(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2003－308319
（P2003－308319A）
(43)公開日　平成15年10月31日(2003.10.31)

(51)Int.Cl.[7]	識別記号	ＦＩ	ﾃｰﾏｺｰﾄﾞ(参考)
Ｇ０６Ｆ　17/28		Ｇ０６Ｆ　17/28	Ｐ　５Ｂ０９１

審査請求　有　　請求項の数32　ＯＬ　（全 16 頁）

(21)出願番号　特願2002－113422(P2002－113422)

(22)出願日　平成14年４月16日(2002.4.16)

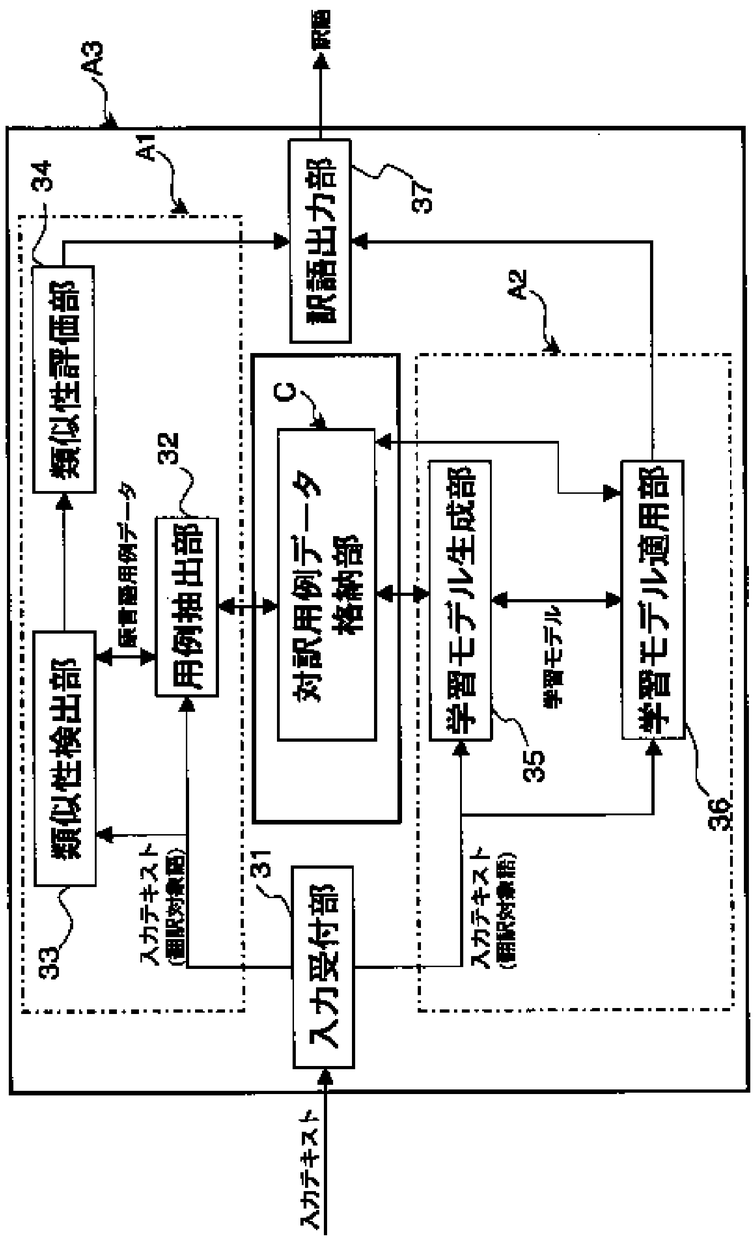
特許法第30条第１項適用申請有り　2001年10月10日　社団法人電子情報通信学会発行の「電子情報通信学会技術研究報告　信学技報　Ｖｏｌ．101　Ｎｏ．351」に発表

(71)出願人　301022471
独立行政法人通信総合研究所
東京都小金井市貫井北町４－２－１
(72)発明者　内元　清貴
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内
(72)発明者　関根　聡
アメリカ合衆国　1003　ニューヨーク州　ニューヨーク　ブロードウェイ　715　セブンスフロア　ニューヨーク大学内
(74)代理人　100085338
弁理士　赤澤　一博　（外１名）

最終頁に続く

(54)【発明の名称】　訳語選択装置、翻訳装置、訳語選択プログラム、及び翻訳プログラム

(57)【要約】
【課題】多量の対訳用例データを収集することなく、従来は適切な訳語選択或いは翻訳が困難であった慣用的な表現に対しても精度の高い訳語選択、並びに機械翻訳を行うことができるようにする。
【解決手段】入力テキストと対訳用例データとの文字列の類似性により、入力テキスト中の翻訳対象語に対応する訳語を出力する方法、又は翻訳対象語に基づいて学習モデルを生成し、そのうち学習データに対し最も精度の高い学習モデルを入力文に適用したうえで最高の確信度が得られたものに対応する訳語候補を出力する方法を、それぞれ単独で又は組み合わせて用いる。



【特許請求の範囲】
【請求項１】第１言語によるテキストからなる原言語用例及びそれに含まれる語とその語の第２言語による訳語及び当該訳語に関する情報とを含む原言語用例データと、前記原言語用例から第２言語で翻訳されたテキストからなる目的言語用例とを対にした対訳用例データを格納する対訳用例データ格納部を利用して、第１言語で入力された入力テキストに含まれる翻訳すべき語である翻訳対象語に対応する第２言語で記述された訳語を選択するものであって、
前記入力テキストの入力を受け付ける入力受付部と、
入力受付部で受け付けた入力テキスト中の前記翻訳対象語に該当する語を含む少なくとも一以上の原言語用例データを、前記対訳用例データ格納部から抽出する用例抽出部と、
前記入力テキスト及び前記用例抽出部で抽出した原言語用例データに基づき、入力テキストと原言語用例との類似性を検出する類似性検出部と、
類似性検出部で検出した原言語用例の類似性を比較評価し、最も高い類似性を有する少なくとも原言語用例データを出力する類似性評価部と、
類似性評価部で出力した原言語用例データに対応する対訳用例データに含まれる目的言語用例中の前記翻訳対象語に対応する訳語を出力する訳語出力部とを具備してなることを特徴とする訳語選択装置。
【請求項２】類似性検出部が、入力テキストと抽出された原言語用例データに含まれる原言語用例とを文字単位で比較して求められる差異に基づき入力テキストと原言語用例との一致した文字列の割合、又は一致した部分が何カ所に分割されて一致しているかを示す分割数の少なくともいずれか一方を用いて計算される類似度を前記類似性として演算する類似度演算部を有している請求項１記載の訳語選択装置。
【請求項３】用例抽出部が、抽出した原言語用例データに含まれる原言語用例に文末処理を施して処理済原言語用例を出力する原言語用例処理部を有するものであり、類似性検出部において前記類似度演算部が、入力テキストと処理済原言語用例との文字単位で比較して求められる差異の演算結果に基づいて、一致した文字列の当該処理済原言語用例の文字列に対する割合、又は一致した部分が何カ所に分割されて一致しているかを示す分割数の少なくともいずれか一方を類似度として演算する請求項２記載の訳語選択装置。
【請求項４】訳語出力部が、類似性検出部の類似度演算部で演算し類似性評価部で評価した結果、類似度が最大となる原言語用例データが複数ある場合に、前記類似度演算部における演算の結果、入力テキストと一致した文字列の割合又は前記分割数が最大の原言語用例を含む対訳用例データにおける前記翻訳対象語に対応する訳語を出力する請求項３記載の訳語選択装置。
【請求項５】入力受付部が、入力テキストを形態素解析により翻訳対象語を自動抽出する入力テキスト処理部を有している請求項１、２、３又は４記載の訳語選択装置。
【請求項６】対訳用例データが、原言語用例に含まれる語に基づいて生成された原言語見出し語を含むものであり、用例抽出部が、少なくとも前記翻訳対象語に該当する原言語見出し語を含む原言語用例データを対訳用例データ格納部から抽出するものである請求項１、２、３、４又は５記載の訳語選択装置。
【請求項７】対訳用例データが、原言語用例に含まれる語に基づいて生成された原言語見出し語とそれに対応する訳語に基づいて生成された目的言語見出し語とを有するものであり、用例抽出部が、前記翻訳対象語に該当する原言語見出し語を含む原言語用例データを少なくとも抽出するものであって、訳語出力部が、類似性評価部において出力された原言語用例データに含まれ且つ前記用例抽出部で抽出された原言語見出し語に対応する目的言語見出し語を出力する請求項１、２、３、４又は５記載の訳語選択装置。
【請求項８】第１言語によるテキストからなる原言語用例及びそれに含まれる語とその語の第２言語による訳語及び当該訳語に関する情報とを含む原言語用例データと、前記原言語用例から第２言語で翻訳されたテキストからなる目的言語用例とを対にした対訳用例データを格納する対訳用例データ格納部を利用して、第１言語で入力された入力テキストに含まれる翻訳すべき語である翻訳対象語に対応する第２言語で記述された訳語を選択するものであって、
前記入力テキストの入力を受け付ける入力受付部と、
対訳用例データ格納部に格納された原言語用例に含まれる語及び当該原言語用例に対応する対訳用例データに基づいて作成された学習データを利用して、入力受付部で受け付けた入力テキスト中の翻訳対象語に対応した学習モデルを生成する学習モデル生成部と、
学習モデル生成部で生成した学習モデルを入力テキスト中の翻訳対象語に適用し、当該翻訳対象語の訳語候補の全てについて確信度を演算し、確信度順に順序付けて訳語候補を出力する学習モデル適用部と、
学習モデル適用部で出力した訳語候補のうち、最も高い確信度が得られた訳語候補を選択して翻訳対象語に対応する訳語として出力する訳語出力部とを具備してなることを特徴とする訳語選択装置。
【請求項９】学習モデル生成部が、入力受付部で受け付けた入力テキスト中の翻訳対象語ごとにそれを含む原言語用例に対応する対訳用例データを前記対訳用例データ格納部から抽出し、その抽出された対訳用例データに基づいて学習モデルを生成するものである請求項８記載の訳語選択装置。
【請求項１０】学習モデル生成部が、学習データごとに

対応して学習モデルを生成するものであり、入力受付部で受け付けた入力テキスト中の前記翻訳対象語ごとに前記学習データで精度が最高となる学習モデルを選択する学習モデル選択部をさらに含むものであり、学習モデル適用部が、前記学習モデル選択部で選択した学習モデルを入力テキスト中の翻訳対象語に適用するものである請求項８又は９記載の訳語選択装置。

【請求項１１】入力受付部が、入力テキストを形態素解析により翻訳対象語を自動抽出する入力テキスト処理部を有している請求項８、９又は１０記載の訳語選択装置。

【請求項１２】対訳用例データが、原言語用例に含まれる語に基づいて生成された原言語見出し語を含むものであり、学習モデル生成部が、少なくとも前記翻訳対象語に該当する原言語見出し語を含む原言語用例データを対訳用例データ格納部から抽出するものである請求項８、９、１０又は１１記載の訳語選択装置。

【請求項１３】第１言語によるテキストからなる原言語用例及びそれに含まれる語とその語の第２言語による訳語及び当該訳語に関する情報とを含む原言語用例データと、前記原言語用例から第２言語で翻訳されたテキストからなる目的言語用例とを対にした対訳用例データを格納する対訳用例データ格納部を利用して、第１言語で入力された入力テキストに含まれる翻訳すべき語である翻訳対象語に対応する第２言語で記述された訳語を選択するものであって、

前記入力テキストの入力を受け付ける入力受付部と、

入力受付部で受け付けた入力テキスト中の前記翻訳対象語に該当する語を含む少なくとも一以上の原言語用例データを、前記対訳用例データ格納部から抽出する用例抽出部と、

前記入力テキスト及び前記用例抽出部で抽出した原言語用例データに基づき、入力テキストと原言語用例との類似性を検出する類似性検出部と、

類似性検出部で検出した原言語用例の類似性を比較評価し、最も高い類似性を有する少なくとも原言語用例データを出力する類似性評価部と、

対訳用例データ格納部に格納された原言語用例に含まれる語及び当該原言語用例に対応する対訳用例データに基づいて作成された学習データを利用して、入力受付部で受け付けた入力テキスト中の翻訳対象語に対応した学習モデルを生成する学習モデル生成部と、

学習モデル生成部で生成した学習モデルを入力テキスト中の翻訳対象語に適用し、当該翻訳対象語の訳語候補の全てについて確信度を演算し、確信度順に順序付けて訳語候補を出力する学習モデル適用部と、

類似性評価部で出力する原言語用例データに対応する対訳用例データに含まれる目的言語用例中の前記翻訳対象語に対応する訳語、又は、学習モデル適用部で出力する訳語候補から、最適なものを選択して翻訳対象語に対応する訳語として出力する訳語出力部とを具備してなることを特徴とする訳語選択装置。

【請求項１４】訳語出力部が、類似性評価部において所定の閾値以上の類似性が得られた対訳用例データの出力がある場合に、当該類似性評価部で出力した結果得られる翻訳対象語に対応する訳語を出力し、類似性評価部において所定の閾値以上の類似性が得られた対訳用例データの出力がない場合に、前記学習モデル適用部で出力した結果得られる翻訳対象語に対応する訳語を出力するものである請求項１３記載の訳語選択装置。

【請求項１５】類似性評価部において所定の閾値以上の類似性が得られた対訳用例データの出力がない場合に、前記学習モデル生成部、学習モデル適用部及び訳語出力部を動作させるようにしている請求項１３記載の訳語選択装置。

【請求項１６】用例抽出部が利用する対訳用例データ格納部と、学習モデル生成部が利用する対訳用例データ格納部とが、それぞれ異なる言語資源に基づいて作成された異なる対訳用例データ格納部である請求項１３、１４又は１５記載の訳語選択装置。

【請求項１７】第１言語によるテキストからなる原言語用例及びそれに含まれる語とその語の第２言語による訳語及び当該訳語に関する情報とを含む原言語用例データと、前記原言語用例から第２言語で翻訳されたテキストからなる目的言語用例とを対にした対訳用例データを格納する対訳用例データ格納部を利用して、第１言語で入力された入力テキストに基づいてその第２言語による翻訳文である対象テキストを出力するものであって、

前記入力テキストの入力を受け付ける入力受付部と、

入力受付部で受け付けた入力テキスト中の各翻訳対象語に該当する語を含む少なくとも一以上の原言語用例データを、前記対訳用例データ格納部から抽出する用例抽出部と、

前記入力テキスト及び前記用例抽出部で抽出した原言語用例データに基づき、入力テキストと原言語用例との類似性を検出する類似性検出部と、

類似性検出部で検出した原言語用例の類似性を比較評価し、最も高い類似性を有する少なくとも原言語用例データを出力する類似性評価部と、

類似性評価部で出力した原言語用例データに対応する対訳用例データに含まれる目的言語用例中の前記翻訳対象語に対応する訳語を出力する訳語出力部と、

訳語出力部で出力した訳語及び当該訳語を含む対訳用例データに基づいて、入力テキストに対応する対象テキストを生成し出力する翻訳文出力部とを具備してなることを特徴とする翻訳装置。

【請求項１８】第１言語によるテキストからなる原言語用例及びそれに含まれる語とその語の第２言語による訳語及び当該訳語に関する情報とを含む原言語用例データと、前記原言語用例から第２言語で翻訳されたテキスト

からなる目的言語用例とを対にした対訳用例データを格納する対訳用例データ格納部を利用して、第１言語で入力された入力テキストに含まれる翻訳すべき語である翻訳対象語に対応する第２言語で記述された訳語を選択するものであって、
前記入力テキストの入力を受け付ける入力受付部と、
対訳用例データ格納部に格納された原言語用例に含まれる語及び当該原言語用例に対応する対訳用例データに基づいて作成された学習データを利用して、入力受付部で受け付けた入力テキスト中の翻訳対象語に対応した学習モデルを生成する学習モデル生成部と、
学習モデル生成部で生成した学習モデルを入力テキスト中の翻訳対象語に適用し、当該翻訳対象語の訳語候補の全てについて確信度を演算し、確信度順に順序付けて訳語候補を出力する学習モデル適用部と、
学習モデル適用部で出力した訳語候補のうち、最も高い確信度が得られた訳語候補を選択して翻訳対象語に対応する訳語として出力する訳語出力部と、
訳語出力部で出力した訳語及び当該訳語を含む対訳用例データに基づいて、入力テキストに対応する対象テキストを生成し出力する翻訳文出力部とを具備してなることを特徴とする翻訳装置。
【請求項１９】第１言語によるテキストからなる原言語用例及びそれに含まれる語とその語の第２言語による訳語及び当該訳語に関する情報とを含む原言語用例データと、前記原言語用例から第２言語で翻訳されたテキストからなる目的言語用例とを対にした対訳用例データを格納する対訳用例データ格納部を利用して、第１言語で入力された入力テキストに含まれる翻訳すべき語である翻訳対象語に対応する第２言語で記述された訳語を選択するものであって、
前記入力テキストの入力を受け付ける入力受付部と、
入力受付部で受け付けた入力テキスト中の前記翻訳対象語に該当する語を含む少なくとも一以上の原言語用例データを、前記対訳用例データ格納部から抽出する用例抽出部と、
前記入力テキスト及び前記用例抽出部で抽出した原言語用例データに基づき、
入力テキストと原言語用例との類似性を検出する類似性検出部と、
類似性検出部で検出した原言語用例の類似性を比較評価し、最も高い類似性を有する少なくとも原言語用例データを出力する類似性評価部と、
対訳用例データ格納部に格納された原言語用例に含まれる語及び当該原言語用例に対応する対訳用例データに基づいて作成された学習データを利用して、入力受付部で受け付けた入力テキスト中の翻訳対象語に対応した学習モデルを生成する学習モデル生成部と、
学習モデル生成部で生成した学習モデルを入力テキスト中の翻訳対象語に適用し、当該翻訳対象語の訳語候補の全てについて確信度を演算し、確信度順に順序付けて訳語候補を出力する学習モデル適用部と、
類似性評価部で出力する原言語用例データに対応する対訳用例データに含まれる目的言語用例中の前記翻訳対象語に対応する訳語、又は、学習モデル適用部で出力する訳語候補から、最適なものを選択して翻訳対象語に対応する訳語として出力する訳語出力部と、
訳語出力部で出力した訳語及び当該訳語を含む対訳用例データに基づいて、入力テキストに対応する対象テキストを生成し出力する翻訳文出力部とを具備してなることを特徴とする翻訳装置。
【請求項２０】第１言語によるテキストからなる原言語用例及びそれに含まれる語とその語の第２言語による訳語及び当該訳語に関する情報とを含む原言語用例データと、前記原言語用例から第２言語で翻訳されたテキストからなる目的言語用例とを対にした対訳用例データを格納する対訳用例データ格納部を利用してコンピュータを動作させ、第１言語で入力された入力テキストに含まれる翻訳すべき語である翻訳対象語に対応する第２言語で記述された訳語を選択するプログラムであって、
前記入力テキストの入力を受け付ける入力受付ステップと、
前記受け付けた入力テキスト中の前記翻訳対象語に該当する語を含む少なくとも一以上の原言語用例データを、前記対訳用例データ格納部から抽出する用例抽出ステップと、
前記入力テキスト及び前記抽出した原言語用例データに基づき、入力テキストと原言語用例との類似性を検出する類似性検出ステップと、
前記検出した原言語用例の類似性を比較評価し、最も高い類似性を有する少なくとも原言語用例データを出力する類似性評価ステップと、
前記出力した原言語用例データに対応する対訳用例データに含まれる目的言語用例中の前記翻訳対象語に対応する訳語を出力する訳語出力ステップとを有することを特徴とする訳語選択プログラム。
【請求項２１】類似性検出ステップが、入力テキストと抽出された原言語用例データに含まれる原言語用例とを文字単位で比較して求められる差異に基づき入力テキストと原言語用例との一致した文字列の割合、又は一致した部分が何カ所に分割されて一致しているかを示す分割数の少なくともいずれか一方を用いて計算される類似度を前記類似性として演算する類似度演算ステップを含む請求項２１記載の訳語選択プログラム。
【請求項２２】用例抽出ステップが、用例抽出ステップで抽出した原言語用例データに含まれる原言語用例に文末処理を施して処理済原言語用例を出力する原言語用例処理ステップを含み、類似度演算ステップにおいて、入力テキストと処理済原言語用例との文字単位で比較して求められる差異の演算結果に基づき、一致した文字列の

当該処理済原言語用例の文字列に対する割合、又は一致した部分が何カ所に分割されて一致しているかを示す分割数の少なくともいずれか一方を類似度として演算する請求項２１記載の訳語選択プログラム。

【請求項２３】訳語出力ステップにおいて、類似度出力ステップで出力し類似性評価ステップで評価した結果、類似度が最大となる原言語用例データが複数ある場合に、前記差異演算ステップにおける演算の結果、入力テキストと一致した文字列又は前記分割数が最大の原言語用例を含む対訳用例データにおける前記翻訳対象語に対応する訳語を出力する請求項２２記載の訳語選択装置。

【請求項２４】第１言語によるテキストからなる原言語用例及びそれに含まれる語とその語の第２言語による訳語及び当該訳語に関する情報とを含む原言語用例データと、前記原言語用例から第２言語で翻訳されたテキストからなる目的言語用例とを対にした対訳用例データを格納する対訳用例データ格納部を利用してコンピュータを動作させ、第１言語で入力された入力テキストに含まれる翻訳すべき語である翻訳対象語に対応する第２言語で記述された訳語を選択するプログラムであって、

前記入力テキストの入力を受け付ける入力受付ステップと、

対訳用例データ格納部に格納された原言語用例に含まれる語及び当該原言語用例に対応する対訳用例データに基づいて作成された学習データを利用して、入力受付部で受け付けた入力テキスト中の翻訳対象語に対応した学習モデルを生成する学習モデル生成ステップと、

前記生成した学習モデルを入力テキスト中の翻訳対象語に適用し、当該翻訳対象語の訳語候補の全てについて確信度を演算し、確信度順に順序付けて訳語候補を出力する学習モデル適用ステップと、

前記出力した訳語候補のうち、最も高い確信度が得られた訳語候補を選択して翻訳対象語に対応する訳語として出力する訳語出力ステップとを有することを特徴とする訳語選択プログラム。

【請求項２５】学習モデル生成ステップにおいて、入力受付ステップで受け付けた入力テキスト中の翻訳対象語ごとにそれを含む原言語用例に対応する対訳用例データを前記対訳用例データ格納部から抽出し、その抽出された対訳用例データに基づいて学習モデルを生成する請求項２４記載の訳語選択プログラム。

【請求項２６】学習モデル生成ステップにおいて、学習データごとに対応して学習モデルを生成し、入力受付ステップで受け付けた入力テキスト中の前記翻訳対象語ごとに前記学習データで精度が最高となる学習モデルを選択する学習モデル選択ステップをさらに含み、学習モデル適用ステップが、前記学習モデル選択ステップで選択した学習モデルを入力テキスト中の翻訳対象語に適用するものである請求項２４又は２５記載の訳語選択プログラム。

【請求項２７】第１言語によるテキストからなる原言語用例及びそれに含まれる語とその語の第２言語による訳語及び当該訳語に関する情報とを含む原言語用例データと、前記原言語用例から第２言語で翻訳されたテキストからなる目的言語用例とを対にした対訳用例データを格納する対訳用例データ格納部を利用してコンピュータを動作させ、第１言語で入力された入力テキストに含まれる翻訳すべき語である翻訳対象語に対応する第２言語で記述された訳語を選択するプログラムであって、

前記入力テキストの入力を受け付ける入力受付ステップと、

前記受け付けた入力テキスト中の前記翻訳対象語に該当する語を含む少なくとも一以上の原言語用例データを、前記対訳用例データ格納部から抽出する用例抽出ステップと、

前記入力テキスト及び前記抽出した原言語用例データに基づき、入力テキストと原言語用例との類似性を検出する類似性検出ステップと、

前記検出した原言語用例の類似性を比較評価し、最も高い類似性を有する少なくとも原言語用例データを出力する類似性評価ステップと、

対訳用例データ格納部に格納された原言語用例に含まれる語及び当該原言語用例に対応する対訳用例データに基づいて作成された学習データを利用して、入力受付部で受け付けた入力テキスト中の翻訳対象語に対応した学習モデルを生成する学習モデル生成ステップと、

前記生成した学習モデルを入力テキスト中の翻訳対象語に適用し、当該翻訳対象語の訳語候補の全てについて確信度を演算し、確信度順に順序付けて訳語候補を出力する学習モデル適用ステップと、

類似性評価ステップで出力する原言語用例データに対応する対訳用例データに含まれる目的言語用例中の前記翻訳対象語に対応する訳語、又は、学習モデル適用ステップで出力する訳語候補から、最適のものを選択して翻訳対象語に対応する訳語として出力する訳語出力ステップとを有することを特徴とする訳語選択プログラム。

【請求項２８】訳語出力ステップにおいて、類似性評価ステップで所定の閾値以上の類似性が得られた対訳用例データの出力がある場合に、当該類似性評価ステップで出力した結果得られる翻訳対象語に対応する訳語を出力し、類似性評価ステップで所定の閾値以上の類似性が得られた対訳用例データの出力がない場合に、前記学習モデル適用部で出力した結果得られる翻訳対象語に対応する訳語を出力するものである請求項２７記載の訳語選択プログラム。

【請求項２９】類似性評価ステップにおいて所定の閾値以上の類似性が得られた対訳用例データの出力がない場合に、前記学習モデル生成ステップ、学習モデル適用ステップ及び訳語出力ステップを経るようにしている請求項２８記載の訳語選択装置。

【請求項３０】第１言語によるテキストからなる原言語用例及びそれに含まれる語とその語の第２言語による訳語及び当該訳語に関する情報とを含む原言語用例データと、前記原言語用例から第２言語で翻訳されたテキストからなる目的言語用例とを対にした対訳用例データを格納する対訳用例データ格納部を利用してコンピュータを動作させ、第１言語で入力された入力テキストに基づいてその第２言語による翻訳文である対象テキストを出力するものであって、
前記入力テキストの入力を受け付ける入力受付ステップと、
前記受け付けた入力テキスト中の各翻訳対象語に該当する語を含む少なくとも一以上の原言語用例データを、前記対訳用例データ格納部から抽出する用例抽出ステップと、
前記入力テキスト及び前記抽出した原言語用例データに基づき、入力テキストと原言語用例との類似性を検出する類似性検出ステップと、
前記検出した原言語用例の類似性を比較評価し、最も高い類似性を有する少なくとも原言語用例データを出力する類似性評価ステップと、
前記出力した原言語用例データに対応する対訳用例データに含まれる目的言語用例中の前記翻訳対象語に対応する訳語を出力する訳語出力ステップと、
前記出力した訳語及び当該訳語を含む対訳用例データに基づいて、入力テキストに対応する対象テキストを生成し出力する翻訳文出力ステップとを有することを特徴とする翻訳プログラム。
【請求項３１】第１言語によるテキストからなる原言語用例及びそれに含まれる語とその語の第２言語による訳語及び当該訳語に関する情報とを含む原言語用例データと、前記原言語用例から第２言語で翻訳されたテキストからなる目的言語用例とを対にした対訳用例データを格納する対訳用例データ格納部を利用してコンピュータを動作させ、第１言語で入力された入力テキストに基づいてその第２言語による翻訳文である対象テキストを出力するものであって、
前記入力テキストの入力を受け付ける入力受付ステップと、
対訳用例データ格納部に格納された原言語用例に含まれる語及び当該原言語用例に対応する対訳用例データに基づいて作成された学習データを利用して、入力受付部で受け付けた入力テキスト中の翻訳対象語に対応した学習モデルを生成する学習モデル生成ステップと、
前記生成した学習モデルを入力テキスト中の翻訳対象語に適用し、当該翻訳対象語の訳語候補の全てについて確信度を演算し、確信度順に順序付けて訳語候補を出力する学習モデル適用ステップと、
前記出力した訳語候補のうち、最も高い確信度が得られた訳語候補を選択して翻訳対象語に対応する訳語として出力する訳語出力ステップと、
前記出力した訳語及び当該訳語を含む対訳用例データに基づいて、入力テキストに対応する対象テキストを生成し出力する翻訳文出力ステップとを有することを特徴とする翻訳プログラム。
【請求項３２】第１言語によるテキストからなる原言語用例及びそれに含まれる語とその語の第２言語による訳語及び当該訳語に関する情報とを含む原言語用例データと、前記原言語用例から第２言語で翻訳されたテキストからなる目的言語用例とを対にした対訳用例データを格納する対訳用例データ格納部を利用してコンピュータを動作させ、第１言語で入力された入力テキストに基づいてその第２言語による翻訳文である対象テキストを出力するものであって、
前記入力テキストの入力を受け付ける入力受付ステップと、
前記受け付けた入力テキスト中の各翻訳対象語に該当する語を含む少なくとも一以上の原言語用例データを、前記対訳用例データ格納部から抽出する用例抽出ステップと、
前記入力テキスト及び前記抽出した原言語用例データに基づき、入力テキストと原言語用例との類似性を検出する類似性検出ステップと、
前記検出した原言語用例の類似性を比較評価し、最も高い類似性を有する少なくとも原言語用例データを出力する類似性評価ステップと、
対訳用例データ格納部に格納された原言語用例に含まれる語及び当該原言語用例に対応する対訳用例データに基づいて作成された学習データを利用して、入力受付部で受け付けた入力テキスト中の翻訳対象語に対応した学習モデルを生成する学習モデル生成ステップと、
前記生成した学習モデルを入力テキスト中の翻訳対象語に適用し、当該翻訳対象語の訳語候補の全てについて確信度を演算し、確信度順に順序付けて訳語候補を出力する学習モデル適用ステップと、
類似性評価ステップで出力する原言語用例データに対応する対訳用例データに含まれる目的言語用例中の前記翻訳対象語に対応する訳語、又は、学習モデル適用ステップで出力する訳語候補から、最適のものを選択して翻訳対象語に対応する訳語として出力する訳語出力ステップと、
前記出力した訳語及び当該訳語を含む対訳用例データに基づいて、入力テキストに対応する対象テキストを生成し出力する翻訳文出力ステップとを有することを特徴とする翻訳プログラム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、ある言語で入力されたテキストを他の言語へ翻訳する際に使用される訳語選択装置、翻訳装置及びそれらのプログラムに関するも

のである。

【０００２】

【従来の技術】機械翻訳において、ある言語で記述された文、句、節、又は単語等の原テキストと、その原テキストを別の言語に翻訳した翻訳テキストとを対にした対訳データを格納したデータベースが使用されることがある。特に最近では、単語だけでなく、単語を含む文や句等の用例のデータベース（以下、「対訳コーパス」と称する）が使用されるようになってきている。現在では、新聞や辞書等を言語資源とした多種多様な対訳コーパスがインターネット等で公開され、利用に供されている。

【０００３】機械翻訳では、訳語選択が重要な技術要素の一つとして考えられるが、対訳用例コーパスを用いた場合、単純には対訳データの量が多ければ多いほど用例の数や種類が多くなると考えられることから、単一の対訳コーパスのみを使用するのではなく、可能な限り多種類の対訳コーパス又は対訳データを収集し、それらを用いて機械翻訳を実行することが考えられている。この場合、翻訳対象となる原言語での入力テキストに基づいて収集された対訳コーパスを参照し、入力テキストと合致する或いは最も類似する用例を含む原テキストに対応する対訳テキストを翻訳結果として出力する、という用例ベースの訳語選択方法が最も単純な手法であると考えられる。この他にも、対訳コーパスに基づいて作成した学習データを学習モデルに適用し、単純な統計的に確からしい訳語を出力するという、学習ベースの訳語選択方法も考えられている。

【０００４】

【発明が解決しようとする課題】用例ベースの訳語選択方法では、多種多様な対訳コーパスを参照しているため、それだけ翻訳の正確さが向上するものと一応は推測することができる。しかしながら、多種類の対応する訳語が存在する多義性を有する原言語の単語についてみれば、上述の方法では、対訳コーパス中に入力テキストと同一又は類似する用例が存在しなければ、正しい訳語を出力することができず、柔軟性に欠けるという不具合がある。一方、学習ベースの訳語選択方法では、統計的に頻度が高い用例で用いられている訳語を優先的に出力するために、数多くの用例で一般的に用いられ出現頻度の高い当該単語の訳語の正確性は向上する一方で、出現頻度が低い訳語については翻訳の正確さが低下する。

【０００５】このような問題は、ある単語が他の語句と結びついて独特の表現となる、「慣用表現」を入力テキスト中に含む場合に生じることが多い。一例として、日本語において多義的な「買う」という単語が原テキストに含まれる場合について考えると、「本を買う」という表現と「反感を買う」という表現とでは、「買う」の意味が異なり、それによって「買う」に対応する英語の訳語が異なる。この場合、日英の対訳コーパスには、「物を買う」という場合における「買う」の訳語と同じ英訳語（ｂｕｙ）が使われる用例は多数あってその英訳語の出現頻度は高いと考えられるのに対して、「反感を買う」というような慣用表現では「買う」の英訳語（ａｎｔｉｐａｔｈｙ）が特殊なものであるためにその英訳語を含む「買う」の用例は少ないものと考えられる。

【０００６】また、いずれの訳語選択方法においても、精度の高い翻訳を実現するには、対訳コーパスを大量に収集する必要があるが、自然言語には多様なバリエーションがあり得るため、単に多数の対訳コーパスを収集する方法ではコンピュータ処理の負荷が高まるだけで、現実にはこのような方法によって短時間で正確な機械翻訳を実施するのは不可能であると考えられる。

【０００７】そこで本発明は、以上のような問題に鑑みて、機械翻訳において、装置に過剰な負荷を掛けることなく、訳語選択並びに翻訳を正確かつ適正に短時間で行うことができるようにすることを主たる目的としている。

【０００８】

【課題を解決するための手段】本発明は、基本的に、第１言語によるテキストからなる原言語用例及びそれに含まれる語とその語の第２言語による訳語及び当該訳語に関する情報とを含む原言語用例データと、原言語用例から第２言語で翻訳されたテキストからなる目的言語用例とを対にした対訳用例データを格納する対訳用例データ格納部を利用して、第１言語で入力された入力テキストに含まれる翻訳すべき語である翻訳対象語に対応する第２言語で記述された訳語を選択するものである。ここで利用する対訳用例データ格納部は、上述したいわゆる対訳コーパスに該当するが、一つ以上を利用すればその数は問わない。但し、複数の対訳用例データ格納部を利用すれば、用例数を増加させて訳語選択の正確性を向上することができる。また、対訳用例データ格納部は、以下に述べる訳語選択装置や翻訳装置の一構成要素としたり、これら訳語選択装置等と通信可能な別の装置に設けることが可能である。

【０００９】このようなものにおいて本発明は、図１に概略構成図を示すように、第１の訳語選択装置Ａ１の基本構成として、入力テキストの入力を受け付ける入力受付部１と、その受け付けた入力テキスト中の前記翻訳対象語に該当する語を含む少なくとも一以上の原言語用例データを対訳用例データ格納部Ｃから抽出する用例抽出部２と、抽出した原言語用例データと前記入力テキストとに基づき入力テキストと原言語用例との類似性を検出する類似性検出部３と、検出した原言語用例の類似性を比較評価して最も高い類似性を有する少なくとも原言語用例データを出力する類似性評価部４と、出力した原言語用例データに対応する対訳用例データに含まれる目的言語用例中の前記翻訳対象語に対応する訳語を出力する訳語出力部５とを有していることを特徴とするものである。

【００１０】このように構成することによって、入力テキスト中に含まれる翻訳対象語に対して、それが用いられている原言語用例との類似性が最も高い訳語を出力することができる。したがって、特に原言語で使用される慣用句等の出現頻度が低い語句の訳語選択に際して、あまりに多くの対訳用例データを利用することなく、またコンピュータ処理に多大な負荷を掛けることなく、適切な訳語選択を行うことが可能となる。

【００１１】特に、類似性検出部３において、好適な類似性の検出を行い得る態様としては、入力テキストと抽出された原言語用例データに含まれる原言語用例とを文字単位で比較して求められる差異に基づき入力テキストと原言語用例との一致した文字列の割合、又は一致した部分が何カ所に分割されて一致しているかを示す分割数の少なくともいずれか一方を用いて計算される類似度を類似性として演算するようにしたものが挙げられる。

【００１２】また、用例抽出部２で抽出した原言語用例についてそれ以後の処理の便宜を図るためには、この用例抽出部２において、抽出された原言語用例データに含まれる原言語用例に文末処理を施して処理済原言語用例を出力するようにすればよく、この場合、類似性検出部３において、入力テキストと処理済原言語用例との文字単位で比較した場合の差異の演算結果に基づいて、一致した文字列の当該処理済原言語用例の文字列に対する割合、又は一致した部分が何カ所に分割されて一致しているかを示す分割数の少なくともいずれか一方を類似度として演算するように構成することが望ましい。

【００１３】さらに、訳語出力部５において、類似性検出部３で演算の上、出力し類似性評価部４で評価した結果、類似度が最大となる原言語用例データが複数ある場合が想定される。この場合、前記演算の結果、入力テキストと一致した文字列又は前記分割数が最大の原言語用例を含む対訳用例データにおける翻訳対象語に対応する訳語を出力することで、最も適していると推定される訳語を出力することができる。

【００１４】また、入力テキストの受付後の処理を簡便化するには、入力受付部１において、入力テキストを形態素解析により翻訳対象語を自動抽出するようにしておくことが好ましい。なお、「形態素解析」とは、入力テキストを単語毎に分割し、それぞれに品詞を割り当てる等の解析処理をいい、所定の解析アルゴリズム及び解析用辞書データが用いられる。

【００１５】さらに対訳用例データが、原言語用例に含まれる語に基づいて生成された原言語見出し語を含むものである場合には、用例抽出部２において、少なくとも翻訳対象語に該当する原言語見出し語を含む原言語用例データを対訳用例データ格納部Ｃから抽出するようにすることで、対訳用例データ格納部Ｃからの原言語用例データの抽出処理を高速化することができる。

【００１６】さらにまた、対訳用例データが、原言語用例に含まれる語に基づいて生成された原言語見出し語とそれに対応する訳語に基づいて生成された目的言語見出し語とを有する場合には、用例抽出部２において、翻訳対象語に該当する原言語見出し語を含む原言語用例データを少なくとも抽出し、訳語出力部５において、類似性評価部４で出力した原言語用例データに含まれ且つ用例抽出部２で抽出した原言語見出し語に対応する目的言語見出し語を出力することで、訳語出力までの処理をさらに高速化することができる。

【００１７】また本発明は、図２に概略構成図を示すように、第２の訳語選択装置Ａ２の基本構成として、入力テキストの入力を受け付ける入力受付部１１と、対訳用例データ格納部に格納された原言語用例に含まれる語及び当該原言語用例に対応する対訳用例データに基づいて作成された学習データを利用して、入力受付部で受け付けた入力テキスト中の翻訳対象語に対応した学習モデルを生成する学習モデル生成部１２と、その生成した学習モデルを入力テキスト中の翻訳対象語に適用し、当該翻訳対象語の訳語候補の全てについて確信度を演算し、確信度順に順序付けて訳語候補を出力する学習モデル適用部１３と、その出力した訳語候補のうち最も高い確信度が得られた訳語候補を選択して翻訳対象語に対応する訳語として出力する訳語出力部１４とを有することを特徴としている。ここで、「学習データ」とは、対訳用例に基づいて作成された第１言語で入力される語、それに対応して第２言語で出力されるべき正解の訳語、及びそれらに付随する属性や素性等の情報をいう。また、「学習モデル」とは、前記学習データを利用して推定されたパラメータを含み機械学習の手法により生成される関数的モデルである。また、確信度の順序づけは、降順又は昇順の何れであるかを問わない。

【００１８】このような構成によれば、一定量の学習データを作成又は収集しておくと、それに基づいて生成した適切な学習モデルを翻訳対象となる目的言語に適用した上で、確信度の最も高い訳語候補、すなわち最も適切であると推測することができる訳語を出力することができる。したがって、このような訳語選択装置Ａ２であれば、訳語選択に際して、翻訳対象となる語句（単語）ごとに学習モデルを生成することで、各語句（単語）に応じた適切なモデルによって訳語を選択することができるようになる。

【００１９】特に学習モデル生成部１２において、入力受付部１１で受け付けた入力テキスト中の翻訳対象語ごとにそれを含む原言語用例に対応する対訳用例データを対訳用例データ格納部Ｃから抽出し、その抽出された対訳用例データに基づいて学習モデルを生成するように構成すれば、迅速且つ正確な訳語出力処理を行うことができる。

【００２０】また、出力する訳語の正確性を高めるためには、学習モデル生成部１２において、学習データを利

用し各学習データごとにそれぞ学習モデルを生成し、さらに入力受付部１１で受け付けた入力テキスト中の翻訳対象語ごとに学習データで精度が最高となる学習モデルを選択し、学習モデル適用部１３において、学習モデル生成部１２で選択した最高の精度を得た学習モデルを入力テキスト中の翻訳対象語に適用するようにするとよい。なお、利用する学習データ数は一つであってもよいし複数であってもよい。

【００２１】また、この訳語選択装置Ａ２においても、入力受付部１１において、入力テキストを形態素解析により翻訳対象語を自動抽出することで、入力テキストの受付後の処理を簡便化することができる。同様に、対訳用例データに、原言語用例に含まれる語に基づいて生成された原言語見出し語が含まれる場合には、学習モデル生成部１２が、少なくとも翻訳対象語に該当する原言語見出し語を含む原言語用例データを対訳用例データ格納部Ｃから抽出するようにすることで、対訳用例データ格納部Ｃからの原言語用例データの抽出処理を高速化することができる。

【００２２】本発明の訳語選択装置はまた、上述した２種類の訳語選択装置Ａ１、Ａ２を組み合わせた態様として、出力される訳語の精度を飛躍的に向上させることもできる。すなわち、本発明は、図３に概略構成図を示すように、第３の訳語選択装置Ａ３の基本構成として、入力テキストの入力を受け付ける入力受付部３１と、入力受付部１で受け付けた入力テキスト中の前記翻訳対象語に該当する語を含む少なくとも一以上の原言語用例データを、前記対訳用例データ格納部Ｃから抽出する用例抽出部３２と、入力テキスト及び用例抽出部で抽出した原言語用例データに基づき入力テキストと原言語用例との類似性を検出する類似性検出部３３と、類似性検出部３で検出した原言語用例の類似性を比較評価し最も高い類似性を有する少なくとも原言語用例データを出力する類似性評価部３４と、対訳用例データ格納部Ｃに格納された原言語用例に含まれる語及び当該原言語用例に対応する対訳用例データに基づいて作成された学習データを利用して、入力受付部３１で受け付けた入力テキスト中の翻訳対象語に対応した学習モデルを生成する学習モデル生成部３５と、学習モデル生成部３５で生成した学習モデルを入力テキスト中の翻訳対象語に適用し、当該翻訳対象語の訳語候補の全てについて確信度を演算し、確信度順に順序付けて訳語候補を出力する学習モデル適用部３６と、類似性評価部３４で出力する原言語用例データに対応する対訳用例データに含まれる目的言語用例中の翻訳対象語に対応する訳語、又は、学習モデル適用部３６で出力する訳語候補から、最適のもの、すなわち前記訳語又は最高の確信度を得た訳語候補のいずれかを選択して翻訳対象語に対応する訳語として出力する訳語出力部３７とを有することを特徴とするものである。

【００２３】すなわち、入力受付部３１で受け付けた入力テキスト及び対訳用例データ格納部Ｃに格納される対訳用例データに基づいて、第１の訳語選択装置Ａ１に該当する用例抽出部３２、類似性検出部３３及び類似性評価部３４により処理された訳語、或いは第２の訳語選択装置Ａ２に該当する学習モデル生成部３５及び学習モデル適用部３６により処理された訳語候補のいずれかを、訳語出力部３７において出力する。なお、第１の訳語選択装置Ａ１該当部分と第２の訳語選択装置Ａ２該当部分とが利用する対訳用例データ格納部Ｃは、同一のものであってもよいし異なっていてもよい。

【００２４】この場合、望ましくは次の二態様の何れかを採用することが好適である。

【００２５】すなわち、まず、第１の訳語選択装置Ａ１該当部分と、第２の訳語選択装置Ａ２該当部分とを並列的に動作させ、訳語出力部３７において、類似性評価部３４で所定の閾値以上の類似性が得られた対訳用例データの出力がある場合に、その結果得られる翻訳対象語に対応する訳語を出力し、所定の閾値以上の類似性が得られた対訳用例データの出力がない場合に、学習モデル適用部３６で出力した結果得られる翻訳対象語に対応する訳語を出力する態様をとることができる。このようにすれば、並列処理により迅速に訳語を出力できることになる。

【００２６】一方、第１の訳語選択装置Ａ１該当部分をまず動作させ、類似性評価部３４において所定の閾値以上の類似性が得られた対訳用例データの出力がない場合に、第２の訳語選択装置該当部分Ａ２である前記学習モデル生成部３５、学習モデル適用部３６を動作させたうえで、訳語出力部３７を動作させるようにする態様をとることもできる。このようにすれば、類似性評価部３４において閾値以上の類似性が得られた対訳用例データがあれば、第２の訳語選択装置該当部分Ａ２を動作させる必要がないためコンピュータ処理に掛かる負荷を低減するとともに、第２の訳語選択装置Ａ２該当部分を動作させる際に、異なる対訳用例データ格納部Ｃを利用するなど、必要に応じて対訳用例データを追加収集又は取捨選択することができる。

【００２７】上記いずれの態様であっても、用例抽出部３２が利用する対訳用例データ格納部と、学習モデル生成部３５が利用する対訳用例データ格納部Ｃとが、それぞれ異なる言語資源に基づいて作成された異なるものであれば、対訳用例の数及び種類をより多様なものとして、最終的に出力される訳語の正確性を向上することが可能となる。

【００２８】また本発明は、以上のような訳語選択装置Ａ１、Ａ２、Ａ３の何れかを利用して、好適な翻訳装置を構成することも可能である。すなわち、当該翻訳装置は、訳語選択装置Ａ１、Ａ２、Ａ３の構成に加えて、それら何れかにおける訳語出力部で出力した訳語及び当該訳語を含む対訳用例データに基づいて、入力テキストに

対応する対象テキストを生成し出力する翻訳文出力部を更に備えたものである。このようにすれば、単に入力テキスト中の翻訳対象語に対応する訳語選択を行うのみならず、第１言語による入力テキストに基づいて第２言語で翻訳された対象テキストを生成して出力することまで可能となる。

【００２９】

【発明の実施の形態】以下、本発明の一実施形態を、図４～図８を参照して説明する。

【００３０】図４に概略構成図を示すこの実施形態は、上述した第３の基本構成を有する訳語選択装置Ａ３である。すなわち、第１の基本構成を有する訳語選択装置Ａ１に該当する部分と、第２の基本構成を有する訳語選択装置Ａ２に該当する部分と、これらに共通する部分とから構成される。また、対訳用例データ格納部Ｃは、この訳語選択装置Ａ３に含まれるものとしているが、必要に応じて通信回線で接続された他の装置に設けてある対訳用例データ格納部Ｃから収集することも可能である。なお、本実施形態では、第１言語（原言語）として日本語を、第２言語（目的言語）として英語を適用した場合について説明するものとする。

【００３１】まず、対訳用例データ格納部Ｃについて説明する。対訳用例データ格納部Ｃは、日本語によるテキストからなる用例（以下、「日本語用例」）及び当該日本語用例に含まれる語とその語の英語による訳語（以下、英訳語）並びに当該英訳語に関する各種情報とを含む日本語用例データと、前記日本語用例に対応して英語に翻訳されたテキストからなる英語用例を含む英語用例データとを対にした日英対訳用例データを格納してあるデータベースである。なお、日英対訳用例データにはさらに、日本語用例毎に翻訳対象語となり得る日本語見出し語が含まれており、場合によっては当該日本語見出し語に対応する正しい訳語となり得る英語見出し語が含まれる場合がある。このような日英対訳用例データとしては、例えば新聞や雑誌等の記事に基づき出現頻度等を考慮して作成されたデータベースや、日英対訳電子辞書データベース、その他オンライン上で利用可能なデータベース等に格納されたデータを利用することができる。

【００３２】ここで、日英対訳用例データの一例の一部を図５に示す。この例では、日本語「遠慮」という語を含む３つの日本語用例と、それらに対応する英語用例とが組になっている。この場合、日本語見出し語には「遠慮」が該当し、英語見出し語には「feel constrained」、「constraint」、「refrain」等が該当する。但し、日本語見出し語に対応する英語見出し語のみ、或いは日本語見出し語と英語見出し語の両方に関しては、既に設定されたものがある場合はそれを利用すればよく、ない場合は人手で設定するか或いはコンピュータ処理により自動的に設定されるようにしておく必要がある。

【００３３】次に訳語選択装置Ａ３の機能について説明する。この訳語選択装置Ａ３は、汎用コンピュータ又は専用コンピュータのＨＤＤ等の記憶装置に記憶させた所定のプログラムに従ってＣＰＵやメモリ等の通常のコンピュータが有する内部及び外部装置が動作することによって、第１の訳語選択装置Ａ１としての機能を奏する用例抽出部３２、類似性検出部３３、類似性評価部３４と、第２の訳語選択装置Ａ２としての機能を奏する学習モデル生成部３５、学習モデル適用部３６と、これらに共通の機能を奏する入力受付部３１、訳語出力部３７としての機能を発揮する。

【００３４】入力受付部３１は、日本語で作成されたテキストデータ（入力テキスト）の入力を受け付ける。この入力受付部３１には、入力テキスト処理部３１１が含まれる。入力テキスト処理部３１１は、前記入力テキストに対して形態素解析を行い、当該入力テキストから翻訳対象語を自動的に抽出する。なお、入力テキストの入力時に、翻訳対象語を指定しておくことができるが、この場合は入力テキスト処理部３１１にて形態素解析のみを行う。

【００３５】用例抽出部３２は、入力受付部３１で得られた翻訳対象語が含まれた日本語用例データを、対訳用例データ格納部Ｃを抽出する。その際、対訳用例データ格納部Ｃに日本語見出し語が含まれている場合にはそれを参照して該当する翻訳対象語を検索のうえ抽出を行う。この用例抽出部３２には、原言語用例処理部たる日本語用例処理部３２１が含まれる。この日本語用例処理部３２１は、対訳用例データ格納部Ｃから抽出した日本語用例データについて、文末処理を行うものである。例えば上述の図５に示す日英対訳用例データのうち、日本語用例データについて文末処理を行うことによりと、「母に遠慮する」、「母への遠慮」、「献金を遠慮してもらう」は、それぞれ「母に遠慮」、「母への遠慮」、「献金を遠慮」となる。

【００３６】類似性検出部３３は、入力受付部３１で受け付けた入力テキストと、用例抽出部３２で抽出した日本語用例データとを対比し、それらの類似性を検出する。具体的にはこの類似性検出部３３に含まれる類似度演算部３３１により演算された入力テキストと日本語用例データとの一致する割合である類似度が前記類似性として検出される。すなわち、類似度は、動的計画法により入力テキストと日本語用例データとを文字単位で比較して両者の差異を求め、一致した文字列の割合として求められる。より具体的に類似度は、例えばＵＮＩＸ（登録商標）のｄｉｆｆコマンドにより次式

【００３７】

【式１】

$$\text{類似度}=\frac{(\text{入力テキストと日本語用例とのdiff処理時に一致した文字数})}{(\text{文末処理した日本語用例の文字数})}$$

【００３８】により求められる。なお、日本語用例データは、日本語用例処理部３２１で文末処理を施したものを利用する。

【００３９】類似性評価部３４は、入力テキストと対比された各日本語用例データについて類似性検出部３３で検出した類似性、すなわち前式で得られた類似度を比較評価し、最も高い類似度ｒが得られた日本語用例データ又はその日本語用例データを含む日英対訳用例データを出力する。このとき、最大の類似度ｒが得られた日本語用例データが複数あった場合は、最長の日本語用例を含む日本語用例データを最も高い類似性を有するものとして出力する。但し、入力テキストと一致した部分が日本語見出し単語の長さよりも長い場合に限られる。

【００４０】学習モデル生成部３５は、学習データを利用して入力受付部３１で受け付けた入力テキスト中の翻訳対象語毎に対応した学習モデルを生成する。学習データは、対訳用例データ格納部Ｃに格納された日本語用例に含まれる語とその日本語用例に対応する英語用例データとに基づいて作成されたものであり、日本語で入力される語、それに対応して英語で出力されるべき正解の訳語、及びそれらに付随する属性や素性等の情報等からなる。また、本実施形態では学習モデルとして、例えばＳＶＭ（Support Vector Machine）、ＭＥ（Maximum Entropy）、ＤＬ（Decision List）等の既知の機械学習モデルを複数種類適用することとしている。そして、これら学習モデルを各翻訳対象語に適用することにより、それぞれの正解の訳語が生成される確率を求める。その際、各学習モデルには、素性を与える必要があるが、本実施形態では素性として、前記学習データから得られた情報である形態素情報、文字n-gram、最大一致となる日本語用例に関する情報、内容語とその訳語候補の出現頻度に関する情報の４種類の情報を用いている。この学習モデル生成部３５には、学習モデル選択部３５１が含まれる。この学習モデル選択部３５１は、各学習モデルについて学習データを用いてクロスバリデーションを行い精度が最高となる学習モデルを選択する。

【００４１】学習モデル適用部３６は、学習モデル生成部３５で生成した学習モデル、具体的には学習モデル選択部３５１で選択した学習モデルを入力テキスト中の翻訳対象語に適用することにより、その翻訳対象語の訳語候補の全てについて確信度を演算し、確信度順に順序付けを行って訳語候補を出力する。この確信度は基本的に、文脈の集合をＢ、分類クラスの集合をＡとした場合、文脈ｂ（∈Ｂ）でクラスａ（∈Ａ）となる事象（ａ，ｂ）の確率分布のスコアｐ（ａ，ｂ）として求められる。なお、学習モデルの種類によってこのような確率分布が得られない場合、例えばＳＶＭを適用した場合、便宜的に最適のクラスに対して確率値を１、その他のクラスに対して確率値を０としている。

【００４２】訳語出力部３７は、入力テキスト中の翻訳対象語に対応する訳語を出力するものであり、訳語選択装置Ａ１のルート又は訳語選択装置Ａ２のルートの何れかから得られる訳語、すなわち、類似性評価部３４で最高の類似性を得た日本語用例データに該当する日英対訳用例データに含まれる訳語、又は、学習モデル適用部３６で出力した訳語候補のうち最高の確信度（スコア）を得た訳語候補、の何れかを選択して出力する。具体的に、本実施形態では、類似性検出部３３における類似性演算部３３１で得られる類似度に閾値を設定しており、類似性評価部３４で出力する日本語用例データが当該閾値以上の場合には、その日本語用例データに対応する訳語を出力する。本実施形態では前記閾値を１としている。一方、閾値以上の日本語用例データがない場合に、学習モデル適用部３６で出力した訳語候補から最高の確信度を得たものを出力する。なお、入力受付部３１で入力テキストを受け付けた際に、訳語選択装置Ａ１のルートと訳語選択装置Ａ２のルートとを同時に動作させてもよいし、訳語選択装置Ａ１のルートを先に動作させてから閾値以上の日本語用例データがない場合にのみ訳語選択装置Ａ２のルートを動作させてもよい。

【００４３】以下、本実施形態の訳語選択装置Ａ３の一利用態様例を、図６及び図７に示した訳語選択装置Ａ３の動作手順を表すフローチャートを用いて説明する。なお、以下の説明は、本発明の発明者が参加した（参加者名、CRL-NYU）単語の多義性解消コンテスト第２回ＳＥＮＳＥＶＡＬ｛以下、「ＳＥＮＳＥＶＡＬ２」、２００１年開催（SENSEVAL-2 Organization Committee）｝の日本語翻訳タスクに本実施形態の訳語選択装置Ａ３を適用したものであり、同コンテストにおいては訳語選択装置Ａ３の改良前のもので参加しているが、極めて高い評価を得ている。

【００４４】前提として、日英対訳用例データ（３２０語の日本語見出し語、一見出し語につき約２０の用例数）は前記コンテスト前に予め与えられたＳＥＮＳＥＶＡＬ２日本語翻訳タスクのものに準ずる。これらのうちから選択された４０語（名詞２０語、動詞２０語）について３０出現ずつのテストデータが用いられ、翻訳対象とされる日本語の単語はのべ１２００語である。また、コンテストのしゃん貨車は、与えられた日英対訳用例データ以外の言語資源から得た対訳辞書や各種新聞記事に基づく日英対訳用例データも用いることも許容されている。さらに、最終的に出力された訳語の正誤を公正に評価するために、所定の入力テキスト及び翻訳対象語と正解の訳語に基づいて、訳語の精度が評価されている。

【００４５】なお、説明を簡素化するため、ここではまず訳語選択装置Ａ１のルートから開始し、当該ルートから訳語が出力されなかった場合に訳語選択装置Ａ２のルートに移行する態様について説明するが、両ルートを同時に進行させてもよいのは上述したとおりである。まず、入力受付部３１が入力テキスト（例えば慣用表現で

ある「一役買う」の表現を含む日本語のテキスト）の入力を受け付ける（図６；ステップＳ１）と、入力テキスト処理部３１１がこの入力テキストを形態素解析することにより、翻訳対象語（例えば＜買う＞）を抽出する（ステップＳ２）。次に、用例抽出部３２が前記抽出された翻訳対象語（＜買う＞）に基づいて対訳用例データ格納部Ｃを検索し、当該翻訳対象語を含む日本語用例データを抽出し（ステップＳ３）、日本語用例処理部３２１が抽出した日本語用例データに含まれる各日本語用例について文末処理を行う（ステップＳ４）。次に、この文末処理が施された各日本語用例と前記入力テキストについて、類似性検出部３３における類似性演算部３３１が前記式１に基づいて類似度ｒを演算する（ステップＳ５）。そして、類似度ｒが最大となる日本語用例数を調べ（ステップＳ６）、その数が１であれば（ステップＳ６；Ｙ）、類似性評価部３４が、当該日本語用例を含む日本語用例データを出力する（ステップＳ７）。一方、ステップＳ６において類似度ｒが最大の日本語用例数が１以上であれば（ステップＳ６；Ｎ）、そのうち類似する文字列が最長の日本語用例を含む日本語用例データを選択し（ステップＳ６ａ）、その日本語用例データを最も高い類似性を有するものとして出力する（ステップＳ７）。ここで、この場合、類似度ｒが最高の日本語用例が、入力テキストに対応する表現（「一役買う」）を含んでおり、この日英対訳用例データにおける前記日本語用例に対応する英語用例に、翻訳対象語に対応する英訳語（＜ｔｏ　ｏｆｆｅｒ　ｔｏ　ｈｅｌｐ＞）が含まれていたものとする。そして、出力された日本語用例データの類似度と所定の閾値（例えば１）とを比較し（ステップＳ８）、類似度が閾値（１）以上であれば（ステップＳ８；Ｙ）、訳語出力部３７が、翻訳対象語（＜買う＞）に対応する英訳語（例えば＜ｏｆｆｅｒ＞）を出力する（ステップＳ９）。なお、「一役買う」という日本語の慣用表現に対応する英語の表現が、「ｔｏ　ｏｆｆｅｒ　ｔｏ　ｈｅｌｐ」であり、この場合、翻訳対象語「買う」に対する正解の英訳語が「ｏｆｆｅｒ」であると与えられていれば、ステップＳ９で出力した英訳語は正解となる。

【００４６】一方、ステップＳ８において、閾値（１）以上の日本語用例データがなかった場合（ステップＳ８；Ｎ）、すなわち、入力テキスト中の翻訳対象語を含む日本語用例と同一又は類似の用例が、いずれの日本語用例データがない場合、訳語翻訳装置Ａ２のルートに移行する｛Ｓ６（Ｎ）｝。この場合、学習モデル生成部３５において、まず入力受付部３１で受け付けた入力テキスト中の翻訳対象語に基づいて、前記訳語選択装置Ａ１のルートで用いたものとは別の日英対訳用例データ格納部Ｃを検索し、該当する語を含む日本語用例データを抽出する（図７、ステップＳ１１）。そして、抽出した各日本語用例データに含まれる日本語用例毎に学習データを適用して学習モデル（ＳＶＭ、ＤＬ、ＭＥのいずれかに基づく）を生成する（ステップＳ１２）。さらに、学習モデル選択部３５１によって、生成された各学習モデルについて、学習データを用いてクロスバリデーションを行ったうえで精度が最高となった学習モデルを選択する（ステップＳ１３）。ここで選択された学習モデルを、学習モデル適用部３６において入力テキスト中の翻訳対象語に適用して、それに対応する訳語候補の全てについて確信度ｐを演算し（ステップＳ１４）、確信度ｐ順に例えば降順で順序付けて訳語候補を出力する（ステップＳ１５）。最後に、出力した訳語候補から、最高の確信度ｐが得られた訳語候補を選択して訳語出力手段３７により出力する（ステップＳ１６）。この出力した訳語候補が、予め与えられた正解の英訳語と合致していれば、当該英訳語が正解となる。

【００４７】参考として、図８に、ＳＥＮＳＥＶＡＬ２のコンテストにおける訳語選択装置Ａ１及びＡ２による結果を一覧表にして示す。この結果は、コンテストで与えられた翻訳対象語である単語（名詞２０、動詞２０）ごとについて出力した英訳語の正解率を精度として示すものである。与えられたのべ１２００の翻訳対象語のうち、１００について訳語選択装置Ａ１を適用した結果、精度は９１．０％であった。また、１１００の翻訳対象語について訳語選択装置Ａ２を適用した結果、精度は６０．９％であった。なお、比較のため、これら訳語選択装置Ａ１、Ａ２による総合的な結果（Ａ１＋Ａ２）も同一覧表に示している。この結果から、訳語選択装置Ａ１について精度が芳しくなかった翻訳対象語については、訳語選択装置Ａ２を適用するという、本実施形態の訳語選択装置Ａ３を適用することが適切であるといえる。すなわち、文字列の類似性に基づく訳語選択装置Ａ１を適用するルートは、慣用的表現を含むなど一般に学習データ数が少ない用例、換言すればそのような日英対訳用例データ数が少ない用例に対して適しているといえ、一方、上記ルートで精度が悪い場合に学習データ及び学習モデルを適用して確信度を得る訳語選択装置Ａ２のルートを適用することで、通常用いられる表現は勿論のこと慣用的表現も含めて、全体として精度の高い訳語選択を実行することが可能であるといえる。

【００４８】本発明は、以上に説明した実施形態に限られるものではない。例えば、訳語選択装置Ａ１、Ａ２を単独で用いたり、訳語出力部で出力される訳語に基づいて入力テキストに対応する対象テキストを生成し出力する翻訳文出力部を設けることによって翻訳装置を構成することも可能である。また、その他、各部の具体的構成についても上記実施形態に限られるものではなく、本発明の趣旨を逸脱しない範囲で種々変形が可能である。

【００４９】

【発明の効果】以上に詳述したように、本発明によれば、多大な人手を掛けずコンピュータに過剰な負荷を掛

けることなく、すなわち、多量の対訳用例データを収集することなく、精度の高い訳語選択、並びに機械翻訳を行うことができる。特に、文字列の類似性に基づく方法と、学習データ及び学習モデルを適用する方法とをそれぞれ別個に用いたり、或いはそれらを併用することで相互に補完しあうことになり、通常用いられる自然言語の表現や、出現頻度の低い慣用的表現に対しても極めて精度の高い訳語選択及び機械翻訳が可能である。

【図面の簡単な説明】

【図１】本発明の第１の態様に対応する訳語選択装置の概略機能構成図。

【図２】本発明の第２の態様に対応する訳語選択装置の概略機能構成図。

【図３】本発明の第３の態様に対応する訳語選択装置の概略機能構成図。

【図４】本発明の一実施形態における訳語選択装置の概略機能構成図。

【図５】同実施形態に用いられる日英対訳用例データの一例を示す図。

【図６】同実施形態の動作手順を示す概略的なフローチャート。

【図７】同実施形態の動作手順を示す概略的なフローチャート。

【図８】本発明を適用したＳＥＮＳＥＶＡＬ２のコンテストにおける訳語選択結果を一覧表にして示す図。

【符号の説明】

Ａ１、Ａ２、Ａ３…訳語選択装置

Ｃ…対訳用例データ格納部

１、１１、２１、３１…入力受付部

２、３２…用例抽出部

３、３３…類似性検出部

４、３４…類似性評価部

５、１４、３７…訳語出力部

１２、３５…学習モデル生成部

１３、３６…学習モデル適用部

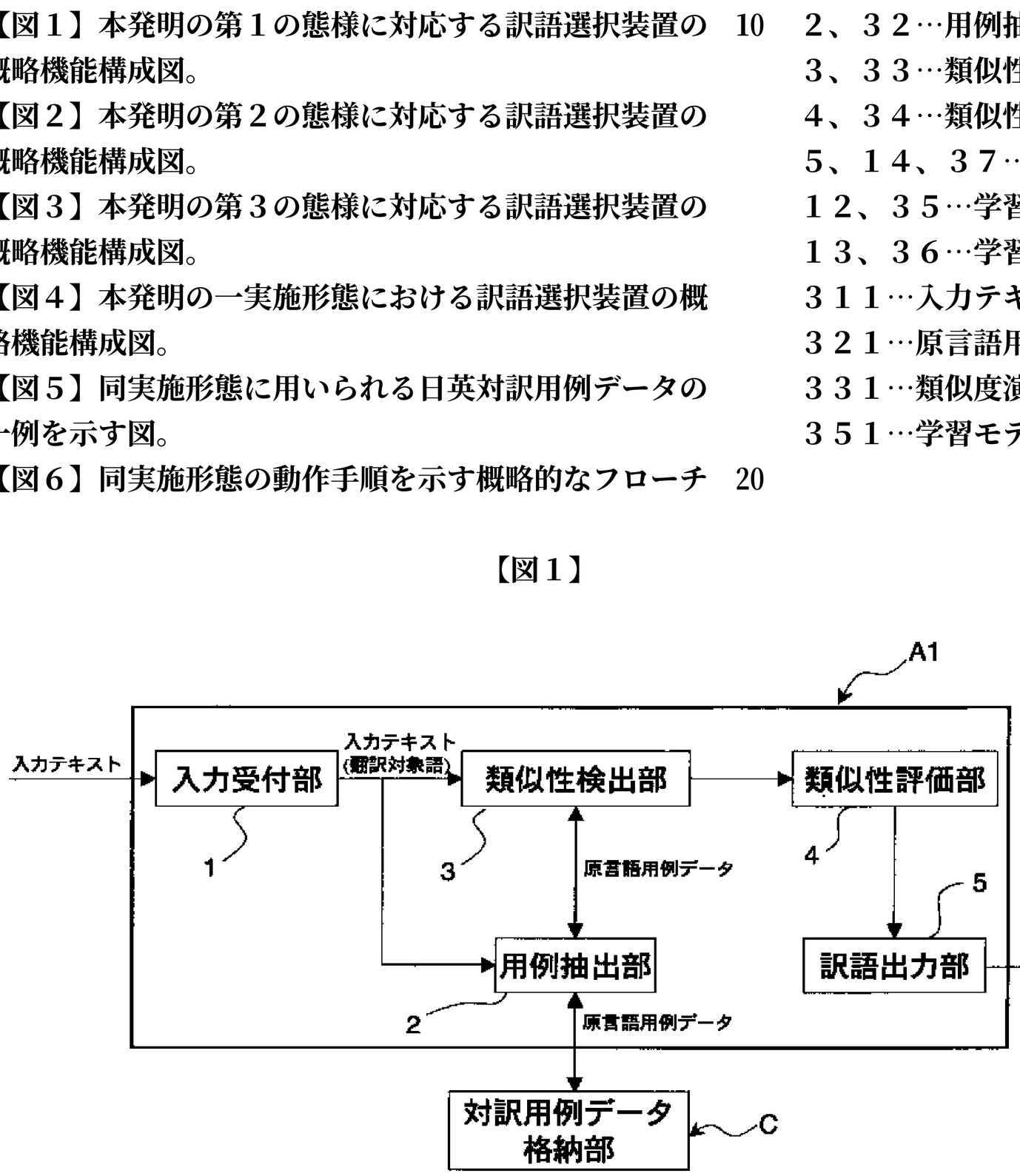
３１１…入力テキスト処理部

３２１…原言語用例処理部（日本語用例処理部）

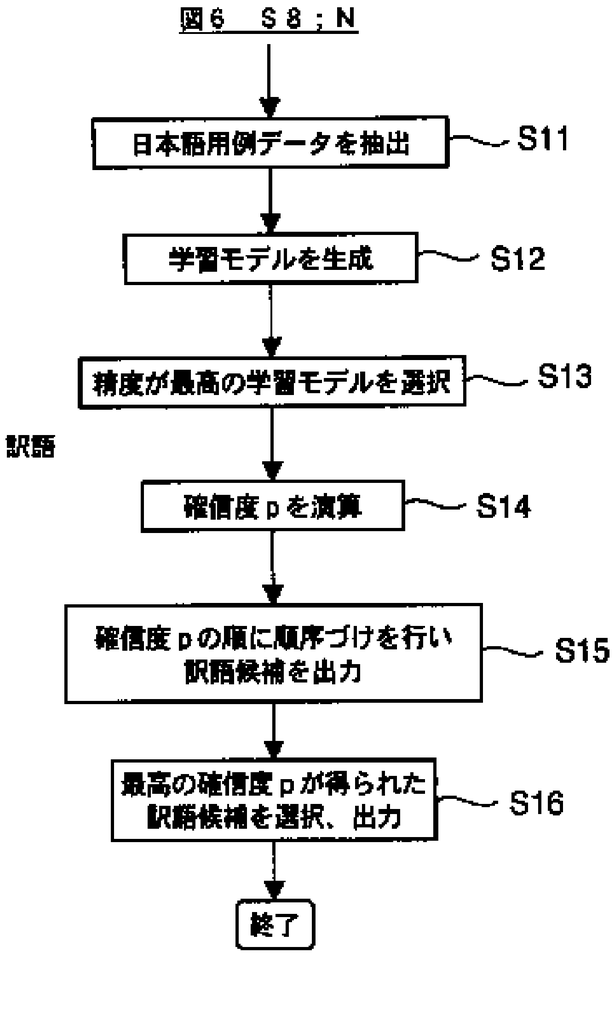
３３１…類似度演算部

３５１…学習モデル選択部

【図１】



【図５】

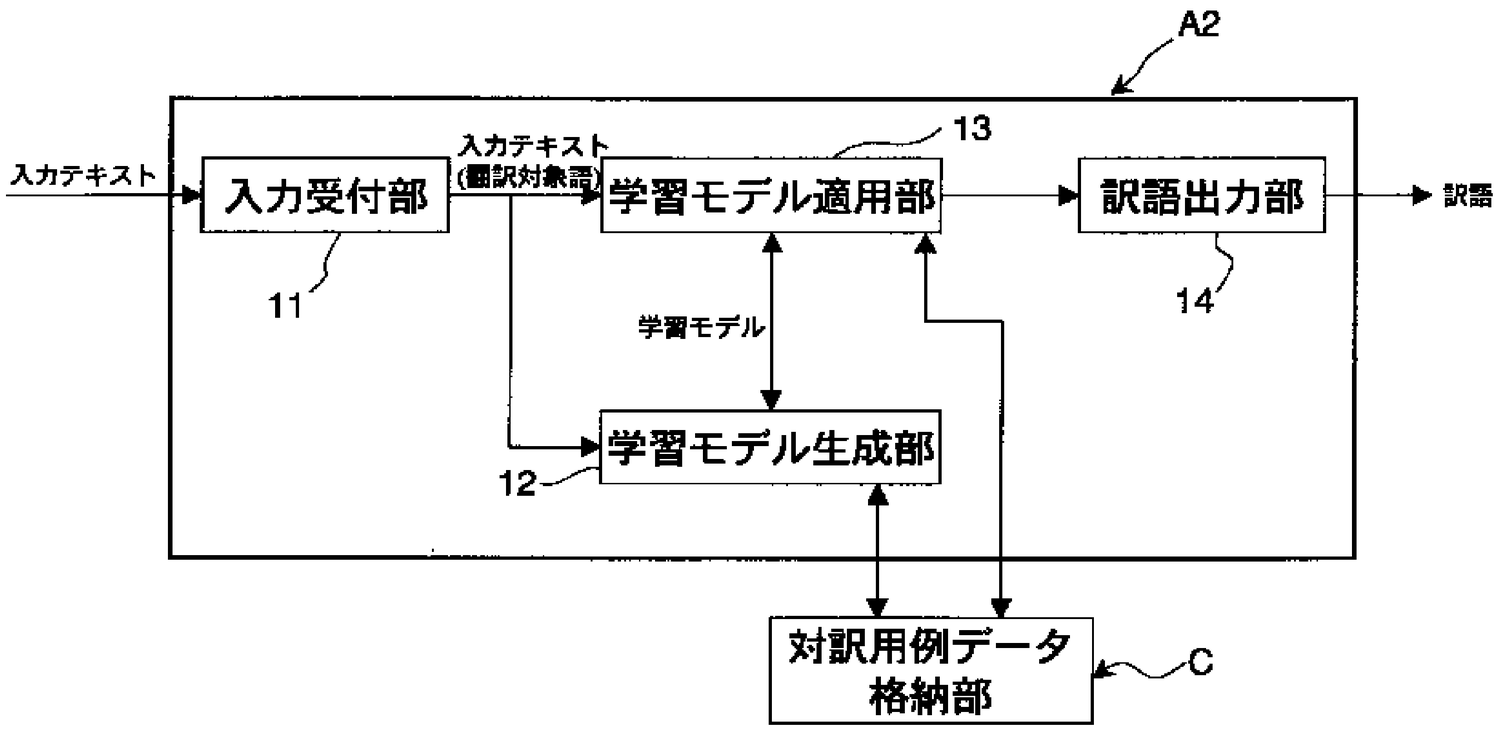


【図７】

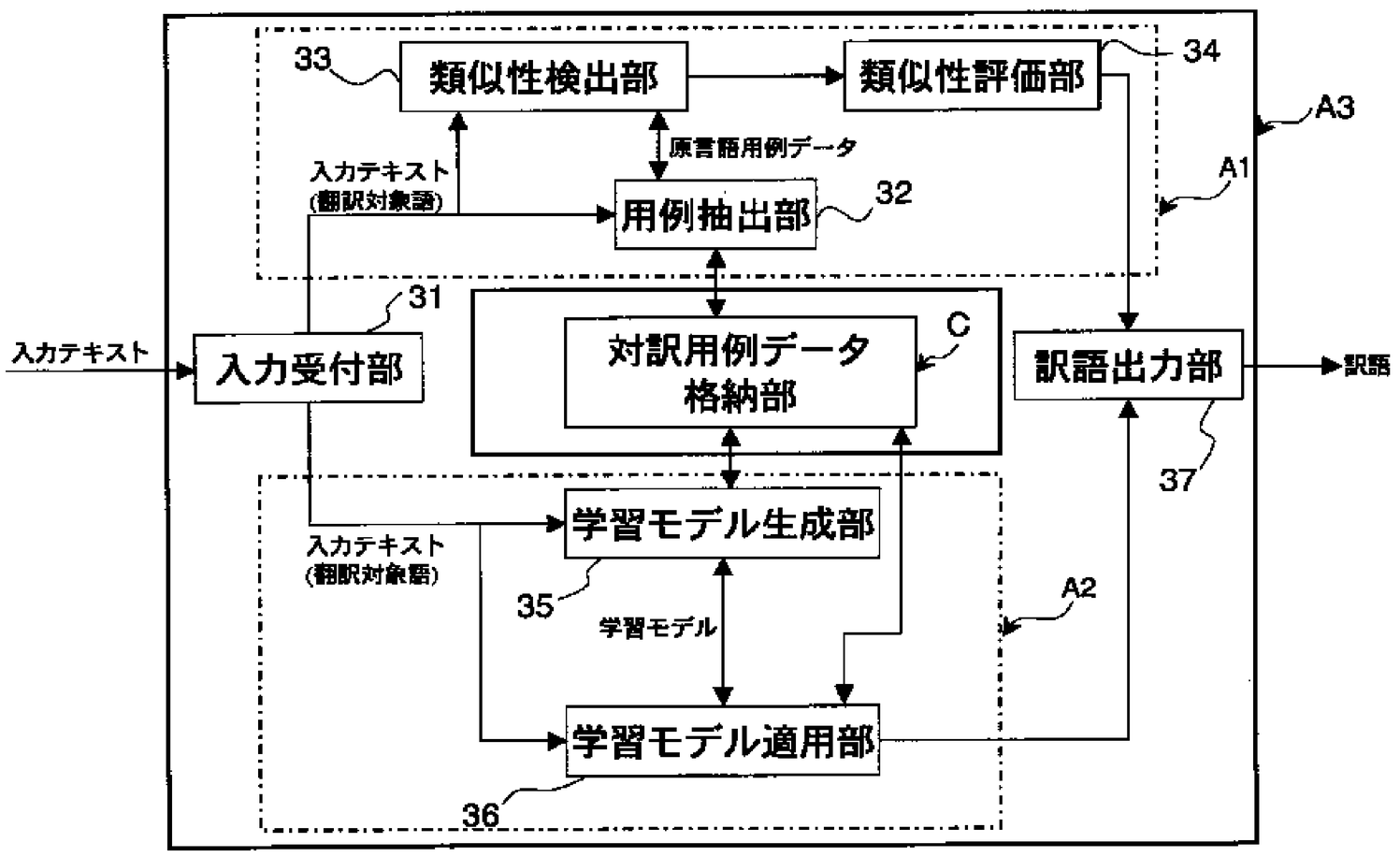
日英対訳用例データの一例（一部）

```
<entry id="1" headword="遠慮">
    <sense id="1-1">
        <jexpression> 母に遠慮する </ jexpression >
        <eexpression> to feel constrained for one's mother </ eexpression >
    </sense>
    <sense id="1-2">
        <jexpression> 母への遠慮 </ jexpression >
        <eexpression> constraint toward one's mother </ eexpression >
        <transmemo>UC</transmemo>
    </sense>
    <sense id="1-3">
        <jexpression> 献金を遠慮してもらう </ jexpression >
        <eexpression> to request to refrain from donation </ eexpression >
    </sense>
    .....
</entry>
```

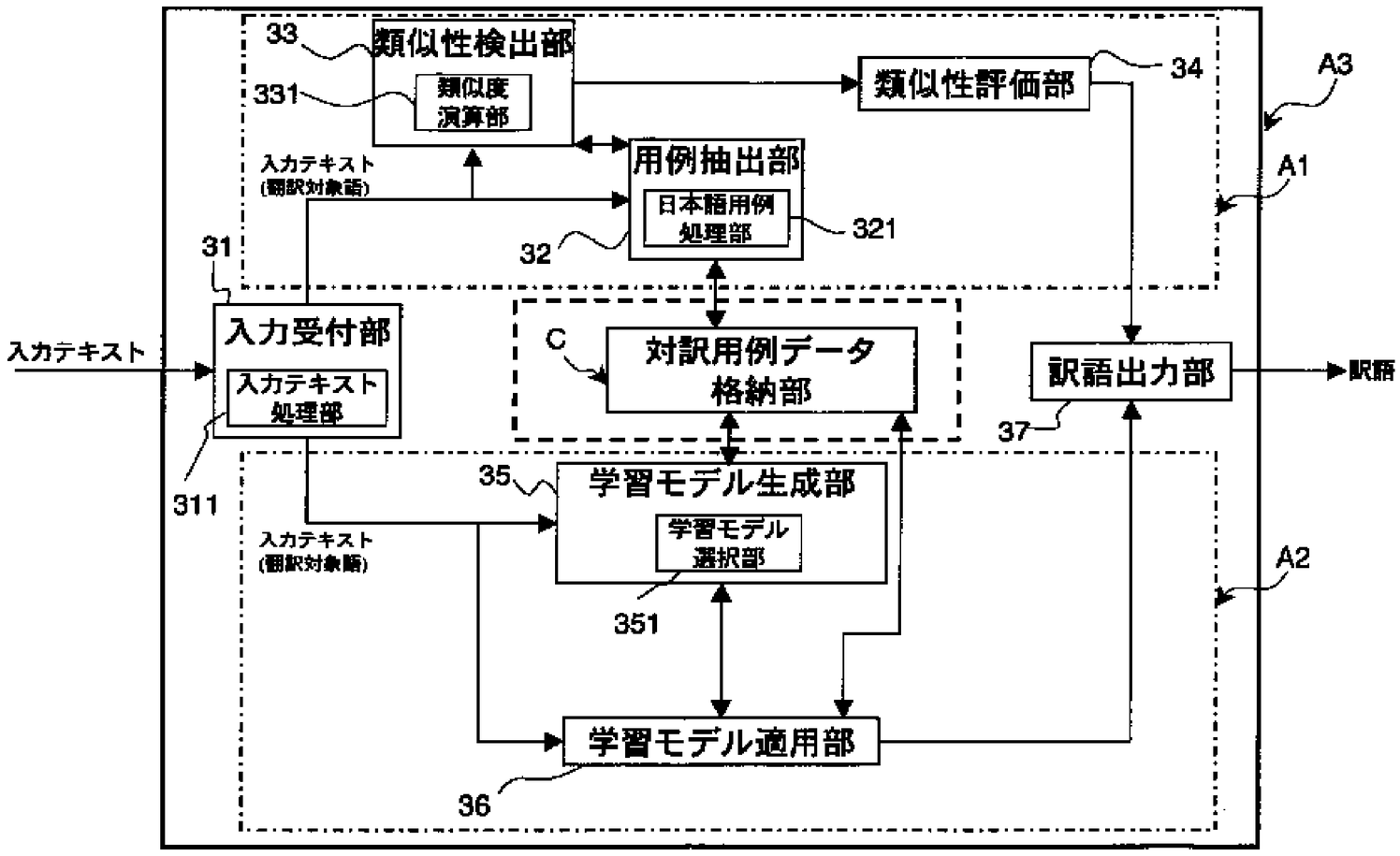
【図２】


A2
入力テキスト
入力受付部
入力テキスト
(翻訳対象語)
学習モデル適用部
訳語出力部
訳語
13
11
14
学習モデル
学習モデル生成部
12
対訳用例データ
格納部
C

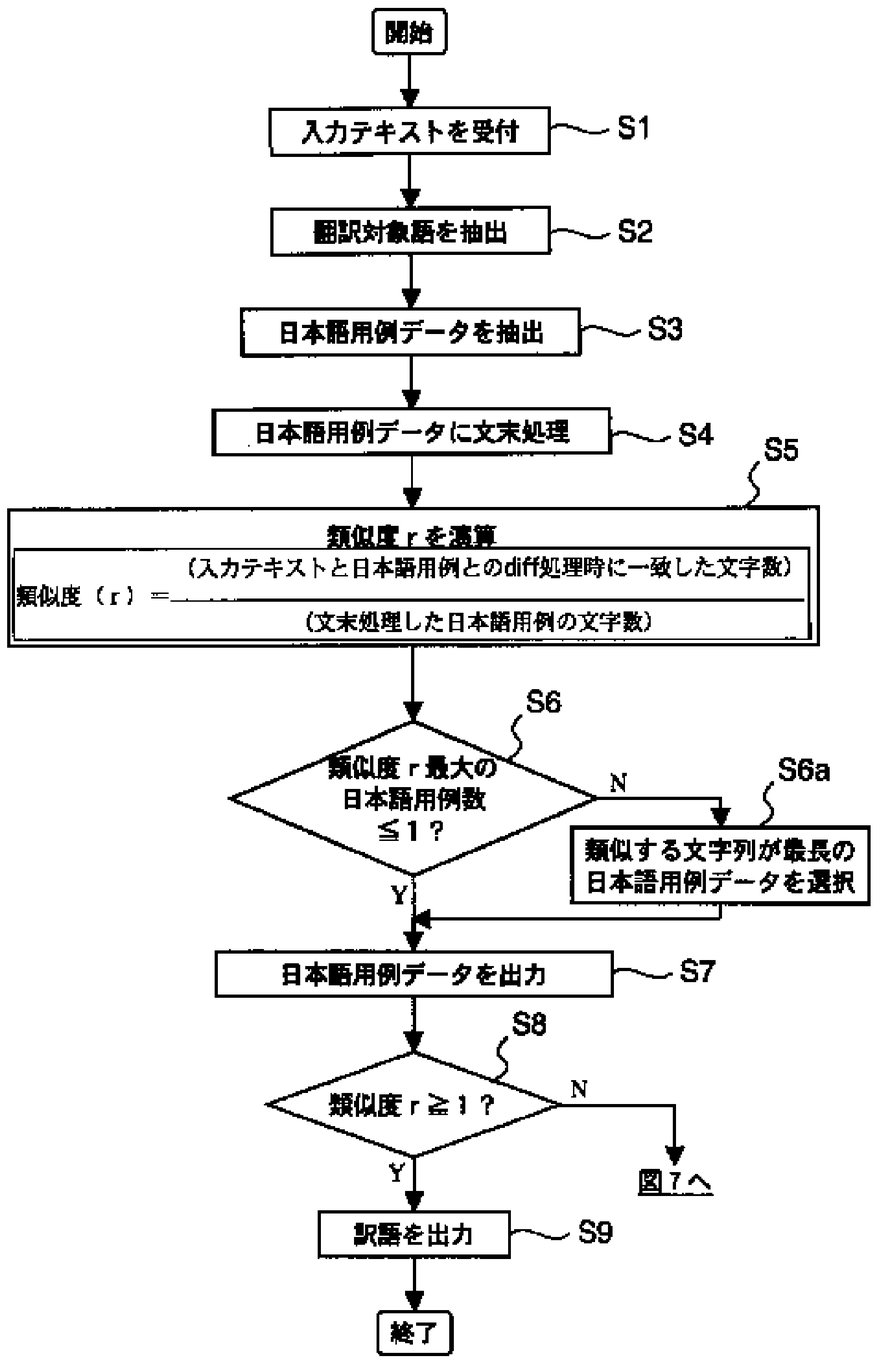
【図３】


33
類似性検出部
類似性評価部
34
A3
A1
原言語用例データ
入力テキスト
(翻訳対象語)
用例抽出部
32
31
入力テキスト
入力受付部
対訳用例データ
格納部
C
訳語出力部
訳語
37
学習モデル生成部
入力テキスト
(翻訳対象語)
35
A2
学習モデル
学習モデル適用部
36

【図４】


33
類似性検出部
331
類似度
演算部
類似性評価部
34
A3
A1
入力テキスト
(翻訳対象語)
用例抽出部
日本語用例
処理部
321
32
31
入力受付部
入力テキスト
処理部
入力テキスト
C
対訳用例データ
格納部
訳語出力部
訳語
37
311
35
学習モデル生成部
学習モデル
選択部
入力テキスト
(翻訳対象語)
A2
351
学習モデル適用部
36

【図６】


開始
入力テキストを受付
S1
翻訳対象語を抽出
S2
日本語用例データを抽出
S3
日本語用例データに文末処理
S4
S5
類似度ｒを演算
類似度（ｒ）＝（入力テキストと日本語用例とのdiff処理時に一致した文字数）／（文末処理した日本語用例の文字数）
S6
類似度ｒ最大の
日本語用例数
≦1？
N
S6a
類似する文字列が最長の
日本語用例データを選択
Y
日本語用例データを出力
S7
S8
類似度ｒ≧１？
N
図７へ
Y
訳語を出力
S9
終了

【図８】

単語ごとの精度

単語（読み）	用例数	学習文数	クラス数	学習文数／クラス	学習モデル	精度（A1+A2）		精度（A1）		精度（A2）	
一般　(ippan)	33	760	16	47.5	SVM	56.7%	(17/30)	66.7%	(2/3)	55.6%	(15/27)
一方　(ippou)	14	172	19	9.1	DL	23.3%	(7/30)	-		23.3%	(7/30)
今　(ima)	15	433	15	28.9	DL	63.3%	(19/30)	-		63.3%	(19/30)
意味　(imi)	22	419	18	23.3	SVM	66.7%	(20/30)	100.0%	(1/1)	65.5%	(19/29)
核　(kaku_n)	8	1,007	8	125.9	SVM	80.0%	(24/30)	100.0%	(3/3)	77.8%	(21/27)
記録　(kiroku)	18	608	11	55.3	SVM	80.0%	(24/30)	100.0%	(1/1)	79.3%	(23/29)
国内　(kokunai)	14	346	6	57.7	SVM	83.3%	(25/30)	75.0%	(3/4)	84.6%	(22/26)
言葉　(kotoba)	35	925	28	33.0	DL	80.0%	(24/30)	0.0%	(0/1)	82.8%	(24/29)
市民　(shimin)	23	187	8	23.4	DL	50.0%	(15/30)	100.0%	(5/5)	40.0%	(10/25)
事業　(jigyou)	17	854	14	61.0	SVM	63.3%	(19/30)	100.0%	(7/7)	52.2%	(12/23)
時代　(jidai)	39	621	10	62.1	DL	83.3%	(25/30)	100.0%	(4/4)	80.8%	(21/26)
姿　(sugata)	28	139	19	7.3	SVM	46.7%	(14/30)	80.0%	(4/5)	40.0%	(10/25)
近く　(chikaku)	15	123	11	11.2	SVM	73.3%	(22/30)	-		73.3%	(22/30)
中心　(chushin)	15	392	16	24.5	SVM	56.7%	(17/30)	-		56.7%	(17/30)
花　(hana)	27	677	20	33.9	SVM	83.3%	(25/30)	100.0%	(2/2)	82.1%	(23/28)
反対　(hantai)	26	480	17	28.2	SVM	93.3%	(28/30)	71.4%	(5/7)	100.0%	(23/23)
場合　(baai)	23	1,167	16	72.9	DL	86.7%	(26/30)	-		86.7%	(26/30)
前　(mae)	25	1,968	26	75.7	DL	63.3%	(19/30)	-		63.3%	(19/30)
胸　(mune)	30	368	26	14.2	DL	53.3%	(16/30)	100.0%	(3/3)	48.1%	(13/27)
問題　(mondai)	32	1,795	10	179.5	SVM	100.0%	(30/30)	100.0%	(2/2)	100.0%	(28/28)
全名詞	459	13,441	304	44.2		69.3%	(416/600)	87.5%	(42/48)	67.8%	(347/552)
与える(ataeru)	36	808	34	23.8	SVM	70.0%	(21/30)	100.0%	(3/3)	66.7%	(18/27)
言う　(iu)	32	2,248	21	107.0	DL	73.3%	(22/30)	50.0%	(1/2)	75.0%	(21/28)
受ける(ukeru)	22	5,143	25	205.7	SB	20.0%	(6/30)	50.0%	(1/2)	17.9%	(5/28)
描く　(egaku)	12	271	14	19.4	SVM	76.7%	(23/30)	100.0%	(1/1)	75.9%	(22/29)
買う　(kau)	31	1,117	19	58.8	SVM	86.7%	(26/30)	100.0%	(3/3)	85.2%	(23/27)
書く　(kaku_v)	15	795	4	198.8	SVM	76.7%	(23/30)	80.0%	(4/5)	76.0%	(19/25)
聞く　(kiku)	25	536	14	38.3	SVM	66.7%	(20/30)	100.0%	(3/3)	63.0%	(17/27)
越える(koeru)	14	109	10	10.9	SVM	63.3%	(19/30)	-		63.3%	(19/30)
使う　(tsukau)	19	1,139	14	81.4	SVM	56.7%	(17/30)	100.0%	(1/1)	55.2%	(16/29)
作る　(tsukuru)	19	834	17	49.1	SB	10.0%	(3/30)	100.0%	(2/2)	3.6%	(1/28)
伝える(tsutaeru)	19	155	15	10.3	DL	80.0%	(24/30)	100.0%	(3/3)	77.8%	(21/27)
出る　(deru)	30	4,705	26	181.0	SB	3.3%	(1/30)	100.0%	(1/1)	0.0%	(0/29)
乗る　(noru)	23	712	17	41.9	DL	53.3%	(16/30)	100.0%	(8/8)	36.4%	(8/22)
図る　(hakaru)	17	3,14	17	187.3	SB	2.7%	(8/30)	100.0%	(8/8)	0.0%	(0/22)
待つ　(matsu)	23	618	15	41.2	SVM	93.3%	(28/30)	100.0%	(1/1)	93.1%	(27/29)
守る　(mamoru)	16	522	19	27.5	SVM	46.7%	(14/30)	100.0%	(3/3)	40.7%	(11/27)
見せる(miseru)	20	285	12	23.8	SVM	90.0%	(27/30)	100.0%	(1/1)	89.7%	(26/29)
認める(mitomeru)	10	929	13	71.5	DL	66.7%	(20/30)	100.0%	(1/1)	65.5%	(19/29)
持つ　(motsu)	59	1,835	23	79.8	SB	46.7%	(14/30)	100.0%	(3/3)	40.7%	(11/17)
求める(motomeru)	10	481	22	21.9	SVM	43.3%	(13/30)	100.0%	(1/1)	41.4%	(12/29)
全動詞	452	26,426	351	75.3		57.5%	(345/600)	94.2%	(49/52)	54.0%	(296/548)
合計	911	39.87	655	60.9		63.5%	(761/1,200)	91.0%	(91/100)	60.9%	(670/1100)

フロントページの続き

(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内

(72)発明者　井佐原　均
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内

Ｆターム(参考)　5B091 AA05 CA02 CA22 CC01 EA02