(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5705472号
(P5705472)
(45)発行日　平成27年4月22日(2015.4.22)　(24)登録日　平成27年3月6日(2015.3.6)

(51)Int.Cl.　F I
*G06F 17/28 (2006.01)*　G06F 17/28 618

請求項の数 3　(全 21 頁)

(21)出願番号　特願2010-166695(P2010-166695)
(22)出願日　平成22年7月26日(2010.7.26)
(65)公開番号　特開2011-28754(P2011-28754A)
(43)公開日　平成23年2月10日(2011.2.10)
審査請求日　平成25年7月23日(2013.7.23)
(31)優先権主張番号　12/509,633
(32)優先日　平成21年7月27日(2009.7.27)
(33)優先権主張国　米国(US)

(73)特許権者　596170170
ゼロックス　コーポレイション
ＸＥＲＯＸ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国、コネチカット州　０６８５６、ノーウォーク、ピーオーボックス　４５０５、グローバー・アヴェニュー　４５
(74)代理人　110001210
特許業務法人ＹＫＩ国際特許事務所
(72)発明者　ミハイル　ザスラフスキ
フランス　ジフ　シュ　リヴェット　リュ　デュ　クロス　1
(72)発明者　マルク　ディムトマン
フランス　グルノーブル　キ　ド　フランス　４

最終頁に続く

(54)【発明の名称】一般化された巡回セールスマン問題としてのフレーズ－ベースの統計的機械翻訳

(57)【特許請求の範囲】
【請求項１】
統計的機械翻訳（ＳＭＴ）および一般化された非対称巡回セールスマン問題（ＧＴＳＰ）グラフを使用してソース言語を目標言語に翻訳する方法であって、
コンピュータが、
ソース言語及び目標言語で表現されたフレーズのペアであるバイ－フレーズを各ノードとし、前記各ノード間のエッジにコストが設定されたＧＴＳＰを形成するステップと、
ソース言語で表現された入力文に含まれる各フレーズを、前記ＧＴＳＰを表すＧＴＳＰグラフ内のノードに対応するバイ－フレーズに基づいて、目標言語に置換するステップと、
前記入力文のブロックを目標言語での表現に置換するために用いた前記ノードを巡回する巡回経路であって、コストの合計が最小となる巡回経路を決定するステップと、
前記巡回経路によって定義される順序で目標言語に置換された各フレーズを出力するステップと、
を実行し、
前記巡回経路を決定するステップは、
ＧＴＳＰグラフの最適巡回を生成するステップと、
前記ＧＴＳＰグラフの３つ以上のノードの接続関係に基づいて決定されたコストに基づいて、前記最適巡回の真のコスト$C_+$を計算するステップと、
前記ＧＴＳＰグラフのエッジに設定されたコストに基づいて、前記最適巡回の見かけの

コスト$C_a$を計算するステップと、
前記真のコスト$C_t$と前記見かけのコスト$C_a$の間の差Ｄを判定するステップと、
前記差Ｄがあらかじめ設定された閾値εより小さいか否かを判定するステップと、
Ｄが前記あらかじめ設定された閾値εより小さい場合には、前記巡回経路として前記最適巡回を出力するステップと、
を包含することを特徴とする方法。
【請求項２】
前記巡回経路を決定するステップは、
前記ＧＴＳＰグラフを非対称巡回セールスマン問題（ＡＴＳＰ）グラフに変換するステップと、
前記ＡＴＳＰグラフを標準巡回セールスマン問題（ＴＳＰ）グラフに変換するステップと、
前記ＴＳＰグラフに対してコンコード（Ｃｏｎｃｏｒｄｅ）ソルバおよびリン・カーニハンのヒューリスティックのうちの少なくとも１つを使用してコストの合計が最小となる巡回経路を決定するステップと、
を包含する請求項１に記載の方法。
【請求項３】
さらに、
Ｄが前記あらかじめ設定された閾値εに等しいか、またはそれを超える場合に、前記グラフ内の少なくとも１つのノードを絞り込み、前記絞り込んだノードを包含する絞り込み済みグラフを生成するステップと、
Ｄがεより小さくなるまで反復的に、１つまたは複数の絞り込み済みグラフについて前記真のコスト$C_t$および見かけのコスト$C_a$を計算し、それらの間の前記差Ｄを判定し、かつ前記差Ｄと前記あらかじめ設定された閾値εを比較するステップと、を包含し、
少なくとも、
前記最適巡回が各ノードを正確に一度だけ訪問し、前記最適巡回内のノードの間の各エッジがそれぞれのバイグラム重みと関連付けられ、かつ前記最適巡回の前記見かけのコスト$C_a$が前記巡回内のすべてのエッジの前記バイグラム重みの合計によって計算されることと、
前記最適巡回の前記真のコスト$C_t$がトリグラム・コストを使用して計算されることと、のうちの一方を含む、
請求項１に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
この出願は、コンピューティング・システム内における統計的機械翻訳（ｓｔａｔｉｓｔｉｃａｌ　ｍａｃｈｉｎｅ　ｔｒａｎｓｌａｔｉｏｎ，ＳＭＴ）に関する。ここで述べる手法は、このほかの翻訳システム、このほかの統計的マッピング応用、および／またはこのほかの翻訳方法の中にも応用を見つけることができることを理解されたい。
【背景技術】
【０００２】
統計的機械翻訳（ＳＭＴ）に対する古典的アプローチは、「バイ－フレーズ（ｂｉ－ｐｈｒａｓｅｓ）」を伴う。「バイ－フレーズ」とは、ソース言語および目標言語の表現またはフレーズのペアであり、この表現またはフレーズのペアは、ソース文から目標（すなわち、翻訳された）文を構成するためのビルディング・ブロックを形成する。
【０００３】
Ｎ－グラム（Ｎ－ｇｒａｍ）言語モデルは、シーケンス内の次の項目を予測するための確率論的モデルの一種である。Ｎ－グラムは、統計的自然言語処理および遺伝子配列の分析の多様な分野において使用されている。Ｎ－グラムは、与えられたシーケンスからのｎ個の項目のサブシーケンスである。懸案の項目は、音素、音節、文字、単語、塩基対等々

とすることができる。
【０００４】
与えられたソース文Ｓを翻訳するために、古典的なフレーズ・ベースのＳＭＴシステムは、次の形式の対数－線形モデル（ｌｏｇ－ｌｉｎｅａｒ　ｍｏｄｅｌ）を用いる。
【数１】

$$p(t,a\mid s)=1/Z_s \exp\sum_k \lambda_k h_k(s,a,t)$$

ここで、$h_k$は、「特徴」であり、ソース文字列ｓ、目標文字列ｔ、および配列ａの関数である。配列ａは、ソース文字列ｓから目標文字列ｔを組み立てるのに用いられる、バイ－フレーズのシーケンスの表現である。$\lambda_k$は、重みであり、$Z_s$は、ｐ（ｔ，ａ｜ｓ）がペア（ｔ，ａ）について適正な条件付き確率分布となることを保証する正規化因子である。
【０００５】
一旦、対数－線形モデルが定義されてしまえば（トレーニング段階を伴う；たとえば、参照によりこれに援用される非特許文献１を参照されたい）、デコーダの役割は、条件付き確率ｐ（ｔ，ａ｜ｓ）を最大化するペア（ｔ，ａ）を見つけること、および対応する目標文字列ｔを出力することになる。
【０００６】
古典的なシステムは、ヒューリスティックな左から右へのサーチの何らかの変形に基づいており、これは、各ステップにおいて、新しいバイ－フレーズを用いて現在の部分翻訳を拡張しつつ、かつ２つのスコア、すなわちこれまでの部分翻訳の既知の要素についてのスコア、および翻訳を完了するための残りのコストのヒューリスティックな評価を計算しつつ、左から右へと漸増的に候補翻訳（ｔ，ａ）の組み立てを試みる。もっとも頻繁に使用される変形は、いくつかの部分的な候補が並列に維持され、現在の評価が低すぎる候補が取り除かれて、より有望な候補を選ぶビーム－サーチの形式である。
【先行技術文献】
【非特許文献】
【０００７】
【非特許文献１】ロペズ・Ａ（Ｌｏｐｅｚ，Ａ．）著、２００８．Ｓｔａｔｉｓｔｉｃａｌ　Ｍａｃｈｉｎｅ　Ｔｒａｎｓｌａｔｉｏｎ．ＡＣＭ　Ｃｏｍｐｕｔ．Ｓｕｒｖ．４０，３（２００８年８月）、ｐ．１－４９
【発明の概要】
【発明が解決しようとする課題】
【０００８】
左から右へのヒューリスティックなサーチによる翻訳を行う従来技術では、サーチの早い段階で生じる誤った選択から過剰な影響を受けることがあり、早い段階での誤りから回復することは困難であり得る。
【課題を解決するための手段】
【０００９】
グラフ内のノードとしてバイ－フレーズをモデリングすることによってバイグラム（または、より高いＮ－グラム）言語モデルを組み込むフレーズ－ベースのモデルのためのＳＭＴを容易にするシステムおよび方法を述べる。たとえば、統計的機械翻訳（ＳＭＴ）および一般非対称巡回セールスマン問題（ｇｅｎｅｒａｌｉｚｅｄ　ａｓｙｍｍｅｔｒｉｃ　ｔｒａｖｅｌｉｎｇ　ｓａｌｅｓｍａｎ　ｐｒｏｂｌｅｍ，ＧＴＳＰ）のグラフを使用して２つの言語を翻訳する方法は、ＧＴＳＰとしてＳＭＴを定義するステップと、ＧＴＳＰを表すＧＴＳＰグラフ内のノードに対応するバイ－フレーズを使用して入力文のブロックを翻訳するステップと、ＧＴＳＰを解決するステップと、ＧＴＳＰ解決によって定義される順序で翻訳されたブロックを出力するステップと、を包含する。
【図面の簡単な説明】

【００１０】
【図１】ＧＴＳＰアプローチを使用して言語間の翻訳において使用するためのグラフを生成する、フレーズ－ベースのＳＭＴを実行するためのシステムを図解したブロック図である。
【図２】それぞれの「コスト」に従って０～６がラベル付けされた複数のエッジを伴う非対称ＴＳＰ（ａｓｙｍｍｅｔｒｉｃ　ＴＳＰ，ＡＴＳＰ）を標準ＴＳＰに変換するための第１の変換を図解した説明図である。
【図３】ＡＴＳＰを標準ＴＳＰに変換するための第２の変換を図解した説明図である。
【図４】ＧＴＳＰをＡＴＳＰに変換する変換を図解した説明図である。
【図５】既存のエッジのサブセット、すなわちノード「ｔｒａｄｕｃｔｉｏｎ－ｍｔ」に入るか、または出るすべてのエッジだけが示された、ソース文「ｃｅｔｔｅ　ｔｒａｄｕｃｔｉｏｎ　ａｕｔｏｍａｔｉｑｕｅ　ｅｓｔ　ｃｕｒｉｅｕｓｅ」についての遷移グラフを図解した説明図である。
【図６Ａ】１つの出力に対応するＧＴＳＰ巡回を図解した説明図である。
【図６Ｂ】他の１つの出力に対応するＧＴＳＰ巡回を図解した説明図である。
【図７】バイ－フレーズ「ｉ」だけが取り除かれたグラフであって、現在は「ｉ」をカプセル化している拡張バイ－フレーズのいくつかだけがグラフを通る１つの有効な巡回を定義するエッジを伴って示されているグラフを図解した説明図である。
【図８】選択的なオンデマンド絞り込みを図解した説明図である。
【図９Ａ】ＴＳＰとしてフレーズ－ベースの統計的機械翻訳を実行するための方法を図解したフローチャートである。
【図９Ｂ】ＴＳＰとしてフレーズ－ベースの統計的機械翻訳を実行するための代替方法を図解したフローチャートである。
【図１０】トリグラム言語モデルを使用してフレーズ－ベースの翻訳を実行するためのグラフを図解した説明図である。
【図１１】差がεより大きいか、または等しい場合に得られるグラフを図解した説明図である。
【発明を実施するための形態】
【００１１】
グラフ内のノードとしてバイ－フレーズをモデリングすることによってフレーズ－ベースのＳＭＴを容易にするシステムおよび方法を述べる。ここで使用するときの「グラム」は、単語を意味するものであり、バイグラム言語モデルは２語のグループを採用し、トリグラム言語モデルは３語のグループを採用し、以下同様とする。
【００１２】
このシステムおよび方法においては、バイ－フレーズが、グラフ内のノードとしてモデル化される。それに加えて翻訳の構成が、グラフ内のノードの間の「巡回」、すなわち各ノードを正確に一度だけ訪問する経路としてモデル化される。巡回の全体的なコストが、その巡回の間に通り抜けたエッジに関連付けされたコストを加算することによって計算される。
【００１３】
したがって、ここで述べるシステムおよび方法は、ＳＭＴ問題をＧＴＳＰ問題に直接マップし、ＧＴＳＰとしてフレーズ－ベースの翻訳を表現する。
【００１４】
図１を参照すると、ＧＴＳＰアプローチを使用してフレーズ－ベースのＳＭＴの実行を容易にするシステム１０が図解されている。このシステムは、ここで述べる多様な手法、方法、応用、アルゴリズム等を実行するためのコンピュータ実行可能命令を実行するプロセッサ１２およびそれらの命令を記憶するメモリ１３を含む。
【００１５】
メモリ１３は、コンピュータ上で実行できるコンピュータ・プログラム製品を包含し得る。コンピュータ・プログラム製品は、コントロール・プログラムが記録されたコンピュ

ータ可読記録媒体（たとえばメモリ１３）であってよく、例えば、ディスク、ハード・ドライブ、またはこれらと同種のものであってよい。一例によれば、ここで述べる手法は、１つまたは複数の汎用コンピュータ、専用コンピュータ（１つまたは複数）、プログラムされたマイクロプロセッサまたはマイクロコントローラおよび周辺集積回路素子、ＡＳＩＣまたはそのほかの集積回路、デジタル信号プロセッサ、ハードワイヤードの電子または論理回路、たとえばディスクリート素子回路、プログラマブル論理デバイス、たとえばＰＬＤ、ＰＬＡ、ＦＰＧＡ、グラフィック・カードＣＰＵ（ＧＰＵ）、またはＰＡＬ、またはこれらの類の上で実装することができる。

【００１６】

システム１０は、さらに、ユーザがシステムによる翻訳のための入力文１５を入力することができるユーザ・インターフェース１４を包含している。入力文１５は、コンピュータ実行可能アルゴリズム（１つまたは複数）を使用してプロセッサ１２によって処理され、オプションとして、翻訳済みデータ１７（たとえば、翻訳された文）がユーザ・インターフェース１４に出力されるまで、１つまたは複数の中間データ段階を通過する。

【００１７】

メモリ１３は、さらに、入力文１５を翻訳するための多様な構成要素（たとえば、コンピュータ実行可能インストラクションまたはその類）を記憶する。それに加えて、あらかじめ生成済みのバイ－フレーズがバイ－フレーズ・ライブラリ２０内に記憶される。

【００１８】

翻訳のための入力文を受け取ると、プロセッサ１２は、入力文と両立するバイ－フレーズをバイ－フレーズ・ライブラリ２０から検索するとともに、言語モデル１９にもアクセスし、検索したバイ－フレーズおよびその言語モデルを利用してＧＴＳＰグラフ２２を組み立てる。

【００１９】

ＧＴＳＰグラフ２２のノードを通る最適巡回を生成するために、プロセッサは、厳密ソルバ・アルゴリズム２４、近似ソルバ・アルゴリズム２５等々のうちの１つまたは複数であり得るＴＳＰソルバ２３を実行する。

【００２０】

システム１０は、ＧＴＳＰグラフ２２を入力文１５に適用することによって入力文のフレーズ－ベースの統計的機械翻訳（ｐｈｒａｓｅ　ｂａｓｅｄ　ｓｔａｔｉｓｔｉｃａｌ　ｍａｃｈｉｎｅ　ｔｒａｎｓｌａｔｉｏｎ，ＰＢＳＭＴ）を容易にする。プロセッサ１２は、ＧＴＳＰとしてＰＢＳＭＴタスクを定義する。プロセッサは、入力文と矛盾のない１つまたは複数のバイ－フレーズをバイ－フレーズ・ライブラリ２０から検索し、それぞれがバイ－フレーズに対応する複数のノードを包含するＧＴＳＰグラフ２２を生成する。ＴＳＰソルバ２３が実行され、ＧＴＳＰグラフ２２の最適巡回が生成される。

【００２１】

１つの実施態様においては、バイグラム言語モデルに代えてＮ－グラム言語モデル（２より大きいＮを用いる）を使用するとき、プロセッサは、最適巡回の真のコスト$C_t$、最適巡回の見かけのコスト$C_a$を計算し、真のコスト$C_t$と見かけのコスト$C_a$の間の差Ｄを決定する。プロセッサは、Ｄがあらかじめ決定済みの閾値εより小さいとき、ＧＴＳＰに対する解として最適巡回を出力し、出力されるＧＴＳＰ解を使用して入力文を第１の言語から第２の言語へ翻訳する。

【００２２】

プロセッサ１２は、Ｄがあらかじめ決定済みの閾値εに等しいか、またはそれを超える場合に、ＧＴＳＰグラフ２２内の少なくとも１つのノードを絞り込み、絞り込んだノードを包含する絞り込み後のグラフを生成する。プロセッサは、グラフの各絞り込みについて真のコスト$C_t$および見かけのコスト$C_a$の計算、それらの間の差Ｄの決定、および差Ｄとあらかじめ決定済みの閾値εの比較を、Ｄがεより小さくなるまで反復的に継続する。

【００２３】

ＴＳＰモデルは、次に述べるとおり、４つの主要な変形を含む。対称ＴＳＰ（ｓｙｍｍ

ｅｔｒｉｃ ＴＳＰ，ＳＴＳＰ）は、Ｎ個のノードについての無向グラフＧを伴い、このグラフは、エッジ（ライン）が実数値コストを持ち、＋∞（正の無限大）のコストが許容される。ＳＴＳＰ問題は、合計コストが最小となる「巡回」を見つけ出すことにあり、それにおいて巡回（ハミルトン閉路（Ｈａｍｉｌｔｏｎｉａｎ Ｃｉｒｃｕｉｔ）とも呼ばれる）は、グラフの各ノードを正確に一度だけ訪問するノード$X_1$，$X_2$，．．．，$X_N$，$X_1$の「循環」シーケンスであり、巡回の合計コストは、対応するエッジの寄与を加算することによって計算される。

【００２４】

ＡＴＳＰはＳＴＳＰの変形であり、基礎をなすグラフＧが有向であり、グラフの２つのノードａおよびｂについて、エッジ（ａ，ｂ）とエッジ（ｂ，ａ）とが異なるコストを持っていてよい。

【００２５】

一般対称ＴＳＰ（Ｇｅｎｅｒａｌｉｚｅｄ Ｓｙｍｍｅｔｒｉｃ ＴＳＰ）またはＳＧＴＳＰは、エッジが実数値コストを持つＮ個のノードの無向グラフＧを伴う。Ｎ個のノードを、Ｍ個の空でない共通の要素を持たないサブセット（クラスタと呼ばれる）へ分割することを考えると、目的は、各クラスタが正確に一度だけ訪問される最小合計コストを伴うＭ個のノード$X_1$，$X_2$，．．．，$X_M$，$X_1$の循環シーケンスを見つけることとなる。

【００２６】

一般非対称ＴＳＰ（Ｇｅｎｅｒａｌｉｚｅｄ Ａｓｙｍｍｅｔｒｉｃ ＴＳＰ）またはＧＴＳＰは、ＳＧＴＳＰに類似であるが、グラフＧが有向グラフである。理解されるものとするが、ここでＧＴＳＰを使用して説明する場合には、特に示さない限りは非対称ＧＴＳＰを意味する。

【００２７】

ＳＴＳＰは、しばしば単にＴＳＰと示され、ＮＰ困難であることが知られているが、それのための効率的な厳密および近似ソルバの開発には多大な関心が存在し、それにおいては「効率」が、いわゆるＴＳＰＬＩＢライブラリに提供されるような大規模ベンチマーク例を解決するために要する時間によって測定される。

【００２８】

ＡＴＳＰ、ＳＧＴＳＰ、およびＧＴＳＰは、すべて、ＳＴＳＰへの単純な（たとえば、問題のインスタンスのサイズにおける多項式または線形増加）変換によってマップが可能である。たとえば以下に述べるとおり、２つの「変換器」（たとえば、メモリ内に記憶され、プロセッサによって実行されるプログラム）が採用される。このうち１つはＧＴＳＰからＡＴＳＰへの変換のための変換器であり、もう１つはＡＴＳＰからＴＳＰへの変換のための変換器である。

【００２９】

引き続き図１を参照するが、図２に、それぞれの「コスト」に従って０～６がラベル付けされた複数のエッジを伴う第１の変換３０を図解する。たとえばアプリゲート（Ａｐｐｌｅｇａｔｅ）ほかの「Ｔｈｅ Ｔｒａｖｅｌｉｎｇ Ｓａｌｅｓｍａｎ Ｐｒｏｂｌｅｍ： Ａ Ｃｏｍｐｕｔａｔｉｏｎａｌ Ｓｔｕｄｙ」Ｐｒｉｎｃｅｔｏｎ ＵＰ、２００６年、ｐ．１２６を参照されたい。元の有向グラフ３２の各ノードＡは、変換された無向グラフ３４の３つのノードＡ、Ａ’、Ａ”によって置換され、２つの「０コスト」のエッジが追加される。無向グラフの任意の巡回に中間ノードＡ’が存在しなければならないことから、０コストのエッジも存在しなければならない。またこのことが、たとえば、最初にコスト２のエッジを通り、続いてコスト５のエッジ（および、同様に「逆」方向に対応することになるあらゆるエッジのペア）を通る巡回を除外する。これは、その後その巡回がノードＡに入射する３つのエッジを含む必要があり、それは不可能であるからである。したがって、無向グラフ３４の任意の最適巡回は、元のグラフ３２の最適巡回に対応する。

【００３０】

図３は、いくつかのエッジへの大きな人工的な重みの導入を対価として、２つのノードを導入するだけで元のグラフの１つのノードを置換する利点を有する第２の変換５０を図解している。元のグラフ５２の各ノードＸ（たとえば、ノードＡ、Ｂ、およびＣ）が、ノードＸおよびＸ’に複製され、大きな負の重み－ＫでＸとＸ’とを接続し、元のグラフ内の有向エッジ（Ｘ，Ｙ）のコストが変換後のグラフ５４内のエッジ（Ｘ’，Ｙ）上に再現される。Ｋが充分に大きい（たとえば、元のグラフ５２内のすべての有限のコストの合計よりＫが大きい）場合には、無向グラフ５４内の任意の最適巡回が、ほかのいずれの構成より（Ｘ，Ｘ’）エッジを通ることを選択する。これらの（Ｘ，Ｘ’）エッジのうち１つでも通らないことがあれば、少なくともＫ単位分のコストを失うことを意味するからである。しかしながらこれは、変換後のグラフ５４のノードの任意の最適巡回において、ＸとＸ’とが常に互いに隣り合い、ＸとＹとの間またはＸ’とＹ’との間にリンクが存在しないことから、その種の巡回だけが$X_1$，$X'_1$，$X_2$，$X'_2$，．．．，$X_N$，$X'_N$，$X_1$、または$X'_1$，$X_2$，$X'_2$，．．．，$X_N$，$X'_N$，$X_1$，$X'_1$という形式になることを意味し、これは、元のグラフ５２における「方向の変更」を禁止する制約条件に対応する。

【００３１】

図４は、ＧＴＳＰをＡＴＳＰに変換する変換７０を図解している。たとえば、ヌーン（Ｎｏｏｎ）ほかの「Ａｎ ｅｆｆｉｃｉｅｎｔ ｔｒａｎｓｆｏｒｍａｔｉｏｎ ｏｆ ｔｈｅ ｇｅｎｅｒａｌｉｚｅｄ ｔｒａｖｅｌｉｎｇ ｓａｌｅｓｍａｎ ｐｒｏｂｌｅｍ」ＩＮＦＯＲ３１（１９９３年）ｐ．３９～４４を参照されたい。この変換においては、元のグラフ７４内の$Y_1$，．．．，$Y_k$が与えられたクラスタ７２のノードであり、ＸおよびＺがほかのクラスタに属する任意のノードであると仮定する。変換後のグラフ７６においては、図に示されるとおりに循環を形成するために$Y_i$’の間にエッジ７８が導入され、各エッジは大きな負のコスト－Ｋを有する。Ｘから$Y_i$へ入るエッジはそのまま残され、$Y_i$からＺへ出るエッジがその起点を$Y_{i-1}$に変更されている。すると、Ｘ，$Y_i$，Ｚを通過する元のＧＴＳＰ問題における実行可能な巡回は、変換後のグラフ７６において最初にＸを通り、続いて$Y_i$，$Y_{i+1}$，．．．，$Y_k$，．．．，$Y_{i-1}$を通り、その後Ｚを通る巡回として「エンコード」される（このエンコードは元のコストから（ｋ－１）Ｋを減じたコストを有する）。それに加えてＫが充分に大きければ、変換後のＡＴＳＰグラフのためのソルバが、可能な限り多くの－Ｋエッジを通り抜ける傾向を有することになり、これは、正確にｋ－１個のその種のエッジ、たとえばそのクラスタに関連付けされた１つのエッジを除くすべてを通り抜けることを意味する（その種のエッジすべてを通り抜けるとグラフ全体のための巡回を見つけることができなくなるため、ソルバがそれを行うことはない）。言い替えると、ＧＴＳＰ問題のいくつかの実行可能な巡回のエンコーディングである巡回を作り出すことになる。

【００３２】

以下の例は、巡回セールスマン問題としてフレーズ－ベースのデコーディングを説明する。この例では、フランス語の文「ｃｅｔｔｅ ｔｒａｄｕｃｔｉｏｎ ａｕｔｏｍａｔｉｑｕｅ ｅｓｔ ｃｕｒｉｅｕｓｅ（この機械翻訳は奇妙である。）」が英語に翻訳される。この文を翻訳のための関連するバイ－フレーズを次の表１に示す。

【表１】

| バイーフレーズ識別子 | ソース | 目標 |
|---|---|---|
| h | *cette* | *this* |
| t | *traduction* | *translation* |
| ht | *cette traduction* | *this translation* |
| mt | *traduction automatique* | *machine translation* |
| a | *automatique* | *automatic* |
| m | *automatique* | *machine* |
| i | *est* | *is* |
| s | *curieuse* | *strange* |
| c | *curieuse* | *curious* |

【００３３】
このモデルにより、次の翻訳が生成される。
ｈ．ｍｔ．ｉ．ｓ → ｔｈｉｓ ｍａｃｈｉｎｅ ｔｒａｎｓｌａｔｉｏｎ ｉｓ ｓｔｒａｎｇｅ（この機械翻訳は奇妙である）
ｈ．ｃ．ｔ．ｉ．ａ → ｔｈｉｓ ｃｕｒｉｏｕｓ ｔｒａｎｓｌａｔｉｏｎ ｉｓ ａｕｔｏｍａｔｉｃ（この好奇心旺盛な翻訳は自動である）
ｈｔ．ｓ．ｉ．ａ → ｔｈｉｓ ｔｒａｎｓｌａｔｉｏｎ ｓｔｒａｎｇｅ ｉｓ ａｕｔｏｍａｔｉｃ（この翻訳奇妙は自動である）
…
上記では、各翻訳を導くバイーフレーズの順序付きシーケンスが、矢印の左側に示されている。そして、デコーディングは、ＧＴＳＰとして、グラフのノードがすべての可能ペア（ｗ，ｂ）を表す態様で公式化される。ここで、ｗはソース文ｓ内のソース単語であり、ｂはこのソース単語を含むバイーフレーズである。ここでは、同じ単語タイプでも出現が異なれば異なる単語と考える。特別なバイーフレーズ$b_{\$}$＝（＄，＄’）が導入され、ここで、＄（または＄’）はソース（または目標）文の開始を示す特別なソース単語となり、かつ、ペア（＄，$b_{\$}$）に関連付けられる、対応する追加のグラフ・ノード＄＄＝（＄，（＄，＄’））が導入される。
【００３４】
グラフ・クラスタは、共通のソース単語ｗを共有するグラフ・ノードのサブセットになり、ノード＄＄は、ソース単語＄に関連付けされたクラスタ内の唯一のノードになる。グラフのノードＭとＮとの間における遷移のコストは、次のように定義される。Ｍが（ｗ，ｂ）の形式であり、Ｎが（ｗ’，ｂ）の形式であり、ｂが単一のバイーフレーズで、ｗおよびｗ’がｂ内で連続する単語である場合、遷移コストは０である（遷移コストがない）。直観的に言えば、ｂの最初の単語の使用に一旦掛かり合えば、ｂによってカバーされるほかのソース単語に移動するための追加のコストはない。Ｍが（ｗ，ｂ）の形式であり、ｗがバイーフレーズｂ内の「一番右のソース単語」であり、Ｎが（ｗ’，ｂ’）の形式であり、ｗ’≠ｗがｂ’内の「一番左のソース単語」である場合、遷移コストは、バイーフレーズｂを選択した直後にバイーフレーズｂ’を選択する実際のコストに対応する。ソース文から見ると、これはｂのソース側を消費した後のｂ’のソース側の「消費」に対応し（ソース文内のそれらの相対的なポジションによらない）、目標側から見ると、これはｂの目標側を生成した直後におけるｂ’の目標側の生成に対応する。
【００３５】
遷移コストは、その場合、バイーフレーズ・ライブラリ２０（図１）内のｂに関連付け

された静的コストを含むいくつかの寄与の和になる。このコストは、順方向および逆方向の条件付き確率、バイーフレーズ内の目標単語の数、およびこれらの類（導入部分の説明を参照されたい）といった構成要素に対応する。「歪み」コストは、ソース単語ｗを消費した直後にソース単語ｗ’を消費する選択に関連付けされる。ｗ’がソース文内においてｗに直接続く単語である場合には、このコストがゼロであり、ｂおよびｂ’の目標側の連続性がそれらの目標によって与えられる状況に対応する。そのほかの場合には、ソース文内のｗおよびｗ’のポジションをｐｏｓ（ｗ）およびｐｏｓ（ｗ’）とするとき、（ｐｏｓ（ｗ）＋１－ｐｏｓ（ｗ’））の絶対値としてコストが計算される。「言語モデル」のコストは、ｂの目標単語をもたらしたばかりの文脈において目標単語ｂ’をもたらすコストである。バイグラム言語モデルが仮定される場合には、ｂおよびｂ’がわかるとすぐにこのコストをあらかじめ計算することが可能になる。というのも、これはｂがそれの目標側に少なくとも１つの単語を含み、そのことがｂの最後の目標単語を知った上でのｂ’の最初の目標単語の寄与の計算を可能にするからである。ｂ’の２番目、３番目等々の目標単語については、ｂ’のみを基礎として寄与が計算される。注意されたいが、このバイグラム・モデルの制限は、ここで論じられているほかの手法を使用して克服できる。

【００３６】

バイーフレーズｂおよびｂ’のうちの１つが＄＄に等しい場合においては、以前の寄与の簡単な適応を容易に実行することができる。ほかのすべての場合には遷移コストが無限となり、換言すればＭとＮの間のグラフ内にエッジが存在しない。

【００３７】

図５は、既存のエッジのサブセット、すなわちノードｔｒａｄｕｃｔｉｏｎ－ｍｔに入るか、または出るすべてのエッジ８２だけが示された、ソース文「ｃｅｔｔｅ　ｔｒａｄｕｃｔｉｏｎ　ａｕｔｏｍａｔｉｑｕｅ　ｅｓｔ　ｃｕｒｉｅｕｓｅ」についての遷移グラフ８０を図解している。注意を要するが、ｔｒａｄｕｃｔｉｏｎ－ｍｔの唯一の後続ノードはａｕｔｏｍａｔｉｑｕｅ－ｍｔであり、ｃｅｔｔｅ－ｈｔは、ｔｒａｄｕｃｔｉｏｎ－ｍｔの先行ノードでない。それに代えて、ａｕｔｏｍａｔｉｑｕｅ－ｍおよびａｕｔｏｍａｔｉｑｕｅ－ａからｔｒａｄｕｃｔｉｏｎ－ｍｔにエッジを引くことは可能であるが、その種のエッジは、実際のところ、ｔｒａｄｕｃｔｉｏｎ－ｍｔからの唯一の出口がａｕｔｏｍａｔｉｑｕｅ－ｍｔであり、このノードがそれのクラスタ内のほかのノードと排他であることから、横切ることができない。

【００３８】

図６Ａおよび図６Ｂは、示された２つの出力に対応する２つのＧＴＳＰ巡回を図解している。図６Ａにおいては、巡回９０が、結果として出力ｈ．ｍｔ．ｉ．ｓをもたらす。図６Ｂにおいては、巡回９２が、結果として出力ｈｔ．ｓ．ｉ．ａをもたらす。

【００３９】

上述の図面に関して述べたモデルは、一般ＴＳＰの非対称バージョンに対応する。この再公式化を前提とすると、追随可能ないくつかのストラテジが存在し、ＧＴＳＰ用に特別に設計されたアルゴリズムを使用してもよいし、ＧＴＳＰをＡＴＳＰに変換し、ＡＴＳＰ用に設計されたアルゴリズムを使用してもよいし、かつ／またはＡＴＳＰをＳＴＳＰに変換し、ＳＴＳＰ用に設計されたアルゴリズムを使用してもよい。各オプションは、それぞれ独自の利点および欠点を有する。コンコード（Ｃｏｎｃｏｒｄｅ）ソルバ（たとえば、ｗｗｗ．ｔｓｐ．ｇａｔｅｃｈ．ｅｄｕ／ｃｏｎｃｏｒｄｅ参照）等の既存の効率的なＴＳＰ用のソルバが使用される場合には、ＳＴＳＰ公式化が採用される。しかしながらＡＴＳＰがＳＴＳＰに変換される場合には、ＴＳＰグラフ内の頂点の数が２倍になる。さらにまたＧＴＳＰからＡＴＳＰへの経路は、より一般的な公式化が採用されることから非能率の潜在的原因である。したがって、コンコード・テクニックとともにＳＴＳＰ再公式化を使用することが望ましい。

【００４０】

別の重要な要因は、厳密な解が望ましいか、または近似解で充分とし得るか、ということである。ＳＴＳＰの場合においては、厳密な解法（たとえば、コンコード・ソルバ）を

採用すること、または近似アルゴリズム（たとえば、リン・カーニハン（Ｌｉｎ－Ｋｅｒｎｉｇｈａｎ）のヒューリスティック）を使用することができる。

【００４１】

言語モデルがバイグラム・タイプである場合、説明してきたモデルは重要な「マルコフの」性質、すなわち経路のコストは、その経路上の２つの連続するノードの間における遷移のコストに関する加法であるという性質を有する。図６Ａにおいて、翻訳候補「ｔｈｉｓ．ｍａｃｈｉｎｅ　ｔｒａｎｓｌａｔｉｏｎ．ｉｓ．ｓｔｒａｎｇｅ」のコストは、単語「ｉｓ」に関する単語「ｓｔｒａｎｇｅ」の条件付き確率を考慮に入れるだけでよく、単語「ｔｒａｎｓｌａｔｉｏｎ」および「ｉｓ」に関しては考慮しなくてよい。

【００４２】

別の実施態様においては、モデルの性能が、バイグラム言語モデルの使用から、３－グラム言語モデル等のより強力なｎ－グラム言語モデルの使用に拡張され、いくつかのアプローチを適用することができる。第１のアプローチは、少なくとも２つの単語の目標側を有するバイ－フレーズだけを保持するために、目標側が１つの単語だけを含むすべてのバイ－フレーズを「編集により除外（ｃｏｍｐｉｌｉｎｇ　ｏｕｔ）」することを包含する。この態様においては、２つのバイ－フレーズｂおよびｂ’の目標側が連結されるとき、ｂが少なくとも２つの単語を含むことから、トリグラム言語モデルが、ｂに関するｂ’の寄与の計算に充分な文脈を有する。適正な機能を保証するためにバイ－フレーズの概念の拡張を採用し、ここでバイ－フレーズの順序付きシーケンス

【数２】

$$[(\tilde{s}_1,\tilde{t}_1).(\tilde{s}_2,\tilde{t}_2). \ldots .(\tilde{s}_k,\tilde{t}_k)]$$

として拡張バイ－フレーズを定義する。ここで、ｋ≧１であり、各

【数３】

$$\tilde{s}_i$$

または、各

【数４】

$$\tilde{t}_i$$

は、ソース単語または目標単語のリストである。ｋ＝１の場合には、この手法はバイ－フレーズのオリジナルの概念に戻る。ソース文ｓの翻訳のための概念の解釈は、オリジナルの場合といくらか異なる。つまり、それぞれの個別の

【数５】

$$\tilde{s}_i$$

内のトークンは、ｓ内において連続的にマッチングされる必要があるけれども、

【数６】

$$\tilde{s}_1,\tilde{s}_2,\ldots,\tilde{s}_k$$

が連続的にマッチングされることは必要ないし、あるいは、ｓの内側におけるその順序でのマッチングさえ必要ない。これに対して目標側においては、

【数７】

$$\tilde{t}_1,\tilde{t}_2,...,\tilde{t}_k$$

内のトークンは、連続的に、かつその順序で作られる。この概念の下に、前述と同じ可能バイーフレーズの表を使用して、次に示す拡張バイーフレーズｍｔｉ、ｔｉ、およびｓｉが提供される。

【００４３】

ｍｔｉ＝［ｍｔ．ｉ］＝［（ｔｒａｄｕｃｔｉｏｎ ａｕｔｏｍａｔｉｑｕｅ，ｍａｃｈｉｎｅ ｔｒａｎｓｌａｔｉｏｎ）．（ｅｓｔ，ｉｓ）］

ｔｉ＝［ｔ．ｉ］＝［（ｔｒａｄｕｃｔｉｏｎ，ｔｒａｎｓｌａｔｉｏｎ）．（ｅｓｔ，ｉｓ）］

ｓｉ＝［ｓ．ｉ］＝［（ｃｕｒｉｅｕｓｅ，ｓｔｒａｎｇｅ）．（ｅｓｔ，ｉｓ）］

【００４４】

これらを使用して、次に示す翻訳を生成することができる。

［ｈ］．［ｍｔ．ｉ］．［ｓ］→ｔｈｉｓ ｍａｃｈｉｎｅ ｔｒａｎｓｌａｔｉｏｎ ｉｓ ｓｔｒａｎｇｅ

［ｈ］．［ｃ］．［ｔ．ｉ］．［ａ］→ｔｈｉｓ ｃｕｒｉｏｕｓ ｔｒａｎｓｌａｔｉｏｎ ｉｓ ａｕｔｏｍａｔｉｃ

［ｈｔ］．［ｓ．ｉ］．［ａ］→ｔｈｉｓ ｔｒａｎｓｌａｔｉｏｎ ｓｔｒａｎｇｅ ｉｓ ａｕｔｏｍａｔｉｃ

【００４５】

翻訳プロセスのオリジナルのアカウントと現在のそれの間の主要な相違は、拡張バイーフレーズが、ユニットのシーケンスを通じて以前に達成できたものを単一のユニットの下にカプセル化することである。ＧＴＳＰグラフとしての翻訳プロセスのエンコードについては、それが簡単になり、グラフのノードがこの場合はペア（ｗ，ｂ）であり、それにおいてｗはソース文の単語、ｂは拡張バイーフレーズ、クラスタは同一のｗを有するノードのサブセットである。オリジナルの規則に対する拡張によって、ｗが

【数８】

$$\tilde{s}_1$$

の最初の単語であり、形式

【数９】

$$(w, b) = (w, [(\tilde{s}_1,\tilde{t}_1).(\tilde{s}_2,\tilde{t}_2). ... .(\tilde{s}_k,\tilde{t}_k)])$$

であるノードに入るグラフ内の経路は、それが拡張バイーフレーズｂを「離れる」（この時点において実際の選択、すなわち次の拡張バイーフレーズの選択がある）前に、

【数１０】

$$\tilde{s}_1$$

のすべての単語を順番に通り、続いて

【数１１】

$$\tilde{s}_2$$

のすべての単語を通り、同様に繰り返して最後に

【数１２】

$$\tilde{s}_k$$

のすべての単語を通らなければならない。拡張バイ－フレーズ
【数１３】

$$[(\tilde{s}_1,\tilde{t}_1).(\tilde{s}_2,\tilde{t}_2). \ldots .(\tilde{s}_k,\tilde{t}_k)]$$

の「内側の」コストは、
【数１４】

$$(\tilde{s}_1,\tilde{t}_1)$$

から
【数１５】

$$(\tilde{s}_2,\tilde{t}_2)$$

等のように遷移するときに生じるであろうコスト（歪みコストを含む）を加算することによって事前編集できる。概して言えば、基本バイ－フレーズの構成要素にわたって対応する経路を考慮することにより生じるコストを「回収する」ことによって拡張バイ－フレーズにわたる経路のコストを計算することは簡単である。
【００４６】
バイグラム言語モデルからトリグラム言語モデルへの移動の問題に戻るが、バイ－フレーズ・ライブラリから単一単語の目標を有するバイ－フレーズｉを取り除くステップ、および拡張バイ－フレーズｍｔｉ、ｔｉ、ｓｉ等々（たとえば、ライブラリ内のバイ－フレーズのｉとの連結からなるすべての拡張バイ－フレーズ）をライブラリに追加するステップが実行される。これらの拡張バイ－フレーズは、すぐ次にもたらされる目標単語（与えられている例においては、それぞれ単語「ｓｔｒａｎｇｅ」、「ａｕｔｏｍａｔｉｃ」、および「ａｕｔｏｍａｔｉｃ」）についてのトリグラム確率を計算する充分な文脈を提供する。これらのステップが、ｉと類似の、すなわち単一単語の目標を有する（手元のソース文に適切な）すべてのバイ－フレーズについて網羅的に実行されると、各ポイントにおいてトリグラム言語モデルの計算を可能にする表現が得られる。
【００４７】
図７は、バイ－フレーズ「ｉ」だけが取り除かれた、現在は「ｉ」をカプセル化している拡張バイ－フレーズのいくつかだけがグラフを通る１つの有効な回路または巡回を定義するエッジ１０２を伴って示されているグラフ１００を図解している。気付かれるであろうが、ｍｔｉが充分に大きな目標文脈を提供することから、２つのノード（ｅｓｔ，ｍｔｉ）および（ｃｕｒｉｅｕｓｅ，ｓ）を接続するエッジが、ここでトリグラムのコストｐ（ｓｔｒａｎｇｅ｜ｔｒａｎｓｌａｔｉｏｎ　ｉｓ）に関連付けされる。
【００４８】
図８は、選択的なオンデマンド絞り込みを伴う第２のアプローチを図解している。今述べたばかりの網羅的な「編集により除外」する方法は、基本的に有効であるが、翻訳されるべき文についてｍ個の関連するバイ－フレーズがあり、そのうちのｋ個が単一単語の目標を有している場合には、ｋ．ｍ個の拡張バイ－フレーズが作られ、ｋがｍに関して大きくなれば直ちにＴＳＰソルバについての有意のオーバーヘッドを表すおそれがある。この効果は、編集により除外する方法がｎ＞３を伴うｎグラム言語モデルに拡張されると悪化することがある。
【００４９】

この効果を緩和するために、第２のアプローチは、２つの構成要素を有する選択的絞り込みを使用する。第１の構成要素は、何らかの広い評価基準（長さ１の目標側を有する等）に関してすべてのノードの文脈を絞り込むのではなく、グラフ内の選択されたノードの文脈を絞り込む能力である。その種の絞り込みは、グラフ内のノードの少数派のためのトリグラム文脈を提供するが、残りのノードについてはバイグラム文脈だけとなる。第２の構成要素は、その種の絞り込みのための最適ＴＳＰ解と、グラフ内のすべてのノードについてトリグラム文脈が使用された場合に到達する「真の」最適解との間の結合不等式の維持からなり、真の最適解への絞り込みプロセスの収斂を保証する。

【００５０】

したがって図８に、ＧＴＳＰグラフ１１０の選択的絞り込みを図解する。ＧＴＳＰグラフ１１０において、ａ、ｂ、およびｃはグラフ内の、異なるクラスタに属するノードであり、ノードｂに出入りするエッジ（ノードを接続するライン）のうちのいくつかが示されている。それに加えて、各エッジについてのコストまたは重みが、α、β、γ、δ、η、およびθとラベル付けされて示されている。変換後のＧＴＳＰグラフ１１２においては、ノードｂが、ｂと同じクラスタに属する（したがって、相互に排他的な）２つの「クローン」ノードｂ１およびｂ２に置き換えられ、ｂのすべての点に関してまったく等しいが、異なる入射エッジを有し、ａに関する入射エッジ（そのうちの１つだけが示されている）は変化しないが、新しくｃに到来するエッジが追加されている。ここではノードｂ１を「ａの直接の後続ノードであるという文脈におけるｂ」として解釈することが可能であり、ｂ２は「そのほかの任意の文脈におけるｂ」として解釈される。

【００５１】

この変換において注意する最初の性質は、コスト$\beta_1$および$\beta_2$がβに等しいと仮定された場合に、変換後のグラフ１１２が１つのノードおよび３つのエッジをオリジナルより多く有するが、注意深い観察によってわかるとおり、最適巡回は正確に同じであり、同一のコストを伴うことである。しかしながらここでは、ｂ１（またはｂ２）がａの直接の後続ノードである（または、直接の後続ノードでない）という文脈に特化され、したがってこれらの特化された文脈が、この追加の知識に関してコスト$\beta_1$および$\beta_2$をより良好に定義するために利用される。特に、この変換がＳＭＴの状況に適用されるとき、$\beta_1$はａに関連付けされた目標単語を承知しており、ｃの最初の目標単語を条件付けするためにトリグラム言語モデルを利用することを可能にする。

【００５２】

ＧＴＳＰグラフを前提として、グラフに関する良好に形成された巡回τが考慮される。ＧＴＳＰグラフのエッジに対して与えられる重みに従って、τが特定のコスト、すなわち見かけのコストを有する。何らかの外部測定に従って、同一の巡回τが異なるコスト、すなわち真のコストを実際に有することがある。この状況の例は、ＧＴＳＰエッジが言語モデルのためのバイグラム・コストを持つが、真のスコアはトリグラムの知識に従って計算されるべきであるときに生じる。より一般的に言えば、巡回の真のコストは、グラフのエッジに局所的な重みによって計算可能であるより、いくぶん局所的でない巡回の性質に依存し得る。

【００５３】

ＧＴＳＰグラフのエッジに関するコストは、グラフに関して任意の良好に形成された巡回τについて、巡回の見かけのコストが真のコストより小さいか、またはそれに等しい場合に限って「楽観的（ｏｐｔｉｍｉｓｔｉｃ）」と定義することができる。「楽観（ｏｐｔｉｍｉｓｍ）」の概念は、ツリー・サーチにおける許容可能なヒューリスティック（たとえばＡ＊）の概念と何らかの類似性を有し、その場合においては「現実性のある」楽観的エッジのコストが注目される。標準サーチ・ヒューリスティクスは、サーチ・ツリーの拡張における局所的な決定を得るために使用されるが、ここで述べているヒューリスティック手順は、問題グラフのより一層正確な明細の反復的な提供に焦点を当てつつ、ＴＳＰソルバの「注意」が焦点されるグラフの部分を強調し、続いて「グローバル」ＴＳＰソルバに現在の最良の見かけの解を見つけさせる。この概念を踏まえて、実行される一般的な

手順を、次に図９Ａおよび図９Ｂに関して説明する。

【００５４】

図９Ａは、巡回セールスマン問題としてフレーズ－ベースの統計的機械翻訳を実行するための方法を図解している。１２０においてＳＭＴがＧＴＳＰとして定義され、ＧＴＳＰグラフが生成される。１２２においては、ソース文に整合するバイ－フレーズが検索され、検索されたバイ－フレーズは、それぞれＧＴＳＰグラフ内のノードに対応する。各バイ－フレーズは、第１の言語のフレーズ（たとえば、入力された文の言語において、入力またはソース文内のフレーズに整合するフレーズ）および第２の言語のフレーズ（たとえば、第２のまたは目標言語に翻訳された入力フレーズ）を含む。１２４においてはＧＴＳＰが解かれる。バイ－フレーズが、ＧＴＳＰの解の関数として選択され、１２６において、選択されたバイ－フレーズの目標または第２の言語のフレーズが、ＧＴＳＰの解によって定義された順序で出力される。

【００５５】

１つの実施態様においては、ＧＴＳＰを解くことは、ＧＴＳＰをＡＴＳＰに変換すること、ＡＴＳＰを標準ＴＳＰに変換すること、およびＴＳＰを解決して入力文のブロックを翻訳することを含む。ＴＳＰの解決は、コンコード・ソルバまたはリン・カーニハンのヒューリスティックまたはこれらの類を使用して実行される。

【００５６】

図９Ｂは、巡回セールスマン問題としてフレーズ－ベースの統計的機械翻訳を実行するための代替または追加の方法を図解している。１３０において、それの巡回の真のコストに関して楽観的なＧＴＳＰグラフ$Ｇ_0$の初期仕様が、ｉ＝０となるように初期化される。グラフ内のノードは、セグメント化された文のブロックに関連付けされたバイ－フレーズを定義する。１３２においては、ＧＴＳＰソルバ・アプリケーション（たとえば、図１のＴＳＰソルバ２３）が起動され、このグラフに関する最適巡回$τ_i$（または、近似ソルバが使用される場合にはその種の最適巡回の近似）が獲得される。１３４においては、$τ_i$の真のコストが（たとえば、$τ_i$のすべてのエッジが既知であることから）計算される。$Ｇ_i$が楽観的であることから、$Ｇ_i$に関する真のコストは、$τ_i$の見かけのコスト$Ｃ_a$より大きくなる。１３６においては、見かけのコストと真のコストの間の差Ｄがあらかじめ定義済みの閾値εより小さいか否かに関しての判定が行われる。

【００５７】

これら２つのコストの間の差Ｄが特定の閾値εより小さい場合には解$τ_i$が出力され、１３８においてこの方法が終了する。そうでなければ１４０において$Ｇ_i$の少なくとも１つのノード、特に、$τ_i$上に（可能性としては、ほかのいくつかにも）現れている特定のノードが、図７の原理に従って絞り込まれる。この種の絞り込みの間はグラフ$Ｇ_i$が楽観的にとどまるが、より厳しい値が、βによって提供された$β_1$および$β_2$のために提供される。すなわち、$β_1$＞β等の制約が提供され、かつ可能性として$β_2$＞β等の制約も提供される。１４２においては、絞り込みの結果として新しいグラフ$Ｇ_{i+1}$が獲得される。方法は１３４に戻るが、ｉ：＝ｉ＋１を伴う。

【００５８】

この図９Ｂの方法は、いくつかの重要な性質を有する。たとえば、任意の反復において、$τ_i$の見かけのコストは、元のグラフにおける真の最適巡回$τ_{true}$の新のコストの下限になる。真の最適巡回$τ_{true}$は、すなわち、すべての巡回の真のコストにわたって真のコストが最小となる巡回である。厳密ＴＳＰソルバの場合においては、ｔｒｕｅ＿ｃｏｓｔ（τ）≧ｔｒｕｅ＿ｃｏｓｔ（$τ_{true}$）であり、すべての巡回τについて（$τ_{true}$の定義により）、ｔｒｕｅ＿ｃｏｓｔ（$τ_{true}$）≧ａｐｐａｒｅｎｔ＿ｃｏｓｔ（$τ_i$）である。なぜなら、$τ_i$は、すべての巡回のコストの楽観的仕様より最適であり、特にｔｒｕｅ＿ｃｏｓｔ（$τ_{true}$）≧ａｐｐａｒｅｎｔ＿ｃｏｓｔ（$τ_{true}$）≧ａｐｐａｒｅｎｔ＿ｃｏｓｔ（$τ_i$）であるからである。

【００５９】

$Ｇ_i$および$τ_i$においてアルゴリズムが終了するときは、ａｐｐａｒｅｎｔ＿ｃｏｓｔ

（$\tau_i$）＋ε≧ｔｒｕｅ＿ｃｏｓｔ（$\tau_i$）である。したがって、ａｐｐａｒｅｎｔ＿ｃｏｓｔ（$\tau_i$）＋ε≧ｔｒｕｅ＿ｃｏｓｔ（$\tau_i$）≧ｔｒｕｅ＿ｃｏｓｔ（$\tau_{true}$）≧ａｐｐａｒｅｎｔ＿ｃｏｓｔ（$\tau_i$）である。言い替えると、反復の間に見つけられた巡回$\tau_i$は、真の最適巡回のそれと無視できる程度に異なる真のコストを伴った真の最適巡回の近似である。

【００６０】

アルゴリズムの終了特性については、有限数の絞り込みしか存在しないためにグラフを無限に絞り込むことが不可能であること、および与えられた巡回におけるノードが充分に絞り込まれるとき、巡回の見かけのコスト（$\beta_1$および$\beta_2$等の絞り込まれた重みに依存する）がそれの真のコストに等しくなることという２つの要因に依存する。

【００６１】

図９Ａおよび図９Ｂに図解されている方法は、ここに述べられているほかの手法またはアルゴリズムに加えて、コンピュータ上において実行できるコンピュータ・プログラム製品として実装できる。コンピュータ・プログラム製品は、ディスク、ハード・ドライブ、またはこれらの類といった、コントロール・プログラムが記録されたコンピュータ可読記録媒体（たとえばメモリ１３）とすることができる。一般的な形式のコンピュータ可読媒体は、たとえば、フロッピー（登録商標）ディスク、フレキシブル・ディスク、ハードディスク、磁気テープまたは任意のそのほかの磁気記憶媒体、ＣＤ－ＲＯＭ、ＤＶＤまたは任意のそのほかの光学媒体、ＲＡＭ、ＰＲＯＭ、ＥＰＲＯＭ、ＦＬＡＳＨ－ＥＰＲＯＭまたはそのほかのメモリ・チップまたはカートリッジまたは任意のそのほかの、コンピュータによる読み出しおよび使用が可能な有体の媒体を含む。それに代えてこの方法を、たとえば無線波および赤外線データ通信の間に生成されるような音響または光の波等の送信媒体およびこれらの類を使用するデータ信号としてコントロール・プログラムが埋め込まれる送信可能な搬送波内において実装することができる。

【００６２】

例示的な方法は、１つまたは複数の汎用コンピュータ、専用コンピュータ（１つまたは複数）、プログラムされたマイクロプロセッサまたはマイクロコントローラおよび周辺集積回路要素、ＡＳＩＣまたはそのほかの集積回路、デジタル信号プロセッサ、ハードワイヤード電子または論理回路、たとえばディスクリート素子回路、ＰＬＤ、ＰＬＡ、ＦＰＧＡ、グラフィック・カードＣＰＵ（ＧＰＵ）、またはＰＡＬといったプログラマブル論理デバイス、またはこれらの類の上において（たとえば、プロセッサ１２により）実装できる。概して言えば、有限状態マシンの実装が可能であり、続いてそれが図９Ａおよび図９Ｂに示されたフローチャートを実装できる任意のデバイスを、ＧＴＳＰモデルを使用するフレーズ－ベースのＳＭＴを実行するための方法の実装に使用することが可能である。

【００６３】

図１０は、トリグラム言語モデルを使用してフレーズ－ベースの翻訳を実行するためのグラフ１５０を図解している。認識されることになろうが、同じアプローチを４－グラム、５－グラム等々に簡単に適用することができる。トリグラム言語モデルは、単語ｚが２つの単語ｘおよびｙに続く確率ｐ（ｚ｜ｘｙ）の評価を提供する手順を容易にする。１つの実施態様においては、言語モデルがすべての３成分要素（ｘ，ｙ，ｚ）を、それらの確率とともに内部的に記憶する。別の実施態様においては、言語モデルが、明示的に特定のトリグラム、バイグラム、およびユニグラムのためのコーパス・カウントを記憶し、それらのテーブルからｐ（ｚ｜ｘｙ）の計算のためにスムージング手法を頼る。

【００６４】

このアプローチは次のとおりとなる。ｐ（ｚ｜ｘｙ）を考慮して目標言語モデルについてのグラウンド・トゥルースが提供される一方、初期グラフのすべてのエッジにトリグラム・コスト

【数１６】

$$\mathrm{tri}(z|xy) \triangleq -\log p(z|xy)$$

がラベル付けされる必要はない。しかし、むしろいくつかのエッジ（ノードが目標側に１つの単語だけを有するエッジ（ａ，ｂ））に、次のとおりに定義される「バイグラム・プロクシ」ｂｉ（ｚ｜ｙ）をラベル付けすることができる。

【数１７】

$$\mathrm{bi}(z|y) \triangleq \min_x -\log p(z|xy)$$

言い替えると、プロクシｂｉ（ｚ｜ｙ）は、ｙに先行し得る単語ｘに関して最大限に楽観的なｙとｚの間における遷移のコストのための評価である。ここで最小（ｍｉｎ）は、翻訳されるべき特定の文に関係するｘに関して求められる。たとえば、その文を翻訳するためのバイ－フレーズの最後の目標単語と同一のｘに関して求められ、語彙内のすべての可能な単語ｘに関して求められることはない。注意を要するが、ｐ（ｚ｜ｙ）がｐ（ｚ｜ｘｙ）から導かれたバイグラム言語モデルを表すとき、ｂｉ（ｚ｜ｙ）＝－ｌｏｇ　ｐ（ｚ｜ｙ）は概して真にならず、したがってｂｉ（ｚ｜ｙ）は、その用語の通常の意味におけるバイグラム確率を表さない。

【００６５】

したがって、図１０のグラフ１５０は、図５とまったく同じＧＴＳＰグラフを伴って開始することによって獲得され、それにおいてエッジ上の言語モデルのコストは、それらのエッジ上で得られる文脈を前提として可能な限り特定的であり、言い替えるとそれらは、先行するバイ－フレーズに応じてｔｒｉ（ｚ｜ｘｙ）の形式またはｂｉ（ｚ｜ｙ）の形式のいずれかとなる。特定の巡回のためのすべてのエッジが示され、それに加えて説明の焦点が当てられるノードｅｓｔ－ｉ上の入射エッジが略式に（破線で）示されたグラフ１５０によって一例を示す。

【００６６】

ＴＳＰソルバが起動され、特定のバイ・コストだけでなくいくつかのトリ・コストも使用して見かけの最適巡回が獲得される。グラフ１５０内の巡回の真のコストが、すべての真のトリグラム・コストｔｒｉ（ｔｈｉｓ｜＄　＄）、ｔｒｉ（ｍａｃｈｉｎｅ｜＄　ｔｈｉｓ）、ｔｒｉ（ｔｒａｎｓｌａｔｉｏｎ｜ｔｈｉｓ　ｍａｃｈｉｎｅ）、ｔｒｉ（ｉｓ｜ｍａｃｈｉｎｅ　ｔｒａｎｓｌａｔｉｏｎ）、ｔｒｉ（ｓｔｒａｎｇｅ｜ｔｒａｎｓｌａｔｉｏｎ　ｉｓ）、およびｔｒｉ（＄｜ｉｓ　ｓｔｒａｎｇｅ）の計算を伴って計算される。真のコストは（より小さい）見かけのコストと比較され、その差がεより小さければ、それ以上の動作がとられる必要はない。

【００６７】

図１１は、図７および図９に関して述べた手順を使用し、差がεより大きいか、または等しい場合に得られる、識別された巡回上のノードの少なくとも１つの３成分要素（ａ，ｂ，ｃ）に適用されるグラフ１６０を示している。たとえば、図７の手順を、（ａ，ｂ，ｃ）＝（ａｕｔｏｍａｔｉｑｕｅ－ｍｔ，ｅｓｔ－ｉ，ｃｕｒｉｅｕｓｅ－ｓ）とともに適用してグラフ１６０を得ることが可能であり、それにおいてｃｕｒｉｅｕｓｅ－ｓにｅｓｔ－ｉ－１をリンクしているエッジ上には、そのエッジのトリ・コストの計算に充分な文脈が存在する。注意を要するが、ｂｉ（ｓｔｒａｎｇｅ｜ｉｓ）は、ｅｓｔ－ｉ－２をｃｕｒｉｅｕｓｅ－ｓにリンクしている破線のエッジ上に保持されるが、ａｕｔｏｍａｔｉｑｕｅ－ｍｔを除外し、ｅｓｔ－ｉ－２に先行できるノードにわたって最小化することによるコストの再計算によって、わずかにより厳しい結合を得ることができる。

【００６８】

グラフ１６０が獲得されると、ＴＳＰソルバが、再起動されて手順が反復的に実行される。新しい見かけの最適巡回がエッジ（ｅｓｔ－ｉ－１，ｃｕｒｉｅｕｓｅ－ｓ）を含む

ときは常に、このエッジのコストが以前のものより正確になる。単に、各反復時に見かけの最適巡回において１つの３成分要素だけを絞り込むことからなるアプローチは、いずれかのポイントで終了する。そうでなければ、いずれかのポイントにおいて、見かけの最適巡回において、必然的に、すべてのエッジがトリ・コストを持つことになり、したがってその見かけのコストが真のコストに等しくなり、そしてε閾値の評価基準を満たすことになるためである。

【００６９】

これは、アルゴリズムの収斂の形式的な証明を提供するが、変換されたグラフがコストの真の状態をより迅速に「模する」ためには、各反復において、単一の３成分要素より多くを絞り込むほうがより効率的なことがある。認識されるであろうが、その種のアプローチのすべての可能な変形および／または組み合わせは、この説明の範囲内となることが意図されている。しかしながら、単純な方法は、現在の見かけの最適巡回上に現れるすべての３成分要素（ａ，ｂ，ｃ）を絞り込むことである（ソース文の長さがｎであれば、多くともｎのその種の絞り込みを行うことが可能である）。上述した編集により除外する手法は、すべてのトリグラムの網羅的な絞り込みに対応し、したがって選択的な絞り込み手法で行うことが可能な１つの極端な場合であることに注意を要する。

【００７０】

この態様においては、ＳＭＴの状況で起こりがちなように、新しい見かけの巡回がいくつかの下位経路を以前の見かけの巡回と共有する場合に、それらの経路に関する絞り込み済みの知識を利用することになる。注意されたいが、たとえ以前の巡回においてε閾値条件が満たされていない場合であっても、新しい巡回が実際に以前の巡回とまったく同じであることが可能であり、これは言語モデルのコスト以外のコストが、この経路上におけるバイ・コストからトリ・コストへの移動に関連付けされる補償より大きくなり得ることによる。

【００７１】

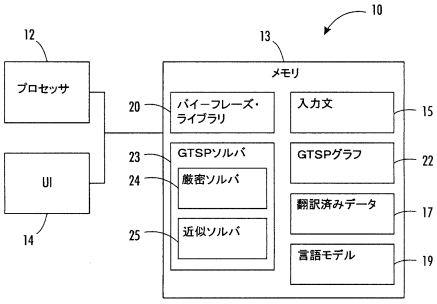
以上の説明はトリグラムの扱いに焦点を当てたが、選択的絞り込みのアプローチがｎグラム（ｎは整数）に拡張できることは容易に理解される。このアプローチは、トリグラム、４－グラム等々のための拡張された状況の提供に採用可能であり、方法は、巡回のいくつかの部分の言語モデルのコストを絞り込むことが、ほかのすべてのＳＭＴ制約を前提に巡回の最適性を先細りにしない限り効果的である。

【符号の説明】

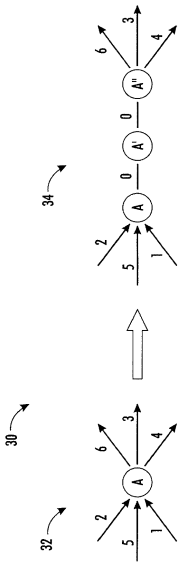
【００７２】

１０　システム、１２　プロセッサ、１３　メモリ、１４　ユーザ・インターフェース、１５　入力文、１７　翻訳済みデータ、１９　言語モデル、２０　バイ－フレーズ・ライブラリ、２２　ＧＴＳＰグラフ、２３　ＧＴＳＰソルバ、２４　厳密ソルバ・アルゴリズム、２５　近似ソルバ・アルゴリズム。

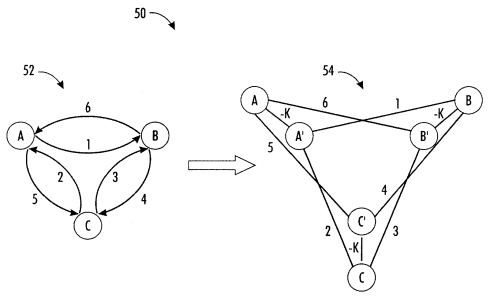
【図１】


10
12
13
メモリ
プロセッサ
UI
14
20
バイーフレーズ・ライブラリ
入力文
15
23
GTSPソルバ
GTSPグラフ
22
24
厳密ソルバ
翻訳済みデータ
17
25
近似ソルバ
言語モデル
19

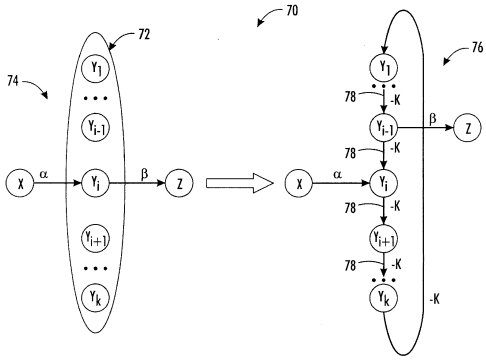
【図２】


34
A''
A'
A
0
0
6
3
4
2
5
1
30
32

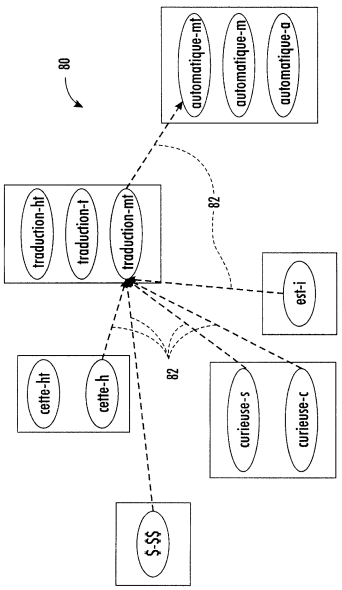
【図３】


50
52
54
A
B
C
A'
B'
C'
6
1
2
3
4
5
-K

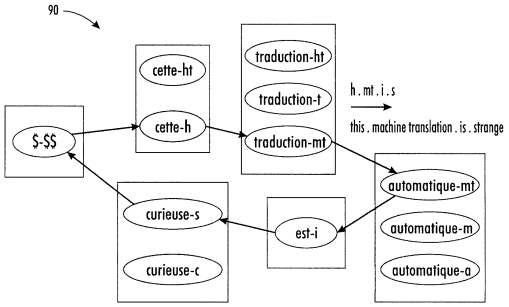
【図４】


70
72
74
76
78
X
Z
α
β
-K

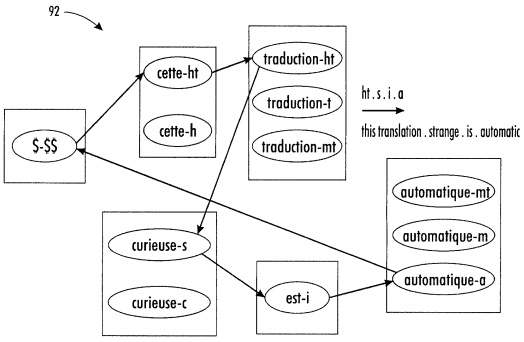
【図５】


80
automatique-mf
automatique-m
automatique-g
traduction-ht
traduction-t
traduction-mt
est-i
cette-ht
cette-h
curieuse-s
curieuse-c
$-$$
82

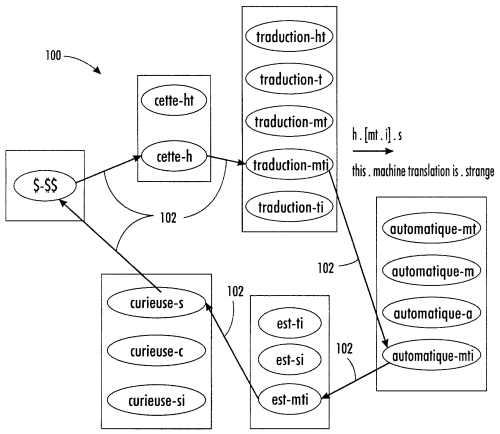
【図６Ａ】



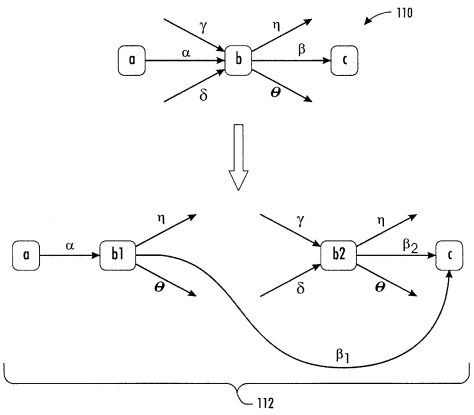

【図６Ｂ】



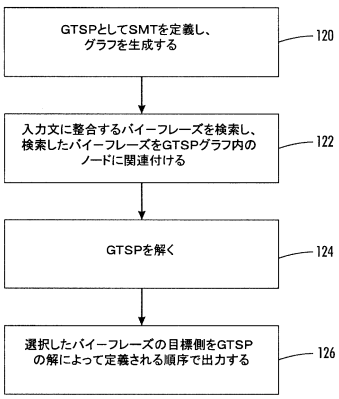

【図７】



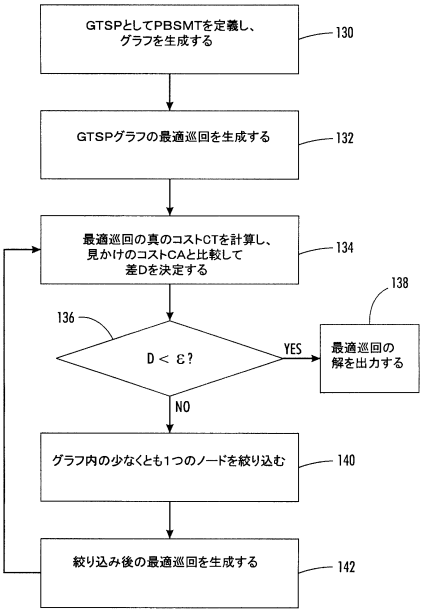

【図８】



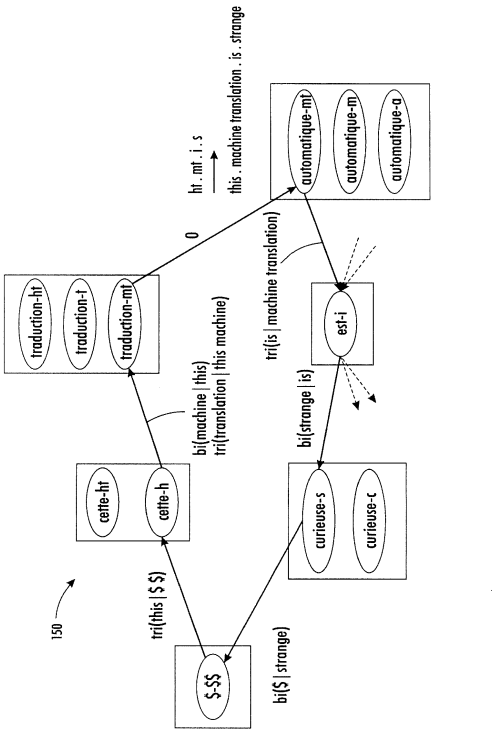

【図９Ａ】



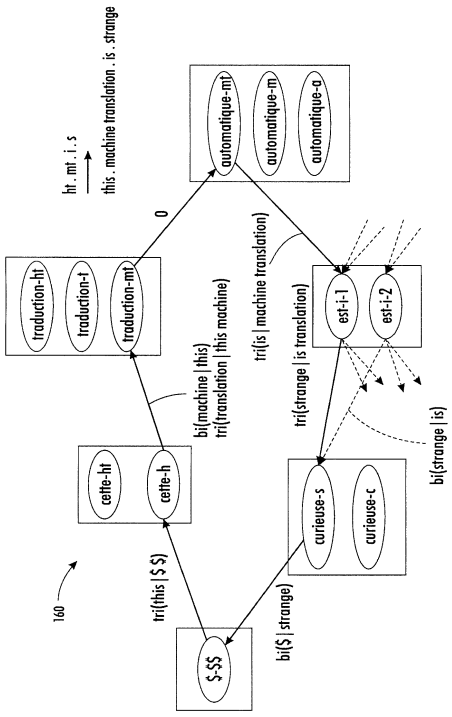
【図９Ｂ】


GTSPとしてPBSMTを定義し、
グラフを生成する
130
GTSPグラフの最適巡回を生成する
132
最適巡回の真のコストCTを計算し、
見かけのコストCAと比較して
差Dを決定する
134
136
D < ε?
YES
138
最適巡回の
解を出力する
NO
グラフ内の少なくとも1つのノードを絞り込む
140
絞り込み後の最適巡回を生成する
142

【図１０】


ht . mt . i . s
this . machine translation . is . strange
automatique-mt
automatique-m
automatique-a
0
traduction-ht
traduction-t
traduction-mt
tri(is | machine translation)
esti-i
bi(machine | this)
tri(translation | this machine)
bi(strange | is)
cette-ht
cette-h
curieuse-s
curieuse-r
tri(this | $ $)
150
$-$$
bi($ | strange)

【図１１】


ht . mt . i . s
this . machine translation . is . strange
automatique-mt
automatique-m
automatique-a
0
traduction-ht
traduction-t
traduction-mt
tri(is | machine translation)
esti-i-1
esti-i-2
bi(machine | this)
tri(translation | this machine)
tri(strange | is translation)
bi(strange | is)
cette-ht
cette-h
curieuse-s
curieuse-r
160
tri(this | $ $)
bi($ | strange)
$-$$

フロントページの続き

(72)発明者　ニコラ　カンチェッダ
フランス　グルノーブル　クール　ジャン　ジョレス　３１

審査官　長　由紀子

(56)参考文献　国際公開第０２／０９７６６３（WO，Ａ１）
特開２００７－３２８４８３（ＪＰ，Ａ）
特開２００４－１０２９４６（ＪＰ，Ａ）
Yookyung Kim　etal., Sehda S2MT :Incorporation of Syntax into Statistical Translation System, International Workshop on Spoken Language Traslation (IWSLT) 2005, 米国，２００５年　8月２５日, p.153-158, ＵＲＬ, http://20.210-193-52.unknown.qala.com.sg/archive/iwslt_05/papers/slt5_153.pdf
本橋　瞬　外2名，Ｌｉｎ－Ｋｅｒｎｉｇｈａｎアルゴリズムをカオス駆動する巡回セールスマン問題の解法，電子情報通信学会技術研究報告，日本，社団法人電子情報通信学会，２００８年　3月２１日，第107巻　第561号，P.43～48
越川　満　外４名，統計的機械翻訳におけるフレーズ対応最適化を用いた翻訳候補のリランキング，言語処理学会第１５回年次大会発表論文集，日本，言語処理学会，２００９年　3月　2日，p.861-864

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