(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2005-292832
(P2005-292832A)
(43)公開日　平成17年10月20日(2005.10.20)

(51)Int.Cl.[7]	F I			テーマコード（参考）
G10L 15/18	G10L	3/00	537D	5D015
G10L 15/06	G10L	3/00	521R	

審査請求 未請求　請求項の数 16　OL　（全 16 頁）

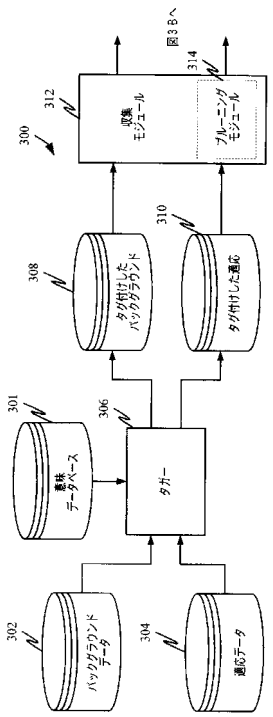
(21)出願番号　特願2005-101016 (P2005-101016)
(22)出願日　平成17年3月31日 (2005.3.31)
(31)優先権主張番号　10/814,906
(32)優先日　平成16年3月31日 (2004.3.31)
(33)優先権主張国　米国 (US)

(71)出願人　500046438
マイクロソフト　コーポレーション
アメリカ合衆国　ワシントン州　９８０５２－６３９９　レッドモンド　ワン　マイクロソフト　ウェイ
(74)代理人　100077481
弁理士　谷　義一
(74)代理人　100088915
弁理士　阿部　和夫
(72)発明者　アレジャンドロ　アチェロ
アメリカ合衆国　９８０５２　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション内

最終頁に続く

(54)【発明の名称】意味管理を用いた言語モデル適応

(57)【要約】
【課題】　言語モデルを適応させるための方法および装置を提供する。
【解決手段】　言語モデルを適応させるための方法および装置が提供される。この方法および装置は、インドメイン意味情報を利用して言語モデルの管理されたクラスベースの適応を提供する。
【選択図】　図３Ａ



【特許請求の範囲】
【請求項１】
Ｎグラム言語モデルを新しいドメインに適応させる方法であって、
前記新しいドメインに向けられていない一般的テキストフレーズを示すバックグラウンドデータを受け取ることと、
前記新しいドメインで使用され、クラスに編成された意味エンティティのセットを受け取ることと、
前記バックグラウンドデータ、前記意味エンティティおよびそのクラスに基づいてバックグラウンドＮグラムクラスカウントデータを生成することと、
前記バックグラウンドＮグラムクラスカウントデータに基づいて言語モデルをトレーニングすることと
を備えることを特徴とする方法。
【請求項２】
前記新しいドメインで使用されるテキストフレーズを示す適応データを受け取ることと、
前記適応データ、前記意味エンティティおよびそのクラスに基づいて適応Ｎグラムクラスカウントデータを生成することと
をさらに備え、
前記言語モデルをトレーニングすることは、前記バックグラウンドＮグラムクラスカウントデータおよび前記適応Ｎグラムクラスカウントデータに基づいてトレーニングすることを備えることを特徴とする請求項１に記載の方法。
【請求項３】
前記バックグラウンドＮグラムクラスカウントデータ、前記意味エンティティおよびそのクラスに基づいてバックグラウンドＮグラムワードデータを生成することと、
前記適応Ｎグラムクラスカウントデータ、前記意味エンティティおよびそのクラスに基づいて適応Ｎグラムワードデータを生成することと
をさらに備え、
前記バックグラウンドＮグラムクラスカウントデータおよび前記適応Ｎグラムクラスカウントデータに基づいて前記言語モデルをトレーニングすることは、バックグラウンドＮグラムワードデータおよび適応Ｎグラムワードデータを用いることを備えることを特徴とする請求項２に記載の方法。
【請求項４】
バックグラウンドＮグラムワードデータを生成することは、各データエントリが選択された数のワードを備えるマルチワード意味エンティティのバックグラウンドＮグラムワードデータを生成することを備えることを特徴とする請求項３に記載の方法。
【請求項５】
適応Ｎグラムワードデータを生成することは、各データエントリが選択された数のワードを備えるマルチワード意味エンティティの適応Ｎグラムワードデータを生成することを備えることを特徴とする請求項４に記載の方法。
【請求項６】
前記バックグラウンドデータ、前記意味エンティティおよびそのクラスに基づいてバックグラウンドＮグラムクラスカウントデータを生成することは、前記意味エンティティおよびそのクラスに基づいてワードレベルのバックグラウンドデータをタグ付けすることを備えることを特徴とする請求項４に記載の方法。
【請求項７】
前記適応データ、前記意味エンティティおよびそのクラスに基づいて適応Ｎグラムクラスカウントデータを生成することは、前記意味エンティティおよびそのクラスに基づいてワードレベルの適応データをタグ付けすることを備えることを特徴とする請求項５に記載の方法。
【請求項８】

前記バックグラウンドデータ、前記意味エンティティおよびそのクラスに基づいてバックグラウンドＮグラムクラスカウントデータを生成することは、前記タグ付けされたバックグラウンドデータの固有のクラスレベルＮグラムをカウントすることを備えることを特徴とする請求項６に記載の方法。

【請求項９】

前記適応データ、前記意味エンティティおよびそのクラスに基づいて適応Ｎグラムクラスカウントデータを生成することは、前記タグ付けされた適応データの固有のクラスレベルＮグラムをカウントすることを備えることを特徴とする請求項７に記載の方法。

【請求項１０】

前記バックグラウンドデータ、前記意味エンティティおよびそのクラスに基づいてバックグラウンドＮグラムクラスカウントデータを生成することは、前記タグ付けされたバックグラウンドデータからいくつかのクラスＮグラムを廃棄することを備えることを特徴とする請求項８に記載の方法。

【請求項１１】

前記適応データ、前記意味エンティティおよびそのクラスに基づいて適応Ｎグラムクラスカウントデータを生成することは、前記タグ付けされた適応データからいくつかのクラスＮグラムを廃棄することを備えることを特徴とする請求項９に記載の方法。

【請求項１２】

言語モデルを生成するステップを行うためのコンピュータ実行可能命令を有するコンピュータ可読媒体であって、前記ステップは、

選択されたドメインで使用され、クラスに編成された意味エンティティのセットを受け取るステップと、

前記意味エンティティのセットのクラスに相関され、一般テキストを示すバックグラウンドデータに基づくバックグラウンドＮグラムクラスカウントデータを受け取るステップと、

前記意味エンティティのセットのクラスに相関され、モデル化される選択されたドメインを示す適応データに基づく適応Ｎグラムクラスカウントデータを受け取るステップと、

前記バックグラウンドＮグラムクラスカウントデータ、前記適応Ｎグラムクラスカウントデータおよび前記意味エンティティのセットに基づいて言語モデルをトレーニングするステップと

を備えたことを特徴とするコンピュータ可読媒体。

【請求項１３】

前記言語モデルをトレーニングするステップは、前記バックグラウンドＮグラムクラスカウントデータおよび前記意味エンティティのセットに基づいてバックグラウンドワードカウントデータを算出するステップを備えたことを特徴とする請求項１２に記載のコンピュータ可読媒体。

【請求項１４】

前記言語モデルをトレーニングするステップは、前記適応Ｎグラムクラスカウントデータおよび前記意味エンティティのセットに基づいて適応ワードカウントデータを算出するステップを備えたことを特徴とする請求項１３に記載のコンピュータ可読媒体。

【請求項１５】

前記言語モデルをトレーニングするステップは、Ｎグラム相対頻度を平滑化するステップを備えたことを特徴とする請求項１４に記載のコンピュータ可読媒体。

【請求項１６】

平滑化するステップは、削除補間アルゴリズムを使用するステップを備えたことを特徴とする請求項１５に記載のコンピュータ可読媒体。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、言語処理で用いられる言語モデルに関する。詳細には、本発明は、所望のド

メインに言語モデルを適応させることに関する。
【背景技術】
【０００２】
自動音声認識（ＡＳＲ：ａｕｔｏｍａｔｉｃ ｓｐｅｅｃｈ ｒｅｃｏｇｎｉｔｉｏｎ）などの言語処理システムは、しばしば、トレーニングおよびテストデータと実際のドメインデータとの間の不一致に由来するエラーによる性能劣化を扱わなければならないことが多い。よく知られているように、音声認識システムは、音響モデルおよび統計言語モデル（ＬＭ：ｌａｎｇｕａｇｅ ｍｏｄｅｌ）を用いて認識を提供する。音響モデルの新しいドメインへの適応は、限られた成功でしか取り組まれておらず、言語モデルの適応は、満足のいく成果を達成していない。
【０００３】
統計言語モデル（ＬＭ）は、単語系列の事前確率推定を提供する。ＬＭは、最も可能性の高い単語系列の仮説探索（ｈｙｐｏｔｈｅｓｉｓ ｓｅａｒｃｈ）をガイドするので、ＡＳＲおよびその他の形態の言語処理において重要な構成要素である。よいＬＭは、優れた言語処理性能のために不可欠であることが知られている。
【０００４】
広く、ＬＭはテストデータに類似していると期待される大量のトレーニングデータから集められた、平滑化されたＮグラム統計（ｎ－ｇｒａｍ ｓｔａｓｔｉｓｔｉｃｓ）を使用する。しかしながら、類似性の定義はゆるく、対象となるある所与のドメインにどのデータソースを使用すべきかの決定は、ほとんどの場合試行錯誤により、普通、モデル作成者に委ねられる。
【０００５】
常に、トレーニングまたはテストデータと実際のドメインまたは「インドメイン（ｉｎ－ｄｏｍａｉｎ）」データとの間には不一致が存在し、それによりエラーがもたらされる。不一致の１つの源は、テストデータ中の語彙外の単語（ｏｕｔ－ｏｆ ｖｏｃａｂｕｌａｒｙ ｗｏｒｄｓ）に由来するものである。例えば、元々１つの航空会社のために設計された飛行機旅行情報システムは、問題の会社によってサービスが提供される都市名、空港名などでの不一致のために、別の会社に対してはうまく機能しないことがある。
【０００６】
別の潜在的な不一致の源は、異なる言語スタイルに由来するものである。例えば、ニュースドメインでの言語スタイルは、飛行機旅行情報ドメインとは異なる。ニュースワイヤやその他の一般的テキストでトレーニングされた言語モデルは、飛行機旅行情報ドメインではあまりうまく機能しないことがある。
【０００７】
【非特許文献１】Frederick Jelinek and Robert Mercer, "Interpolated Estimation of Markov Source Parameters from Sparse Data," In E. Gelsema and L. Kanal, editors, Pattern Recognition in Practice, pages 381-397, 1980
【発明の開示】
【発明が解決しようとする課題】
【０００８】
異なる技法を用いて大量のバックグラウンドデータでトレーニングされたＬＭを適応させるために様々なアプローチが試されてきたが、どれも優れた結果を達成しておらず、ゆえに、ＬＭ適応における改善が絶えず必要とされている。前述の問題の１つまたは複数に対処する方法があれば役立つであろう。
【課題を解決するための手段】
【０００９】
言語モデルを適応させるための方法および装置が提供される。この方法および装置は、インドメイン意味情報（ｉｎ－ｄｏｍａｉｎ ｓｅｍａｎｔｉｃ ｉｎｆｏｒｍａｔｉｏｎ）を利用して言語モデルの管理されたクラスベースの適応（ｓｕｐｅｒｖｉｓｅｄ ｃｌａｓｓ－ｂａｓｅｄ ａｄａｐｔａｔｉｏｎ）を提供する。

【００１０】
一般に、適応を行うのに使用されるリソースは、一般のテキストを示すバックグラウンドデータと、選択されたドメインで使用され、クラスに編成された意味エンティティのセットから導出される。別の実施形態では、モデル化される選択されたドメインを示す適応データも使用される。
【００１１】
前記の別の実施形態では、そのようなデータは、意味エンティティのセットのクラスに相関され、一般のテキストを示すバックグラウンドデータに基づくバックグラウンドＮグラムクラスカウントデータと、意味エンティティのセットのクラスに相関され、モデル化される選択されたドメインを示す適応データに基づく適応Ｎグラムクラスカウントデータとを備える。このデータから、そして意味エンティティのセットを使用して、バックグラウンドワードカウントデータおよび適応ワードカウントデータを計算し、適応データおよび意味項目のセットのドメインに言語モデルを適応させるためのベースとして使用することができる。
【発明を実施するための最良の形態】
【００１２】
本発明は、言語モデル適応のシステムおよび方法に関する。しかしながら、本発明をより詳細に論じる前に、本発明を使用することのできる１つの例示的な環境について論じることにする。
【００１３】
図１に、本発明を実施することのできる適したコンピューティング環境１００の一例を示す。コンピューティングシステム環境１００は、適したコンピューティング環境の一例にすぎず、本発明の使用または機能の範囲に関していかなる限定を示唆するものではない。また、コンピューティング環境１００は、例示的な動作環境１００に示す構成要素のいずれか１つまたはその組合せに関するいかなる依存性または要件を有するものと解釈されるべきではない。
【００１４】
本発明は、数多くの他の汎用または専用のコンピューティングシステム環境または構成で動作する。本発明と共に使用するのに適すると考えられる周知のコンピューティングシステム、環境、および／または構成の例には、それだけに限られないが、パーソナルコンピュータ、サーバコンピュータ、ハンドヘルドまたはラップトップデバイス、マルチプロセッサシステム、マイクロプロセッサベースのシステム、セットトップボックス、プログラム可能な民生用電子機器、ネットワークＰＣ、ミニコンピュータ、メインフレームコンピュータ、上記のシステムまたはデバイスのいずれかを含む分散コンピューティング環境などが含まれる。
【００１５】
本発明は、コンピュータにより実行される、プログラムモジュールなどのコンピュータ実行可能命令の一般的コンテキストで説明することができる。一般に、プログラムモジュールには、特定のタスクを行い、または特定の抽象データ型を実装するルーチン、プログラム、オブジェクト、コンポーネント、データ構造などが含まれる。当業者は、本明細書の説明および／または図をコンピュータ実行可能命令として実施することができ、それは、以下で論じる任意の形態のコンピュータ可読媒体上で具体化することができる。
【００１６】
本発明は、通信ネットワークを介してリンクされたリモート処理装置によってタスクが行われる分散コンピューティング環境で実施することもできる。分散コンピューティング環境では、プログラムモジュールは、メモリ記憶装置を含むローカルとリモート両方のコンピュータ記憶媒体に位置することがある。
【００１７】
図１を参照すると、本発明を実施するための例示的なシステムは、コンピュータ１１０の形態の汎用コンピューティングデバイスを含む。コンピュータ１１０の構成要素には、

それだけに限られないが、処理装置１２０、システムメモリ１３０、およびシステムメモリを含む様々なシステム構成要素を処理装置１２０に結合するシステムバス１２１が含まれ得る。システムバス１２１は、メモリバスまたはメモリコントローラ、周辺バス、および様々なバスアーキテクチャのいずれかを使用したローカルバスを含むいくつかのタイプのバス構造のいずれでもよい。例として、限定ではなく、そのようなアーキテクチャには、ＩＳＡ（Ｉｎｄｕｓｔｒｙ Ｓｔａｎｄａｒｄ Ａｒｃｈｉｔｅｃｔｕｒｅ）バス、ＭＣＡ（Ｍｉｃｒｏ Ｃｈａｎｎｅｌ Ａｒｃｈｉｔｅｃｕｔｒｅ）バス、拡張ＩＳＡ（ＥＩＳＡ）バス、ＶＥＳＡ（Ｖｉｄｅｏ Ｅｌｅｃｔｒｏｎｉｃｓ Ｓｔａｎｄａｒｄｓ Ａｓｓｏｃｉａｔｉｏｎ）ローカルバス、およびメザニンバスとも呼ばれるＰＣＩ（Ｐｅｒｉｐｈｅｒａｌ Ｃｏｍｐｏｎｅｎｔ Ｉｎｔｅｒｃｏｎｎｅｃｔ）バスが含まれる。

【００１８】

コンピュータ１１０は、通常、様々なコンピュータ可読媒体を含む。コンピュータ可読媒体は、コンピュータ１１０によってアクセスすることができる任意の利用可能な媒体とすることができ、それには揮発性媒体と不揮発性媒体の両方、リムーバブル媒体と非リムーバブル媒体の両方が含まれる。例として、限定ではなく、コンピュータ可読媒体には、コンピュータ記憶媒体および通信媒体が含まれ得る。コンピュータ記憶媒体には、コンピュータ可読命令、データ構造、プログラムモジュールまたはその他のデータなどの情報を記憶するための任意の方法または技術で実施された、揮発性および不揮発性の両方、リムーバブルおよび非リムーバブルの両方の媒体が含まれる。コンピュータ記憶媒体には、それだけに限られないが、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ、フラッシュメモリなどのメモリ技術、ＣＤ－ＲＯＭ、デジタル多用途ディスク（ＤＶＤ）などの光ディスク記憶、磁気カセット、磁気テープ、磁気ディスク記憶などの磁気記憶装置、または所望の情報の格納に使用でき、コンピュータ１１０によってアクセスすることができる他の任意の媒体が含まれる。通信媒体は、通常、コンピュータ可読命令、データ構造、プログラムモジュールまたはその他のデータを、搬送波ＷＡＶまたは他のトランスポート機構などの変調データ信号に具体化し、任意の情報配信媒体を含む。「変調データ信号」という用語は、信号に情報を符号化するような形でその特性の１つまたは複数が設定または変更された信号を意味する。例として、限定ではなく、通信媒体には、有線ネットワークや直接配線接続などの有線媒体、および音響、ＲＦ、赤外線、その他の無線媒体などの無線媒体が含まれる。上記のいずれの組合せも、コンピュータ可読媒体の範囲内に含まれるべきである。

【００１９】

システムメモリ１３０は、読取り専用メモリ（ＲＯＭ）１３１やランダムアクセスメモリ（ＲＡＭ）１３２などの揮発性および／または不揮発性メモリの形態のコンピュータ記憶媒体を含む。起動時などに、コンピュータ１１０内の要素間の情報転送を助ける基本ルーチンが入った基本入出力システム（ＢＩＯＳ）１３３は、通常、ＲＯＭ１３１に格納される。ＲＡＭ１３２は、通常、処理装置１２０によって直ちにアクセス可能であり、そして／または現在処理されているデータおよび／またはプログラムモジュールを収容する。例として、限定ではなく、図１に、オペレーティングシステム１３４、アプリケーションプログラム１３５、その他のプログラムモジュール１３６、およびプログラムデータ１３７を示す。

【００２０】

コンピュータ１１０は、他のリムーバブル／非リムーバブル、揮発性／不揮発性コンピュータ記憶媒体を含むことができる。例にすぎないが、図１に、非リムーバブルの不揮発性磁気媒体との間で読取りまたは書込みを行うハードディスクドライブ１４１、リムーバブルの不揮発性磁気ディスク１５２との間で読取りまたは書込みを行う磁気ディスクドライブ１５１、およびＣＤ－ＲＯＭや他の光媒体など、リムーバブルの不揮発性光ディスク１５６との間で読取りまたは書込みを行う光ディスクドライブ１５５を示す。例示的な動作環境で使用することのできる他のリムーバブル／非リムーバブル、揮発性／不揮発性のコンピュータ記憶媒体には、それだけに限られないが、磁気テープカセット、フラッシュメモリカード、デジタル多用途ディスク、デジタルビデオテープ、ソリッドステートＲＡ

M、ソリッドステートＲＯＭなどが含まれる。ハードディスクドライブ１４１は、通常、インターフェース１４０などの非リムーバブルメモリインターフェースを介してシステムバス１２１に接続され、磁気ディスクドライブ１５１および光ディスクドライブ１５５は、通常、インターフェース１５０などのリムーバブルメモリインターフェースによってシステムバス１２１に接続される。

【００２１】

前述し、図１に示すドライブおよび関連するコンピュータ記憶媒体は、コンピュータ可読命令、データ構造、プログラムモジュールおよびコンピュータ１１０のその他のデータのストレージを提供する。図１では、例えば、ハードディスクドライブ１４１は、オペレーティングシステム１４４、アプリケーションプログラム１４５、その他のプログラムモジュール１４６、およびプログラムデータ１４７を格納するものとして示されている。これらのコンポーネントは、オペレーティングシステム１３４、アプリケーションプログラム１３５、その他のプログラムモジュール１３６、およびプログラムデータ１３７と同じでも、異なっていてもよいことに留意されたい。オペレーティングシステム１４４、アプリケーションプログラム１４５、その他のプログラムモジュール１４６、およびプログラムデータ１４７には、少なくともそれらが異なるコピーであることを示すために、図では異なる番号が付与されている。

【００２２】

ユーザは、キーボード１６２や、マイクロフォン１６３などの入力装置、マウス、トラックボール、タッチパッドといったポインティングデバイス１６１を介してコンピュータ１１０にコマンドおよび情報を入力することができる。他の入力装置（図示せず）には、ジョイスティック、ゲームパッド、衛星アンテナ、スキャナなどが含まれることがある。上記およびその他の入力装置は、システムバスに結合されたユーザ入力インターフェース１６０を介して処理装置１２０に接続されるが、パラレルポート、ゲームポート、ユニバーサルシリアルバス（ＵＳＢ）といった他のインターフェースおよびバス構造によって接続することもできる。モニタ１９１または他の種類の表示装置もビデオインターフェース１９０などのインターフェースを介してシステムバス１２１に接続される。モニタ１９１に加えて、コンピュータはスピーカ１９７やプリンタ１９６など他の周辺出力装置を含むこともでき、それらは、出力周辺インターフェース１９５を介して接続することができる。

【００２３】

コンピュータ１１０は、リモートコンピュータ１８０など、１つまたは複数のリモートコンピュータへの論理接続を使用してネットワーク化された環境で動作することができる。リモートコンピュータ１８０は、パーソナルコンピュータ、ハンドヘルド装置、サーバ、ルータ、ネットワークＰＣ、ピアデバイスおよびその他の共通ネットワークノードとすることができ、通常、コンピュータ１１０に関連して上述した要素の多くまたはすべてを含む。図１に示す論理接続には、ローカルエリアネットワーク（ＬＡＮ）１７１およびワイドエリアネットワーク（ＷＡＮ）１７３が含まれるが、他のネットワークを含むこともある。そのようなネットワーク環境は、オフィス、企業規模のコンピュータネットワーク、イントラネットおよびインターネットで一般的である。

【００２４】

ＬＡＮネットワーク環境で使用されるとき、コンピュータ１１０はネットワークインターフェースまたはアダプタ１７０を介してＬＡＮ１７１に接続される。ＷＡＮネットワーク環境で使用されるとき、コンピュータ１１０は、通常、モデム１７２、またはインターネットなどのＷＡＮ１７３を介して通信を確立するための他の手段を含む。モデム１７２は、内蔵でも外付けでもよく、ユーザ入力インターフェース１６０または他の適当な機構を介してシステムバス１２１に接続することができる。ネットワーク化環境では、コンピュータ１１０に関連して示すプログラムモジュール、またはその部分は、リモートメモリ記憶装置に格納することもできる。例として、限定ではなく、図１に、リモートアプリケーションプログラム１８５をリモートコンピュータ１８０上にあるものとして示す。図示

したネットワーク接続は例示的なものであり、コンピュータ間で通信を確立する他の手段を使用できることが理解されるであろう。

【００２５】

本発明は、図１との関連で説明したようなコンピュータシステム上で実行できることに留意すべきである。しかしながら、本発明は、サーバや、メッセージ処理専用のコンピュータでも、本発明の異なる部分が分散コンピューティングシステムの異なる部分で実行される分散システムでも実行することができる。

【００２６】

上に示したように、本発明は、言語モデル適応のためのシステムおよび方法に関するものである。適応を行うのに使用されるリソースは、適応される必要のあるバックグラウンドＬＭを含む。一般に、バックグラウンドＬＭは、それだけに限られないが、ニュース記事など、バックグラウンドトレーニングデータの大きなコーパスから得られる。このバックグラウンドトレーニングデータを使用して、バックグラウンド言語モデルのためのＮグラム統計が得られる。

【００２７】

意味データベースまたは意味情報は、適応のための管理された情報（ｓｕｐｅｒｖｉｓｅｄ　ｉｎｆｏｒｍａｔｉｏｎ）を提供する。この説明では、意味データベースは意味エンティティ（クラス）のリストを広く、模式的に表し、それらがインドメイン自然言語テキストにおいて遭遇するのと同じ形態であると想定される具現化（ｒｅａｌｉｚａｔｉｏｎｓ）のリストを各エンティティが伴う。例えば、意味データベースは、複数のクラスについて一般的に明確に定義された意味エンティティのリストの形態とすることができる。例えば、以下に一例として使用するように、予約を行う旅行情報を得るために航空会社によって使用される音声認識における言語モデルの意味項目には、その航空会社によってサービスが提供される都市および飛行先の様々な空港のリストが含まれ得る。意味エンティティおよびクラスの別の例は、企業の従業員、当月の日、当年の月のリストとすることができ、それらは、おそらく、スケジューリング用途のインドメインに含まれるであろう。

【００２８】

意味クラスは、開いたクラスと閉じたクラスに分類することができる。開いたクラス中のクラスメンバは、ドメイン間で変化するが、閉じたクラス中のクラスメンバは変化しない。例えば、飛行機旅行用途での意味クラスは以下のものとすることができるであろう。

【００２９】

開いたクラス：｛ＡＩＲＬＩＮＥ（航空会社）、ＡＩＲＰＯＲＴ（空港）、ＣＩＴＹ（都市）、ＳＴＡＴＥ（州）｝

閉じたクラス：｛ＤＡＹＳ（日）、ＭＯＮＴＨ（月）、ＩＮＴＥＲＶＡＬ（間隔）、ＣＬＡＳＳ　ＯＦ　ＳＥＲＶＩＣＥ（サービスクラス）、ＺＯＮＥ（ゾーン）、ＦＯＯＤ　ＳＥＲＶＩＣＥ（フードサービス）、ＧＲＯＵＮＤ　ＳＥＲＶＩＣＥ（地上サービス）｝

用途ごとに、意味クラスの数およびタイプは大きく変化する可能性がある。しかしながら、飛行機旅行用途などいくつかの用途では、その意味クラスが特定されると、その意味エンティティまたは単語（ｗｏｒｄ）レベルの具現化だけで、別の航空会社による使用のためにその言語モデルを本質的に適応させるために変更する必要のあるすべてとすることができる。

【００３０】

言語モデル適応で使用されるオプションの第３のリソースは適応データである。適応データは、インドメインのアプリケーションでのクラスの使用例となりうるセンテンス（ｓｅｎｔｅｎｃｅｓ）、フレーズ（ｐｈｒａｓｅｓ）、テキストセグメント（ｔｅｘｔ　ｓｅｇｅｍｅｎｔｓ）などの形態の実際のまたはインドメインのデータを備える。バックグラウンドデータに比べて、適応データは、普通、バックグラウンドデータより何桁も少ない。一実施形態では、インドメインデータは、適応開発データ（ａｄａｐｔａｔｉｏｎ　ｄｅｖｅｌｏｐｅｍｅｎｔ　ｄａｔａ）と適応トレーニングデータ（ａｄａｐｔａｔｉｏ

ｎ　ｔｒａｉｎｉｎｇ　ｄａｔａ）にサブ分割される。適応トレーニングデータは、バックグラウンドトレーニングセットと組み合わされてより大きなトレーニングセットになり、両セットからのＮグラムカウントは等しい重みで混合される（ただし、他の混合スキームも可能であり、Ｎグラムカウントは、ＭＡＰ適応などのように、異なる重みで混合することもできる）。適応開発データは、厳密に、バックグラウンド言語モデルおよび適応言語モデルの両方を平滑化するために使用される。開発セットからのＮグラムは、バックグラウンド／適応言語モデルには含められない。

【００３１】

例示的な実施形態では、すべてのデータセットは単語レベルの自然言語テキストである。

【００３２】

（クラスベースの適応）

管理された意味情報（ｓｕｐｅｒｖｉｓｅｄ　ｓｅｍａｎｔｉｃ　ｉｎｆｏｒｍａｔｉｏｎ）がクラスベースの言語モデルの使用を通じて言語モデルに組み込まれる。簡潔には、単一の意味クラスc_3に属する新しい単語w_3の確率推定は以下のように行うことができる。

【００３３】

$$Pr(w_3 | w_2 w_1) = Pr(w_3 | c_3) \cdot Pr(c_3 | w_2 w_1) \quad (1)$$

ここで、$Pr(w_3 | c_3 w_2 w_1) = Pr(w_3 | c_3)$というモデルリングを前提とする。

【００３４】

例えば、Ｐｒ（ｃｉｔｙ　ｎａｍｅ（都市名）｜ｆｌｙ　ｔｏ（へ飛行））は以下を用いて推定される。

【００３５】

Ｐｒ（ｃｉｔｙ　ｎａｍｅ｜ｆｌｙ　ｔｏ）＝Ｐｒ（ｃｉｔｙ　ｎａｍｅ｜ＣＩＴＹ（都市））・Ｐｒ（ＣＩＴＹ｜ｆｌｙ　ｔｏ）

ここで、Ｐｒ（ＣＩＴＹ｜ｆｌｙ　ｔｏ）は、意味クラスでタグ付けされたトレーニングデータを用いて推定され、Ｐｒ（ｃｉｔｙ　ｎａｍｅ｜ＣＩＴＹ）は、インドメイン意味データベースを用いて適応される。以前のインドメイン知識が利用可能である場合、よくある都市名は、まれな都市名より高い確率で割り当てることができ、そうでなければ、都市名の一様分布が想定される。クラスベースの適応アプローチを用いることの利点は以下の通りである。

【００３６】

単語コンテキストを仮定した意味クラスの確率はうまく推定することができる。上記の例では、Ｐｒ（ｃｉｔｙ　ｎａｍｅ｜ｆｌｙ　ｔｏ）は、トレーニングデータおよび適応データで非常に類似していると考えられる。

【００３７】

インドメイン意味データを用いて$Pr(w_3 | c_3)$を適応させることにより、高速ＬＭ適応を行なうことができる。適応された確率$Pr(w_3 | c_3)$は、ドメイン特有の言語モデルを再トレーニングする新しいトレーニングテキストを収集することなく、カウント「$w_1 w_2 w_3$」と組み合わされる。

【００３８】

語句（ｗｏｒｄ　ｐｈｒａｓｅｓ）が意味クラスにカプセル化されるので、より広範な単語コンテキストで確率推定を実現することができる。例えば、５グラム「ｌｏｓ　ａｎｇｅｌｅｓ　ｔｏ　ｎｅｗ　ｙｏｒｋ」は、トライグラム（ｔｒｉｇｒａｍｓ）「ｌｏｓ　ａｎｇｅｌｅｓ　ｔｏ」、「ａｎｇｅｌｅｓ　ｔｏ　ｎｅｗ」および「ｔｏ　ｎｅｗ　ｙｏｒｋ」のシーケンスとしてモデル化されるよりも直感的に満足の行くクラストライグラム「ＣＩＴＹ　ｔｏ　ＣＩＴＹ」としてモデル化される。

【００３９】

（適応手順）

図２に例示的な適応手順２００を示す。図３Ａおよび３Ｂに、手順２００を行うための例示的システム３００を示す。上に示したように、適応データの使用はオプションであるが、本発明の別の実施形態である。両方を用いた実施形態について以下に説明するが、これは必須または限定とみなすべきではない。また、説明に進む前に、手順２００およびシステム３００が、一般に同時にバックグラウンドデータおよびオプションの適応データ上で動作するものとして記述されていることに留意されたい。しかしながら、これは、理解を簡単にするためであり、必要または限定とみなすべきではない。

【００４０】

ステップ２０２は、一般に、バックグラウンドデータおよび適応データの両方のためにタグ付けデータを得ることを表している。例示した実施形態では、これは、２０２に示すように単語レベルのデータをタグ付けすることを含む。特に、トレーニング（バックグラウンドおよび適応）データが、図２のステップ２０２で意味クラスと先ずタグ付けされる。当然ながら、タグ付けされたデータが存在する場合、このステップは不要である。図３Ａでは、意味データベースが３０１で示され、トレーニングデータはコーパス３０２および３０４にあり、この場合、タグ付けはタガー３０６によって行われる。

【００４１】

タガー３０６は、コーパス３０２および３０４によって提供された単語レベルのテキストを変更し、そこで認識された意味エンティティのクラスを示すタグを付加する。例えば、「ｆｌｙ ｆｒｏｍ ｓａｎ ｆｒａｎｃｉｓｃｏ ｔｏ」が与えられ、「ｓａｎ ｆｒａｎｃｉｓｃｏ」が意味クラス「ＣＩＴＹ」に属すると知っている場合、タガー３０６からの出力は、「ｆｌｙ ｆｒｏｍ ＣＩＴＹ ｔｏ」になることになる。意味エンティティのいくつかが対応する意味クラスで置き換えられた単語レベルのトレーニングデータは３０８および３１０で示されている。

【００４２】

一実施形態では、タグ付けにヒューリスティックス（ｈｅｕｒｉｓｔｉｃｓ）を適用することができる。そのようなヒューリスティックスは、タグ付けのための単純な文字列マッチングアプローチを含んでもよい。タガー３０６は、所与のデータベースエントリをテキスト中の単語のシーケンスと合致させ、そのようにして特定された最長のフレーズにクラスラベルを割り当てる。別の実施形態では、単語のあいまい性が異なるクラス間で発生した場合、その語句はタグなしのままとされる。別の実施形態では、各意味クラス候補に確率を割り当てることによって、ソフトなタグ付けを行うことができるであろう。

【００４３】

ステップ２０２でタグ付けが行われた後、タグ付けデータが、別途、提供されない場合、手順はステップ２０４に進んですべてのトレーニングテキストからクラスＮグラムカウントを収集し、そうでない場合は、タグ付けデータに含まれる固有のＮグラムをカウントする。図３Ａで、このステップは収集モジュール３１２によって行われる。

【００４４】

クラスＮグラムカウントのプルーニングを備えるオプションのステップ２０６を必要に応じて行うこともできる。クラスベースの適応では、クラスＮグラムが単語Ｎグラムに拡張されると、言語モデルのサイズは、各意味クラス中の要素数によって強く影響される。例えば、クラストライグラム「ＰＥＲＳＯＮ ｊｏｉｎｓ ＣＯＭＰＡＮＹ（「人」が「会社」に入社する）」（この場合、「ＰＥＲＳＯＮ」および「ＣＯＭＰＡＮＹ」が意味クラスを含む）は、「ＰＥＲＳＯＮ」と「ＣＯＭＰＡＮＹ」がそれぞれ何千ものクラス要素を含むとき何百万もの単語トライグラムを生じる。それゆえ、言語モデルプルーニングが、言語モデルのサイズを扱いやすくするために必要になることがある。一実施形態では、複数の意味クラスを含むＮグラムは廃棄される。計算リソースが利用可能であれば、それらを保持することもできるであろう。加えて、単語Ｎグラムに拡張する前に、クラスＮグラムのカウントカットオフプルーニングを用いることができる。図３Ａには、収集モジュール３１２がプルーニングモジュール３１４を使用することによってこの機能を行うもの

してと示されている。収集モジュール３１２からの出力は、図３Ｂに示すバックグラウンドＮグラムカウントデータ３１６および適応Ｎグラムカウントデータ３１８を備える。

【００４５】

ステップ２０８で、クラスＮグラムが意味データベース３０１を用いて単語Ｎグラムに拡張される。図３Ｂでは、このステップはワードＮグラムジェネレータ３２０によって行われる。一実施形態では、ワードＮグラムジェネレータ３２０は、以下の拡張アルゴリズムを実施し、バックグラウンドＮグラムワードカウントデータ３２２および適応Ｎグラムワードカウントデータ３２４を生成することができる。

【００４６】

（ａ）クラスＮグラムを仮定し、クラスタグをそのクラス要素のそれぞれによって置き換える。

【００４７】

例えば、クラストライグラム「ａｎａｌｙｓｔ　ｆｏｒ　ＣＯＭＰＡＮＹ」は、単語４グラム「ａｎａｌｙｓｔ　ｆｏｒ　ｘ．ｙ．」を作成することができ、ここで、「ｘ．ｙ．」は意味データベース中の会社名（Ｖｅｒｉｚｏｎ　Ｗｉｒｅｌｅｓｓなど）である。

【００４８】

（ｂ）クラスＮグラムカウントから単語Ｎグラムカウントを算出する。

【００４９】

単語Ｎグラムカウントは、Ｐｒ（ｗｏｒｄ｜ｃｌａｓｓ）に応じて、その対応するクラスＮグラムカウントの一部として算出される。

【００５０】

意味クラス「ＣＯＭＰＡＮＹ」の確率が、

Ｐｒ（ｍｉｃｒｏｓｏｆｔ｜ＣＯＭＰＡＮＹ）＝０．５

Ｐｒ（ｏｒａｃｌｅ｜ＣＯＭＰＡＮＹ）＝０．２５

Ｐｒ（ｖｅｒｉｚｏｎ　ｗｉｒｅｌｅｓｓ｜ＣＯＭＰＡＮＹ）＝０．２５

であり、

Ｎグラム「ａｎａｌｙｓｔ　ｆｏｒ　ＣＯＭＰＡＮＹ」が５カウントであったと想定すると、

単語レベルのＮグラムカウントデータは、

「ａｎａｌｙｓｔ　ｆｏｒ　ｍｉｃｒｏｓｏｆｔ」＝２．５

「ａｎａｌｙｓｔ　ｆｏｒ　ｏｒａｃｌｅ」＝１．２５

「ａｎａｌｙｓｔ　ｆｏｒ　ｖｅｒｉｚｏｎ　ｗｉｒｅｌｅｓｓ」＝１．２５

になるであろう。

【００５１】

上記の例では、生成された単語４グラム「ａｎａｌｙｓｔ　ｆｏｒ　ｘ．ｙ．」のカウントは、

＃（「ａｎａｌｙｓｔ　ｆｏｒ　ＣＯＭＰＡＮＹ」）・Ｐｒ（「ｘ．ｙ．」｜ＣＯＭＰＡＮＹ）

に等しい。

【００５２】

（ｃ）しかしながら、クラスベースのＮグラムは、マルチワード（ｍｕｌｔｉ－ｗｏｒｄ）意味エントリのために特定のＮグラムのトレーニングと動作しない単語レベルのＮグラムを生成することができることに留意されたい。例えば、３ワードのＮグラム言語モデルが望まれていると想定すると、「ａｎａｌｙｓｔ　ｆｏｒ　ｖｅｒｉｚｏｎ　ｗｉｒｅｌｅｓｓ」は正しい形態のものではない。この状況では、スライディングウィンドウを用いてより低次の単語Ｎグラムが生成される。上記の例では、「ａｎａｌｙｓｔ　ｆｏｒ　ｖｅｒｉｚｏｎ」も１．２５のカウントを有するであろうし、「ｆｏｒ　ｖｅｒｉｚｏｎ　ｗｉｒｅｌｅｓｓ」も１．２５のカウントを有するであろう。

【００５３】

しかしながら、クラスがＮグラム中の他の場所、すなわち、右端の位置以外に現れた場合、マルチワード意味項目拡張（ｍｕｌｔｉ－ｗｏｒｄ　ｓｅｍａｎｔｉｃ　ｉｔｅｍ　ｅｘｐａｎｓｉｏｎ）について二重カウンティングを避けるのに以下のステップを行うことができる。先の例と同様に、拡張に関するステップ（ａ）と計算に関するステップ（ｂ）が同じ方式で行われる。しかし、ステップ（ｃ）は行われず、むしろ、Ｎグラムのコンテキストは、拡張後に所望の数の右端の単語だけを取ることにより短縮される。

【００５４】

例として、カウント５を有する「ＣＯＭＰＡＮＹ　ａｎａｌｙｓｔ　ｓａｉｄ」のクラストライグラムを、

Ｐｒ（ｍｉｃｒｏｓｏｆｔ｜ＣＯＭＰＡＮＹ）＝０．５

Ｐｒ（ｏｒａｃｌｅ｜ＣＯＭＰＡＮＹ）＝０．２５

Ｐｒ（ｖｅｒｉｚｏｎ　ｗｉｒｅｌｅｓｓ｜ＣＯＭＰＡＮＹ）＝０．２５

の意味クラス「ＣＯＭＰＡＮＹ」の同じ確率で想定すると、その単語レベルのＮグラムデータは、

「ｍｉｃｒｏｓｏｆｔ　ａｎａｌｙｓｔ　ｓａｉｄ」＝２．５

「ｏｒａｃｌｅ　ａｎａｌｙｓｔ　ｓａｉｄ」＝１．２５

「ｗｉｒｅｌｅｓｓ　ａｎａｌｙｓｔ　ｓａｉｄ」＝１．２５

になり、ここで、「ｗｉｒｅｌｅｓｓ　ａｎａｌｙｓｔ　ｓａｉｄ」は、トライグラムに右端の３ワードだけを取ることにより実現されたものである。

【００５５】

意味データベース３０１がタガー３０６およびワードＮグラムジェネレータ３２０と共に動作可能である場合が示されているが、データベース３０１のインスタンスのそれぞれでの内容は、多くの用途で異なることがあり、それによりこの方法がより役立つことを理解されたい。

【００５６】

ステップ２１０で、言語モデル３２６が、バックグラウンドデータおよびオプションの適応データの生成された単語Ｎグラムカウントを用いてトレーニングされ、ここでは、トレーニングモジュール３２８によって行われる。必要ならば、単語Ｎグラムに関してカウントカットオフプルーニングを行って言語モデルのサイズをさらに低減することもできる。

【００５７】

トレーニングは、Ｎグラム相対頻度推定（ｎ－ｇｒａｍ　ｒｅｌａｔｉｖｅ　ｆｒｅｑｕｅｎｃｙ　ｅｓｔｉｍａｔｅｓ）を平滑化することを含むことができる。例えば、参照により本明細書に組み込まれる非特許文献１に記載されている削除補間法（ｄｅｌｅｔｅｄ－ｉｎｔｅｒｐｏｌａｔｉｏｎ　ｍｅｔｈｏｄ）を、Ｎグラム相対頻度推定を平滑化するために使用することができる。簡潔には、この再帰的削除補間式（ｒｅｃｕｒｓｉｖｅ　ｄｅｌｅｔｅｄ－ｉｎｔｅｒｐｏｌａｔｉｏｎ　ｆｏｒｍｕｌａ）は以下のように定義される。

【００５８】

【数１】

$$\Pr_I\left(w\middle|w_1^{n-1}\right)=\left(1-\lambda_{w_1^{n-1}}\right)\cdot f\left(w\middle|w_1^{n-1}\right)+\lambda_{w_1^{n-1}}\cdot\Pr_I\left(w\middle|w_2^{n-1}\right)$$

$$\Pr_I(w)=(1-\lambda)\cdot f(w)+\lambda\cdot\frac{1}{V}$$

【００５９】

ここで、

【００６０】

【数２】
$f\left(w\middle|w_{1_k}^{n-1}\right)$

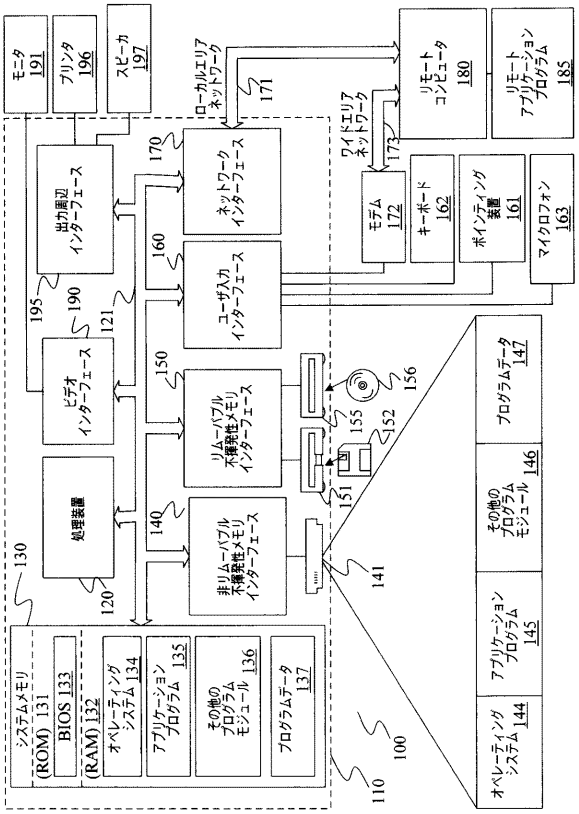
【００６１】
は単語Ｎグラムの相対頻度を表し、
【００６２】
【数３】
$w_{1_k}^{n-1}$

【００６３】
は前のｎ－１語に及ぶ単語履歴である。異なるコンテキスト順で均一な単語分布１／ＶのＮグラムモデルが線形に補間される。補間重み
【００６４】
【数４】
$\lambda_{w_1^{n-1}}$

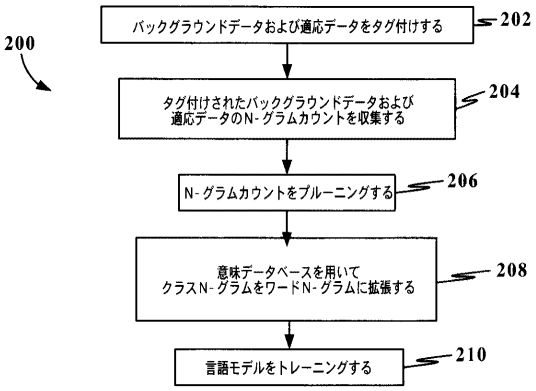
【００６５】
は、周知の最尤法を用いて推定することができる。データのまばらさのために、補間重みは、普通、単語コンテキストをクラスにグループ化することにより推定パラメータ数を低減するために結び付けられる。１つの可能な方法は、ある所与の単語コンテキストの出現数に基づいてパラメータをバケット化することである。
【００６６】
ステップ２１０により、管理された言語モデル適応を完了し、この例では、削除補間された言語モデルが提供される。言語処理システムにおける削除補間された言語モデルの実装は、標準ＡＲＰＡ形式のバックオフ言語モデルへの変換を含んでもよい。２００４年３月２６日に出願された「ＲＥＰＲＥＳＥＮＴＡＴＩＯＮ　ＯＦ　Ａ　ＤＥＬＥＴＥＤ　ＩＮＴＥＲＰＯＬＡＴＩＯＮ　Ｎ－ＧＲＡＭ　ＬＡＮＧＵＡＧＥ　ＭＯＤＥＬ　ＩＮ　ＡＲＰＡ　ＳＴＡＮＤＡＲＤ　ＦＯＲＭＡＴ」というタイトルの同時継続中の米国特許出願に、ＡＲＰＡ形式への変換の一例について記載されており、それをワンパスシステムに用いることができる。
【００６７】
本発明を特定の実施形態を参照して説明してきたが、本発明の精神および範囲を逸脱することなく、形態および詳細に変更を加えることができることを当業者は理解するであろう。
【図面の簡単な説明】
【００６８】
【図１】本発明を実施することのできる一般的なコンピューティング環境を示すブロック図である。
【図２】言語モデルを適応させるための流れ図である。
【図３Ａ】言語モデルを適応させるためのシステムを示すブロック図である。
【図３Ｂ】言語モデルを適応させるためのシステムを示すブロック図である。
【符号の説明】
【００６９】
１００　コンピューティング環境
１１０　コンピュータ
１２０　処理装置
１３０　システムメモリ
１３１　ＲＯＭ
１３２　ＲＡＭ
１３３　ＢＩＯＳ

１３４　オペレーティングシステム
１３５　アプリケーションプログラム
１３６　その他のプログラムモジュール
１３７　プログラムデータ
１４０　非リムーバブル不揮発性メモリインターフェース
１４１　ハードディスクドライブ
１４４　オペレーティングシステム
１４５　アプリケーションプログラム
１４６　その他のプログラムモジュール
１４７　プログラムデータ
１５０　リムーバブル不揮発性メモリインターフェース
１５１　磁気ディスクドライブ
１５２　リムーバブルの不揮発性磁気ディスク
１５５　光ディスクドライブ
１５６　リムーバブルの不揮発性光ディスク
１６０　ユーザ入力インターフェース
１６１　ポインティング装置
１６２　キーボード
１６３　マイクロフォン
１７０　ネットワークインターフェース
１７１　ローカルエリアネットワーク
１７２　モデム
１７３　ワイドエリアネットワーク
１８０　リモートコンピュータ
１８５　リモートアプリケーションプログラム
１９０　ビデオインターフェース
１９１　モニタ
１９５　出力周辺インターフェース
１９６　プリンタ
１９７　スピーカ
３００　例示的システム
３０１　意味データベース
３０２　バックグラウンドデータ
３０４　適応データ
３０６　タガー
３０８　タグ付けしたバックグラウンド
３１０　タグ付けした適応
３１２　収集モジュール
３１４　プルーニングモジュール
３１６　バックグラウンドＮ－グラムクラスカウントデータ
３１８　適応Ｎ－グラムクラスカウントデータ
３２０　ワードＮ－グラムジェネレータ
３２２　バックグラウンドＮ－グラムワードカウントデータ
３２４　適応Ｎ－グラムワードカウントデータ
３２６　適応された言語モデル
３２８　トレーニングモジュール

【図１】


モニタ 191
プリンタ 196
スピーカ 197
ローカルエリアネットワーク 171
リモートコンピュータ 180
リモートアプリケーションプログラム 185
ワイドエリアネットワーク 173
出力周辺インターフェース 195
ネットワークインターフェース 170
モデム 172
キーボード 162
ポインティング装置 161
マイクロフォン 163
ユーザ入力インターフェース 160
ビデオインターフェース 190
121
リムーバブル不揮発性メモリインターフェース 150
処理装置 120
非リムーバブル不揮発性メモリインターフェース 140
130
システムメモリ (ROM) 131
BIOS 133
(RAM) 132
オペレーティングシステム 134
アプリケーションプログラム 135
その他のプログラムモジュール 136
プログラムデータ 137
110
100
オペレーティングシステム 144
アプリケーションプログラム 145
その他のプログラムモジュール 146
プログラムデータ 147
141
151
152
155
156

【図２】


200
バックグラウンドデータおよび適応データをタグ付けする 202
タグ付けされたバックグラウンドデータおよび適応データのN-グラムカウントを収集する 204
N-グラムカウントをプルーニングする 206
意味データベースを用いてクラスN-グラムをワードN-グラムに拡張する 208
言語モデルをトレーニングする 210

【図３Ａ】

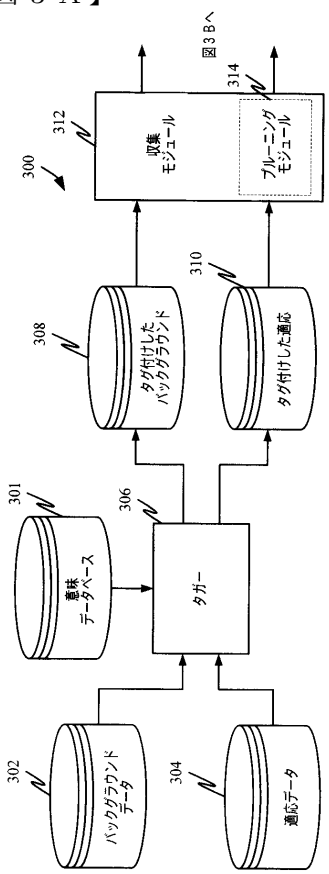
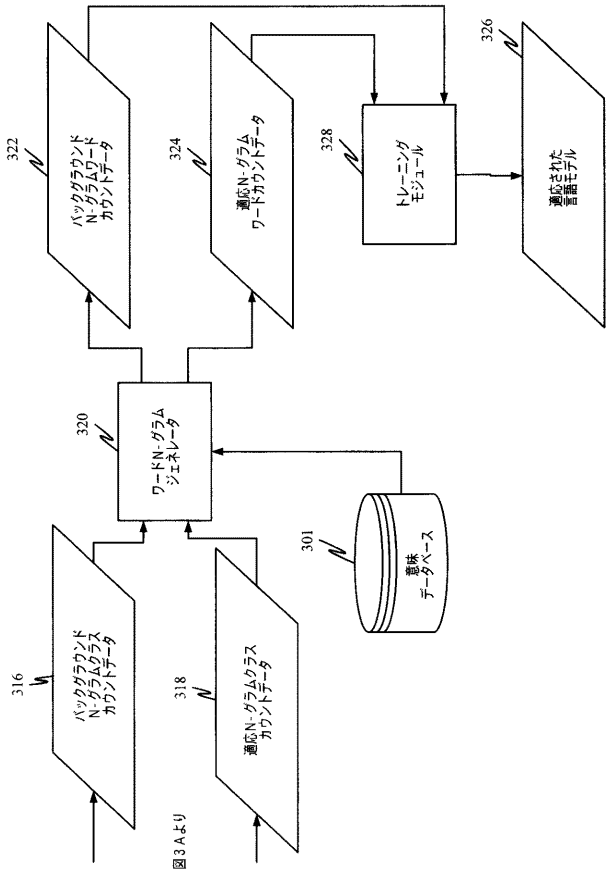


図３Ｂへ
300
収集モジュール 312
プルーニングモジュール 314
タグ付けしたバックグラウンド 308
タグ付けした適応 310
意味データベース 301
タガー 306
バックグラウンドデータ 302
適応データ 304

【図３Ｂ】


バックグラウンドN-グラムカウントデータ 322
適応N-グラムワードカウントデータ 324
トレーニングモジュール 328
適応された言語モデル 326
ワードN-グラムジェネレータ 320
意味データベース 301
バックグラウンドN-グラムクラスカウントデータ 316
適応N-グラムクラスカウントデータ 318
図３Ａより

フロントページの続き

(72)発明者　シプリアン　アイ．ケルバ
アメリカ合衆国　９８０５２　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション内
(72)発明者　ミライインド　マハジャン
アメリカ合衆国　９８０５２　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション内
(72)発明者　イーチュン　タム
アメリカ合衆国　９８０５２　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション内

Ｆターム(参考) 5D015 HH23