補語が新たな語であると判定するステップと、
を備えることを特徴とするコンピュータによって実施される方法。
【請求項３４】
語コーパスを識別するステップは、ウェブ文書を識別するステップを備えることを特徴とする請求項33に記載の方法。

【請求項３５】
第1の語確率を算出するステップは、前記第1のコーパスにおける前記既存の語および前記候補語の語確率に関して前記第1のコーパス上で言語モデルを訓練するステップを備え、さらに
第2の語確率を算出するステップは、前記構成要素語および候補語のそれぞれに関して語カウント値を算出するステップを備えることを特徴とする請求項33に記載の方法。
【請求項３６】
第1のエントロピー関連値を算出するステップは、
前記候補語および前記構成要素語の前記第1の語確率に基づいて、第1の対数値を算出するステップと、
前記候補語の前記語カウント値、および前記第1の対数値に基づいて、前記第1のエントロピー関連値を算出するステップと、
を備え、
第2のエントロピー関連値を算出するステップは、
前記候補語および前記構成要素語の前記第１の語確率に基づいて、第2の対数値を算出するステップと、
前記構成要素語の前記語カウント、および前記第2の対数値に基づいて、前記第2のエントロピー関連値を算出するステップと、
を備えることを特徴とする請求項35に記載の方法。
【請求項３７】
前記語はそれぞれ、1つまたは複数のHanzi文字を備えることを特徴とする請求項33に記載の方法。
【請求項３８】
ウェブ文書のコレクションを訓練コーパスと開発コーパスに分割するステップと、
前記訓練コーパスにおける語の第1の語確率に関して前記訓練コーパス上で言語モデルを訓練するステップと、
前記開発コーパスにおける前記候補語および前記2つ以上の対応する語の出現回数をカウントするステップと、
前記開発コーパスにおける前記候補語の前記出現回数、および前記第1の語確率に基づいて、第1の値を算出するステップと、
前記開発コーパスにおける前記2つ以上の対応する語の前記出現回数、および前記第1の語確率に基づいて、第2の値を算出するステップと、
前記第1の値を前記第2の値と比較するステップと、
前記比較に基づいて、前記候補語が新たな語であるかどうかを判定するステップと、
を備え、
前記訓練コーパスにおける語は、辞書の中の既存の語である前記訓練コーパスの中の2つ以上の対応する語の系列によって定義される候補語を含むことを特徴とするコンピュータによって実施される方法。
【請求項３９】
前記候補語が新たな語であると判定された場合、前記候補語を前記辞書に追加するステップをさらに備えることを特徴とする請求項38に記載の方法。
【請求項４０】
前記訓練コーパスにおける語の第1の語確率に関して前記訓練コーパス上で言語モデルを訓練するステップは、nグラム言語モデルを訓練するステップを備えることを特徴とする請求項38に記載の方法。
【請求項４１】
前記開発コーパスにおける前記候補語の前記出現回数、および前記第1の語確率に基づいて、第1の値を算出するステップは、
前記候補語に関する前記第1の語確率、および前記2つ以上の対応する語の前記第1の語確率に基づいて、第1の対数値を算出するステップと、

前記第1の対数値に前記候補語の前記カウントされた出現回数を掛けるステップと、
を備え、さらに
前記開発コーパスにおける前記2つ以上の対応する語、および前記第1の語確率に基づいて、第2の値を算出するステップは、
前記候補語の前記第1の語確率、および前記2つ以上の対応する語の前記第1の語確率に基づいて、第2の対数値を算出するステップと、
前記第2の対数値に前記2つ以上の対応する語の前記カウントされた出現回数を掛けるステップと、
を備えることを特徴とする請求項40に記載の方法。
【請求項４２】
前記語はそれぞれ、1つまたは複数のHanzi文字を備えることを特徴とする請求項41に記載の方法。
【請求項４３】
コンピュータ可読媒体の中に格納されているコンピュータ命令を備え、該コンピュータ命令がコンピュータデバイスによって実行されると、語コーパスにアクセスして、該語コーパスを訓練コーパスと開発コーパスに分割し、さらに、
2つ以上の対応する語を備える候補語を含む、前記訓練コーパスの中に格納された語に関する第1の語確率と、
前記開発コーパスにおける前記語に関する第2の語確率と、
を生成するように構成されている語処理モジュールと、
コンピュータ可読媒体の中に格納されているコンピュータ命令を備え、該コンピュータ命令がコンピュータデバイスによって実行されると、前記第1の語確率、および前記第2の語確率を処理し、さらに、
前記候補語および前記2つ以上の対応する語に関する前記第1の語確率、および前記候補語に関する前記第2の語確率に基づく第1の値、および、
前記候補語および前記2つ以上の対応する語に関する前記第1の語確率、および前記2つ以上の対応する語に関する前記第2の語確率に基づく第2の値と、
を生成するように構成されている新語アナライザモジュールと、
を具備し、
前記第1の値を前記第2の値と比較し、前記比較に基づいて、前記候補語が新たな語であるかどうかを判定するようにさらに構成されていることを特徴とするシステム。
【請求項４４】
コンピュータ可読媒体の中に格納され、さらにコンピュータデバイスによって実行されると、辞書を識別された新たな語で更新するように構成されるコンピュータ命令を備える辞書アップデータモジュールをさらに備えていることを特徴とする請求項43に記載のシステム。
【請求項４５】
前記語処理モジュールは、nグラム言語モデルを備えていることを特徴とする請求項43に記載のシステム。
【請求項４６】
前記第1の値および前記第2の値は、エントロピー関連値であることを特徴とする請求項43に記載のシステム。
【請求項４７】
前記語コーパスは、ウェブ文書を備えていることを特徴とする請求項44に記載のシステム。
【請求項４８】
前記語処理モジュールは、Hanzi文字処理モジュールを備えていることを特徴とする請求項43に記載のシステム。
【請求項４９】
各語は、1つまたは複数のHanzi文字を備えていることを特徴とする請求項48に記載のシ

ステム。
【請求項５０】
コンピュータ可読媒体の中に格納されたソフトウェアを備える装置であって、
前記ソフトウェアは、コンピュータ処理デバイスによって実行可能であるコンピュータ可読命令を有し、
該コンピュータ可読命令が実行されると、前記コンピュータ処理デバイスに、
訓練コーパスにおける既存の語、およびそれぞれが辞書の中の既存の語である構成要素語の系列によって定義される候補語に関する第1の語頻度を算出させ、
開発コーパスにおける前記構成要素語および前記候補語に関する第2の語頻度を算出させ、
前記候補語の前記第2の語頻度、および前記構成要素語および前記候補語の前記第1の語頻度に基づいて、候補語エントロピー関連測度を算出させ、
前記構成要素語の前記第2の語頻度、および前記構成要素語および前記候補語の前記第1の語頻度に基づいて、既存語エントロピー関連測度を算出させ、さらに
前記候補語エントロピー関連測度が前記既存語エントロピー関連測度を超えている場合、前記候補語が新たな語であると判定させることを特徴とする装置。
【請求項５１】
第1のコーパスにおける既存の語、およびそれぞれが辞書の中の既存の語である構成要素語によって定義される候補語に関する第1の語確率を算出するための手段と、
第2のコーパスにおける前記構成要素語および前記候補語に関する第2の語確率を算出するための手段と、
前記候補語の前記第2の語確率、ならびに前記候補語および前記構成要素語の前記第1の語確率に基づいて、第1のエントロピー関連値を算出するための手段と、
前記構成要素語の前記第2の語確率、ならびに前記候補語および前記構成要素語の前記第1の語確率に基づいて、第2のエントロピー関連値を算出するための手段と、
前記第1のエントロピー関連値と前記第2のエントロピー関連値の間の比較に基づいて、候補語が新たな語であるかどうかを判定するための手段と、
を備えることを特徴とするシステム。
【請求項５２】
語コーパスにアクセスして、さらに該語コーパスを訓練コーパスと開発コーパスに分割するとともに、
2つ以上の対応する語を備える候補語を含む、前記訓練コーパスの中に格納された語に関する第1の語確率と、
前記開発コーパスにおける前記語に関する第2の語確率と、
を生成するように構成されている語処理手段と、
前記第1の語確率および前記第2の語確率を受け取るとともに、
前記候補語および前記2つ以上の対応する語に関する前記第1の語確率、および前記候補語に関する前記第2の語確率に基づく第1の値と、
前記候補語および前記2つ以上の対応する語に関する前記第1の語確率、および前記2つ以上の対応する語に関する前記第2の語確率に基づく第2の値と、
を生成するように構成されている新語アナライザ手段と、
を備え、
前記第1の値と前記第2の値を比較して、前記比較に基づいて、前記候補語が新たな語であるかどうかを判定するようにさらに構成されていることを特徴とするシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、機械翻訳、非ローマ字言語の語のセグメント化、音声認識、およびインプットメソッドエディタなどの自然言語処理アプリケーションのための辞書に関する。
【０００２】

本出願は、ともに2007年8月23日に出願した米国特許出願第11/844,067号および米国特許出願第11/844,153号の優先権を主張する。これらの先出願の開示は、本出願の開示の一部と考えられる(さらに参照により本出願の開示に組み込まれている)。

【背景技術】

【０００３】

ますます進んだ自然言語処理技術が、音声処理システム、手書き/光学文字認識システム、自動翻訳システムなどのデータ処理システムにおいて、またはワードプロセッシングシステムにおけるスペル/文法検査のために使用されている。これらの自然言語処理技術は、例えば、非ローマ字言語の語のセグメント化、機械翻訳、自動校正、音声認識、インプットメソッドエディタなどと関係する自然言語アプリケーションのための辞書の自動更新を含むことが可能である。

【０００４】

1つまたは2つの文字、例えば、象形文字が、1つの語もしくは意味に対応する表語文字書記体系を使用する非ローマ字言語は、移動デバイスキーパッド上のコンピュータキーボードなどの標準入力デバイス上のキーより多くの文字を有する。例えば、中国語は、基本音声、つまりピンインと5つの声調によって定義される数万の表意文字を含む。これら多くを1つの関連付けにマップすることは、入力デバイス上で見つからない文字および記号の入力を円滑にするインプットメソッドによって実施されることが可能である。したがって、西洋スタイルのキーボードを使用して、中国語の文字、日本語の文字、または朝鮮語の文字が入力されることが可能である。

【０００５】

インプットメソッドエディタが、インプットメソッドを実現するのに使用されることが可能である。そのようなインプットメソッドエディタは、語および/または句の辞書を含む、またはそのような辞書にアクセスすることが可能である。しかし、言語の語彙は、常に進化しており、このため、インプットメソッドエディタのための辞書は、頻繁な更新を要求する可能性がある。例えば、新たな語が、或る言語に急速に導入されることが可能であり、例えば、ポップカルチャーリファレンス、または或る商品に関する新たな商品名が、或る語彙に導入されることが可能である。このため、インプットメソッドエディタ辞書を適時に更新するのを怠ることにより、ユーザが、その新たな語を入力フィールドに入力するのにインプットメソッドエディタを利用することができない、またはそのように利用することに苦労する可能性があるので、ユーザ体験が低下する可能性がある。例えば、ユーザが、新たな語、例えば、新たな商品名を、検索エンジンに検索クエリとしてサブミットすることを所望することが可能である。しかし、インプットメソッドエディタが、その新たな語を認識しない場合、ユーザは、検索エンジンにその新たな語を入力することに困難を経験する可能性がある。

【０００６】

中国語、日本語、タイ語、および朝鮮語などの一部の言語において、文の中で語の境界は、全く存在しない。したがって、新たな語は、それらの新たな語が、文字、または既存の語の複合した連続であるので、テキストの中で容易に識別され得ない。このことは、それらの言語に関して、新たな語の検出を困難な作業にする。したがって、新たな語が識別されると、それらの新たな語、および他の既存の語が関係するトピックを識別することが、望ましい。そのようなトピックの識別は、文の中に境界を有さない言語、または他の言語に関して、言語モデル、および/またはその言語モデルを使用するシステムもしくはデバイスのパフォーマンスを向上させることが可能である。

【発明の概要】

【発明が解決しようとする課題】

【０００７】

本明細書で開示されるのは、自動的にトピック領域を識別するため、およびそれらのトピック領域と関係する領域辞書を作成するための方法、システム、および装置である。

【課題を解決するための手段】

【０００８】
或る実施形態において、方法は、トピック文書コーパスにおける第1のトピック語分布の、文書コーパスにおける第2のトピック語分布に対する比に実質的に比例するトピック相違値を算出することを含む。トピック文書コーパスは、トピックと関係するトピック文書のコーパスであり、さらに文書コーパスは、それらのトピック文書、およびその他の文書を含む文書のコーパスである。また、この方法は、候補トピック語に関する候補トピック語相違値を算出することも含む。候補トピック語相違値は、トピック文書コーパスにおける候補トピック語の第1の分布の、文書コーパスにおける候補トピック語の第2の分布に対する比に実質的に比例する。この方法は、候補トピック語相違値およびトピック相違値に基づいて、候補トピック語が新たなトピック語であるかどうかを判定する。

【０００９】
別の実施形態において、方法は、トピックと関係するトピック語を備えるトピック辞書を選択すること、ならびにトピック語、文書コーパス、およびトピック文書コーパスに基づいて、トピック語相違値を算出することを含む。トピック文書コーパスは、トピックと関係するトピック文書のコーパスであり、さらにこの文書コーパスは、それらのトピック文書、およびその他の文書を含む文書のコーパスである。トピック語は、トピックと関係する語である。また、この方法は、文書コーパスおよびトピック文書コーパスに基づいて、候補トピック語に関する候補トピック語相違値を算出すること、および候補トピック語相違値およびトピック語相違値に基づいて、候補トピック語が新たなトピック語であるかどうかを判定することも含む。

【００１０】
別の実施形態において、システムは、データストア、トピック語処理モジュール、および辞書アップデータモジュールを含む。データストアは、トピックと関係するトピック語を備えるトピック辞書を格納する。トピック語処理モジュールは、トピック語、文書コーパス、およびトピック文書コーパスに基づいて、トピック語相違値を算出するように構成される。トピック文書コーパスは、トピックと関係するトピック文書のコーパスであり、さらに文書コーパスは、それらのトピック文書、およびその他の文書を含む文書のコーパスである。トピック語は、トピックと関係するトピック辞書の中の語である。また、トピック語処理モジュールは、候補トピック語を選択し、さらに文書コーパスおよびトピック文書コーパスに基づいて、その候補トピック語に関する候補トピック語相違値を算出し、さらに候補トピック語相違値およびトピック語相違値に基づいて、候補トピック語がトピック語であるかどうかを判定するようにも構成される。辞書アップデータモジュールは、候補トピック語がトピック語であると判定された場合、トピック辞書の中にその候補トピック語を格納するように構成される。

【００１１】
本開示において提供される方法、システム、および装置によれば、言語モデル、例えば、文の中に境界を有さない言語に関する言語モデルを使用するシステムのデータ処理パフォーマンスを向上させることが可能である。例えば、そのシステムまたはデバイスは、自動的に更新されるトピック辞書の使用によって、音声処理、手書き/光学文字認識、自動翻訳、自動分類、自動抽象化、および/またはワードプロセッシングシステムにおけるスペル/文法検査において向上したパフォーマンスを有することが可能である。

【００１２】
本明細書で説明される主題の1つまたは複数の実施形態の詳細は、添付の図面、および後段の説明において示される。主題のその他の特徴、態様、および利点は、その説明、それらの図面、および特許請求の範囲から明白となる。

【図面の簡単な説明】

【００１３】
【図１Ａ】インプットメソッドエディタを実施するのに利用されることが可能である例示的なデバイス100を示すブロック図である。
【図１Ｂ】例示的なインプットメソッドエディタシステム120を示すブロック図である。

【図２Ａ】例示的な語検出システムを示すブロック図である。
【図２Ｂ】図2Aのシステムの例示的な実施形態を示すブロック図である。
【図３】語コーパスにおける新たな語を識別するための例示的なプロセスを示す流れ図である。
【図４】候補語および既存の語に関するエントロピー関連測度を算出するための例示的なプロセスを示す流れ図である。
【図５】語コーパスにおける新たな語を識別するための別の例示的なプロセスを示す流れ図である。
【図６】或る語コーパスにおける新たな語を、別の語コーパスからの語確率に基づいて、識別するための別の例示的なプロセスを示す流れ図である。
【図７Ａ】例示的なトピック語識別システムを示すブロック図である。
【図７Ｂ】図7Aのシステムのより詳細なブロック図である。
【図８】トピック語を識別するための例示的なプロセスを示す流れ図である。
【図９】トピック語相違値を算出するための例示的なプロセスを示す流れ図である。
【図１０】例示的な文書/語クラスタリングプロセスを示す流れ図である。
【図１１】トピック語を識別するための別の例示的なプロセスを示す流れ図である。
【発明を実施するための形態】
【００１４】
これらの様々な図面における同様の符号、および同様の名称は、同様の要素を示す。
【００１５】
図1Aは、IME(インプットメソッドエディタ)を実施するのに利用されることが可能である例示的なデバイス100のブロック図である。デバイス100は、例えば、パーソナルコンピュータデバイス、ネットワークサーバ、遠隔通信スイッチなどのコンピュータデバイス、あるいは移動電話機、移動通信デバイス、PDA(パーソナルデジタルアシスタント)、ゲームボックスなどの他の電子デバイスにおいて実施されることが可能である。
【００１６】
例示的なデバイス100は、処理デバイス102、第1のデータストア104、第2のデータストア106、入力デバイス108、出力デバイス110、およびネットワークインタフェース112を含む。例えば、データバスおよびマザーボードを含むバスシステム114が、構成要素102、104、106、108、110、および112の間でデータ通信を確立し、制御するのに使用されることが可能である。また、他の例示的なシステムアーキテクチャが使用されることも可能である。
【００１７】
処理デバイス102は、例えば、1つまたは複数のマイクロプロセッサを含むことが可能である。第1のデータストア104は、例えば、ダイナミックランダムアクセスメモリなどのランダムアクセスメモリストレージデバイス、または他のタイプのコンピュータ可読媒体メモリデバイスを含むことが可能である。第2のデータストア106は、例えば、1つまたは複数のハードドライブ、フラッシュメモリ、および/または読み取り専用メモリ、または他のタイプのコンピュータ可読媒体メモリデバイスを含むことが可能である。
【００１８】
例示的な入力デバイス108は、キーボード、マウス、スタイラス、タッチスクリーンディスプレイなどを含むことが可能であり、さらに例示的な出力デバイス110は、ディスプレイデバイス、オーディオデバイスなどを含むことが可能である。ネットワークインタフェース112は、例えば、ネットワーク116にデータを通信するとともに、ネットワーク116からデータを通信するように動作可能な有線または無線のネットワークデバイスを含むことが可能である。ネットワーク116は、1つまたは複数のLAN(ローカルエリアネットワーク)、および/またはインターネットなどのWAN(ワイドエリアネットワーク)を含むことが可能である。
【００１９】
一部の実施形態において、デバイス100は、データストア106のようなデータストアの中

にインプットメソッドエディタコード101を含むことが可能である。インプットメソッドエディタコード101は、実行されると、インプットメソッド編集機能を処理デバイス102に実行させる命令によって定義されることが可能である。或る実施形態において、インプットメソッドエディタコード101は、ウェブブラウザ環境において実行されることが可能なスクリプト命令、例えばJava(登録商標)Script命令やECMAScript命令などの解釈される命令を、例えば、備えることが可能である。他の実施形態、例えば、コンパイルされる命令、スタンドアロンアプリケーション、アプレット、プラグインモジュールなどが、使用されることも可能である。

【００２０】

インプットメソッドエディタコード101の実行は、インプットメソッドエディタインスタンス103を生成または起動する。インプットメソッドエディタインスタンス103は、インプットメソッドエディタ環境、例えば、ユーザインタフェースを定義することが可能であり、さらにデバイス100における1つまたは複数のインプットメソッドの処理を円滑にすることが可能であり、この処理の間、デバイス100は、例えばHanzi文字などの入力文字、表意文字、または記号に関する構成入力を受け取ることができる。例えば、ユーザは、Hanzi文字の識別のために構成入力を入力するのに入力デバイス108の1つまたは複数(例えば、西洋スタイルのキーボードなどのキーボード、手書き認識エンジンを有するスタイラスなど)を使用することができる。一部の実施例において、Hanzi文字は、複数の構成入力に関連付けられることが可能である。

【００２１】

第1のデータストア104および/または第2のデータストア106は、構成入力と文字の関連付けを格納することができる。ユーザ入力に基づいて、インプットメソッドエディタインスタンス103は、データストア104および/またはデータストア106の中の情報を使用して、入力によって表される1つまたは複数の候補文字を識別することができる。一部の実施形態において、複数の候補文字が識別された場合、それらの候補文字が、出力デバイス110上に表示される。入力デバイス108を使用して、ユーザは、それらの候補文字から、ユーザが入力することを所望するHanzi文字を選択することができる。

【００２２】

一部の実施形態において、デバイス100上のインプットメソッドエディタインスタンス103は、1つまたは複数のピンイン構成入力を受け取り、さらにこれらの構成入力をHanzi文字に変換することができる。インプットメソッドエディタインスタンス103は、例えば、キーストロークから受け取られたピンイン音節または文字の構成を使用して、Hanzi文字を表すことができる。各ピンイン音節は、例えば、西洋スタイルキーボードにおけるキーに対応することが可能である。ピンインインプットメソッドエディタを使用して、ユーザは、Hanzi文字を、そのHanzi文字の音を表す1つまたは複数のピンイン音節を含む構成入力を使用することによって、入力することができる。また、ピンインIMEを使用して、ユーザは、2つ以上のHanzi文字を含む語を、それらのHanzi文字の音を表す2つ以上のピンイン音節を含む構成入力を使用することによって、入力することもできる。しかし、他の言語に関するインプットメソッドが、円滑にされることも可能である。

【００２３】

また、ウェブブラウザ、ワードプロセッシングプログラム、電子メールクライアントなどを含む他のアプリケーションソフトウェア105が、データストア104および/または106の中に格納されることも可能である。これらのアプリケーションのそれぞれは、対応するアプリケーションインスタンス107を生成することができる。各アプリケーションインスタンスは、ユーザにデータを提示すること、およびユーザからのデータ入力を円滑にすることによって、ユーザ体験を円滑にすることができる環境を定義することができる。例えば、ウェブブラウザソフトウェアが、検索エンジン環境を生成することが可能であり、電子メールソフトウェアが、電子メール環境を生成することが可能であり、ワードプロセッシングプログラムが、エディタ環境を生成することが可能であるといった具合である。

【００２４】

一部の実施形態において、デバイス100へのアクセスを有する遠隔コンピューティングシステム118が、表語文字書記体系を編集するのに使用されることも可能である。例えば、デバイス100は、ネットワーク116を介して表語文字書記体系編集能力を提供するサーバであることが可能である。一部の実施例において、ユーザは、遠隔のコンピューティングシステム、例えば、クライアントコンピュータを使用して、データストア104および/またはデータストア106の中に格納された表語文字書記体系を編集することができる。代替として、ユーザは、デバイス100にアクセスを有して、遠隔システム118上に格納された表語文字書記体系を編集することができ、例えば、デバイス100が、クライアントコンピュータによって利用されることが可能なウェブベースのインプットメソッドエディタを提供することが可能である。デバイス100は、例えば、或る文字を選択し、ネットワークインタフェース112を介してユーザから構成入力を受け取ることが可能である。処理デバイス102が、例えば、その選択された文字に隣接する1つまたは複数の文字を識別し、さらにその受け取られた構成入力およびそれらの隣接する文字に基づいて、1つまたは複数の候補文字を識別することができる。デバイス100は、それらの候補文字を含むデータ通信を遠隔のコンピューティングシステムに送り返すことができる。

【００２５】

また、他の実施形態が使用されることも可能である。例えば、インプットメソッドエディタ機能は、アプレットまたはスクリプトの形態でクライアントデバイスに供給されることも可能である。

【００２６】

図1Bは、例示的なインプットメソッドエディタシステム120のブロック図である。インプットメソッドエディタシステム120は、例えば、インプットメソッドエディタコード101、および関連するデータストア104および106を使用して実施されることが可能である。インプットメソッドエディタシステム120は、インプットメソッドエディタエンジン122、辞書124、および構成入力データストア126を含む。また、他の実施アーキテクチャおよびストレージアーキテクチャが使用されることも可能である。一部の実施形態において、構成入力データストア126は、或る言語モデルを含むことが可能である。例えば、この言語モデルは、少なくとも1つの前の語を所与とした、現在の語の確率行列であることが可能である(例えば、ユニグラムモデル)。

【００２７】

中国語を対象とする実施形態において、ユーザは、IMEシステム120を使用して、ピンイン文字をタイプ入力することによって中国語の語または句を入力することができる。IMEエンジン122は、辞書124を検索して、それらのピンイン文字と合致する1つまたは複数の中国語の語および句をそれぞれが含む候補辞書エントリを識別することができる。辞書124は、1つまたは複数の言語モデルにおいて使用される表語文字書記体系の文字、語、または句、ならびに、例えば、英語、ドイツ語、スペイン語などのローマ字ベースの、もしくは西洋スタイルのアルファベットにおける文字、語、および句に対応するエントリ128を含む。

【００２８】

語は、1つのHanzi文字、または連続するHanzi文字の系列を含むことが可能である。連続するHanzi文字の系列は、辞書124の中の複数の語を構成することが可能である。例えば、「リンゴ」という意味を有する語(「苹果」)は、ピンイン入力、「ping」および「guo」にそれぞれ対応する2つの構成要素Hanzi文字「苹」および「果」を含む。また、文字「果」は、「果物」を意味する構成要素語でもある。同様に、「全球定位系統」という語は、辞書124の中の3つの語から成る。構成要素語には、(1)「地球全体の」を意味する「全球」、(2)「測位」を意味する「定位」、および(3)「システム」を意味する「系統」が含まれることが可能である。これらの語、「全球」、「定位」、および「系統」のそれぞれも同様に、辞書124の中に存在する2つの構成要素語から成る。

【００２９】

辞書エントリ128は、それぞれが1つまたは複数の文字を含む、慣用句(例えば、「胸有

成竹」)、固有名詞(例えば、「オーストリア共和国」を意味する「奥地利共和国」)、歴史上の人物もしくは有名人の名前(例えば、「チンギスハン」を意味する「成吉思汗」)、技術用語(例えば、「全地球測位システム」を意味する「全球定位系統」)、句(「一去不復返」)、書名(例えば、「Dream of the Red Chamber」を意味する「紅楼梦」)、美術作品の題名(例えば、「Upper River During the Qing Ming Festival」を意味する「清明上河図」)、および映画の題名(例えば、「Crouching Tiger, Hidden Dragon」を意味する「臥虎藏龍」)などを含むことが可能である。同様に、辞書エントリ128は、例えば、地理的エンティティまたは政治的エンティティの名前、企業の名前、教育機関の名前、動物または植物の名前、機械の名前、曲名、演劇の題名、ソフトウェアプログラムの名前、消費者製品の名前などを含むことが可能である。辞書124は、例えば、数千の文字、語、および句を含むことが可能である。

【００３０】

一部の実施形態において、辞書124は、文字の間の関係についての情報を含む。例えば、辞書124は、文字に、その文字に隣接する文字に応じて割り当てられたスコアまたは確率値を含むことが可能である。辞書124は、辞書エントリ128の1つにそれぞれ関連付けられて、エントリ128が一般にどれだけ頻繁に使用されるかを示すエントリスコアまたはエントリ確率値を含むことが可能である。

【００３１】

構成入力データストア126は、構成入力と、辞書124の中に格納されたエントリ128との関連付けを含む。一部の実施形態において、構成入力データストア126は、辞書124の中のエントリのそれぞれを、インプットメソッドエディタエンジン122によって使用される構成入力(例えば、ピンイン入力)にリンクすることができる。例えば、インプットメソッドエディタエンジン122が、辞書124および構成入力データストア126の中の情報を使用して、辞書124の中の1つまたは複数のエントリを、構成入力データストア126の中の1つまたは複数の構成入力に関連付け、さらに/またはそのような入力として識別することができる。また、他の関連付けが使用されることも可能である。IMEシステム120における候補選択は、格付けされ、さらにこの格付けに応じてインプットメソッドエディタにおいて提示されることが可能である。

【００３２】

一部の実施形態において、インプットメソッドエディタエンジン122が、構成入力データストア126の言語モデルを使用して、エントリを関連付け、さらに/または識別することができる。例えば、IMEシステム120が、言語モデルを使用して、前の1つまたは複数の入力語に基づいて、候補関連付けを格付けすることができる。

【００３３】

辞書124の中に格納された語および句の一部は、語彙の中で長い履歴を有することが可能である一方で、他の語および句は、比較的新しいことが可能である。言語の語彙は常に進化しているため、辞書124は、頻繁な更新を要求する可能性がある。正確で適時の更新を円滑にするのに、語検出システムが、利用されることが可能である。

【００３４】

図2Aは、例示的な語検出システム200のブロック図である。語検出システム200は、辞書、例えば、辞書124、語処理モジュール206、新語アナライザモジュール208、および辞書アップデータモジュール210を含む。語検出システムは、インターネットなどのネットワーク、例えば、WAN(ワイドエリアネットワーク)202を介して語コーパス204にアクセスすることができる。語検出システム200は、語コーパス204の中の新たな語を検出するように構成されることが可能である。例えば、語検出システム200は、語コーパス204からのHanzi文字によって定義される新たな中国語の語を識別することができる。一部の実施形態において、語検出システム200は、識別された新たな語を辞書124の中に格納することによって、辞書124を更新する。例えば、語検出システム200は、これらの新たな中国語の語を表すエントリを辞書124に追加することができる。次に、辞書124が、辞書124と適合するインプットメソッドエディタを利用するコンピュータデバイスに供給される、さらに/また

はそのようなデバイスによってアクセスされることが可能である。

【００３５】

語処理モジュール206、新語アナライザモジュール208、および辞書アップデータモジュール210は、語コーパス204の中の新たな語を検出するように構成されたソフトウェアおよび/またはハードウェアの処理モジュールであることが可能である。これらのモジュールの例示的なソフトウェア実施形態には、実体のあるコンピュータ可読媒体の中に格納され、さらにこの実体のあるコンピュータ可読媒体とデータ通信状態にあるコンピュータ処理デバイスによって実行可能である命令が含まれる。そのような命令には、オブジェクトコード、コンパイルされるコード、解釈される命令などが含まれることが可能である。一部の実施形態において、語処理モジュール206、新語アナライザモジュール208、および辞書アップデータモジュール210は、1つまたは複数のネットワーク化されたサーバコンピュータ、例えば、サーバファームにおいて実施されることが可能であり、さらに大量の語コーパス、例えば、数千もしくは数百万さえものウェブベースの文書にアクセスし、そのような文書を処理するように構成されることが可能である。また、他の実施形態が使用されることも可能である。

【００３６】

語コーパス204は、様々なソースからの語を含む。例示的な語コーパスは、ウェブページおよびウェブファイル、クエリログ、ブログ、電子メールメッセージ、あるいは語データを含む他のデータなどの、ウェブ文書を含むことが可能である。図示される実施例において、語コーパス204は、ウェブ文書214、電子通信216、データストア218、および他の語ソース220からのHanzi文字を含むことが可能である。ウェブ文書214は、WAN202を介してアクセス可能な、公開されたウェブページを含むことが可能である。例えば、語コーパス204は、個人ウェブサイトもしくは会社ウェブサイト、ソーシャルネットワーキングウェブサイトにおけるプロファイルページ、ブログエントリ、オンラインニュース記事、および/またはインターネット上で公開される他のテキストからの語を含むことが可能である。電子通信216は、電子メール、SMS(ショートメッセージサービス)、検索クエリ、または他の通信方法などのネットワーク通信を含むことが可能である。例えば、語コーパス204は、電子メールメッセージ、SMSメッセージ、および検索クエリの中で使用されるテキストを含むことが可能である。一部の実施形態において、語コーパス204は、他のIMEデバイスに関連するオンライン辞書、ユーザファイルなどの他のデータストア218からの語を含むことも可能である。一部の実施例において、語コーパス204は、電子ブック、電子辞書、電子形態の様々なデバイスのユーザマニュアル、または語データの他の任意の電子ソースなどの、他の語ソース220の中で使用される語を含むことも可能である。

【００３７】

一部の実施形態において、語コーパス204は、1つまたは複数の言語における文書の中の語を含むことが可能である。例えば、コーパス204の中の単一の文書が、複数の言語を含むことが可能である(例えば、英国の政治についての中国語新聞における社説が、中国語と英語の両方を含むことが可能である)。一部の実施形態において、語処理モジュール206が、語検出のために語コーパス204から或る言語に関する文字、例えば、Hanzi文字を抽出することが可能である。

【００３８】

一部の実施形態において、語処理モジュール206が、Hanzi文字処理モジュールを含むことが可能である。一実施例において、Hanzi文字処理モジュールは、語コーパス204の中のHanzi文字を処理することができる。一部の実施例において、語処理モジュール206は、日本語文字処理モジュール、朝鮮語文字処理モジュール、および/または他の表語文字処理モジュールなどの、他の表語文字言語を処理する処理モジュールを含むことが可能である。

【００３９】

一部の実施形態において、語検出システム200は、パーティションデータストア212を含む。パーティションデータストア212は、語コーパス204のコピー、または語コーパスの大

部分、例えば、ソフトウェアエージェントが巡回するウェブページのコピーを含むことが可能であり、さらに語処理モジュール206が、パーティションデータストア212の中に格納されたデータを分割することができる。例えば、語処理モジュール206は、語コーパス204と関係するデータを訓練コーパスと開発コーパスに分割することができる。一部の実施形態において、訓練コーパスおよび開発コーパスの中のデータは、パーティションデータストア212の中に格納されることが可能である。一部の実施形態において、2つ以上のパーティションが生成されて、パーティションデータストア212の中に格納されることが可能である。

【００４０】

一部の実施形態において、語処理モジュール206は、語コーパス204の中の文書を識別し、さらにパーティションデータストア212の中のパーティションデータに従って文書識別子、例えば、URL(ユニフォームリソースロケータ)を格納することができる。これらの実施形態において、パーティションデータストア212は、語コーパス204のコピー、または語コーパス204の大部分のコピーを含まなくてもよい。また、語コーパス204を管理するための他のデータ記憶技術および/またはデータ割り当て技術が使用されることも可能である。

【００４１】

語処理モジュール206は、言語モデルを含むことが可能である。例えば、語処理モジュール206は、語コーパス204の中のデータを利用して、nグラム言語モデルを生成することができる。nグラム言語モデルは、所与の系列からのn個の語の部分系列の確率を含むことが可能である。nグラム言語モデルは、n=1であるユニグラム言語モデル、n=2であるバイグラム言語モデル、および/またはn=3であるトライグラム言語モデル、あるいは他のnグラムモデルを含むことが可能である。いくつかの実施形態において、語処理モジュール206は、パーティションデータストア212の中の分割されたデータセットの1つまたは複数、例えば、訓練コーパスに関するnグラム言語モデルを生成することができる。

【００４２】

一部の実施形態において、語処理モジュール206が、区切り記号なしに語コーパス204の中の語を識別することが可能である。例えば、語処理モジュール206は、辞書124、および既存の1つまたは複数の言語モデルを使用して、語コーパス204の中の語を識別することができる。一実施例において、語コーパス204の中の所与の文に関して、語処理モジュール206は、その文を形成する語の1つまたは複数の組合せを識別することができる。その言語モデルに基づいて、語処理モジュール206は、例えば、組合せを格付けし、さらに最高の格付けを有する語の組合せを選択することができる。

【００４３】

語処理モジュール206は、訓練コーパスの中の語と、辞書124の中の語とを比較して、1つまたは複数の潜在的な新しい語、例えば、訓練コーパスの中に現れるが、辞書124の中には現れない候補語を識別することができる。一部の実施例において、システム200は、分割されたデータストア212の中のデータを使用して、候補語が新たな語であるかどうかを検証することができる。語処理モジュール206は、例えば、訓練コーパス(例えば、訓練コーパス232)の中のnグラム言語モデルに基づく、候補語の第1の確率、および候補語を構成する語の確率、ならびに、例えば、開発コーパスの中で候補語が出現する回数、および開発コーパスの中の語の総数に基づく第2の確率を算出する。

【００４４】

第1の確率、および第2の確率を使用して、新語アナライザモジュール208が、候補語が新たな語であるかどうかを判定することができる。一実施例において、新語アナライザモジュール208は、第1の確率、および第2の確率を使用して、候補語に関して、開発コーパスにおける不確かさ、例えば、エントロピー値が、減少するかどうかを判定することができる。一部の実施形態において、新語アナライザモジュール208は、第1の確率、および第2の確率に基づいて、第1のエントロピー関連値、および第2のエントロピー関連値を生成する。例えば、第1のエントロピー関連値、および第2のエントロピー関連値は、それぞれ

、候補語を伴う言語モデルの不確かさ、および候補語を伴わない言語モデルの不確かさを表すことが可能である。一部の実施形態において、新語アナライザモジュール208は、第1のエントロピー関連値が第2のエントロピー関連値より小さい場合、候補語が新たな語であると判定する。エントロピーの低下は、新たな語を正しく検出したことからもたらされるIG(情報利得)を示すことが可能である。

【００４５】

候補語が新たな語であると判定された場合、新語アナライザモジュール208は、その新たな語で辞書124を更新するよう辞書アップデータモジュール210に通知することができる。

【００４６】

一部の実施形態において、エントロピー関連値は、実際のエントロピー値の近似であることが可能である。例えば、訓練コーパスおよび開発コーパスの中の語の数は、言語モデルの中に候補語を含めることによって、わずかに変化する可能性があり、例えば、「全球」という語は、1つの語としてカウントされることが可能であり、あるいは構成要素文字、全と球が別々に考慮される場合、2つの語としてカウントされることが可能である。

【００４７】

一実施形態において、新語アナライザモジュール208は、例えば、候補語、および候補語を定義する構成要素語だけに関する確率を調整することによって、訓練コーパスおよび開発コーパスの固定サイズを使用してエントロピー関連値を生成することができる。このため、これらのエントロピー関連値は、実際のエントロピー値の良好な近似である。新語アナライザモジュール208は、これらのエントロピー関連値を、訓練コーパスおよび/または開発コーパスのエントロピー値として使用することができる。

【００４８】

図2Bは、図2Aのシステム200の例示的な実施形態のブロック図である。図2Bに示されるとおり、システム200は、訓練コーパス232および開発コーパス234を含む。一部の実施形態において、語処理モジュール206は、語コーパス204を分割して、訓練コーパス232および開発コーパス234を生成する。例えば、訓練コーパス232および開発コーパス234は、パーティションデータストア212の中に格納される、またはストア212の中で表されることが可能である。

【００４９】

一部の実施形態において、語処理モジュール206は、語の間にスペースのない原文を語系列にセグメント化するセグメント化モジュールを含むことが可能である。語処理モジュールの中のセグメント化モジュールは、例えば、辞書および言語モデルを利用して、語系列のセグメントを生成することができる。

【００５０】

前述したとおり、語処理モジュール206は、訓練コーパス232の中にnグラム言語モデルを含めることが可能である。一部の実施形態において、語処理モジュール206は、訓練コーパス232の中の既存の2つ以上の語を組み合わせることによって、候補語を識別することができる。例えば、語処理モジュール206は、既存の2つの語、xとyを組み合わせることによって、候補語(x，y)を識別することができる。

【００５１】

一部の実施形態において、システム200は、語コーパス204からの語データ、例えば、訓練コーパス232および開発コーパス234の中のウェブページデータを利用して、候補語が新たな語であるかどうかを判定することができる。例えば、語処理モジュール206は、識別された候補語(x，y)を含めるように訓練コーパス232の中に格納されたデータからnグラム言語モデルを生成することができる。ユニグラムモデルは、候補語の確率、P(x，y)を含むことが可能であり、さらに語処理モジュール206は、候補語、xyを構成する語xとyの対応する確率P(x)およびP(y)を算出することもできる。さらに、語処理モジュール206は、開発コーパス234から候補語の語カウント値、D(x，y)、および構成要素語、D(x)およびD(y)の語カウント値を生成する。例えば、D(x)、D(y)、およびD(x，y)は、開発コーパス234

の中で、それぞれ、x、y、および(x, y)が出現する回数であることが可能である。語カウント値を使用して、システム200は、開発コーパス234におけるx、y、および(x, y)の確率を算出することができる。例えば、開発コーパス234における(x, y)の確率は、

【００５２】

【数１】

$$\frac{D(x,y)}{\|D\|}$$

【００５３】

によって算出されることが可能であり、ただし、‖D‖は、開発コーパス234の中の語の総数である。

【００５４】

確率、p(x)、p(y)、およびp(x, y)、ならびに語カウント値、D(x)、D(y)、およびD(x, y)を受け取った後、新語アナライザモジュール208は、候補語が新たな語であるかどうかを判定する。一部の実施形態において、新語アナライザモジュール208は、その候補語を新たな語として含めることによって、開発コーパス234の不確かさが減少する場合、その候補語が新たな語であると判定することができる。一部の実施例において、エントロピー値を使用して、開発コーパス234の不確かさが測定されることが可能である。例えば、開発コーパス234のエントロピー値は、

【００５５】

【数２】

$$H = -\sum_{w \in V} \frac{D(w)}{\|D\|} \cdot \log p(w)$$

【００５６】

によって算出されることが可能であり、ただし、Vは、エントロピー、Hを計算するのに考慮される語セット全体であり、wは、開発コーパス234の中の語であり、p(w)は、開発コーパスの中の、この語の確率であり、さらにD(w)は、開発コーパスの中でwが出現する回数である。

【００５７】

一部の実施形態において、新語アナライザモジュール208は、開発コーパス234に関するエントロピー値、HおよびH'を生成することができ、ただし、HおよびH'は、それぞれ、言語モデルに候補語を含めることを伴わない開発コーパス234のエントロピー値、および言語モデルに候補語を含めることを伴う開発コーパス234のエントロピー値である。一部の実施形態において、新語アナライザモジュール208は、それぞれ、候補語を伴わないコーパスの実際のサイズ、および候補語を伴うコーパスの実際のサイズを使用して、実際のエントロピー値、HおよびH'を生成する。一部の実施形態において、新語アナライザモジュール208は、これらの実際のエントロピー値を近似することができる1つまたは複数のエントロピー関連値を使用することもできる。例えば、新語アナライザモジュール208は、候補語を伴わないコーパス232、234のサイズを使用して、H'を生成することができる。訓練コーパス232、および開発コーパス234のサイズは、語彙の中に新たな語として(x, y)を含めた後、小さくなる可能性があるものの、その違いは、候補語(x, y)を伴うコーパス232、234のエントロピーを計算することに関して、無視できるほど小さいことが可能である。例えば、n個の構成要素語の系列、W1W2...Wnが、潜在的に新しい語と考えられる場合、コーパスのサイズは、W1W2...Wnの出現の回数、例えば、mにn-1を掛けた値、すなわち、m*(n-1)の分だけしか小さくならない。

【００５８】
HとH'を比較することによって、新語アナライザモジュール208は、候補語が新たな語であるかどうかを判定することができる。例えば、H'-H<0である場合、新語アナライザモジュール208は、その候補語を含めることによって開発コーパス234のエントロピー値が小さくなるため、その候補語が新たな語であると判定することが可能である。
【００５９】
一部の実施例において、新語アナライザモジュール208は、確率、p(x)、p(y)、p(x，y)、および語カウント値、D(x)、D(y)、およびD(x，y)を使用して、エントロピー値、HとH'を比較する。候補語、および構成要素語以外の語の語頻度は、候補語の追加による影響を受けないため、HとH'の差を生成するための式は、単純化された式を使用して生成されることが可能である。等しい項を消去することによって、HとH'の差を計算する以下の式が、導き出されることが可能である。すなわち、
【００６０】
【数３】

$$Z = H' - H = -\left[\frac{D(x,y)}{\|D\|}\cdot\log p'(x,y) + \frac{D(x)-D(x,y)}{\|D\|}\cdot\log p'(x) + \frac{D(y)-D(x,y)}{\|D\|}\cdot\log p'(y)\right] + \left[\frac{D(x)}{\|D\|}\cdot\log p(x) + \frac{D(y)}{\|D\|}\cdot\log p(y)\right]$$

【００６１】
ただし、p'(x)、p'(y)、p'(x，y)、p(x)、およびp(y)は、訓練コーパス232の言語モデルの確率である。p'(x)、p'(y)、p'(x，y)の値は、それぞれ、文字の系列xyが候補語と考えられる場合の言語モデルにおけるx、y、および(x，y)の確率である。逆に、p(x)およびp(y)の値は、それぞれ、文字の系列xyが候補語と考えられない場合の言語モデルにおけるx、およびyの確率である。このため、系列xyの各回の出現が、p(x)およびp(y)のそれぞれの確率を増加させるにつれ、p(x)の値> p'(x)であり、さらにp(y)の値>p'(y)である。
【００６２】
或る実施形態において、新語アナライザモジュール208は、以下の条件と等価であるZ<0である場合、候補語(x，y)が新たな語であると判定することができる。すなわち、
【００６３】
【数４】

$$\frac{D(x,y)}{\|D\|}\cdot\log\frac{p'(x,y)}{p'(x)\cdot p'(y)} > \frac{D(x)}{\|D\|}\cdot\log\frac{p(x)}{p'(x)} + \frac{D(y)}{\|D\|}\cdot\log\frac{p(y)}{p'(y)}$$

【００６４】
したがって、候補語(x，y)は、以上の不等式が成立する場合、新たな語であると判定される。
【００６５】
一部の実施形態において、確率p(x)、p(y)、p'(x)、およびp'(y)は、訓練コーパス232の中の語の総数で割った、訓練コーパス232の中でx、y、および(x，y)が出現する回数を使用して表される。例えば、
【００６６】

【数５】

$$p'(x)=\frac{T(x)-T(x,y)}{\|T\|}=p(x)-p(x,y)$$

$$p'(y)=\frac{T(y)-T(x,y)}{\|T\|}=p(y)-p(x,y)$$

$$p(x)=\frac{T(x)}{\|T\|}$$

$$p(y)=\frac{T(y)}{\|T\|}$$

【００６７】

ただし、T(x)、T(y)、およびT(x，y)は、訓練コーパス232の中で、それぞれ、x、y、および(x，y)が出現する回数であり、さらに‖Ｔ‖は、訓練コーパス232の中の語の総数である。このため、新語アナライザモジュール208は、以下の不等式に従って前出の不等式を評価することができる。

【００６８】

【数６】

$$\frac{D(x,y)}{\|D\|}\cdot\log\frac{p'(x,y)}{p'(x)\cdot p'(y)}>\frac{D(x)}{\|D\|}\cdot\log\frac{p(x)}{p(x)-p(x,y)}+\frac{D(y)}{\|D\|}\cdot\log\frac{p(y)}{p(y)-p(x,y)}$$

【００６９】

この不等式は、候補語が妥当であるかどうかを判定するように、以下のとおり書き換えられることが可能である。すなわち、

【００７０】

【数７】

$$\frac{D(x,y)}{\|D\|}\cdot\log\frac{p'(x,y)}{p'(x)\cdot p'(y)}>\frac{D(x)}{\|D\|}\cdot\log\frac{T(x)}{T(x)-T(x,y)}+\frac{D(y)}{\|D\|}\cdot\log\frac{T(y)}{T(y)-T(x,y)}$$

【００７１】

或る実施形態において、新語アナライザモジュール208が、開発コーパス234の中の候補語の語頻度(例えば、

【００７２】

【数８】

$$\frac{D(x,y)}{\|D\|}$$

【００７３】

)、ならびに訓練コーパス232の中の候補語、および構成要素語の語頻度(例えば、p(x)、p

(y)、およびp(x, y))を使用して、第1の値を生成することができる。これらの値に基づく第1のエントロピー様の値V1が、以下の式に基づいて計算されることが可能である。すなわち、

【００７４】

【数９】

$$V1 = \frac{D(x,y)}{\|D\|} \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)}$$

【００７５】

同様に、新語アナライザモジュール208は、開発コーパス234の中の構成要素語の語頻度(例えば、

【００７６】

【数１０】

$$\frac{D(x)}{\|D\|} \text{および} \frac{D(y)}{\|D\|}$$

【００７７】

)、ならびに訓練コーパス232の中の候補語、および構成要素語の語頻度を使用して、第2のエントロピー値を生成することができる。これらの値に基づく第2のエントロピー様の値V2が、以下の式に基づいて計算されることが可能である。すなわち、

【００７８】

【数１１】

$$V2 = \frac{D(x)}{\|D\|} \cdot \log \frac{p(x)}{p(x) - p(x,y)} + \frac{D(y)}{\|D\|} \cdot \log \frac{p(y)}{p(y) - p(x,y)}$$

【００７９】

一部の実施形態において、新語アナライザモジュール208は、V1>V2である場合、候補語が新たな語であると判定する。新たな語をより多く包含するように、またはより少なく包含するように他の不等式、例えば、Sがスカラー値であるV1>S^{*}V2が使用されることも可能である。このスカラー値は、固定、例えば、0.9であることも、応用先に応じて調整されることも可能である。

【００８０】

辞書アップデータモジュール210は、新語アナライザモジュール208からの判定を示すデータを受け取る。一部の実施形態において、新語アナライザモジュール208が、候補語が新語であると判定した場合、辞書アップデータモジュール210が、その新語を辞書124に追加することができる。

【００８１】

システム200は、語コーパス204を処理し、さらにスケジュールされた仕方で複数の候補語を処理することができる。例えば、コーパスの中の新たな語を検出するプロセスが、毎日、毎週、または毎月、実施されることが可能である。また、他のトリガするイベントが、使用されることも可能であり、例えば、ウェブベースのインプットメソッドエディタに

関する新語検出プロセスが、認識されない語が統計的に有意であるのに十分な頻度で入力として受け取られた場合、実行されることが可能である。

【００８２】

図3は、語コーパス(例えば、語コーパス204)の中の新たな語を識別するための例示的なプロセス300の流れ図である。プロセス300は、例えば、1つまたは複数のコンピュータを含むシステムにおいて実施されることが可能である。例えば、語検出システム200が、プロセス300における動作の一部またはすべてを実行するのに使用されることが可能である。

【００８３】

プロセス300は、訓練コーパスの中の既存の語、および候補語に関する第1の語頻度を算出することから始まる(302)。候補語は、構成要素語の系列によって定義されることが可能であり、さらに各構成要素語は、辞書の中の既存の語であることが可能である。例えば、語処理モジュール206が、訓練コーパス232における候補語(例えば、(x，y))、およびその候補語を構成する既存の語(例えば、xおよびy)の確率(例えば、p(x)、p(y)、p(x，y))を算出することができる。一部の実施形態において、語処理モジュール206は、訓練コーパス232の中でnグラム言語モデルを生成して、これらの語頻度を算出することができる。

【００８４】

次に、プロセス300は、開発コーパスの中の構成要素語および候補語に関する第2の語頻度を算出する(304)。例えば、語処理モジュール206が、開発コーパス234の中の識別された新語および構成要素語の語カウント値(例えば、D(x，y)、D(x)、およびD(y))を算出することができる。一部の実施形態において、開発コーパス234の中の語の語頻度は、開発コーパス234の中のその語の語カウントを開発コーパス234の中の語の総数で割ることによって、算出されることが可能である。例えば、語処理モジュール206は、

【００８５】

【数１２】

$$\frac{D(w)}{\|D\|}$$

【００８６】

を計算することによって、開発コーパスの中のwの語頻度を算出することができる。

【００８７】

語頻度を算出した後、プロセス300は、候補語の第2の語頻度、ならびに構成要素語および候補語の第1の語頻度に基づいて、候補語エントロピー関連測度を算出する(306)。例えば、新語アナライザモジュール208が、D(x，y)、p(x)、p(y)、およびp(x，y)を使用して、候補語エントロピー関連測度V1を算出することができる。

【００８８】

プロセス300は、構成要素語の第2の語頻度、ならびに構成要素語および候補語の第1の語頻度に基づいて、既存語エントロピー関連測度を算出する(308)。例えば、新語アナライザモジュール208が、D(x)、D(y)、p(x)、p(y)、およびp(x，y)を使用して、既存語エントロピー関連測度V2を算出することができる。

【００８９】

次に、プロセス300は、候補語エントロピー関連測度が、既存語エントロピー関連測度を超えているかどうかを判定する(310)。例えば、新語アナライザモジュール208が、V1とV2を比較して、V1がV2より大きいかどうかを判定することができる。

【００９０】

プロセス300が、候補語エントロピー関連測度が既存語エントロピー関連測度を超えていると判定した場合、その候補語は、新たな語であると判定される(312)。例えば、新語アナライザモジュール208が、V1>V2である場合、その候補語が新たな語であると判定する

ことができる。

【００９１】

プロセス300が、候補語エントロピー関連測度が既存語エントロピー関連測度を超えていないと判定した場合、その候補語は、新たな語であるとは判定されない(314)。例えば、新語アナライザモジュール208が、V1≦V2である場合、その候補語が新たな語ではないと判定することができる。

【００９２】

一部の実施形態において、エントロピー関連測度は、図2A～図2Bを参照して説明されるとおり、エントロピー測度を計算することによって、またはコーパスの固定サイズを使用してエントロピー測度を近似することによって、算出される。

【００９３】

図4は、候補語および既存の語に関するエントロピー関連測度を算出するための例示的なプロセス400の流れ図である。例えば、プロセス400は、1つまたは複数のコンピュータを含むシステムにおいて実施されることが可能である。例えば、語検出システム200が、プロセス400における動作の一部またはすべてを実行するのに使用されることが可能である。

【００９４】

プロセス400は、候補語および構成要素語の確率に基づいて、第1の対数値を算出することから始まる(402)。例えば、新語アナライザモジュール208は、p(x)、p(y)、およびp(x, y)を使用して、第1の対数値を算出することができる。一実施例において、第1の対数値は、

【００９５】

【数１３】

$$\log\frac{p(x,y)}{p(x)\cdot p(y)}$$

【００９６】

であることが可能である。

【００９７】

次に、プロセス400は、候補語の語カウント値、および第1の対数値に基づいて、候補語エントロピー測度を算出する(404)。例えば、新語アナライザモジュール208が、候補語の語カウントD(x, y)、および第1の対数値を使用して、値V1を生成することができる。

【００９８】

プロセス400は、候補語および構成要素語の確率に基づいて、第2の対数値を算出する(406)。例えば、新語アナライザモジュール208が、p(x)、p(y)、およびp(x, y)を使用して、第2の対数値を算出することができる。例えば、第2の対数値は、

【００９９】

【数１４】

$$\log\frac{p(x)}{p(x)-p(x,y)}\text{および}\log\frac{p(y)}{p(y)-p(x,y)}$$

【０１００】

を含むことが可能である。

【０１０１】

次に、プロセス400は、構成要素語の語カウント、および第2の対数値に基づいて、既存語エントロピー測度を算出する(408)。例えば、新語アナライザモジュール208が、候補語の語カウントD(x)、D(y)、および第2の対数値を使用して値V2を生成することができる。

【０１０２】

図5は、語コーパスの中の新たな語を識別するための別の例示的なプロセス500の流れ図である。例えば、プロセス500が、システム200において実施されることが可能である。プロセス500は、第1のコーパスの中の既存の語、および候補語に関する第1の語確率を算出することから始まる(502)。例えば、語処理モジュール206が、訓練コーパス232におけるp(x)、p(y)、およびp(x，y)を算出することができる。

【０１０３】

プロセス500は、第2のコーパスにおける構成要素語および候補語に関する第2の語確率を算出する(504)。候補語は、構成要素語の系列によって定義されることが可能であり、さらに各構成要素語は、辞書の中の既存の語であることが可能である。例えば、語処理モジュール206が、開発コーパス234における構成要素語、xおよびy、ならびに候補語(x，y)の確率を算出することができる。例えば、語処理モジュール206が、開発コーパス234におけるD(x)、D(y)、およびD(x，y)、ならびに‖D‖を使用して、開発コーパス234におけるx、y、および(x，y)の確率を算出することができる。

【０１０４】

次に、プロセス500は、候補語の第2の候補語確率、および構成要素語の第1の語確率に基づいて、第1のエントロピー関連値を算出する(506)。例えば、新語アナライザモジュール208が、D(x，y)、p(x)、p(y)、およびp(x，y)を使用して、V1を算出することができる。

【０１０５】

プロセス500は、候補語の第2の構成要素語確率、および構成要素語の第1の語確率に基づいて、第2のエントロピー関連値を算出する(508)。例えば、新語アナライザモジュール208が、D(x)、D(y)、ならびにp(x)、p(y)、およびp(x，y)を使用して、V2を算出することができる。

【０１０６】

エントロピー関連値を算出した後、プロセス500が、第1のエントロピー関連値が第2のエントロピー関連値を超えているかどうかを判定する(510)。例えば、新語アナライザモジュール208が、V1>V2であるかどうかを判定することができる。

【０１０７】

プロセス500が、第1のエントロピー関連値V1が第2のエントロピー関連値V2を超えていると判定した場合、その候補語は、新たな語であると判定される(512)。例えば、新語アナライザモジュール208が、V1>V2である場合、その候補語が新たな語であると判定することができる。

【０１０８】

プロセス500が、第1のエントロピー関連値が第2のエントロピー関連値を超えていないと判定した場合、その候補語は、新たな語ではないと判定される(514)。例えば、新語アナライザモジュール208が、V1≦V2である場合、その候補語が新たな語ではないと判定することができる。

【０１０９】

図6は、或る語コーパスの中の新たな語を、別の語コーパスからの語確率に基づいて、識別するための別の例示的なプロセス600の流れ図である。例えば、プロセス400が、1つまたは複数のコンピュータを含むシステムにおいて実施されることが可能である。

【０１１０】

プロセス600は、ウェブ文書のコレクションを訓練コーパスと開発コーパスに分割することから始まる(602)。例えば、語処理モジュール206が、語コーパス204を訓練コーパス232と開発コーパス234に分割することができる。

【０１１１】

次に、プロセス600は、訓練コーパスにおける語の第1の語確率に関して訓練コーパス上で言語モデルを訓練する(604)。例えば、語処理モジュール206が、訓練コーパス232のnグラム言語モデルを訓練し、さらに訓練コーパス232における語の確率(例えば、p(x)、p(y)、およびp(x，y))を獲得することができる。

【０１１２】
プロセス600は、開発コーパスにおける候補語、および2つ以上の対応する語の出現回数をカウントする(606)。例えば、語処理モジュール206が、開発コーパス234における候補語の出現回数D(x, y)、ならびに候補語の構成要素語の出現回数D(x)およびD(y)をカウントすることができる。
【０１１３】
次に、プロセス600は、開発コーパスにおける候補語の出現回数、および第1の語確率に基づいて、第1の値を算出する(608)。例えば、新語アナライザモジュール208が、D(x, y)、ならびにp(x)、p(y)、およびp(x, y)に基づいて、V1を算出する。
【０１１４】
プロセス600は、開発コーパスにおける2つ以上の対応する語の出現回数、および第1の語確率に基づいて、第2の値を算出する(610)。例えば、新語アナライザモジュール208が、D(x)およびD(y)、ならびにp(x)、p(y)、およびp(x, y)に基づいて、V2を算出する。
【０１１５】
第1の値、および第2の値を算出した後、プロセス600は、第1の値を第2の値と比較することによって、その候補語が新たな語であるかどうかを判定する(612)。例えば、新語アナライザモジュール208が、V1とV2を比較することができる。プロセス600が、その候補語が新たな語であると判定した場合、プロセス600は、その候補語を辞書に追加する(614)。例えば、辞書アップデータモジュール210が、その新たな語を辞書124に追加することができる。プロセス600が、その候補語が新たな語ではないと判定した場合、プロセス600は、別の候補語(616)を識別し、さらにステップ606が繰り返される。例えば、語処理モジュール206が、語コーパス204から別の候補語を識別することが可能である。
【０１１６】
新たな語を検出することの例は、既存の2つの語に関連して前段で説明されるものの、語検出システム200は、既存の2より多くの語を構成する新たな語を検出することができる。例えば、語検出システム200は、既存の3つの語、x、y、およびzから成る候補語(x, y, z)を識別することができる。新語アナライザモジュール208が、
【０１１７】
【数１５】

$$V1 = \frac{D(x,y,z)}{\|D\|} \cdot \log \frac{p(x,y,z)}{p(x) \cdot p(y) \cdot p(z)}$$

【０１１８】
を計算することによって、第1のエントロピー関連値V1を生成することができ、さらに
【０１１９】
【数１６】

$$V2 = \frac{D(x)}{\|D\|} \cdot \log \frac{p(x)}{p(x) - p(x,y,z)} + \frac{D(y)}{\|D\|} \cdot \log \frac{p(y)}{p(y) - p(x,y,z)} + \frac{D(z)}{\|D\|} \cdot \log \frac{p(z)}{p(z) - p(x,y,z)}$$

【０１２０】
を計算することによって、第2のエントロピー関連値V2を生成することができる。V1>V2である場合、新語アナライザモジュール208が、その候補語(x, y, z)が新たな語であると判定することができ、さらに辞書アップデータモジュール210が、その新たな語を辞書124の中に格納することができる。例えば、システム200は、或る言語語彙に導入されている以下の新たな3文字語/句および4文字語/句、すなわち、「丁俊暉」(ding junhui)、「本賽季」(今季)、「世錦賽」(世界選手権)、「季后賽」(プレーオフ)、「范甘迪」(Van Cundy)、「国際足聯」(FIFA)、「反傾鎖」(アンチ低価格ダンピング)、「浄利潤」(純利益)、「証監会」(SEC)、「国資委」(中国国有資産監督管理委員会)、「美聯儲」(FED)、および

「非流通股」(非取引株式)を識別することができる。

【０１２１】

一部の実施形態において、コンピュータシステムは、1つまたは複数の特定のトピックと関係する1つまたは複数のトピック辞書を含むことが可能である。例えば、図1Bの辞書124が、1つまたは複数のトピック辞書を含むことが可能であり、さらに各トピック辞書が、或る特定のトピックに対応して、その特定のトピックと関係するトピック語を含むことが可能である。特定のトピックの例には、スポーツトピック、音楽トピック、法律トピック、医療トピックなどが含まれることが可能である。或るスポーツトピックと関係するトピック辞書には、そのスポーツと関係する語および句、例えば、「サッカー」、「フットボール」、「ゴール」、「赤い旗」などが含まれることが可能である。これらの語のいくつか、例えば、「サッカー」は、言語辞書の中の既存の語であることが可能であり、さらにこれらの語のいくつか、例えば、新しい選手の名前、新しい会場の名前などは、新たな語であることが可能である。

【０１２２】

一部の実施形態において、トピック語は、これらの新たな語および/または既存の語から識別されることが可能である。一実施例において、これらの新たな語の1つまたは複数は、これらの新たな語がシステム200を使用して識別された後、或る特定のトピックと関係付けられるように分類されることが可能である。一部の実施形態において、トピック語識別システムが、語コーパス204からトピック語を識別することができる。識別されたトピック語は、トピック辞書の1つまたは複数の辞書の中に含められることが可能である。

【０１２３】

図7Aは、トピック語を識別するための例示的なトピック語識別システム700のブロック図である。トピック語識別システム700は、トピック分類モジュール702、トピック語処理モジュール704、辞書アップデータモジュール706、およびトピック辞書708を含む。トピック分類モジュール702、トピック語処理モジュール704、および辞書アップデータモジュール706は、1つまたは複数のコンピュータ、例えば、単一のコンピュータ、またはWAN202などのネットワークを介して通信状態にある1つまたは複数のコンピュータの上に統合されることが可能である。同様に、WAN202を介して、トピック分類モジュール702は、語コーパス204の中の文書、例えば、文書コーパス710を取り出すことができる。一部の実施例において、トピック語識別システム700は、語コーパス204の中のトピック語を識別し、識別されたトピック語をトピック辞書708に更新することができる。

【０１２４】

文書コーパス710は、語コーパス204からの文書を含むことが可能であり、例えば、文書コーパス710は、語コーパス204のコピー、または語コーパス204の大部分、例えば、ソフトウェアエージェントが巡回するウェブページのコピーを含むことが可能である。この実施例において、文書コーパス710は、n個のトピック714を含み、さらに各トピックは、文書コーパス710からのトピック関連の文書、例えば、トピック文書コーパスを含む。例えば、文書コーパス710が、スポーツ関連文書、医療関連文書などを含むことが可能であり、さらにスポーツトピックが、スポーツトピック文書コーパスとしてスポーツ関連文書を含むことが可能であり、医療トピックが、医療トピック文書コーパスとして医療関連文書を含むことが可能である、といった具合である。一部の実施形態において、トピック714のそれぞれは、システム700において事前定義されることが可能である。さらに、これらのトピックの一部は、別のトピックのサブトピックであることも可能である。例えば、「テニス」および「バスケットボール」というトピックが、「スポーツ」というトピックのサブトピックであることが可能である。

【０１２５】

一部の実施形態において、トピック分類モジュール702は、文書コーパス710の中の文書をクラスタ化して、トピック文書クラスタを生成する。例えば、トピック分類モジュール702は、トピック714の1つと関係する文書をクラスタ化して、そのトピックのトピック文書クラスタを形成することができる。トピック分類モジュール702は、様々なトピック検

出方法を使用して文書を分類することができる。例えば、トピック分類モジュール702は、いくつかのクラスタ化技術(例えば、SVD(特異値分解)、K平均クラスタ化など)を使用して、文書コーパス710の中の文書からトピック文書のクラスタを生成することができる。或る実施例において、トピック分類モジュール702は、文書のそれぞれに関連度値を割り当てることができる。一実施形態において、関連度値は、文書とトピック714の各トピックの重心との類似度値であることが可能である。これらの関連度値に基づき、トピック分類モジュール702は、最も関連のあるトピックにそれらの文書を割り当てる。これらの文書割り当てに基づき、トピック分類モジュール702は、トピック714のそれぞれに関するトピック文書クラスタを生成することができる。

【０１２６】

システム700は、新語データストア712を含むことが可能である。一部の実施形態において、新語データストア712は、語コーパス204から識別された新たな語を含む。例えば、新語データストア712は、システム200を使用して識別された新たな語を格納することができる。

【０１２７】

トピック語処理モジュール704が、新語データストア712の中に格納された、識別された新たな語、および/または文書コーパス710の中で識別された既存の語を、トピック文書クラスタのそれぞれに関する候補トピック語として選択し、さらに選択された候補語が或るトピックに属するかどうかを判定することができる。選択された候補トピック語が、或る特定のトピックに属すると判定された場合、対応するトピック辞書708が、その候補トピック語で更新されることが可能である。

【０１２８】

一実施形態において、トピック語処理モジュール704は、新語データストア712およびトピック辞書708を使用して、候補トピック語を選択することができる。トピック語処理モジュール704は、対応するトピック文書の中の語のそれぞれを、新たな語、トピック語、または非トピック語として識別することができる。例えば、新たな語は、トピック辞書708のいずれにも含まれていない可能性がある新語データストア712の中に含まれる語であることが可能であり、トピック語は、関係のあるトピック辞書の中に存在する語であることが可能であり、さらに非トピック語は、関係のあるトピック辞書の中に存在しない既存の語であることが可能である。トピック語処理モジュール704は、それらの新たな語、およびそれらの非トピック語を候補トピック語として選択することができる。

【０１２９】

トピック辞書708の中に格納されたトピック文書クラスタおよびデータに基づき、トピック語処理モジュール704は、候補トピック語がトピック辞書708の1つの辞書のトピック語であると判定することができる。例えば、トピック語処理モジュール704が、文書コーパス710の中の既存の語である候補トピック語、Weが、トピック2に関連していると判定した場合、トピック語処理モジュール704は、候補トピック語、Weをトピック2辞書の中に格納するよう辞書アップデータモジュール706に通知することができる。同様に、トピック語処理モジュール704が、新たな語である候補トピック語、Wnが、トピックnに関連していると判定した場合、トピック語処理モジュール704は、候補トピック語Wnをトピックn辞書の中に格納するよう辞書アップデータモジュール706に通知することができる。

【０１３０】

図7Bは、図7Aのシステム700の例示的な実施形態のより詳細なブロック図である。図7Bに示されるとおり、トピック分類モジュール702は、クラスタ化モジュール722、重心モジュール724、および類似度モジュール726を含む。トピック分類モジュール702は、モジュール722、724、および726を使用して、文書コーパス710の中のトピック文書クラスタを生成することができる。

【０１３１】

トピック語処理モジュール704は、相違値モジュール732および閾値評価モジュール734を含む。トピック語処理モジュール704は、文書コーパス710の中の生成されたトピック文

書クラスタから、さらに/または新語データストア712から候補トピック語を識別し、さらにモジュール732および734を利用して、候補トピック語がトピック語であるかどうかを判定することができる。

【０１３２】

一部の実施形態において、トピック分類モジュール702は、文書コーパス710の中の文書のそれぞれに関するTF-IDF(用語頻度/逆文書頻度)ベクトルを生成することができる。例えば、クラスタ化モジュール722は、以下の数式に従って、文書jの中の語w_iに関するTF-IDFユニグラム頻度m_{ij}を算出することができる。すなわち、

【０１３３】

【数１７】

$$m_{ij} = f_j(w_i) \cdot \log \frac{D}{D_{w_i}}$$

【０１３４】

この式において、DおよびD_{wi}は、それぞれ、文書の総数、およびw_iを含む文書の数であり、さらに$f_j(w_i)$は、文書jの中のw_iの頻度である。文書jの中の語のTF-IDF頻度を使用して、クラスタ化モジュール722は、TF-IDFベクトルX_jを生成することによって、文書jを表すことができる。例えば、文書jは、

【０１３５】

【数１８】

$$x_j = \begin{bmatrix} m_{1j} & m_{2j} & \cdots & m_{|V|j} \end{bmatrix}^T$$

【０１３６】

として表されることが可能であり、ただし、|V|は、システム700における識別された語の数である。一部の実施形態において、クラスタ化モジュール722は、文書ベクトルm_{ij}を使用して、共起行列Mを生成することができる。

【０１３７】

同様に、トピック分類モジュール702は、例えば、トピックの文書のTF-IDFベクトルと関係する重心ベクトルを使用して、トピックのそれぞれを表すことができる。例えば、重心モジュール724が、トピック1、2、...nをそれぞれ表すトピック重心Y_1、Y_2、...Y_nを算出することができる。一部の実施形態において、重心モジュール724は、或るトピックに割り当てられた文書のTF-IDFベクトルを組み合わせることによって、トピック重心を算出することができる。一実施形態において、重心モジュール724は、以下の式に従ってトピックk(T_k)に関するトピック重心Y_kを算出することができる。

【０１３８】

【数１９】

$$Y_k = \sum_{X_i \in T_k} X_i$$

【０１３９】

一部の実施形態において、類似度モジュール726が、文書X_jと重心Y_1、Y_2、...Y_nの間の類似度距離、例えば、コサイン類似度距離を算出することができる。文書Xとトピック重心Yの間の距離D(X, Y)は、以下の式に従って算出されることが可能である。すなわち、

【０１４０】

【数２０】

$$D(X,Y)=1-\frac{X\cdot Y+\varepsilon\sum_{x_i>0}x_i+\varepsilon\sum_{y_i>0}y_i+\varepsilon^2}{(\|X\|+\varepsilon)\cdot(\|Y\|+\varepsilon)}$$

【０１４１】
ただし、x_iは、TF-IDFベクトルXの成分であり、y_iは、TF-IDFベクトルYの成分であり、さらにεは、1より小さい正の実数である。
【０１４２】
文書と重心のそれぞれとの間の距離に基づいて、クラスタ化モジュール722は、文書に最も近いトピックに文書を割り当てることによって、文書を文書クラスタの中に再クラスタ化することができる。例えば、クラスタ化モジュール722は、文書とトピック重心の間の距離を比較し、さらに最も近い重心を決定する。
【０１４３】
トピック分類モジュール702は、トピック文書を繰り返し分類することができる。最初、トピック分類モジュール702は、n個の初期クラスタ、およびこれらのクラスタのn個の初期重心を生成することができる。一実施例において、クラスタ化モジュール722は、共起行列Mに関するSVD(特異値分解)を実行して初期文書クラスタを識別することができる。例えば、文書のそれぞれが、$C^0(X_i)$によって表される初期クラスタの1つに割り当てられることが可能である。他の実施形態において、初期クラスタは、トピックに文書をランダムに割り当てることによって、生成されることも可能である。初期文書クラスタに基づいて、重心モジュール724は、以下を計算することによって初期重心を生成することができる。すなわち、
【０１４４】
【数２１】

$$Y_j^0=\sum_{i:C^0(X_i)=j}X_i \qquad j=1,2,3,\ldots,n$$

【０１４５】
これらの初期重心を使用して、類似度モジュール726は、重心のそれぞれと、文書のそれぞれとの間の類似度距離D(X, Y)を生成することができる。
【０１４６】
初期設定の後、クラスタ化モジュール722は、各回で、現在、最も近いトピック重心に基づいて、文書を再割り当てすることができる。一実施例において、$D(X_{14}, Y_2)$が、現行の回に、j=1,2,...,nに関してすべての$D(X_{14}, Y_j)$の中で最小である場合、クラスタ化モジュール722は、ドキュメント14をトピック2に割り当てることができる。文書を再割り当てした後、重心モジュール724は、その新たな割り当てに基づいて、トピックの重心を更新する。例えば、ステップnで、重心モジュール724は、
【０１４７】
【数２２】

$$Y_j^n=\sum_{i:C^n(X_i)=j}X_i \qquad j=1,2,3,\ldots,n$$

【０１４８】
を計算することによって、新たな重心を計算することができる。
【０１４９】
更新された重心を使用して、類似度モジュール726は、文書と、更新された重心の間の

新たな類似度距離を算出することができる。次に、これらの算出された距離を使用して、次の回に文書が再割り当てされることが可能である。例えば、トピック分類モジュール702が、トピック文書クラスタが収束するまで、文書をクラスタに割り当てる動作、トピック重心を更新する動作、および更新された重心と文書の間の距離を計算する動作を繰り返し実行することができる。例えば、現行の回で(例えば、第n回で)、クラスタ化モジュール722が、前のステップで(例えば、第n-1回で)計算された距離を使用して、文書を或るトピックに割り当てることができる。一実施例において、クラスタ化モジュール722は、式

【０１５０】

【数２３】

$$C^n(X_i) = \arg\min_{j=1}^{n} D\left(X_i, Y_j^{n-1}\right)$$

【０１５１】

を使用して、X_iをクラスタ$C^n(X_i)$(例えば、第nのステップにおけるX_iの割り当てられたクラスタ)に再割り当てすることができる。

【０１５２】

トピック分類モジュール702は、重心の位置が収束するまで、これらの動作を繰り返すことができる。一実施例において、トピック分類モジュール702は、

【０１５３】

【数２４】

$$\left\| Y_j^n - Y_j^{n-1} \right\| < L$$

【０１５４】

である場合、重心Y_jの位置が収束すると判定することができ、ただし、Lは、正の実数である。

【０１５５】

別の実施形態において、文書は、人間による注釈、例えば、トピックIDと関係する注釈またはメタデータに従って、初期クラスタに割り当てられることが可能である。別の実施形態において、トピックキーワードリストが、文書クラスタおよびトピッククラスタの識別のために各トピッククラスタに種を入れるのに使用されることが可能である。また、他のクラスタ化技術が使用されることも可能である。

【０１５６】

トピック文書クラスタが生成された後、トピック語処理モジュール704が、これらの文書クラスタの中の候補トピック語を選択する。例えば、トピック語処理モジュール704は、トピック文書クラスタの各クラスタからの1つまたは複数の非トピック語および新たな語を、候補トピック語として識別することができる。

【０１５７】

相違値モジュール732が、或るトピックにおける語の語相違値を算出する。一部の実施形態において、トピック語分類モジュール704が、選択されたトピックおよびトピック語に関するトピック語相違値を算出することができる。例えば、トピック語処理モジュール704は、選択されたトピックのトピック辞書からトピック語を選択することができる。いくつかの実施形態において、相違値モジュール732は、文書コーパス710、ならびに選択されたトピックのトピック文書クラスタに属する文書におけるトピック語分布に基づいて、トピック語相違値を算出することができる。例えば、トピック語相違値は、或るトピックに関するトピック文書におけるトピック語の確率分布と、文書コーパス710の中のすべての文書に関するトピック語の確率分布との比に実質的に比例することができる。一実施例において、トピック語wのトピック語相違値Qが、

【０１５８】
【数２５】

$$Q = \frac{P_d(w)}{P(w)} \cdot \log P_d(w)$$

【０１５９】
によって算出されることが可能であり、ただし、$P_d(w)$は、文書コーパス710の中のトピックdと関係する文書における選択されたトピック語wの確率であり、さらにP(w)は、文書コーパス710の中のすべての文書における選択されたトピック語の確率である。
【０１６０】
閾値評価モジュール734が、1つまたは複数のトピック語相違値に基づいて、トピック相違値を算出することができる。一部の実施形態において、閾値評価モジュール734は、トピック語相違値の中心傾向に基づいて、トピック相違値を算出することができる。例えば、閾値評価モジュール734は、トピック語相違値の平均値を計算し、さらにこの平均値をトピック相違値として使用することができる。また、トピック語相違値に基づく他の値が使用されることも可能である。例えば、閾値評価モジュール734は、算出されたトピック語相違値を比較すること、およびそれらのトピック語相違値の最大値を、トピック相違値として選択することによって、トピック相違値を算出することができる。
【０１６１】
一部の実施形態において、閾値評価モジュール734は、トピック相違値をスケール変更することができる。例えば、閾値評価モジュール734は、式
T=(1+t)・S
に従ってトピック相違値をスケール変更することができ、ただし、Tは、スケール変更されたトピック相違値であり、tは、実数であり、さらにSは、トピック相違値である。
【０１６２】
同様に、相違値モジュール732は、候補トピック語の候補語相違値を算出することができる。トピックに関する候補トピック語は、既存の語、またはそのトピックに関するトピック辞書の中のトピック語ではない新たな語である。候補語相違値は、文書コーパス710、および選択されたトピックのトピック文書クラスタに属する文書における候補トピック語の確率分布に基づくことが可能である。一実施例において、候補トピック語w_cの候補トピック語相違値Rは、
【０１６３】
【数２６】

$$R = \frac{P_d(w_c)}{P(w_c)} \cdot \log P_d(w_c)$$

【０１６４】
によって算出されることが可能であり、ただし、$P_d(w_c)$は、文書コーパス710の中のトピックdと関係する文書における候補トピック語w_cの確率であり、さらに$P(w_c)$は、文書コーパス710のすべての文書における候補トピック語の確率である。
【０１６５】
トピック語処理モジュール704は、トピック相違値および候補語相違値に基づいて、候補トピック語がトピック語であるかどうかを判定することができる。例えば、候補相違値がトピック相違値と比較されて、候補トピック語がトピック語であるかどうかが判定されることが可能である。或る実施形態において、閾値評価モジュール734は、R>Sである、すなわち、
【０１６６】

【数２７】

$$\frac{P_d(w_c)}{P(w_c)} \cdot \log P_d(w_c) > S$$

【０１６７】
である場合、候補トピック語w_cがトピック語であると判定し、ただし、Sは、トピック相違値である。
【０１６８】
代替として、Tのスケール変更された値は、候補語相違値Rと比較されることも可能であり、ただし、$T=(1+t)^*S$である。別の実施形態において、Tの値は、対応するトピックの具体性に応じて、さらにスケール変更されることが可能である。例えば、非常に一般的なトピック、例えば、「スポーツ」というトピックに関して、Tの値は、トピック語の判定が、より包含的であるように、Sよりはるかに小さい大きさにスケール変更されることが可能である。逆に、非常に具体的なトピック、例えば、「ウェーブレット数学」に関して、Tの値は、トピック語の判定が、それほど包含的ではないように、実質的にS以上である大きさにスケール変更されることが可能である。また、他のスケール変更技術が、使用されることも可能である。
【０１６９】
候補トピック語が、或るトピックに関するトピック語であると判定された場合、辞書アップデータモジュール706は、その候補トピック語を含めるように、そのトピックに関するトピック辞書708を更新する。例えば、閾値評価モジュール734が、既存の語である候補トピック語、Weが、例えば、トピック2のトピック語であると判定した場合、トピック語処理モジュール704が、候補トピック語、Weをトピック2辞書の中に格納するよう辞書アップデータモジュール706に通知することができる。同様に、閾値評価モジュール734が、新たな語である候補トピック語、Wnが、例えば、トピックnのトピック語であると判定した場合、トピック語処理モジュール704が、候補トピック語、Wnをトピックn辞書の中に格納するよう辞書アップデータモジュール706に通知することができる。
【０１７０】
相違値と関係する他の関数が、使用されることも可能である。例えば、一対の単調関数f(x)およびg(x)、例えば、
【０１７１】
【数２８】

$$Q = f\left[\frac{P_d(w)}{P(w)}\right] \cdot g[P_d(w)]$$

【０１７２】
を使用して、相違値Qが算出されることが可能である。前述の例示的な実施形態において、f(x)=xであり、さらにg(x)=log(x)である。しかし、他の単調関数が使用されることも可能である。
【０１７３】
図8は、トピック語を識別するための例示的なプロセス800の流れ図である。プロセス800は、図7Aおよび図7Bのシステム700を実施する1つまたは複数のコンピュータを含むシステムにおいて実施されることが可能である。一部の実施例において、トピック語処理モジュール704は、語コーパス204から候補トピック語を識別し、さらにプロセス800を使用して、その候補トピック語が新たなトピック語であるかどうかを判定することができる。
【０１７４】
プロセス800は、トピック相違値を算出する(802)。例えば、相違値モジュール732が、選択されたトピックの1つまたは複数のトピック語相違値に基づいて、トピックのトピッ

ク相違値を算出することができる。一部の実施形態において、トピック相違値は、トピック文書コーパスにおける第1のトピック語分布(例えば、トピック文書コーパスにおけるトピック語の分布)の、文書コーパスにおける第2のトピック語分布(例えば、文書コーパス710におけるトピック語の分布)に対する比に実質的に比例することが可能である。トピック文書コーパスは、或るトピックと関係するトピック文書のコーパス、例えば、文書コーパス710の中の文書のサブセットであることが可能であり、さらに文書コーパスは、トピック文書、およびその他の文書を含む文書のコーパス、例えば、文書コーパス710であることが可能である。

【０１７５】

次に、プロセス800は、候補トピック語に関する候補トピック語相違値を算出する(804)。一部の実施形態において、候補トピック語相違値は、トピック文書コーパスにおける候補トピック語の第1の分布の、文書コーパスにおける候補トピック語の第2の分布に対する比に実質的に比例することが可能である。例えば、相違値モジュール732は、

【０１７６】

【数２９】

$$R = \frac{P_d(w_c)}{P(w_c)} \cdot \log P_d(w_c)$$

【０１７７】

を計算することによって、候補トピック語相違値Rを算出することができ、ただし、W_cは、候補トピック語であり、$P_d(W_c)$は、トピック文書コーパスにおける候補トピック語wの確率であり、さらに$P(W_c)$は、文書コーパス710における候補トピック語の確率である。

【０１７８】

トピック相違値および候補語相違値を算出した後、プロセス800は、候補トピック語相違値がトピック相違値を超えているかどうかを判定する(806)。例えば、トピック語処理モジュール704が、候補トピック語相違値とトピック相違値を比較することができる。

【０１７９】

候補トピック語相違値が、トピック相違値より大きい場合、プロセス800は、その候補トピック語を新たなトピック語として識別する(808)。例えば、候補トピック語相違値が、トピック相違値より大きい場合、トピック語処理モジュール704が、その候補トピック語が新たなトピック語であると判定することができる。

【０１８０】

候補トピック語相違値が、トピック相違値より大きくはない場合、プロセス800は、その候補トピック語を新たなトピック語として識別しない(810)。例えば、候補トピック語相違値が、トピック相違値より大きくはない場合、トピック語処理モジュール704が、その候補トピック語が新たなトピック語ではないと判定することができる。

【０１８１】

図9は、トピック語相違値を算出するための例示的なプロセス900の流れ図である。プロセス900は、図7Aおよび図7Bのシステム700を実施する1つまたは複数のコンピュータを含むシステムにおいて実施されることが可能である。一部の実施形態において、相違値モジュール732が、プロセス900を使用してトピック相違値を算出することができる。

【０１８２】

プロセス900は、トピック語を選択する(902)。例えば、相違値モジュール732が、トピック714のうちの1つのトピックから1つまたは複数のトピック語を選択することができる。

【０１８３】

次に、プロセス900は、それらのトピック語のそれぞれに関するトピック語相違値を算出する(904)。例えば、各トピック語相違値は、トピック文書コーパスにおける各トピック語の第1の分布の、文書コーパスにおける各トピック語の第2の分布に対する比に実質的

に比例する。一実施例において、相違値モジュール732は、

【０１８４】

【数３０】

$$Q = \frac{P_d(w)}{P(w)} \cdot \log P_d(w)$$

【０１８５】

を計算することによって、選択されたトピック語(w)のそれぞれに関するトピック語相違値を算出することができ、ただし、$P_d(w)$は、トピックdにおける選択されたトピック語wの確率であり、さらにP(w)は、文書コーパスにおける選択されたトピック語の確率である。

【０１８６】

トピック語相違値を算出した後、プロセス900は、トピック語相違値の中心傾向に基づいて、トピック相違値を算出する(906)。例えば、相違値モジュール732が、トピック語相違値の平均を算出することによって、トピック相違値を算出することができる。

【０１８７】

図10は、例示的な文書/語クラスタ化プロセス1000の流れ図である。プロセス1000は、図7Aおよび図7Bのシステム700を実施する1つまたは複数のコンピュータを含むシステムにおいて実施されることが可能である。

【０１８８】

プロセス1000は、文書コーパスの中でトピックと関係する文書を識別する(1002)。例えば、トピック分類モジュール702が、文書のTF-IDFベクトルと、トピックの重心ベクトルとの間の距離に基づいて、文書コーパス710の中の文書がトピック714の1つと関係していると識別することができる。一実施例において、トピック分類モジュール702は、図7Bを参照して説明されるとおり、繰り返しプロセスを使用して、文書を識別することができる。

【０１８９】

プロセス1000は、トピックと関係する文書クラスタを生成する(1004)。文書とトピックの間の識別された関係に基づき、トピック分類モジュール702は、トピックと関係する文書を文書クラスタの中に含めることによって、各トピックに関する文書クラスタを生成することができる。

【０１９０】

次に、プロセス1000は、文書クラスタのそれぞれの中の語を識別する(1006)。例えば、トピック語処理モジュール704が、トピック辞書708および/または新語データストア712を使用して、トピック文書クラスタのそれぞれの中のトピック語、非トピック語、および/または新たな語を識別することができる。

【０１９１】

プロセス1000は、文書クラスタのそれぞれの中の識別された語から候補トピック語を選択する(1008)。例えば、トピック語処理モジュール704が、文書コーパス710の中の識別されたトピック文書クラスタから候補トピック語を選択することができる。

【０１９２】

図11は、トピック語を識別するための別の例示的なプロセスの流れ図である。プロセス1100は、図7Aおよび図7Bのシステム700を実施する1つまたは複数のコンピュータを含むシステムにおいて実施されることが可能である。一部の実施形態において、トピック分類モジュール704が、プロセス1100における動作の一部またはすべてを使用して、新たなトピック語を識別することができる。

【０１９３】

プロセス1100は、或るトピックと関係するトピック語を備えるトピック辞書を選択する(1102)。例えば、トピック分類モジュール704が、選択されたトピック(例えば、トピック

1、トピック2、...またはトピックn)と関係するトピック辞書708の1つを選択することができる。

【０１９４】

プロセス1100は、トピック語、文書コーパス、およびトピック文書コーパスに基づいて、トピック語相違値を算出する(1104)。例えば、トピック文書コーパスが、トピック分類モジュール702によって生成されたトピック文書クラスタの1つに属する文書を備えることが可能である。トピック分類モジュール704が、選択されたトピック辞書から或るトピック語を選択することができる。このトピック語、ならびに文書クラスタおよび文書コーパスにおける、このトピック語のトピック語分布を使用して、相違値モジュール732が、トピック語相違値を算出することができる。例えば、相違値モジュール732は、選択されたトピックにおける、選択されたトピック語の確率、および文書コーパス710における、選択されたトピック語の確率に基づいて、トピック語相違値を計算することができる。

【０１９５】

プロセス1100が、文書コーパスおよびトピック文書コーパスに基づいて、候補トピック語に関する候補トピック語相違値を算出する(1106)。例えば、相違値モジュール732が、或る候補トピック語を選択し、さらに選択されたトピックにおける、選択された候補トピック語の確率、および文書コーパス710における、選択された候補トピック語の確率に基づいて、候補トピック語相違値を計算することによって、候補トピック語相違値を算出することができる。

【０１９６】

プロセス1100は、候補トピック語相違値がトピック語相違値より大きいかどうかを判定する(1108)。例えば、トピック分類モジュール704が、候補トピック語相違値とトピック語相違値を比較することができる。

【０１９７】

候補トピック語相違値が、トピック語相違値より大きい場合、その候補トピック語は、新たなトピック語であると判定される(1110)。例えば、トピック語処理モジュール704が、候補トピック語相違値がトピック語相違値より大きいと判定した場合、その候補トピック語は、新たなトピック語である。

【０１９８】

候補トピック語相違値が、トピック語相違値より大きくはない場合、候補トピック語は、新たなトピック語であるとは判定されない(1112)。例えば、トピック語処理モジュール704が、候補トピック語相違値がトピック語相違値より大きいと判定した場合、その候補トピック語は、新たなトピック語ではない。

【０１９９】

システム200によって新たな語として識別された3文字の語/句および4文字の語/句を再び参照すると、システム700は、各語を候補トピック語として識別し、さらに前述したとおり相違値を算出することができる。例示的な評価において、「丁俊暉」(ding junhui)、「本賽季」(今季)、「世錦賽」(世界選手権)、「季后賽」(プレーオフ)、「范甘迪」(Van Cundy)、および「国際足聯」(FIFA)という語が、スポーツトピックに割り当てられることが可能であり、さらに「反傾鎖」(アンチ低価格ダンピング)、「浄利潤」(純利益)、「証監会」(SEC)、「国資委」(中国国有資産監督管理委員会)、「美聯儲」(FED)、および「非流通股」(非取引株式)という語が、金融トピックに割り当てられることが可能である。

【０２００】

本明細書で説明される主題および機能上の動作の実施形態は、本明細書で開示される構造体、および構造上の均等物、あるいはこれらの1つまたは複数の構造体の組合せを含む、デジタル電子回路において、あるいはコンピュータソフトウェア、コンピュータファームウェア、またはコンピュータハードウェアにおいて実施されることが可能である。本明細書で説明される主題の実施形態は、1つまたは複数のコンピュータプログラム製品として、すなわち、データ処理装置によって実行されるように、またはデータ処理装置の動作

を制御するように実体のあるプログラムキャリア上に符号化されたコンピュータプログラム命令の1つまたは複数のモジュールとして実施されることが可能である。実体のあるプログラムキャリアは、伝搬される信号、またはコンピュータ可読媒体であることが可能である。伝搬される信号は、コンピュータによって実行されるように適切な受信機装置に伝送するために情報を符号化するように生成された、人工的に生成された信号、例えば、マシンによって生成された電気信号、光信号、または電磁信号である。コンピュータ可読媒体は、マシン可読ストレージデバイス、マシン可読ストレージ基板、メモリデバイス、マシン可読の伝搬される信号をもたらす材料の合成、または以上の1つまたは複数の要素の組合せであることが可能である。

【０２０１】

「データ処理装置」という用語は、例として、プログラマブルプロセッサ、コンピュータ、または複数のプロセッサもしくはコンピュータを含む、データを処理するためのすべての装置、デバイス、およびマシンを包含する。装置は、ハードウェアに加えて、当該のコンピュータプログラムのための実行環境を作るコード、例えば、プロセッサファームウェア、プロトコルスタック、データベース管理システム、オペレーティングシステム、または以上の1つまたは複数の要素の組合せを構成するコードを含むことが可能である。

【０２０２】

コンピュータプログラム(プログラム、ソフトウェア、ソフトウェアアプリケーション、スクリプト、またはコードとしても知られる)は、コンパイルされる言語もしくは解釈される言語、または宣言型言語もしくは手続き型言語を含む、任意の形態のプログラミング言語で書かれることが可能であり、さらにコンピュータプログラムは、スタンドアロンのプログラムとして、あるいはモジュール、コンポーネント、サブルーチン、またはコンピューティング環境において使用されるのに適した他のユニットとしての形態を含め、任意の形態で展開されることが可能である。コンピュータプログラムは、ファイルシステムにおけるファイルに必ずしも対応しない。プログラムは、他のプログラムもしくはデータを保持するファイルの一部分(例えば、マークアップ言語文書の中に格納された1つまたは複数のスクリプト)の中に、当該のプログラムに専用の単一のファイルの中に、または調整された複数のファイル(例えば、1つまたは複数のモジュール、サブプログラム、またはコードの部分を格納する複数のファイル)の中に格納されることが可能である。コンピュータプログラムは、1つのコンピュータ上で、あるいは1つのサイトに配置された、または複数のサイトにわたって分散されて、通信ネットワークによって互いに接続された複数のコンピュータ上で実行されるように展開されることが可能である。

【０２０３】

本明細書で説明されるプロセスおよび論理フローは、入力データを操作すること、および出力を生成することによって機能を実行するように1つまたは複数のコンピュータプログラムを実行する1つまたは複数のプログラマブルプロセッサによって実行されることが可能である。また、これらのプロセスおよび論理フローは、専用論理回路、例えば、FPGA(フィールドプログラマブルゲートアレイ)またはASIC(特定用途向け集積回路)によって実行されることも可能であり、さらに装置が、専用論理回路、例えば、FPGA(フィールドプログラマブルゲートアレイ)またはASIC(特定用途向け集積回路)として実施されることも可能である。

【０２０４】

コンピュータプログラムの実行に適したプロセッサには、例として、汎用マイクロプロセッサと専用マイクロプロセッサの両方、および任意の種類のデジタルコンピュータの任意の1つまたは複数のプロセッサが含まれる。一般に、プロセッサは、読み取り専用メモリまたはランダムアクセスメモリから、あるいはこの両方から命令およびデータを受け取る。コンピュータの基本的な要素は、命令を実行するためのプロセッサ、ならびに命令およびデータを格納するための1つまたは複数のメモリデバイスである。一般に、コンピュータは、データを格納するための1つまたは複数の大容量ストレージデバイス、例えば、磁気ディスク、光磁気ディスク、または光学ディスクも含み、あるいはそのようなデバイ

スからデータを受け取る、またはそのようなデバイスにデータを転送する、あるいはその両方を行うように動作上、結合される。しかし、コンピュータは、そのようなデバイスを有さなくてもよい。さらに、コンピュータは、別のデバイス、例えば、いくつかだけを挙げると、移動電話機、PDA(パーソナルデジタルアシスタント)、移動オーディオプレーヤもしくは移動ビデオプレーヤ、ゲームコンソール、GPS(全地球測位システム)受信機に埋め込まれることが可能である。

【０２０５】

コンピュータプログラム命令およびデータを格納するのに適したコンピュータ可読媒体は、例として、半導体メモリデバイス、例えば、EPROM、EEPROM、およびフラッシュメモリデバイス、磁気ディスク、例えば、内部ハードディスクまたはリムーバブルディスク、光磁気ディスク、ならびにCD-ROMディスクおよびDVD-ROMディスクを含む、すべての形態の不揮発性のメモリ、媒体、およびメモリデバイスを含む。プロセッサおよびメモリは、専用論理回路によって補足される、または専用論理回路に組み込まれることが可能である。

【０２０６】

ユーザとの対話をもたらすのに、本明細書で説明される主題の実施形態は、ユーザに情報を表示するためのディスプレイデバイス、例えば、CRT(陰極線管)モニタまたはLCD(液晶ディスプレイ)モニタ、ならびにユーザがコンピュータに入力を与えることができるキーボードおよびポインティングデバイス、例えば、マウスまたはトラックボールを有するコンピュータ上で実施されることが可能である。他の種類のデバイスが、ユーザとの対話をもたらすのに使用されることも可能であり、例えば、ユーザに与えられるフィードバックは、任意の形態の知覚フィードバック、例えば、視覚的フィードバック、聴覚的フィードバック、または触覚的フィードバックであることが可能であり、さらにユーザからの入力は、音響入力、音声入力、または触覚入力を含む任意の形態で受け取られることが可能である。

【０２０７】

本明細書で説明される主題の実施形態は、例えば、データサーバなどのバックエンド構成要素を含む、またはミドルウェア構成要素、例えば、アプリケーションサーバを含む、またはフロントエンド構成要素、例えば、ユーザが、本明細書で説明される主題の実施形態と対話することができるグラフィカルユーザインタフェースまたはウェブブラウザを有するクライアントコンピュータを含む、あるいはそのようなバックエンド構成要素、ミドルウェア構成要素、またはフロントエンド構成要素の任意の組合せを含むコンピューティングシステムにおいて実施されることが可能である。システムのこれらの構成要素は、任意の形態もしくは媒体のデジタルデータ通信、例えば、通信ネットワークによって互いに接続されることが可能である。通信ネットワークの例には、ローカルエリアネットワーク(「LAN」)およびワイドエリアネットワーク(「WAN」)、例えば、インターネットが含まれる。

【０２０８】

コンピューティングシステムは、クライアントおよびサーバを含むことが可能である。クライアントおよびサーバは、一般に、互いに遠隔であり、さらに通常、通信ネットワークを介して対話する。クライアントとサーバの関係は、それぞれのコンピュータ上で実行され、さらに互いにクライアント/サーバ関係を有するコンピュータプログラムのお陰で生じる。

【０２０９】

本明細書は、多くの特定の実施上の詳細を含むが、これらの詳細は、任意の本発明の範囲、または主張される可能性があることの限定として解釈されるべきではなく、むしろ、特定の発明の特定の実施形態に固有である可能性がある特徴の説明として解釈されるべきである。別々の実施形態の文脈において本明細書で説明されるいくつかの特徴は、単一の実施形態において組合せで実施されることも可能である。逆に、単一の実施形態の文脈において説明される様々な特徴が、複数の実施形態において別々に、または任意の適切な部

分的組合せで実施されることも可能である。さらに、特徴は、いくつかの組合せで作用するものとして前段で説明される可能性があり、さらに当初、そのようなものとして主張さえされる可能性があるものの、主張される組合せからの1つまたは複数の特徴は、一部の事例において、その組合せから切り離されることが可能であり、さらに主張される組合せは、部分的組合せまたは部分的組合せの変種に向けられることも可能である。

【０２１０】

同様に、動作は、図面において或る特定の順序で示されるが、このことは、望ましい結果を実現するのに、そのような動作が、示される特定の順序、または順番に実行されること、または図示されるすべての動作が実行されることを要求するものと理解されるべきではない。いくつかの状況において、マルチタスキングおよび並行処理が、有利である可能性がある。さらに、前段で説明される実施形態における様々なシステム構成要素の分離は、すべての実施形態においてそのような分離を要求するものと解釈されるべきではなく、さらに説明されるプログラム構成要素およびプログラムシステムは、一般に、単一のソフトウェア製品の中に一緒に統合される、または複数のソフトウェア製品にパッケージ化されることが可能であることを理解されたい。

【０２１１】

本明細書で説明される主題の特定の実施形態が、説明されてきた。その他の実施形態も、添付の特許請求の範囲に含まれる。例えば、特許請求の範囲に記載されるアクションは、異なる順序で実行されて、それでも、望ましい結果を実現することが可能である。一例として、添付の図に示されるプロセスは、望ましい結果を実現するのに、示される特定の順序、または順番を必ずしも要求しない。いくつかの実施形態において、マルチタスキングおよび並行処理が、有利である可能性がある。

【符号の説明】

【０２１２】

100 デバイス
101 インプットメソッドエディタコード
102 処理デバイス
103 インプットメソッドエディタインスタンス
104 データストア
105 アプリケーションソフトウェア
106 データストア
107 アプリケーションインスタンス
108 入力デバイス
110 出力デバイス
112 ネットワークインタフェース
114 バスシステム
116 ネットワーク
118 コンピューティングシステム
120 インプットメソッドエディタシステム
122 インプットメソッドエディタエンジン
124 辞書
126 構成入力データストア
128 辞書エントリ
200 語検出システム
202 ワイドエリアネットワーク
204 語コーパス
206 語処理モジュール
208 新語アナライザモジュール
210 辞書アップデータモジュール
212 パーティションデータストア

214　ウェブ文書
216　電子通信
218　データストア
220　語ソース
232　訓練コーパス
234　開発コーパス
700　トピック語識別システム
702　トピック分類モジュール
704　トピック語処理モジュール
706　辞書アップデータモジュール
708　トピック辞書
710　文書コーパス
712　新たな語
714　トピック
722　クラスタ化モジュール
724　重心モジュール
726　類似度モジュール
732　相違値モジュール
734　閾値評価モジュール

【図１Ａ】

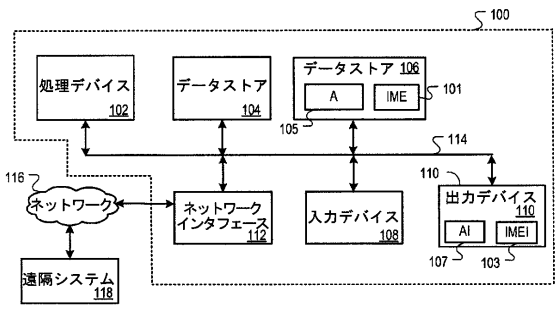

FIG. 1A

【図１Ｂ】

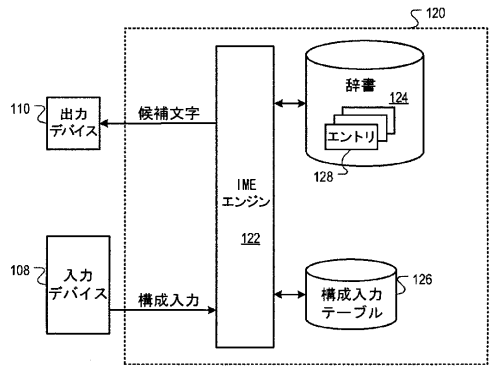

FIG. 1B

【図２Ａ】

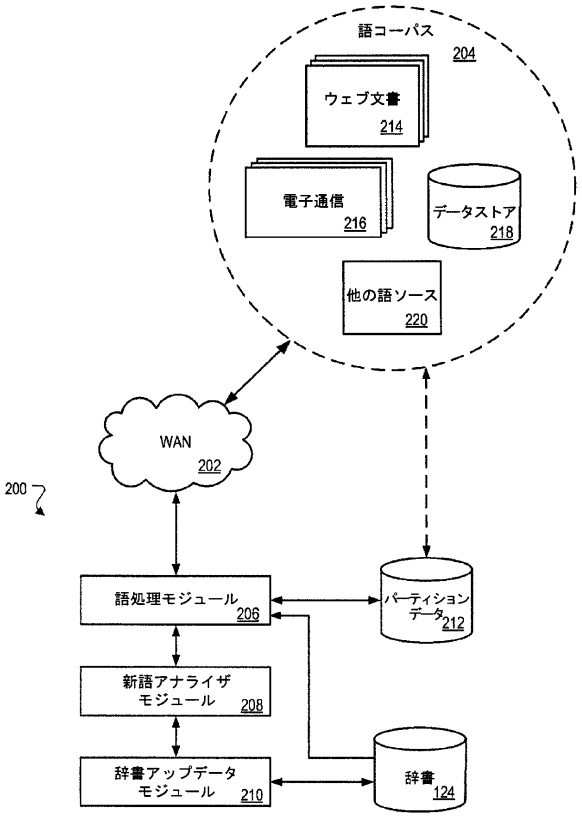

FIG. 2A

【図２Ｂ】

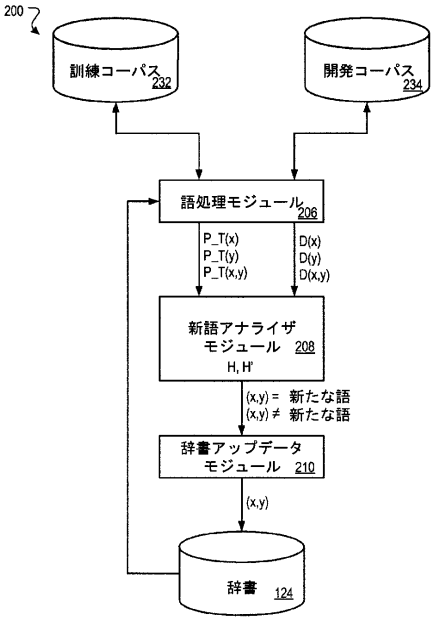

FIG. 2B

【図３】

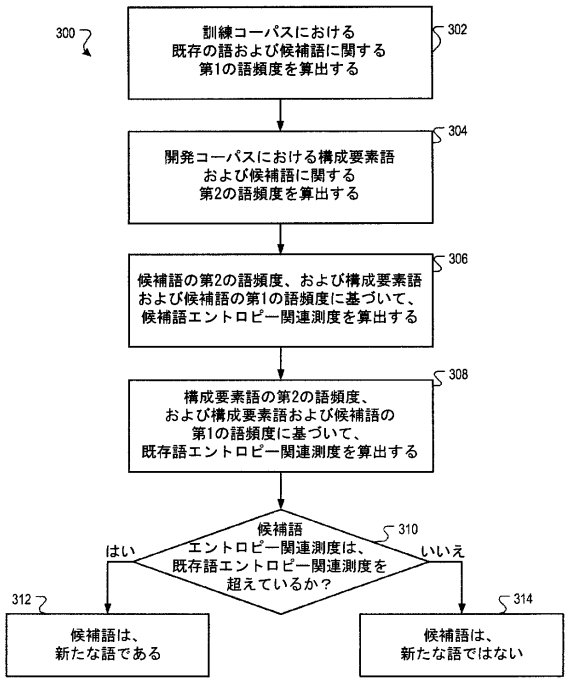

FIG. 3

【図４】

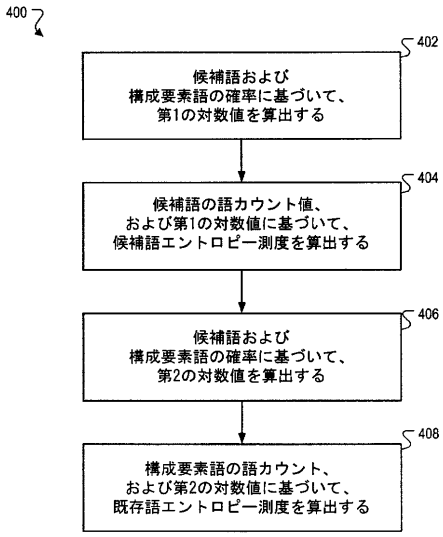

FIG. 4

【図５】

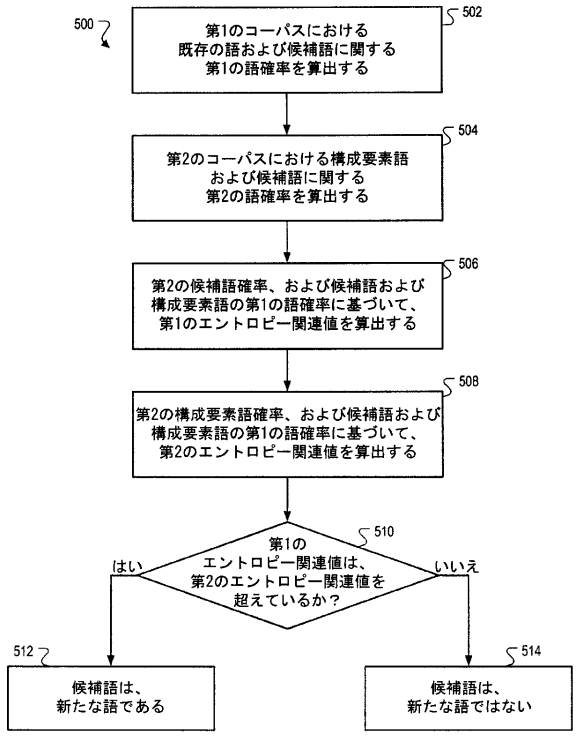

FIG. 5

【図６】

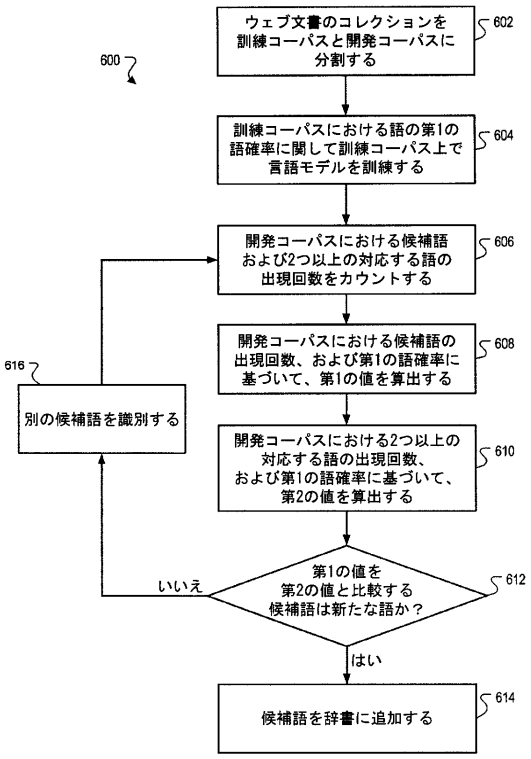

FIG. 6

【図７Ａ】

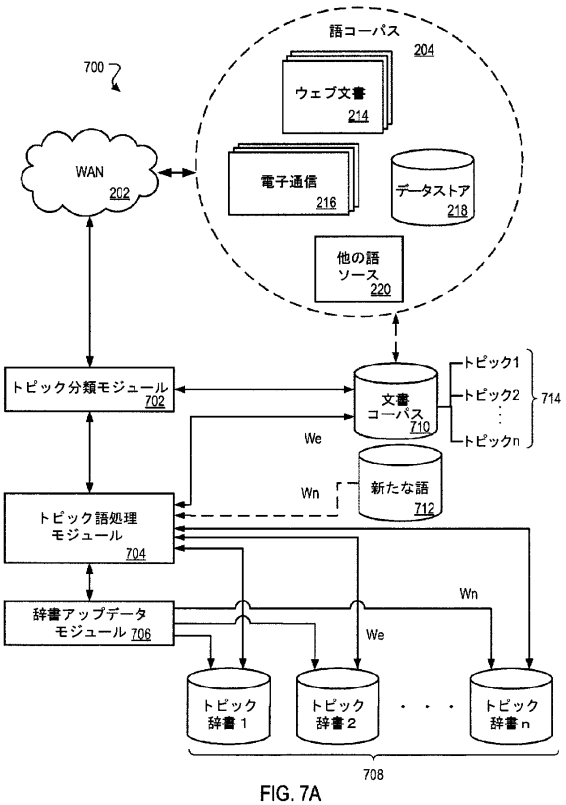

FIG. 7A

【図７Ｂ】

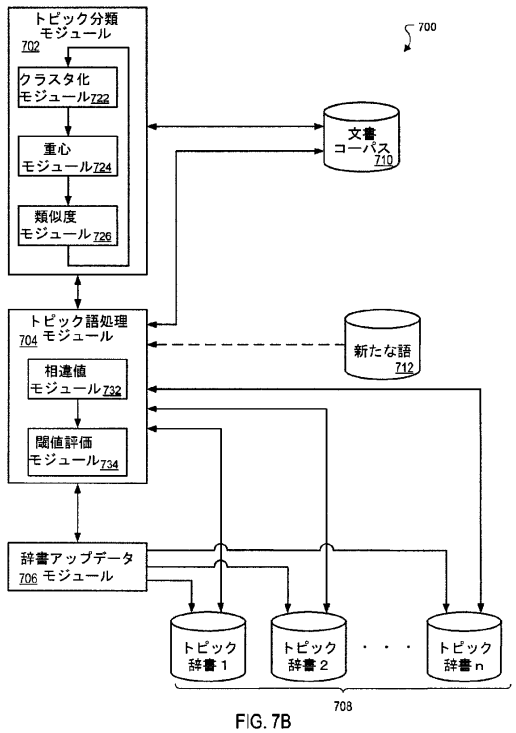

FIG. 7B

【図８】

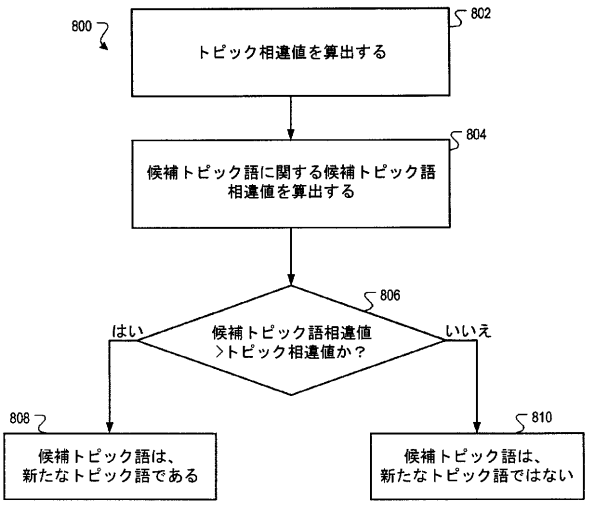

FIG. 8

【図９】

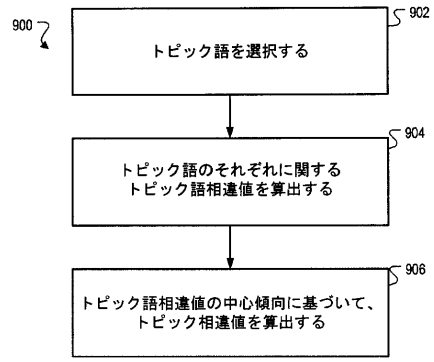

FIG. 9

【図１０】

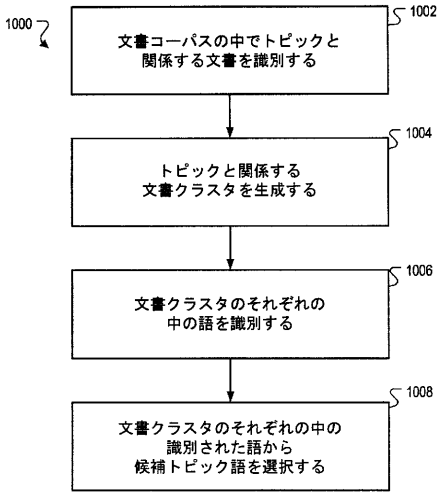

FIG. 10

【図１１】

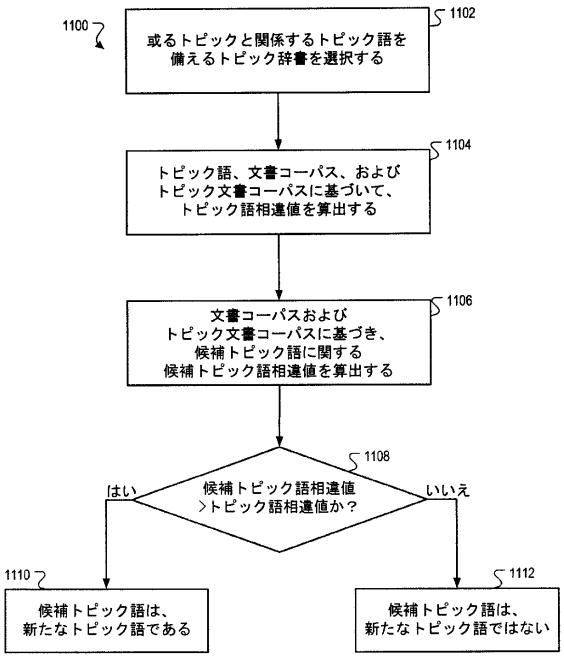

FIG. 11

フロントページの続き

(72)発明者　ジュン・ウ
アメリカ合衆国・カリフォルニア・９５０７０・サラトガ・フランクリン・アヴェニュー・２０３９８
(72)発明者　タン・シー・リウ
中華人民共和国・１０００８４・ベイジン・ハイディアン・ディストリクト・ジョングアンツゥン・イースト・ロード・ナンバー・１・ツィンファ・サイエンス・パーク・ビルディング・６
(72)発明者　フェン・ホン
アメリカ合衆国・カリフォルニア・９４４０４・フォスター・シティー・ラム・レーン・８０７
(72)発明者　ヨンガン・ワン
中華人民共和国・１０００８８・ベイジン・ハイディアン・ディストリクト・ジーチュン・ロード・ナンバー・６・９－１９０６
(72)発明者　ボー・ヤン
中華人民共和国・１０００８４・ベイジン・ハイディアン・ディストリクト・ジョングアンツゥン・イースト・ロード・ナンバー・１・ツィンファ・サイエンス・パーク・ビルディング・６
(72)発明者　レイ・ジャン
中華人民共和国・１０００８１・ベイジン・ハイディアン・ジョングアンツゥン・ノース・ロード・ナンバー・１・ケーユアン・サブディストリクト・ビー５・６０６

審査官　梅本　達雄

(56)参考文献　特開平０３－２８６３７２（ＪＰ，Ａ）
国際公開第２００７／０１０９３６（WO，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／２０　－　１７／２８