

(19) 日本国特許庁(JP)

(12) 特許公報(B2)

(11) 特許番号

特許第5126737号
(P5126737)

(45) 発行日 平成25年1月23日(2013.1.23)

(24) 登録日 平成24年11月9日(2012.11.9)

(51) Int.Cl. F I
G06F 17/28 (2006.01) G O 6 F 17/28 Z
G06F 17/21 (2006.01) G O 6 F 17/21 5 5 0 A

請求項の数 8 (全 35 頁)

<p>(21) 出願番号 特願2007-165752 (P2007-165752) (22) 出願日 平成19年6月25日(2007.6.25) (65) 公開番号 特開2009-3818 (P2009-3818A) (43) 公開日 平成21年1月8日(2009.1.8) 審査請求日 平成22年4月19日(2010.4.19)</p> <p>特許法第30条第1項適用 平成19年3月28日 社団法人情報処理学会発行の「情報処理学会研究報告会情処研報 Vol. 2007, No. 35」に発表〔刊行物等〕平成19年3月31日 http://chasen.org/daiti-m/paper/n1178vpylm.pdf及びhttp://chasen.org/daiti-m/paper/n1178vpylm-slides.pdfを通じて発表</p>	<p>(73) 特許権者 301022471 独立行政法人情報通信研究機構 東京都小金井市貫井北町4-2-1 (74) 代理人 100115749 弁理士 谷川 英和 (72) 発明者 持橋 大地 東京都小金井市貫井北町4-2-1 独立行政法人情報通信研究機構内 (72) 発明者 隅田 英一郎 東京都小金井市貫井北町4-2-1 独立行政法人情報通信研究機構内 審査官 梅本 達雄</p>
---	---

最終頁に続く

(54) 【発明の名称】 情報処理装置、情報処理方法、及びプログラム

(57) 【特許請求の範囲】

【請求項1】

記号の並びを示すデータである訓練データが記憶される訓練データ記憶部と、前記訓練データに含まれる各記号に対応するグラム長を示す情報であるグラム長情報と、前記訓練データに含まれる各記号に対応するグラム長情報の示すグラム長より短いグラム長を有する代理の記号に関する情報である代理情報とが記憶される可変長情報記憶部と、前記訓練データと前記グラム長情報と前記代理情報とに対応する、前記訓練データに含まれる記号の接尾辞木を示す情報である接尾辞木情報が記憶される接尾辞木情報記憶部と、前記訓練データを用いて、前記接尾辞木情報を更新しながら前記各記号の前記グラム長情報と前記代理情報とをギブスサンプリングにより算出して前記可変長情報記憶部に蓄積する処理を繰り返して実行するギブスサンプリング処理を行うギブスサンプリング部と、を備え、
前記ギブスサンプリング部は、
前記訓練データに含まれるすべての記号をランダムに順次選択する選択手段と、
前記選択手段が選択した記号がNグラム(Nは0以上の整数)となる確率を算出する確率算出手段と、
前記確率算出手段が算出した確率の分布に応じて、前記選択手段が選択した記号に対応するグラム長を選択し、当該選択したグラム長を前記グラム長情報に設定すると共に、前記選択手段が選択した記号に対応する代理の記号のグラム長を、当該選択されたグラム長に応じて階層Pitman-Yor過程により決定し、当該決定したグラム長を前記代理情

10

20

報に設定する設定手段と、
前記選択手段が選択した記号に関する情報が削除されるように前記接尾辞木情報を更新すると共に、前記設定手段によるグラム長の選択及び代理の記号のグラム長の決定に応じて、当該選択されたグラム長の記号に関する情報が追加されるように前記接尾辞木情報を更新する接尾辞木情報更新手段と、
前記選択手段による訓練データに含まれるすべての記号のランダムな選択と、当該選択された記号に関するグラム長情報の設定、代理情報の設定、及び接尾辞木情報の更新とが、繰り返して実行されるように制御する制御手段と、を備え、
前記確率算出手段は、前記記号がNグラムとなる確率を、当該確率を算出する記号の階層 P i t m a n - Y o r 過程におけるNグラム確率と、前記確率を算出する記号以外の前記接尾辞木情報における記号によってNグラム長に到達する確率とを掛け合わせることでより算出する、 情報処理装置。

10

【請求項2】

記号の並びを示すデータであるテストデータが記憶されるテストデータ記憶部と、
 前記訓練データと、前記グラム長情報と、前記代理情報とに対応する接尾辞木情報を用いて、前記テストデータに含まれる記号の可変長Nグラム確率を、当該確率を算出する記号の階層 P i t m a n - Y o r 過程におけるNグラム確率と、前記確率を算出する記号がNグラム長に到達する確率との積を各Nについて足しあわせることによって算出する確率算出部と、
 前記確率算出部が算出した可変長Nグラム確率を出力する出力部と、をさらに備えた、請求項1記載の情報処理装置。

20

【請求項3】

前記ギブスサンプリング部は、前記確率算出部が前記テストデータに含まれる記号の確率を算出することに、ギブスサンプリング処理を実行し、
 前記確率算出部は、前記訓練データと、前記グラム長情報と、前記代理情報とに対応する接尾辞木情報を用いて、前記テストデータに含まれる記号の可変長Nグラム確率を、当該確率を算出する記号の階層 P i t m a n - Y o r 過程におけるNグラム確率と、前記確率を算出する記号がNグラム長に到達する確率との積を各Nについて足しあわせた値を、複数回のギブスサンプリングについて平均をとることによって算出する、請求項2記載の情報処理装置。

30

【請求項4】

前記訓練データと、前記グラム長情報と、前記代理情報とに対応する接尾辞木情報を用いて、記号の並びに含まれる記号の確率を、当該確率を算出する記号の階層 P i t m a n - Y o r 過程におけるNグラム確率と、前記確率を算出する記号がNグラム長に到達する確率とを掛け合わせることでより算出する確率算出部と、
 前記確率算出部が算出した複数の記号の確率において、他の記号の確率に比べて大きい確率を有する記号を含むK個（Kは、その記号の確率が算出された際のNグラム長の値である）の記号の並びである記号列を選択する記号列選択部と、
 前記記号列選択部が選択した記号列を出力する出力部と、をさらに備えた、請求項1記載の情報処理装置。

40

【請求項5】

前記確率算出部が確率を算出する記号の含まれる記号の並びは、前記訓練データに含まれる記号の並びである、請求項4記載の情報処理装置。

【請求項6】

前記記号は単語であり、前記訓練データは、単語の並びを示す文書である、請求項1から請求項5いずれか記載の情報処理装置。

【請求項7】

記号の並びを示すデータである訓練データが記憶される訓練データ記憶部と、前記訓練データに含まれる各記号に対応するグラム長を示す情報であるグラム長情報と、前記訓練データに含まれる各記号に対応するグラム長情報の示すグラム長より短いグラム長を有する

50

代理の記号に関する情報である代理情報とが記憶される可変長情報記憶部と、前記訓練データと前記グラム長情報と前記代理情報とに対応する、前記訓練データに含まれる記号の接尾辞木を示す情報である接尾辞木情報が記憶される接尾辞木情報記憶部と、ギブスサンプリング部と、を用いて処理される情報処理方法であって、

前記ギブスサンプリング部が、前記訓練データを用いて、前記接尾辞木情報を更新しながら、前記各記号の前記グラム長情報と前記代理情報とをギブスサンプリングにより算出して前記可変長情報記憶部に蓄積する処理を繰り返して実行するギブスサンプリング処理を行うギブスサンプリングステップを備え、

前記ギブスサンプリングステップは、

前記訓練データに含まれるすべての記号をランダムに順次選択する選択ステップと、

前記選択ステップで選択した記号がNグラム（Nは0以上の整数）となる確率を算出する確率算出ステップと、

前記確率算出ステップで算出した確率の分布に応じて、前記選択ステップで選択した記号に対応するグラム長を選択し、当該選択したグラム長を前記グラム長情報に設定すると共に、前記選択ステップで選択した記号に対応する代理の記号のグラム長を、当該選択されたグラム長に応じて階層Pitman-Yor過程により決定し、当該決定したグラム長を前記代理情報に設定する設定ステップと、

前記選択ステップで選択した記号に関する情報が削除されるように前記接尾辞木情報を更新すると共に、前記設定ステップにおけるグラム長の選択及び代理の記号のグラム長の決定に応じて、当該選択されたグラム長の記号に関する情報が追加されるように前記接尾辞木情報を更新する接尾辞木情報更新ステップと、

前記選択ステップにおける訓練データに含まれるすべての記号のランダムな選択と、当該選択された記号に関するグラム長情報の設定、代理情報の設定、及び接尾辞木情報の更新とが、繰り返して実行されるように制御する制御ステップと、を備え、

前記確率算出ステップでは、前記記号がNグラムとなる確率を、当該確率を算出する記号の階層Pitman-Yor過程におけるNグラム確率と、前記確率を算出する記号以外の前記接尾辞木情報における記号によってNグラム長に到達する確率とを掛け合わせる事により算出する、情報処理方法。

【請求項8】

コンピュータを、

訓練データ記憶部で記憶される、記号の並びを示すデータである訓練データを用いて、接尾辞木情報記憶部で記憶される、前記訓練データと前記訓練データに含まれる各記号に対応するグラム長を示す情報であり可変長情報記憶部で記憶されるグラム長情報と前記訓練データに含まれる各記号に対応するグラム長情報の示すグラム長より短いグラム長を有する代理の記号に関する情報であり前記可変長情報記憶部で記憶される代理情報とに対応する、前記訓練データに含まれる記号の接尾辞木を示す情報である接尾辞木情報を更新しながら、前記各記号の前記グラム長情報と前記代理情報とをギブスサンプリングにより算出して前記可変長情報記憶部に蓄積する処理を繰り返して実行するギブスサンプリング処理を行うギブスサンプリング部として機能させ、

前記ギブスサンプリング部は、

前記訓練データに含まれるすべての記号をランダムに順次選択する選択手段と、

前記選択手段が選択した記号がNグラム（Nは0以上の整数）となる確率を算出する確率算出手段と、

前記確率算出手段が算出した確率の分布に応じて、前記選択手段が選択した記号に対応するグラム長を選択し、当該選択したグラム長を前記グラム長情報に設定すると共に、前記選択手段が選択した記号に対応する代理の記号のグラム長を、当該選択されたグラム長に応じて階層Pitman-Yor過程により決定し、当該決定したグラム長を前記代理情報に設定する設定手段と、

前記選択手段が選択した記号に関する情報が削除されるように前記接尾辞木情報を更新すると共に、前記設定手段によるグラム長の選択及び代理の記号のグラム長の決定に応じて

10

20

30

40

50

、当該選択されたグラム長の記号に関する情報が追加されるように前記接尾辞木情報を更新する接尾辞木情報更新手段と、
前記選択手段による訓練データに含まれるすべての記号のランダムな選択と、当該選択された記号に関するグラム長情報の設定、代理情報の設定、及び接尾辞木情報の更新とが、
繰り返して実行されるように制御する制御手段と、を備え、
前記確率算出手段は、前記記号がNグラムとなる確率を、当該確率を算出する記号の階層
Pitman - Yor過程におけるNグラム確率と、前記確率を算出する記号以外の前記
接尾辞木情報における記号によってNグラム長に到達する確率とを掛け合わせることに
より算出する、プログラム。

【発明の詳細な説明】

10

【技術分野】

【0001】

本発明は、記号の並びを示す訓練データに基づいて、各記号について可変長のグラム長を設定する処理等を行う情報処理装置等に関する。

【背景技術】

【0002】

従来、単語間のマルコフ過程によって文の確率を計算するnグラムモデルが用いられてきていた。そのnグラムモデルでは、単語の種類数をV個とした場合に、状態数がVの(n-1)乗のオーダーとなるため、大きなnのnグラムモデルを扱うことは困難であった。通常、n=3(トライグラム)であり、n=4,5程度が限界であった。

20

【0003】

なお、近年、nグラム分布を階層的に生成する確率モデルが研究されてきている。その確率モデルでは、階層Pitman - Yor(ピットマン・ヨー)過程と呼ばれるノンパラメトリックな確率過程によって、適切にスムージングされたnグラム分布を0グラム 1グラム 2グラムとベイズ統計の枠組みから階層的に生成及び推定できることが知られている(例えば、非特許文献1、非特許文献2参照)。

【非特許文献1】Yee Whye Teh, 「A Bayesian Interpretation of Interpolated Kneser - Ney」 Technical Report TRA2/06, School of Computing, NUS, 2006年

30

【非特許文献2】Yee Whye Teh, 「A Hierarchical Bayesian Language Model based on Pitman - Yor Processes」 In Proc. of COLING/ACL 2006, p. 985 - 992, 2006年

【発明の開示】

【発明が解決しようとする課題】

【0004】

しかしながら、そのような階層Pitman - Yor過程によるnグラム言語モデルであっても、nグラムの「n」の値は固定であるため、種々の問題が発生する。例えば、現実の言語データには、3グラムや4グラムを超える長い系列が頻繁に現れる。具体的には、「the united states of america」や、「京都 大学 大学院 情報 学 研究科」等である。前述のように、通常、n=3程度であるため、そのような長い系列を扱うことができないという問題がある。また一方、nグラムの「n」の値を大きく設定すると、前述のように状態数が多くなるばかりではなく、ノイズも増えてしまうという問題もある。

40

【0005】

そのため、可変長nグラム言語モデルについての研究がなされてきているが、従来の可変長nグラム言語モデルは、巨大なnグラム言語モデルを生成し、その枝刈りをすることによって、可変長モデルにする方法がとられていた。そのような方法では、まず、巨大なnグラム言語モデルを生成する必要があるため、前述のように、やはりnの値には限界

50

があることになる。

【0006】

すなわち、一般的に言うと、記号の並びは、さまざまに長さの異なるnグラムから生成されていると考えられる。ここで、記号とは、単語や文字、バイオインフォマティクス等における塩基やアミノ酸等に対応する。したがって、そのような異なるnグラムの記号の並びを、前述の枝刈り等の処理によるのではなく、はじめから可変のnグラム長を許容する枠組みである可変長nグラムによって適切に記述することが期待されている。

【0007】

本発明は、上記問題点を解決するためになされたものであり、可変長nグラムを適切に扱うことができる情報処理装置等を提供することを目的とする。

10

【課題を解決するための手段】

【0008】

上記目的を達成するため、本発明による情報処理装置は、記号の並びを示すデータである訓練データが記憶される訓練データ記憶部と、前記訓練データに含まれる各記号に対応するグラム長を示す情報であるグラム長情報と、前記訓練データに含まれる各記号に対応するグラム長情報の示すグラム長より短いグラム長を有する代理の記号に関する情報である代理情報とが記憶される可変長情報記憶部と、前記訓練データと前記グラム長情報と前記代理情報とに対応する、前記訓練データに含まれる記号の接尾辞木を示す情報である接尾辞木情報が記憶される接尾辞木情報記憶部と、前記訓練データを用いて、前記接尾辞木情報を更新しながら前記各記号の前記グラム長情報と前記代理情報とをギブスサンプリングにより算出して前記可変長情報記憶部に蓄積する処理を繰り返して実行するギブスサンプリング処理を行うギブスサンプリング部と、を備えたものである。

20

【0009】

このような構成により、枝刈り等の行うことなく、可変長nグラムを適切に扱うことができる。その結果、可変長nグラムを許容したモデルを生成することができるようになる。例えば、記号が単語であり、訓練データが文書である場合に、言語は長さの異なる種々のnグラム文脈からなると考えられる。可変長nグラムを扱うことができることによって、そのような言語の特徴を適切に表すことのできるモデルを生成することができる。

【0010】

また、本発明による情報処理装置では、前記ギブスサンプリング部は、前記訓練データに含まれるすべての記号をランダムに順次選択する選択手段と、前記選択手段が選択した記号がNグラム(Nは0以上の整数)となる確率を算出する確率算出手段と、前記確率算出手段が算出した確率の分布に応じて、前記選択手段が選択した記号に対応するグラム長を選択し、当該選択したグラム長を前記グラム長情報に設定すると共に、前記選択手段が選択した記号に対応する代理の記号のグラム長を、当該選択されたグラム長に応じて階層Piman-Yor過程により決定し、当該決定したグラム長を前記代理情報に設定する設定手段と、前記選択手段が選択した記号に関する情報が削除されるように前記接尾辞木情報を更新すると共に、前記設定手段によるグラム長の選択及び代理の記号のグラム長の決定に応じて、当該選択されたグラム長の記号に関する情報が追加されるように前記接尾辞木情報を更新する接尾辞木情報更新手段と、前記選択手段による訓練データに含まれるすべての記号のランダムな選択と、当該選択された記号に関するグラム長情報の設定、代理情報の設定、及び接尾辞木情報の更新とが、繰り返して実行されるように制御する制御手段と、を備えてもよい。

30

40

【0011】

このような構成により、ギブスサンプリング処理によって、順次、各記号のグラム長を選択して設定し、また、代理の記号に関するグラム長を決定して設定し、それに応じて接尾辞木情報を更新していくことができ、訓練データの特徴を示すように、グラム長情報、代理情報、及び接尾辞木情報を変更していくことができる。

【0012】

また、本発明による情報処理装置では、前記確率算出手段は、前記記号がNグラムとな

50

る確率を、当該確率を算出する記号の階層 P i m a n - Y o r 過程における N グラム確率と、前記確率を算出する記号以外の前記接尾辞木情報における記号によって N グラム長に到達する確率とを掛け合わせるにより算出してもよい。

このような構成により、種々の N の値について、各記号が N グラムとなる確率を算出することができる。

【 0 0 1 3 】

また、本発明による情報処理装置では、記号の並びを示すデータであるテストデータが記憶されるテストデータ記憶部と、前記訓練データと、前記グラム長情報と、前記代理情報とに対応する接尾辞木情報を用いて、前記テストデータに含まれる記号の可変長 N グラム確率を、当該確率を算出する記号の階層 P i m a n - Y o r 過程における N グラム確率と、前記確率を算出する記号が N グラム長に到達する確率との積を各 N について足しあわせることによって算出する確率算出部と、前記確率算出部が算出した可変長 N グラム確率を出力する出力部と、をさらに備えてもよい。

10

【 0 0 1 4 】

このような構成により、テストデータについて、可変長 N グラム確率を算出することができる。前述のように、例えば、記号が単語であり、訓練データが文書である場合に、言語は長さの異なる種々の n グラム文脈からなると考えられる。可変長 n グラムを扱うことができることによって、そのような言語の特徴を適切に取り入れた可変長 N グラム確率を算出することができる。

【 0 0 1 5 】

20

また、本発明による情報処理装置では、前記ギブスサンプリング部は、前記確率算出部が前記テストデータに含まれる記号の確率を算出するごとに、ギブスサンプリング処理を実行し、前記確率算出部は、前記訓練データと、前記グラム長情報と、前記代理情報とに対応する接尾辞木情報を用いて、前記テストデータに含まれる記号の可変長 N グラム確率を、当該確率を算出する記号の階層 P i m a n - Y o r 過程における N グラム確率と、前記確率を算出する記号が N グラム長に到達する確率との積を各 N について足しあわせた値を、複数回のギブスサンプリングについて平均をとることによって算出してもよい。

このような構成により、複数回のギブスサンプリングについての平均をとることによって、より正確な可変長 N グラム確率を算出することができるようになる。

【 0 0 1 6 】

30

また、本発明による情報処理装置では、前記訓練データと、前記グラム長情報と、前記代理情報とに対応する接尾辞木情報を用いて、記号の並びに含まれる記号の確率を、当該確率を算出する記号の階層 P i m a n - Y o r 過程における N グラム確率と、前記確率を算出する記号が N グラム長に到達する確率とを掛け合わせるにより算出する確率算出部と、前記確率算出部が算出した複数の記号の確率において、他の記号の確率に比べて大きい確率を有する記号を含む K 個 (K は、その記号の確率が算出された際の N グラム長の値である) の記号の並びである記号列を選択する記号列選択部と、前記記号列選択部が選択した記号列を出力する出力部と、をさらに備えてもよい。

【 0 0 1 7 】

このような構成により、慣用的に用いられている記号の並びを示す記号列を選択して出力することができる。例えば、記号が単語であり、訓練データが文書である場合に、慣用的なフレーズや、熟語等を抽出することが可能となる。特に、可変長 n グラムを扱うことによって、そのフレーズや熟語の単語の個数が限定されないというメリットがある。

40

【 0 0 1 8 】

また、本発明による情報処理装置では、前記確率算出部が確率を算出する記号の含まれる記号の並びは、前記訓練データに含まれる記号の並びであってもよい。

このような構成により、一般に規模が大きいと考えられる訓練データを用いて、確率を算出することができ、より多くの適切な記号列を選択することができると考えられる。

【 発明の効果 】

【 0 0 1 9 】

50

本発明による情報処理装置等によれば、可変長 n グラムを適切に扱うことができる情報処理装置等を提供することができる。したがって、本発明による情報処理装置等によれば、例えば、枝刈り等の手法を用いることなく、可変長 n グラムを扱うことができる。

【発明を実施するための最良の形態】

【0020】

以下、本発明による情報処理装置について、実施の形態を用いて説明する。なお、以下の実施の形態において、同じ符号を付した構成要素及びステップは同一または相当するものであり、再度の説明を省略することができる。

【0021】

(実施の形態1)

本発明の実施の形態1による情報処理装置について、図面を参照しながら説明する。

図1は、本実施の形態による情報処理装置の構成を示すブロック図である。図1において、本実施の形態による情報処理装置1は、訓練データ記憶部11と、可変長情報記憶部12と、接尾辞木情報記憶部13と、ギブスサンプリング部14と、テストデータ記憶部15と、確率算出部16と、出力部17とを備える。

【0022】

訓練データ記憶部11では、訓練データが記憶される。ここで、訓練データとは、記号の並びを示すデータである。記号とは、例えば、「あ」「い」「う」「a」「b」「c」等の文字であってもよく、「大阪」「会社」「行く」「sing」「she」等の単語であってもよく、「A」「G」「C」等の塩基を示す情報であってもよく、「Ala」「Arg」「Asn」等のアミノ酸を示す情報であってもよく、ISBN(International Standard Book Number)や製品ID、サービスを識別する情報等の販売対象を識別する情報であってもよく、その他の情報であってもよい。記号が文字や単語である場合には、訓練データは文書となる。その文書は、例えば、大規模なコーパスであることが好適である。また、記号が塩基を示す情報である場合には、訓練データは、塩基配列となる。また、記号がアミノ酸を示す情報である場合には、訓練データは、アミノ酸の配列となる。また、訓練データが販売対象を識別する情報である場合には、訓練データは、販売対象の販売履歴や、販売対象がウェブサイト等において顧客に選択されたり表示されたりした履歴等である。

【0023】

訓練データに含まれる記号は、結果として文字や単語等を特定できるのであれば、その記号そのものは文字や単語等でなくてもよい。すなわち、訓練データに含まれる記号は、例えば、文字や単語、塩基を示す情報等のそのものであってもよく、あるいは、文字や単語、塩基を示す情報等を識別する識別情報であってもよい。前者の場合には、記号自体が、例えば、「あ」「い」「う」であることになる。後者の場合には、記号自体が、例えば、「001」「002」「003」であり、その「001」等が「あ」「い」「う」等に対応付けられていることになる。

【0024】

訓練データ記憶部11は、所定の記録媒体(例えば、半導体メモリや磁気ディスク、光ディスクなど)によって実現されうる。また、訓練データ記憶部11に訓練データが記憶される過程は問わない。例えば、記録媒体を介して訓練データが訓練データ記憶部11で記憶されるようになってよく、通信回線等を介して送信された訓練データが訓練データ記憶部11で記憶されるようになってよく、あるいは、入力デバイスを介して入力された訓練データが訓練データ記憶部11で記憶されるようになってよくよい。

【0025】

可変長情報記憶部12では、グラム長情報と、代理情報とが記憶される。ここで、グラム長情報とは、訓練データ記憶部11で記憶されている訓練データに含まれる各記号に対応するグラム長を示す情報である。このグラム長情報の示す値は、その性質上、0以上の整数値である。本実施の形態による情報処理装置1では、可変長の n グラムを扱うため、各記号に対応するグラム長は一定ではなく、種々の値がとられることになる。なお、この

10

20

30

40

50

グラム長に上限値が設定されていてもよい。そのグラム長の上限值は、例えば、3グラムや5グラム、8グラム等であってもよい。上限値が設定されている場合であっても、その上限値までの範囲内で可変長の n グラムとなる。また、このグラム長情報の示す値は、グラム長と等価な情報であってもよい。例えば、グラム長の代わりにマルコフ過程のオーダーを用いてもよい。マルコフ過程のオーダーは、後述する接尾辞木の深さに対応する値であり、マルコフ過程のオーダー「 k 」は、 n グラム長「 $k + 1$ 」に対応している。

【0026】

代理情報とは、訓練データ記憶部11で記憶されている訓練データに含まれる各記号に対応するグラム長情報の示すグラム長より短いグラム長を有する代理の記号に関する情報である。前述のように、階層Pitman-Yor過程では、 n グラムを階層的に生成・推定するが、その際に、代理の記号を用いることになる。例えば、ある記号が3グラムである場合に、その記号が、ある確率で2グラムの代理の記号を生成することになる。代理情報は、その代理の記号のグラム長を示す情報である。なお、代理の記号のことを代理記号と呼ぶこともある。また、代理の記号でない記号のことを、代理の記号と区別するために実記号と呼ぶこともある。この代理情報の示す値は、0以上の整数値であり、対応するグラム長情報の示す値未満の値である。代理情報は、2以上の値を有することがありうる。この場合には、2以上の代理がなされていることになる。なお、ある記号に対応する代理の記号が存在しない場合には、代理情報には、代理の記号のグラム長を示す情報が設定されないことになる。また、前述のグラム長情報の場合と同様に、この代理情報の示す値も、グラム長と等価な情報、例えば、マルコフ過程のオーダー（接尾辞木の深さ）であってもよい。

【0027】

また、可変長情報記憶部12では、訓練データ記憶部11で記憶されている訓練データに含まれる各記号に、グラム長情報の示す記号のグラム長（あるいは、接尾辞木の深さ）と、代理情報の示す代理の記号のグラム長（あるいは、接尾辞木の深さ）とが対応付けられていることになる。したがって、可変長情報記憶部12を参照することによって、所望の記号のグラム長（あるいは、接尾辞木の深さ）と、その記号に対応する代理の記号のグラム長（あるいは、接尾辞木の深さ）とを知ることができる。

【0028】

可変長情報記憶部12は、所定の記録媒体（例えば、半導体メモリや磁気ディスク、光ディスクなど）によって実現されうる。また、可変長情報記憶部12において、後述する1回目のギブスサンプリング処理が行われるまでは、グラム長情報や代理情報に値は設定されていない。ギブスサンプリング処理が行われることによって、グラム長情報や代理情報が順次、設定・更新されることになる。

【0029】

接尾辞木情報記憶部13では、接尾辞木情報が記憶される。ここで、接尾辞木情報とは、訓練データに含まれる記号の接尾辞木(Suffix Tree)を示す情報である。この接尾辞木情報は、訓練データとグラム長情報と代理情報とに対応するものである。「接尾辞木情報が訓練データに対応する」とは、接尾辞木情報が、訓練データに含まれる記号の並びに対応した接尾辞木を示していることを意味している。また、「接尾辞木情報がグラム長情報に対応する」とは、接尾辞木情報が、グラム長情報の示すグラム長（あるいは、接尾辞木の深さ）を有する記号に関する情報を含むことを意味している。また、「接尾辞木情報が代理情報に対応する」とは、接尾辞木情報が、代理情報の示す代理の記号に関する情報を含むことを意味している。接尾辞木情報の具体例については後述する。なお、ここでは、接尾辞木情報が接尾辞木を示す情報であるとしているが、接尾辞木情報は、その接尾辞木と等価な別のデータ構造を示す情報であってもよい。結果として、同じ情報を示すことになるからである。

【0030】

接尾辞木情報記憶部13は、所定の記録媒体（例えば、半導体メモリや磁気ディスク、光ディスクなど）によって実現されうる。また、接尾辞木情報記憶部13において、後述

10

20

30

40

50

する 1 回目のギブスサンプリング処理が行われるまでは、記号の配置が何も設定されていない。ギブスサンプリング処理が行われることによって、記号の配置が順次、設定・更新されることになる。

【 0 0 3 1 】

ギブスサンプリング部 1 4 は、ギブスサンプリング処理を行う。ここで、ギブスサンプリング処理とは、接尾辞木情報を更新しながら、各記号のグラム長情報と代理情報とをギブスサンプリングにより算出して可変長情報記憶部 1 2 に蓄積する一連の処理を、繰り返して実行する処理である。このギブスサンプリング処理は、訓練データ記憶部 1 1 で記憶されている訓練データを用いて実行される。なお、ギブスサンプリング処理において、接尾辞木情報を更新しながら、各記号のグラム長情報と代理情報とをギブスサンプリングにより算出して可変長情報記憶部 1 2 に蓄積する一連の処理を繰り返す回数は、例えば、あらかじめ決められており、その回数になるまで、その一連の処理を繰り返して実行してもよい。

10

【 0 0 3 2 】

図 2 は、ギブスサンプリング部 1 4 の構成を示すブロック図である。ギブスサンプリング部 1 4 は、選択手段 2 1 と、確率算出手段 2 2 と、設定手段 2 3 と、接尾辞木情報更新手段 2 4 と、制御手段 2 5 とを備える。

【 0 0 3 3 】

選択手段 2 1 は、訓練データに含まれるすべての記号をランダムに順次選択する。訓練データに含まれる記号の総数が T である場合に、選択手段 2 1 は、例えば、1 から T までの数字の列をランダムに置換する処理を行い、その置換後の数字の列の値に対応する記号を、順次、選択するようにしてもよい。

20

【 0 0 3 4 】

確率算出手段 2 2 は、選択手段 2 1 が選択した記号が N グラム (N は 1 以上の整数) となる確率を算出する。また、確率算出手段 2 2 は、記号が N グラムとなる確率を、その確率を算出する記号の階層 $P_{i \text{ man} - Y \text{ or}}$ 過程における N グラム確率と、その確率を算出する記号以外の接尾辞木情報における記号によって N グラム長に到達する確率とを掛け合わせるにより算出してよい。この処理の詳細については後述する。

【 0 0 3 5 】

なお、確率算出手段 2 2 は、選択された記号が 1 グラムとなる確率から、グラムとなる確率までのすべての確率を算出することが理想的であるが、現実の計算量の問題から、所定の N の値で確率の算出を止めてもよい。例えば、あらかじめ N の上限値が設定されており、確率算出手段 2 2 は、1 から上限値までの N の値について、選択された記号が N グラムとなる確率を算出してよい。また、例えば、確率算出手段 2 2 は、選択された記号が N グラムとなる確率を、1, 2 ... の各 N の値について算出し、前述の N グラム長に到達する確率の値があらかじめ設定されているしきい値よりも小さくなった場合に、その算出をやめるようにしてもよい。なお、詳細については後述するが、確率算出手段 2 2 が算出する確率は、規格化のなされていないものであってもよい。すなわち、確率算出手段 2 2 は、定数倍の任意性のある確率を算出してよい。確率の値そのものを得ることが目的ではなく、各 N グラムについての確率の相対的な大小を得ることが目的だからである。

30

40

【 0 0 3 6 】

設定手段 2 3 は、確率算出手段 2 2 が算出した確率の分布に応じて、選択手段 2 1 が選択した記号に対応するグラム長を選択し、その選択したグラム長を可変長情報記憶部 1 2 で記憶されているグラム長情報に設定する。この選択や設定等の処理は、ギブスサンプリング処理としてすでに広く知られており、その詳細な説明を省略する。この設定手段 2 3 による処理によって、各記号のグラム長が設定されることになる。

【 0 0 3 7 】

また、設定手段 2 3 は、その選択したグラム長を可変長情報記憶部 1 2 で記憶されているグラム長情報に設定すると共に、選択手段 2 1 が選択した記号に対応する代理の記号のグラム長を、その選択されたグラム長に応じて階層 $P_{i \text{ man} - Y \text{ or}}$ 過程により決定し

50

、その決定したグラム長に応じて代理情報を設定する。この設定手段 2 3 による処理によって、各代理の記号のグラム長が設定されることになる。なお、この処理は、厳密には、階層 P i t m a n - Y o r 過程における処理であって、ギブスサンプリング処理ではないが、本実施の形態では、広い意味においてギブスサンプリング処理に含まれるものとして説明する。

【 0 0 3 8 】

接尾辞木情報更新手段 2 4 は、選択手段 2 1 が選択した記号に関する情報が削除されるように接尾辞木情報を更新すると共に、設定手段 2 3 によるグラム長の選択及び代理の記号のグラム長の決定に応じて、その選択されたグラム長の記号に関する情報が追加されるように接尾辞木情報を更新する。なお、選択手段 2 1 が選択した記号に関する情報には、
10 選択手段 2 1 が選択した実記号に関する情報と、その実記号に対応する代理記号に関する情報が含まれるものとする。同様に、選択されたグラム長の記号に関する情報には、選択されたグラム長の実記号に関する情報と、その実記号に対応する代理記号に関する情報が含まれるものとする。

【 0 0 3 9 】

また、ギブスサンプリング処理を初めて行う際には、前述のように、可変長情報記憶部 1 2 では、グラム長情報や代理情報が設定されていない。したがって、ギブスサンプリング処理を初めて行う際には、接尾辞木情報更新手段 2 4 は、選択手段 2 1 が選択した記号に関する情報が削除されるように接尾辞木情報を更新しなくてもよい。「記号に関する情報が削除されるように接尾辞木情報を更新する」とは、その記号（実記号、代理記号）が
20 存在しない状況を示す接尾辞木情報となるように、接尾辞木情報に含まれる情報を更新することである。また、「記号に関する情報が追加されるように接尾辞木情報を更新する」とは、その記号（実記号、代理記号）が存在する状況を示す接尾辞木情報となるように、接尾辞木情報に含まれる情報を更新することである。具体的な接尾辞木情報の更新については、後述する。

【 0 0 4 0 】

制御手段 2 5 は、選択手段 2 1 による訓練データに含まれるすべての記号のランダムな選択と、その選択された記号に関するグラム長情報の設定、代理情報の設定、及び接尾辞木情報の更新とが、繰り返して実行されるように制御する。制御手段 2 5 は、例えば、これらの処理が所定の回数だけ繰り返されるように制御してもよく、あるいは、何らかの条件を満
30 たす回数まで繰り返されるように制御してもよい。本実施の形態では、前者の場合について説明する。この制御手段 2 5 による制御が行われることによって、ギブスサンプリング処理における一連の処理が繰り返して実行されることになる。

【 0 0 4 1 】

テストデータ記憶部 1 5 では、テストデータが記憶される。ここで、テストデータとは、記号の並びを示すデータである。なお、このテストデータは、後述する確率算出部 1 6 における確率の算出で用いられるデータである。

【 0 0 4 2 】

テストデータ記憶部 1 5 は、所定の記録媒体（例えば、半導体メモリや磁気ディスク、光ディスクなど）によって実現されうる。テストデータ記憶部 1 5 にテストデータが記憶される過程は問わない。例えば、記録媒体を介してテストデータがテストデータ記憶部 1 5 で記憶されるようになって
40 もよく、通信回線等を介して送信されたテストデータがテストデータ記憶部 1 5 で記憶されるようになってよく、あるいは、入力デバイスを介して入力されたテストデータがテストデータ記憶部 1 5 で記憶されるようになってよくよい。

【 0 0 4 3 】

確率算出部 1 6 は、訓練データと、グラム長情報と、代理情報とに対応する接尾辞木情報を用いて、テストデータに含まれる記号の可変長 N グラム確率を、その確率を算出する記号の階層 P i t m a n - Y o r 過程における N グラム確率と、確率を算出する記号が N グラム長に到達する確率との積を各 N について足しあわせることによって算出する。ここで、確率算出部 1 6 は、訓練データと、グラム長情報と、代理情報とに対応する接尾辞木
50

情報として、例えば、接尾辞木情報記憶部 13 で記憶されている接尾辞木情報を用いてもよく、あるいは、訓練データと、グラム長情報と、代理情報とを用いて新たに生成した接尾辞木情報を用いてもよい。確率算出部 16 が確率を算出する処理の詳細については後述する。

【0044】

なお、確率算出部 16 は、訓練データと、グラム長情報と、代理情報とに対応する接尾辞木情報を用いて、テストデータに含まれる記号の可変長 N グラム確率を、その確率を算出する記号の階層 Pitman - Yor 過程における N グラム確率と、その確率を算出する記号が N グラム長に到達する確率との積を各 N について足しあわせた値を、複数回のギブスサンプリングについて平均をとることによって算出してもよい。本実施の形態では、この場合について説明する。なお、この場合には、ギブスサンプリング部 14 は、確率算出部 16 がテストデータに含まれる記号の確率を算出するごとに、ギブスサンプリング処理を実行してもよい。また、ここで実行されるギブスサンプリング処理における繰り返し回数（すなわち、制御手段 25 が制御する繰り返し回数）は、例えば、1 回であってもよく、あるいは、2 回以上であってもよい。

10

【0045】

出力部 17 は、確率算出部 16 が算出した可変長 N グラム確率を出力する。ここで、この出力は、例えば、表示デバイス（例えば、CRT や液晶ディスプレイなど）への表示でもよく、所定の機器への通信回線を介した送信でもよく、プリンタによる印刷でもよく、スピーカによる音声出力でもよく、記録媒体への蓄積でもよい。なお、出力部 17 は、出力を行うデバイス（例えば、表示デバイスやプリンタなど）を含んでもよく、あるいは含まなくてもよい。また、出力部 17 は、ハードウェアによって実現されてもよく、あるいは、それらのデバイスを駆動するドライバ等のソフトウェアによって実現されてもよい。

20

【0046】

なお、訓練データ記憶部 11、可変長情報記憶部 12、接尾辞木情報記憶部 13、テストデータ記憶部 15 での記憶は、RAM 等における一時的な記憶でもよく、磁気ディスク等における長期的な記憶でもよい。

【0047】

また、訓練データ記憶部 11、可変長情報記憶部 12、接尾辞木情報記憶部 13、テストデータ記憶部 15 は、同一の記録媒体によって実現されてもよく、あるいは、別々の記録媒体によって実現されてもよい。前者の場合には、例えば、訓練データを記憶している領域が訓練データ記憶部 11 となり、グラム長情報と代理情報とを記憶している領域が可変長情報記憶部 12 となる。

30

【0048】

なお、前述のように、本実施の形態による情報処理装置 1 における各種の処理等において、グラム長に代えて、それと等価な情報であるマルコフ過程のオーダー（接尾辞木の深さ）を用いてもよい。

【0049】

次に、可変長 N グラム確率を算出する方法について、具体的な式を用いて説明する。ここでは、まず、階層 Pitman - Yor 過程について説明し、次に、その階層 Pitman - Yor 過程の可変長 N グラムへの拡張について説明する。

40

【0050】

[階層 Pitman - Yor 過程]

ここでは、例として、トライグラムの言語モデルについて説明する。すなわち、記号 = 単語の場合について説明する。階層 Pitman - Yor 過程ではない通常のトライグラムの場合には、例えば、図 3 で示されるように深さ 2 の接尾辞木によって示すことができる。この接尾辞木は、訓練データから作成される。例えば、訓練データに「she will sing」が含まれる場合には、深さ 0 のノード（図 3 中の ）から深さ 1 の「will」でラベルされるノード、深さ 2 の「she」でラベルされるノードを順にたどって、その深さ 2 の「she」でラベルされるノードの sing に対応するカウント値を

50

1だけカウントアップする。なお、それらのノードが存在しない場合には、新たなノードを作成するものとする。ここで、「will」でラベルされるノードのことをノード「will」と呼ぶこともある。他の記号、ノードについても同様である。

【0051】

そのように作成された接尾辞木を用いて、she willに続くsingを予測したい場合には、ノード「」、ノード「will」、ノード「she」の順に枝をたどり、到達したノード「she」のカウント分布を用いて、 $p(\text{sing} | \text{she will})$ を計算する。なお、図3の場合には、ノード「she」に「like」のカウントはないものとする。すると、 $p(\text{like} | \text{she will})$ は0となる。

【0052】

階層Pitman-Yor過程においては、単語をノードに追加する際、例えば、「sing」を「she」でラベルされるノードに追加する際に、ある確率でその単語の代理（コピー）が親ノードに送られる。したがって、階層Pitman-Yor過程の接尾辞木は、図4で示されるようになり、深さ2以外のノードにも、代理の単語が存在することになる。したがって、ノード「she」に単語「like」が存在しなくても、ノード「he」から親ノード「will」に単語「like」が代理の単語として送られている場合には、 $p(\text{like} | \text{she will})$ は、3グラム確率（これは0となる）と、2グラム確率である $p(\text{like} | \text{will})$ とを用いて計算される。なお、 $p(\text{like} | \text{will})$ も、2グラム確率と、1グラム確率とを用いて計算され、このことが0グラム確率まで再帰的に繰り返される。ここで、0グラム確率は、 $1/V$ で与えられるものであり、その確率の推定はなされない。ただし、 V は訓練データにおける記号の種類の数である。

【0053】

より具体的には、階層Pitman-Yor過程では、次のようになる。

【数1】

$$p(w|h) = \frac{c(w|h) - d \cdot t_{hw}}{\theta + c(h)} + \frac{\theta + d \cdot t_h}{\theta + c(h)} p(w|h') \quad (1)$$

【0054】

ただし、 $h = w_{t-n} \dots w_{t-1}$ であり、 w は h に続く記号である。すなわち、記号の並びは、 hw である。また、 $c(w|h)$ は、ノード h での w のカウントである。 $c(h) = \sum_w c(w|h)$ であり、ノード h での w のカウントの総和である。 t_{hw} は、 w が $p(w|h)$ からではなく、親ノード $p(w|h')$ から生成されたと推定された回数である。 $t_h = \sum_w t_{hw}$ である。 d 、 θ は、階層Pitman-Yor過程のパラメータであり、接尾辞木上のすべての記号の分布から、それぞれベータ事後分布、ガンマ事後分布によって推定できる。 $h' = w_{t-n+1} \dots w_{t-1}$ であり、 h よりも1つオーダーを落とした記号の並びである。

【0055】

なお、式(1)は、階層Pitman-Yor過程として、すでに広く知られており、その詳細な説明を省略する。式(1)の詳細については、前述の非特許文献1または非特許文献2を参照されたい。

【0056】

この式(1)を $p(w|h')$ が0グラム確率となるまで再帰的に用いることによって、 $p(w|h)$ を計算することができる。

【0057】

[可変長Nグラムへの拡張]

ここでは、接尾辞木を用いて説明を行う便宜上、グラム長に代えて、接尾辞木の深さ（マルコフ過程のオーダー）を用いて説明を行う。前述のように、グラム長から1だけ減算したものが、接尾辞木の深さに対応している。

【0058】

10

20

30

40

50

接尾辞木の各ノード i に、接尾辞木を根 () からたどる際に、そのノード i で止まる確率 q_i があるとする。すなわち、 $(1 - q_i)$ は、図 5 で示されるように、ノード i を通過する確率となる。各 q_i は次式のように共通のベータ事前分布から生成されていると仮定する。

【 0 0 5 9 】

【 数 2 】

$$q_i \sim \text{Be}(\alpha, \beta) \quad (2)$$

ここで、 $\text{Be}(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} q^{\alpha-1} (1-q)^{\beta-1}$ は、

二項確率 q の確率分布であるベータ分布の確率密度関数であり、期待値は $E[q] = \alpha / (\alpha + \beta)$ である。

【 0 0 6 0 】

記号の並び $h = w_t \dots w_{t-2} w_{t-1}$ に続いて記号 w_t が観測されたとする、接尾辞木を根 から初めて、 $w_{t-1} w_{t-2} \dots$ の順にたどることになるが、この際は、可変長 N グラムであるため、深さ n で必ず止まるのではなく、パス上の q_i をそれぞれ $q_0, q_1, q_2 \dots$ として、次式の確率にしたがって深さ l で停止し、その深さに記号を追加する。

【 0 0 6 1 】

【 数 3 】

$$p(n = l | h) = q_l \prod_{i=0}^{l-1} (1 - q_i) \quad (3)$$

【 0 0 6 2 】

この式からわかるように、非常に深いノードであっても、そのパスに沿った q_i が小さければ、すなわち、通過する確率が高ければ、接尾辞木において到達する深さは深くなる。逆に、浅いノードでも、 q_i が大きければ、すなわち、通過する確率が低ければ、接尾辞木において到達する深さは浅くなる。

【 0 0 6 3 】

上記式 (3) より、深さ n のノードに到達する確率は、 n が大きくなるにしたがっておよそ指数的に減少するが、その度合いは接尾辞木の枝によって異なり、高頻度の長い系列に対応する深いノードを許すことのできるモデルとなっている。

なお、接尾辞木の各ノードが持つ真の q_i の値はわからないため、ギブスサンプリング処理を用いる。そのギブスサンプリング処理について説明する。

【 0 0 6 4 】

まず、可変長 N グラムモデルでは、訓練データ $w = w_1 w_2 \dots w_T$ について、それぞれの単語が生成された隠れた深さ $n = n_1 n_2 \dots n_T$ が存在していると仮定する。したがって、訓練データ w の確率は、次式のようなになる。

【 0 0 6 5 】

【 数 4 】

$$p(\mathbf{w}) = \sum_{\mathbf{n}} \sum_{\mathbf{s}} p(\mathbf{w}, \mathbf{n}, \mathbf{s}) \quad (4)$$

【 0 0 6 6 】

ここで、 s は代理の記号の配置を示す隠れ変数である。この s によって、前述のように、代理の記号の接尾辞木における深さが示されることになる (詳細については、非特許文献 1 または非特許文献 2 参照)。この n, s をギブスサンプリング処理によって推定する。具体的には、記号 w_t の持つ隠れた深さのオーダー (マルコフ過程のオーダー) n_t を

10

20

30

40

50

次式のようにサンプリングして更新していく。前述のように、 n_t のサンプリングの際に、 s_t も階層 Pitman - Yor 過程によって算出して更新していく。

【0067】

【数5】

$$n_t \sim p(n_t | \mathbf{w}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) \quad (5)$$

【0068】

ここで、 \mathbf{n}_{-t} 、 \mathbf{s}_{-t} は、それぞれ、 \mathbf{n} 、 \mathbf{s} から n_t 、 s_t を除いたベクトルである。上記式(5)は、ベイズの定理から次のように展開することができる。

【0069】

【数6】

$$p(n_t | \mathbf{w}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) \propto p(w_t | \mathbf{w}_{-t}, \mathbf{n}, \mathbf{s}_{-t}) p(n_t | \mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) \quad (6)$$

【0070】

式(6)の右辺第一項は、深さのオーダーが n_t と決まったときの w_t の n グラム確率であり、上記式(1)を用いて算出することができる。ただし、深さ k であれば、グラム長が $k+1$ であるとして上記式(1)を用いるものとする。また、右辺第二項は、この記号の並びでの深さ n_t のノードに到達する事前確率である。ここで、 w_t 以外の他の記号がノード i で止まった回数を a_i 、通過した回数を b_i とすると、 q_i の期待値は、ベータ事後分布の期待値として、次式のように推定される。

【0071】

【数7】

$$E[q_i] = \frac{a_i + \alpha}{a_i + b_i + \alpha + \beta} \quad (7)$$

【0072】

したがって、式(3)、式(7)より、式(6)の右辺第二項は、次のように計算することができる。

【0073】

【数8】

$$p(n_t = l | \mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) = \frac{a_l + \alpha}{a_l + b_l + \alpha + \beta} \prod_{i=0}^{l-1} \frac{b_i + \beta}{a_i + b_i + \alpha + \beta} \quad (8)$$

【0074】

次に、テストデータに関する予測について説明する。接尾辞木の深さ (n グラム長) を固定していないため、予測の際にも n を隠れ変数と見なして、記号の並び h に対して次式のように予測を行う。

【0075】

【数9】

$$p(\mathbf{w} | \mathbf{h}) = \sum_n p(\mathbf{w}, n | \mathbf{h}) \quad (9)$$

$$= \sum_n p(\mathbf{w} | n, \mathbf{h}) p(n | \mathbf{h}) \quad (10)$$

【0076】

ここで、式(10)の第二項 $p(n | \mathbf{h})$ は、記号の並び h の持つ n グラム長分布であり、式(8)で与えられる。また、式(10)の第一項 $p(\mathbf{w} | n, \mathbf{h})$ は、オーダーを n とした階層 Pitman - Yor 過程の予測確率であり、式(1)で求められる。この場合にも、深さ k であれば、グラム長が $k+1$ であるとして上記式(1)を用いるものとする。なお、確率算出部16についての説明でも述べたように、訓練データから生成され

10

20

30

40

50

たモデルを用いた式(10)の値の算出と、訓練データに対するギブスサンプリング処理によるモデルの更新との処理を繰り返して実行し、その式(10)の値を平均化して予測を行ってもよい。

【0077】

次に、本実施の形態による情報処理装置1のモデルを作成する動作について、図6のフローチャートを用いて説明する。なお、この処理が実行されるまでに、訓練データ記憶部11において訓練データが記憶されているものとする。

【0078】

(ステップS101)ギブスサンプリング部14の制御手段25は、カウンタ*i*を1に設定する。

10

(ステップS102)選択手段21は、訓練データ記憶部11で記憶されている訓練データから、ある記号 w_t を選択する。なお、この記号 w_t の選択において、あるカウンタ*i*に対しては、同じ記号を2回以上選択することはないものとする。

【0079】

(ステップS103)制御手段25は、カウンタ*i*が1より大きいかどうか判断する。そして、1より大きい場合には、ステップS104に進み、そうでない場合には、ステップS105に進む。したがって、カウンタ*i*が1の場合には、後述するステップS104の処理を経由しないでステップS105に進むが、これは、選択された記号 w_t に関する情報が接尾辞木情報にまだ含まれていないからである。

【0080】

20

(ステップS104)接尾辞木情報更新手段24は、選択手段21が選択した記号 w_t に関する情報が削除されるように接尾辞木情報を更新する。具体的には、接尾辞木情報から記号 w_t を削除する。また、必要があれば、その記号 w_t にいたるノードの通過回数や、停止回数を更新してもよい。さらに、その記号 w_t に対応する代理の記号が存在する場合には、接尾辞木情報更新手段24は、その代理の記号に関する情報も削除されるように接尾辞木情報を更新する。その代理の記号は、 s_t の値によって特定することができる。

【0081】

(ステップS105)確率算出手段22は、式(6)で示される、記号 w_t が深さ*N*(*N*+1グラム)となる確率を算出する。確率算出手段22は、一般に、複数の*N*の値についてこの確率を算出する処理を行う。この処理の詳細については、図7のフローチャートを用いて後述する。

30

【0082】

(ステップS106)設定手段23は、確率算出手段22が算出した確率の分布に応じて、記号 w_t に対応する深さ n_t を選択し、その選択した深さ n_t を可変長情報記憶部12で記憶されているグラム長情報に設定する。また、それと同時に、記号 w_t に対応する代理の記号が存在するかどうかについても計算し、代理の記号が存在する場合には、その代理の記号に応じた情報(s_t)を可変長情報記憶部12で記憶されている代理情報に設定する。なお、代理の記号は、2以上存在する場合もある。

【0083】

(ステップS107)接尾辞木情報更新手段24は、設定手段23による深さ n_t の選択に応じて、その選択された深さ n_t の記号 w_t に関する情報が追加されるように接尾辞木情報を更新する。具体的には、接尾辞木情報に記号 w_t を追加する。また、必要があれば、その記号 w_t にいたるノードの通過回数や停止回数を更新してもよい。さらに、その記号 w_t に対応する代理の記号が存在する場合には、その代理の記号に関しても、実記号 w_t と同様の接尾辞木情報の更新を行う。

40

【0084】

(ステップS108)選択手段21は、まだ選択していない記号 w_t が存在するかどうか判断する。まだ選択していないかどうかは、カウンタ*i*の各値について行われる。そして、選択していない記号 w_t が存在する場合には、ステップS102に戻り、そうでない場合には、ステップS109に進む。

50

【 0 0 8 5 】

(ステップS 1 0 9) 制御手段 2 5 は、カウンタ i が所定の回数以上であるかどうか判断する。この所定の回数は、前述のように、ギブスサンプリング処理における一連の処理の繰り返し回数としてあらかじめ決められている値である。その所定の値は、所定の記録媒体に記憶されており、制御手段 2 5 は、その記録媒体から、その所定の値を読み出して、この判断を行ってもよい。そして、所定の回数以上である場合には、モデルを生成する一連の処理は終了となり、そうでない場合には、ステップ S 1 1 0 に戻る。

(ステップ S 1 1 0) 制御手段 2 5 は、カウンタ i を 1 だけインクリメントする。そして、ステップ S 1 0 2 に戻る。

【 0 0 8 6 】

図 7 は、図 6 のフローチャートにおけるステップ S 1 0 5 の処理の詳細を示すフローチャートである。

(ステップ S 2 0 1) 確率算出手段 2 2 は、カウンタ k を 0 に設定する。

【 0 0 8 7 】

(ステップ S 2 0 2) 確率算出手段 2 2 は、 $n_t = k$ とした式 (6) の右辺第一項の値を算出する。具体的には、式 (1) を用いて、その値を算出する。その際に、ここでの $n_t = k$ は深さであるので、グラム長を「 $k + 1$ 」として式 (1) を用いる。確率算出手段 2 2 は、その算出した値を、図示しない記録媒体において一時的に記憶しておいてもよい。

【 0 0 8 8 】

(ステップ S 2 0 3) 確率算出手段 2 2 は、 $n_t = k$ とした式 (6) の右辺第二項の値を算出する。具体的には、式 (8) を用いて、その値を算出する。確率算出手段 2 2 は、その算出した値を、図示しない記録媒体において一時的に記憶しておいてもよい。

【 0 0 8 9 】

(ステップ S 2 0 4) 確率算出手段 2 2 は、ステップ S 2 0 2 , S 2 0 3 で算出した値を掛け合わせるにより、式 (6) の右辺の値を算出する。

(ステップ S 2 0 5) 確率算出手段 2 2 は、ステップ S 2 0 4 で算出した値を図示しない記録媒体において一時的に記憶する。

【 0 0 9 0 】

(ステップ S 2 0 6) 確率算出手段 2 2 は、確率の算出を終了するかどうか判断する。確率算出手段 2 2 は、前述のように、例えば、 k の値があらかじめ定められている値を越えた場合に、確率の算出を終了すると判断してもよく、ステップ S 2 0 3 で算出した値があらかじめ定められているしきい値より小さくなった場合に、確率の算出を終了すると判断してもよい。そして、終了する場合には、図 6 のフローチャートに戻り、そうでない場合には、ステップ S 2 0 7 に進む。

(ステップ S 2 0 7) 確率算出手段 2 2 は、カウンタ k を 1 だけインクリメントする。そして、ステップ S 2 0 2 に戻る。

【 0 0 9 1 】

なお、式 (1) 、式 (8) から明らかなように、図 7 のフローチャートのステップ S 2 0 2 , S 2 0 3 の値の算出において、それぞれ、カウンタ k の値が 1 だけ小さい場合のステップ S 2 0 2 , S 2 0 3 の値を利用することができる。したがって、ステップ S 2 0 2 , S 2 0 3 の値を一時的に記憶しておき、その値を、カウンタ k が 1 だけカウントアップされた後のステップ S 2 0 2 , S 2 0 3 の値の計算で用いるようにしてもよい。

【 0 0 9 2 】

次に、本実施の形態による情報処理装置 1 がテストデータについて確率を算出する処理について、図 8 のフローチャートを用いて説明する。なお、図 8 のフローチャートでは、式 (1 0) の値の算出と、訓練データに対するギブスサンプリング処理によるモデルの更新との処理を繰り返して実行し、その式 (1 0) の値を平均化する場合について説明する。

【 0 0 9 3 】

10

20

30

40

50

(ステップS301) 確率算出部16は、カウンタ*i*を1に設定する。

(ステップS302) 確率算出部16は、 $S^{(i)}$ を0に設定する。

(ステップS303) 確率算出部16は、カウンタ*k*を0に設定する。

【0094】

(ステップS304) 確率算出部16は、 $n = k$ とした式(10)の第一項の値を算出する。具体的には、式(1)を用いて、その値を算出する。その際に、ここでの $n = k$ は深さであるので、グラム長を「 $k + 1$ 」として式(1)を用いる。確率算出部16は、その算出した値を、図示しない記録媒体において一時的に記憶しておいてもよい。

【0095】

なお、この確率の算出の際に、例えば、確率を算出する*hw*の記号の並びが指定されてもよく、あるいは、テストデータそのものが、*hw*の記号の並びであってもよい。*hw*の記号の並びが指定される場合には、あらかじめテストデータにおいてその指定がなされていてもよく、あるいは、情報処理装置1への入力によって、*hw*の記号の並びが指定されてもよい。

10

【0096】

(ステップS305) 確率算出部16は、 $n = k$ とした式(10)の第二項の値を算出する。具体的には、式(8)と同様にして、その値を算出する。確率算出部16は、その算出した値を、図示しない記録媒体において一時的に記憶しておいてもよい。

【0097】

(ステップS306) 確率算出部16は、ステップS304, S305で算出した値を掛け合わせて $S^{(i)}$ に足すことによって、 $S^{(i)}$ の値を更新する。確率算出部16は、その更新した $S^{(i)}$ の値を、図示しない記録媒体において一時的に記憶しておいてもよい。

20

【0098】

(ステップS307) 確率算出部16は、カウンタ*k*をカウントアップして、 $S^{(i)}$ の値を更新する処理を継続するか、処理を終了するか判断する。そして、処理を終了する場合には、ステップS309に進み、そうでない場合には、ステップS308に進む。

【0099】

なお、このステップS307の判断は、図7のフローチャートのステップS206の判断、すなわち、訓練データからモデルを作成する際に確率の算出を打ち切る判断と同様にしてもよい。例えば、ステップS206において、*k*の値があらかじめ定められている値を越えた際に、確率の算出を終了すると判断した場合には、ステップS307においても、*k*の値があらかじめ定められている値(ステップS206の判断で用いられる値と同じ値)を超えた場合に、 $S^{(i)}$ の値を更新する処理を終了すると判断してもよい。また、例えば、ステップS206において、ステップS203で算出した値があらかじめ定められているしきい値より小さくなった際に、確率の算出を終了すると判断した場合には、ステップS307においても、ステップS305で算出した値があらかじめ定められているしきい値(ステップS206の判断で用いられるしきい値と同じしきい値)よりも小さくなった場合に、 $S^{(i)}$ の値を更新する処理を終了すると判断してもよい。

30

【0100】

(ステップS308) 確率算出部16は、カウンタ*k*を1だけインクリメントする。そして、ステップS304に戻る。

40

(ステップS309) 確率算出部16は、ギブスサンプリング処理を継続するか、ギブスサンプリング処理を行わないか判断する。確率算出部16は、例えば、カウンタ*i*があらかじめ定められている値を超えた場合に、ギブスサンプリング処理を行わないと判断してもよく、あるいは、何らかの条件が満たされる場合に、ギブスサンプリング処理を行わないと判断してもよい。ギブスサンプリング処理を行わないと判断した場合には、ステップS312に進み、そうでない場合、すなわち、ギブスサンプリング処理を継続すると判断した場合には、ステップS310に進む。

【0101】

50

(ステップS310) ギブスサンプリング部14は、ギブスサンプリング処理を行う。この処理の詳細については、図9のフローチャートを用いて後述する。

(ステップS311) 確率算出部16は、カウンタ*i*を1だけインクリメントする。そして、ステップS302に戻る。

【0102】

(ステップS312) 確率算出部16は、それまでに算出した $S^{(i)}$ (ここで、 $i = 0, 1, 2, \dots, i_{max}$) の平均値を算出する。この平均値が可変長*N*グラム確率となる。具体的には、 $S^{(i)}$ ($i = 0, 1, 2, \dots, i_{max}$) をすべて加算した値を、($i_{max} + 1$) で割ればよい。

【0103】

(ステップS313) 出力部17は、確率算出部16がステップS312で算出した可変長*N*グラム確率を出力する。そして、可変長*N*グラム確率を算出する一連の処理は終了となる。

【0104】

図9は、図8のフローチャートにおけるステップS310の処理の詳細を示すフローチャートである。なお、図9のフローチャートにおける各処理は、図6のフローチャートの対応する処理と同様のものであり、その説明を省略する。なお、図9のフローチャートでは、ギブスサンプリング処理における繰り返し回数が1回である場合について示しているが、前述のように、この繰り返し回数は、2回以上であってもよい。繰り返し回数が2回以上である場合には、図6のフローチャートと同様に、図9のフローチャートの処理を繰り返すことになる。

【0105】

次に、本実施の形態による情報処理装置1の動作について、具体例を用いて説明する。

まず、訓練データ、グラム長情報、代理情報について説明する。図10は、訓練データと、グラム長情報と、代理情報との対応の一例を示す図である。図10では、訓練データに含まれる記号 w_t と、グラム長情報の示す接尾辞木の深さ n_t と、代理情報の示す代理の記号の接尾辞木の深さ s_t とが対応付けられている。例えば、記号 w_t に対応付けられて、 $n_t = 3$ 、 $s_t = 1, 2$ が設定されている場合には、記号 w_t に関する接尾辞木は、図11で示されるようになる。すなわち、記号 w_t は深さが3(4グラム)であるため、記号 w_{t-3} でラベルされるノードに実記号 w_t が存在する。また、代理記号の深さが1, 2(グラム長が2, 3)であるため、記号 w_{t-2} 、 w_{t-1} でラベルされるノードに、それぞれ代理記号 w_t が存在する。なお、接尾辞木の深さの代わりにグラム長を用いてもよいことは言うまでもない。

【0106】

なお、図10で示される訓練データ、グラム長情報、代理情報は、同一の記録媒体で構成される訓練データ記憶部11と、可変長情報記憶部12とにおいて記憶されていてもよく、あるいは、別々の記録媒体で構成される訓練データ記憶部11と、可変長情報記憶部12とにおいて記憶されていてもよい。

【0107】

次に、接尾辞木情報と接尾辞木とについて説明する。この具体例では、記号は単語であるとす。また、図12で示されるように、記号と、その記号を識別する情報である記号IDとが対応付けられており、情報処理装置1は、記号を識別する記号IDを用いて処理を行うものとする。

【0108】

また、例えば、図13で示される接尾辞木に対応する接尾辞木情報は、図14で示されるようになる。図14の接尾辞木情報において、ノードのアドレスと、記号IDと、親アドレスと、子アドレスと、実記号の記号ID及び個数と、代理記号の記号ID及び個数と、停止回数「a」と、通過回数「b」とが対応付けられている。ノードのアドレスは、ノードを識別する情報である。記号IDは、そのノードをラベルする記号を識別する情報である。親アドレスは、そのノードにつながっている1だけ階層の浅いノードを識別

10

20

30

40

50

する情報である。親アドレスは、根 () のノード以外、1 個だけ存在する。子アドレスは、そのノードにつながっている 1 だけ階層の深いノードを識別する情報である。子アドレスは、1 以上存在する場合もあり、あるいは、存在しない場合もある。実記号の記号 ID は、そのノードに存在する実記号を識別する情報である。その個数は、実記号が存在する個数を示す情報である。代理記号の記号 ID は、そのノードに存在する代理記号を識別する情報である。その個数は、代理記号が存在する個数を示す情報である。停止回数 a は、そのノードで実記号が停止した回数、すなわち、そのノードに存在する実記号の全個数である。通過回数 b は、そのノードで実記号が停止せずに、さらに深い階層のノードにまで到達した回数、すなわち、そのノードにつながっている 1 だけ階層の深いすべてのノードの停止回数 a と、通過回数 b との和をとった値である。

10

【0109】

図 13 より、アドレス「A0002」のノードをラベルする記号は「will」であるため、図 12 より、アドレス「A0002」のレコードにおける記号 ID は、記号「will」に対応している「2051」となる。また、そのノードに対応する親アドレスは、図 13 で示されるように「A0001」であり、そのノードに対応する子アドレスは、図 13 で示されるように「A0003」「A0004」である。また、そのノードには、少なくとも、記号 ID「3210」で識別される 5 個の実記号「like」と、記号 ID「4321」で識別される 4 個の実記号「sing」が存在する。また、そのノードには、少なくとも、記号 ID「4321」で識別される 4 個の代理記号「sing」と、記号 ID「5432」で識別される 2 個の代理記号が存在する。また、そのノードでの実記号の停止回数 a は 40 回である。すなわち、そのノードには、40 個の実記号が存在することになる。また、そのノードでの実記号の通過回数 b は 80 回である。すなわち、そのノードと直接的に、または、間接的につながっている子ノードに 80 個の実記号が存在することになる。このようにして、図 13 で示される接尾辞木を、図 14 で示される接尾辞木情報によって表現することができる。

20

【0110】

なお、図 14 の接尾辞木情報において、子アドレスは、必ずしも必要ではない。親アドレスがわかれば、それをを用いることによって、あるノードにつながっている 1 だけ階層の深い子ノードを特定することができるからである。また、図 14 の接尾辞木情報において、停止回数 a や通過回数 b は、必ずしも必要ではない。そのノードに存在する実記号の個数や、そのノードと直接的に、または、間接的につながっている子ノードに存在する実記号の個数を数えることによって算出することができるからである。このように、接尾辞木情報には、ある程度の任意性がある。

30

【0111】

また、この具体例では、 (a, b) が実記号に関する停止回数と通過回数である場合について説明するが、 (a, b) は、実記号と代理記号との両方に関する停止回数と通過回数であってもよい。

【0112】

次に、モデルを生成する処理について説明する。まず、図 10 で示される訓練データ、グラム長情報、代理情報において、 n, s は、何も設定されていないものとする。その状況において、ギブスサンプリング部 14 が、モデルを生成する処理を開始したとする。すると、選択手段 21 は、1 から T までの整数をランダムに置換した整数の列 $randperm(1 \dots T)$ を算出する。そして、その 1 番目の数字に対応する w_t を選択する (ステップ S101, S102)。ここでは、接尾辞木情報の更新は行われない (ステップ S103, S104)。まだ更新すべき接尾辞木情報が存在しないからである。その後、確率算出手段 22 は、 w_t が n_t の深さとなる確率を算出する (ステップ S105, S201 ~ S207)。ここでは、 $k = 0 \sim 7$ までの各確率が算出されたものとする。すると、設定手段 23 は、その算出された確率の分布に応じて、いずれかの n_t を選択し、可変長情報記憶部 12 で記憶されているグラム長情報に設定する (ステップ S106)。また、設定手段 23 は、階層 Pitman-Yor 過程の処理によって、記号 w_t に代理が存在す

40

50

るかどうかについても計算し、代理が存在する場合には、その代理に応じて、可変長情報記憶部12で記憶されている代理情報に s_t を設定する。例えば、 $n_t = 3$ であり、実記号 w_t の親ノードと、さらにその親ノードとに代理記号 w_t が存在する場合には、 $s_t = 1, 2$ となる。

【0113】

接尾辞木情報更新手段24は、設定手段23が設定した n_t, s_t の値に応じて、接尾辞木情報記憶部13で記憶されている接尾辞木情報における、実記号と代理記号に関する情報を更新する(ステップS107)。具体的には、対応するノードにおける実記号の個数や代理記号の個数をインクリメントしたり、必要であれば、新たな実記号の記号IDや、代理記号の記号IDをレコードに登録したり、新たなノード(レコード)を作成したりする。また、実記号や代理記号の追加に応じて、停止回数 a 、通過回数 b も更新する。

10

【0114】

このような処理が、 $randperm(1 \dots T)$ で算出された各数字の列に対応する各 w_t に対して、順次、実行されていく(ステップS102~S108)。そして、 $randperm(1 \dots T)$ で算出されたすべての数字に対応する記号 w_t に対して処理が実行されると、さらに繰り返して、その一連の処理が実行される(ステップS102~S110)。なお、 $randperm(1 \dots T)$ を2回目に実行した際(すなわち、カウンタ $i = 2$ の際)には、すでに接尾辞木情報が設定されているため、選択された記号 w_t に関する情報が削除されるように接尾辞木情報が更新される(ステップS103, S104)。

【0115】

20

例えば、接尾辞木情報が図14で示される場合に、記号ID「4321」で識別される記号「sing」が選択されたとする(ステップS102)。その選択された記号「sing」に対応する n_t は「2」であり、 s_t は、「1」であったとする。また、その選択された記号 $w_t = sing$ よりも1個前の記号 w_{t-1} と、2個前の記号 w_{t-2} とは、それぞれ「will」と、「she」であったとする。

【0116】

すると、接尾辞木情報更新手段24は、接尾辞木情報を根のノード()から、各レコードの子アドレスを用いることによって、will、sheとたどり、選択された記号「sing」の存在するアドレス「A0003」のノードを特定する。そして、接尾辞木情報更新手段24は、選択された記号「sing」を削除するために、接尾辞木情報のアドレス「A0003」のレコードにおける実記号の記号ID「4321」に対応する個数と、そのレコードにおける停止回数 a とを1だけデクリメントする。また、接尾辞木情報更新手段24は、停止回数 a を1だけデクリメントしたことに伴って、そのアドレス「A0003」のノードの親ノードから根のノード()までのそれぞれの通過回数 b を、1だけデクリメントする。

30

【0117】

また、 $s_t = 1$ であるため、その選択された記号「sing」の存在するノードよりも1個浅い親ノードに代理記号「sing」が存在する。したがって、接尾辞木情報更新手段24は、アドレス「A0002」のノードのレコードにおける代理記号の記号ID「4321」に対応する個数を1だけデクリメントする。その結果、接尾辞木情報は、図15で示されるようになる。

40

【0118】

次に、確率算出手段22は、選択された記号「sing」が深さ0から深さ7となる確率をそれぞれ算出する(ステップS105, S201~S207)。この算出の際に、式(8)を用いるが、そのときの a_i, b_i の値は、接尾辞木情報に含まれる a, b の値を用いることができる。すでに、記号 w_t に関する情報が削除されるように接尾辞木情報を更新しているからである。また、ベータ事前分布のパラメータは、(,) = (4, 1)を用いたが、後述するように、(,)の値による性能の差はほとんどない。

【0119】

ここで、式(8)の具体的な計算の一例について説明する。式(8)において深さ n_t

50

= 2であり、(,) = (4 , 1)であり、ノードA 0 0 0 1において、(a , b) = (1 0 0 , 9 0 0)であり、ノードA 0 0 0 2 , A 0 0 0 3は、図15で示されるようになっていたとする。すると、式(8)の値は、次のように計算される。

【 0 1 2 0 】

【数10】

$$p(n_t = 2 | \mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) = \frac{29+4}{29+4+1} \times \frac{79+1}{119+4+1} \times \frac{900+1}{1000+4+1}$$

10

【 0 1 2 1 】

その後、設定手段23は、その算出された確率の分布に応じて、 $n_t = 1$ を選択し、グラム長情報に設定したとする(ステップS106)。なお、この場合には、記号「sing」の代理は存在しないことになったとする。したがって、 s_t には何も設定されない。

【 0 1 2 2 】

すると、接尾辞木情報更新手段24は、接尾辞木情報を根のノード()から、willまでたどり、選択された記号「sing」の存在するべきノードを特定する。そのノードは、アドレス「A0002」のノードである。そして、接尾辞木情報更新手段24は、そのアドレス「A0002」のノードのレコードにおける実記号「4321」に対応する個数を1だけインクリメントし、停止回数aを1だけインクリメントする。また、接尾辞木情報更新手段24は、そのノードの親ノードであるアドレス「A0001」のノード、すなわち、根のノード()のレコードにおける通過回数bを1だけインクリメントする。このようにして、設定手段23が選択した深さ「1」の記号「sing」に関する情報が追加されるように接尾辞木情報が更新される(ステップS107)。

20

【 0 1 2 3 】

このような処理が繰り返されることにより、グラム長情報や、代理情報が適切な値に近づいていくことになる。例えば、カウンタiが $i_{max} = 200$ となるまでこの処理を繰り返してもよい。このようにして、訓練データからモデルの生成が行われる。

【 0 1 2 4 】

訓練データに、例えば、「シャンソンの響きに日本語を乗せるつまり言葉と音の融合というところで遊んでいたEOS」という文が含まれており、その訓練データから生成された単語と深さ n_t との対応は、次のようになる。

30

【 0 1 2 5 】

シャンソン	1
の	2
響き	6
に	3
日本語	4
を	3
乗せる	5
つまり	3
言葉	2
と	2
音	4
の	3
融合	5
という	3
ところ	5
で	4
遊ん	5

40

50

で 1
 い 4
 た 3
 E O S 5

【0126】

この結果からわかるように、「...遊んで...」の「で」は、深さ1(2グラム)であるため、1個前の単語、すなわち、「遊ん」にのみ依存していることがわかる。

【0127】

次に、生成されたモデルから、テストデータの可変長Nグラム確率を算出する処理について説明する。テストデータ記憶部15で記憶されているテストデータから、確率を算出するべきhとwが特定されると、確率算出部16は、式(10)の第一項の値と、第二項の値とを算出し(ステップS301~S305)、それらを掛け合わせて、順次、加算していく(ステップS304~S308)。モデルを用いた式(10)の値の算出が終了すると、訓練データに対するギブスサンプリング処理によるモデルの更新(ステップS310, S102~108, S311)が行われ、再度、式(10)の値の算出が行われる(ステップS302~S304)。この式(10)の値の算出と、ギブスサンプリング処理によるモデルの更新との処理が繰り返して実行され、その式(10)の値を平均化することによって、可変長Nグラム確率が算出され(ステップS312)、出力される(ステップS313)。これらは単なる計算の処理であるため、具体的な数値については省略する。

【0128】

この確率の計算が行われることによって、例えば、記号が単語である場合に、ある単語の並びの次に出現する単語の確率を算出することができる。また、記号が販売対象を識別する情報である場合に、あるユーザが、販売対象を順番に購入した次に購入する販売対象の確率を算出することができる。したがって、その結果に基づいて、他の販売対象と比較して確率の高い販売対象をレコメンドすることにより、そのユーザの販売対象の購入に関するレコメンドを行うことができる。

【0129】

なお、ステップS305の計算において、式(8)と同様にして計算することができる。と前述したが、式(8)における a_i 、 b_i の値は、記号 w_t に関する情報が削除されるように接尾辞木情報を更新された後の値であったが、このステップS305の計算においては、そのような更新を行わないで、接尾辞木情報に含まれる a 、 b の値そのものを用いればよい。

【0130】

また、この具体例において、訓練データからモデルを生成する際に、深さ7までの確率を算出したため、この可変長Nグラム確率の算出の際にも、深さ7までの確率を算出する。すなわち、ステップS307において、 $k>7$ であれば、ステップS309に進み、そうでなければ、ステップS308に進めばよい。

【0131】

また、式(10)の値を算出する処理を、訓練データに対してギブスサンプリングを行いながら、複数回繰り返す(ステップS302~S311)。例えば、50回繰り返す場合には、ステップS309において、 $i=50$ かどうか判断し、 $i=50$ の場合には、すでに50回繰り返されているのでステップS312に進み、そうでない場合には、ステップS310に進めばよい。

このようにして、訓練データからのモデルの生成と、その生成されたモデルを用いた予測とが行われることになる。

【0132】

次に、本実施の形態による情報処理装置1での実験結果について説明する。この実験では、英語については、NAB(North American Business News)コーパスのWSJセットよりランダムに選択した409,246文、10,007

、108語を訓練データとして用い、さらに10,000文をテストデータとした。単語はすべて小文字とし、全体で頻度10未満の単語は同じ特別な語にまとめた。総語彙数は26,497語である。日本語については、毎日新聞2000年度のテキスト(約150万文)からランダムに選んだ520,000文、10,079,410語を訓練データとして用い、さらに10,000文を評価データとした。それらの日本語のデータでは、MeCabで分かち書きを行い、頻度10未満の語をまとめた。その結果、総語彙は32,783語となった。

【0133】

モデルの生成においては、本実施の形態1による情報処理装置1のモデル(VPYLM)の生成と、比較例としての、階層Pitman-Yor過程を用いたモデル(HPYLM)の生成とを行った。それぞれのモデルの生成において、 $N = 200$ のギブスサンプリング処理を行った。すなわち、図6のフローチャートにおいて、カウンタ $i = 200$ となるまで処理を行った。また、モデルを用いた予測においては、50回のギブスサンプリング処理を行った。すなわち、図8のフローチャートにおいて、カウンタ $i = 50$ となるまで処理を行った。さらに、ベータ事前分布のパラメータは、 $(\alpha, \beta) = (4, 1)$ とした。実験はすべて、Xeon 3.2GHz、メモリ4GBのLinux上で行った。

【0134】

図16、図17は、その実験結果を示す図である。図16が英語のNABコーパスを訓練データとした場合の実験結果を示しており、図17が日本語の毎日新聞コーパスを訓練データとした場合の実験結果を示している。図16、図17において、HPYLM、VPYLMの値はそれぞれ、階層Pitman-Yor過程を用いたモデルと、本実施の形態による情報処理装置1が生成したモデルのパープレキシティを示している。図16、図17における n はグラム長である。そのグラム長は、階層Pitman-Yor過程を用いたモデルにおいては、固定の n グラム長であり、本実施の形態による情報処理装置1が生成したモデルにおいては、 n グラム長の最大値である。また、Nodes(H)、Nodes(V)の値はそれぞれ、階層Pitman-Yor過程のモデルにおけるノード数と、本実施の形態による情報処理装置1が生成したモデルのノード数とを示している。なお、N/Aは、メモリアーフローを示す。

【0135】

この結果から、VPYLMはHPYLMとほぼ同等の性能を40%以上少ないノード数で達成し、HPYLMでは推定できない $n = 7, 8$ のような高次 n グラムについても、必要なもののみを選択的にモデルに加えることで推定が可能であり、より高い性能を持つことがわかる。ノード数が少なくてもよいと、使用するメモリ容量を少なくすることが可能となりうる。

【0136】

また、同じ(最大)オーダー n の場合でも、VPYLMはHPYLMより20%程度学習が高速である。これは、VPYLMにおいて n グラムオーダーをサンプリングする計算コストよりも、不必要に深いノードを追加しないことによる計算量の削減が大きいからだと考えられる。図18に、8グラムVPYLMにおいて推定された深さ n のオーダーの、データ全体での分布を示す。なお、図18における n は接尾辞木の深さ(マルコフ過程のオーダー)である。文脈長を長くするメリットと、深いノードに到達するコストの間で適切なトレードオフが行われ、 $n = 3, 4$ 程度をピークに指数的な減衰が起きていることがわかる。

【0137】

以上のように、本実施の形態による情報処理装置1によれば、可変長 N グラムのモデルを生成することができる。また、その可変長 N グラムのモデルは、従来の階層Pitman-Yor過程によるモデルよりもメモリ量が少なくよく、より高い性能を持つことが確認された。さらに、可変長 N グラムのモデルを生成する方が、従来の階層Pitman-Yor過程によるモデルを生成するよりも高速であることも確認された。

【0138】

10

20

30

40

50

なお、図19は、VPYLMのパラメータ()と、テストデータのパープレキシティ(PPL)との関係を示す図である。図19において、() (0.1~1.0) × (0.1~1.0)の範囲でパラメータ()の値を変化させている。この図19からわかるように、 となる場合を除いて、性能はほぼ一定であることがわかる。

【0139】

また、本実施の形態では、情報処理装置1がテストデータに関する確率の算出をも行う場合について説明したが、このテストデータに関する確率の算出は、情報処理装置1が生成した接尾辞木情報を用いて、情報処理装置1以外で行われてもよい。その場合には、情報処理装置1は、テストデータ記憶部15、確率算出部16、出力部17を備えていなくてもよい。

10

【0140】

また、本実施の形態では、ギブスサンプリング部14によるギブスサンプリング処理を開始する際に、 n_t 、 s_t に何らかの初期値が設定されていてもよい。例えば、 n_t の各値に2が設定されており、深さ2(3グラム)の場合の階層Pitman-Yor過程によって算出される s_t の値が設定されていてもよい。

また、本実施の形態では、接尾辞木の深さを用いて主に説明したが、前述のように、接尾辞木の深さに代えてグラム長を用いてもよい。

【0141】

(実施の形態2)

本発明の実施の形態2による情報処理装置について、図面を参照しながら説明する。本実施の形態による情報処理装置は、記号の並びがモデルから生成される事前確率を計算することにより、慣用的な記号の並び(例えば、記号が単語や文字である場合には、慣用的なフレーズとなる)を抽出するものである。

20

【0142】

図20は、本実施の形態による情報処理装置2の構成を示すブロック図である。図20において、本実施の形態による情報処理装置2は、訓練データ記憶部11と、可変長情報記憶部12と、接尾辞木情報記憶部13と、ギブスサンプリング部14と、確率算出部31と、記号列選択部32と、出力部33とを備える。なお、確率算出部31、記号列選択部32、出力部33以外の構成及び動作は、実施の形態1と同様であり、その説明を省略する。

30

【0143】

なお、本実施の形態においても、グラム長と、それと等価な情報であるマルコフ過程のオーダー(接尾辞木の深さ)のどちらを用いてもよいものとする。本実施の形態では、接尾辞木を用いて説明を行う便宜上、主に接尾辞木の深さを用いて説明を行う。

【0144】

確率算出部31は、訓練データと、グラム長情報と、代理情報とに対応する接尾辞木情報を用いて、記号の並びに含まれる記号の確率を、その確率を算出する記号の階層Pitman-Yor過程における接尾辞木の深さがNとなる確率(N+1グラム確率)と、その確率を算出する記号が接尾辞木の深さNに到達する確率(N+1グラム長に到達する確率)とを掛け合わせることによって算出する。この確率を算出する記号の含まれる記号の並びは、訓練データであってもよく、訓練データとは別のデータであってもよい。本実施の形態では、前者の場合について説明する。なお、後者の場合には、その訓練データとは別の記号の並びを示すデータが記憶されるデータ記憶部(図示せず)を、情報処理装置2が有するものとする。

40

【0145】

ここで、その確率を算出する記号の階層Pitman-Yor過程における接尾辞木の深さがNとなる確率とは、その確率を算出する記号をwとした場合における、実施の形態1の式(10)の第一項の確率である。また、その確率を算出する記号が接尾辞木の深さNに到達する確率とは、その確率を算出する記号をwとした場合における、実施の形態1の式(10)の第二項の確率である。したがって、確率算出部31は、式(10)の第一

50

項と第二項とを単に掛け合わせただけの値、すなわち、式(10)の和をとらない値を算出することになる。この算出方法は、実施の形態1で説明したとおりであり、その詳細な説明を省略する。

【0146】

記号列選択部32は、確率算出部31が算出した複数の記号の確率において、他の記号の確率に比べて大きい確率を有する記号を含むK個の記号の並びである記号列を選択する。ここで、Kは、その記号の確率が算出された際の接尾辞木の深さの値である。すなわち、記号列選択部32は、確率算出部31が算出した確率 $p(w|n, h)p(n|h)$ から値の大きいものを特定し、その特定した確率に対応するhwの記号の並びを選択する。ここで、大きい確率を有する記号を含むK個の記号の並びが、hwとなる。なお、確率の値の大きいものとは、例えば、確率の値の降順になるようにソートして、確率の値の最大値からあらかじめ決められている個数のものであってもよく、あるいは、所定のしきい値があらかじめ決められており、そのしきい値以上の確率の値であってもよい。

10

【0147】

出力部33は、記号列選択部32が選択した記号列を出力する。ここで、この出力は、例えば、表示デバイス(例えば、CRTや液晶ディスプレイなど)への表示でもよく、所定の機器への通信回線を介した送信でもよく、プリンタによる印刷でもよく、スピーカによる音声出力でもよく、記録媒体への蓄積でもよい。なお、出力部33は、出力を行うデバイス(例えば、表示デバイスやプリンタなど)を含んでもよく、あるいは含まなくてもよい。また、出力部33は、ハードウェアによって実現されてもよく、あるいは、それらのデバイスを駆動するドライバ等のソフトウェアによって実現されてもよい。

20

【0148】

次に、本実施の形態による情報処理装置2の動作について、フローチャートを用いて説明する。本実施の形態による情報処理装置2が訓練データからモデルを生成する処理は、実施の形態1における図6のフローチャートと同様の処理であり、その説明を省略する。

【0149】

図21は、本実施の形態による情報処理装置2が確率の高い記号列を選択して出力する動作を示すフローチャートである。

(ステップS401) 確率算出部31は、訓練データから記号wを選択する。なお、この記号の選択において、毎回異なる選択を行うものとする。したがって、例えば、訓練データの先頭の記号から、順番に選択を行っていてもよい。

30

【0150】

(ステップS402) 確率算出部31は、カウンタkを0に設定する。

(ステップS403) 確率算出部31は、確率 $p(w|n=k, h)$ の値を算出する。具体的には、式(1)を用いて、その値を算出する。確率算出部31は、その算出した値を、図示しない記録媒体において一時的に記憶しておいてもよい。なお、hは、訓練データにおけるwより前の記号の並びである。したがって、訓練データには、hwの記号の並びが存在している。ただし、kが接尾辞木の深さであれば、グラム長がk+1であるとして上記式(1)を用いるものとする。

【0151】

(ステップS404) 確率算出部31は、確率 $p(n|h)$ の値を算出する。具体的には、式(8)を用いて、その値を算出する。確率算出部31は、その算出した値を、図示しない記録媒体において一時的に記憶しておいてもよい。

40

【0152】

(ステップS405) 確率算出部31は、ステップS403, ステップS404で算出した値を掛け合わせる。

(ステップS406) 確率算出部31は、ステップS405で算出した値を、記号の並びhwに対応付けて図示しない記録媒体において一時的に記憶する。

【0153】

(ステップS407) 確率算出部31は、確率 $p(w|n, h)p(n|h)$ を算出す

50

る処理を継続するか、終了するか判断する。そして、処理を終了する場合には、ステップ S 4 0 9 に進み、そうでない場合には、ステップ S 4 0 8 に進む。

【 0 1 5 4 】

なお、このステップ S 4 0 7 の判断は、図 7 のフローチャートのステップ S 2 0 6 の判断、すなわち、訓練データからモデルを作成する際に確率の算出を打ち切る判断と同様にしてもよい。例えば、ステップ S 2 0 6 において、 k の値があらかじめ定められている値を越えた際に、確率の算出を終了すると判断した場合には、ステップ S 4 0 7 においても、 k の値があらかじめ定められている値（ステップ S 2 0 6 の判断で用いられる値と同じ値）を超えた場合に、確率を算出する処理を終了すると判断してもよい。また、例えば、ステップ S 2 0 6 において、ステップ S 2 0 3 で算出した値があらかじめ定められているしきい値より小さくなった際に、確率の算出を終了すると判断した場合には、ステップ S 4 0 7 においても、ステップ S 4 0 4 で算出した値があらかじめ定められているしきい値（ステップ S 2 0 6 の判断で用いられるしきい値と同じしきい値）よりも小さくなった場合に、確率を算出する処理を終了すると判断してもよい。

10

【 0 1 5 5 】

（ステップ S 4 0 8）確率算出部 3 1 は、カウンタ k を 1 だけインクリメントする。そして、ステップ S 4 0 3 に戻る。

（ステップ S 4 0 9）確率算出部 3 1 は、訓練データにおいて、未選択の記号 w が存在するかどうか判断する。そして、存在する場合には、ステップ S 4 0 1 に戻り、そうでない場合には、ステップ S 4 1 0 に進む。なお、例えば、訓練データの先頭の記号から順番に選択している場合には、訓練データの最後の記号について一連の処理（ステップ S 4 0 2 ~ 4 0 8 の処理）を実行した後に、未選択の記号 w が存在しないと判断してもよい。

20

【 0 1 5 6 】

（ステップ S 4 1 0）記号列選択部 3 2 は、確率算出部 3 1 が一時的に記憶した、互いに対応付けられている記号の並び hw と確率の値とを、確率の値の降順となるようにソートする。

【 0 1 5 7 】

（ステップ S 4 1 1）記号列選択部 3 2 は、ソート後の記号の並び hw と確率の値との対応において、大きい確率の値に対応する記号列 hw を選択する。記号列選択部 3 2 は、前述のように、例えば、確率の値が最大のものから順番に、所定の個数の記号列を選択してもよく、確率の値が所定のしきい値以上の記号列を選択してもよい。

30

【 0 1 5 8 】

（ステップ S 4 1 2）出力部 3 3 は、選択した記号列を出力する。なお、この出力時には、確率の値に対応付けて記号列を出力してもよく、そうでなくてもよい。このようにして、記号列を選択して出力する一連の処理は終了となる。

【 0 1 5 9 】

なお、図 2 1 のフローチャートでは、すべての記号の並びと、確率の値との対応を一時的に記憶する場合について説明したが、そうでなくてもよい。例えば、所定のしきい値を設けて、そのしきい値以下の確率の値の場合には、一時的な記憶を行わないようにしてもよい。また、確率の値を一時的に記憶するたびに、あるいは、その一時的な記憶の処理を複数回行うたびに、ステップ S 4 1 0 と同様のソートの処理を行い、記号列と確率の値との対応を所定の個数だけ一時記憶するように残し、残りは削除するようにしてもよい。このようにすることで、一時記憶のための記録媒体の容量を節約することができる。

40

【 0 1 6 0 】

また、この実施の形態 2 による情報処理装置 2 の動作は、確率の値のソートや、記号列の選択、出力以外、実施の形態 1 における処理と同様であり、具体例の説明を省略する。

【 0 1 6 1 】

次に、実験結果について説明する。実施の形態 1 の実験と同様にして訓練データによる学習を行い、その訓練データに含まれる記号列と、その記号列について求めた確率 p との対応を図 2 2 に示す。図 2 2 は、英語の N A B コーパスで学習した 8 グラムの V P Y L M

50

から得られた記号列である。BOS (Beginning Of Sentence) は、文頭を示す記号である。また、NUM (Number) は、任意の数字を示す記号である。

【 0 1 6 2 】

以上のように、本実施の形態による情報処理装置 2 によれば、例えば、図 2 2 で示されるように、記号の並びにおいて慣用的に使用されている記号の並びを抽出することができる。例えば、文書に対して用いると、慣用的に使用されているフレーズを抽出することが可能となる。特に、可変長 N グラムによる学習を行っているため、従来例のように、グラム長 (マルコフ過程のオーダー) が限定されず、種々の長さの記号列を抽出することが可能となる。

10

【 0 1 6 3 】

なお、上記各実施の形態において、記号が単語である場合に、その単語の言語が英語である場合について説明したが、これは一例であって、日本語や中国語、ドイツ語等の他の言語の単語であってよい。

【 0 1 6 4 】

また、上記各実施の形態において、各処理または各機能は、単一の装置または単一のシステムによって集中処理されることによって実現されてもよく、あるいは、複数の装置または複数のシステムによって分散処理されることによって実現されてもよい。

【 0 1 6 5 】

また、上記各実施の形態において、各構成要素は専用のハードウェアにより構成されてもよく、あるいは、ソフトウェアにより実現可能な構成要素については、プログラムを実行することによって実現されてもよい。例えば、ハードディスクや半導体メモリ等の記録媒体に記録されたソフトウェア・プログラムを CPU 等のプログラム実行部が読み出して実行することによって、各構成要素が実現され得る。なお、上記各実施の形態における情報処理装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、コンピュータを、訓練データ記憶部で記憶される、記号の並びを示すデータである訓練データを用いて、接尾辞木情報記憶部で記憶される、前記訓練データと前記訓練データに含まれる各記号に対応するグラム長を示す情報であり可変長情報記憶部で記憶されるグラム長情報と前記訓練データに含まれる各記号に対応するグラム長情報の示すグラム長より短いグラム長を有する代理の記号に関する情報であり前記可変長情報記憶部で記憶される代理情報とに対応する、前記訓練データに含まれる記号の接尾辞木を示す情報である接尾辞木情報を更新しながら、前記各記号の前記グラム長情報と前記代理情報とをギブスサンプリングにより算出して前記可変長情報記憶部に蓄積する処理を繰り返して実行するギブスサンプリング処理を行うギブスサンプリング部として機能させるためのものである。

20

30

【 0 1 6 6 】

このプログラムは、コンピュータを、さらに、前記訓練データと、前記グラム長情報と、前記代理情報とに対応する接尾辞木情報を用いて、テストデータ記憶部で記憶される、記号の並びを示すデータであるテストデータに含まれる記号の可変長 N グラム確率を、当該確率を算出する記号の階層 P i m a n - Y o r 過程における N グラム確率と、前記確率を算出する記号が N グラム長に到達する確率との積を各 N について足しあわせることによって算出する確率算出部と、前記確率算出部が算出した可変長 N グラム確率を出力する出力部として機能させてもよい。

40

【 0 1 6 7 】

なお、上記プログラムにおいて、上記プログラムが実現する機能には、ハードウェアでしか実現できない機能は含まれない。例えば、情報を出力する出力部などにおけるモデムやインターフェースカードなどのハードウェアでしか実現できない機能は、上記プログラムが実現する機能には少なくとも含まれない。

【 0 1 6 8 】

また、このプログラムは、サーバなどからダウンロードされることによって実行されて

50

もよく、所定の記録媒体（例えば、CD-ROMなどの光ディスクや磁気ディスク、半導体メモリなど）に記録されたプログラムが読み出されることによって実行されてもよい。

【0169】

また、このプログラムを実行するコンピュータは、単数であってもよく、複数であってもよい。すなわち、集中処理を行ってもよく、あるいは分散処理を行ってもよい。

【0170】

図23は、上記プログラムを実行して、上記実施の形態による情報処理装置を実現するコンピュータの外観の一例を示す模式図である。上記実施の形態は、コンピュータハードウェア及びその上で実行されるコンピュータプログラムによって実現される。

【0171】

図23において、コンピュータシステム100は、CD-ROM（Compact Disk Read Only Memory）ドライブ105、FD（Flexible Disk）ドライブ106を含むコンピュータ101と、キーボード102と、マウス103と、モニタ104とを備える。

【0172】

図24は、コンピュータシステムを示す図である。図24において、コンピュータ101は、CD-ROMドライブ105、FDドライブ106に加えて、CPU（Central Processing Unit）111と、ブートアッププログラム等のプログラムを記憶するためのROM（Read Only Memory）112と、CPU111に接続され、アプリケーションプログラムの命令を一時的に記憶すると共に、一時記憶空間を提供するRAM（Random Access Memory）113と、アプリケーションプログラム、システムプログラム、及びデータを記憶するハードディスク114と、CPU111、ROM112等を相互に接続するバス115とを備える。なお、コンピュータ101は、LANへの接続を提供する図示しないネットワークカードを含んでいてもよい。

【0173】

コンピュータシステム100に、上記実施の形態による情報処理装置の機能を実行させるプログラムは、CD-ROM121、またはFD122に記憶されて、CD-ROMドライブ105、またはFDドライブ106に挿入され、ハードディスク114に転送されてもよい。これに代えて、そのプログラムは、図示しないネットワークを介してコンピュータ101に送信され、ハードディスク114に記憶されてもよい。プログラムは実行の際にRAM113にロードされる。なお、プログラムは、CD-ROM121やFD122、またはネットワークから直接、ロードされてもよい。

【0174】

プログラムは、コンピュータ101に、上記実施の形態による情報処理装置の機能を実行させるオペレーティングシステム（OS）、またはサードパーティプログラム等を必ずしも含んでいなくてもよい。プログラムは、制御された態様で適切な機能（モジュール）を呼び出し、所望の結果が得られるようにする命令の部分のみを含んでいてもよい。コンピュータシステム100がどのように動作するのかについては周知であり、詳細な説明は省略する。

【0175】

また、本発明は、以上の実施の形態に限定されることなく、種々の変更が可能であり、それらも本発明の範囲内に包含されるものであることは言うまでもない。

【産業上の利用可能性】

【0176】

以上より、本発明による情報処理装置等によれば、可変長Nグラムを適切に扱うことができ、例えば、可変長Nグラムのモデルを生成する装置や、そのモデルによって、可変長Nグラム確率を算出する装置等として有用である。

【図面の簡単な説明】

【0177】

10

20

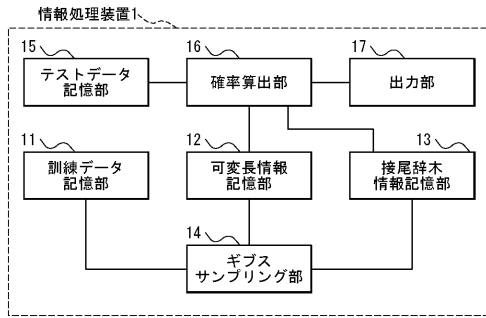
30

40

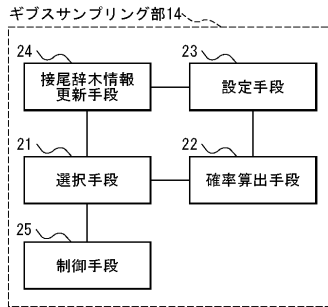
50

- 【図 1】本発明の実施の形態 1 による情報処理装置の構成を示すブロック図
- 【図 2】同実施の形態による情報処理装置におけるギブスサンプリング部の構成を示すブロック図
- 【図 3】同実施の形態における接尾辞木の一例を示す図
- 【図 4】同実施の形態における接尾辞木の一例を示す図
- 【図 5】同実施の形態における接尾辞木のノードを通過する確率について説明するための図
- 【図 6】同実施の形態による情報処理装置の動作を示すフローチャート
- 【図 7】同実施の形態による情報処理装置の動作を示すフローチャート
- 【図 8】同実施の形態による情報処理装置の動作を示すフローチャート 10
- 【図 9】同実施の形態による情報処理装置の動作を示すフローチャート
- 【図 10】同実施の形態における訓練データ、グラム長情報、代理情報に対応の一例を示す図
- 【図 11】同実施の形態における接尾辞木の実記号と代理記号との一例を示す図
- 【図 12】同実施の形態における記号と記号 ID との対応の一例を示す図
- 【図 13】同実施の形態における接尾辞木の一例を示す図
- 【図 14】同実施の形態における接尾辞木情報の一例を示す図
- 【図 15】同実施の形態における接尾辞木情報の一例を示す図
- 【図 16】同実施の形態による実験結果の一例を示す図
- 【図 17】同実施の形態による実験結果の一例を示す図 20
- 【図 18】同実施の形態における実験での n グラムオーダーの分布の一例を示す図
- 【図 19】同実施の形態におけるパラメータとパープレキシティとの関係の一例を示す図
- 【図 20】本発明の実施の形態 2 による情報処理装置の構成を示す図
- 【図 21】同実施の形態による情報処理装置の動作を示すフローチャート
- 【図 22】同実施の形態における、訓練データに含まれる記号列と、その記号列について求めた確率との対応の一例を示す図
- 【図 23】同実施の形態におけるコンピュータシステムの外観一例を示す模式図
- 【図 24】同実施の形態におけるコンピュータシステムの構成の一例を示す図
- 【符号の説明】
- 【0178】 30
- 1、2 情報処理装置
 - 11 訓練データ記憶部
 - 12 可変長情報記憶部
 - 13 接尾辞木情報記憶部
 - 14 ギブスサンプリング部
 - 15 テストデータ記憶部
 - 16 確率算出部
 - 17 出力部
 - 21 選択手段
 - 22 確率算出手段 40
 - 23 設定手段
 - 24 接尾辞木情報更新手段
 - 25 制御手段
 - 31 確率算出部
 - 32 記号列選択部
 - 33 出力部

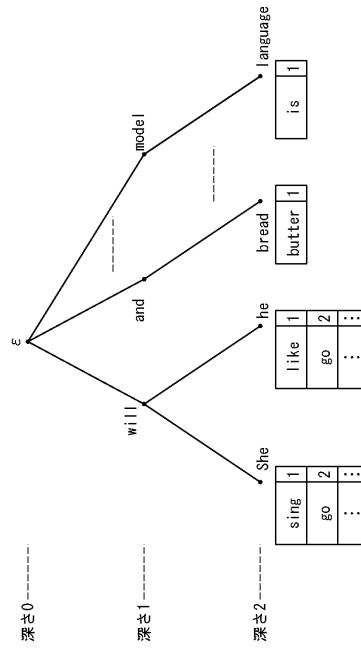
【図1】



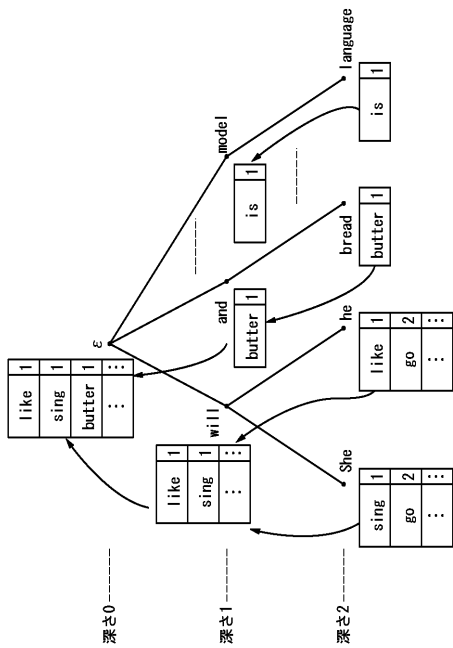
【図2】



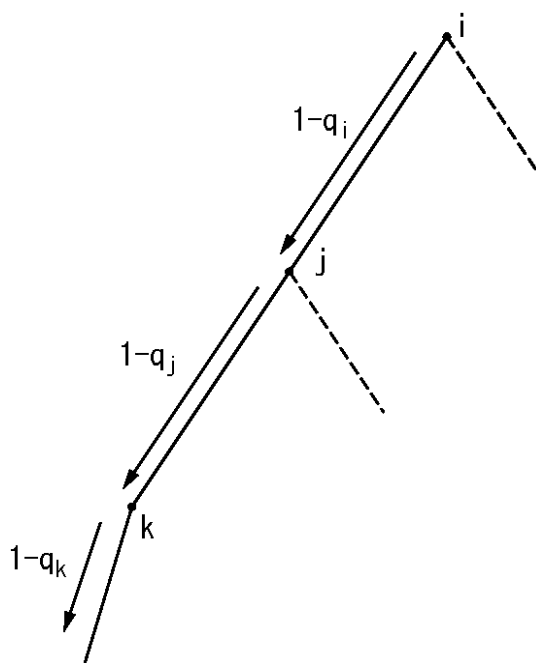
【図3】



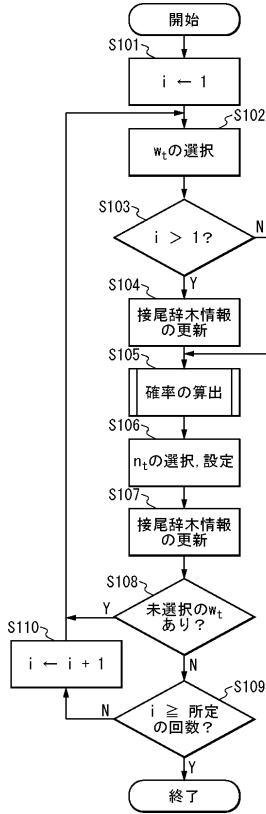
【図4】



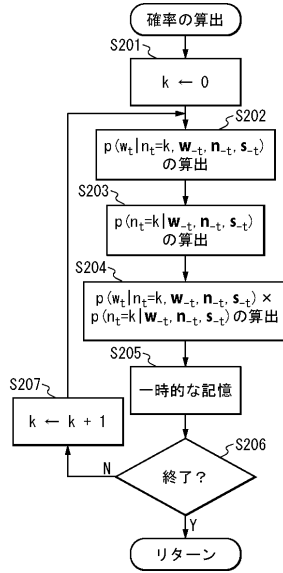
【図5】



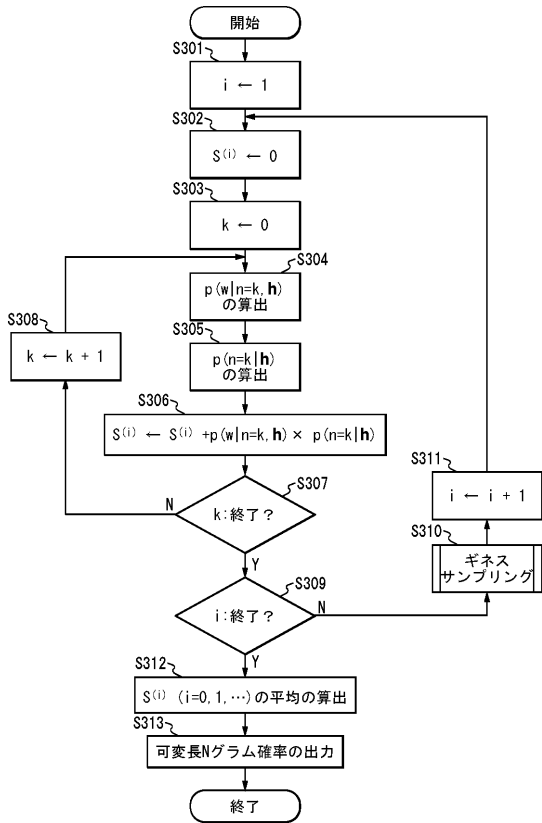
【図6】



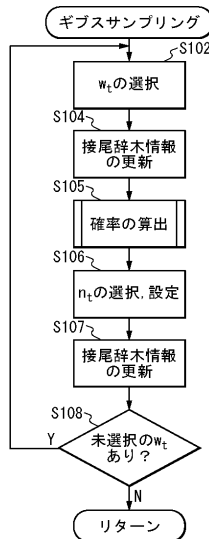
【図7】



【図8】



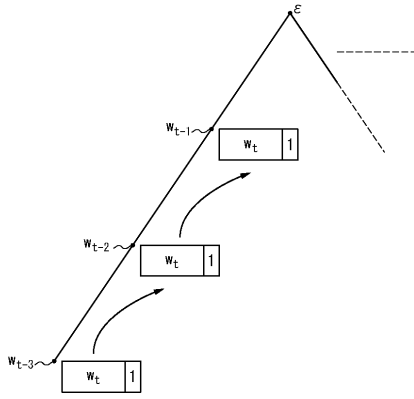
【図9】



【図10】

w	w ₁	w ₂	----	w _t	----	w _T	⇐ 訓練データ
n	n ₁	n ₂	----	n _t	----	n _T	⇐ グラム長情報
s	s ₁	s ₂	----	s _t	----	s _T	⇐ 代理情報

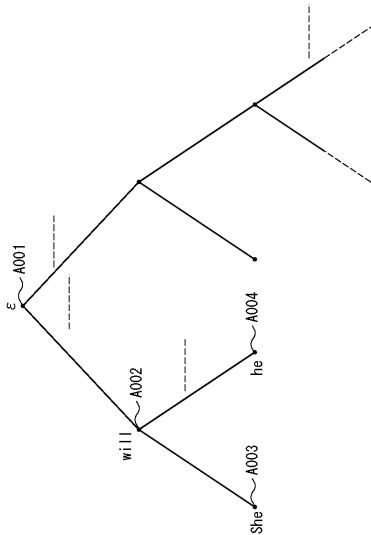
【図 1 1】



【図 1 2】

記号	記号ID
---	---
she	1001
he	1002
---	---
will	2051
---	---
like	3210
---	---
sing	4321
---	---

【図 1 3】



【図 1 4】

ノードのアドレス	記号ID	親アドレス	子アドレス	表記号		代理記号		停止回数 a	通過回数 b
				記号ID	値数	記号ID	値数		
---	---	---	---	---	---	---	---	---	---
A0002	2051	A0001	A0003 A0004	3210 4321	5 4	4321 5432	4 2	40	80
A0003	1001	A0002	-	4321 5432	7 5	-	-	30	0
---	---	---	---	---	---	---	---	---	---

接尾辞本情報

【図15】

接尾辞本情報	ノードのアドレス	---	---	---	---	---	---	---	---
	記号ID	---	2051	1001	---	---	---	---	---
	親アドレス	---	A0001	A0002	---	---	---	---	---
	子アドレス	---	A0003 A0004	-	---	---	---	---	---
	実記号	記号ID	---	3210	---	---	---	---	---
		個数	---	5	---	---	---	---	---
	代理記号	記号ID	---	4321	---	---	---	---	---
		個数	---	4	---	---	---	---	---
	停止回数 a	---	---	40	---	---	---	---	---
		---	---	29	---	---	---	---	---
	通過回数 b	---	---	79	---	---	---	---	---
		---	---	0	---	---	---	---	---

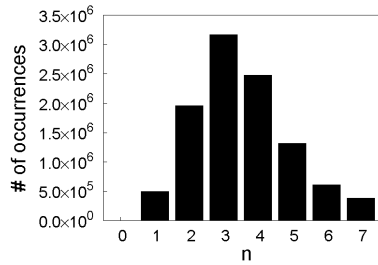
【図16】

n	HPYLM	VPYLM	Nodes (H)	Nodes (V)
3	113.60	113.74	1,417K	1,344K
5	101.08	101.69	12,699K	7,466K
7	N/A	100.68	N/A	10,182K
8	N/A	100.58	N/A	10,434K
∞	-	161.68	-	6,837K

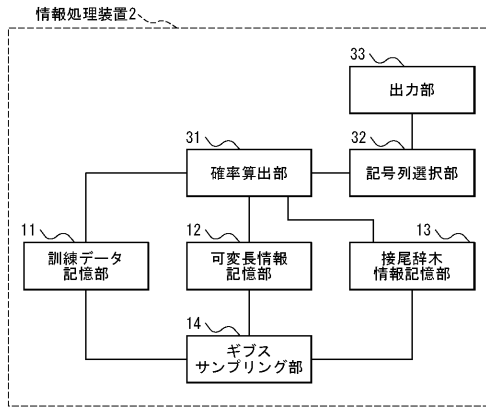
【図17】

n	HPYLM	VPYLM	Nodes (H)	Nodes (V)
3	78.06	78.22	1,341K	1,243K
5	68.36	69.35	12,140K	6,705K
7	N/A	68.63	N/A	9,134K
8	N/A	68.60	N/A	9,490K
∞	-	141.81	-	5,396K

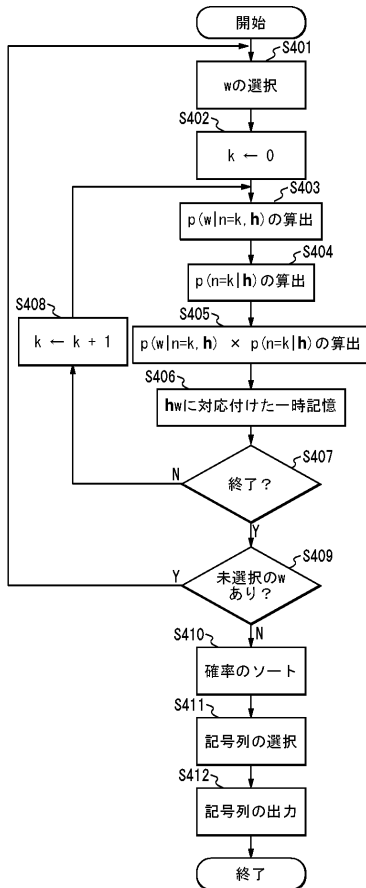
【図18】



【図20】



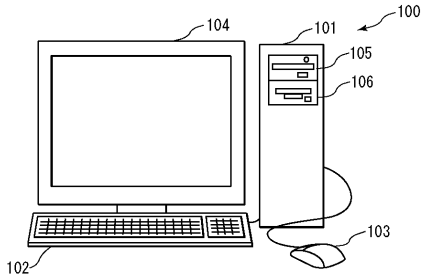
【図21】



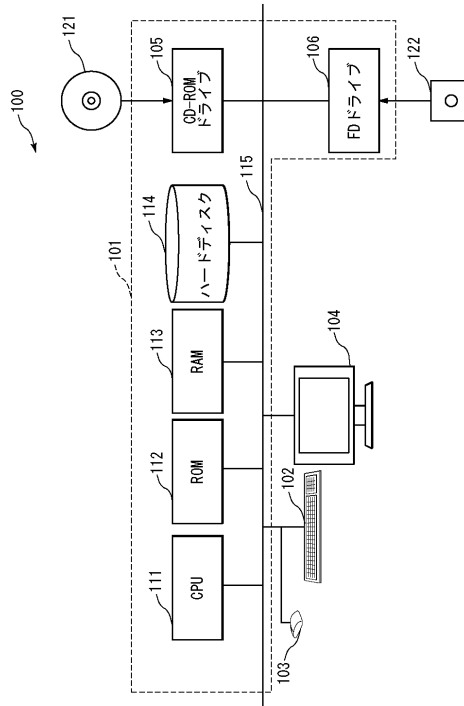
【 2 2 】

p	記号別
0.9784	primary new issues
0.9726	BOS at the same time
0.9556	american telephone &
0.9512	is a unit of
0.9394	to NUM % from NUM %
0.8896	in a number of
0.8831	in new york stock exchange composite trading
0.8696	a merrill lynch & co.
0.7566	mechanism of the european monetary
0.7134	increase as a result of
0.6617	tiffany & co.
.	.

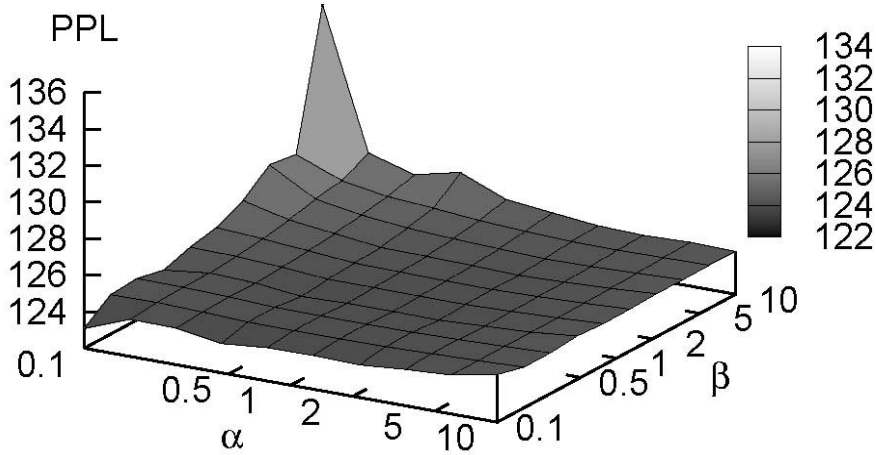
【 2 3 】



【 2 4 】



【 1 9 】



フロントページの続き

- (56)参考文献 加藤 智之 森 康真 荒木 康太郎 黒木 進 北上 始, ギブスサンプリングを用いた可変長配列パターンの抽出, 第69回(平成19年)全国大会講演論文集(4) インタフェース コンピュータと人間社会, 日本, 社団法人情報処理学会, 2007年 3月 6日, 4-671~4-672
中川 哲治, ギブスサンプリングを用いた係り受け解析, 情報処理学会研究報告 Vol. 2007 No. 47, 日本, 社団法人情報処理学会, 2007年 5月24日, 19~24
坪井 裕太, 頻出部分文字列のマイニング, 情報処理学会研究報告 Vol. 2003 No. 108, 日本, 社団法人情報処理学会, 2003年11月 7日, 147~154

(58)調査した分野(Int.Cl., DB名)

G06F 17/20 - 17/28