(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5479581号
(P5479581)

(45)発行日　平成26年4月23日(2014.4.23)　(24)登録日　平成26年2月21日(2014.2.21)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 220Z
G06F 17/28 (2006.01)　G06F 17/30 380E
G06F 17/28 Z

請求項の数 15　(全 21 頁)

(21)出願番号　特願2012-511920 (P2012-511920)
(86)(22)出願日　平成22年5月14日(2010.5.14)
(65)公表番号　特表2012-527701 (P2012-527701A)
(43)公表日　平成24年11月8日(2012.11.8)
(86)国際出願番号　PCT/US2010/035033
(87)国際公開番号　WO2010/135204
(87)国際公開日　平成22年11月25日(2010.11.25)
審査請求日　平成25年5月7日(2013.5.7)
(31)優先権主張番号　12/470,492
(32)優先日　平成21年5月22日(2009.5.22)
(33)優先権主張国　米国(US)

(73)特許権者　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　100107766
弁理士　伊東　忠重
(74)代理人　100070150
弁理士　伊東　忠彦
(74)代理人　100091214
弁理士　大貫　進介

最終頁に続く

(54)【発明の名称】構造化されていないリソースからの句対のマイニング

(57)【特許請求の範囲】
【請求項１】
複数のクエリを構築するステップであって、前記複数のクエリは、当該複数のクエリについての１つの言語で表現されるクエリ用語を有する、構築するステップと、
前記複数のクエリを取出しモジュールに提示するステップであって、前記取出しモジュールは、探索動作を実行して前記複数のクエリのクエリ用語に合致する複数の文書を識別するように構成されている、提示するステップと、
前記取出しモジュールから結果セットを受信するステップであって、前記結果セットは、前記複数のクエリのクエリ用語に合致する文書から前記取出しモジュールによって識別された結果項目を提供し、該結果項目もまた前記クエリについての１つの言語で表現されている、受信するステップと、
構造化されたトレーニングセットを作成するために、前記結果セットを処理するステップであって、前記トレーニングセットは、前記結果セット内の前記結果項目の複数組の対を示し、前記トレーニングセットの個々の対は、
前記取り出しモジュールから受信した第1の結果項目であって、前記クエリについての前記言語で表現された第1の複数の単語を有する第1の結果項目と、
前記取り出しモジュールから受信した第２の結果項目であって、同様に前記クエリについての前記言語で表現された第２の複数の単語を有する第２の結果項目と
を含む、処理するステップとを備え、
前記トレーニングセットは、それによって電気トレーニングシステムが前記統計的翻訳

モデルを学習できる基礎を提供することを特徴とするコンピュータ実装方法。
【請求項２】
前記取出しモジュールは、前記複数のクエリを処理して複数の結果項目を識別する探索エンジンであり、個々の前記結果項目は個々の関連の文書の概要または要約を含むことを特徴とする請求項１に記載の方法。
【請求項３】
前記結果セットは広域ネットワークを通じて受信し、該広域ネットワークは、該広域ネットワークを通じて前記文書を取り出すためのリソースロケータを有することを特徴とする請求項２に記載の方法。
【請求項４】
前記処理するステップは、少なくとも１つの要件に基づいて、前記結果セット内の前記複数の結果項目を制約するステップを含むことを特徴とする請求項１に記載の方法。
【請求項５】
前記制約するステップは、前記複数の結果項目に関連するランキングスコアに基づいて、ペアワイズのマッチングに関する候補を前記複数の結果項目から識別するステップを含むことを特徴とする請求項４に記載の方法。
【請求項６】
前記制約するステップは、複数の候補と前記結果セットに関連するそれぞれの語彙的な署名との間の合意に基づいて、ペアワイズのマッチングに関する候補を前記複数の結果項目から識別するステップを含むことを特徴とする請求項４に記載の方法。
【請求項７】
前記制約するステップは、他の個々の結果項目に対する個々の結果項目の類似性を表す類似性スコアに基づいて、ペアワイズのマッチングに関する候補を識別するステップを含むことを特徴とする請求項４に記載の方法。
【請求項８】
前記制約するステップは、前記複数の結果項目と前記複数の結果項目の識別されたクラスタとの間の関連性に基づいて、ペアワイズでマッチングに関する候補を識別するステップを含むことを特徴とする請求項４に記載の方法。
【請求項９】
前記トレーニングセットを使用して前記統計的翻訳モデルをトレーニングし、前記クエリの言語の第２の句を得るために前記クエリの言語の第１の句を書き換えるステップをさらに含むことを特徴とする請求項１に記載の方法。
【請求項１０】
前記トレーニングセット内の複数の句の試行的なペアワイズのアラインメントを得るために、終了状態に満足するまで前記統計的翻訳モデルを相互作用的に適用するステップをさらに含むことを特徴とする請求項９に記載の方法。
【請求項１１】
前記終了状態の満足に応答して前記統計的翻訳モデルを決定する統計パラメータを生成するステップをさらに含むことを特徴とする請求項１０に記載の方法。
【請求項１２】
前記複数のクエリは共通の主題に関連し、前記複数のクエリの各々が前記共通の主題を目的とすることを特徴とする請求項１に記載の方法。
【請求項１３】
前記トレーニングセットに基づいて前記統計的翻訳モデルを生成するステップと、前記統計的翻訳モデルを適用するステップとをさらに備え、前記適用するステップは、
前記統計的翻訳モデルを使用して、探索クエリを拡張するステップ、
前記統計的翻訳モデルを使用して、文書索引付け決定を円滑にするステップ、
前記統計的翻訳モデルを使用して、テキストコンテンツを改正するステップ、または
前記統計的翻訳モデルを使用して、広告情報を拡張するステップのうちの１つ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１４】
請求項１乃至１３のいずれかの項に記載の方法をコンピュータに実行させるためのコンピュータプログラム。
【請求項１５】
請求項１４に記載のコンピュータプログラムを格納するコンピュータ読み取り可能記録媒体を有するコンピュータ。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、構造化されていないリソースからの句対のマイニングに関する。
【背景技術】
【０００２】
近年、統計的機械翻訳技術に相当な関心が寄せられている。この技術は、まずトレーニングセットを確立することによって動作する。伝統的には、このトレーニングセットは、第１の言語のテキスト本体および対応する第２の言語のテキスト本体など、テキストの並列コーパスを提供する。トレーニングモジュールは、テキストの第１の本体がテキストの第２の本体にマップする可能性が高い様式を決定するために統計技術を使用する。この解析は、結果として、翻訳モデルを生成する。復号化段階において、第１の言語のテキストのインスタンスを対応する第２の言語のインスタンスにマップするために翻訳モデルを使用することが可能である。
【０００３】
統計的翻訳モデルの有効性は、多くの場合、翻訳モデルを作成するために使用されるトレーニングセットの頑強さに依存する。しかし、高い品質のトレーニングセットを提供することは困難な課題である。１つには、これは、トレーニングモジュールは一般に、大量のトレーニングデータを必要とするが、かかる情報を供給するための事前に確立された並列コーパスタイプのリソースは不足しているためである。伝統的な事例では、トレーニングセットは、例えば、人間の翻訳者を使用することによって、並列テキストを手動で生成することによって取得可能である。しかし、これらのテキストの手動生成は、非常に時間のかかる任務である。
【０００４】
より自動化された形で並列テキストを識別するためのいくつかの技術が存在する。例えば、ウェブサイトが、情報のそれぞれのバージョンが別個のネットワークアドレス（例えば、別個のＵＲＬ）に関連している同じ情報を複数の異なる言語で伝える事例を検討する。一技術では、取出しモジュールは、例えば、ＵＲＬ内の特徴情報に基づいて、これらの並列文書の識別を試みる際に探索索引を調査することが可能である。しかし、この技術は、比較的に制限された数の並列テキストにアクセス可能である。さらに、この手法は推定に依存する場合があり、これは多くの事例に当てはまらない可能性がある。
【０００５】
上の例は、２つの異なる自然言語間でテキストを変換するモデルとの関連で構成される。単一言語モデルも提案されている。かかるモデルは、入力テキストを書き換えて、入力テキストと同じ言語で出力テキストを作成することを試みる。一応用例では、例えば、このタイプのモデルは、例えば、探索クエリを表現するための追加の様式を識別することによって、ユーザの探索クエリを修正するために使用可能である。
【０００６】
単一言語モデルは、上記と同じ欠点を受ける。実際に、同じ言語内に事前に存在する並列コーパスを見出すことは特に困難な場合がある。すなわち、二言語コンテキストで、異なる言語で並列テキストを生成して、異なる読み手の母語に対処する必要があらかじめ存在する。同じ言語でテキストの並列バージョンを生成する、よりいっそう制限された必要が存在する。
【０００７】

それでもなお、かかる単一言語情報は少数存在する。例えば、従来のシソーラスは、類似の意味を有する同じ言語の語に関する情報を提供する。もう１つの例では、一部の書籍は異なる翻訳者によって同じ言語に翻訳されている。これらの異なる翻訳は、並列単一言語コーパスとして役立つ場合がある。しかし、このタイプの並列情報は、より一般的な状況で効果的に使用されるにはあまりにも専門化され過ぎている可能性がある。さらに、述べたように、このタイプの情報は比較的少数だけ存在する。

【発明の概要】

【発明が解決しようとする課題】

【０００８】

同じ主題に関する単一言語文書の本体を自動的に識別し、次いで、並列文の存在に関して、それらの文書をマイニングすることも試みられている。しかし、場合によっては、これらの手法は、その有効性および一般性を制限する可能性がある、コンテキスト特定の推定に依存している。これらの困難に加えて、テキストは非常に多くの様式で書き換えられることが可能であり、したがって、単一言語コンテキストで並列性を識別することは、二言語コンテキストで関係するテキストを識別するよりも潜在的により複雑な任務である。

【課題を解決するための手段】

【０００９】

構造化されていないリソースから構造化されたトレーニングセットを選び取るマイニングシステムが本明細書で説明される。すなわち、構造化されていないリソースは、反復コンテンツ内および交番タイプのコンテンツ内で潜在的に豊富な可能性がある。反復コンテンツは、構造化されていないリソースがテキストの同じインスタンスの多くの反復を含むことを意味する。交番タイプのコンテンツは、構造化されていないリソースが、形態は異なるが、類似の意味内容を表現するテキストの多くのインスタンスを含むことを意味する。このマイニングシステムは、構造化されていないリソースのこれらの特性を露出および抽出し、そのプロセスを通じて、翻訳モデルをトレーニングする際に使用するために、未加工の構造化されていないコンテンツを構造化されたコンテンツに変換する。一事例では、この構造化されていないリソースは、ネットワークアクセス可能なリソース項目（例えば、インターネットアクセス可能なリソース項目）のリポジトリに対応する。

【００１０】

１つの例示的な実施形態によれば、マイニングシステムは、クエリを取出しモジュールに提出することによって動作する。この取出しモジュールは、それらのクエリを使用して、構造化されていないリソース内で探索を実行し、その時点で、この取出しモジュールは結果項目を提供する。これらの結果項目は、構造化されていないリソース内で提供された関連するリソース項目を要約するテキスト区分に対応し得る。このマイニングシステムは、それらの結果項目をフィルタリングして、結果項目のそれぞれの対を識別することによって、構造化されたトレーニングセットを作成する。トレーニングシステムは、トレーニングセットを使用して、統計的翻訳モデルを作成することが可能である。

【００１１】

１つの例示的な態様によれば、このマイニングシステムは、同じ主題に対処するリソース項目のグループを事前に識別せずに、クエリの提出だけに基づいて、結果項目を識別することが可能である。すなわち、このマイニングシステムは、概して、リソース項目（例えば、文書）の主題に関するアグノスティック（ａｇｎｏｓｔｉｃ）手法をとることが可能であり、このマイニングシステムは、構造化されていないリソース内の構造をサブドキュメント（ｓｕｂ－ｄｏｃｕｍｅｎｔ）断片レベルで露出する。

【００１２】

もう１つの例示的な態様によれば、このトレーニングセットは、文の断片に対応する項目を含むことが可能である。すなわち、（このトレーニングシステムは完全文を含むトレーニングセットを成功裏に処理することも可能であるが）このトレーニングシステムは、文レベルの並列性の識別および利用に依存しない。

【００１３】

もう１つの例示的な態様によれば、この翻訳モデルは、単一言語内で入力句を出力句に変換するために、単一言語コンテキストで使用することが可能であり、この場合、入力句および出力句は、類似の意味内容を有するが、異なる形態の表現を有する。すなわち、入力句のパラフレーズされた（ｐａｒａｐｈｒａｓｅｄ）バージョンを提供するためにこの翻訳モデルを使用することが可能である。第１の言語の入力句を第２の言語の出力句に翻訳するために、この翻訳モデルを二言語コンテキストで使用することも可能である。

【００１４】

もう１つの例示的な態様によれば、翻訳モデルの様々な応用例が説明される。

【００１５】

上記の手法は、様々なタイプのシステム、構成要素、方法、コンピュータ可読媒体、データ構造、製品などの形で表すことが可能である。

【００１６】

この課題を解決するための手段は、精選された概念を簡素化された形態で紹介するために提供され、これらの概念は下で発明を実施するための形態においてさらに説明される。この課題を解決するための手段は、特許請求される主題の主な特徴または必須の特徴を識別することが意図されず、特許請求される主題の範囲を限定するために使用されることも意図されない。

【図面の簡単な説明】

【００１７】

【図１】統計的機械翻訳モデルを作成および適用するための１つの例示的なシステムを示す図である。

【図２】ネットワーク関連環境内の図１のシステムの一実施形態を示す図である。

【図３】１つの結果セット内の一連の結果項目の一例を示す図である。図１のシステムは、クエリを取出しモジュールに提出することに応答して、その結果セットを戻す。

【図４】図１のシステムが結果セット内の結果項目の対をどのように確立できるかを明示する一例を示す。

【図５】図１のシステムが、異なる結果セットに関して実行された解析に基づいて、トレーニングセットをどのように作成できるかを明示する一例を示す図である。

【図６】図１のシステムの動作の概要を提示する１つの例示的な手順を示す図である。

【図７】図６の手順内でトレーニングセットを確立するための１つの例示的な手順を示す図である。

【図８】図１のシステムを使用して作成された翻訳モデルを適用するための１つの例示的な手順を示す図である。

【図９】前述の図面に示される特徴の任意の態様を実施するために使用可能な例示的な処理機能性を示す図である。

【発明を実施するための形態】

【００１８】

類似の構成要素および特徴を参照するために、本開示および図面の全体にわたって同じ番号が使用される。１００の連番は、図１において当初見出される特徴を指し、２００の連番は、図２において当初見出される特徴を指し、３００の連番は、図３において当初見出される特徴を指す、等々である。

【００１９】

本開示は、統計的翻訳モデルを確立するために使用可能なトレーニングセットを生成するための機能性を記載する。本開示は、統計的翻訳モデルを生成および適用するための機能性も記載する。

【００２０】

本開示は以下のように組織される。セクションＡは、上で要約された機能を実行するための１つの例示的なシステムを説明する。セクションＢは、セクションＡのシステムの動作を説明する例示的な方法を説明する。セクションＣは、セクションＡおよびＢで説明される特徴の任意の態様を実施するために使用可能な例示的な処理機能性を説明する。

【００２１】
予備事項として、図面うちのいくつかは、機能性、モジュール、特徴、要素など、様々に称される、１つまたは複数の構造的な構成要素との関連で概念を説明する。図面に示される様々な構成要素は、例えば、ソフトウェア、ハードウェア（例えば、ディスクリート論理構成要素など）、ファームウェアなど、またはこれらの実施形態の任意の組合せによって、いかなるようにも実施可能である。１つの事例では、図面の様々な構成要素を別個のユニットに例示的に分離することは、実際の実施形態において、対応する別個の構成要素を使用することを反映する場合がある。代わりに、または加えて、図に例示される任意の単一の構成要素は、複数の実際の構成要素によって実施可能である。代わりに、または加えて、図面の任意の２つ以上の別個の構成要素の描写は、単一の実際の構成要素によって実行される異なる機能を反映する場合もある。次に説明される図９は、図面に示される機能の１つの例示的な実施形態に関して追加の詳細を提供する。
【００２２】
その他の特徴は、流れ図の形態でこれらの概念を説明する。この形態では、ある順序で実行される別個のブロックを構成するとして、いくつかの動作が説明される。かかる実施形態は、例示的であり、限定的ではない。本明細書で説明されるいくつかのブロックは、一緒にグループ化されて、単一の動作の形で実行されることが可能であり、いくつかのブロックは、分裂されて複数の構成要素ブロックにされることが可能であり、いくつかのブロックは、（ブロックを実行する並列様式を含めて）本明細書で例示される順序とは異なる順序で実行されることも可能である。流れ図に示されるブロックは、ソフトウェア、ハードウェア（例えば、ディスクリート論理構成要素など）、ファームウェア、手動処理など、またはこれらの実施形態の任意の組合せによって実施可能である。
【００２３】
専門用語に関して、「するように構成された」という句は、識別される動作を実行するために任意の種類の機能性を構築することが可能な任意の様式を包括する。この機能性は、例えば、ソフトウェア、ハードウェア（例えば、ディスクリート論理構成要素など）、ファームウェアなど、および／またはそれらの任意の組合せを使用して、動作を実行するように構成可能である。
【００２４】
「論理」という用語は、任務を実行するための任意の機能性を包括する。例えば、流れ図に例示されるそれぞれの動作は、その動作を実行するための論理に対応する。動作は、例えば、ソフトウェア、ハードウェア（例えば、ディスクリート論理構成要素など）、ファームウェアなど、および／またはそれらの任意の組合せを使用して実行可能である。
【００２５】
Ａ．例示的なシステム
図１は、翻訳モデル１０２を生成および適用するための１つの例示的なシステム１００を示す。翻訳モデル１０２は、入力句を出力句にマップするための統計的機械翻訳（ＳＭＴ）モデルに対応し、この場合、「句」はここでは任意の１つまたは複数のテキストストリングを指す。翻訳モデル１０２は、規則ベースの手法ではなく、統計技術を使用してこの動作を実行する。しかし、もう１つの実施形態では、翻訳モデル１０２は、規則ベースの手法の１つまたは複数の特徴を組み込むことによって、その統計解析を補完することが可能である。
【００２６】
一事例では、翻訳モデル１０２は、単一言語コンテキストで動作する。この場合、翻訳モデル１０２は、入力句と同じ言語で表現された出力句を生成する。すなわち、出力句は、入力句のパラフレーズされたバージョンと見なすことができる。もう１つの事例では、翻訳モデル１０２は、二言語（または、多言語）コンテキストで動作する。この場合、翻訳モデル１０２は、入力句と比べて異なる言語で出力句を生成する。さらに別の事例では、翻訳モデル１０２は、翻字コンテキストで動作する。この場合、この翻訳モデルは、入力句と同じ言語で出力句を生成するが、出力句は入力句と比べて異なる書式で表現される

。翻訳モデル１０２は、さらに他の翻訳シナリオにも適用可能である。すべてのかかるコンテキストで、「翻訳」という用語は、テキスト情報の１つの状態から別の状態への任意のタイプの変換を指し、広く解釈されるべきである。

【００２７】

システム１００は、３つの主な構成要素、すなわち、マイニングシステム１０４と、トレーニングシステム１０６と、アプリケーションモジュール１０８とを含む。概要として、マイニングシステム１０４は、翻訳モデル１０２をトレーニングする際に使用するためのトレーニングセットを作成する。トレーニングシステム１０６は、反復手法を適用して、そのトレーニングセットに基づいて翻訳モデル１０２を導出する。アプリケーションモジュール１０８は、翻訳モデル１０２を適用して、特定の使用関連のシナリオにおいて、入力句を出力句にマップする。

【００２８】

一事例では、単一のシステムは、単一のエンティティまたは複数のエンティティの任意の組合せによって管理されるように、図１に示される構成要素のすべてを実施することが可能である。もう１つの事例では、任意の２つ以上の別個のシステムは、この場合も、単一のエンティティまたは複数のエンティティの任意の組合せによって管理されるように、図１に示される任意の２つ以上の構成要素を実施することが可能である。いずれの事例においても、図１に示される構成要素は、単一のサイトに配置可能であり、または複数のそれぞれのサイトに分散されてもよい。以下の説明は、図１に示される構成要素に関して追加の詳細を提供する。

【００２９】

マイニングシステム１０４から始めると、この構成要素は、構造化されていないリソース１１０から結果項目を取り出すことによって動作する。構造化されていないリソース１１０は、リソース項目の任意の局在化されたソースまたは分散されたソースを表す。これらのリソース項目は、今度は、テキスト情報の任意のユニットに対応し得る。例えば、構造化されていないリソース１１０は、インターネットなど、広域ネットワークによって提供されたリソース項目の分散されたリポジトリを表すことができる。この場合、これらのリソース項目は、ネットワークアクセス可能なページおよび／または任意のタイプの関連文書に対応し得る。

【００３０】

構造化されていないリソース１１０は並列コーパスの様式のように先験的に構成されないため、構造化されていないと見なされる。すなわち、構造化されていないリソース１１０は、任意の包括的なスキームに従って、そのリソース項目を互いに関連付けない。それでもなお、構造化されていないリソース１１０は、反復コンテンツ内および交番タイプのコンテンツ内で潜在的に豊富な可能性がある。反復コンテンツは、構造化されていないリソース１１０がテキストの同じインスタンスの多くの反復を含むことを意味する。交番タイプのコンテンツは、構造化されていないリソース１１０が、形式の点で異なるが、類似の意味内容を表現するテキストの多くのインスタンスを含むことを意味する。これは、トレーニングセットを構築する際に使用するためにマイニング可能な構造化されていないリソース１１０の基礎となる特徴が存在することを意味する。

【００３１】

マイニングシステム１０４の１つの目的は、構造化されていないリソース１１０の上述の特性を露出し、そのプロセスを介して、翻訳モデル１０２をトレーニングする際に使用するために、未加工の構造化されていないコンテンツを構造化されたコンテンツに変換することである。マイニングシステム１０４は、１つには、取出しモジュール１１６と共に、クエリ準備モジュール１１２とインターフェースモジュール１１４とを使用して、この目的を達成する。クエリ準備モジュール１１２は、クエリのグループを公式化する。それぞれのクエリは、ターゲット主題に関する１つまたは複数のクエリ用語を含むことが可能である。インターフェースモジュール１１４は、それらのクエリを取出しモジュール１１６に提出する。取出しモジュール１１６は、クエリを使用して、構造化されていないリソ

ース１１０内で探索を実行する。この探索に応答して、取出しモジュール１１６は、異なるそれぞれのクエリに関して複数の結果セットを戻す。それぞれの結果セットは、今度は、１つまたは複数の結果項目を含む。これらの結果項目は、構造化されていないリソース１１０内のそれぞれのリソース項目を識別する。

【００３２】

一事例では、マイニングシステム１０４および取出しモジュール１１６は、同じエンティティまたは異なるそれぞれのエンティティによって管理された同じシステムによって実施される。もう１つの事例では、マイニングシステム１０４および取出しモジュール１１６は、この場合も、同じエンティティまたは異なるそれぞれのエンティティによって管理された、２つのそれぞれのシステムによって実施される。例えば、一実施形態では、取出しモジュール１１６は、限定されないが、ワシントン州、レドモンドのＭｉｃｒｏｓｏｆｔ　Ｃｏｒｐｏｒａｔｉｏｎによって提供されるＬｉｖｅ　Ｓｅａｒｃｈエンジンなどの探索エンジンを表す。ユーザは、その探索エンジンによって提供されたインターフェース（例えば、ＡＰＩなど）など、任意の機構を介して探索エンジンにアクセスすることができる。この探索エンジンは、任意の探索方策およびランキング方策を使用して、提出されたクエリに応答して、結果セットを識別および公式化することが可能である。

【００３３】

一事例では、結果セット内の結果項目は、それぞれのテキスト区分に対応する。異なる探索エンジンは、クエリの提出に応答して、テキスト区分を公式化する際に異なる方策を使用することが可能である。多くの例において、これらのテキスト区分は、提出されたクエリに関するリソース項目の関連性を伝えるリソース項目の代表的な部分（例えば、抜粋）を提供する。説明のために、これらのテキスト区分は、その関連する完全なリソース項目の短い摘要または要約と見なすことができる。より詳細には、一事例では、これらのテキスト区分は、基礎となる完全なリソース項目から取り出された１つまたは複数の文に対応し得る。１つのシナリオでは、インターフェースモジュール１１４および取出しモジュール１１６は、文の断片を含むリソース項目を公式化することが可能である。もう１つのシナリオでは、インターフェースモジュール１１４および取出しモジュール１１６は、完全な文（または、完全な段落など、テキストのより大きな単位）を含むリソース項目を公式化することが可能である。インターフェースモジュール１１４は、ストア１１８内にそれらの結果セットを格納する。

【００３４】

トレーニングセット準備モジュール１２０（簡潔にするために「準備モジュール」）は、トレーニングセットを作成するために、それらの結果セット内の未加工データを処理する。この動作は、２つの構成要素動作、すなわち、別個にまたは一緒に実行可能なフィルタリングとマッチングとを含む。フィルタリング動作に関して、準備モジュール１２０は、１つまたは複数の制約要件に基づいて、結果項目の元のセットをフィルタリングする。この処理の目的は、ペアワイズの（ｐａｉｒｗｉｓｅ）マッチングに適した候補である結果項目のサブセットを識別し、それによって、それらの結果セットから「ノイズ」を除去することである。このフィルタリング動作は、フィルタリングされた結果セットを作成する。マッチング動作に関して、準備モジュール１２０は、フィルタリングされた結果セットに関してペアワイズのマッチングを実行する。このペアワイズのマッチングは、結果セット内の結果項目の対を識別する。準備モジュール１２０は、上で説明された動作によって作成されたトレーニングセットをストア１２２内に格納する。準備モジュール１２０の動作に関する追加の詳細は、この説明の後の時点で提供される。

【００３５】

トレーニングシステム１０６は、翻訳モデル１０２をトレーニングするために、ストア１２２内のトレーニングセットを使用する。このために、トレーニングシステム１０６は、句タイプのＳＭＴ機能性など、任意のタイプの統計的機械翻訳（ＳＭＴ）機能性１２４を含むことが可能である。ＳＭＴ機能性１２４は、トレーニングセット内のパターンを識別するための統計技術を使用することによって動作する。ＳＭＴ機能性１２４は、これら

のパターンを使用して、トレーニングセット内の句の相関関係を識別する。

【００３６】

より詳細には、ＳＭＴ機能性１２４は、反復様式でそのトレーニング動作を実行する。それぞれの段階で、ＳＭＴ機能性１２４は、ＳＭＴ機能性１２４がトレーニングセット内の句のペアワイズのアラインメントに関する一時的な仮定に達することを可能にする統計解析を実行する。ＳＭＴ機能性１２４は、これらの一時的な仮定を使用して、その統計解析を繰り返し、ＳＭＴ機能性１２４が更新された一時的な仮定に達することを可能にする。ＳＭＴ機能性１２４は、終了条件が満たされたと見なされるまで、この反復動作を繰り返す。ストア１２６は、ＳＭＴ機能性１２４によって実行された処理の間に（例えば、翻訳表などの形態で）暫定的なアラインメント情報の作業セットを維持することが可能である。その処理の終了時に、ＳＭＴ機能性１２４は、翻訳モデル１０２を画定する統計パラメータを作成する。ＳＭＴ機能性１２４に関する追加の詳細は、この説明の後の時点で説明される。

【００３７】

アプリケーションモジュール１０８は、翻訳モデル１０２を使用して、入力句を意味的に関係する出力句に変換する。上記のように、入力句および出力句は、同じ言語で表現されてよく、または異なるそれぞれの言語で表現されてもよい。アプリケーションモジュール１０８は、様々なアプリケーションシナリオとの関連でこの変換を実行することが可能である。アプリケーションモジュール１０８およびこれらのアプリケーションシナリオに関する追加の詳細は、この説明の後の時点で提供される。

【００３８】

図２は、図１のシステム１００の１つの代表的な実施形態を示す。この場合、マイニングシステム１０４およびトレーニングシステム１０６を実施するためにコンピューティング機能性２０２を使用することが可能である。コンピューティング機能性２０２は、単一のエンティティもしくは複数のエンティティの組合せによって維持されるように、単一のサイトに維持された、または複数のサイトの全域に分散された任意の処理機能性を表すことが可能である。１つの代表的な事例では、コンピューティング機能性２０２は、パーソナルデスクトップコンピューティングデバイス、サーバタイプのコンピューティングデバイスなど、任意のタイプのコンピュータデバイスに対応する。

【００３９】

一事例では、構造化されていないリソース１１０は、ネットワーク環境２０４によって提供されたリソース項目の分散されたリポジトリによって実施可能である。ネットワーク環境２０４は、任意のタイプのローカルエリアネットワークまたは広域ネットワークに対応し得る。例えば、限定なしに、ネットワーク環境２０４は、インターネットに対応し得る。かかる環境は、例えば、ネットワークアクセス可能なページおよびリンクされたコンテンツ項目に対応する、潜在的に膨大な数のリソース項目に対するアクセスを提供する。取出しモジュール１１６は、従来の様式で、例えば、ネットワーククローリング機能性などを使用して、ネットワーク環境２０４内で利用可能なリソース項目の索引を維持することが可能である。

【００４０】

図３は、クエリ３０４の提出に応答して、取出しモジュール１１６によって戻されることが可能な仮説結果セット３０２の一部の一例を示す。この例は、図１のマイニングシステム１０４の概念的な基礎のうちのいくつかを説明する手段として役立つ。

【００４１】

クエリ３０４「ｓｈｉｎｇｌｅｓ　ｚｏｓｔｅｒ（帯状疱疹）」は、よく知られている疾病に関する。このクエリは、大量の無関係な情報を排除することに十分注目しながら、ターゲット主題を正確に示すために選択されている。この例では、「ｓｈｉｎｇｌｅｓ」は、疾病の一般的な名称を指し、一方、「ｚｏｓｔｅｒ」は、（例えば、ｈｅｒｐｅｓ　ｚｏｓｔｅｒ（帯状ヘルペス）の場合など）疾病のより正式な名称を指す。クエリ用語のこの組合せは、したがって、「ｓｈｉｎｇｌｅｓ（帯状疱疹）」という用語の無関係な意

味および意図されない意味に関する結果項目の取り出しを削減することができる。

【００４２】

結果セット３０２は、Ｒ１～ＲＮとラベル付けされた一連の結果項目を含み、図３は、これらの結果項目の小さな例を示す。それぞれの結果項目は、対応するリソース項目から抽出されたテキスト区分を含む。この事例では、これらのテキスト区分は、文の断片を含む。しかし、インターフェースモジュール１１４および取出しモジュール１１６は、完全文（または、完全な段落など）を含むリソース項目を提供するように構成されることも可能である。

【００４３】

帯状疱疹の疾病は顕著な特性を有する。例えば、帯状疱疹は、水疱瘡を引き起こすのと同じウィルス（帯状ヘルペス）の再活性化によって引き起こされる疾病である。再度活気づくと、このウィルスは身体の神経に沿って移動し、小さな水ぶくれの群れを特徴とする、見た目が赤く、痛みを伴う発疹をもたらす。この疾病は、免疫システムが低下した場合に発生することが多く、したがって、身体外傷、他の疾病、ストレスなどによってトリガされる場合がある。この疾病は、お年寄りを悩ますことが多い、等々である。

【００４４】

異なる結果項目は、この疾病の顕著な特徴に注目するコンテンツを含むことが予想できる。結果として、これらの結果項目は、ある種の示唆に富む句を繰り返すことが予想できる。例えば、インスタンス３０６によって表示されるように、結果項目のうちのいくつかは、様々に表現されるように、痛みを伴う発疹の発生を述べている。インスタンス３０８によって表示されるように、結果項目のうちのいくつかは、この疾病は、様々に表現されるように、弱まった免疫システムに関連することを述べている。インスタンス３１０によって表示されるように、結果項目のうちのいくつかは、この疾病は、結果として、様々に表現されるように、ウィルスが体内の神経に沿って進むことを述べている、等々である。これらの例は、単なる例である。その他の結果項目は、概して、ターゲット主題に無関係な可能性がある。例えば、結果項目３１２は、建材との関連で「Ｓｈｉｎｇｌｅｓ（屋根板）」という用語を使用し、したがって、この主題に関係がない。しかし、この無関係な結果項目３１２すら、他の結果項目と共有される句を含む場合がある。

【００４５】

結果セット３０２内で明らかにされるパターンから、様々な洞察を得ることができる。これらの洞察のうちのいくつかは、ターゲット主題、すなわち、帯状疱疹の疾病に辛うじて関係する。例えば、マイニングシステム１０４は、結果セット３０２を使用して、「ｓｈｉｎｇｌｅｓ」と「ｈｅｒｐｅｓ　ｚｏｓｔｅｒ」が同義語であると推定できる。その他の洞察は、一般に、医療分野に関する。例えば、マイニングシステム１０４は、「痛みを伴う発疹」という句は、「痛みのある発疹」という句に有意義に置換可能であると推定することができる。さらに、マイニングシステム１０４は、免疫システム（および、潜在的に、その他の主題）を説明する場合、「損なわれた」という句は、「弱まった」または「低下した」に有意義に置換可能であると推定することができる。その他の洞察は、全世界的な範囲または領域独立範囲を有し得る。例えば、マイニングシステム１０４は、「に沿って移動する」という句が、「を移動する」または「を進む」に有意義に置換可能であり、「お年寄り」という句は、「年配者」、もしくは「老人」、または「高齢者」に置換可能であるなどを推定することができる。これらの等価は、結果セット３０２内で医療のコンテキストで示されるが、これらは他のコンテキストにも適用可能である。例えば、人は、職場までの移動を、道路「を移動する」または道路「を進む」と説明することができる。

【００４６】

図３は、それによってトレーニングシステム１０６が句同士の間で、有意義な類似点を識別することができる一機構を例示するためにも有用である。例えば、結果項目は、「発疹」、「お年寄り」、「神経」、「免疫システム」など、同じ語の多くを繰り返す。これらの頻繁に出現する語は、意味的に関係する句の存在に関するテキスト区分を調査するた

めのアンカーポイントとして役立つ場合がある。例えば、一般に発生する「免疫システム」という句に関連するアンカーポイントに注目することによって、トレーニングシステム１０６は、「損なわれた」、「弱まった」、および「低下した」は意味的に交換可能な語に対応し得るという結論を導出することができる。トレーニングシステム１０６は、個別の形でこの調査に取りかかることが可能である。すなわち、トレーニングシステム１０６は、句のアラインメントに関して一時的な仮定を導出することが可能である。それらの仮定に基づいて、トレーニングシステム１０６は、その調査を繰り返して、新しい一時的な仮定を導出することが可能である。任意の時点で、これらの一時的な仮定は、トレーニングシステム１０６が、結果項目の関連性に対する追加の洞察を導出することを可能にでき、代わりに、これらの仮定は、後退を表し、さらなる解析を分かりにくくする可能性もある（その場合、これらの仮定は改正可能である）。このプロセスを通じて、トレーニングシステム１０６は、結果セット内の句の関連性に関する仮定の安定したセットに達することを試みる。

【００４７】

より一般には、この例は、マイニングシステム１０４が、同じ主題に対処するリソース項目のグループ（例えば、基礎となる文書）を事前に識別せずに、クエリの提出だけに基づいて結果項目を識別できることも例示する。すなわち、マイニングシステム１０４は、全体としてリソース項目の主題に関してアグノスティック手法をとることが可能である。図３の例では、リソース項目の大部分は、実際に、同じ主題（疾病のｓｈｉｎｇｌｅｓ）に関する可能性が高い。しかし、（１）この類似性は、文書のメタレベル解析ではなく、クエリだけに基づいて露出され、（２）これらのリソース項目が同じ主題に関係するという要件は存在しない。

【００４８】

図４に進むと、この図は、結果セット（R_A）内の結果項目（R_{A1}～R_{AN}）の初期のペアリング（ｐａｉｒｉｎｇ）を確立するために（図１の）準備モジュール１２０を使用することが可能な様式を示す。この場合、準備モジュール１２０は、（結果項目の自己同一的なペアリングを除いて）結果セット内のそれぞれの結果項目と他のすべての結果項目との間のリンクを確立することが可能である。例えば、第１の対は、結果項目R_{AI}を結果項目R_{A2}に接続する。第２の対は、結果項目R_{AI}を結果項目R_{A3}に接続する、等々である。実際には、準備モジュール１２０は、１つまたは複数のフィルタリング要件に基づいて、結果項目同士の間の関連性を制約することができる。セクションＢは、準備モジュール１２０が結果項目のペアワイズのマッチングを制約できる様式に関して追加の情報を提供することになる。

【００４９】

繰り返すと、上記の様式でペアリングされた結果項目は、文の断片を含めて、それらのそれぞれのリソース項目の任意の部分に対応し得る。これは、マイニングシステム１０４は、並列文を識別する明示的な任務なしに、トレーニングセットを確立できることを意味する。すなわち、トレーニングシステム１０６は、文レベルの並列性の活用に依存しない。しかし、トレーニングシステム１０６は、結果項目が完全文（または、テキストのより大きな単位）を含むトレーニングセットを成功裏に処理することも可能である。

【００５０】

図５は、異なる結果のセットからのペアワイズのマッピングを組み合わせて、ストア１２２内にトレーニングセットを形成する様式を例示する。すなわち、クエリQ_Aは結果セットR_Aをもたらし、結果セットR_Aは、今度は、ペアワイズにマッチングされた結果セットTS_Aをもたらす。クエリQ_Bは結果セットR_Bをもたらし、結果セットR_Bは、今度は、ペアワイズにマッチングされた結果セットTS_Bをもたらす、等々である。準備モジュール１２０は、これらの異なるペアワイズにマッチングされた結果セットを組み合わせて、連結させて、トレーニングセットを作成する。全体として、このトレーニングセットは、さらなる調査のために、結果項目同士の間の暫定的なアラインメントの初期セットを確立する。トレーニングシステム１０６は、反復様式でトレーニングセットに関して動作して

、真に関係するテキスト区分を明らかにするアラインメントのサブセットを識別する。最終的に、トレーニングシステム１０６は、それらのアラインメント内に示された意味的に関係する句を識別することを追求する。

【００５１】

このセクションの最後の要点として、図１では、システム１００の異なる構成要素同士の間に破線が引かれている点に留意されたい。これは、任意の構成要素によって下された結論は、他の構成要素の動作を修正するために使用可能であることを図で表す。例えば、ＳＭＴ機能性１２４は、準備モジュール１２０が結果セットのその初期のフィルタリングおよびペアリングを実行する様式に関係するある種の結論を下すことが可能である。準備モジュール１２０は、このフィードバックを受信して、それに応答して、そのフィルタリング行動またはマッチング行動を修正することが可能である。もう１つの事例では、ＳＭＴ機能性１２４または準備モジュール１２０は、例えば、反復コンテンツ内および交番タイプのコンテンツ内で豊富な結果セットを抽出するためのクエリ公式化方策の能力に関係するなど、ある種のクエリ公式化方策の有効性に関する結論を下すことが可能である。クエリ準備モジュール１１２は、このフィードバックを受信して、それに応答して、その行動を修正することが可能である。より詳細には、一事例では、ＳＭＴ機能性１２４または準備モジュール１２０は、別の一連のクエリ内に含むために有用であり得る主要用語または主要句を発見して、解析のための追加の結果セットをもたらすことが可能である。フィードバックに関するさらに他の機会がシステム１００内に存在し得る。

【００５２】

Ｂ．例示的なプロセス

図６～８は、図１のシステム１００の動作の一様式を説明する手順（６００，７００、８００）を示す。システム１００の動作の基礎となる原理は、セクションＡですでに紹介されているため、このセクションでは、いくつかの動作は要約の形で対処される。

【００５３】

図６から始めると、この図は、マイニングシステム１０４およびトレーニングシステム１０６の動作の概要を表す手順６００を示す。より詳細には、動作の第１の段階は、マイニングシステム１０４によって実行されるマイニング動作６０２を説明し、一方、動作の第２の段階は、トレーニングシステム１０６によって実行されるトレーニング動作６０４を説明する。

【００５４】

ブロック６０６において、マイニングシステム１０４は、クエリのセットを構築することによってプロセス６００を開始する。マイニングシステム１０４は、異なる方策を使用して、この任務を実行することが可能である。一事例では、マイニングシステム１０４は、例えば、クエリログなどから取得されるような、ユーザによって探索エンジンにこれまで提出された実際のクエリのセットを抽出することが可能である。もう１つの事例では、マイニングシステム１０４は、任意の参照ソースまたは参照ソースの組合せに基づいて、「人工」クエリを構築することが可能である。例えば、マイニングシステム１０４は、Ｗｉｋｉｐｅｄｉａなどの百科事典的参照ソースの分類索引から、またはシソーラスなどから、クエリ用語を抽出することが可能である。単なる一例を挙げると、マイニングシステム１０４は、参照ソースを使用して、異なる病名を含むクエリの収集物を生成することが可能である。マイニングシステム１０４は、１つまたは複数のその他の用語を用いて、それらの病名を補完して、戻された結果セットに注目することを助けることが可能である。例えば、マイニングシステム１０４は、「ｓｈｉｎｇｌｅｓおよびｚｏｓｔｅｒ」におけるように、その正式な医療同等物を用いてそれぞれの一般的な病名を結合させることが可能である。または、マイニングシステム１０４は、「ｓｈｉｎｇｌｅｓおよびｐｒｅｖｅｎｔｉｏｎ（予防）」など、その病名に若干関係しない別のクエリ用語を用いてそれぞれの病名を結合させることが可能である、等々である。

【００５５】

より広く考えると、ブロック６０６におけるクエリ選択は、異なる包括的な目的によっ

て支配される場合がある。一事例では、マイニングシステム１０４は、特定の領域に注目するクエリの準備を試みることが可能である。この方策は、その特定の領域に向けて多少重み付けられた句を表面化させる際に有効な場合がある。もう１つの事例では、マイニングシステム１０４は、より広い範囲の領域を詳細に調べるクエリの準備を試みることが可能である。この方策は、本質的により領域独立である句を表面化させる際に有効な場合がある。いずれの場合も、マイニングシステム１０４は、上で説明されたように、反復コンテンツ内および交番タイプのコンテンツ内の両方において豊富な結果項目を取得することを追求する。さらに、これらのクエリ自体は、依然として、リソース項目同士の間の類似の主題の任意のタイプの先験的解析ではなく、構造化されていないリソースから並列性を抽出するための主な手段である。

【００５６】

最終的に、マイニングシステム１０４は、そのクエリの選択の有効性を明らかにするフィードバックを受信することが可能である。このフィードバックに基づいて、マイニングシステム１０４は、マイニングシステム１０４がどのようにクエリを構築するかを支配する規則を修正することが可能である。加えて、このフィードバックは、クエリを公式化するために使用可能な特定のキーワードまたは主要句を識別することが可能である。

【００５７】

ブロック６０８において、マイニングシステム１０４は、それらのクエリを取出しモジュール１１６に提出する。取出しモジュール１１６は、今度は、これらのクエリを使用して、構造化されていないリソース１１０内の探索動作を実行する。

【００５８】

ブロック６１０において、マイニングシステム１０４は、取出しモジュール１１６から結果セットを受信し戻す。これらの結果セットは、結果項目のそれぞれのグループを含む。それぞれの結果項目は、構造化されていないリソース１１０内の対応するリソース項目から抽出されたテキスト区分に対応し得る。

【００５９】

ブロック６１２において、マイニングシステム１０４は、トレーニングセットを作成するために、それらの結果セットの初期の処理を実行する。上述のように、この動作は、２つの構成要素を含むことが可能である。フィルタリング構成要素において、マイニングシステム１０４は、それらの結果セットを制約して、意味的に関係する句を識別する際に有用な可能性が低い情報を除去するかまたは無視する。マッチング構成要素において、マイニングシステム１０４は、例えば、セット単位ベースで、結果項目の対を識別する。図４は、１つの例示的な結果セットとの関連でこの動作を図で示す。図７は、ブロック６１２において実行される動作に関する追加の詳細を提供する。

【００６０】

ブロック６１４において、トレーニングシステム１０６は、トレーニングセットに関して動作するために統計的技術を使用して、翻訳モデル１０２を導出する。任意のタイプの句指向の手法など、任意の統計的機械翻訳手法を使用して、この動作を実行することが可能である。一般に、翻訳モデル１０２は、出力句ｙが所与の入力句ｘを表す確率を画定するＰ（ｙ｜ｘ）として表現可能である。ベイズ規則を使用すると、これはＰ（ｙ｜ｘ）＝Ｐ（ｘ｜ｙ）Ｐ（ｙ）／Ｐ（ｘ）として表現可能である。トレーニングシステム１０６は、Ｐ（ｘ｜ｙ）Ｐ（ｙ）を最大化する傾向にある入力句ｘから学習マッピングするために、トレーニングセットの調査に基づいて、この表現によって画定された確率を明らかにするために動作する。上述のように、この調査は本質的に反復的である。動作のそれぞれの段階で、トレーニングシステム１０６は、トレーニングセット内の句（および、全体としてテキスト区分）のアラインメントに関する一時的な結論を下すことが可能である。句指向のＳＭＴ手法において、これらの一時的な結論は、翻訳表などを使用して表現可能である。

【００６１】

ブロック６１６において、トレーニングシステム６１６は、満足のいくアラインメント

結果が達成されていることを表示する終了条件に達しているかどうかを決定する。この決定を行うために、よく知られているバイリンガルエバリュエーションアンダースタディ（Ｂｉｌｉｎｇｕａｌ Ｅｖａｌｕａｔｉｏｎ Ｕｎｄｅｒｓｔｕｄｙ）（ＢＬＥＵ）スコアなど、任意の測定基準を使用することが可能である。

【００６２】

ブロック６１８において、満足のいく結果に達していない場合、トレーニングシステム１０６は、トレーニングの際に使用されるその仮定のうちのいずれかを修正する。これは、結果項目内の句が互いにどのように関係するか（および、テキスト区分が全体として互いにどのように関係するか）に関して一般的な作業仮説を修正する効果を有する。

【００６３】

終了条件が満たされている場合、トレーニングシステム１０６は、そのトレーニングセット内の意味的に関係する句同士の間に識別されたマッピングを有することになる。これらのマッピングを画定するパラメータは、翻訳モデル１０２を確立する。かかる翻訳モデル１０２の使用に内在する推定は、テキストの新たに遭遇されたインスタンスはそのトレーニングセット内で発見されたパターンに類似することになるというものである。

【００６４】

図６の手順は、異なる様式において異なってよい。例えば、代替の実施形態では、ブロック６１４におけるトレーニング動作は、統計解析および規則ベースの解析の組合せを使用して、翻訳モデル１０２を導出することが可能である。もう１つの修正では、ブロック６１４内のトレーニング動作は、そのトレーニング任務を複数の副次的任務に分けて、実質的に、複数の翻訳モデルを作成することが可能である。このトレーニング動作は、次いで、それらの複数の翻訳モデルを単一の翻訳モデル１０２に結合することが可能である。もう１つの修正では、シソーラスから取得された情報など、参照ソースを使用して、ブロック６１４内のトレーニング動作を開始することまたは「準備すること」が可能である。さらに他の修正が可能である。

【００６５】

図７は、図６のブロック６１２においてマイニングシステム１０４によって実行されたフィルタリング処理およびマッチング処理に関する追加の詳細を提供する手順７００を示す。

【００６６】

ブロック７０２において、マイニングシステム１０４は、１つまたは複数の要件に基づいて、元の結果セットをフィルタリングする。この動作は、ペアワイズのマッチングに最も適した候補と見なされる結果項目のサブセットを識別する効果を有する。この動作は、（例えば、低い関連性を有すると評価された結果項目を除去または無視することによって）トレーニングセットの複雑さ、およびトレーニングセット内のノイズ量を低減するのに役立つ。

【００６７】

一事例では、マイニングシステム１０４は、結果項目に関連するランキングスコアに基づいて、ペアワイズのマッチングに適した候補として、それらの結果項目を識別することが可能である。反対に述べると、マイニングシステム１０４は、所定の関連性しきい値未満のランキングスコアを有する結果項目を除去することが可能である。

【００６８】

代わりに、または加えて、マイニングシステム１０４は、（例えば、それらの結果セット内に出現する語の共通性に基づいて）それらの結果セット内で見出された典型的なテキスト特徴を表現するそれぞれの結果セットに関して語彙的な署名を生成することが可能である。マイニングシステム１０４は、次いで、それぞれの結果項目をその結果セットに関連する語彙的な署名と比較することが可能である。マイニングシステム１０４は、この比較に基づいて、ペアワイズのマッチングに適した候補として結果項目を識別することが可能である。反対に述べると、マイニングシステム１０４は、所定の量だけそれらの語彙的な署名とは異なる結果項目を除去することが可能である。それほど正式でない述べ方をす

ると、マイニングシステム１０４は、それらのそれぞれの結果セット内で「突出している」結果項目を除去することが可能である。

【００６９】

代わりに、または加えて、マイニングシステム１０４は、それぞれの結果項目が結果セット内のそれぞれの他の結果項目とどれだけ類似するかを識別する類似性スコアを生成することが可能である。マイニングシステム１０４は、この決定を行うために、これに限定されないが、コサイン類似性測定基準（ｃｏｓｉｎｅ　ｓｉｍｉｌａｒｉｔｙ　ｍｅｔｒｉｃ）など、任意の類似性測定基準に依存することも可能である。マイニングシステム１０４は、それらの類似性スコアに基づいて、ペアワイズのマッチングに適した候補として結果項目を識別することが可能である。反対に述べると、マイニングシステム１０４は、類似性スコアによって明らかにされた、所定の量を超える量だけ互いと異なるため、マッチングに関する良好な候補ではない結果項目の対を識別することが可能である。

【００７０】

代わりに、または加えて、マイニングシステム１０４は、例えば、ｋ最近傍クラスタリング技術または任意のその他のクラスタリング技術を使用して、類似の結果項目のグループを決定するために、結果セット内の結果項目に関してクラスタ解析を実行することが可能である。マイニングシステム１０４は、次いで、異なるクラスタ全域の候補としてではなく、ペアワイズのマッチングに適した候補として、それぞれのクラスタ内の結果項目を識別することができる。

【００７１】

マイニングシステム１０４は、さらに他の動作を実行して、構造化されていないリソース１１０から収集された結果項目をフィルタリングまたは「処分する」ことが可能である。ブロック７０２は、結果として、フィルタリングされた結果セットの生成をもたらす。

【００７２】

ブロック７０４において、マイニングシステム１０４は、フィルタリングされた結果セット内の対を識別する。既に説明されたように、図４は、例示的な結果セットとの関連でこの動作をどのように実行できるかを示す。

【００７３】

ブロック７０６において、マイニングシステム１０４は、（個々の結果セットに関連する）ブロック７０４の結果を組み合わせて、トレーニングセットを提供することが可能である。既に説明されたように、図５は、この動作をどのように実行できるかを示す。

【００７４】

ブロック７０４は、説明を容易にするために、ブロック７０２とは別として示されるが、ブロック７０２および７０４は、統合された動作として実行可能である。さらに、ブロック７０２および７０４のフィルタリング動作ならびにマッチング動作は、動作の複数の段階にわたって分散可能である。例えば、マイニングシステム１０４は、ブロック７０６に続き、それらの結果項目にさらなるフィルタリングを実行できる。さらに、トレーニングシステム１０６は、（図６のブロック６１４～６１８によって表されるように）その反復処理の過程で、それらの結果項目にさらなるフィルタリングを実行できる。

【００７５】

別の変形形態として、ブロック７０４は、個々の結果セット内の結果項目の対を確立する関連で説明された。しかし、もう１つのモードでは、マイニングシステム１０４は、異なる結果セットの全体で候補の対を確立することが可能である。

【００７６】

図８は、翻訳モデル１０２の例示的な応用を説明する手順８００を示す。

【００７７】

ブロック８０２において、アプリケーションモジュール１０８は入力句を受信する。

【００７８】

ブロック８０４において、アプリケーションモジュール１０８は、翻訳モデル１０２を使用して、入力句を出力句に変換する。

【００７９】
ブロック８０６において、アプリケーションモジュール１０８は、その出力句に基づいて出力結果を生成する。異なるアプリケーションモジュールは、異なるそれぞれの利益を達成するために、異なるそれぞれの出力結果を提供することができる。
【００８０】
１つのシナリオでは、アプリケーションモジュール１０８は、翻訳モデル１０２を使用して、クエリ修正動作を実行することが可能である。この場合、アプリケーションモジュール１０８は、探索クエリとしてこの入力句を扱う。アプリケーションモジュール１０８は、この出力句を使用して、探索クエリを置換または補完することが可能である。例えば、この入力句が「ｓｈｉｎｇｌｅｓ」である場合、アプリケーションモジュール１０８は、その出力句「ｚｏｓｔｅｒ」を使用して、「ｓｈｉｎｇｌｅｓおよびｚｏｓｔｅｒ」の補完されたクエリを生成することが可能である。アプリケーションモジュール１０８は、次いで、拡張されたクエリを探索エンジンに提示できる。
【００８１】
もう１つのシナリオでは、アプリケーションモジュール１０８は、翻訳モデル１０２を使用して、索引付け分類決定を行うことが可能である。この場合、アプリケーションモジュール１０８は、いずれかのテキストコンテンツを分類されることになる文書から抽出して、入力句としてそのテキストコンテンツを扱うことができる。アプリケーションモジュール１０８は、その出力句を使用して、その文書の主題に関する追加の洞察を集めることが可能であり、今度は、その文書の適切な分類を実現するために、その追加の洞察を使用することが可能である。
【００８２】
もう１つのシナリオでは、アプリケーションモジュール１０８は、翻訳モデル１０２を使用して、任意のタイプのテキスト改正動作を実行できる。この場合、アプリケーションモジュール１０８は、テキスト改正に関する候補としてその入力句を扱うことができる。アプリケーションモジュール１０８は、その出力句を使用して、その入力句が改正され得る様式を示唆することが可能である。例えば、その入力句が、「痛みのある発疹」という、どちらかといえば冗長なテキストに対応すると仮定する。アプリケーションモジュール１０８は、この入力句をより簡潔な「痛みを伴う発疹」に置換することが可能であることを示唆できる。この示唆を行う際に、アプリケーションモジュール１０８は、（その出力句が文法的誤りおよび／または綴り誤りを含まないと仮定して）元の句のいかなる文法的誤りおよび／または綴り誤りも修正することが可能である。一事例では、アプリケーションモジュール１０８は、ユーザが異なる改正の妥当性を評価することを可能にする何らかのタイプの情報に加えて、ユーザが入力句をどのように改正できるかに関する複数の選択肢をユーザに提供することが可能である。例えば、アプリケーションモジュール１０８は、（代表的な例を単に挙げると）あなたの考えを表現するこの方法は著者の８０％によって使用されていると表示することによって、特定の改正に注釈をつけることができる。代わりに、アプリケーションモジュール１０８は、１つまたは複数の要件に基づいて、自動的に改正を行うことが可能である。
【００８３】
もう１つのテキスト改正事例では、アプリケーションモジュール１０８は、翻訳モデル１０２を使用して、テキスト切断動作を実行できる。例えば、アプリケーションモジュール１０８は、移動体電話デバイスなど、小型スクリーン表示デバイス上に提示するために元のテキストを受信することが可能である。アプリケーションモジュール１０８は、翻訳モデル１０２を使用して、入力句として扱われるテキストをそのテキストの省略バージョンに変換することが可能である。もう１つの事例では、アプリケーションモジュール１０８は、この手法を使用して、元の句がＴｗｉｔｔｅｒのような通信機構など、そのメッセージにサイズ制約を課す任意のメッセージ送信機構と互換性を持つように、その元の句を短縮することが可能である。
【００８４】

もう１つのテキスト改正事例では、アプリケーションモジュール１０８は、翻訳モデル１０２を使用して、文書または句を要約することが可能である。例えば、アプリケーションモジュール１０８は、この手法を使用して、元の要約の長さを削減することが可能である。もう１つの事例では、アプリケーションモジュール１０８は、この手法を使用して、テキストのより長い節に基づいてタイトルを提案することが可能である。代わりに、アプリケーションモジュール１０８は、翻訳モデル１０２を使用して、文書または句を拡張することが可能である。

【００８５】

もう１つのシナリオでは、アプリケーションモジュール１０８は、翻訳モデル１０２を使用して、広告情報の拡張を実行できる。この場合、例えば、広告主は、広告コンテンツ（例えば、ウェブページまたはその他のネットワークアクセス可能なコンテンツ）に関連する初期のトリガキーワードを選択した可能性がある。エンドユーザがこれらのトリガキーワードを入力した場合、またはユーザが、それとも、これらのトリガキーワードに関連するコンテンツを消費している場合、広告機構は、そのユーザをそれらのトリガキーワードに関連する広告コンテンツに向けることができる。この場合、アプリケーションモジュール１０８は、翻訳モデル１０２を使用して拡張されることになる入力句として、トリガキーワードの初期のセットを考慮することが可能である。代わりに、または加えて、アプリケーションモジュール１０８は、広告コンテンツ自体を入力句として扱うこともできる。アプリケーションモジュール１０８は、次いで、翻訳モデル１０２を使用して、広告コンテンツに関係するテキストを示唆することが可能である。広告主は、その示唆されたテキストに基づいて、１つまたは複数のトリガキーワードを提供することが可能である。

【００８６】

上述のアプリケーションは、代表的なものであり、包括的ではない。その他のアプリケーションが可能である。

【００８７】

上の説明では、出力句は入力句と同じ言語で表現されるという仮定が立てられた。この場合、出力句は、入力句のパラフレーズと見なすことができる。もう１つの事例では、マイニングシステム１０４およびトレーニングシステム１０６は、第１の言語の句を別の言語（または複数の他の言語）の対応する別の言語の句に変換する翻訳モデル１０２を作成するために使用可能である。

【００８８】

二言語コンテキストまたは多言語コンテキストで動作するために、マイニングシステム１０４は、二言語情報または多言語情報に関する上述の同じ基本的な動作を実行できる。一事例では、マイニングシステム１０４は、ネットワーク環境内で並列クエリを提出することによって、二言語の結果セットを確立することが可能である。すなわち、マイニングシステム１０４は、第１の言語で表現されたクエリのあるセットと、第２の言語で表現されたクエリの別のセットとを提出することが可能である。例えば、マイニングシステム１０４は、「ｒａｓｈ　ｚｏｓｔｅｒ」という句を提出して、英語の結果セットを生成し、「ｚｏｓｔｅｒ　ｅｒｕｐｃｉｏｎ　ｄｅ　ｐｉｅｌ」という句を提出して、英語の結果セットのスペイン語の同等物を生成することが可能である。マイニングシステム１０４は、次いで、英語の結果項目をスペイン語の結果項目にリンクする対を確立することが可能である。このマッチング動作の目的は、トレーニングシステム１０６が英語およびスペイン語の意味的に関係する句の間のリンクを識別することを可能にするトレーニングセットを提供することである。

【００８９】

もう１つの事例では、マイニングシステム１０４は、「ｓｈｉｎｇｌｅｓ　ｒａｓｈ　ｅｒｕｐｃｉｏｎ　ｄｅ　ｐｉｅｌ」というクエリの場合など、英語およびスペイン語の主要用語の両方を組み合わせるクエリを提出することが可能である。この手法では、取出しモジュール１１６は、英語で表現された結果項目とスペイン語で表現された結果項目とを組み合わせる結果セットを提供することが予測できる。マイニングシステム１０４は、

次いで、それらの結果項目が英語で表現されているかまたはスペイン語で表現されているかを区別せずに、この混合された結果セット内の異なる結果項目間のリンクを確立することが可能である。トレーニングシステム１０６は、混合されたトレーニングセット内の基礎となるパターンに基づいて、単一の翻訳モデル１０２を生成することが可能である。使用の際、翻訳モデル１０２は、単一言語モードで適用可能であり、この場合、翻訳モデル１０２は、入力句と同じ言語で出力句を生成するように制約される。または、翻訳モデル１０２は、二言語モードで動作することも可能であり、その場合、翻訳モデル１０２は、入力句と比べて異なる言語で出力句を生成するように制約される。または、翻訳モデル１０２は、制約されないモードで動作することが可能であり、その場合、翻訳モデル１０２は、結果を両方の言語で提案する。

【００９０】

Ｃ．代表的な処理機能性

図９は、上述の機能の任意の態様を実施するために使用可能な例示的な電気データ処理機能性９００を記載する。図１および２を参照すると、例えば、システム１００またはコンピューティング機能性２０２の任意の態様などを実施するために、図９に示される処理機能性９００のタイプを使用することが可能である。一事例では、処理機能性９００は、１つまたは複数の処理デバイスを含む、任意のタイプのコンピューティングデバイスに対応し得る。

【００９１】

処理機能性９００は、ＲＡＭ９０２およびＲＯＭ９０４などの揮発性メモリならびに不揮発性メモリと同様に、１つまたは複数の処理デバイス９０６を含むことが可能である。処理機能性９００はまた、ハードディスクモジュール、光ディスクモジュールなど、様々な媒体デバイス９０８をオプションで含む。処理機能性９００は、（１つまたは複数の）処理デバイス９０６がメモリ（例えば、ＲＡＭ９０２、ＲＯＭ９０４、またはその他の場所）によって維持された命令を実行する場合、上で識別された様々な動作を実行できる。より一般的には、命令およびその他の情報は、静的メモリ記憶デバイス、磁気記憶デバイス、光記憶デバイスなどを含むが、これらに限定されない、任意のコンピュータ可読媒体９１０上に格納可能である。コンピュータ可読媒体という用語は、複数の記憶デバイスも包括する。コンピュータ可読媒体という用語は、例えば、有線伝送、ケーブル伝送、無線伝送など、第１の位置から第２の位置まで送信される信号も包括する。

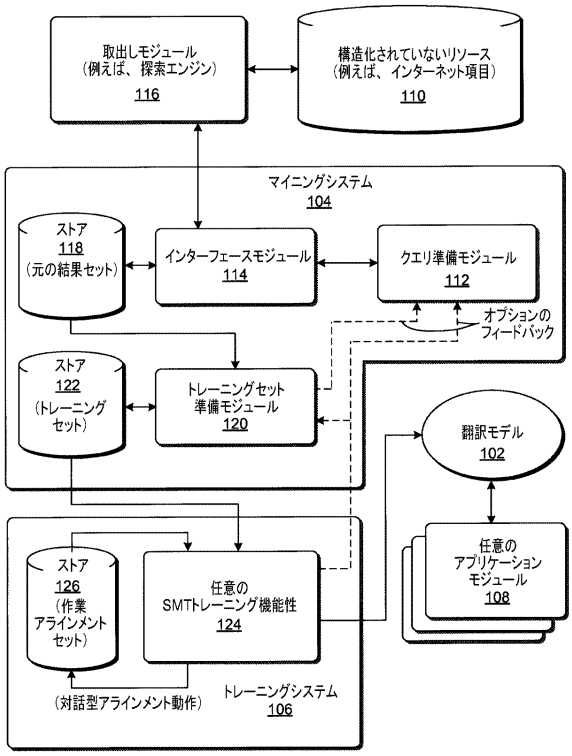
【００９２】

処理機能性９００は、（入力モジュール９１４を経由して）ユーザから様々な入力を受信して、（出力モジュールを経由して）ユーザに様々な出力を提供するための入出力モジュール９１２も含む。１つの特定の出力機構は、提示モジュール９１６および関連するグラフィカルユーザインターフェイス（ＧＵＩ）９１８を含むことが可能である。処理機能性９００は、１つまたは複数の通信導管９２２を経由して他のデバイスとデータを交換するための１つまたは複数のネットワークインターフェース９２０を含むことも可能である。１つまたは複数の通信バス９２４は、上述の構成要素を通信可能に一緒に結合する。

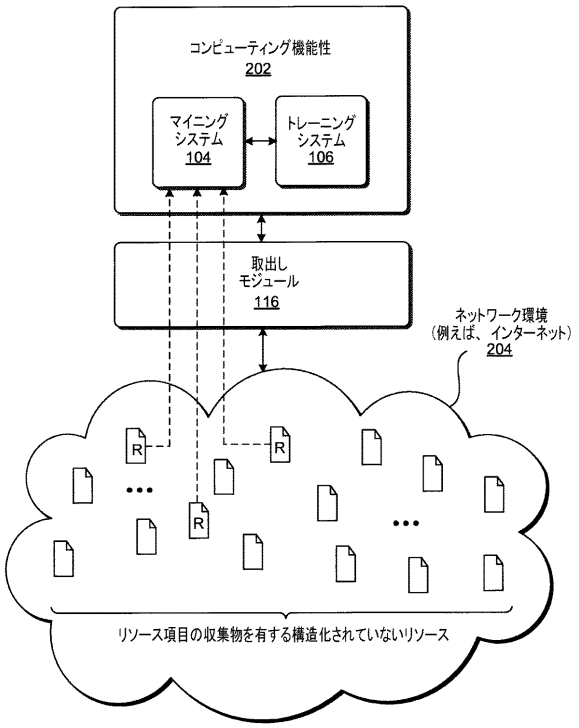
【００９３】

本主題は、構造的特徴および／または方法論的動作に特定の言語で説明されているが、添付の特許請求の範囲内で画定される本主題は、上述の特定の特徴または動作に限定されるとは限らない点を理解されたい。むしろ、上述の特定の特徴および動作は、特許請求の範囲を実施する例示的な形態として開示される。

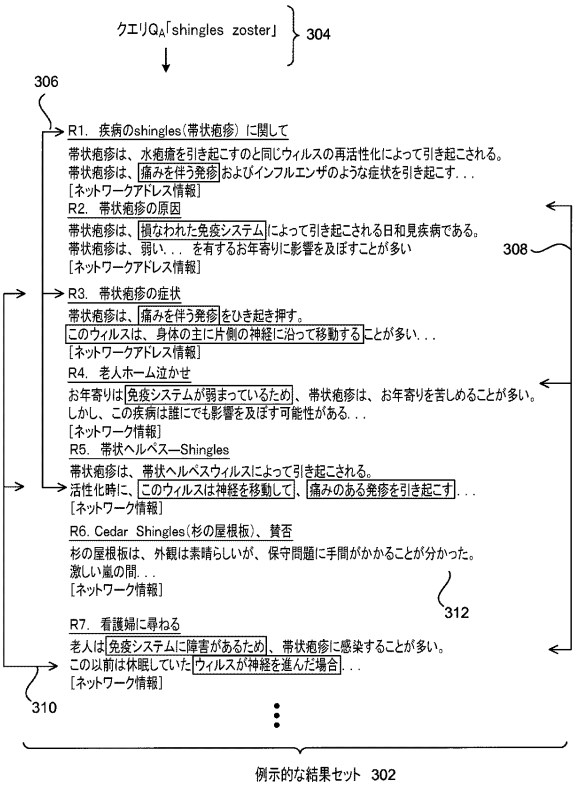
【図１】


取出しモジュール
（例えば、探索エンジン）
116
構造化されていないリソース
（例えば、インターネット項目）
110
マイニングシステム
104
ストア
118
（元の結果セット）
インターフェースモジュール
114
クエリ準備モジュール
112
オプションの
フィードバック
ストア
122
（トレーニング
セット）
トレーニングセット
準備モジュール
120
翻訳モデル
102
ストア
126
（作業
アラインメント
セット）
任意の
SMTトレーニング機能性
124
任意の
アプリケーション
モジュール
108
（対話型アラインメント動作）
トレーニングシステム
106

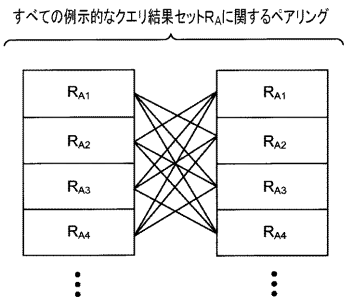
【図２】


コンピューティング機能性
202
マイニング
システム
104
トレーニング
システム
106
取出し
モジュール
116
ネットワーク環境
（例えば、インターネット）
204
R
R
R
リソース項目の収集物を有する構造化されていないリソース

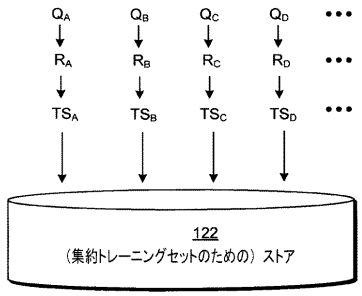
【図３】


クエリQA「shingles zoster」
304
306
R1. 疾病のshingles（帯状疱疹）に関して
帯状疱疹は、水疱瘡を引き起こすのと同じウィルスの再活性化によって引き起こされる。
帯状疱疹は、痛みを伴う発疹およびインフルエンザのような症状を引き起こす. . .
[ネットワークアドレス情報]
R2. 帯状疱疹の原因
帯状疱疹は、損なわれた免疫システムによって引き起こされる日和見疾病である。
帯状疱疹は、弱い. . . を有するお年寄りに影響を及ぼすことが多い
[ネットワークアドレス情報]
308
R3. 帯状疱疹の症状
帯状疱疹は、痛みを伴う発疹をひき起き押す。
このウィルスは、身体の主に片側の神経に沿って移動することが多い. . .
[ネットワークアドレス情報]
R4. 老人ホーム泣かせ
お年寄りは免疫システムが弱まっているため、帯状疱疹は、お年寄りを苦しめることが多い。
しかし、この疾病は誰にでも影響を及ぼす可能性がある. . .
[ネットワーク情報]
R5. 帯状ヘルペス—Shingles
帯状疱疹は、帯状ヘルペスウィルスによって引き起こされる。
活性化時に、このウィルスは神経を移動して、痛みのある発疹を引き起こす. . .
[ネットワーク情報]
R6. Cedar Shingles（杉の屋根板）、賛否
杉の屋根板は、外観は素晴らしいが、保守問題に手間がかかることが分かった。
激しい嵐の間. . .
[ネットワーク情報]
312
R7. 看護婦に尋ねる
老人は免疫システムに障害があるため、帯状疱疹に感染することが多い。
この以前は休眠していたウィルスが神経を進んだ場合. . .
[ネットワーク情報]
310
例示的な結果セット 302

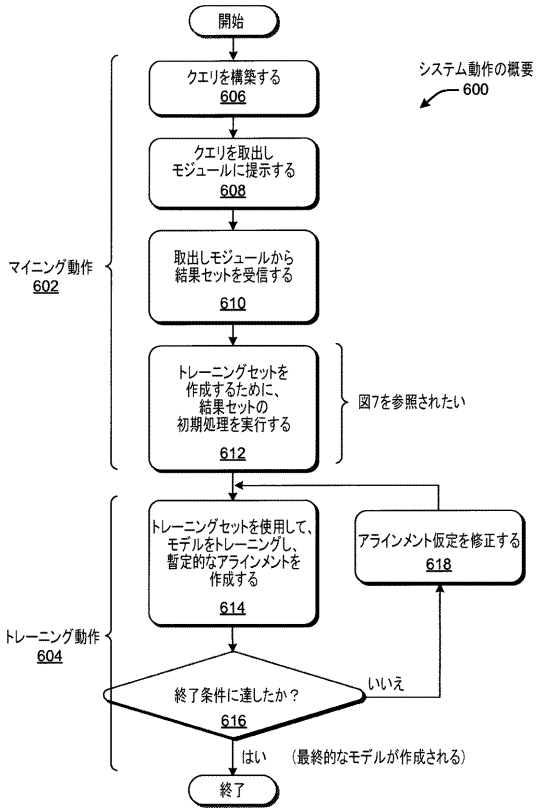
【図４】


すべての例示的なクエリ結果セットRAに関するペアリング
RA1
RA2
RA3
RA4
RA1
RA2
RA3
RA4

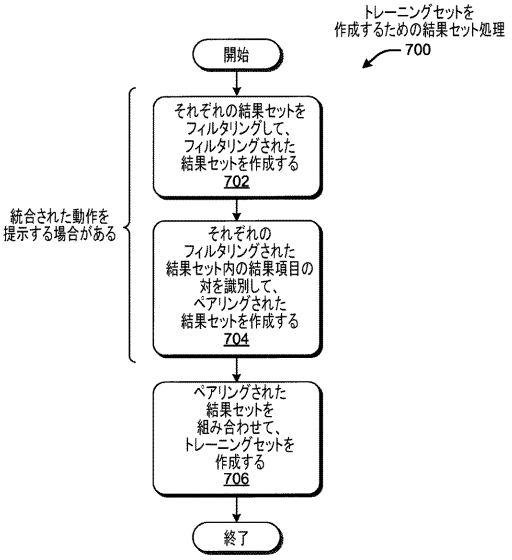
【図５】


QA
QB
QC
QD
RA
RB
RC
RD
TSA
TSB
TSC
TSD
122
（集約トレーニングセットのための）ストア

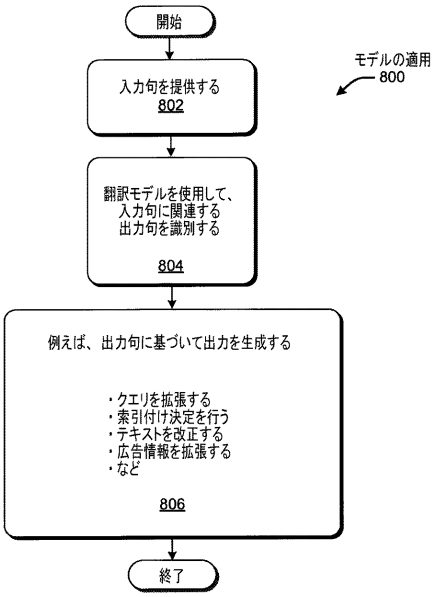
【図６】


開始
クエリを構築する
606
クエリを取出し
モジュールに提示する
608
取出しモジュールから
結果セットを受信する
610
トレーニングセットを
作成するために、
結果セットの
初期処理を実行する
612
トレーニングセットを使用して、
モデルをトレーニングし、
暫定的なアラインメントを
作成する
614
終了条件に達したか？
616
アラインメント仮定を修正する
618
いいえ
はい
（最終的なモデルが作成される）
終了
システム動作の概要
600
マイニング動作
602
図7を参照されたい
トレーニング動作
604

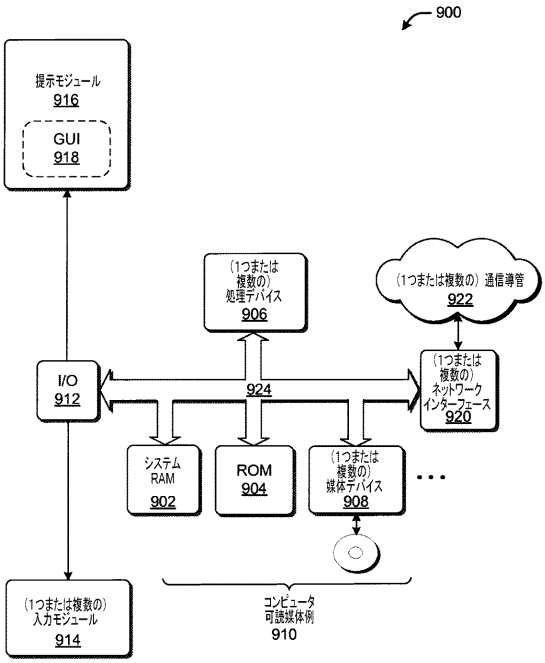
【図７】


トレーニングセットを
作成するための結果セット処理
700
開始
それぞれの結果セットを
フィルタリングして、
フィルタリングされた
結果セットを作成する
702
それぞれの
フィルタリングされた
結果セット内の結果項目の
対を識別して、
ペアリングされた
結果セットを作成する
704
ペアリングされた
結果セットを
組み合わせて、
トレーニングセットを
作成する
706
終了
統合された動作を
提示する場合がある

【図８】


開始
入力句を提供する
802
翻訳モデルを使用して、
入力句に関連する
出力句を識別する
804
例えば、出力句に基づいて出力を生成する
・クエリを拡張する
・索引付け決定を行う
・テキストを改正する
・広告情報を拡張する
・など
806
終了
モデルの適用
800

【図９】


900
提示モジュール
916
GUI
918
（1つまたは
複数の）
処理デバイス
906
（1つまたは複数の）通信導管
922
I/O
912
924
（1つまたは
複数の）
ネットワーク
インターフェース
920
システム
RAM
902
ROM
904
（1つまたは
複数の）
媒体デバイス
908
（1つまたは複数の）
入力モジュール
914
コンピュータ
可読媒体例
910

フロントページの続き

(72)発明者　ウィリアム　ビー．ドーラン
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　エルシーエー－インターナショナル　パテンツ内
(72)発明者　クリストファー　ジェイ．ブロケット
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　エルシーエー－インターナショナル　パテンツ内
(72)発明者　ジュリオ　ジェイ．カスティーリョ
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　エルシーエー－インターナショナル　パテンツ内
(72)発明者　ルクレティア　エイチ．ヴァンダーヴェンデ
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　エルシーエー－インターナショナル　パテンツ内

審査官　齊藤　貴孝

(56)参考文献　米国特許出願公開第２００５／０１０２６１４（ＵＳ，Ａ１）
米国特許出願公開第２００７／００６７２８１（ＵＳ，Ａ１）
特開２００４－２５２４９５（ＪＰ，Ａ）
特開平１０－０７４２１０（ＪＰ，Ａ）
特開２００２－２４５０７０（ＪＰ，Ａ）
特開２００５－２８５１２９（ＪＰ，Ａ）
特開２００１－０４３２３６（ＪＰ，Ａ）
特開２００６－２８５９８２（ＪＰ，Ａ）
特開２００４－２０６５１７（ＪＰ，Ａ）
米国特許出願公開第２００２／０１９８７０１（ＵＳ，Ａ１）
米国特許出願公開第２００３／０２０４４００（ＵＳ，Ａ１）
米国特許出願公開第２００４／０１０２９５７（ＵＳ，Ａ１）
永田　昌明、外２名，機械翻訳最新事情，情報処理，日本，社団法人情報処理学会，２００８年１月１５日，第４９巻，第１号，ｐ．８９－９５

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０
Ｇ０６Ｆ　１７／２８