(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2009-3818
(P2009-3818A)
(43)公開日　平成21年1月8日(2009.1.8)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/28	(2006.01)	G06F	17/28	Z	5B009
G10L 15/18	(2006.01)	G10L	15/18	200D	5B091
G10L 15/06	(2006.01)	G10L	15/06	300C	5B109
G06F 17/21	(2006.01)	G10L	15/18	200F	5D015
		G06F	17/21	550A	

審査請求 未請求　請求項の数 10　ＯＬ　（全 34 頁）

(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)

特許法第３０条第１項適用申請有り　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／ｎｌ１７８ｖｐｙｌｍ．ｐｄｆ及びｈｔｔｐ：／／ｃｈａｓｅｎ．ｏｒｇ／~ｄａｉｔｉ－ｍ／ｐａｐｅｒ／ｎｌ１７８ｖｐｙｌｍ－ｓｌｉｄｅｓ．ｐｄｆを通じて発表

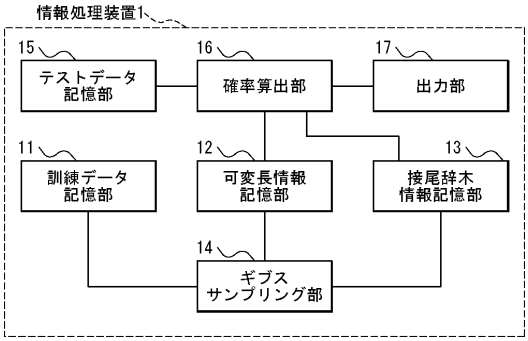
（特許庁注：以下のものは登録商標）
１．Ｌｉｎｕｘ

(71)出願人　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

Ｆターム(参考)　5B009 QA01
5B091 CA01　EA24
5B109 QA01
5D015 HH23

(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)【要約】
【課題】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を提供する。
【解決手段】記号の並びを示す訓練データが記憶される訓練データ記憶部１１と、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グラム長情報と、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代理情報とが記憶される可変長情報記憶部１２と、訓練データとグラム長情報と代理情報とに対応する、訓練データに含まれる記号の接尾辞木を示す接尾辞木情報が記憶される接尾辞木情報記憶部１３と、訓練データを用いて、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１２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１４と、を備える。
【選択図】図１



【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、を備えた情報処理装置。
【請求項２】
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、を備えた、請求項１記載の情報処理装置。
【請求項３】
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Ｐｉｍａｎ－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、請求項２記載の情報処理装置。
【請求項４】
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Ｐｉｍａｎ－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から請求項３いずれか記載の情報処理装置。
【請求項５】
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Ｐｉｍａｎ－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４記載の情報処理装置。
【請求項６】
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Ｐｉｍａｎ－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がＮ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から請求項３いずれか記載の情報処理装置。
【請求項７】
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６記載の情報処理装置。
【請求項８】
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７いずれか記載の情報処理装置。
【請求項９】
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、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た情報処理方法。
【請求項１０】
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プログラム。
【発明の詳細な説明】
【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、単語の種類の数を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。

【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

【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。

【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、それ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。

【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。

【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。

【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グラムとなる確率を、当該確率を算出する記号の階層Ｐｉｍａｎ－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。

【００１３】

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Ｐｉｍａｎ－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。

【００１４】

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。

【００１５】

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Ｐｉｍａｎ－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。

【００１６】

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Ｐｉｍａｎ－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がＮ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。

【００１７】

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。

【００１８】

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】

【００１９】

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ある。

【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もよく、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もよく、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もよい。

【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値は、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。

【００３１】

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１２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１１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１２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。

【００３２】

図２は、ギブスサンプリング部１４の構成を示すブロック図である。ギブスサンプリング部１４は、選択手段２１と、確率算出手段２２と、設定手段２３と、接尾辞木情報更新手段２４と、制御手段２５とを備える。

【００３３】
選択手段２１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２１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。
【００３４】
確率算出手段２２は、選択手段２１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２２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Ｐｉｍａｎ－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ことにより算出してもよい。この処理の詳細については後述する。
【００３５】
なお、確率算出手段２２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２２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もよい。また、例えば、確率算出手段２２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２２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２２は、定数倍の任意性のある確率を算出してもよい。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。
【００３６】
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、選択手段２１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１２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２３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。
【００３７】
また、設定手段２３は、その選択したグラム長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と共に、選択手段２１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Ｐｉｍａｎ－Ｙｏｒ過程により決定し、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２３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Ｐｉｔｍａｎ－Ｙｏｒ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。
【００３８】
接尾辞木情報更新手段２４は、選択手段２１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２３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２１が選択した記号に関する情報には、選択手段２１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。
【００３９】
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１２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２４は、選択手段２１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。

【００４０】

制御手段２５は、選択手段２１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２５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２５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。

【００４１】

テストデータ記憶部１５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１６における確率の算出で用いられるデータである。

【００４２】

テストデータ記憶部１５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１５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１５で記憶されるようになってもよく、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１５で記憶されるようになってもよく、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１５で記憶されるようになってもよい。

【００４３】

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、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。

【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。

【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。

【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。

【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うる。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ε）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「ε」、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（ｓｉｎｇ｜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（ｌｉｋｅ｜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（ｌｉｋｅ｜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（ｌｉｋｅ｜ｗｉｌｌ）とを用いて計算される。なお、ｐ（ｌｉｋｅ｜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】

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

【００５４】

ただし、$h=w_{t-n}\cdots w_{t-1}$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）＝Σ_wｃ（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。$t_{h\cdot}=\Sigma_w t_{hw}$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。$h'=w_{t-n+1}\cdots w_{t-1}$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】

接尾辞木の各ノードｉに、接尾辞木を根（ε）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。

【００５９】

【数２】

$$q_i\sim \mathrm{Be}(\alpha,\beta) \qquad (2)$$

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

【００６０】

記号の並び$h=w_{t-\infty}\cdots w_{t-2}w_{t-1}$に続いて記号$w_t$が観測されたとすると、接尾辞木を根εから初めて、$\varepsilon\rightarrow w_{t-1}\rightarrow w_{t-2}\cdots$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$_{2}$…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。

【００６１】

【数３】

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

【００６２】

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。

【００６３】

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。

【００６４】

まず、可変長Ｎグラムモデルでは、訓練データｗ＝ｗ$_{1}$ｗ$_{2}$…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$_{2}$…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、次式のようになる。

【００６５】

【数４】

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

【００６６】

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】

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

【００６８】

ここで、ｎ$_{-t}$，ｓ$_{-t}$は、それぞれ、ｎ，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】

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

【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】

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

【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】

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

【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】

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

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

【００７６】

ここで、式（１０）の第二項ｐ（ｎ｜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（ｗ｜ｎ，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。

（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】

（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。

【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。

【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。

【００８５】

（ステップＳ１０９）制御手段２５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２５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Ｓ１１０に戻る。

（ステップＳ１１０）制御手段２５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Ｓ１０２に戻る。

【００８６】

図７は、図６のフローチャートにおけるステップＳ１０５の処理の詳細を示すフローチャートである。

（ステップＳ２０１）確率算出手段２２は、カウンタｋを０に設定する。

【００８７】

（ステップＳ２０２）確率算出手段２２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

２２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。

【００８８】

（ステップＳ２０３）確率算出手段２２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２２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。

【００８９】

（ステップＳ２０４）確率算出手段２２は、ステップＳ２０２，Ｓ２０３で算出した値を掛け合わせることにより、式（６）の右辺の値を算出する。

（ステップＳ２０５）確率算出手段２２は、ステップＳ２０４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。

【００９０】

（ステップＳ２０６）確率算出手段２２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２２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Ｓ２０７に進む。

（ステップＳ２０７）確率算出手段２２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Ｓ２０２に戻る。

【００９１】

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Ｓ２０２，Ｓ２０３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Ｓ２０２，Ｓ２０３の値を利用することができる。したがって、ステップＳ２０２，Ｓ２０３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Ｓ２０２，Ｓ２０３の値の計算で用いるようにしてもよい。

【００９２】

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（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）の値を平均化する場合について説明する。

【００９３】

（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。

【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。

【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。

【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。

（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】

（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，…，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，…，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る。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とする。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する情報である。親アドレスは、根（ε）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。

【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。

【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。

【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…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＝０～７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。

【０１１４】

このような処理が、ｒａｎｄｐｅｒｍ（１…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…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…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】

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（ε）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（ε）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。

【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。

【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、（α，β）＝（４，１）を用いたが、後述するように、（α，β）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＝２であり、（α，β）＝（４，１）であり、ノードＡ０００１において、（ａ，ｂ）＝（１００，９００）であり、ノードＡ０００２，Ａ０００３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。

【０１２０】

【数１０】

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

【０１２１】

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。

【０１２２】

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（ε）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（ε）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）。

【０１２３】

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。

【０１２４】

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。

【０１２５】

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
で	1
い	4
た	3
ＥＯＳ	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うる。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うる。

【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、（α，β）＝（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、メモリオーバーフローを示す。

【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よいため、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てよく、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】

なお、図１９は、ＶＰＹＬＭのパラメータ（α，β）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、（α，β）∈（０．１～１０）×（０．１～１０）の範囲でパラメータ（α，β）の値を変化させている。この図１９からわかるように、β≫α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。

【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。
【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。
【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。
【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。
【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。

【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ってもよい。

【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。

【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Ｓ４０９に進み、そうでない場合には、ステップＳ４０８に進む。

【０１５４】

なお、このステップＳ４０７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Ｓ４０７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、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Ｓ４０７においても、ステップＳ４０４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、確率を算出する処理を終了すると判断してもよい。

【０１５５】

（ステップＳ４０８）確率算出部３１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Ｓ４０３に戻る。

（ステップＳ４０９）確率算出部３１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Ｓ４０１に戻り、そうでない場合には、ステップＳ４１０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Ｓ４０２～４０８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。

【０１５６】

（ステップＳ４１０）記号列選択部３２は、確率算出部３１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。

【０１５７】

（ステップＳ４１１）記号列選択部３２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３２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。

【０１５８】

（ステップＳ４１２）出力部３３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。

【０１５９】

なお、図２１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Ｓ４１０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。

【０１６０】

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。

【０１６１】

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２２に示す。図２２は、英語のＮＡＢコーパスで学習した８グラムのＶＰＹＬＭ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。

【０１６２】

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２２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。

【０１６３】

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もよい。

【０１６４】

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。

【０１６５】

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。

【０１６６】

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Ｐｉｍａｎ－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。

【０１６７】

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。

【０１６８】

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（Ｃｅｎｔｒ

ａｌ　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】

【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

【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

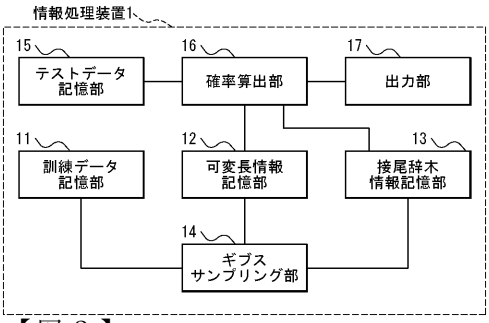
【図１２】同実施の形態における記号と記号ＩＤとの対応の一例を示す図

【図１３】同実施の形態における接尾辞木の一例を示す図

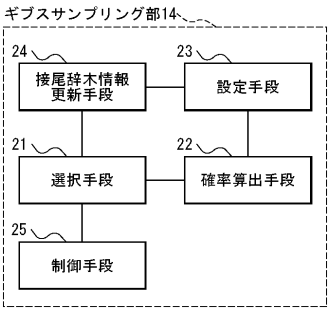
【図１４】同実施の形態における接尾辞木情報の一例を示す図

【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
【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】
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
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

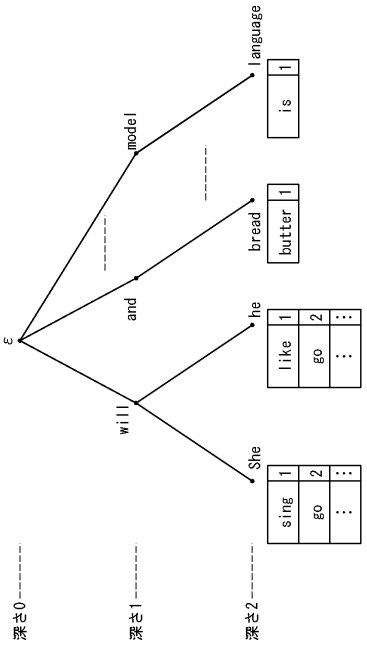
【図１】


情報処理装置1
15
テストデータ記憶部
16
確率算出部
17
出力部
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

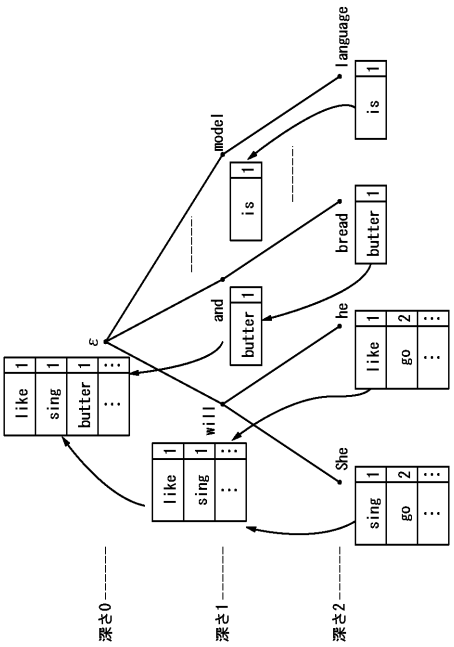
【図２】


ギブスサンプリング部14
24
接尾辞木情報更新手段
23
設定手段
21
選択手段
22
確率算出手段
25
制御手段

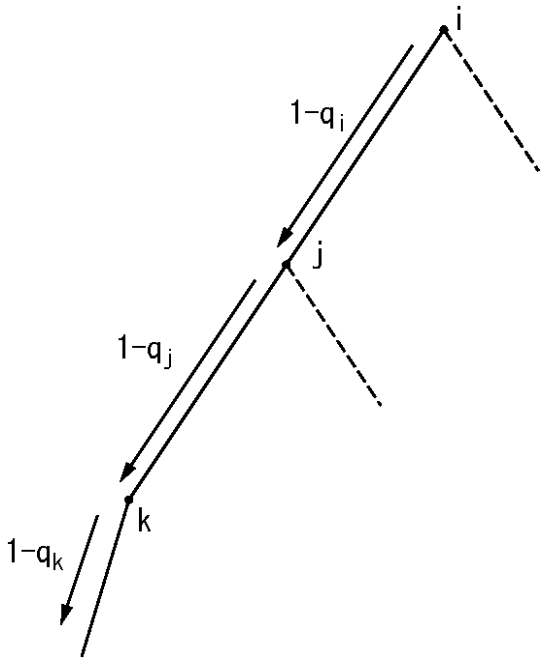
【図３】


ε
model
and
will
language
is
bread
butter
he
like
go
She
sing
深さ0
深さ1
深さ2

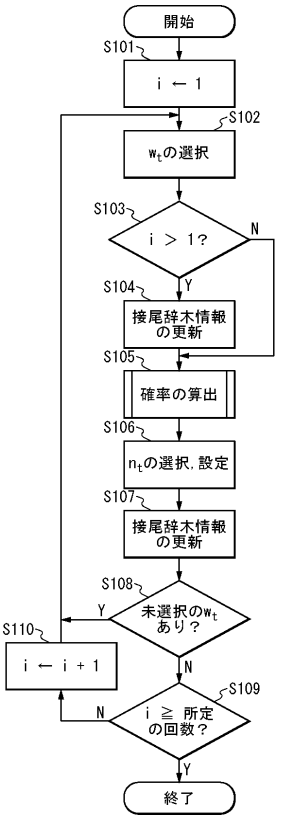
【図４】


language
model
bread
and
he
will
She
ε
like
sing
butter
go
深さ0
深さ1
深さ2

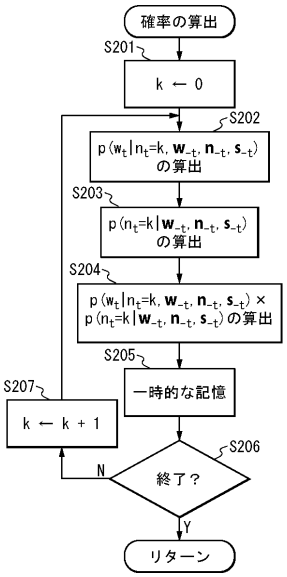
【図５】


i
j
k
1−q_i
1−q_j
1−q_k

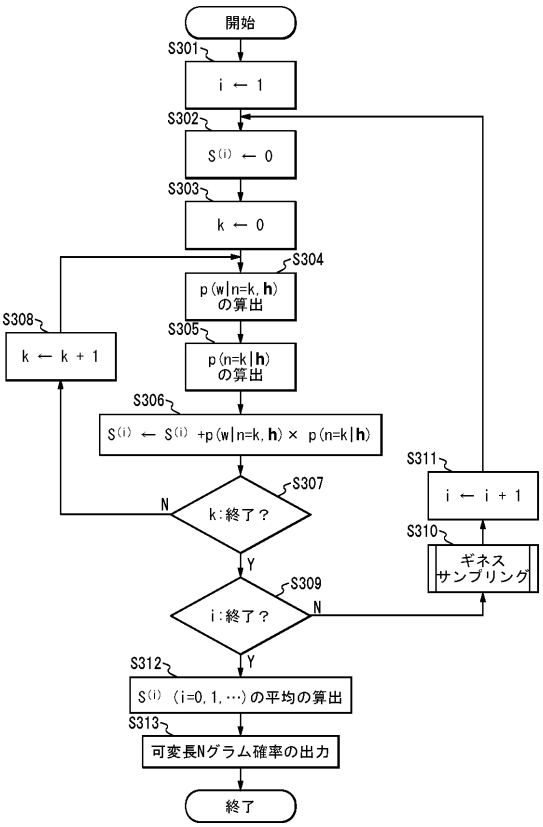
【図６】



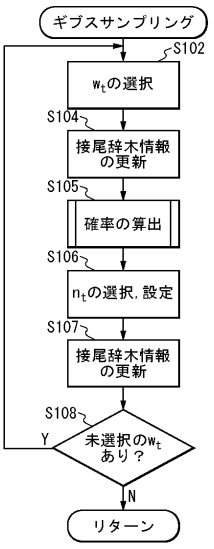
【図７】



【図８】



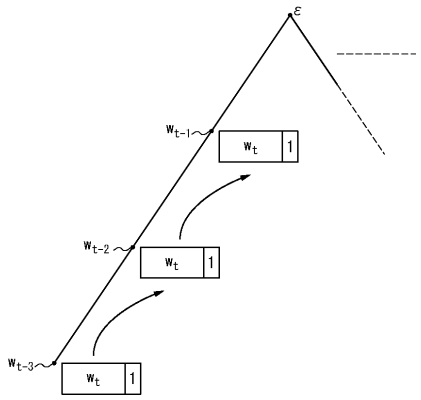
【図９】



【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

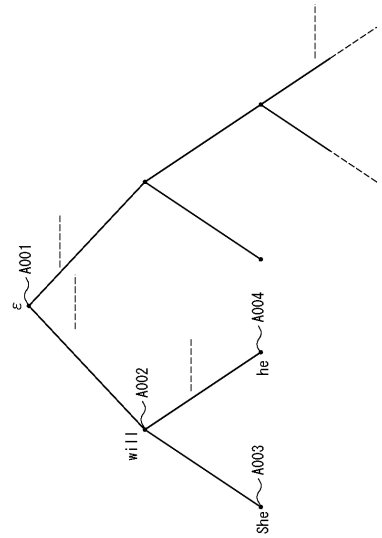
【図１１】



【図１２】

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

【図１３】



【図１４】

接尾辞木情報

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

【図１５】

接尾辞木情報

ノードのアドレス	記号ID	親アドレス	子アドレス	実記号 記号ID	実記号 個数	代理記号 記号ID	代理記号 個数	停止回数 a	通過回数 b
---	---	---	---	---	---	---	---	---	---
A0002	2051	A0001	A0003, A0004	3210	5	4321	3	40	79
				4321	4	5432	2		
					---	---	---		
A0003	1001	A0002	-	4321	6	-	-	29	0
				5432	5				

------	------	------	------	------	------	------	------	------	------

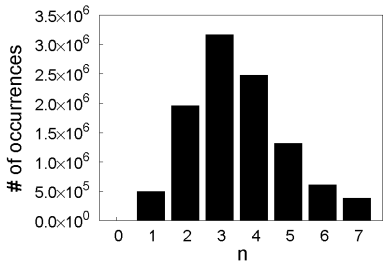
【図１６】

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

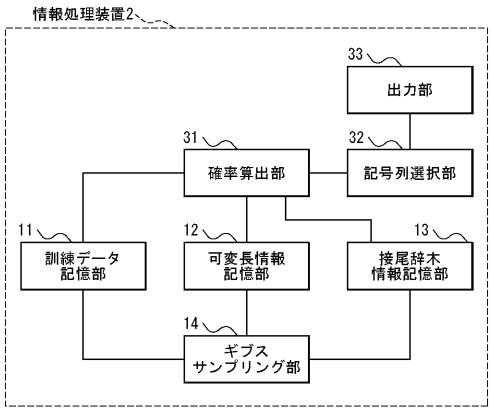
【図１７】

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

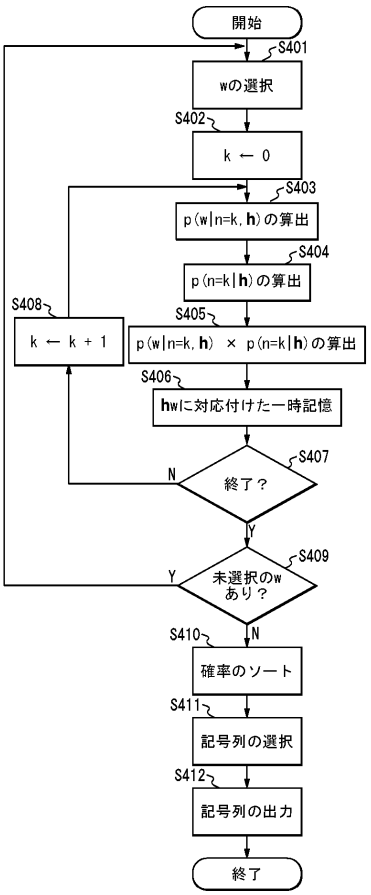
【図１８】



【図２０】



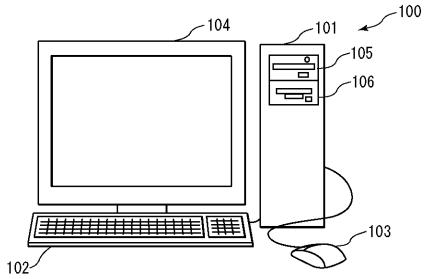
【図２１】



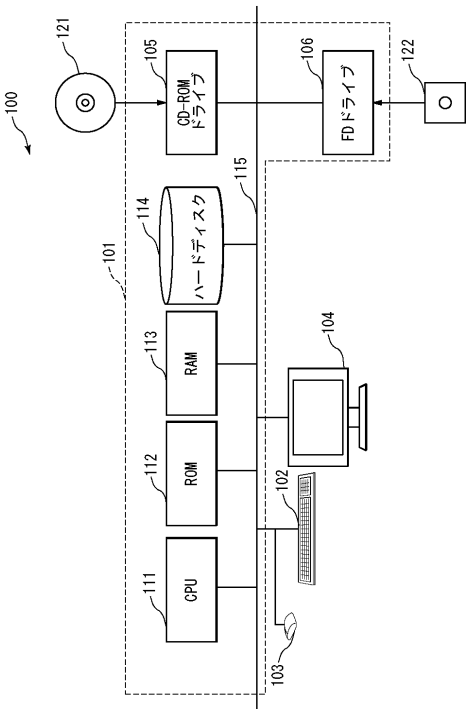
【図２２】

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.
：	

【図２３】



【図２４】



【図１９】