(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2005-234800
(P2005-234800A)
(43)公開日　平成17年9月2日(2005.9.2)

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

審査請求 未請求　請求項の数 16　ＯＬ　（全 20 頁）

(21)出願番号　特願2004-41751 (P2004-41751)
(22)出願日　平成16年2月18日 (2004.2.18)

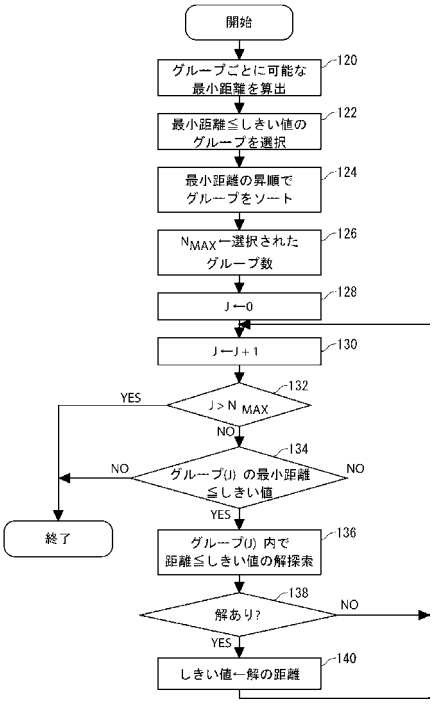
（出願人による申告）平成１５年度通信・放送機構、研究テーマ「大規模コーパスベース音声対話翻訳技術の研究開発」に関する委託研究、産業活力再生特別措置法第３０条の適用を受ける特許出願

(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　土居　誉生
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　山本　博史
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 AA03　AB17　BA04　CA21　CC01
CC05　CC15　EA24

(54)【発明の名称】用例機械翻訳装置及び用例翻訳コンピュータプログラム、並びに用例検索装置及び用例検索コンピュータプログラム

(57)【要約】
【課題】、高速で単語列編集距離を基準にして入力文と所定の関係を持つ用例を高速に検索可能にする。
【解決手段】用例機械翻訳装置で用いられる用例検索装置は、各用例に含まれる第１の言語の単語列に含まれる内容語数及び機能語数に基づいて複数のグループに分割されたコーパスから、第１の言語の入力文に最も近い用例を検索するために、各グループに含まれる第１の言語の単語列と入力文との間で可能な最小の単語列編集距離を算出する最小距離算出部１２０と、距離の下限値がしきい値以下のものを選択するグループ選択部１２２と、選択されたグループに含まれる用例の中で、単語列編集距離により定義される入力文との距離が最小の解を探索する探索部１２４－１４０とを含む。探索にはＡ＊アルゴリズムを使用しても良い。
【選択図】　図４



【特許請求の範囲】
【請求項１】
コンピュータ読取可能な、第１及び第２の言語の対訳関係にある用例からなる用例コーパスを記憶するための第１の記憶手段と、
それぞれコンピュータ読取可能な、前記第１及び第２の言語の対訳辞書並びに前記第１及び第２の言語のシソーラスを記憶するための第２の記憶手段と、
前記第１の言語の入力文を受け、前記シソーラスを用いて前記用例コーパスから前記入力文と所定の関係を有する用例を検索するための用例検索手段と、
前記用例検索手段により検索された用例と、前記対訳辞書及び前記シソーラスとを用いて前記第１の言語の入力文を前記第２の言語の文に翻訳するための翻訳手段とを含み、
前記用例コーパスは、各用例に含まれる前記第１の言語の単語列に含まれる内容語数及び機能語数に基づいて複数のグループに分割され、
前記用例検索手段は、
前記複数のグループの各々について、予め定義された単語列編集距離を用い、当該グループに含まれる前記第１の言語の単語列と前記入力文との間で距離の下限値を算出するための最小距離算出手段と、
前記複数のグループのうち、前記最小距離算出手段により算出された距離の下限値が所定のしきい値以下のもののみを選択するためのグループ選択手段と、
前記グループ選択手段により選択されたグループに含まれる用例の中で、前記単語列編集距離により定義される入力文との距離が最小の解を探索するための探索手段とを含む、用例機械翻訳装置。
【請求項２】
前記用例検索手段はさらに、前記探索手段により解が見出されたことに応答して、前記探索手段による探索と並行して、前記所定のしきい値を前記見出された解の距離で置換するための手段を含む、請求項１に記載の用例機械翻訳装置。
【請求項３】
前記複数のグループの各々に含まれる用例の、前記第１の言語の単語列は、一つの単語グラフ形式で表現され、
前記探索手段は、前記単語グラフの先頭ノードから最終ノードまでの可能な全経路について、当該経路に現れる単語列と入力単語列との照合を行なうことによって、前記入力単語列との間の単語列編集距離が最小となる経路を選択するための経路選択手段を含む、請求項１又は請求項２に記載の用例機械翻訳装置。
【請求項４】
前記経路選択手段は、各グループに含まれる単語グラフの各経路に現れる単語列と、入力文との間の照合の途中経過を状態とする問題状態集合の中から、Ａ＊アルゴリズムを用いて単語列編集距離を最小にする目標状態を探索するための手段を含む、請求項３に記載の用例機械翻訳装置。
【請求項５】
コンピュータ読取可能な、第１及び第２の言語の対訳関係にある用例からなる用例コーパスを記憶するための第１の記憶装置と、それぞれコンピュータ読取可能な、前記第１及び第２の言語の対訳辞書並びに前記第１及び第２の言語のシソーラスを記憶するための第２の記憶装置とを備えたコンピュータ上で実行されると、当該コンピュータを、前記用例コーパスを用いた用例機械翻訳装置として動作させる、用例翻訳コンピュータプログラムであって、
前記用例機械翻訳装置は、
前記第１の言語の入力文を受け、前記シソーラスを用いて前記用例コーパスから前記入力文と所定の関係を有する用例を検索するための用例検索手段と、
前記用例検索手段により検索された用例と、前記対訳辞書及び前記シソーラスとを用いて前記第１の言語の入力文を前記第２の言語の文に翻訳するための翻訳手段とを含み、
前記用例コーパスは、各用例に含まれる前記第１の言語の単語列に含まれる内容語数及

び機能語数に基づいて複数のグループに分割され、
前記用例検索手段は、
前記複数のグループの各々について、予め定義された単語列編集距離を用い、当該グループに含まれる前記第１の言語の単語列と前記入力文との間で距離の下限値を算出するための最小距離算出手段と、
前記複数のグループのうち、前記最小距離算出手段により算出された距離の下限値が所定のしきい値以下のもののみを選択するためのグループ選択手段と、
前記グループ選択手段により選択されたグループに含まれる用例の中で、前記単語列編集距離により定義される入力文との距離が最小の解を探索するための探索手段とを含む、用例翻訳コンピュータプログラム。

【請求項６】
前記用例検索手段はさらに、前記探索手段により解が見出されたことに応答して、前記探索手段による探索と並行して、前記所定のしきい値を前記見出された解の距離で置換するための手段を含む、請求項５に記載の用例翻訳コンピュータプログラム。

【請求項７】
前記複数のグループの各々に含まれる用例の、前記第１の言語の単語列は、一つの単語グラフ形式で表現され、
前記探索手段は、前記単語グラフの先頭ノードから最終ノードまでの可能な全経路について、当該経路に現れる単語列と入力単語列との照合を行なうことによって、前記入力単語列との間の単語列編集距離が最小となる経路を選択するための経路選択手段を含む、請求項５又は請求項６に記載の用例翻訳コンピュータプログラム。

【請求項８】
前記経路選択手段は、各グループに含まれる単語グラフの各経路に現れる単語列と、入力文との間の照合の途中経過を状態とする問題状態集合の中から、Ａ＊アルゴリズムを用いて単語列編集距離を最小にする目標状態を探索するための手段を含む、請求項７に記載の用例翻訳コンピュータプログラム。

【請求項９】
コンピュータ読取可能な、第１及び第２の言語の対訳関係にある用例からなる用例コーパスを記憶するための第１の記憶手段と、それぞれコンピュータ読取可能な、前記第１及び第２の言語の対訳辞書並びに前記第１及び第２の言語のシソーラスを記憶するための第２の記憶手段と、前記用例コーパスから検索された用例と、前記対訳辞書及び前記シソーラスとを用いて前記第１の言語の入力文を前記第２の言語の文に翻訳するための翻訳手段とを含む用例機械翻訳装置で使用され、前記第１の言語の入力文を受け、前記シソーラスを用いて前記用例コーパスから前記入力文と所定の関係を有する用例を検索するための用例検索装置であって、
前記用例コーパスは、各用例に含まれる前記第１の言語の単語列に含まれる内容語数及び機能語数に基づいて複数のグループに分割され、
前記用例検索装置は、
前記複数のグループの各々について、予め定義された単語列編集距離を用い、当該グループに含まれる前記第１の言語の単語列と前記入力文との間で距離の下限値を算出するための最小距離算出手段と、
前記複数のグループのうち、前記最小距離算出手段により算出された距離の下限値が所定のしきい値以下のもののみを選択するためのグループ選択手段と、
前記グループ選択手段により選択されたグループに含まれる用例の中で、前記単語列編集距離により定義される入力文との距離が最小の解を探索するための探索手段とを含む、用例検索装置。

【請求項１０】
さらに、前記探索手段により解が見出されたことに応答して、前記探索手段による探索と並行して、前記所定のしきい値を前記見出された解の距離で置換するための手段を含む、請求項９に記載の用例検索装置。

【請求項１１】
前記複数のグループの各々に含まれる用例の、前記第１の言語の単語列は、一つの単語グラフ形式で表現され、
前記探索手段は、前記単語グラフの先頭ノードから最終ノードまでの可能な全経路について、当該経路に現れる単語列と入力単語列との照合を行なうことによって、前記入力単語列との間の単語列編集距離が最小となる経路を選択するための経路選択手段を含む、請求項９又は請求項１０に記載の用例検索装置。
【請求項１２】
前記経路選択手段は、各グループに含まれる単語グラフの各経路に現れる単語列と、入力文との間の照合の途中経過を状態とする問題状態集合の中から、Ａ＊アルゴリズムを用いて単語列編集距離を最小にする目標状態を探索するための手段を含む、請求項１１に記載の用例検索装置。
【請求項１３】
コンピュータ読取可能な、第１及び第２の言語の対訳関係にある用例からなる用例コーパスを記憶するための第１の記憶装置と、それぞれコンピュータ読取可能な、前記第１及び第２の言語の対訳辞書並びに前記第１及び第２の言語のシソーラスを記憶するための第２の記憶装置と、前記対訳コーパスから検索された用例を用いて、前記第１の言語の入力文を前記第２の言語の文に翻訳する用例機械翻訳手段とを備えたコンピュータ上で実行されると、当該コンピュータを、前記第１の言語の入力文を受け、前記シソーラスを用いて前記用例コーパスから前記入力文と所定の関係を有する用例を検索する用例検索装置として動作させる、用例検索コンピュータプログラムであって、
前記用例コーパスは、各用例に含まれる前記第１の言語の単語列に含まれる内容語数及び機能語数に基づいて複数のグループに分割され、
前記用例検索装置は、
前記複数のグループの各々について、予め定義された単語列編集距離を用い、当該グループに含まれる前記第１の言語の単語列と前記入力文との間で距離の下限値を算出するための最小距離算出手段と、
前記複数のグループのうち、前記最小距離算出手段により算出された距離の下限値が所定のしきい値以下のもののみを選択するためのグループ選択手段と、
前記グループ選択手段により選択されたグループに含まれる用例の中で、前記単語列編集距離により定義される入力文との距離が最小の解を探索するための探索手段とを含む、用例検索コンピュータプログラム。
【請求項１４】
前記用例検索装置はさらに、前記探索手段により解が見出されたことに応答して、前記探索手段による探索と並行して、前記所定のしきい値を前記見出された解の距離で置換するための手段を含む、請求項１３に記載の用例検索コンピュータプログラム。
【請求項１５】
前記複数のグループの各々に含まれる用例の、前記第１の言語の単語列は、一つの単語グラフ形式で表現され、
前記探索手段は、前記単語グラフの先頭ノードから最終ノードまでの可能な全経路について、当該経路に現れる単語列と入力単語列との照合を行なうことによって、前記入力単語列との間の単語列編集距離が最小となる経路を選択するための経路選択手段を含む、請求項１３又は請求項１４に記載の用例検索コンピュータプログラム。
【請求項１６】
前記経路選択手段は、各グループに含まれる単語グラフの各経路に現れる単語列と、入力文との間の照合の途中経過を状態とする問題状態集合の中から、Ａ＊アルゴリズムを用いて単語列編集距離を最小にする目標状態を探索するための手段を含む、請求項１５に記載の用例検索コンピュータプログラム。

【発明の詳細な説明】

【技術分野】
【０００１】
この発明は機械翻訳装置に関し、特に、用例を用いて機械翻訳を行なう用例機械翻訳装置並びに当該装置において用例を高速に検索するための用例検索装置に関する。
【背景技術】
【０００２】
アナロジーに基づく機械翻訳の概念が提唱されて以来、このアイデアを具体化した用例翻訳が数多く提案されてきた。用例翻訳では、あらかじめ二つの言語で同じ意味を表す対訳からなる対訳コーパスを準備する。そしてこの対訳コーパスから対訳表現を抽出し、翻訳すべき文を構成する各部分に適合する対訳表現を見つけ、それらを組み合わせて翻訳文を生成する。
【０００３】
一方、用例翻訳以前の代表的な翻訳方式としてルールベース翻訳がある。ルールベース翻訳では、言語現象について人間が内省することにより作成した翻訳のためのルールに基づいて翻訳を実行する。
【０００４】
ルールベース翻訳では、翻訳ルールの精密な構築が翻訳システムの能力を決める。これに対し、用例翻訳では対訳コーパスからの学習によって能力が決定する。用例翻訳の大きな利点は、人手による翻訳ルールの記述が不要であり、翻訳システムの保守を含めた開発の効率が高いことである。この特長により、異なるドメインへの移植や新たな言語対への適用が容易になる。
【０００５】
特許文献１において、用例翻訳の一つとしてＤＰ（Ｄｙｎａｍｉｃ　Ｐｒｏｇｒａｍｍｉｎｇ）マッチに基づいた翻訳方式が提案されている。
【０００６】
この方式では、用例は原言語文と目的言語文との対である。各文は単語列として表現される。翻訳実行時には入力文との類似度の最も高いと判定される原言語文を持つ用例を検索する。検索された用例の原言語文と入力文との差異、及び差異部分と対応する目的言語文の部分を求め、翻訳パターンを生成する。
【０００７】
他の用例翻訳方式と比べたときのこの方式の特徴は、以下の通りである。
【０００８】
（１）多くの方式ではあらかじめ翻訳パターンを作成しているが、この方式では翻訳実行時に翻訳パターンを作成する。
【０００９】
（２）多くの方式では構文解析又はツリーバンクの利用を仮定しているが、この方式ではそれらを利用しない。
【００１０】
このように特許文献１において提案された翻訳方式では、用例を抽象化せず単語列の形のまま保持し、検索する。さらに訳文を生成する際にも用例の目的言語文の変更を最低限にとどめる。そのため、入力文に近い用例が存在すれば自然な表現の翻訳結果が得られるという特徴がある。また、多くの言語に高精度パーザを期待することができない現状を考えると、この方式は解析知識を使わないため用例翻訳方式の中でも特に多言語に適用することが容易であり、その点で他方式と比較して優れるものと考えられる。
【００１１】
【特許文献１】特開２００３－６１９３号公報
【特許文献２】特開平８－１８５４８２号公報
【非特許文献１】ラップ　Ｒ．、「翻訳メモリのための品詞に基づく検索アルゴリズム」、ＬＲＥＣ２００２予稿集、ｐｐ．４６６－４７２、２００２年（Ｒａｐｐ，Ｒ．：　Ａ　Ｐａｒｔ－ｏｆ－Ｓｐｅｅｃｈ－Ｂａｓｅｄ　Ｓｅａｒｃｈ　Ａｌｇｏｒｉｔｈｍ　ｆ

ｏｒ Ｔｒａｎｓｌａｔｉｏｎ Ｍｅｍｏｒｉｅｓ、Ｐｒｏｃ． ｏｆ ＬＲＥＣ ２００２、ｐｐ．４６６－４７２、２００２）

【発明の開示】

【発明が解決しようとする課題】

【００１２】

特許文献１に記載の用例翻訳方式では、記憶した文の集合の中から最も入力文に類似したものを選び出す仕組みがその中核にある。特に大量の用例を利用する場合、その仕組みの効率的な実装が必須となる。

【００１３】

用例検索処理の効率的な実装は、翻訳メモリの課題と共通である。翻訳メモリに関する非特許文献１に記載の研究では、品詞レベルでの文の完全一致に基づいた検索が提案されている。しかし、この用例翻訳方式では単語列の間の編集距離を基準にして用例を検索している。そのために、用例の検索では単語の挿入及び削除をも考慮しなければならない。そのため非特許文献１において提案された方式は用例翻訳方式には採用できない。

【００１４】

特許文献２では、編集距離最小文を検索する方法を提示しているが、個々の候補に対する計算を繰返すので大量の用例を使った場合の高速化には限界があると考えられる。

【００１５】

それゆえにこの発明の目的は、高速で単語列編集距離を基準にして用例を検索することができる用例検索装置及びそれを用いた用例機械翻訳装置、ならびにそれらのコンピュータプログラムを提供することである。

【課題を解決するための手段】

【００１６】

本発明の第１の局面に係る用例機械翻訳装置は、コンピュータ読取可能な、第１及び第２の言語の対訳関係にある用例からなる用例コーパスを記憶するための第１の記憶手段と、それぞれコンピュータ読取可能な、第１及び第２の言語の対訳辞書並びに第１及び第２の言語のシソーラスを記憶するための第２の記憶手段と、第１の言語の入力文を受け、シソーラスを用いて用例コーパスから入力文と所定の関係を有する用例を検索するための用例検索手段と、用例検索手段により検索された用例と、対訳辞書及びシソーラスとを用いて第１の言語の入力文を第２の言語の文に翻訳するための翻訳手段とを含み、用例コーパスは、各用例に含まれる第１の言語の単語列に含まれる内容語数及び機能語数に基づいて複数のグループに分割され、用例検索手段は、複数のグループの各々について、予め定義された単語列編集距離を用い、当該グループに含まれる第１の言語の単語列と入力文との間で距離の下限値を算出するための最小距離算出手段と、複数のグループのうち、最小距離算出手段により算出された距離の下限値が所定のしきい値以下のもののみを選択するためのグループ選択手段と、グループ選択手段により選択されたグループに含まれる用例の中で、単語列編集距離により定義される入力文との距離が最小の解を探索するための探索手段とを含む。

【００１７】

好ましくは、用例検索手段はさらに、探索手段により解が見出されたことに応答して、探索手段による探索と並行して、所定のしきい値を見出された解の距離で置換するための手段を含む。

【００１８】

さらに好ましくは、複数のグループの各々に含まれる用例の、第１の言語の単語列は、一つの単語グラフ形式で表現されており、探索手段は、単語グラフの先頭ノードから最終ノードまでの可能な全経路について、当該経路に現れる単語列と入力単語列との照合を行なうことによって、入力単語列との間の単語列編集距離が最小となる経路を選択するための経路選択手段を含む。

【００１９】

経路選択手段は、各グループに含まれる単語グラフの各経路に現れる単語列と、入力文

との間の照合の途中経過を状態とする問題状態集合の中から、Ａ＊アルゴリズムを用いて単語列編集距離を最小にする目標状態を探索するための手段を含んでもよい。

【００２０】

本発明の第２の局面に係る用例翻訳コンピュータプログラムは、コンピュータ読取可能な、第１及び第２の言語の対訳関係にある用例からなる用例コーパスを記憶するための第１の記憶装置と、それぞれコンピュータ読取可能な、第１及び第２の言語の対訳辞書並びに第１及び第２の言語のシソーラスを記憶するための第２の記憶装置とを備えたコンピュータ上で実行されると、当該コンピュータを、用例コーパスを用いた用例機械翻訳装置として動作させる、用例翻訳コンピュータプログラムである。このプログラムにより実現される用例機械翻訳装置は、第１の言語の入力文を受け、シソーラスを用いて用例コーパスから入力文と所定の関係を有する用例を検索するための用例検索手段と、用例検索手段により検索された用例と、対訳辞書及びシソーラスとを用いて第１の言語の入力文を第２の言語の文に翻訳するための翻訳手段とを含み、用例コーパスは、各用例に含まれる第１の言語の単語列に含まれる内容語数及び機能語数に基づいて複数のグループに分割され、用例検索手段は、複数のグループの各々について、予め定義された単語列編集距離を用い、当該グループに含まれる第１の言語の単語列と入力文との間で距離の下限値を算出するための最小距離算出手段と、複数のグループのうち、最小距離算出手段により算出された距離の下限値が所定のしきい値以下のもののみを選択するためのグループ選択手段と、グループ選択手段により選択されたグループに含まれる用例の中で、単語列編集距離により定義される入力文との距離が最小の解を探索するための探索手段とを含む。

【００２１】

好ましくは、用例検索手段はさらに、探索手段により解が見出されたことに応答して、探索手段による探索と並行して、所定のしきい値を見出された解の距離で置換するための手段を含む。

【００２２】

さらに好ましくは、複数のグループの各々に含まれる用例の、第１の言語の単語列は、一つの単語グラフ形式で表現され、探索手段は、単語グラフの先頭ノードから最終ノードまでの可能な全経路について、当該経路に現れる単語列と入力単語列との照合を行なうことによって、入力単語列との間の単語列編集距離が最小となる経路を選択するための経路選択手段を含む。

【００２３】

経路選択手段は、各グループに含まれる単語グラフの各経路に現れる単語列と、入力文との間の照合の途中経過を状態とする問題状態集合の中から、Ａ＊アルゴリズムを用いて単語列編集距離を最小にする目標状態を探索するための手段を含んでもよい。

【００２４】

本発明の第３の局面に係る用例検索装置は、コンピュータ読取可能な、第１及び第２の言語の対訳関係にある用例からなる用例コーパスを記憶するための第１の記憶手段と、それぞれコンピュータ読取可能な、第１及び第２の言語の対訳辞書並びに第１及び第２の言語のシソーラスを記憶するための第２の記憶手段と、用例コーパスから検索された用例と、対訳辞書及びシソーラスとを用いて第１の言語の入力文を第２の言語の文に翻訳するための翻訳手段とを含む用例機械翻訳装置で使用され、第１の言語の入力文を受け、シソーラスを用いて用例コーパスから入力文と所定の関係を有する用例を検索するための用例検索装置である。用例コーパスは、各用例に含まれる第１の言語の単語列に含まれる内容語数及び機能語数に基づいて複数のグループに分割されている。用例検索装置は、複数のグループの各々について、予め定義された単語列編集距離を用い、当該グループに含まれる第１の言語の単語列と入力文との間で距離の下限値を算出するための最小距離算出手段と、複数のグループのうち、最小距離算出手段により算出された距離の下限値が所定のしきい値以下のもののみを選択するためのグループ選択手段と、グループ選択手段により選択されたグループに含まれる用例の中で、単語列編集距離により定義される入力文との距離が最小の解を探索するための探索手段とを含む。

【００２５】
好ましくは、用例検索装置はさらに、探索手段により解が見出されたことに応答して、探索手段による探索と並行して、所定のしきい値を見出された解の距離で置換するための手段を含む。
【００２６】
さらに好ましくは、複数のグループの各々に含まれる用例の、第１の言語の単語列は、一つの単語グラフ形式で表現され、探索手段は、単語グラフの先頭ノードから最終ノードまでの可能な全経路について、当該経路に現れる単語列と入力単語列との照合を行なうことによって、入力単語列との間の単語列編集距離が最小となる経路を選択するための経路選択手段を含む。
【００２７】
経路選択手段は、各グループに含まれる単語グラフの各経路に現れる単語列と、入力文との間の照合の途中経過を状態とする問題状態集合の中から、Ａ＊アルゴリズムを用いて単語列編集距離を最小にする目標状態を探索するための手段を含んでもよい。
【００２８】
本発明の第４の局面に係る用例検索コンピュータプログラムは、コンピュータ読取可能な、第１及び第２の言語の対訳関係にある用例からなる用例コーパスを記憶するための第１の記憶装置と、それぞれコンピュータ読取可能な、第１及び第２の言語の対訳辞書並びに第１及び第２の言語のシソーラスを記憶するための第２の記憶装置と、対訳コーパスから検索された用例を用いて、第１の言語の入力文を第２の言語の文に翻訳する用例機械翻訳手段とを備えたコンピュータ上で実行されると、当該コンピュータを、第１の言語の入力文を受け、シソーラスを用いて用例コーパスから入力文と所定の関係を有する用例を検索する用例検索装置として動作させる、用例検索コンピュータプログラムである。用例コーパスは、各用例に含まれる第１の言語の単語列に含まれる内容語数及び機能語数に基づいて複数のグループに分割されている。当該用例検索装置は、複数のグループの各々について、予め定義された単語列編集距離を用い、当該グループに含まれる第１の言語の単語列と入力文との間で距離の下限値を算出するための最小距離算出手段と、複数のグループのうち、最小距離算出手段により算出された距離の下限値が所定のしきい値以下のもののみを選択するためのグループ選択手段と、グループ選択手段により選択されたグループに含まれる用例の中で、単語列編集距離により定義される入力文との距離が最小の解を探索するための探索手段とを含む。
【００２９】
好ましくは、用例検索装置はさらに、探索手段により解が見出されたことに応答して、探索手段による探索と並行して、所定のしきい値を見出された解の距離で置換するための手段を含む。
【００３０】
さらに好ましくは、複数のグループの各々に含まれる用例の、第１の言語の単語列は、一つの単語グラフ形式で表現され、探索手段は、単語グラフの先頭ノードから最終ノードまでの可能な全経路について、当該経路に現れる単語列と入力単語列との照合を行なうことによって、入力単語列との間の単語列編集距離が最小となる経路を選択するための経路選択手段を含む。
【００３１】
経路選択手段は、各グループに含まれる単語グラフの各経路に現れる単語列と、入力文との間の照合の途中経過を状態とする問題状態集合の中から、Ａ＊アルゴリズムを用いて単語列編集距離を最小にする目標状態を探索するための手段を含んでもよい。
【発明を実施するための最良の形態】
【００３２】
－構成－
図１に、本発明の一実施の形態に係る用例に基づく用例機械翻訳装置３０のブロック図を示す。図１を参照して、この用例機械翻訳装置３０は、言語資源としてコンピュータ読

取可能な対訳コーパス４２、第１の言語のシソーラス４４、第２の言語のシソーラス４５及び対訳辞書５０を含む。これらはいずれもハードディスク等の記憶装置に格納される。

【００３３】

対訳コーパス４２は、翻訳方向における原言語と目的言語との文の対の集合である。両言語の文は互いに対訳関係にある。対訳コーパス４２中の文は単語に分割されそれぞれ品詞情報が付与されている。

【００３４】

この用例機械翻訳装置３０では、特許文献１に記載の方式と同様に、対訳コーパス４２中の対訳関係にある文のペアを利用して翻訳を実行する。以下、この文のペアを「用例」と呼ぶ。

【００３５】

対訳辞書５０は、後述するように翻訳パターンの抽出と訳語の置換処理において使用される。

【００３６】

シソーラスとしては、第１の言語（原言語）のシソーラス４４と第２の言語（目的言語）のシソーラス４５とが用意されている。シソーラスは、単語を単語間の意味の近さに基づいてツリー状の階層関係に配置したものである。第１の言語のシソーラス４４は、用例検索及び翻訳パターン抽出処理において使用される。第２の言語の言語のシソーラス４５は、翻訳パターンの抽出に用いられる。

【００３７】

用例機械翻訳装置３０はさらに、原言語の入力文４０を受け、対訳コーパス４２及びシソーラス４４を用いて、対訳コーパス４２の中で入力文４０と最も類似した原言語の文を持つ用例を検索するための用例検索部４６と、用例検索部４６により検索された用例から、第１のシソーラス４４、第２のシソーラス４５及び対訳辞書５０を参照して翻訳パターンを抽出するための翻訳パターン抽出部４８と、用例検索部４６により検索された用例が複数ある場合、翻訳パターン抽出部４８から出力される複数の翻訳パターンの中で所定のものを選択するための翻訳パターン選択部５２と、翻訳パターン選択部５２により選択された翻訳パターンの中で、変数に束縛された単語の訳語を対訳辞書５０から引き、その訳語でもって目的言語パターンの変数を具体化する処理を行なって出力文５６を生成するための訳語置換部５４とを含む。

【００３８】

以下簡単に用例検索部４６、翻訳パターン抽出部４８、翻訳パターン選択部５２及び訳語置換部５４について説明する。用例検索部４６の詳細については、後にさらに詳述する。

【００３９】

用例検索部４６は、全用例の原言語文を走査する。入力文と用例原言語文の単語列間の距離を測り、最小距離の用例を選び出す。ただしこの最小距離が大きければ、検索された用例は翻訳処理に有用ではない。そのため距離にしきい値を設ける。しきい値以内の距離の用例が存在しなければ用例検索及び翻訳処理は失敗に終わる。

【００４０】

単語列間の距離には意味距離の加味された単語列編集距離が使われる。この単語列編集距離ｄｉｓｔは次の式で表される。

【００４１】

【数１】

$$dist = \frac{I + D + 2\sum SEMDIST}{L_{input} + L_{example}}$$

ここでL_{input}は入力文の単語数、$L_{example}$は用例原言語文の単語数、Ｉは用例原言語文を入力文に変換するために必要な挿入単語数、Ｄは同じく必要な削除単語数、ＳＥＭＤ

ＩＳＴは同じく必要な置換により置換される語の間の意味距離を示す。

【００４２】

この式に従って、挿入語と削除語の数、及び置換語の意味距離が足し合わされ、入力文と用例原言語文の長さの和でもって正規化して単語列編集距離ｄｉｓｔが算出される。２単語が同品詞の内容語である場合のみ置換の対象となる。この場合には２単語間の意味距離が単語列編集距離の計算に使われる。意味距離計算においては、二つの単語に関して第１の言語のシソーラス４４中の概念階層における意味概念間の位置関係によって意味距離を計算する。意味距離は０～１までを値域とし、０に近いほど２単語が意味的に類似していることを示す。

【００４３】

以下、日英翻訳における用例検索の例を示す。（１‐ｊ）は入力文、（２‐ｊ）は用例の原言語文とする。このうち入力文（１‐ｊ）の「色」と用例の原言語文（２‐ｊ）の「デザイン」の部分が両文の差分となる。

【００４４】

（１‐ｊ）色／が／気／に／入り／ません

（２‐ｊ）デザイン／が／気／に／入り／ません

ここで、「色」と「デザイン」とがシソーラス４４上で完全に異なった語であるものとすると、単語間の意味距離は１となる。従ってこの２文間の単語列編集距離は（０＋０＋２＊１）／（６＋６）＝０．１６７となる。

【００４５】

翻訳パターン抽出部４８は、用例検索部４６により検索された全ての用例に対し、原言語文中の、入力文と異なる箇所を変数で置換し、用例目的言語文中の対応する箇所に同じ変数を当てはめた翻訳パターンを生成する。両言語の文の間で対応をとる際は、変数となる単語のみ対象とし、全ての単語の対応をとる必要はない。つまり、変数部分以外の箇所は全体として対応していると仮定する。このため用例のほとんどの部分は変更されず、訳文の組み合わせ時に発生する誤りや不自然さの回避が期待される。

【００４６】

この原言語と目的言語の単語間の対応をとるには、様々な単語アライメント手法を適用できる。本実施の形態では、対訳辞書５０、第１の言語のシソーラス４４及び第２の言語のシソーラス４５に基づいて単語間の対応関係を判断している。

【００４７】

先の例の（２‐ｊ）に対応する目的言語文を（２‐ｅ）とする。

【００４８】

（２‐ｊ）デザイン／が／気／に／入り／ません

（２‐ｅ）Ｉ　ｄｏ　ｎｏｔ　ｌｉｋｅ　ｔｈｅ　ｄｅｓｉｇｎ．

このフェーズでは目的言語文（２‐ｅ）中で「デザイン」に対応する箇所が探し出され、「デザイン」と「ｄｅｓｉｇｎ」との間の対応が取られる。この結果、以下に示すような原言語パターン（２‐ｊ‐ｐ）と目的言語パターン（２‐ｅ‐ｐ）とからなる翻訳パターンが作られる。入力文によるその変数束縛は（１‐ｊ‐ｂ）となる。

【００４９】

（２‐ｊ‐ｐ）Ｘ／が／気／に／入り／ません

（２‐ｅ‐ｐ）Ｉ　ｄｏ　ｎｏｔ　ｌｉｋｅ　ｔｈｅ　Ｘ．

（１‐ｊ‐ｂ）Ｘ＝「色」

翻訳パターン選択部５２は、用例検索部４６により複数の用例が検索され、その結果翻訳パターン抽出部４８によって複数の翻訳パターンが抽出された際に、その中の一つを選択するための処理を行なうものである。翻訳パターン選択部５２は、複数の翻訳パターンから一つを選択するために、（１）より多くの用例検索結果から同じ翻訳パターンが抽出された方を選ぶ、（２）翻訳パターン中に現れる単語のコーパスでの出現頻度の合計が大きい方を選ぶ、というヒューリスティクスを使用して翻訳パターンを選択する。これらで一つの翻訳パターンを決定できない場合には、翻訳パターン選択部５２は任意の一つの翻

訳パターンを選ぶ。

【００５０】

訳語置換部５４は、翻訳パターン選択部５２により選択された翻訳パターンの変数に束縛された単語の訳を対訳辞書５０から引き、その訳語でもって目的言語パターンの変数を具体化する。先の例に基づいて説明すると、目的言語側の変数束縛は以下の（１‐ｅ‐ｂ）となり、訳文（１‐ｅ）が得られる。

【００５１】

（１‐ｅ‐ｂ）Ｘ＝“ｃｏｌｏｒ”

（１‐ｅ）Ｉ ｄｏ ｎｏｔ ｌｉｋｅ ｔｈｅ ｃｏｌｏｒ．

－用例検索部４６の詳細－

図１に示す用例機械翻訳装置３０の各処理の中で翻訳実行時間の大きな割合を占めるのは、用例検索部４６による用例検索である。用例の選択基準には、前述した単語列編集距離が使われる。

【００５２】

用例検索処理は、用例の原言語文を候補文とし、入力文との単語列編集距離がしきい値以内で最小の候補文をすべて求める。単語列編集距離は２文間の関係で定義され、二つの単語列のＤＰマッチにより計算可能である。従って各候補文と入力文間のＤＰマッチを逐次的に繰り返すことで単語列編集距離が最小の候補文を求めることができる。

【００５３】

しかし単純にこの方法を使おうとすれば、用例数に比例した処理時間がかかる。そのため大規模コーパスを利用したリアルタイムの翻訳処理をそうした方法で実現することは通常のコンピュータでは困難である。そこで本実施の形態では、以下に述べるような実装により用例検索を効率的に短時間で行なえるようにした。

【００５４】

［候補文集合の分割］

図２を参照して、対訳コーパス４２に含まれる候補文を、その内容語数と機能語数とを元に複数（Ｍ個）のグループ７０－１、７０－２、…７０－Ｍにグループ分けする。このようにグループ分けすることにより、入力文の内容語数と機能語数及び距離しきい値を用いて検索対象の候補文数を絞ることができる。具体的には以下のように用例検索を行なう。

【００５５】

まず、機能語同士、内容語同士はすべて一致すると仮定した場合の、グループごとに可能な最小距離を求める。最小距離が距離しきい値の範囲内で小さいグループから順に、単語列編集距離が最初に定めたしきい値以内で最小の候補文を検索する。あるグループ中に解が見つかれば、その解の距離を新たなしきい値として検索対象のグループをさらに絞ることができる。グループ内での用例検索については後述するが、グループ内では全ての候補文の内容語数と機能語数が等しい、つまりは単語数も等しいという事実が用例検索の前提条件として利用されている。

【００５６】

［単語グラフ］

内容語数と機能語数とを基準に分けられたグループ７０－１，７０－２，…，７０－Ｍの各々に対し、複数の候補文を一つの単語グラフ７２－１，７２－２，…，７２－Ｍにまとめる。すなわち一つのグループごとに一つの単語グラフが作成される。図３に単語グラフの例を示す。

【００５７】

図３に示されるように、単語グラフ８０は有向グラフであり、先頭ノード９０から最終ノード１０８に至る可能な道筋がそれぞれ一つの候補文に対応する。図３に示す単語グラフ８０は、先頭ノード９０及び最終ノード１０８を含めて、全部で１０個のノード９０、９２、９４、９６、９８、１００、１０２、１０４、１０６及び１０８を含む。ノードの間を結ぶリンクが単語に対応する。

【００５８】
例えば「全部売り切れました」という文はノード９０、９２、９６、１０２及び１０８という道筋に対応する。ノード列だけで見れば「全部届きました」という単語列も同じ道筋に対応する。ただし先の文では「売り切れる」というリンクによりノード９２及び９６が接続されていたが、この文では「届く」というリンクでこれらノードが接続されている点が異なる。このようにして、複数の文で共通な単語列をグラフ中では一つにまとめていくことにより、１グループに含まれる全ての候補文を一つの単語グラフ８０で表すことができる。
【００５９】
単語グラフを利用することにより、グループ内の全候補文を同時並行的に調べながら、入力文との距離が最小の候補文を検索することができる。
【００６０】
［Ａ＊アルゴリズム］
二つの単語列を照合した結果を示す単語の一致、置換、挿入、削除の列を「照合列」と呼ぶこととする。グループ内において単語列編集距離を最小とする候補文又は解を検索することは、単語グラフの先頭ノードから最終ノードまでの可能な全経路について、各経路に現れる単語列と入力単語列との照合列の中から単語列編集距離を最小にするものを探索することである。本実施の形態では、この探索問題の解放にＡ＊アルゴリズムを用いている。
【００６１】
一般にＡ＊アルゴリズムでは、問題状態集合の中から最小コストの下限の推定値が最小のものが選ばれ、継続状態に展開される。ここで対象とする問題では、状態は、単語グラフの経路と入力文との照合の途中経過を意味する。また「継続状態に展開」とは、選ばれた状態から遷移可能な全ての状態を生成し考慮の対象とすることを意味する。
【００６２】
［探索］
ここでは、ある単語グラフを用いた、単語列編集距離最小の経路を探索する処理について説明する。図３に示すように、ある単語グラフはノードとリンクとを含む。リンクは単語をラベルとして持ち、一つの始点ノードと一つの終点ノードとを結ぶ。例えば図３に示す「売り切れる」というリンクはノード９２と９６とを結んでいる。単語グラフ全体で一つの先頭ノードと一つの最終ノードとを持つ。図３ではこれらはノード９０及び１０８に相当する。
【００６３】
対象となる問題状態空間は以下で説明する状態、作用素、初期状態及び目標状態により構成される。
【００６４】
（１）状態
状態はｐａｔｈｓ，ｎｏｄｅ，ｉｎｐｕｔ，ｔｒａｎｓという属性を持つ。各属性の内容は以下の通りである。
・ｐａｔｈｓ：その時点までの照合列のリスト
・ｎｏｄｅ：単語グラフのノード。このノードまで照合が進んだことを示す。
・ｉｎｐｕｔ：入力単語列のうち、まだ照合に使われていない部分
・ｔｒａｎｓ：適用可能な作用素。
【００６５】
ｐａｔｈｓ内の各照合列の一致、置換、挿入、削除をそれぞれ（Ｅ：単語）、（Ｓ：グラフ側単語、入力側単語）、（Ｉ：入力側単語）、（Ｄ：グラフ側単語）の形式で表し、それぞれＥレコード、Ｓレコード、Ｉレコード及びＤレコードと呼ぶ。
【００６６】
状態のコストはｐａｔｈｓ内の任意の一つの照合列のコストである。ｐａｔｈｓ内のどの照合列も等しいコストを持つ。照合列のコストは、それに含まれるレコードのコストの

和である。Ｅレコードのコストは０、Ｉレコードのコストは１、Ｄレコードのコストは１と定義する。Ｓレコードのコストは、置換される２単語間の意味距離を２倍した値であるが、意味距離が０の場合には０でなく小さな正の値を与える。これは、類似語関係と同一語関係とを区別するためである。この値がＳレコードの最小コストとなる。

【００６７】

状態に作用素を適用することにより継続状態が生成される。一般に一つの状態に複数の作用素を適用することが可能である。従って一つの状態からいくつかの継続状態が生成される。

【００６８】

（２）作用素

５種類の作用素、Ｔ作用素、Ｅ作用素、Ｓ作用素、Ｉ作用素及びＤ作用素を以下のように定義する。Ｔ作用素とＩ作用素は状態に適用されるが、Ｅ、Ｓ、Ｄの各作用素は状態、及びその状態のｎｏｄｅを始点とするリンクの組に適用される。Ｔ作用素は実際に照合を進める作用素ではなく、ｔｒａｎｓ属性とともにＥ、Ｓ、Ｉ、Ｄの各作用素の適用順序を制御する役目を持つ。なお以下の説明では、作用素が適用される状態をｓ、リンクをｌ、生成される継続状態をｓ´と表し、各作用素について適用条件とどのような継続状態が生成されるかを示す。

【００６９】

・Ｔ作用素：

－条件：ｓ．ｔｒａｎｓがＥ作用素又はＳ作用素である。

【００７０】

－生成：ｓ´．ｔｒａｎｓ＝ｓ．ｔｒａｎｓがＥ作用素ならばＳ作用素とＮＩＬとから選択（説明は後述）、ｓ．ｔｒａｎｓがＳ作用素ならばＮＩＬ。ｓ´の他の属性値はｓと同じ。

【００７１】

・Ｅ作用素：

－条件：ｓ．ｔｒａｎｓがＥ作用素である。

【００７２】

かつ、ｓ．ｉｎｐｕｔが空リストでない。

【００７３】

かつ、ｌのラベルと、ｓ．ｉｎｐｕｔの先頭とが同一語である。

【００７４】

－生成：ｓ´．ｐａｔｈｓ＝ｓ．ｐａｔｈｓの各要素にＥレコードを追加した値
ｓ´．ｎｏｄｅ＝ｌの終点
ｓ´．ｉｎｐｕｔ＝ｓ．ｉｎｐｕｔから先頭を消去した値
ｓ´．ｔｒａｎｓ＝Ｅ作用素とＳ作用素とＮＩＬとから選択（説明は後述）

・Ｓ作用素：

－条件：ｓ．ｔｒａｎｓがＳ作用素である。

【００７５】

かつ、ｓ．ｉｎｐｕｔが空リストでない。

【００７６】

かつ、ｓ．ｉｎｐｕｔの先頭と、ｌのラベルとが同品詞の内容語であり、かつ同一語ではない。

【００７７】

かつ、これら２単語の意味距離は１未満である。

【００７８】

－生成：ｓ´．ｐａｔｈｓ＝ｓ．ｐａｔｈｓの各要素にＳレコードを追加した値
ｓ´．ｎｏｄｅ＝ｌの終点
ｓ´．ｉｎｐｕｔ＝ｓ．ｉｎｐｕｔから先頭を消去した値
ｓ´．ｔｒａｎｓ＝Ｅ作用素とＳ作用素とＮＩＬとから選択

・Ｉ作用素：
－条件：ｓ．ｔｒａｎｓがＮＩＬである。
【００７９】
かつ、ｓ．ｉｎｐｕｔが空リストでない。
【００８０】
－生成：ｓ´．ｐａｔｈｓ＝ｓ．ｐａｔｈｓの各要素にＩレコードを追加した値
ｓ´．ｎｏｄｅ＝ｓ．ｎｏｄｅ
ｓ´．ｉｎｐｕｔ＝ｓ．ｉｎｐｕｔから先頭を消去した値
ｓ´．ｔｒａｎｓ＝Ｅ作用素とＳ作用素とＮＩＬとから選択
・Ｄ作用素：
－条件：ｓ．ｔｒａｎｓがＮＩＬである。
【００８１】
かつ、ｓ．ｐａｔｈｓに最新レコードがＩレコードでない要素がある。
【００８２】
－生成：ｓ´．ｐａｔｈｓ＝ｓ．ｐａｔｈｓから最新レコードがＩレコードである要素を除き、残った要素にＤレコードを追加した値
ｓ´．ｎｏｄｅ＝ｌの終点
ｓ´．ｉｎｐｕｔ＝ｓ．ｉｎｐｕｔ
ｓ´．ｔｒａｎｓ＝Ｅ作用素とＳ作用素とＮＩＬとから選択
上の記載において、「Ｓ作用素とＮＩＬとから選択」とは、ｓ´にＳ作用素を適用できる可能性があればｓ´．ｔｒａｎｓの値をＳ作用素とし、可能性がなければＮＩＬとすることを意味する。本実施の形態では、ｓ´．ｉｎｐｕｔの先頭が内容語であり、その語と同一の語を除く同品詞語をラベルとしｓ´．ｎｏｄｅを始点とするリンクが存在する場合にＳ作用素を適用できる可能性があると判断する。
【００８３】
また、「Ｅ作用素とＳ作用素とＮＩＬとから選択」とは、ｓ´．ｉｎｐｕｔの先頭語をラベルとし、ｓ´．ｎｏｄｅを始点とするリンクが存在すればｓ´．ｔｒａｎｓの値をＥ作用素とし、そうでなければＳ作用素とＮＩＬとから選択する。
【００８４】
Ｄ作用素の２番目の条件はＩレコードの後にＤレコードが来るのを防いでいる。つまり、ＩレコードとＤレコードとが連続する場合、Ｄレコードが先に来るようにし、実質的に同じ削除と挿入が入れ替わっただけの異なる状態が現れる冗長性を排除する。
【００８５】
（３）初期状態と目標状態
初期状態では、ｐａｔｈｓは空リストを要素とするリスト、ｎｏｄｅは先頭ノード、ｉｎｐｕｔは入力単語列全体、ｔｒａｎｓはＥ作用素である。目標状態は、ｎｏｄｅが最終ノード、かつｉｎｐｕｔが空リストであるような状態である。
【００８６】
［探索アルゴリズム］
上記の初期状態、作用素及び目標状態で表現される状態空間からコスト最小の目標状態を探索する。初期条件としてコスト上限値が与えられる。コスト上限値は入力文長と候補文長の和を距離しきい値に乗じた値である。
【００８７】
［評価関数］
状態空間探索時に使用する評価関数ｆ＊を次のように定義する。
【００８８】
ｆ＊（ｓ）＝ｇ（ｓ）＋ｈ＊（ｓ）
ｇ（ｓ）は初期状態から状態ｓに達するまでにかかったコストを示す。つまりｇ（ｓ）は先に定義した状態のコストであり、ｓ．ｐａｔｈｓから計算できる。目標状態ではｆ＊（ｓ）＝ｇ（ｓ）となる。ｈ＊（ｓ）は状態ｓから目標状態までにかかるコストの下限で

ある。
【００８９】
一つの単語グラフを構成する全候補文の内容語数及び機能語数はそれぞれ同一である。従って状態ｓにおいて入力文側とグラフ側の未処理の内容語数及び機能語数がそれぞれ一意に決まる。それぞれの個数をC_{input}、C_{graph}、F_{input}、F_{graph}として、残り語数に基づく最小コストｈ´（ｓ）を次のように計算する。
【００９０】
ｈ´（ｓ）＝｜$C_{input}-C_{graph}$｜＋｜$F_{input}-F_{graph}$｜
さらに、Ｔ作用素の適用が先行する場合を含めて状態ｓに最初に適用可能なＥ、Ｓ、Ｉ及びＤの各作用素について、それが適用されたと仮定したときの目標状態までにかかるコストの下限を次の値とする。
【００９１】
・Ｅ作用素：ｈ´（ｓ）
・Ｓ作用素：ｈ´（ｓ）にＳレコードの最小コストを加えた値。
【００９２】
・Ｉ作用素：ｓ．ｉｎｐｕｔの先頭が内容語の場合は、｜（$C_{input}-1$）$-C_{graph}$｜＋｜$F_{input}-F_{graph}$｜に１を加えた値、機能語の場合は｜$C_{input}-C_{graph}$｜＋｜（$F_{input}-1$）$-F_{graph}$｜に１を加えた値。
【００９３】
・Ｄ作用素：｜C_{input}－（$C_{graph}-1$）｜＋｜$F_{input}-F_{graph}$｜と｜$C_{input}-C_{graph}$｜＋｜F_{input}－（$F_{graph}-1$）｜を求め、その小さいほうの値に１を加えた値。ただし、ｓ．ｎｏｄｅを始点とするリンクのラベルが内容語のみであるか機能語のみであれば、対応する一方の値に１を加えた値。
【００９４】
これらを使ってｈ＊（ｓ）を以下のように計算する。
【００９５】
（１）ｓ．ｔｒａｎｓがＥ作用素のときは、Ｅ作用素が適用されたときのコストの下限
（２）ｓ．ｔｒａｎｓがＳ作用素のときは、Ｓ作用素、Ｉ作用素又はＤ作用素が適用されたときのコストの下限の最小値
（３）ｓ．ｔｒａｎｓがＮＩＬのときは、Ｉ作用素又はＤ作用素が適用されたときのコストの下限の最小値。
【００９６】
［探索処理］
探索処理は以下のように行なわれる。以下に示す探索処理は、コンピュータ上で実行されるプログラムにより上記した用例検索を行なう際のプログラムの制御の流れを示すものである。なお以下の説明中において、ＯＰＥＮは未展開状態を、ＣＬＯＳＥＤは展開済状態を、それぞれ保持するためのリストを示す。また「同じ状態」とは、ｐａｔｈｓを除く属性値が等しい状態を意味する。
【００９７】
図４は、この処理全体のフローチャートである。まずステップ１２０で、グループごとに可能な最小距離を算出する。ステップ１２２で、可能な最小距離がしきい値以下のグループを選択する。ステップ１２４で、選択されたグループを最小距離の昇順でソートする。ステップ１２６で、選択されたグループの数をグループ数の値N_{MAX}に代入する。
【００９８】
ステップ１２８以下がグループごとの繰り返し処理である。まずステップ１２８で繰り返し制御変数Ｊに０を代入する。ステップ１３０で変数Ｊに１を加算する。ステップ１３２でＪがグループ数N_{MAX}を超えたか否かを判定する。超えていれば処理を終了する。超えていなければステップ１３４でグループ（Ｊ）の可能な最小距離がしきい値以下か否かを判定する。しきい値以下であればステップ１３６に進み、それ以外の場合には処理を終了する。

【００９９】
ステップ１３６ではグループ（ｊ）内で距離≦しきい値の解を探索する処理を行なう。この処理の詳細については図５を参照して説明する。
【０１００】
ステップ１３８で解が存在したか否かについての判定が行なわれる。解が存在していれば制御はステップ１４０に進み、それ以外の場合にはステップ１３０に戻る。
【０１０１】
ステップ１４０では、求められた解の距離を新たなしきい値に代入し、ステップ１３０に戻る。
【０１０２】
図５に、ステップ１３６で行なわれる処理の詳細について示す。ステップ１６０で、コスト上限C_{MAX}に、所与の値（入力文長と候補文長の和を距離しきい値に乗じた値）を代入する。
【０１０３】
ステップ１６２でＯＰＥＮに初期状態のみを入れる。
【０１０４】
ステップ１６４でＯＰＥＮ内にコストがC_{MAX}以下の状態があるか否かについての判定を行なう。条件を満たす状態がなければ処理を終了する。条件を満たす状態があればステップ１６６に進む。
【０１０５】
ステップ１６６ではＯＰＥＮから評価関数f^*を最小にする状態ｓを取り除き、ＣＬＯＳＥＤに入れる。
【０１０６】
ステップ１６８で、状態ｓが目標状態か否かを判定する。目標状態であればステップ１７４に進み、それ以外の場合にはステップ１７０に進む。
【０１０７】
ステップ１７４では状態ｓを解の一つとし、続くステップ１７６でコスト上限C_{MAX}を状態ｓのコストで置換し、ステップ１６４に戻る。
【０１０８】
一方ステップ１７０では、状態ｓの全ての継続状態を生成する。そしてステップ１７２で、各継続状態ｓ´について図６に示す処理を実行する。
【０１０９】
図６を参照して、ステップ１９０でf^*（ｓ´）がコスト上限C_{MAX}以下か否かを判定する。条件が充足されていなければ処理を終了する。条件が充足されていればステップ１９２で、ＯＰＥＮ及びＣＬＯＳＥＤ中の同じ状態と比較し、条件により以下の処理を行なう。
【０１１０】
（ａ）同じ状態がなければ、状態ｓ´をＯＰＥＮに追加する（ステップ１９４）
（ｂ）状態ｓ´よりコストの大きい同じ状態がＯＰＥＮ又はＣＬＯＳＥＤに既存であれば、この既存状態を消去し（ステップ１９６）、状態ｓ´をＯＰＥＮに追加する（ステップ１９８）。
【０１１１】
（ｃ）状態ｓ´とコストの等しい同じ状態がＣＬＯＳＥＤに既存であれば、この既存状態を消去し（ステップ２００）、状態ｓ´をＯＰＥＮに追加する（ステップ２０２）。
【０１１２】
（ｄ）コストの等しい同じ状態がＯＰＥＮに既存であれば、この既存状態のｐａｔｈｓにｓ´．ｐａｔｈｓをマージする（ステップ２０４）。
【０１１３】
以上の処理が終了したら図６の処理を終了し、図５のステップ１６４に戻る。
【０１１４】

［単語グラフの特徴の利用］

単語グラフの形状の特徴として、開始ノードを始点とするリンク数が他のノードを始点とするリンク数よりも圧倒的に大きくなる傾向がある。そのためｎｏｄｅ属性に開始ノードを持つ状態にＤ作用素が適用されると、多くの継続状態が作られることとなり計算時間が大きくなる。これは、照合列の先頭要素がＤレコードとなる場合である。この展開数の増大を避けるため、単語グラフ中、先頭ノードから数段階の仮のリンクとノードを加える。先頭ノードを持つ状態からＤ作用素によって第１の仮のノードを持つ状態へ遷移する。第１の仮のノードは、全候補文について２番目の語をラベルとするリンクの始点となり、通常の単語グラフのノードに合流する。第１の仮のノードにある状態はＥ作用素又はＳ作用素の適用により通常のノードの状態、Ｄ作用素によって第２の仮のノードを持つ状態に遷移する。

【０１１５】

何段階まで仮のノードを用意するかは、用例検索時に使われる可能性のある距離しきい値の最大値から計算できる。候補文の長さをＬとすると、照合列の先頭にＤレコードがｄ個並ぶという条件で、候補文とその距離を最小にする入力文は、候補文から先頭ｄ語を除いた文である。そのときの距離はｄ／（（Ｌ－ｄ）＋Ｌ）である。この値が距離しきい値の最大値を越える場合は探索する必要はない。距離しきい値の最大値をΘとすると、ｄ／（（Ｌ－ｄ）＋Ｌ）≦Θからｄ≦２ΘＬ／（１＋Θ）が導かれる。この式を満たすｄの最大の整数値が用意すべき仮のノードの段数である。

【０１１６】

－動作－

以下、本実施の形態に係る用例機械翻訳装置３０の動作について説明する。図１において、翻訳パターン抽出部４８、翻訳パターン選択部５２及び訳語置換部５４の動作は特許文献１において提案されている用例機械翻訳装置の動作と同様である。従ってここではそれらについての詳しい説明は繰り返さず、用例検索部４６による用例検索の詳細について説明する。

【０１１７】

用例検索部４６の動作を説明するために、用例検索の実行例を示す。ここでは、図３に示す単語グラフ８０から入力文「全部揃いました」の類似文を検索することにする。以下の説明では状態を［ｐａｔｈｓ，ｎｏｄｅ，ｉｎｐｕｔ，ｔｒａｎｓ，ｆ＊関数値］の形式で記述する。ｎｏｄｅ値には図３中でノードにつけた参照番号（９０－１０８）を用いる。δをＳレコードの最小コストとする。また「揃う」と「売り切れる」の意味距離を１．０、「揃う」と「届く」との意味距離を０．７であると仮定する。

【０１１８】

初期状態ｓ０は次のようになる。

【０１１９】

ｓ０＝［（（）），ノード９０，（全部，揃う，ます，た），Ｅ作用素，０］

状態ｓ０に適用可能な作用素はＥ作用素とＴ作用素とである。これらの作用素を適用して継続状態ｓ１とｓ２とが得られる。ＯＰＥＮは｛ｓ１，ｓ２｝となる。

【０１２０】

ｓ１＝［（（（Ｅ，全部））），ノード９２，（揃う，ます，た），Ｓ作用素，δ］

ｓ２＝［（（）），ノード９０，（全部，揃う，ます，た），ＮＩＬ，２］

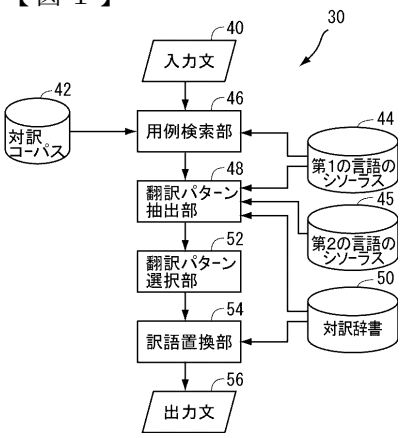
ここでＯＰＥＮの中から、ｆ＊関数値の小さな状態ｓ１が選ばれ展開される。状態ｓ１に適用可能な作用素はＳ作用素とＴ作用素とである。「売り切れる」と「届く」とをラベルとする二つのリンクに関してＳ作用素の適用条件がテストされる。ここでは、１未満の意味距離の得られる「届く」のリンクについてのみテストが成功する。Ｓ作用素の適用により継続状態ｓ３が得られる。またＴ作用素の適用により継続状態ｓ４が生成される。ＯＰＥＮは｛ｓ２，ｓ３，ｓ４｝となる。

【０１２１】

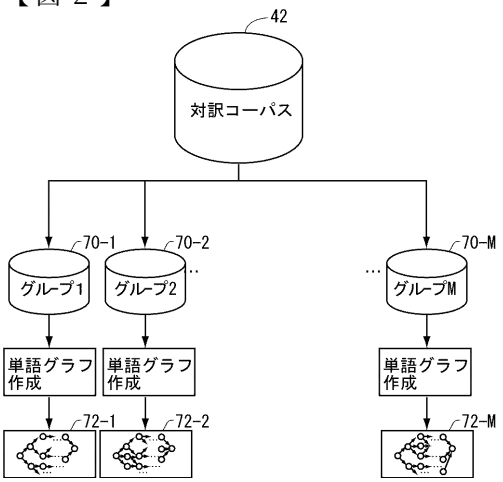
ｓ３＝［（（（Ｅ，全部），（Ｓ，届く，揃う））），ノード９６，（ます，た），Ｅ

作用素，１．４］
ｓ４＝［（（（Ｅ，全部））），ノード９２，（揃う，ます，た），ＮＩＬ，２］
ここでＯＰＥＮの中から、ｆ＊関数値が最小の状態ｓ３が選ばれ展開される。状態ｓ３にＥ作用素を２回適用した状態ｓ７が解となる。
【０１２２】
ｓ７＝［（（（Ｅ，全部），（Ｓ，届く，揃う），（Ｅ，ます），（Ｅ，た））），ノード１０８，（），ＮＩＬ，１．４］
この例の状態遷移の様子を図７に示す。図７に示されるように、Ｔ作用素の適用による遷移を除き、解である状態ｓ７に向かって一直線に探索が進んでいる。
【０１２３】
以上のようにこの実施の形態に係る用例機械翻訳装置３０によれば、大規模なコーパスから短時間で入力文との編集距離が最小の用例を探索することができる。探索された用例から翻訳パターンを抽出し、翻訳パターン中の変数部分を訳語置換することにより、入力文４０（図１参照）に対応する翻訳出力文５６を得ることができる。
【０１２４】
特に、本実施の形態は、（１）シソーラスを使わず、単語列編集距離の定義において意味距離を定数とする、（２）内容語・機能語の分類を行なわない（全ての語が一方に属するとする。）、又は他の分類を使う（文字種で分ける場合など。）、並びに（３）１文字ずつを１単語として扱う、という場合に編集距離に基づいたストリング検索装置を提供し、特許文献２が挙げている各応用課題において、大量の候補の中から解を効率よく検索する手段となる。
【０１２５】
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および範囲内でのすべての変更を含む。
【図面の簡単な説明】
【０１２６】
【図１】本発明の一実施の形態に係る用例機械翻訳装置３０のブロック図である。
【図２】本実施の形態における対訳コーパス４２のグループへの分割と単語グラフの作成とを模式的に示す図である。
【図３】単語グラフの例を示す図である。
【図４】本実施の形態に係る用例機械翻訳装置３０の用例検索部４６で実行される処理のプログラムフローチャートである。
【図５】用例グループ内で解を探索する処理のプログラムフローチャートである。
【図６】探索処理において継続状態ごとに行なわれる処理のプログラムフローチャートである。
【図７】本発明の一実施の形態に係る用例機械翻訳装置３０の用例検索部４６の動作例を説明するための図である。
【符号の説明】
【０１２７】
４０　入力文、４２　対訳コーパス、４４　第１の言語のシソーラス、４５　第２の言語のシソーラス、４６　用例検索部、４８　翻訳パターン抽出部、５０　対訳辞書、５２　翻訳パターン選択部、５４　訳語置換部、５６　出力文、８０　単語グラフ

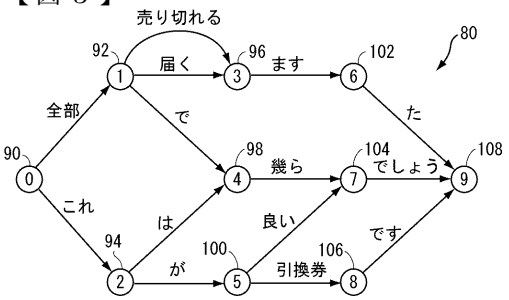
【図１】


30
40
入力文
42
対訳コーパス
46
用例検索部
44
第1の言語のシソーラス
48
翻訳パターン抽出部
45
第2の言語のシソーラス
52
翻訳パターン選択部
50
対訳辞書
54
訳語置換部
56
出力文

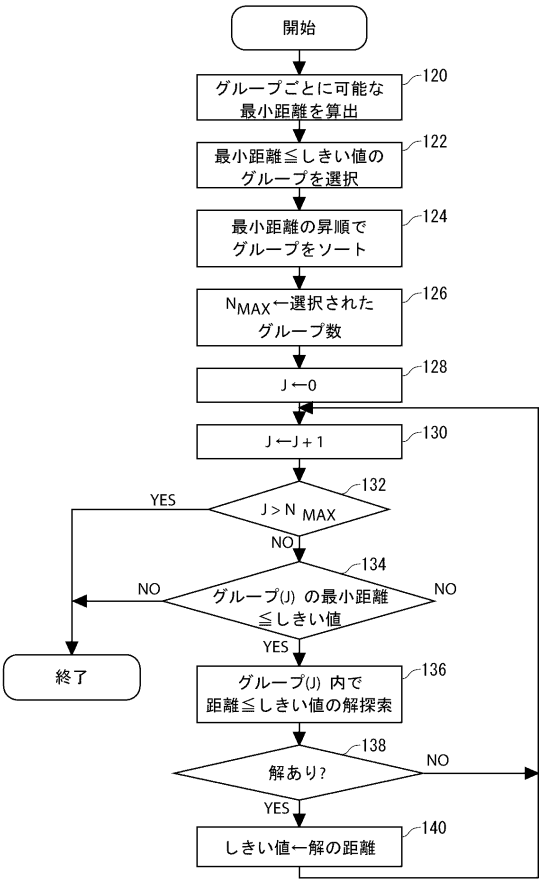
【図２】


42
対訳コーパス
70-1
70-2
70-M
グループ1
グループ2
グループM
単語グラフ作成
単語グラフ作成
単語グラフ作成
72-1
72-2
72-M

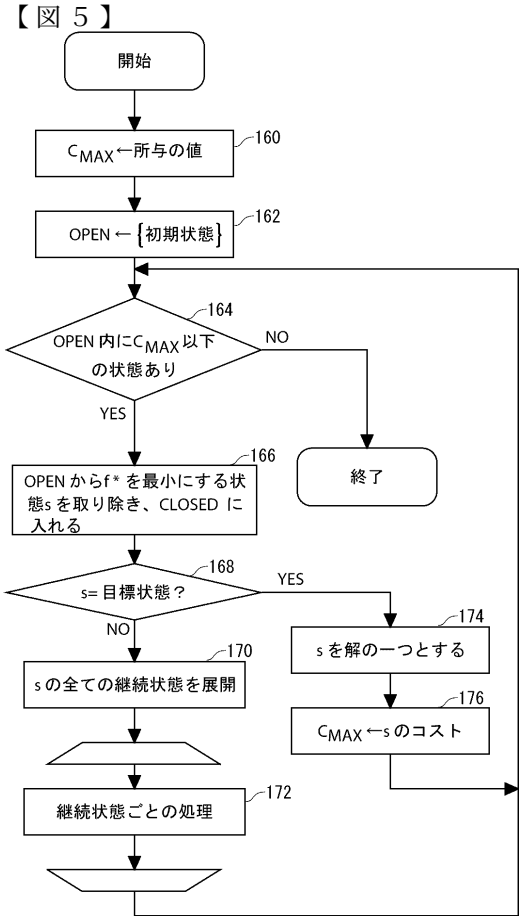
【図３】


80
売り切れる
92
届く
96
ます
102
全部
で
た
90
98
幾ら
104
でしょう
108
これ
は
良い
です
94
が
100
引換券
106

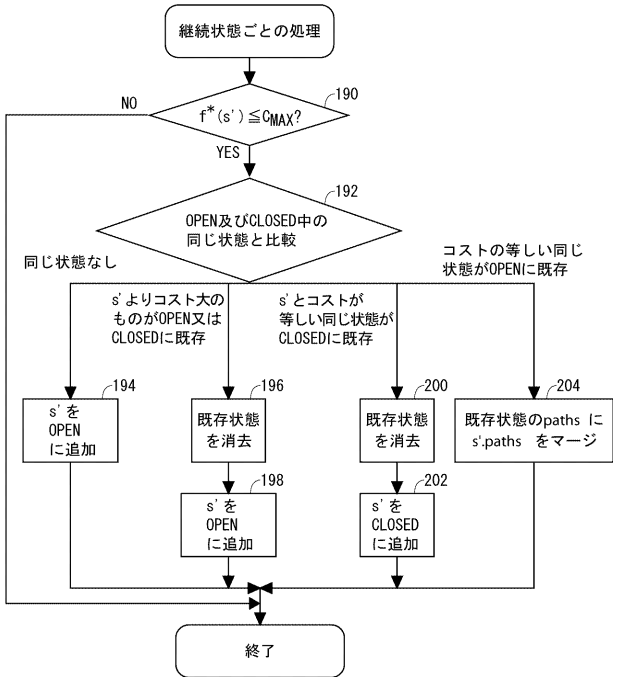
【図４】


開始
グループごとに可能な最小距離を算出
120
最小距離≦しきい値のグループを選択
122
最小距離の昇順でグループをソート
124
NMAX←選択されたグループ数
126
J←0
128
J←J+1
130
J > NMAX
132
YES
NO
グループ(J) の最小距離≦しきい値
134
NO
NO
YES
終了
グループ(J) 内で距離≦しきい値の解探索
136
解あり?
138
NO
YES
しきい値←解の距離
140

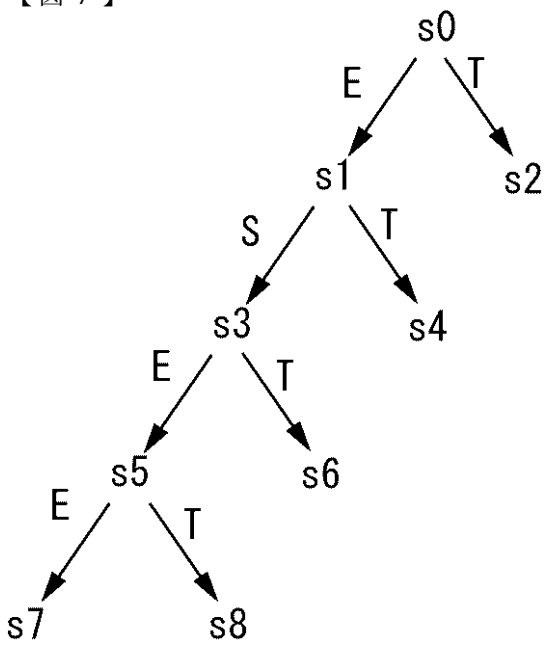
【図５】


開始
CMAX←所与の値
160
OPEN ←{初期状態}
162
OPEN 内にCMAX以下の状態あり
164
NO
YES
終了
OPEN からf* を最小にする状態s を取り除き、CLOSED に入れる
166
s= 目標状態？
168
YES
NO
s を解の一つとする
174
s の全ての継続状態を展開
170
CMAX←s のコスト
176
継続状態ごとの処理
172

【図６】


継続状態ごとの処理
190
NO
f*(s')≦CMAX?
YES
192
OPEN及びCLOSED中の
同じ状態と比較
同じ状態なし
コストの等しい同じ
状態がOPENに既存
s'よりコスト大の
ものがOPEN又は
CLOSEDに既存
s'とコストが
等しい同じ状態が
CLOSEDに既存
194
s'を
OPEN
に追加
196
既存状態
を消去
200
既存状態
を消去
204
既存状態のpaths に
s'.paths をマージ
198
s'を
OPEN
に追加
202
s'を
CLOSED
に追加
終了

【図７】


s0
E
T
s1
s2
S
T
s3
s4
E
T
s5
s6
E
T
s7
s8