(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2004-362249
(P2004-362249A)
(43)公開日　平成16年12月24日(2004.12.24)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/28	G06F 17/28	U	5B091

審査請求 有　請求項の数 17　ＯＬ　（全 29 頁）

(21)出願番号　特願2003-159662(P2003-159662)
(22)出願日　平成15年6月4日(2003.6.4)

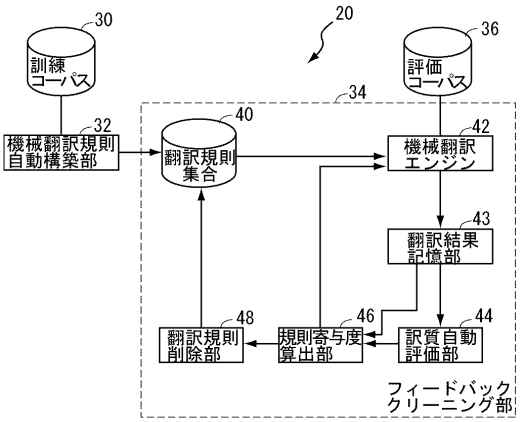
（出願人による申告）国等の委託研究の成果に係る特許出願（平成１５年度通信・放送機構、研究テーマ「大規模コーパスベース音声対話翻訳技術の研究開発」に関する委託研究、産業活力再生特別措置法第３０条の適用を受けるもの）

(71)出願人　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(74)代理人　100099933
弁理士　清水　敏
(72)発明者　今村　賢治
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　隅田　英一郎
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内

Ｆターム(参考)　5B091 BA14 CA21 CC03 CC05 CD11 EA01

(54)【発明の名称】翻訳知識最適化装置、翻訳知識最適化のためのコンピュータプログラム、コンピュータ及び記憶媒体

(57)【要約】
【課題】対訳コーパスから自動獲得された翻訳規則をクリーニングしてより翻訳品質を向上させる事ができる翻訳知識最適化装置を提供する。
【解決手段】翻訳知識最適化装置３４は、翻訳知識を記憶する翻訳規則集合記憶部４０と、評価コーパス３６と、翻訳規則集合記憶部４０に記憶された翻訳知識を利用して、評価コーパス３６中の原言語の文を目的言語に翻訳する機械翻訳エンジン４２と、機械翻訳エンジン４２による翻訳結果の品質を、評価コーパス３６を参照して自動的に評価する訳質自動評価部４４と、訳質自動評価部４４の出力する評価値が極値をとる様に、翻訳規則集合記憶部４０内の翻訳知識の集合の最適化を行なう規則寄与度算出部４６及び翻訳規則削除部４８とを含む。
【選択図】　図１



【特許請求の範囲】
【請求項１】
機械翻訳のための翻訳知識を最適化するための翻訳知識最適化装置であって、
翻訳知識の集合を記憶するための翻訳知識記憶手段と、
原言語と目的言語との対訳文を複数個含む、機械読取可能な対訳コーパスを記憶するための手段と、
前記翻訳知識記憶手段に記憶された前記翻訳知識の集合を利用して、前記対訳コーパス中の前記原言語の文を前記目的言語に機械翻訳するための機械翻訳手段と、
前記機械翻訳手段による翻訳結果の品質を、前記対訳コーパスを参照して自動的に評価して評価値を出力するための訳質自動評価手段と、
前記訳質自動評価手段の出力する評価値が極値をとる様に、前記翻訳知識の集合の最適化を行なうための最適化手段とを含む、翻訳知識最適化装置。
【請求項２】
前記翻訳知識は、前記原言語の構文パターンから前記目的言語の構文パターンへの構文変換規則を含む、請求項１に記載の翻訳知識最適化装置。
【請求項３】
前記最適化手段は、
前記翻訳知識の集合に含まれる翻訳知識の各々について、その規則寄与度を算出するための手段と、
前記規則寄与度が予め定める条件を満足する翻訳知識を、前記翻訳知識の集合から削除するための手段とを含む、請求項１に記載の翻訳知識最適化装置。
【請求項４】
前記規則寄与度を算出するための手段は、
前記翻訳知識の集合の全体を用いて、前記機械翻訳手段による翻訳、及び当該翻訳の結果の前記訳質自動評価手段による訳質評価を行ない、初期評価値を得るための手段と、
前記翻訳知識の集合中の翻訳知識ごとに、前記翻訳知識の集合から当該翻訳知識を削除して得られる部分集合を用いて、前記機械翻訳による翻訳、及びその翻訳結果の前記訳質自動評価手段による訳質評価を行ない、削除後評価値を得るための手段と、
前記削除後評価値と前記初期評価値との差分を、前記ある翻訳知識の前記規則寄与度として算出するための手段とを含む、請求項３に記載の翻訳知識最適化装置。
【請求項５】
前記最適化手段は、
前記翻訳知識の集合の全体を用いて、前記機械翻訳手段による翻訳、及び当該翻訳の結果の前記訳質自動評価手段による訳質評価を行ない、初期評価値を得るための手段と、
予め定められた方法に従って、前記翻訳知識の集合から複数の部分集合を作成するための手段と、
前記複数の部分集合の各々を用いて前記機械翻訳手段による翻訳、及びその翻訳結果の前記訳質自動評価手段による訳質評価を行ない、その評価値が前記初期評価値に対し所定の条件を満足するか否かを判定するための判定手段と、
前記判定するための手段により前記評価値が前記所定の条件を満足すると判定された部分集合の各々について、その補集合に属する翻訳知識を前記翻訳知識の集合から削除するための手段とを含む、請求項１に記載の翻訳知識最適化装置。
【請求項６】
前記部分集合を作成するための手段は、前記翻訳知識の集合から予め定められる数の翻訳知識を除いて得られる部分集合を複数個作成するための手段を含む、請求項５に記載の翻訳知識最適化装置。
【請求項７】
前記部分集合を複数個作成するための手段は、前記翻訳知識の集合から一つの翻訳知識を除いて得られる部分集合を複数個作成するための手段を含む、請求項６に記載の翻訳知識最適化装置。

【請求項８】
前記部分集合を作成するための手段は、前記翻訳知識の集合から予め定められる数の翻訳知識を除いて得る事が可能な全ての部分集合を作成するための手段を含む、請求項５に記載の翻訳知識最適化装置。
【請求項９】
前記機械翻訳手段は、原言語の文を機械翻訳する際に、前記翻訳知識の集合内のどの翻訳知識を使用したかについての情報を出力する機能を持ち、
前記翻訳知識最適化装置はさらに、前記初期評価値を得る際に翻訳された文ごとに、前記機械翻訳手段から出力される、翻訳の際に使用した翻訳規則を特定する情報を記憶するための手段を含み、
前記判定手段は、
前記記憶するための手段に記憶されている、前記翻訳規則を特定する情報を参照して、前記複数の部分集合の各々について、当該部分集合の補集合に含まれる翻訳規則を用いて翻訳された前記原言語の文の集合を特定するための手段と、
前記部分集合の各々を用いて、当該部分集合の補集合に含まれる翻訳規則を用いて翻訳された前記原言語の文の集合を前記機械翻訳手段により再び機械翻訳するための手段と、
前記部分集合の各々に対し、前記初期翻訳結果のうち、当該部分集合の補集合に含まれる翻訳規則を用いて翻訳された翻訳結果を、前記再び機械翻訳するための手段による翻訳結果で置換え、当該置換え後の初期翻訳結果に対して前記訳質自動評価手段による訳質評価を行なって、当該部分集合による翻訳結果の評価値を得るための手段と、
前記部分集合の各々に対し、当該部分集合による翻訳結果の評価値が前記初期評価値に対し前記所定の条件を満足しているか否かを判定するための手段とを含む、請求項５に記載の翻訳知識最適化装置。
【請求項１０】
前記判定するための手段は、前記部分集合の各々に対し、当該部分集合による翻訳結果の評価値が、前記初期評価値を上回っているか否かを判定するための手段を含む、請求項９に記載の翻訳知識最適化装置。
【請求項１１】
予め準備された、前記原言語と前記目的言語との対訳文からなる訓練コーパスから、各々が訓練サブコーパス及び評価サブコーパスを含む複数個のサブコーパス対を作成するための手段と、
予め定められる翻訳規則の構築方式に従って、与えられる対訳コーパスから翻訳規則を自動的に構築するための翻訳知識自動構築手段と、
前記翻訳知識自動構築手段を用いて前記訓練コーパスから翻訳知識を自動構築し、基本翻訳知識として記憶するための基本翻訳知識記憶手段と、
前記複数個のサブコーパス対の各々に対して、前記訓練サブコーパスから前記翻訳知識自動構築手段を用いて翻訳知識の集合を自動構築し、当該翻訳知識の集合に対し、前記評価サブコーパスを前記機械読取可能な対訳コーパスとして、前記翻訳知識記憶手段、前記機械読取可能な対訳コーパスを記憶するための手段、前記機械翻訳手段、前記訳質自動評価手段、及び前記最適化手段による最適化を行なうための手段と、
前記最適化を行なうための手段によって最適化された、前記複数個のサブコーパス対の各々に対して得られる翻訳知識の集合を、一つの翻訳知識の集合に集約するための手段とをさらに含む、請求項１に記載の翻訳知識最適化装置。
【請求項１２】
前記集約するための手段は、
前記基本翻訳知識記憶手段に記憶された前記基本翻訳知識に含まれる翻訳知識の各々について、前記最適化手段により算出された差分を、前記複数個のサブコーパス対の全てにわたって合計するための差分合計手段と、
前記差分合計手段により合計された差分が所定の条件を満足する翻訳知識を削除する様に前記基本翻訳知識記憶手段に記憶されている前記基本翻訳知識を更新するための手段とを

含む、請求項１１に記載の翻訳知識最適化装置。
【請求項１３】
前記基本翻訳知識を更新するための手段は、前記差分合計手段により合計された差分が負となる翻訳知識を削除する様に前記基本翻訳知識記憶手段に記憶されている前記基本翻訳知識を更新するための手段を含む、請求項１２に記載の翻訳知識最適化装置。
【請求項１４】
前記複数個のサブコーパス対を作成するための手段は、
前記訓練コーパスを予め定める個数に実質的に等分して前記予め定める個数の評価サブコーパスを作成するための手段と、
前記予め定める個数の評価サブコーパスの各々に対して、前記訓練コーパスから当該評価サブコーパスを除いたコーパスを作成し、当該評価サブコーパスと対となる訓練サブコーパスを作成するための手段とを含む、請求項１１に記載の翻訳知識最適化装置。
【請求項１５】
コンピュータにより実行されると、当該コンピュータを、請求項１から請求項１４のいずれかに記載の翻訳知識最適化装置として動作させる、翻訳知識最適化のためのコンピュータプログラム。
【請求項１６】
請求項１５に記載のコンピュータプログラムによりプログラムされたコンピュータ。
【請求項１７】
請求項１５に記載のコンピュータプログラムを記録した、コンピュータ読取可能な記憶媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
この発明は翻訳規則等の翻訳知識を用いた機械翻訳装置のための翻訳知識の作成装置に関し、特に、訓練コーパスから自動構築した翻訳知識等の様に誤り又は冗長な情報を含む知識を取捨選択する事により、的確な翻訳知識の集合を自動的に作成するための装置及びそのためのコンピュータプログラムに関する。
【０００２】
【従来の技術】
機械翻訳の一手法として、構文トランスファ方式のものが知られている。構文トランスファ方式は、予め原言語の単語又は句から目的言語への単語又は句などへのマッピング規則（変換規則）及び単語の対訳等を準備しておき、原言語の入力文を解析した後にこのマッピング規則と単語の対訳とを適用して目的言語の翻訳文を得ようとするものである。構文トランスファ方式の機械翻訳システムの構築において最も手間がかかる作業は、この翻訳規則及び単語の対訳の様な翻訳知識の作成である。
【０００３】
元々は翻訳規則は人手で準備されるものであった。しかし、原言語と目的言語との対訳文の集合である対訳コーパスの充実に伴い、翻訳規則を対訳コーパスから自動的に獲得する手法が提案されている。翻訳規則を自動的に獲得できれば、機械翻訳システムの構築のための作業量が大いに削減される。
【０００４】
翻訳規則を対訳コーパスから自動的に獲得する手法として複数のものが提案されている。しかし、その様に自動獲得した規則には以下の様な問題がある。
【０００５】
例えば、従来の翻訳規則の自動構築方法は不完全であり、構築された翻訳規則にはどうしても誤りが含まれる事が避けられない。たとえば、後掲の非特許文献１では対訳コーパスから翻訳規則の基になる句の対応関係を自動抽出しているが、約８％の対応関係が誤っていると報告されている。誤りを含む規則が翻訳時に使用されると誤訳を生じる。
【０００６】

また、通常は一つの原文でも複数通りの翻訳を行なう事が可能である。対訳コーパスにその様な対訳群が含まれていると、その多様性のために多数の冗長な規則が獲得される。その結果、互いに競合する複数の規則が獲得されてしまう。

【０００７】

例えば言換え表現が存在すると、それらの表現ごとに異なる翻訳規則が作成される。その結果、機械翻訳を行なう際のあいまい性が増大する。あいまい性が増大すると、適切な翻訳を生成する事が困難になる。すなわち、対訳コーパス中の言換え表現により、機械翻訳の精度が低下する。

【０００８】

また、対訳コーパス中に、文脈に依存する訳又は状況に依存する訳が含まれていると、過剰な省略を行なったり、湧き出し語を生じたりする翻訳規則が得られてしまう。そうした翻訳規則は誤訳の原因となる。

【０００９】

従来、こうした冗長／競合規則を処理するためのアプローチとして、以下の二つが提案されている。第１のアプローチは、翻訳時に適切な規則を選択する事によりあいまい性を解消する方法である。第２のアプローチは、翻訳規則を自動獲得した後に、後処理として競合規則を取捨選択し、翻訳規則をより的確なものにするという方法である。本発明は、この第２のアプローチをとる。

【００１０】

この第２のアプローチによる競合規則の整理及び最適化（これを以後「翻訳規則のクリーニング」又は単に「クリーニング」と呼ぶ。）として知られているものに、後掲の非特許文献２において提案されているものと、非特許文献３において提案されているものとがある。

【００１１】

非特許文献２において提案されている手法は、自動獲得された翻訳規則のうち、同じパターンの出現する頻度が所定の値（例えば２）以上の規則のみを採用するという、規則の出現頻度に基づく手法である。非特許文献３において提案されている手法は、特に多数出現するパターンのみを処理対象とし、さらにχ二乗検定による仮説検定を行なう事によって統計的に信頼性が高い規則のみを抽出するという手法である。

【００１２】

【非特許文献１】

イマムラ、Ｋ.（２００１）.構文解析と融合した階層的句アライメント.第６回自然言語処理パシフィックリムシンポジウム（ＮＬＰＲＳ２００１）予稿集３７７頁から３８４頁（Ｈｉｅｒａｒｃｈｉｃａｌ ｐｈｒａｓｅ ａｌｉｇｎｍｅｎｔ ｈａｒｍｏｎｉｚｅｄ ｗｉｔｈ ｐａｒｓｉｎｇ. In Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ ６ｔｈ Ｎａｔｕｒａｌ Ｌａｎｇｕａｇｅ Ｐｒｏｃｅｓｓｉｎｇ Ｐａｃｉｆｉｃ Ｒｉｍ Ｓｙｍｐｏｓｉｕｍ（ＮＬＰＲＳ２００１）、ｐｐ. ３７４－３８４）

【００１３】

【非特許文献２】

メネツェス、Ａ.、リチャードソン、スティーブン Ｄ.（２００１）.バイリンガルコーパスからの変換マッピングの自動抽出のための最良優先アルゴリズム.ＭＴサミットＶＩＩＩにおける『用例ベース機械翻訳ワークショップ』予稿集、３５頁から４２頁（Ｍｅｎｅｚｅｓ, Ａ.、Ｒｉｃｈａｒｄｓｏｎ, Ｓｔｅｐｈｅｎ Ｄ.（２００１）Ａ ｂｅｓｔ ｆｉｒｓｔ ａｌｉｇｎｍｅｎｔ ａｌｇｏｒｉｔｈｍ ｆｏｒ ａｕｔｏｍａｔｉｃ ｅｘｔｒａｃｔｉｏｎ ｏｆ ｔｒａｎｓｆｅｒ ｍａｐｐｉｎｇｓ ｆｒｏｍ ｂｉｌｉｎｇｕａｌ ｃｏｒｐｏｒａ. In Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ ’Ｗｏｒｋｓｈｏｐ ｏｎ Ｅｘａｍｐｌｅ－Ｂａｓｅｄ Ｍａｃｈｉｎｅ Ｔｒａｎｓｌａｔｉｏｎ’ ｉｎ ＭＴ Ｓｕｍｍｉｔ ＶＩＩＩ、ｐｐ. ３５－４２）

【００１４】

【非特許文献３】

イマムラ、Ｋ.（２００２）.パターンベース機械翻訳のための句アライメントにより得られた翻訳知識の応用.第９回機械翻訳における理論的及び方法論的問題に関する会議予稿集、７４頁から８４頁（Ｉｍａｍｕｒａ，Ｋ.（２００２）.Ａｐｐｌｉｃａｔｉｏｎ ｏｆ ｔｒａｎｓｌａｔｉｏｎ ｋｎｏｗｌｅｄｇｅ ａｃｕｉｒｅｄ ｂｙ ｈｉｅｒａｒｃｈｉｃａｌ ｐｈｒａｓｅ ａｌｉｇｎｍｅｎｔ ｆｏｒ ｐａｔｔｅｒｎ－ｂａｓｅｄ ＭＴ.Ｉｎ Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ ９ｔｈ Ｃｏｎｆｅｒｅｒｎｃｅ Ｏｎ Ｔｈｅｏｒｅｔｉｃａｌ ａｎｄ Ｍｅｔｈｏｄｏｌｏｇｉｃａｌ Ｉｓｓｕｅｓ ｉｎ Ｍａｃｈｉｎｅ Ｔｒａｎｓｌａｔｉｏｎ（ＴＭＩ－２００２）、ｐｐ.７４－８４）

【００１５】

【発明が解決しようとする課題】

上記した非特許文献２に記載の手法では、規則の数はクリーニング前の１／９程度になり、かつ翻訳品質の若干の向上が見られたという例が非特許文献３に報告されている。しかし、冗長規則が大幅に削減されたにもかかわらず、それに見合う様な翻訳品質の向上は得られていない。

【００１６】

また、非特許文献３で提案された手法では、統計的に信頼できる規則として得られるものの数が、コーパスサイズに比べて少ない。そのため、十分な数の翻訳規則を得るためには超大規模コーパスを必要とする問題点がある。また統計的に信頼でき、かつ機械翻訳に十分な数の規則を作成可能な超大規模コーパスは現在は存在しない。

【００１７】

それゆえにこの発明の目的は、対訳コーパスから自動獲得された翻訳規則をクリーニングしてより翻訳品質を向上させる事ができる翻訳知識最適化装置及びそのためのコンピュータプログラムを提供する事である。

【００１８】

この発明のほかの目的は、通常規模の対訳コーパスから自動獲得された翻訳規則をクリーニングして、より翻訳品質を向上させる事ができる翻訳知識最適化装置及びそのためのコンピュータプログラムを提供する事である。

【００１９】

この発明のほかの目的は、通常規模の対訳コーパスから自動獲得された翻訳規則を比較的短時間にクリーニングして、より翻訳品質を向上させる事ができる翻訳知識最適化装置及びそのためのコンピュータプログラムを提供する事である。

【００２０】

【課題を解決するための手段】

本発明の第１の局面に係る翻訳知識最適化装置は、機械翻訳のための翻訳知識を最適化するための翻訳知識最適化装置であって、翻訳知識の集合を記憶するための翻訳知識記憶手段と、原言語と目的言語との対訳文を複数個含む、機械読取可能な対訳コーパスを記憶するための手段と、翻訳知識記憶手段に記憶された翻訳知識の集合を利用して、対訳コーパス中の原言語の文を目的言語に機械翻訳するための機械翻訳手段と、機械翻訳手段による翻訳結果の品質を、対訳コーパスを参照して自動的に評価して評価値を出力するための訳質自動評価手段と、訳質自動評価手段の出力する評価値が極値をとる様に、翻訳知識の集合の最適化を行なうための最適化手段とを含む。

【００２１】

好ましくは、翻訳知識は、原言語の構文パターンから目的言語の構文パターンへの構文変換規則を含む。

【００２２】

さらに好ましくは、最適化手段は、翻訳知識の集合に含まれる翻訳知識の各々について、その規則寄与度を算出するための手段と、規則寄与度が予め定める条件を満足する翻訳知識を、翻訳知識の集合から削除するための手段とを含む。

【００２３】

規則寄与度を算出するための手段は、翻訳知識の集合の全体を用いて、機械翻訳手段による翻訳、及びその翻訳の結果の訳質自動評価手段による訳質評価を行ない、初期評価値を得るための手段と、翻訳知識の集合から、ある翻訳知識を削除して得られる翻訳知識の集合を用いて、機械翻訳による翻訳、及びその翻訳の結果の訳質自動評価手段による訳質評価を行ない、削除後評価値を得るための手段と、削除後評価値と初期評価値との差分を、ある翻訳知識の規則寄与度として算出するための手段とを含んでもよい。

【００２４】

さらに好ましくは、最適化手段は、翻訳知識の集合の全体を用いて、機械翻訳手段による翻訳、及びその翻訳結果の訳質自動評価手段による訳質評価を行ない、初期評価値を得るための手段と、予め定められた方法に従って、翻訳知識の集合から複数の部分集合を作成するための手段と、複数の部分集合の各々を用いて機械翻訳手段による翻訳、及びその翻訳の訳質自動評価手段による訳質の評価を行ない、その評価値が初期評価値に対して所定の条件を満足するか否かを判定するための判定手段と、判定するための手段により評価値が所定の条件を満足すると判定された部分集合の各々について、その補集合に属する翻訳知識を翻訳知識の集合から削除するための手段とを含む。

【００２５】

部分集合を作成するための手段は、翻訳知識の集合から予め定められる数の翻訳知識を除いて得られる部分集合を複数個作成するための手段を含んでもよい。

【００２６】

好ましくは、部分集合を複数個作成するための手段は、翻訳知識の集合から一つの翻訳知識を除いて得られる部分集合を複数個作成するための手段を含む。

【００２７】

さらに好ましくは、部分集合を作成するための手段は、翻訳知識の集合から予め定められる数の翻訳知識を除いて得る事が可能な全ての部分集合を作成するための手段を含む。

【００２８】

機械翻訳手段は、原言語の文を機械翻訳する際に、翻訳知識の集合内のどの翻訳知識を使用したかについての情報を出力する機能を持ち、翻訳知識最適化装置はさらに、初期評価値を得る際に翻訳された文ごとに、機械翻訳手段から出力される、翻訳の際に使用した翻訳規則を特定する情報を記憶するための手段を含み、判定手段は、記憶するための手段に記憶されている、翻訳規則を特定する情報を参照して、複数の部分集合の各々について、当該部分集合の補集合に含まれる翻訳規則を用いて翻訳された原言語の文の集合を特定するための手段と、部分集合の各々を用いて、当該部分集合の補集合に含まれる翻訳規則を用いて翻訳された原言語の文の集合を機械翻訳手段により再び機械翻訳するための手段と、部分集合の各々に対し、初期翻訳結果のうち、当該部分集合の補集合に含まれる翻訳規則を用いて翻訳された翻訳結果を、再び機械翻訳するための手段による翻訳結果で置換え、当該置換え後の初期翻訳結果に対して訳質自動評価手段による訳質評価を行なって、当該部分集合による翻訳結果の評価値を得るための手段と、部分集合の各々に対し、当該部分集合による翻訳結果の評価値と初期評価値とが所定の条件を満足しているか否かを判定するための手段とを含んでもよい。

【００２９】

好ましくは、判定するための手段は、部分集合の各々に対し、当該部分集合による翻訳結果の評価値が、初期評価値を上回っているか否かを判定するための手段を含む。

【００３０】

好ましくは、翻訳知識最適化装置はさらに、予め準備された、原言語と目的言語との対訳文からなる訓練コーパスから、各々が訓練サブコーパス及び評価サブコーパスを含む複数個のサブコーパス対を作成するための手段と、予め定められる翻訳規則の構築方式に従って、与えられる対訳コーパスから翻訳規則を自動的に構築するための翻訳知識自動構築手段と、翻訳知識自動構築手段を用いて訓練コーパスから翻訳知識を自動構築し、基本翻訳知識として記憶するための基本翻訳知識記憶手段と、複数個のサブコーパス対の各々に対して、訓練サブコーパスから翻訳知識自動構築手段を用いて翻訳知識の集合を自動構築し

、当該翻訳知識の集合に対し、評価サブコーパスを機械読取可能な対訳コーパスとして、翻訳知識記憶手段、機械読取可能な対訳コーパスを記憶するための手段、機械翻訳手段、訳質自動評価手段、及び最適化手段による最適化を行なうための手段と、最適化を行なうための手段によって最適化された、複数個のサブコーパス対の各々に対して得られる翻訳知識の集合を、一つの翻訳知識の集合に集約するための手段とを含む。

【００３１】

さらに好ましくは、集約するための手段は、基本翻訳知識記憶手段に記憶された基本翻訳知識に含まれる翻訳知識の各々について、最適化手段により算出された差分を、複数個のサブコーパス対の全てにわたって合計するための差分合計手段と、差分合計手段により合計された差分が所定の条件を満足する翻訳知識を削除する様に基本翻訳知識記憶手段に記憶されている基本翻訳知識を更新するための手段とを含む。

【００３２】

基本翻訳知識を更新するための手段は、差分合計手段により合計された差分が負となる翻訳知識を削除する様に基本翻訳知識記憶手段に記憶されている基本翻訳知識を更新するための手段を含んでもよい。

【００３３】

好ましくは、複数個のサブコーパス対を作成するための手段は、訓練コーパスを予め定める個数に実質的に等分して予め定める個数の評価サブコーパスを作成するための手段と、予め定める個数の評価サブコーパスの各々に対して、訓練コーパスから当該評価サブコーパスを除いたコーパスを作成し、当該評価サブコーパスと対となる訓練サブコーパスを作成するための手段とを含む。

【００３４】

本発明の第２の局面に係るコンピュータプログラムは、コンピュータにより実行されると、当該コンピュータを、上記したいずれかの翻訳知識最適化装置として動作させるものである。

【００３５】

本発明の第３の局面に係るコンピュータは、上記したコンピュータプログラムによりプログラムされたコンピュータプログラムによりプログラムされたコンピュータである。

【００３６】

本発明の第４の局面に係る記憶媒体は、上記したコンピュータプログラムを記録した、コンピュータ読取可能な記憶媒体である。

【００３７】

【発明の実施の形態】

以下、本発明の実施の形態について説明する。以下の説明では、同じ部品には同じ参照番号を付す。それらの機能も同一である。従って、それらについての詳細な説明は繰返さない。

【００３８】

なお以下の説明では、第１及び第２の実施の形態を説明する。これらの実施の形態の基本的な考え方は以下の通りである。すなわち、自動構築された翻訳規則を用いて評価コーパス中の原言語の文を機械翻訳する。機械翻訳した結果に対し、非特許文献４に記載されている様な訳質の自動評価を行ない、自動評価値を得る。この自動評価値を向上させる様に翻訳規則の取捨選択を行なう事により、最適な翻訳規則の組合せ（最適な翻訳規則集合）を得る。

【００３９】

以下の実施の形態では、最適な翻訳規則の組合せには山登り法を使用する。この際、組合せごとに得られる自動評価値を評価関数の出力とみなす。

【００４０】

特に以下の実施の形態では、自動構築された翻訳規則集合に対し規則の削除だけを行なう事により、翻訳規則集合の最適化を行なう。この様に規則の削除に限定する事により、クリーニングのための処理が早くなるという効果がある。

【００４１】
また、以下の実施の形態では英語から日本語に翻訳する際の翻訳規則集合を最適化する場合について説明する。しかし、本発明はこうした言語の組合せに限定されるわけではなく、翻訳規則を適用する事により翻訳できる言語の組合せであればどの様なものに対しても適用できる。
【００４２】
［第１の実施の形態］
－構成－
図１は本発明の第１の実施の形態に係る翻訳規則抽出装置２０の機能的構成を示すブロック図である。図１を参照して、翻訳規則抽出装置２０は、原言語（英語）と目的言語（日本語）との対訳文を多数含む訓練コーパス３０と、訓練コーパス３０から機械翻訳規則を自動的に構築するための機械翻訳規則自動構築部３２と、機械翻訳規則自動構築部３２が構築した翻訳規則集合に対して、後述する様なフィードバッククリーニング処理を行なうためのフィードバッククリーニング部３４と、フィードバッククリーニング部３４がフィードバッククリーニングを行なう際に、訳質評価のために参照する評価コーパス３６とを含む。評価コーパス３６中の対訳文は、英語の原文と、原文を人が日本語に翻訳した結果（参照訳と呼ぶ。）とからなる。
【００４３】
フィードバッククリーニング部３４は、機械翻訳規則自動構築部３２により訓練コーパス３０から自動的に構築された翻訳規則の集合を記憶するための翻訳規則集合記憶部４０と、翻訳規則集合記憶部４０に記憶された翻訳規則を用いて評価コーパス３６中の全ての英語の原文を目的言語の文に翻訳するための機械翻訳エンジン４２とを含む。機械翻訳エンジン４２は構文トランスファ方式のものであえる。
【００４４】
フィードバッククリーニング部３４はさらに、機械翻訳エンジン４２による翻訳結果を、各文の翻訳の際に使用された翻訳規則を特定する情報とともに記憶するための翻訳結果記憶部４３を含む。翻訳結果記憶部３５はまた、翻訳結果ともに各文の翻訳の際に使用された翻訳規則を特定する情報も記憶する。
【００４５】
フィードバッククリーニング部３４はこれに加えて、翻訳結果記憶部４３に記憶されている日本語の文（翻訳文）の訳の品質（訳質）を、評価コーパス３６を用いて自動的に評価するための訳質自動評価部４４と、翻訳規則集合記憶部４０に含まれる規則ごとに、その規則を削除した後の自動評価値を算出し、削除前の自動評価値との差分（この差分をその規則の「規則寄与度」と呼ぶ。）を算出するための規則寄与度算出部４６とを含む。規則寄与度算出部４６は、寄与度の算出の際には、訳質自動評価部４４による評価値と、翻訳結果記憶部４３に記憶されている、翻訳の際に使用された翻訳規則を特定する情報とを用いる。
【００４６】
フィードバッククリーニング部３４はさらに、翻訳規則のうち、寄与度算出部４６が算出した規則寄与度が所定の条件を充足した翻訳規則（本実施の形態では規則寄与度が負の値である翻訳規則）を翻訳規則集合記憶部４０中の翻訳規則の集合から削除するための翻訳規則削除部４８を含む。
【００４７】
本実施の形態では、機械翻訳規則自動構築部３２による翻訳規則の自動構築には、前述の非特許文献３に記載された方法を使用する。
【００４８】
本実施の形態では、機械翻訳エンジン４２は、構文トランスファ方式であって、後掲の参考文献１に記載されたものを使用する。機械翻訳エンジン４２は、英語の構文構造を日本語の構文構造に変換する翻訳規則を使用する。機械翻訳エンジン４２が使用する翻訳規則の例を図２に示す。この例では、１つの規則は、構文規則と、原言語パターンと、目的言

語パターンと、用例とを含む。
【００４９】
構文カテゴリは、この規則が適用される英語構文ノードのカテゴリを表す。
【００５０】
原言語パターンは、この規則が適用される英語構文構造のパターンを示す。原言語パターンは、Ｘ，Ｙなどの非終端記号（変数）と、単語又はマーカなどの終端記号との列である。
【００５１】
目的言語パターンは、この規則が適用された場合に生成される日本語構文構造のパターンを示す。原言語パターンに対応する変数（Ｘ’、Ｙ’など）と、単語で表現された終端記号との列である。
【００５２】
用例は、訓練コーパス中に現れた変数の実例である。変数の数と一致する主辞単語の組である。本実施の形態における翻訳規則集合記憶部４０中の各規則の用例は、訓練コーパス３０中での出現例となる。
【００５３】
翻訳規則集合記憶部４０が記憶する翻訳規則は機械翻訳エンジン４２が使用する翻訳規則のフォーマットに従ったものである。
【００５４】
図２に示す規則のうち、たとえば規則番号１のものは、英語の”ｐｒｅｓｅｎｔ　ａｔ　ｔｈｅ　ｃｏｎｆｅｒｅｎｃｅ”という句に適用され、「会議（ｃｏｎｆｅｒｅｎｃｅの訳）で発表する（ｐｒｅｓｅｎｔの訳）」という訳を生成する事を表している。
【００５５】
訳質自動評価部４４は、後掲の参考文献２に記載のＢＬＥＵを用いる。ＢＬＥＵの様な機械翻訳の自動評価法についてはこの他にもいくつか提案されている。これらは、機械翻訳システムの開発時、従来主観評価を行なっていた部分を置換える事により、開発サイクルのスピードアップを狙ったものである。これらは全自動で行なわれるため、従来考えられていた様な開発支援ばかりではなく、本実施の形態の様に翻訳システムの自動チューニングにも利用できる。
【００５６】
本実施の形態で訳質自動評価に使用するＢＬＥＵは、評価コーパスの原文を機械翻訳エンジン４２により機械翻訳した結果と、評価コーパス３６中の参照訳との類似度を計算し、訳質をスコア（ＢＬＥＵスコア）として出力する。類似度は、両者のＮ－ｇｒａｍ一致数で測定される。Ｎは可変であるが、本実施の形態では１－ｇｒａｍから４－ｇｒａｍまでを用いる。
【００５７】
ここで注意すべきは、ＢＬＥＵスコアを本実施の形態の様に機械翻訳規則集合の評価に用いるためには、ある程度の大きさを持った文集合を用いる必要がある事である。ＢＬＥＵスコアを１文ごとに算出する事も可能ではあるが、そのままでは主観評価とのずれが大きい。個々の類似度を翻訳結果集合に含まれる翻訳文の全体について計算し総和をとる事により、個々の誤差を相殺できる。
【００５８】
規則寄与度算出部４６は、次の様にして規則ごとに規則寄与度を算出する。まず、機械翻訳エンジン４２による評価コーパス３６の原言語の全ての文の翻訳結果に対し訳質自動評価部４４が算出したスコアを用い、基準となる自動評価値を得る。この値を削除前自動評価値と呼ぶ。この翻訳により、どの文の翻訳にどの規則が使用されたかという情報も得られる。
【００５９】
次に、翻訳規則集合記憶部４０内の翻訳規則ごとに、翻訳規則集合記憶部４０からその規則を削除して得られる部分集合を用いて評価コーパス３６の原言語の全ての文の翻訳を行

なった場合のスコアを計算する。このスコアと削除前自動評価値の差分が規則寄与度である。本実施の形態では、削除後のスコアの計算を以下の考え方に従って行なう。なお、この例では、当然の事ながら、削除される一つの翻訳規則からなる集合と、その翻訳規則を削除する事により形成される部分集合とは互いに補集合の関係にある。

【００６０】

基本的考え方に従って、翻訳規則集合記憶部４０内の規則の組合せ（部分集合）ごとに評価コーパス３６を全て翻訳する事も理論的には考えられる。しかしその場合には翻訳回数が非常に多くなる。よほど計算機資源に恵まれていないと合理的な時間内に結果を得る事ができない。そこで、以下の様にして計算量を少なくする。

【００６１】

機械翻訳エンジン４２による機械翻訳では、１文を翻訳すると、その翻訳に使用された翻訳規則を特定できる。その情報は翻訳結果記憶部４３に記憶されている。逆にいうと、評価コーパス３６全体を翻訳すると、各規則が使われる文を特定できる。

【００６２】

ある規則を翻訳規則集合から削除して得られる部分集合を用いて機械翻訳エンジン４２により翻訳を行なうとき、それによって変化する翻訳文は、そのある規則の削除前にその規則を使用して翻訳された翻訳文だけである。他の文は別の規則を使用して翻訳されたので、削除対象の規則が削除された後の翻訳規則集合を用いた翻訳を行なっても翻訳結果は変化しない。

【００６３】

従って、翻訳規則集合からある規則を削除した場合、削除前にその規則を使用して翻訳した文のみを削除後の翻訳規則集合を用いて翻訳し、他の訳文とあわせて参照訳との類似度を求めれば削除後のＢＬＥＵスコアが得られる。全ての文を翻訳する必要はない。

【００６４】

以上から、翻訳規則の削除のみを行なう事により、合理的な時間内に結果を得る事が可能になる。

【００６５】

すなわち規則寄与度算出部４６は、訳質自動評価部４４による削除前自動評価値と、翻訳にどの規則が使用されたか（どの規則がどの文の翻訳に使用されたか）に関する、翻訳結果記憶部４３に記憶されている情報を得る。規則ごとに、その規則を用いて翻訳された文を、その規則以外の規則を用いて再翻訳した場合の、訳文全体の自動評価値を算出する。この評価値と削除前自動評価値との差分（削除前自動表価値－削除後の評価値）を算出し、それをその規則の規則寄与度とする。規則寄与度算出部４６はさらに、こうして算出された規則寄与度が負となる（つまり、削除する事により自動評価値が大きくなる）規則の規則番号を翻訳規則削除部４８に与える機能を持つ。なお、規則寄与度算出部４６では、その処理の収束を早めるため、削除される規則同士は互いに独立であると仮定し、１回の繰返しで、削除すべき規則を全て決定し削除している。

【００６６】

より具体的には、規則寄与度算出部４６は以下の様にして規則寄与度を算出する。翻訳規則集合のうち、機械翻訳エンジン４２による翻訳の際に使用された翻訳規則の各々について、その規則を翻訳の際に使用した文の集合を求める。その文の集合が空集合でなければ、基の規則集合からその翻訳規則を取除いて得られる部分集合を用いて、その文の集合内の各文について機械翻訳エンジン４２による翻訳を再度行なう。翻訳結果記憶部４３に記憶された翻訳結果のうち、この翻訳規則を用いて翻訳が行なわれたものを、再翻訳したものと置換える。そして再度訳質自動評価部４４によって訳質の自動評価を行なう。こうして得られた削除後の評価値と削除前自動評価値との差分がこの翻訳規則の規則寄与度となる。

【００６７】

この処理を、翻訳規則集合記憶部４０内の全ての翻訳規則に対して行ない、規則寄与度が負の規則を特定する事により、削除すべき翻訳規則が決定される。

【００６８】
翻訳規則削除部４８は、翻訳規則集合記憶部４０内の規則のうち、規則寄与度算出部４６から与えられた情報に対応する翻訳規則を削除する機能を持つ。
【００６９】
－動作－
第１の実施の形態に係る翻訳規則抽出装置２０は以下の様に動作する。訓練コーパス３０及び評価コーパス３６は予め準備されているものとする。機械翻訳規則自動構築部３２は、訓練コーパス３０内の各対訳文から翻訳規則を自動構築し、翻訳規則集合記憶部４０に記憶させる。
【００７０】
機械翻訳エンジン４２は、評価コーパス３６に含まれる対訳文のうちの原文の全てを、翻訳規則集合記憶部４０に記憶されている翻訳規則を用いて翻訳する。翻訳結果は、翻訳の際に使用された翻訳規則を特定する情報とともに翻訳結果記憶部４３に記憶される。
【００７１】
訳質自動評価部４４は、翻訳結果記憶部４３に記憶されている翻訳文の訳質を、評価コーパス３６に記憶されている参照訳を用いて自動的にＢＬＥＵスコアとして評価し、その結果を規則寄与度算出部４６に与える。
【００７２】
規則寄与度算出部４６は、訳質自動評価部４４から与えられたＢＬＥＵスコアを削除前自動評価値とする。次に規則寄与度算出部４６は、翻訳規則集合記憶部４０内の各翻訳規則について、上記した方法に従って規則寄与度を算出する。そして、規則寄与度が負となる規則を特定し、その情報を翻訳規則削除部４８に与える。
【００７３】
翻訳規則削除部４８は、この情報に従って翻訳規則集合記憶部４０に記憶されている翻訳規則集合内の規則を削除する。削除処理後の翻訳規則集合記憶部４０に記憶されている翻訳規則集合は、クリーニングされ最適化されたものとなる。
【００７４】
－具体例－
翻訳例及び規則寄与度の算出の具体例を示す。なお、削除前自動評価値は０．２３３３６３とする。
【００７５】
－翻訳例１－
図２の規則５は、文脈依存訳から作成された誤り規則の例である。”ｔｈｅ　ｎｅａｒｅｓｔ　ｓｕｂｗａｙ　ｓｔａｔｉｏｎ”と「最寄りの地下鉄」から作成された規則であり、原文の”ｓｔａｔｉｏｎ”の訳が日本語では省略されている。
【００７６】
英語”Ｐｌｅａｓｅ　ｔｅｌｌ　ｍｅ　ｗｈｅｒｅ　ｔｈｅ　ｎｅａｒｅｓｔｒａｉｌｒｏａｄ　ｓｔａｔｉｏｎ　ｉｓ．”を翻訳すると，この規則５が適用されて，日本語「最寄りの鉄道はどこにありますか、教えていただけますか。」と翻訳される。
【００７７】
規則５を削除すると、この翻訳は「最寄りの鉄道の駅はどこにありますか、教えていただけますか」に変化する。削除後自動評価値は０．２３３５４９となる。
【００７８】
従って、規則５の規則寄与度は０．２３３３６３－０．２３３５４９＝－０．０００１８６となる。従って規則５は削除される。削除の結果、”ｔｈｅ　ｎｅａｒｅｓｔ　ｒａｉｌｒｏａｄ　ｓｔａｔｉｏｎ”は「最寄りの鉄道の駅」と正しく翻訳されるようになる。
【００７９】
－翻訳例２－
図２の規則６は、翻訳規則自動構築誤りによって作成された誤った規則の例である。自動構築時、”ｒｅｎｔ　ｔｗｏ　ｂｉｃｙｃｌｅｓ”を解析した結果、”ｒｅｎｔ　ｔｗｏ

”が動詞句、”ｂｉｃｙｃｌｅｓ”が名詞句になった例である。正しくは、”ｒｅｎｔ”が動詞句、”ｔｗｏ　ｂｉｃｙｃｌｅｓ”が名詞句であるが、翻訳規則の自動構築の際にはこの種の誤りの発生を完全に防止する事はできない。

【００８０】

英語“Ｉ　ｗａｎｔ　ｔｏ　ｒｅｎｔ　ｔｗｏ　ｒａｃｋｅｔｓ”を翻訳すると、規則６が適用されて「ラケットを２借りたいのですが」と翻訳される。規則６を削除すると、この翻訳は「ラケットを２本借りたいのですが」に変化する。すると、規則６の削除後の自動評価値は０．２３３５２９となる。規則６の規則寄与度は－０．０００１６６となり、規則６は削除される。

【００８１】

－翻訳例３－

図２の規則７及び規則８は、言換え表現から作られた規則の例である。どちらも正しい規則であるが、互いに競合する規則である。

【００８２】

英語“Ｐｌｅａｓｅ　ｃａｓｈ　ｔｈｉｓ　ｔｒａｖｅｌｅｒ’ｓ　ｃｈｅｃｋ．”を翻訳する際には、規則７又は規則８のいずれかが適用される。今回は規則７が選ばれたものとする。翻訳結果は「このトラベラーズチェックを現金にしたいのですが」となる。

【００８３】

規則７を削除すると、この翻訳は「このトラベラーズチェックを現金にしてください」に変化する。すると削除後自動評価値は０．２３３５８５となる。これは、評価コーパス３６中に、規則８に一致する対訳文が、規則７に一致する対訳文よりも多く含まれている事を示す。

【００８４】

規則７の規則寄与度はこの場合－０．０００２２２となる。その結果、規則７が削除され、評価コーパス３６中により多く出現する表現に一致する翻訳が行なわれる様になる。

【００８５】

－実施の形態１の効果－

以上の第１の実施の形態の翻訳規則抽出装置２０では、フィードバッククリーニング部３４の機能により、対訳コーパスから自動構築された翻訳規則群を、訳質自動評価部を用いて自動的にクリーニングする事ができる。その結果、機械翻訳結果に悪影響を及ぼす翻訳規則が排除されるので、自動構築された翻訳規則を用いる翻訳システムの翻訳結果の品質が向上するという効果が得られる。現実に、クリーニング後の翻訳規則を用いて翻訳を行なった結果に対しては、未クリーニングの翻訳規則を用いた翻訳結果よりもよい評価が得られた。

【００８６】

－コンピュータによる実現－

以上述べた第１の実施の形態に係る翻訳規則抽出装置２０は、コンピュータ及びその上で実行されるソフトウェアによっても実現される。図３に翻訳規則抽出装置２０を構成するコンピュータの外観図を、図４にそのブロック図を、それぞれ示す。

【００８７】

図３を参照して、翻訳規則抽出装置２０を構成するコンピュータシステムは、ＣＤ－ＲＯＭ（Ｃｏｍｐａｃｔ　Ｄｉｓｃ　Ｒｅａｄ－Ｏｎｌｙ　Ｍｅｍｏｒｙ）ドライブ７０及びＦＤ（Ｆｌｅｘｉｂｌｅ　Ｄｉｓｋ）ドライブ７２を有するコンピュータ６０と、いずれもコンピュータ６０に接続されたモニタ６２、キーボード６６、及びマウス６８とを含む。

【００８８】

図４を参照して、コンピュータ６０はさらに、ＣＰＵ（中央演算処理装置：Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）７６と、ＣＰＵ７６に接続されたバス８６と、バス８６を介してＣＰＵ７６と相互に接続されたＲＡＭ７８、ＲＯＭ８０、及びハードディスク７４とを含む。バス８６にはＣＤ－ＲＯＭドライブ７０及びＦＤドライブ７２も接続

される。ＣＤ－ＲＯＭドライブ７０にはＣＤ－ＲＯＭ８２が、ＦＤドライブ７２にはＦＤ８４が、それぞれ装着され、ＣＰＵ７６等との間のデータの入出力を行なう事ができる。

【００８９】

図３及び図４に示すコンピュータは、以下に述べる様な制御構造を有するコンピュータプログラム（以下単に「プログラム」と呼ぶ。）を実行する事により、図１に示す翻訳規則抽出装置２０として動作する。このプログラムは、たとえばＣＤ－ＲＯＭ８２上にコンピュータ読取可能なデータとして記録されて流通する。このＣＤ－ＲＯＭ８２をＣＤ－ＲＯＭドライブ７０に装着し、プログラムを読出してハードディスク７４に記憶する事により、コンピュータ６０はいつでもこのプログラムを実行する事ができる。なお、訓練コーパス３０、評価コーパス３６などはハードディスク７４に記憶しておく。ＣＰＵ７６はまた、必要なデータはハードディスク７４から読出してＲＡＭ７８に格納する。

【００９０】

プログラムの実行時には、ハードディスク７４に記憶されているプログラムをＲＯＭ８０にロードする。ＣＰＵ７６は、図示しないプログラムカウンタにより示されるアドレスの命令をＲＯＭ８０から読出して実行する。ＣＰＵ７６は、実行結果を所定のアドレスに出力し、あわせて実行結果に従ってプログラムカウンタの内容を更新する。

【００９１】

こうした処理を繰返し行なう事により、最終的な翻訳規則の集合が得られる。得られた結果は、本実施の形態では最終的にハードディスク７４に格納される。

【００９２】

なお、コンピュータ６０の動作自体は周知であるので、ここではその詳細については繰返さない。

【００９３】

－プログラムの制御構造－

図５を参照して、フィードバッククリーニング部３４を実現するプログラムは以下の制御構造を有する。まず、このプログラムは、起動されるとステップ１００で削除規則集合Ｒｒｅｍｏｖｅを空集合とする。ステップ１０２で、機械翻訳エンジン４２を用いて評価コーパス３６の全ての原文を翻訳規則集合記憶部４０の翻訳規則を参照して翻訳し、翻訳結果集合Ｄｏｃを得る。このとき同時に、翻訳するためにどの規則が使われたかを記録する。この記録に基づき、ある規則ｒを用いて翻訳された原文集合を求める。この原文集合を、規則ｒに対してＳ［ｒ］とする。続いてステップ１０４で、この翻訳結果集合Ｄｏｃから、訳質自動評価部４４を用いて初期（削除前）自動評価値ｓｃｏｒｅを算出する。

【００９４】

続いて以下に述べるステップ１０８～ステップ１２０までの処理を、翻訳規則集合記憶部４０内の全ての翻訳規則ｒについて繰返す。まずステップ１０８では、規則ｒを用いた原文集合Ｓ［ｒ］が空集合か否かを判定する。空集合の場合にはこの規則ｒに対しては何も行なわない。Ｓ［ｒ］が空集合でない場合、制御はステップ１１０に進む。

【００９５】

ステップ１１０では、原文集合Ｓ［ｒ］に含まれる原文の全てを、翻訳規則集合から規則ｒを取除いたものを用いて、機械翻訳エンジン４２により翻訳する。その結果得られる訳文の集合をＴ［ｒ］とする。続くステップ１１２で、ステップ１０２で求めた翻訳結果集合Ｄｏｃ中の、規則ｒを用いて翻訳された文の集合を集合Ｔ［ｒ］で置換えた新たな翻訳結果集合Ｄｏｃ［ｒ］を求める。ステップ１１４で、この翻訳結果集合Ｄｏｃ［ｒ］に対する、訳質自動評価部４４による自動評価値ｓｃｏｒｅ［ｒ］を算出する。この自動評価値ｓｃｏｒｅ［ｒ］が削除後自動評価値である。ステップ１１６で、初期自動評価値ｓｃｏｒｅからこの削除後自動評価値ｓｃｏｒｅ［ｒ］を減算し、その結果を規則寄与度ｃｏｎｔｒｉｂ［ｒ］に代入する。

【００９６】

ステップ１１８では、規則寄与度ｃｏｎｔｒｉｂ［ｒ］が負か否かを判定する。規則寄与度ｃｏｎｔｒｉｂ［ｒ］が負であれば、制御はステップ１２０に進み、この規則ｒを削除

規則集合Ｒｒｅｍｏｖｅに追加する。規則寄与度ｃｏｎｔｒｉｂ［ｒ］が負でなければその規則については何もしない。

【００９７】

以上のステップ１０８～１２０の処理を全ての規則ｒについて繰返し行なった後、制御はステップ１２４に進む。ステップ１２４では、削除規則集合Ｒｒｅｍｏｖｅが空集合でないか判定する。削除規則集合Ｒｒｅｍｏｖｅが空集合であればこのプログラムの実行を終了する。削除規則集合Ｒｒｅｍｏｖｅが空集合でない場合には、ステップ１２６でこの削除規則集合Ｒｒｅｍｏｖｅに含まれる規則を翻訳規則集合記憶部４０に含まれる翻訳規則集合から削除する。この後、制御は先頭のステップ１００に戻り、ステップ１２４で削除規則集合Ｒｒｅｍｏｖｅが空集合であると判定されるまで、以上の処理を繰返す。

【００９８】

以上の様な制御構造を有するプログラムを図３及び図４に示すコンピュータ６０で実行する事により、図１に示す第１の実施の形態の翻訳規則抽出装置２０を実現する事ができる。

【００９９】

－変形例－

上記した第１の実施の形態では、翻訳規則の全てについてその規則寄与度を算出して削除するか否かを判定している。しかし、全ての翻訳規則についてこうした処理を行なう必要はなく、一部の規則のみに対して行なってもそれなりの効果が得られる。しかし、翻訳規則の全てについて規則寄与度を算出して削除するか否かを判定した方が、明らかに最終的に得られる翻訳規則に誤った規則又は冗長な規則が含まれる可能性が低くなる。従って、翻訳規則の全てについて上記した処理を行なう方が好ましい。

【０１００】

また上記した実施の形態では、一度に一つずつの翻訳規則についてその規則寄与度を算出している。この様にすると、翻訳規則の各々について削除すべきか否かを判定できるので、翻訳規則の最適化を目指す上では好ましい。しかし、この判定を翻訳規則の一つずつについて行なう事が必須というわけではない。原理的には、一度に複数の翻訳規則を削除した場合を想定してその寄与度を算出し、その結果に従ってそれら複数の翻訳規則をまとめて削除する事も可能であり、そうした方法によってもある程度は上記した実施の形態と同様の効果を奏すると考えられる。

【０１０１】

また、削除すべきか否かを決定する翻訳規則の数は、上記した実施の形態では「１」に固定されている。この様に数を固定する事により、処理が簡単になるので、実際にはこうした形で本発明を実施する事が多いと思われる。しかしこの数も常に同じ数である必要はない。たとえば何らかの基準によってその都度決められる数の翻訳規則を処理対象として、その規則寄与度を算出する様にしてもよい。

【０１０２】

本発明では、翻訳規則の集合の任意の部分集合（当初の翻訳規則内の翻訳規則の任意の組合せ）を取出し、どの部分集合を用いて機械翻訳を行なえば翻訳結果の訳質として最もよい評価値が得られるか、を確認し、その結果によって最終的な翻訳規則の集合を決定する、という考え方を基本的枠組みとしている。その基本的枠組みの中で、さらに計算機資源を節約しつつどの様にすれば効率的にある程度好ましい基本規則の集合が得られるか、についての一つの実施の形態が上記した第１の実施の形態である。この基本的枠組みの中で、第１の実施の形態とは細部で異なる実施の形態が他にもあり得る事、及びそうした実施の形態が上記した第１の実施の形態についての詳細な説明に基づいて容易に実施する事ができる事は、当業者であれば容易に理解できるであろう。

【０１０３】

［第２の実施の形態］

－概略－

第１の実施の形態の装置によりクリーニングした翻訳規則集合を用いる事により、翻訳の

品質はかなり向上する。しかし、未だ向上の余地があると思われる。また、第１の実施の形態では、訓練コーパスとは別に評価コーパスを準備する必要がある。評価コーパスについては、原文に対する参照訳が必要となるため、できれば評価コーパスを特に準備する必要がないほうが望ましい。

【０１０４】

また、一般的には、訓練コーパスに比べ、評価コーパスはサイズが小さい場合が多い。そのため、たとえ大域最適解を発見する事ができても、評価コーパスではすべての規則をテストできず、クリーニング漏れが発生する。その様なクリーニング漏れの発生を防止できる事が望ましい。

【０１０５】

そこでこの第２の実施の形態の装置では、第１の実施の形態の装置で用いたフィードバッククリーニング部３４によるクリーニング結果に対し、交差検定と同様な考え方を用い、より最適解に近いものを得るためのクリーニングを行なう。本明細書では、こうしたクリーニングの仕方を「交差クリーニング」と呼ぶ。

【０１０６】

一般的にＮ分割交差検定とは、データをＮ個のサブデータにほぼ等分し、一つをあるモデルのパラメータ推定に用い、推定されたモデルの当てはまりのよさを残りのデータで評価する事をＮ個のサブデータの全てについて行なう、という方法である。この交差クリーニングにより、上記した様なクリーニング漏れを防止する事ができる。

【０１０７】

図６に、この実施の形態で行なわれる交差クリーニングの概要を示す。以下、この処理の概要を説明する。

【０１０８】

ステップ１．　訓練コーパス１４０をＮ個に分割する。

【０１０９】

ステップ２．　分割によって得られたＮ個のサブコーパスを評価サブコーパス１６２Ａ、１６２Ｂ、…とする。元の訓練コーパス１４０から一つの評価サブコーパス（例えば評価サブコーパス１６２Ａ）を除いたＮ－１個のサブコーパス（評価サブコーパス１６２Ａの場合、評価サブコーパス１６２Ｂ、１６２Ｃ、…）を一つにまとめ、訓練サブコーパス１６０Ａを作成する。評価サブコーパス１６２Ａと訓練サブコーパス１６０Ａとを対にする。

【０１１０】

同様に、各評価サブコーパス１６２Ｂ、１６２Ｃ、…に対し、訓練サブコーパス１６０Ｂ，１６０Ｃ，…を作成し、それらを元の評価サブコーパス１６２Ｂ、１６２Ｃ、…と対にする。

【０１１１】

以上の処理の結果、Ｎ個のサブコーパス対１５０Ａ，１５０Ｂ、…が形成される。これらＮ個のサブコーパス対１５０Ａ、１５０Ｂ，…に含まれる訓練サブコーパス１６０Ａ，１６０Ｂ，…の各々から、第１の実施の形態と同様にして翻訳規則の自動構築１５１を行なう。その結果、Ｎ個の自動構築翻訳規則集合１５２Ａ，１５２Ｂ、…が得られる。

【０１１２】

ステップ３．　さらに、これら自動構築翻訳規則集合１５２Ａ，１５２Ｂ，に対し、それぞれ評価サブコーパス１６２Ａ，１６２Ｂ，…を用いて、第１の実施の形態と同様のフィードバッククリーニング１５３を行なう。その結果、Ｎ個のクリーニング後規則集合１５４Ａ，１５４Ｂ，…が得られる。

【０１１３】

ステップ４．　最後に、Ｎ個のクリーニング後規則集合１５４Ａ，１５４Ｂ，…に対して機械翻訳規則集約処理１５６を行ない、最終的な交差クリーニング後翻訳規則集合１５８を作成する。

【０１１４】

通常の交差検定との相違点はステップ４である。本実施の形態では、規則毎に規則寄与度の総和を算出し、それが０以上である場合に最終翻訳規則集合にその規則を出力する。逆にいえば、規則寄与度の総和が０未満の規則は翻訳規則集合から削除する。

【０１１５】

－構成－

図７にこの第２の実施の形態の翻訳規則抽出装置１８０の機能的ブロック図を示す。図７を参照して、この翻訳規則抽出装置１８０は、訓練コーパス１４０と、訓練コーパス１４０から自動的に翻訳規則を構築するための機械翻訳規則自動構築部１９８と、機械翻訳規則自動構築部１９８により自動構築された翻訳規則の集合（これを「基本翻訳規則集合」と呼ぶ。）を記憶するための基本規則集合記憶部１９６とを含む。機械翻訳規則自動構築部１９８は第１の実施の形態で使用されている機械翻訳規則自動構築部３２と全く同一の機能を持つ。

【０１１６】

翻訳規則抽出装置１８０はさらに、訓練コーパス１４０をＮ個に分割し、その一つからなる評価サブコーパス１６２と、他のＮ－１個からなる一つの訓練サブコーパス１６０とに分ける機能を持つ訓練コーパス分割部１９０と、訓練サブコーパス１６０から翻訳規則を自動構築するための機械翻訳規則自動構築部３２と、機械翻訳規則自動構築部３２の出力する翻訳規則集合を評価サブコーパス１６２を用いて第１の実施の形態と同様にしてフィードバッククリーニングするためのフィードバッククリーニング部３４とを含む。フィードバッククリーニング部３４及びその各部の機能は、第１の実施の形態におけるフィードバッククリーニング部３４及びその各部の機能と同じである。従ってそれらの詳細な説明はここでは繰返さない。

【０１１７】

翻訳規則抽出装置１８０はさらに、機械翻訳規則自動構築部３２による翻訳規則の自動構築及びフィードバッククリーニング部３４による翻訳規則のフィードバッククリーニングをＮ回繰返して実行する様に、訓練コーパス分割部１９０、機械翻訳規則自動構築部３２、及びフィードバッククリーニング部３４を制御するための繰返制御部１９２を含む。繰返制御部１９２による繰返は、訓練コーパス分割部１９０により選択される評価サブコーパス１６２を一つずつ入替えながら行なわれる。

【０１１８】

翻訳規則抽出装置１８０はこれに加えて、フィードバッククリーニング部３４の規則寄与度算出部４６により算出された規則寄与度を規則ごと及び繰返しごとに記憶するための規則寄与度記憶部２０２と、機械翻訳規則自動構築部３２及びフィードバッククリーニング部３４により作成されたＮ個のフィードバッククリーニング済みの翻訳規則集合を集約し、最終的な一つの交差クリーニング後翻訳規則集合を基本規則集合記憶部１９６内に作成するための翻訳規則集約部１９４とを含む。翻訳規則集約部１９４は、規則寄与度記憶部２０２に記憶されている規則ごと及び繰返しごとの規則寄与度を用いて、基本規則集合記憶部１９６に記憶されている基本翻訳規則集合から不要な規則を削除する事により規則の集約を行なう。

【０１１９】

機械翻訳規則自動構築部３２及びフィードバッククリーニング部３４の機能はそれぞれ第１の実施の形態で説明したものと同じである。

【０１２０】

訓練コーパス分割部１９０は、訓練コーパス１４０を以下の様に繰返しごとに異なる形で分割する。まず、前述の様に訓練コーパス１４０はＮ個のサブコーパスにほぼ等分に分割される。それらをそれぞれ第１のサブコーパス、第２のサブコーパス、…第Ｎのサブコーパスと呼ぶ事にする。

【０１２１】

繰返しの第１回目では、訓練コーパス分割部１９０は第１のサブコーパスを評価サブコーパス１６２とし、第２のサブコーパスから第Ｎのサブコーパスまでをまとめて訓練サブコ

ーパス１６０とする。繰返しの第２回目では訓練コーパス分割部１９０は、第２のサブコーパスを評価サブコーパス１６２とし、第１のサブコーパス、及び第３のサブコーパスから第Ｎのサブコーパスまでをまとめて訓練サブコーパス１６０とする。繰返しの第３回目では訓練コーパス分割部１９０は、第３のサブコーパスを評価サブコーパス１６２とし、第１のサブコーパス、第２のサブコーパス、及び第４のサブコーパスから第Ｎのサブコーパスまでをまとめて訓練サブコーパス１６０とする。以下同様にして、繰返しの第Ｎ回目では訓練コーパス分割部１９０は、第Ｎのサブコーパスを評価サブコーパス１６２とし、第１のサブコーパスから第Ｎ－１のサブコーパスまでをまとめて訓練サブコーパス１６０とする。

【０１２２】

以上が訓練コーパス分割部１９０の機能である。

【０１２３】

翻訳規則集約部１９４は、次の様にしてフィードバッククリーニング後の翻訳規則を集約する。機械翻訳規則自動構築部１９８により、訓練コーパス１４０の全体から基本翻訳規則集合が自動構築される。この基本翻訳規則集合は基本規則集合記憶部１９６に記憶される。

【０１２４】

次に、繰返制御部１９２によるＮ回のフィードバッククリーニングにより、訓練コーパス１４０のＮ個の訓練サブコーパス１６０よりＮ個の翻訳規則集合が得られる。これらを第１の翻訳規則集合、第２の翻訳規則集合、…第Ｎの翻訳規則集合と呼ぶ事とする。そして、これらの翻訳規則集合を作成する際に規則寄与度算出部４６により計算された各規則の規則寄与度が規則寄与度記憶部２０２に繰返しごとに別々に記憶される。規則ｒについてのｉ回目の繰返しの際に計算された規則寄与度を$contrib[i][r]$と表す（$1 \leqq i \leqq N$、$1 \leqq r \leqq$基本規則数）。

【０１２５】

翻訳規則集約部１９４は、全てのフィードバッククリーニングが終了すると、規則寄与度記憶部２０２を参照して、翻訳規則ｒごとに、規則寄与度記憶部２０２に記憶されている規則寄与度の総和$contrib[r]=\Sigma_i\ contib[i][r]$を計算する。そして、総和$contrib[r]$が負であればその規則ｒを基本規則集合記憶部１９６に記憶されている基本規則集合から削除する。この処理を全ての規則ｒに対して実行する事により、基本規則集合記憶部１９６に記憶されている基本規則集合に対するクリーニングが行なわれ、最終的な交差フィードバッククリーニング後の翻訳規則集合が得られる。

【０１２６】

－動作－

この第２の実施の形態に係る翻訳規則抽出装置１８０は以下の様に動作する。訓練コーパス１４０は最初に準備されているものとする。また訓練コーパス１４０をＮ個にほぼ等分する方法も予め決定されているものとする。まず機械翻訳規則自動構築部１９８が訓練コーパス１４０から翻訳規則を自動構築する。構築された翻訳規則集合（基本規則集合）は基本規則集合記憶部１９６に記憶される。

【０１２７】

以下の繰返し処理は、繰返制御部１９２による制御の下で実行される。まず訓練コーパス分割部１９０は、訓練コーパス１４０から第１のサブコーパスを選び、それを評価サブコーパス１６２とする。訓練コーパス分割部１９０はさらに、残りのＮ－１個のサブコーパスをまとめて訓練サブコーパス１６０とする。機械翻訳規則自動構築部３２は、訓練サブコーパス１６０から翻訳規則を自動構築する。構築された翻訳規則集合は翻訳規則集合記憶部４０に記憶される。

【０１２８】

機械翻訳エンジン４２は、翻訳規則集合記憶部４０に記憶されている翻訳規則を用いて、評価サブコーパス１６２中の原文集合に対する翻訳を行なう。訳質自動評価部４４は、機械翻訳エンジン４２による翻訳結果の訳質を自動評価し、スコアとして規則寄与度算出部

４６に与える。

【０１２９】

規則寄与度算出部４６は、第１の実施の形態で説明した通り、翻訳規則集合記憶部４０に記憶されている各規則について、規則寄与度を算出する。算出された規則寄与度は、規則寄与度記憶部２０２に規則ごと、繰返しごとにｃｏｎｔｒｉｂ［ｉ］［ｒ］として記憶される。

【０１３０】

上記した処理をＮ回繰返す事により、規則寄与度記憶部２０２には、規則寄与度ｃｏｎｔｒｉｂ［ｉ］［ｒ］（１≦ｉ≦Ｎ、１≦ｒ≦基本翻訳規則数）が記憶される。

【０１３１】

翻訳規則集約部１９４は、基本規則集合記憶部１９６に記憶されている各規則について、前述した通り規則寄与度の総和$contrib[r] = \Sigma_i contib[i][r]$を計算する。ｃｏｎｔｒｉｂ［ｒ］が負の場合、その規則は基本規則集合記憶部１９６内の基本規則集合から削除される。

【０１３２】

翻訳規則集約部１９４が、基本規則集合記憶部１９６に記憶されている全ての翻訳規則に対して以上の処理を実行する事により、最終的に基本規則集合記憶部１９６には、交差クリーニング後の基本規則集合が得られる。

【０１３３】

－第２の実施の形態の効果－

この第２の実施の形態の翻訳規則抽出装置１８０によって交差クリーニングした後の翻訳規則集合を用いて機械翻訳を行なったところ、第１の実施の形態により得られたものよりもさらによい結果が得られた。また、第１の実施の形態の翻訳規則抽出装置２０では、訓練コーパスとは別に評価コーパスを準備する必要があった。それに対してこの第２の実施の形態の翻訳規則抽出装置１８０では、訓練コーパス１４０のみを使用し、それと別に評価コーパスを用意する必要はない。従って、翻訳規則のクリーニングが、限られた対訳コーパスを用いて行なえ、その結果得られた翻訳規則集合を用いて、精度の高い機械翻訳を行なう事が可能になる。

【０１３４】

－コンピュータによる実現－

この第２の実施の形態に係る翻訳規則抽出装置１８０も、図３及び図４に示すコンピュータと、その上で実行されるプログラムとにより実現可能である。図８に、この第２の実施の形態に係る翻訳規則抽出装置１８０を実現するためのプログラムの制御構造をフローチャート形式で示す。

【０１３５】

図８を参照して、このプログラムは、訓練コーパス１４０から基本規則集合を自動構築するステップ２１０と、訓練コーパス１４０を均等にＮ個のサブコーパスに分類するステップ２１２とを含む。これらＮ個のサブコーパスをＥＣ［ｉ］（１≦ｉ≦Ｎ）とする。

【０１３６】

このプログラムはさらに、以下のステップ２１６からステップ２２０を、変数ｉを１からＮまで１ずつ増加させながら繰返すステップを含む。まずステップ２１６では、訓練コーパス１４０からサブコーパスＥＣ［ｉ］を取除き、訓練サブコーパス１６０を作成する。この訓練サブコーパスをＴＣ［ｉ］とする。

【０１３７】

続いてステップ２１８で、訓練サブコーパスＴＣ［ｉ］から翻訳規則集合Ｒ［ｉ］を自動構築する。さらにステップ２２０で、サブコーパスＥＣ［ｉ］を評価コーパスとみなして翻訳規則集合Ｒ［ｉ］をフィードバッククリーニングする。このフィードバッククリーニング処理の内容は、図５に示した第１の実施の形態のものと同様である。ただしこの際、図５のステップ１１６で算出された規則寄与度ｃｏｎｔｒｉｂ［ｒ］をｃｏｎｔｒｉｂ［ｉ］［ｒ］として記憶しておく事に注意する必要がある。

【０１３８】
ステップ２１６からステップ２２０までの処理をＮ回繰返した後、今度は以下に説明するステップ２２６からステップ２３２の処理を、ステップ２１０で自動構築された基本規則集合内の全ての規則ｒについて繰返し行なう（１≦ｒ≦基本規則集合内の規則数）。
【０１３９】
ステップ２２６では、翻訳規則集合Ｒ［ｉ］（１≦ｉ≦Ｎ）から、規則ｒの規則寄与度ｃｏｎｔｒｉｂ［ｉ］［ｒ］を取得する。具体的には、前述した通り図５のステップ１１６で記憶されていた規則寄与度を記憶領域から取出す。ステップ２２８で、基本規則ｒの寄与度ｃｏｎｔｒｉｂ［ｒ］＝Σ_i ｃｏｎｔｉｂ［ｉ］［ｒ］を算出する。
【０１４０】
続くステップ２３０では、ステップ２２８で算出された寄与度ｃｏｎｔｒｉｂ［ｒ］が負か否かを判定する。負であればステップ２３２でこの規則ｒを基本規則集合から取除く。負でない場合には何もしない。
【０１４１】
以上のステップ２２６からステップ２３２までの処理を、基本規則集合内の全ての規則に対して行なう事により、最終的に交差フィードバッククリーニングが行なわれた翻訳規則が得られる事については前述した通りである。この交差クリーニングにより、第２の実施の形態の説明の冒頭で説明した様なクリーニング漏れを防止する事ができる。
【０１４２】
－第２の実施の形態の変形例－
上記した第２の実施の形態の装置では、機械翻訳規則自動構築部３２とは別に機械翻訳規則自動構築部１９８を設けている。しかしこれらは必ずしも別個のものとする必要はない。同じ機械翻訳規則自動構築部を用いて、その入力及び出力の接続先を切替える様にしてもよい。
【０１４３】
また、上記した実施の形態の装置では訓練コーパス１４０をＮ個のサブコーパスにほぼ等分する事により、訓練サブコーパスと評価サブコーパスとを作成している。しかし本発明はその様な実施の形態に限定されるわけではない。例えば、訓練コーパス１４０を必ずしも等分する必要はない。実質的に大きさの異なったコーパスに分割し、後は上記した通りの処理を行なう様にしてもよい。ただしその場合には、翻訳規則集約部１９４で規則を集約する際の規則寄与度の総和計算において、コーパスの大きさに従った重みを各寄与度に乗じた後に加算する事が望ましい。
【０１４４】
－共通の変形例－
上記した二つの実施の形態では、機械翻訳エンジン４２として参考文献１に記載されたものを使用している。しかし本発明はその様な実施の形態に限定されるわけではない。翻訳規則を用いた構文トランスファ方式の機械翻訳エンジンであればどの様なものを用いてもよい。
【０１４５】
さらに、上記した二つの実施の形態では、訳質自動評価部４４による訳質の自動評価にＢＬＥＵを用いた。しかし訳質の自動評価にはＢＬＥＵのみが使用可能なわけではない。例えば、後掲の参考文献３又は参考文献４に記載のものを用いる事も可能である。
【０１４６】
自動評価値として、本実施の形態では評価コーパス内の訳文との類似度が高い場合に評価値が高くなるものを使用した。しかし自動評価値としてはその様なものには限定されず、類似度が高い場合に評価値が低くなる様なものでもよい。また、評価コーパス内の訳文との類似度が高くなるほど、特定の値に近くなる様な評価値を用いてもよい。
【０１４７】
なお、ソフトウェアの流通形態は上記した様に記憶媒体に固定された形には限定されない。たとえば、ネットワークを通じて接続された他のコンピュータからデータを受取る形で

流通する事もあり得る。また、ソフトウェアの一部が予めハードディスク５４中に格納されており、ソフトウェアの残りの部分をネットワーク経由でハードディスク５４に取込んで実行時に統合する様な形の流通形態もあり得る。

【０１４８】

一般的に、現代のプログラムはコンピュータのオペレーティングシステム（ＯＳ）によって提供される汎用の機能を利用し、それらを所望の目的に従って組織化した形態で実行する事により前記した所望の目的を達成する。従って、以下に述べる本実施の形態の各機能のうち、ＯＳ又はサードパーティが提供する汎用的な機能を含まず、それら汎用的な機能の実行順序の組合せだけを指定するプログラム（群）であっても、それらを利用して全体的として所望の目的を達成する制御構造を有するプログラム（群）である限り、それらが本発明の技術的範囲に含まれる事は明らかである。

【０１４９】

-参考文献リスト-

［参考文献１］　古瀬蔵、山本和英、及び山田節夫（１９９９）．構成素境界解析を用いた多言語話し言葉翻訳．自然言語処理、６（５）：６３－９１。

【０１５０】

［参考文献２］　ペネニ、Ｋ．，ルーコス、Ｓ．，ウォード、Ｔ，及びツー、Ｗ．－Ｊ．（２００２）．Ｂｌｅｕ：機械翻訳の自動評価方法．第４０回計算言語学学会第４０回年次大会予稿集、３１１頁から３１８頁（Ｐａｉｎｅｎｉ，Ｋ．，Ｒｏｕｋｏｓ，Ｓ．，Ｗａｒｄ、Ｔ．，ａｎｄ　Ｚｈｕ，Ｗ．－Ｊ．（２００２）．Ｂｌｅｕ：ａ　ｍｅｔｈｏｄ　ｆｏｒ　ａｕｔｏｍａｔｉｃ　ｅｖａｌｕａｔｉｏｎ　ｏｆ　ｍａｃｈｉｎｅ　ｔｒａｎｓｌａｔｉｏｎ．Ｉｎ　Ｐｒｏｃｅｅｄｉｎｇｓ　ｏｆ　ｔｈｅ　４０ｔｈ　Ａｎｎｕａｌ　Ｍｅｅｔｉｎｇ　ｏｆ　ｔｈｅ　Ａｓｓｏｃｉａｔｉｏｎ　ｆｏｒ　ＣｏｍｐｕｔａｔｉｏｎａｌＬｉｎｇｕｉｓｔｉｃｓ（ＡＣＬ），ｐｐ．３１１－３１８）

【０１５１】

［参考文献３］　ヤスダ、Ｋ．，スガヤ、Ｆ．，タケザワ、Ｔ．，ヤマモト、Ｓ．，及びヤナギダ、Ｍ．、（２００１）．パラレルコーパスから検索された翻訳解候補を用いた翻訳品質の自動評価法、機械翻訳サミット予稿集ＶＩＩＩ，３７３頁から３７８頁（Ｙａｓｕｄａ，Ｋ．，Ｓｕｇａｙａ、Ｆ．，Ｔａｋｅｚａｗａ，Ｔ．，Ｙａｍａｍｏｔｏ，Ｓ．，ａｎｄ　Ｙａｎａｇｉｄａ，Ｍ．，（２００１）．Ａｎ　ａｕｔｏｍａｔｉｃ　ｅｖａｌｕａｔｉｏｎ　ｍｅｔｈｏｄ　ｏｆ　ｔｒａｎｓｌａｔｉｏｎ　ｑｕａｌｉｔｙ　ｕｓｉｎｇ　ｔｒａｎｓｌａｔｉｏｎ　ａｎｓｗｅｒ　ｃａｎｄｉｄａｔｅｓ　ｑｕｅｒｉｅｄ　ｆｒｏｍ　ａ　ｐａｒａｌｌｅｌ　ｃｏｒｐｕｓ．Ｉｎ　Ｐｒｏｃｅｅｄｉｎｇｓ　ｏｆ　Ｍａｃｈｉｎｅ　Ｔｒａｎｓｌａｔｉｏｎ　Ｓｕｍｍｉｔ　ＶＩＩＩ，ｐｐ．３７３－３７８）

【０１５２】

［参考文献４］　アキバ、Ｙ．，イマムラ、Ｋ．，及びスミタ、Ｅ．，（２００１）（Ａｋｉｂａ，Ｙ．，Ｉｍａｍｕｒａ，Ｋ．，ａｎｄ　Ｓｕｍｉｔａ，Ｅ．，（２００１）．複数編集距離を用いた機械翻訳の自動評価．機械翻訳サミット予稿集ＶＩＩＩ、１５頁から２０頁（Ｕｓｉｎｇ　ｍｕｌｔｉｐｌｅ　ｅｄｉｔ　ｄｉｓｔａｎｃｅｓ　ｔｏ　ａｕｔｏｍａｔｉｃａｌｌｙ　ｒａｎｋ　ｍａｃｈｉｎｅ　ｔｒａｎｓｌａｔｉｏｎ　ｏｕｔｐｕｔ．Ｉｎ　Ｐｒｏｃｅｅｄｉｎｇｓ　ｏｆ　Ｍａｃｈｉｎｅ　Ｔｒａｎｓｌａｔｉｏｎ　Ｓｕｍｍｉｔ　ＶＩＩＩ，ｐｐ．１５－２０）

【０１５３】

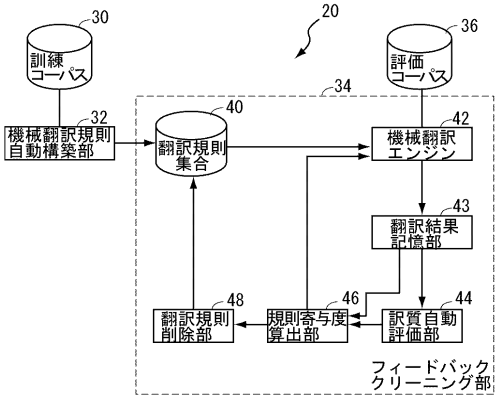
今回開示された実施の形態は単に例示であって、本発明が上記した実施の形態のみに制限されるわけではない。本発明の範囲は、発明の詳細な説明の記載を参酌した上で、特許請求の範囲の各請求項によって示され、そこに記載された文言と均等の意味及び範囲内でのすべての変更を含む。

【図面の簡単な説明】

【図１】本発明の第１の実施の形態に係る翻訳規則抽出装置２０の機能的ブロック図であ

る。
【図２】翻訳規則の例を示す図である。
【図３】翻訳規則抽出装置２０を実現するコンピュータの外観図である。
【図４】図３に示すコンピュータの回路構成を概略的に示す図である。
【図５】第１の実施の形態に係る翻訳規則抽出装置２０をコンピュータで実現するためのプログラムの制御構造を示すフローチャートである。
【図６】本発明の第２の実施の形態における交差クリーニング法の概略を説明するための図である。
【図７】第２の実施の形態の翻訳規則抽出装置１８０の機能的ブロック図である。
【図８】翻訳規則抽出装置１８０を実現するためのプログラムの制御構造を示すフローチャートである。
【符号の説明】
２０，１８０　翻訳規則抽出装置、３０，１４０　訓練コーパス、３２，１９８　機械翻訳規則自動構築部、３４　フィードバッククリーニング部、３６　評価コーパス、４０　翻訳規則集合記憶部、４２　機械翻訳エンジン、４３　翻訳結果記憶部、４４　訳質自動評価部、４６　規則寄与度算出部、４８　翻訳規則削除部、１６０　訓練サブコーパス、１６２　評価サブコーパス、１９０　訓練コーパス分割部、１９２　繰返制御部、１９４　翻訳規則集約部、１９６　基本規則集合記憶部、２０２　規則寄与度記憶部

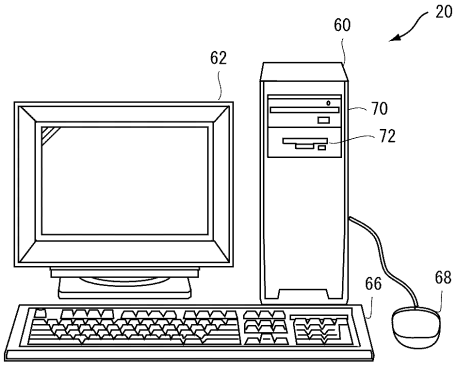
【図１】



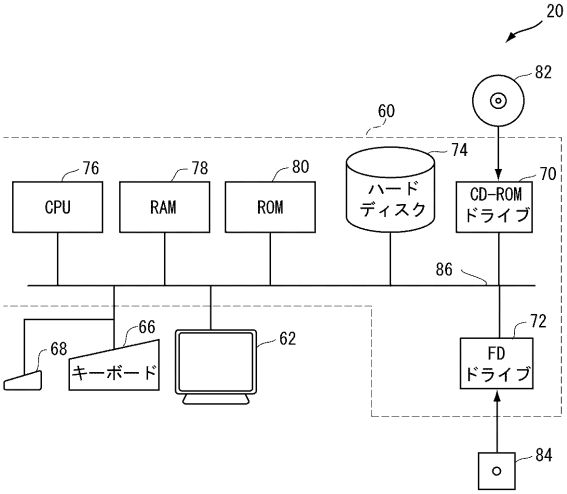
【図２】

規則番号	構文カテゴリ	原言語パターン		目的言語パターン	用例
1	VP	X_{VP} at Y_{NP}	⇒	Y'でX'	((present, conference), ...)
2	VP	X_{VP} at Y_{NP}	⇒	Y'にX'	((stay, hotel), (arrive, p.m.), ...)
3	VP	X_{VP} at Y_{NP}	⇒	Y'をX'	((look, it), ...)
4	NP	X_{NP} at Y_{NP}	⇒	Y'のX'	((man, front desk), ...)
5	NP	the X_{NMP} Y_{NP} station	⇒	X'Y'	((near, subway),...)
6	VP	X_{VP} <num-noun> Y_{NP}	⇒	Y'をX'	((rent, bycycle),...)
7	S	please X_{VP}	⇒	X'たいのですが	((send), (reserve),...)
8	S	pelase X_{VP}	⇒	X'てください	((give), ...)

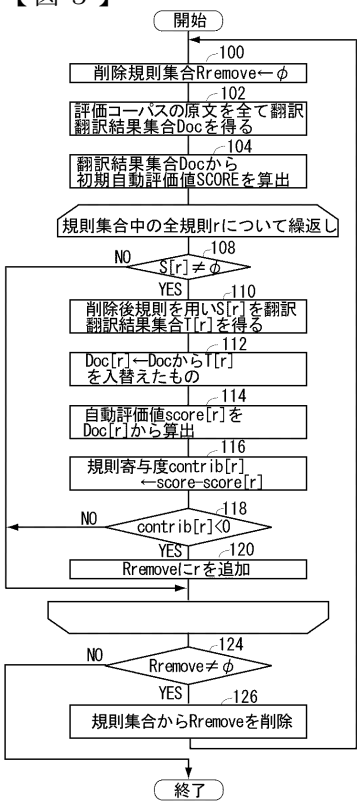
【図３】



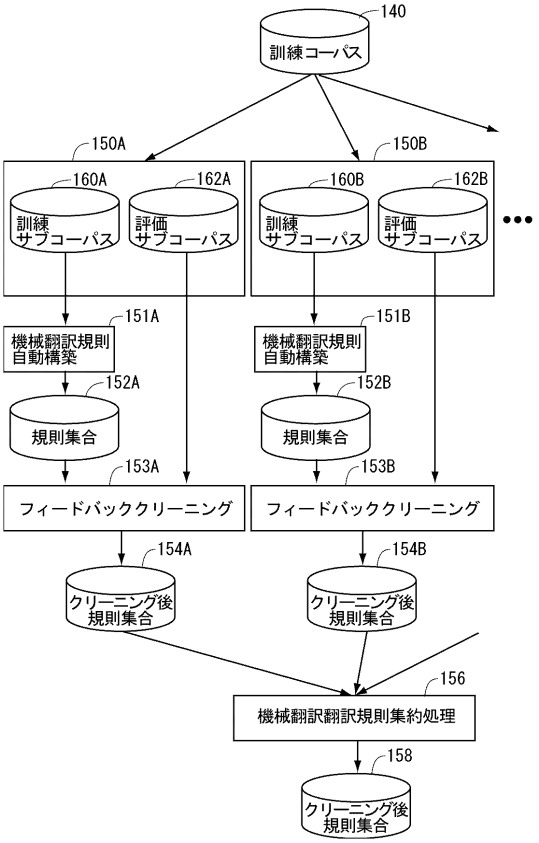
【図４】



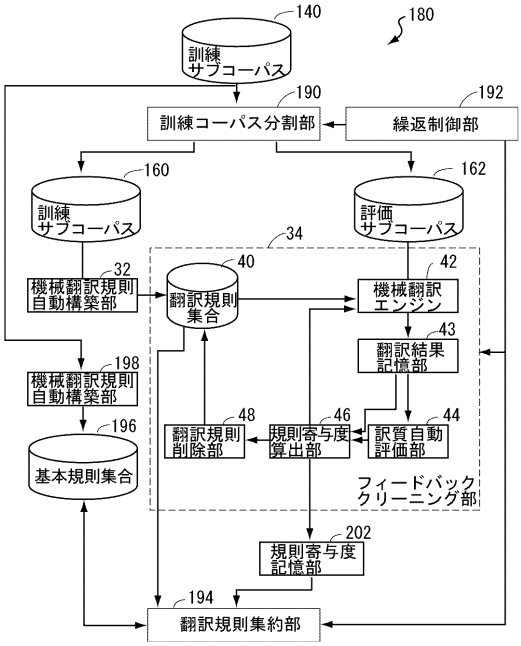
【図５】


開始
100
削除規則集合Rremove←φ
102
評価コーパスの原文を全て翻訳
翻訳結果集合Docを得る
104
翻訳結果集合Docから
初期自動評価値SCOREを算出
規則集合中の全規則rについて繰返し
108
NO
S[r]≠φ
YES
110
削除後規則を用いS[r]を翻訳
翻訳結果集合T[r]を得る
112
Doc[r]←DocからT[r]
を入替えたもの
114
自動評価値score[r]を
Doc[r]から算出
116
規則寄与度contrib[r]
←score-score[r]
118
NO
contrib[r]<0
YES
120
Rremoveにrを追加
124
NO
Rremove≠φ
YES
126
規則集合からRremoveを削除
終了

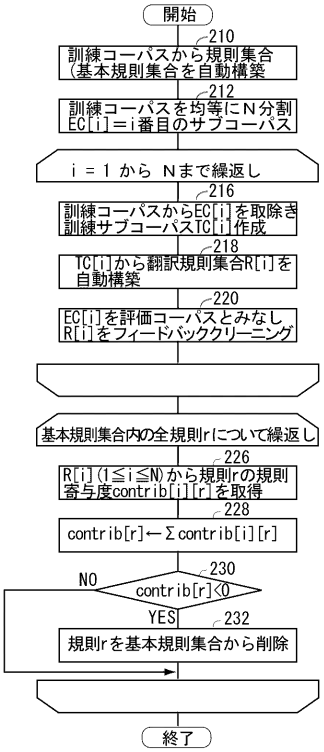
【図６】


140
訓練コーパス
150A
150B
160A
訓練
サブコーパス
162A
評価
サブコーパス
160B
訓練
サブコーパス
162B
評価
サブコーパス
151A
機械翻訳規則
自動構築
151B
機械翻訳規則
自動構築
152A
規則集合
152B
規則集合
153A
フィードバッククリーニング
153B
フィードバッククリーニング
154A
クリーニング後
規則集合
154B
クリーニング後
規則集合
156
機械翻訳翻訳規則集約処理
158
クリーニング後
規則集合

【図７】


140
180
訓練
サブコーパス
190
訓練コーパス分割部
192
繰返制御部
160
訓練
サブコーパス
162
評価
サブコーパス
34
32
機械翻訳規則
自動構築部
40
翻訳規則
集合
42
機械翻訳
エンジン
43
翻訳結果
記憶部
198
機械翻訳規則
自動構築部
48
翻訳規則
削除部
46
規則寄与度
算出部
44
訳質自動
評価部
196
基本規則集合
フィードバック
クリーニング部
202
規則寄与度
記憶部
194
翻訳規則集約部

【図８】


開始
210
訓練コーパスから規則集合
(基本規則集合を自動構築
212
訓練コーパスを均等にＮ分割
EC[i]＝i番目のサブコーパス
ｉ＝１から Ｎまで繰返し
216
訓練コーパスからEC[i]を取除き
訓練サブコーパスTC[i]作成
218
TC[i]から翻訳規則集合R[i]を
自動構築
220
EC[i]を評価コーパスとみなし
R[i]をフィードバッククリーニング
基本規則集合内の全規則rについて繰返し
226
R[i](1≦i≦N)から規則rの規則
寄与度contrib[i][r]を取得
228
contrib[r]←Σcontrib[i][r]
230
NO
contrib[r]<0
YES
232
規則rを基本規則集合から削除
終了

【手続補正書】
【提出日】平成16年6月24日(2004.6.24)
【手続補正１】
【補正対象書類名】明細書
【補正対象項目名】００３５
【補正方法】変更
【補正の内容】
【００３５】
本発明の第３の局面に係るコンピュータは、上記したコンピュータプログラムによりプログラムされたコンピュータである。
【手続補正２】
【補正対象書類名】明細書
【補正対象項目名】００３８
【補正方法】変更
【補正の内容】
【００３８】
なお以下の説明では、第１及び第２の実施の形態を説明する。これらの実施の形態の基本的な考え方は以下の通りである。すなわち、自動構築された翻訳規則を用いて評価コーパス中の原言語の文を機械翻訳する。機械翻訳した結果に対し、後掲の参考文献２に記載されている様な訳質の自動評価を行ない、自動評価値を得る。この自動評価値を向上させる様に翻訳規則の取捨選択を行なう事により、最適な翻訳規則の組合せ（最適な翻訳規則集合）を得る。
【手続補正３】
【補正対象書類名】明細書
【補正対象項目名】００４３
【補正方法】変更
【補正の内容】
【００４３】
フィードバッククリーニング部３４は、機械翻訳規則自動構築部３２により訓練コーパス３０から自動的に構築された翻訳規則の集合を記憶するための翻訳規則集合記憶部４０と、翻訳規則集合記憶部４０に記憶された翻訳規則を用いて評価コーパス３６中の全ての英語の原文を目的言語の文に翻訳するための機械翻訳エンジン４２とを含む。機械翻訳エンジン４２は構文トランスファ方式のものである。
【手続補正４】
【補正対象書類名】明細書
【補正対象項目名】００４４
【補正方法】変更
【補正の内容】
【００４４】
フィードバッククリーニング部３４はさらに、機械翻訳エンジン４２による翻訳結果を、各文の翻訳の際に使用された翻訳規則を特定する情報とともに記憶するための翻訳結果記憶部４３を含む。翻訳結果記憶部４３はまた、翻訳結果とともに各文の翻訳の際に使用された翻訳規則を特定する情報も記憶する。
【手続補正５】
【補正対象書類名】明細書
【補正対象項目名】０１２４
【補正方法】変更
【補正の内容】
【０１２４】
次に、繰返制御部１９２によるＮ回のフィードバッククリーニングにより、訓練コーパス

１４０のＮ個の訓練サブコーパス１６０よりＮ個の翻訳規則集合が得られる。これらを第１の翻訳規則集合、第２の翻訳規則集合、…第Ｎの翻訳規則集合と呼ぶ事とする。そして、これらの翻訳規則集合を作成する際に規則寄与度算出部４６により計算された各規則の規則寄与度が規則寄与度記憶部２０２に繰返しごとに別々に記憶される。規則ｒについてのｉ回目の繰返しの際に計算された規則寄与度をｃｏｎｔｒｉｂ［ｉ］［ｒ］と表す（１≦ｉ≦Ｎ、１≦ｒ≦基本翻訳規則数）。

【手続補正６】
【補正対象書類名】明細書
【補正対象項目名】０１２５
【補正方法】変更
【補正の内容】
【０１２５】
翻訳規則集約部１９４は、全てのフィードバッククリーニングが終了すると、規則寄与度記憶部２０２を参照して、翻訳規則ｒごとに、規則寄与度記憶部２０２に記憶されている規則寄与度の総和$contrib[r]=\Sigma_i cont_r ib[i][r]$を計算する。そして、総和ｃｏｎｔｒｉｂ［ｒ］が負であればその規則ｒを基本規則集合記憶部１９６に記憶されている基本規則集合から削除する。この処理を全ての規則ｒに対して実行する事により、基本規則集合記憶部１９６に記憶されている基本規則集合に対するクリーニングが行なわれ、最終的な交差フィードバッククリーニング後の翻訳規則集合が得られる。

【手続補正７】
【補正対象書類名】明細書
【補正対象項目名】０１３１
【補正方法】変更
【補正の内容】
【０１３１】
翻訳規則集約部１９４は、基本規則集合記憶部１９６に記憶されている各規則について、前述した通り規則寄与度の総和$contrib[r]=\Sigma_i cont_r ib[i][r]$を計算する。ｃｏｎｔｒｉｂ［ｒ］が負の場合、その規則は基本規則集合記憶部１９６内の基本規則集合から削除される。

【手続補正８】
【補正対象書類名】明細書
【補正対象項目名】０１３９
【補正方法】変更
【補正の内容】
【０１３９】
ステップ２２６では、翻訳規則集合Ｒ［ｉ］（１≦ｉ≦Ｎ）から、規則ｒの規則寄与度ｃｏｎｔｒｉｂ［ｉ］［ｒ］を取得する。具体的には、前述した通り図５のステップ１１６で記憶されていた規則寄与度を記憶領域から取出す。ステップ２２８で、基本規則ｒの寄与度$contrib[r]=\Sigma_i cont_r ib[i][r]$を算出する。

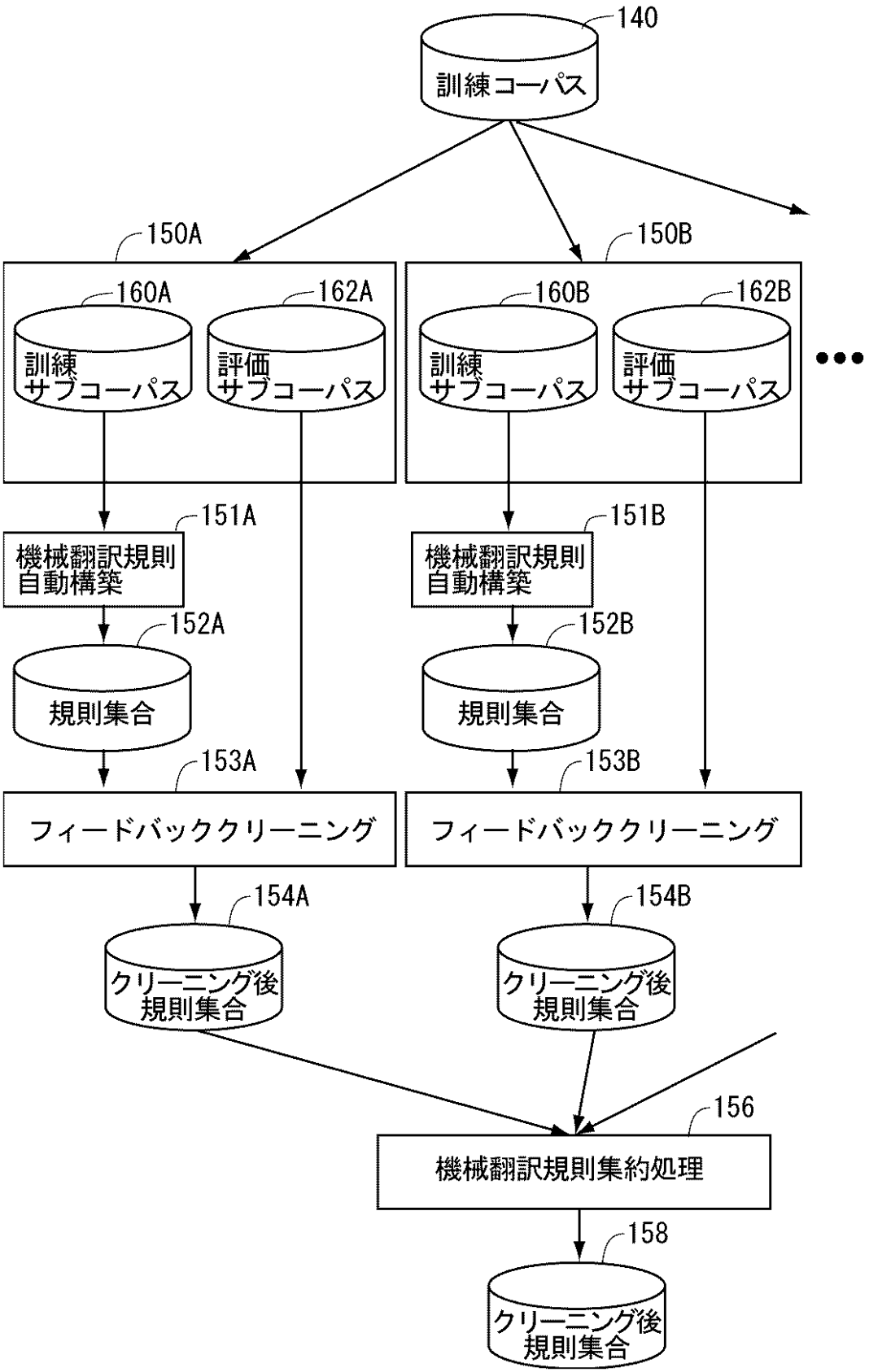
【手続補正９】
【補正対象書類名】明細書
【補正対象項目名】０１４７
【補正方法】変更
【補正の内容】
【０１４７】
なお、ソフトウェアの流通形態は上記した様に記憶媒体に固定された形には限定されない。たとえば、ネットワークを通じて接続された他のコンピュータからデータを受取る形で流通する事もあり得る。また、ソフトウェアの一部が予めハードディスク５４中に格納されており、ソフトウェアの残りの部分をネットワーク経由でハードディスク７４に取込んで実行時に統合する様な形の流通形態もあり得る。

【手続補正１０】
【補正対象書類名】明細書
【補正対象項目名】０１５０
【補正方法】変更
【補正の内容】
【０１５０】
［参考文献２］　パピネニ、Ｋ．，ルーコス、Ｓ．，ウォード、Ｔ，及びツー、Ｗ．－Ｊ．（２００２）．Ｂｌｅｕ：機械翻訳の自動評価方法．第４０回計算言語学学会第４０回年次大会予稿集、３１１頁から３１８頁（Ｐａｐｉｎｅｎｉ，Ｋ．，Ｒｏｕｋｏｓ，Ｓ．，Ｗａｒｄ、Ｔ．，ａｎｄ　Ｚｈｕ，Ｗ．－Ｊ．（２００２）．Ｂｌｅｕ：ａ　ｍｅｔｈｏｄ　ｆｏｒ　ａｕｔｏｍａｔｉｃ　ｅｖａｌｕａｔｉｏｎ　ｏｆ　ｍａｃｈｉｎｅ　ｔｒａｎｓｌａｔｉｏｎ．Ｉｎ　Ｐｒｏｃｅｅｄｉｎｇｓ　ｏｆ　ｔｈｅ　４０ｔｈ　Ａｎｎｕａｌ　Ｍｅｅｔｉｎｇ　ｏｆ　ｔｈｅ　Ａｓｓｏｃｉａｔｉｏｎ　ｆｏｒ　Ｃｏｍｐｕｔａｔｉｏｎａｌ　Ｌｉｎｇｕｉｓｔｉｃｓ（ＡＣＬ），ｐｐ．３１１－３１８）
【手続補正１１】
【補正対象書類名】明細書
【補正対象項目名】０１５２
【補正方法】変更
【補正の内容】
【０１５２】
［参考文献４］　アキバ、Ｙ．，イマムラ、Ｋ．，及びスミタ、Ｅ．，（２００１）（Ａｋｉｂａ，Ｙ．，Ｉｍａｍｕｒａ，Ｋ．，ａｎｄ　Ｓｕｍｉｔａ，Ｅ．，（２００１））．複数編集距離を用いた機械翻訳の自動評価．機械翻訳サミット予稿集ＶＩＩＩ、１５頁から２０頁（Ｕｓｉｎｇ　ｍｕｌｔｉｐｌｅ　ｅｄｉｔ　ｄｉｓｔａｎｃｅｓ　ｔｏ　ａｕｔｏｍａｔｉｃａｌｌｙ　ｒａｎｋ　ｍａｃｈｉｎｅ　ｔｒａｎｓｌａｔｉｏｎ　ｏｕｔｐｕｔ．Ｉｎ　Ｐｒｏｃｅｅｄｉｎｇｓ　ｏｆ　Ｍａｃｈｉｎｅ　Ｔｒａｎｓｌａｔｉｏｎ　Ｓｕｍｍｉｔ　ＶＩＩＩ，ｐｐ．１５－２０）
【手続補正１２】
【補正対象書類名】図面
【補正対象項目名】図２
【補正方法】変更
【補正の内容】
【図２】

規則番号	構文カテゴリ	原言語パターン		目的言語パターン	用例
1	VP	X_{VP} at Y_{NP}	⇒	Y'でX'	((present, conference), ...)
2	VP	X_{VP} at Y_{NP}	⇒	Y'にX'	((stay, hotel),(arrive, p.m.), ...)
3	VP	X_{VP} at Y_{NP}	⇒	Y'をX'	((look,it), ...)
4	NP	X_{NP} at Y_{NP}	⇒	Y'のX'	((man, front desk), ...)
5	NP	the X_{NMP} Y_{NP} station	⇒	X'Y'	((near, subway),...)
6	VP	X_{VP} <num-noun> Y_{NP}	⇒	Y'をX'	((rent, bicycle),...)
7	S	please X_{VP}	⇒	X'たいのですが	((send),(reserve),...)
8	S	please X_{VP}	⇒	X'てください	((give), ...)

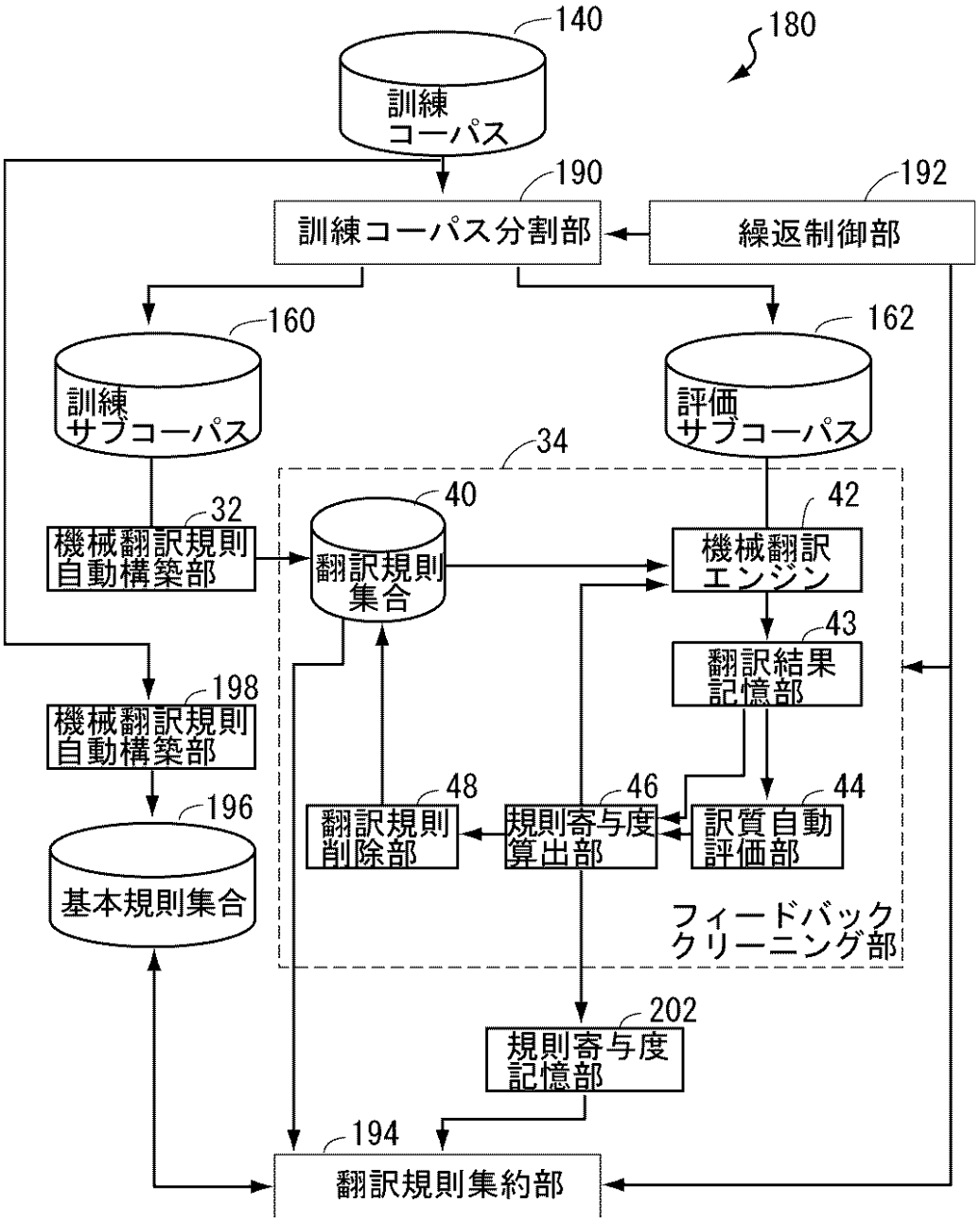
【手続補正１３】
【補正対象書類名】図面
【補正対象項目名】図６
【補正方法】変更

【補正の内容】
【図６】



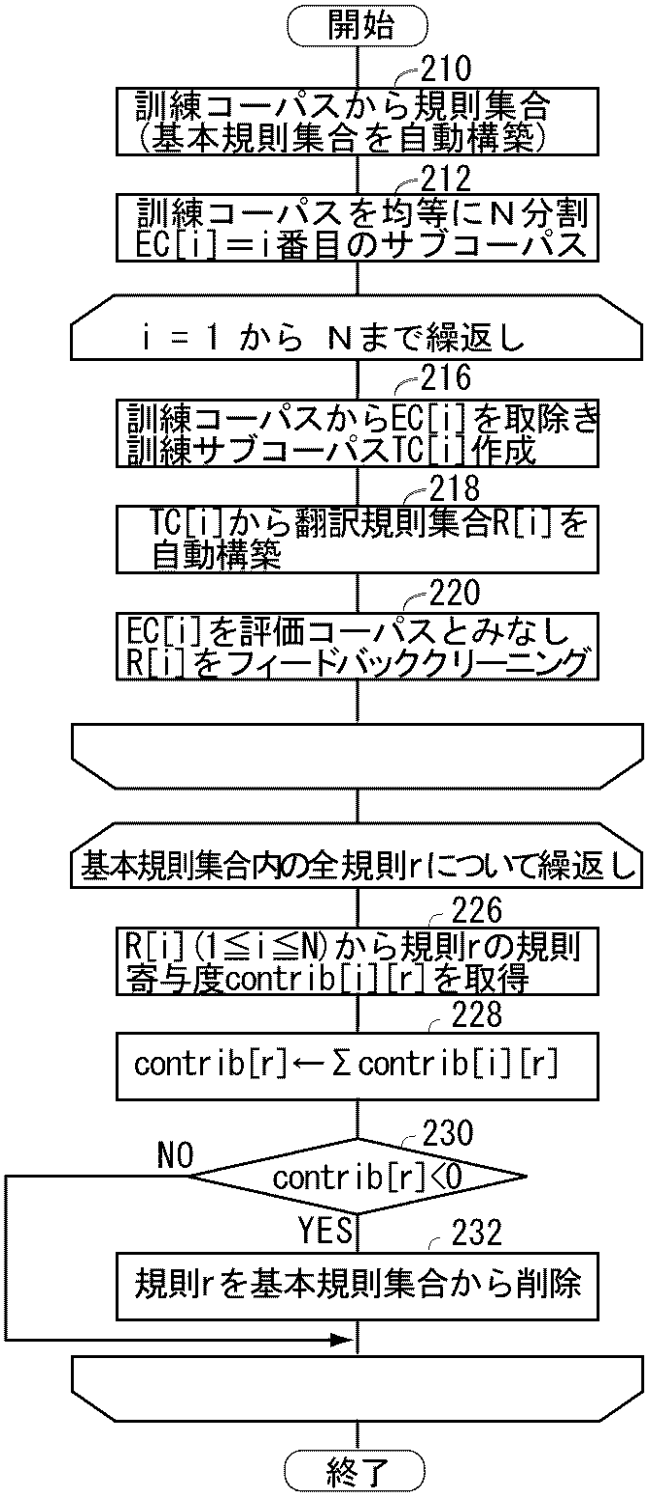
【手続補正１４】
【補正対象書類名】図面
【補正対象項目名】図７
【補正方法】変更

【補正の内容】
【図７】



【手続補正１５】
【補正対象書類名】図面
【補正対象項目名】図８
【補正方法】変更
【補正の内容】

【図８】


開始
210
訓練コーパスから規則集合
(基本規則集合を自動構築)
212
訓練コーパスを均等にＮ分割
EC[i]＝i番目のサブコーパス
ｉ＝１ から Ｎまで繰返し
216
訓練コーパスからEC[i]を取除き
訓練サブコーパスTC[i]作成
218
TC[i]から翻訳規則集合R[i]を
自動構築
220
EC[i]を評価コーパスとみなし
R[i]をフィードバッククリーニング
基本規則集合内の全規則rについて繰返し
226
R[i](1≦i≦N)から規則rの規則
寄与度contrib[i][r]を取得
228
contrib[r]←Σcontrib[i][r]
230
NO
contrib[r]<0
YES
232
規則rを基本規則集合から削除
終了