(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7194759号
(P7194759)

(45)発行日　令和4年12月22日(2022.12.22)

(24)登録日　令和4年12月14日(2022.12.14)

(51)国際特許分類

Ｇ０６Ｆ　40/44　(2020.01)		Ｇ０６Ｆ	40/44	
Ｇ０６Ｎ　20/00　(2019.01)		Ｇ０６Ｎ	20/00	１３０

ＦＩ

請求項の数　6　（全17頁）

(21)出願番号　特願2020-572078(P2020-572078)
(86)(22)出願日　令和1年10月4日(2019.10.4)
(86)国際出願番号　PCT/JP2019/039337
(87)国際公開番号　WO2020/166125
(87)国際公開日　令和2年8月20日(2020.8.20)
審査請求日　令和3年6月8日(2021.6.8)
(31)優先権主張番号　特願2019-22411(P2019-22411)
(32)優先日　平成31年2月12日(2019.2.12)
(33)優先権主張国・地域又は機関
日本国(JP)

(73)特許権者　392026693
株式会社ＮＴＴドコモ
東京都千代田区永田町二丁目１１番１号
(74)代理人　100088155
弁理士　長谷川 芳樹
(74)代理人　100113435
弁理士　黒木 義樹
(74)代理人　100121980
弁理士　沖山 隆
(74)代理人　100128107
弁理士　深石 賢治
(74)代理人　100183438
弁理士　内藤 泰史
(72)発明者　村上 聡一朗
東京都千代田区永田町二丁目１１番１号
最終頁に続く

(54)【発明の名称】　翻訳用データ生成システム

(57)【特許請求の範囲】

【請求項１】

原言語テキストにノイズを付与してノイズ付与原言語テキストを得るノイズ付与部と、

前記ノイズ付与原言語テキストと、該ノイズ付与原言語テキストのノイズ付与前の原言語テキストに対応する目的言語テキストとを対応付けた疑似対訳コーパスを構築するコーパス構築部と、

ノイズを含んだ原言語テキスト群である訓練データを用いて、原言語テキストの各単語の次にノイズが入る場合に各単語に対してノイズのタイプを示すノイズラベルを予測するように学習されたノイズモデルを学習するノイズモデル学習部と、を備え、

前記ノイズ付与部は、

前記ノイズモデルを用いて、原言語テキストの各単語の特徴に応じて、ノイズのタイプを示すノイズラベルを付与し、該ノイズラベルを該ノイズラベルに対応する単語へ置き換えることにより、原言語テキストにノイズを付与し、

前記ノイズラベルの付与について、原言語テキストの各単語の特徴を入力として前記ノイズモデルから出力される各ノイズラベルのスコアに基づく各ノイズラベルの確率分布に従ってノイズラベルをサンプリングし、原言語テキストに付与するノイズラベルを決定する、翻訳用データ生成システム。

【請求項２】

前記疑似対訳コーパスを用いて翻訳モデルを学習する翻訳モデル学習部を更に備える、請求項１記載の翻訳用データ生成システム。

【請求項３】
前記ノイズ付与部は、１つの前記ノイズラベルに対して置き換える単語を複数パターン導出し、１つの原言語テキストから複数パターンの前記ノイズ付与原言語テキストを得る、請求項１又は２記載の翻訳用データ生成システム。
【請求項４】
前記ノイズ付与部は、前記ノイズモデルを用いて、各単語に対応する前記ノイズラベルを複数パターン導出し、１つの原言語テキストから複数パターンの前記ノイズ付与原言語テキストを得る、請求項１～３のいずれか一項記載の翻訳用データ生成システム。
【請求項５】
前記ノイズ付与部は、前記ノイズモデルを用いて、原言語テキストの各単語の特徴である、形態素、品詞、及び単語の読みの少なくとも一つに応じて、前記ノイズラベルを付与する、請求項４記載の翻訳用データ生成システム。
【請求項６】
前記ノイズモデルは、条件付き確率場又はニューラルネットワークを用いた手法により構築されている、請求項１～５のいずれか一項記載の翻訳用データ生成システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明の一態様は、翻訳用データ生成システムに関する。
【背景技術】
【０００２】
機械翻訳システムにおいて、利用者の自然発話入力に、言い淀み、言い直し、又はフィラー等（以下、これらを総称して「ノイズ」と記載する場合がある）が含まれることによって、翻訳精度が低下する場合がある。
【０００３】
このような課題に対して、例えば特許文献１及び特許文献２等に示されるように、発話における言い直し箇所等を特定し利用者の発話内容を修正する技術が知られている。
【先行技術文献】
【特許文献】
【０００４】
【文献】特開２０１０－０７９６４７号公報
特開２００７－０５７８４４号公報
【発明の概要】
【発明が解決しようとする課題】
【０００５】
しかしながら、ノイズ箇所を特定して修正することは容易ではなく、上述した技術によっても翻訳精度を十分に担保することは困難である。
【０００６】
本発明の一態様は上記実情に鑑みてなされたものであり、ノイズが含まれる自然発話に対しても高精度に翻訳を行うことを目的とする。
【課題を解決するための手段】
【０００７】
本発明の一態様に係る翻訳用データ生成システムは、原言語テキストにノイズを付与してノイズ付与原言語テキストを得るノイズ付与部と、ノイズ付与原言語テキストと、該ノイズ付与原言語テキストのノイズ付与前の原言語テキストに対応する目的言語テキストとを対応付けた疑似対訳コーパスを構築するコーパス構築部と、を備える。
【０００８】
本発明の一態様に係る翻訳用データ生成システムでは、原言語テキストにノイズが付与され、ノイズ付与原言語テキストとノイズ付与前の原言語テキストに対応する目的言語テキストとを対応付けた疑似対訳コーパスが構築される。このように、ノイズ付与原言語テ

キストがノイズ付与前の原言語テキストに対応する目的言語テキストに対応付けられた対訳コーパスが構築されることにより、このような対訳コーパスを利用して、例えば自然発話入力にフィラー等のノイズが含まれている場合においても、ノイズ付与前の原言語テキストに対応する目的言語テキストを適切に導出することが可能となる。すなわち、本発明の一態様に係る翻訳用データ生成システムによれば、ノイズが含まれる自然発話に対して頑健なコーパス（疑似対訳コーパス）を構築することができ、ノイズが含まれる自然発話に対しても高精度に翻訳を行うことができる。

【０００９】

上記翻訳用データ生成システムは、疑似対訳コーパスを用いて翻訳モデルを学習する翻訳モデル学習部を更に備えていてもよい。構築したコーパスに基づいて翻訳モデルが学習されることにより、ノイズが含まれる自然発話に対してより高精度に翻訳を行うことができる。

【００１０】

上記翻訳用データ生成システムは、ノイズを含んだ原言語テキスト群である訓練データを用いて、原言語テキストに対するノイズの付与に係るノイズモデルを学習するノイズモデル学習部を更に備え、ノイズ付与部は、ノイズモデルを用いて、原言語テキストにノイズを付与してもよい。予めノイズが含まれている原言語テキスト群に基づきノイズモデルが学習され、該ノイズモデルに基づいてノイズの付与が行われることによって、実際に含まれる可能性が高いノイズが付与され易くなり、翻訳精度をより向上させることができる。

【００１１】

上記翻訳用データ生成システムにおいて、ノイズ付与部は、原言語テキストの各単語に、ノイズのタイプを示すノイズラベルを付与し、該ノイズラベルを該ノイズラベルに対応する単語へ置き換えることにより、原言語テキストにノイズを付与してもよい。原言語テキストの各単語に応じたノイズラベルが付与された後に該ノイズラベルに応じた単語（ノイズ）が導出されることにより、ノイズ付与の容易性及び妥当性を担保することができる。

【００１２】

上記翻訳用データ生成システムにおいて、ノイズ付与部は、１つのノイズラベルに対して置き換える単語を複数パターン導出し、１つの原言語テキストから複数パターンのノイズ付与原言語テキストを得てもよい。これにより、１つの原言語テキストから効率的に疑似対訳コーパスを充実させ、翻訳精度をより向上させることができる。

【００１３】

上記翻訳用データ生成システムにおいて、ノイズ付与部は、各単語に対応するノイズラベルを複数パターン導出し、１つの原言語テキストから複数パターンのノイズ付与原言語テキストを得てもよい。これにより、１つの原言語テキストから効率的に疑似対訳コーパスを充実させ、翻訳精度をより向上させることができる。

【００１４】

上記翻訳用データ生成システムにおいて、ノイズ付与部は、原言語テキストの各単語の特徴に応じて、ノイズラベルを付与してもよい。これにより、各単語に関連して含まれやすいノイズに係るノイズラベルを、各単語に適切に付与することができる。

【００１５】

上記翻訳用データ生成システムにおいて、ノイズ付与部は、原言語テキストの各単語の特徴である、形態素、品詞、及び単語の読みの少なくとも一つに応じて、ノイズラベルを付与してもよい。これにより、各単語に関連して含まれやすいノイズに係るノイズラベルを、各単語に適切に付与することができる。

【００１６】

上記翻訳用データ生成システムにおいて、ノイズ付与部は、原言語テキストの各単語の特徴を入力としてノイズモデルから出力される各ノイズラベルのスコアに基づく各ノイズレベルの確率分布に従ってノイズラベルをサンプリングし、原言語テキストに付与するノイズラベルを決定してもよい。これにより、例えばノイズモデルから出力されたスコアが高いノイズラベルを付与することが可能となり、各単語に関連して含まれやすいノイズに

係るノイズラベルを、各単語に適切に付与することができる。
【００１７】
上記翻訳用データ生成システムにおいて、ノイズモデルは、条件付き確率場又はニューラルネットワークを用いた手法により構築されていてもよい。これにより、機械学習によってノイズモデルを適切に構成することができる。
【発明の効果】
【００１８】
本発明の一態様によれば、ノイズが含まれる自然発話に対しても高精度に翻訳を行うことができる。
【図面の簡単な説明】
【００１９】
【図１】本実施形態に係る翻訳用データ生成システムの処理イメージを模式的に示す図である。
【図２】本実施形態に係る翻訳用データ生成システムの機能構成を示す図である。
【図３】ノイズモデルの概要を説明する図である。
【図４】ノイズラベルを説明する図である。
【図５】疑似対訳コーパスの構築イメージを示す図である。
【図６】翻訳用データ生成システムが実行する処理を示すフローチャートである。
【図７】本実施形態及び比較例の翻訳例を示す表である。
【図８】翻訳用データ生成装置のハードウェア構成を示す図である。
【発明を実施するための形態】
【００２０】
以下、添付図面を参照しながら本発明の実施形態を詳細に説明する。図面の説明において、同一又は同等の要素には同一符号を用い、重複する説明を省略する。
【００２１】
最初に、図１を参照して、本実施形態に係る翻訳用データ生成システム１の処理イメージを説明する。図１は、本実施形態に係る翻訳用データ生成システム１の処理イメージを模式的に示す図である。翻訳用データ生成システム１は、既存の対訳コーパス（一般的に用いられる対訳コーパス）の原言語側のテキスト（原言語テキスト）に対してノイズを付与すると共に、ノイズが付与された原言語テキスト（ノイズ付与原言語テキスト）と、ノイズ付与前の原言語テキストに対応する目的言語側のテキスト（目的言語テキスト）とを対応付けた疑似対訳コーパスを構築し、該疑似対訳コーパスを用いて機械翻訳モデル（例えばＮＭＴ（Neural Machine Translation）モデル）を学習（構築）するシステムである。ここでのノイズとは、利用者の自然発話入力に含まれ得る言い淀み、言い直し、又はフィラー等である。
【００２２】
図１に示される例では、既存の対訳コーパスにおける「主要な高速道路よりも観光ルートの方を走りたいです」との原言語テキストに対して、所定のルールに従って複数パターンのノイズを付与し（詳細は後述）、「えー主要な高速道路よりもまー観光ルートの方を走りたいです」「あ主要な高速道路よりもえー観光ルートの方を走りたいです」「えーっと主要な高速道路よりもまー観光ルートの方を走りたいです」という３パターンのノイズ付与原言語テキスト得ている。そして、ノイズ付与前の原言語テキストに対応する目的言語テキストである「I would rather take a scenic route than a main highway.」と、上述した３パターンのノイズ付与原言語テキストとを対応付けた疑似対訳コーパスが構築されて、該疑似対訳コーパスを用いて機械翻訳モデルが学習（構築）されている。このように、ノイズ付与原言語テキストがノイズ付与前の原言語テキストに対応する目的言語テキストに対応付けられた疑似対訳コーパスが構築されることにより、このような疑似対訳コーパスを利用して、例えば自然発話入力にフィラー等のノイズが含まれている場合においても、ノイズ付与前の原言語テキストに対応する目的言語テキストを適切に導出することが可能となる。以下、翻訳用データ生成システム１の機能の詳細について説明する。

【００２３】
図２は、本実施形態に係る翻訳用データ生成システム１の機能構成を示す図である。図２に示されるように、翻訳用データ生成システム１は、翻訳用データ生成装置１０と、対訳コーパスＤＢ２０と、訓練情報ＤＢ３０と、ノイズモデル学習装置４０（ノイズモデル学習部）と、翻訳モデル学習装置５０（翻訳モデル学習部）と、を備えている。なお、翻訳用データ生成システム１は、必ずしも上記の各構成を備えるものでなくてもよく、例えば翻訳用データ生成装置１０のみで構成されていてもよいし、翻訳用データ生成装置１０及びノイズモデル学習装置４０のみで構成されていてもよいし、翻訳用データ生成装置１０、ノイズモデル学習装置４０、及び翻訳モデル学習装置５０のみで構成されていてもよいし。

【００２４】
対訳コーパスＤＢ２０は、対訳コーパスを記憶しているデータベースである。対訳コーパスとは、原言語テキスト及び目的言語テキストの組み合わせを構造化したものである。対訳コーパスＤＢ２０が記憶する対訳コーパスは、通常利用されるものでよく、例えばＫＦＴＴ（Kyoto Free Translation Task）又はＢＴＥＣ等の日本語・英語の対訳コーパスである。本実施形態では、翻訳用データ生成装置１０によって、対訳コーパスＤＢ２０が記憶する対訳コーパスの原言語テキストにノイズが付与され、疑似対訳コーパスが生成される（詳細は後述）。

【００２５】
訓練情報ＤＢ３０は、ノイズモデル（詳細は後述）を学習するための訓練情報（訓練データ）を記憶しているデータベースである。訓練情報とは、予めノイズがアノテーションされた原言語テキスト群（自然発話の書き起こしコーパス。学習用発話データ）である。このような訓練情報は、例えば通常のコーパスに含まれる原言語テキストにノイズがアノテーションされることによって構築されている。

【００２６】
ノイズモデル学習装置４０は、訓練情報ＤＢ３０に記憶されている訓練情報（ノイズを含んだ原言語テキスト群である訓練データ）を用いて、原言語テキストに対するノイズの付与に係るノイズモデルを学習する。ノイズモデルの学習データ（訓練データ）としては、例えば、日本語はなし言葉コーパス（ＣＳＪ）又はＳｗｉｔｃｈ Ｂｏａｒｄ Ｃｏｒｐｕｓ等の自然発話コーパスの書き起こしコーパスが用いられてもよい。ノイズモデルは、原言語テキストが入力された場合に、該原言語テキストに係るノイズラベルの情報を出力するものである。ノイズラベルとは、ノイズのタイプ（種別）を示す情報である。図４は、ノイズラベルを説明する図である。図４に示されるように、本実施形態では、ノイズラベルとして、＜Ｆ＞、＜Ｄ＞、０の３種類がある。＜Ｆ＞は、フィラーを示すノイズラベルである。＜Ｄ＞は言い淀み又は言い直しを示すノイズラベルである。０はノイズ無しを示すノイズラベルである。ノイズラベルの情報とは、ノイズラベルの種類（上述した＜Ｆ＞、＜Ｄ＞、０）と各ノイズラベルが対応付けられる単語（詳細には形態素）とが紐づいた情報であり、例えば後述するノイズラベル系列である。

【００２７】
図３は、ノイズモデルの概要を説明する図である。図３に示されるように、ノイズモデルは、例えば、品詞タグ付けや固有表現抽出タスク等で広く用いられている双方向再帰的ニューラルネットワーク（ＢｉＲＮＮ：Bi-directional Recurrent Neural Networks）を用いて構築されている。なお、ノイズモデルは、ＲＮＮ等のその他のニューラルネットワークを用いた手法や、ＣＲＦ（Conditional random field）等の条件付き確率場を用いた手法により構築されていてもよい。ノイズモデルは、入力された原言語テキストの各入力要素（単語詳細には形態素）の次にノイズが入る場合、その要素に対して適当なノイズラベルを予測するように学習されている。ノイズモデルを用いたノイズ付与においては、原言語テキストの形態素系列$w=(w_0, w_1, \cdots, w_n)$からノイズラベル系列$l=(l_0, l_1, \cdots, l_n)$を予測する系列ラベリング問題として考える。

【００２８】

いま、「＜Ｆえー＞それでは会議を＜Ｄを＞始め＜Ｆあー＞ます」という学習用発話データを例にノイズモデルを学習する方法を説明する。ここで、学習用発話データに含まれる「＜Ｆえー＞」は「えー」がフィラー＜Ｆ＞に相当することを表している。この場合、まず、学習用発話データから形態素系列ｗ＝（＜ＢＯＳ＞，それでは，会議，を，始め，ます，＜ＥＯＳ＞）が抽出される。図３に示されるように形態素系列は、ｔ＝０～６までのタイムステップに各形態素（＜ＢＯＳ＞及び＜ＥＯＳ＞を含む）が対応付いている。次に、同じ学習用発話データであってノイズがアノテーションされた情報に基づいて、ノイズラベル系列ｌ＝（＜Ｆ＞，０，０，＜Ｄ＞，＜Ｆ＞，０，０）が生成される。最後に、形態素系列ｗからノイズラベル系列ｌを予測する系列ラベリング問題としてＢｉＲＮＮを学習する。ＢｉＲＮＮでは、入力系列に対する出力系列の予測誤差が用いられ、パラメータ学習が行われる。

【００２９】

図２に戻り、翻訳モデル学習装置５０は、翻訳用データ生成装置１０において構築された疑似対訳コーパスを用いて翻訳モデルを学習する。翻訳モデルとしては、Ｔｒａｎｓｆｏｒｍｅｒ又はＲＮＮ－ｂａｓｅｄ　Ｓｅｑｕｅｎｃｅ－ｔｏ－Ｓｅｑｕｅｎｃｅモデル等を用いてもよい。

【００３０】

翻訳用データ生成装置１０は、その機能として、解析部１１と、ノイズ付与部１２と、コーパス構築部１３と、記憶部１４とを備えている。

【００３１】

解析部１１は、対訳コーパスＤＢ２０から原言語テキストを取得し、取得した原言語テキストに対して形態素解析を行う。すなわち、例えば、解析部１１は、「主要な高速道路よりも観光ルートの方を走りたいです。」という原言語テキストを取得すると、該原言語テキストについて形態素系列ｗ＝（主要，な，高速，道路，より，も，観光，ルート，の，方，を，走り，たい，です）を抽出する。

【００３２】

ノイズ付与部１２は、原言語テキスト（詳細には解析部１１が抽出した形態素系列）にノイズを付与してノイズ付与原言語テキストを得る。ノイズ付与部１２は、ノイズモデル学習装置４０によって学習されたノイズモデルを用いて、原言語テキストにノイズを付与する。ノイズ付与部１２は、原言語テキストの各単語の特徴（具体的には形態素）に応じて、各形態素にノイズラベルを付与し、該ノイズラベルを該ノイズラベルに対応する単語（ノイズとしての単語）へ置き換えることにより、原言語テキストにノイズを付与する。ノイズ付与部１２は、ノイズモデルを用いることにより、入力された原言語テキストの形態素系列に対応するノイズラベル系列を予測し、対応する形態素系列の次にノイズラベルを挿入する。そして、ノイズ付与部１２は、挿入したノイズラベルを、ノイズを表す単語に置換し、最終的な出力であるノイズが付与された原言語テキストであるノイズ付き原言語テキストを得る。なお、ノイズ付与部１２は、原言語テキストの形態素に応じてノイズラベルを付与するとして説明したがこれに限定されず、原言語テキストの各単語の品詞や読み（発音）に応じてノイズラベルを付与してもよい。また、ノイズ付与部１２は、単語の形態素、品詞、及び読み等の２つ以上の情報に応じて、ノイズラベルを付与してもよい。

【００３３】

ノイズ付与部１２は、具体的には、まず、原言語テキストの形態素系列をノイズモデルに入力し、各タイムステップ（各形態素系列）におけるノイズモデルの出力ベクトルh_tを取得する。本実施形態では、各タイムステップにおけるノイズラベルについて、単純にノイズラベルの事後確率が最大となるものを推定結果とするのではなく、出力ベクトルh_tに指数をとった値$exp(h_t/\tau)$で定義される多項分布に基づくサンプリングにより決定する。すなわち、各タイムステップにおけるノイズラベルl_tは以下の（１）式に基づき推定される。

$$l_t \sim exp(h_t/\tau) \quad \cdots (1)$$

上記（１）式において、l_tはノイズラベルの推定結果、h_tはノイズモデルの出力ベク

トル、τは温度パラメータである。出力ベクトルh_tは、３種類のラベルタイプ（＜Ｆ＞，＜Ｄ＞，０）についての３次元ベクトルで示される。温度パラメータτは、ノイズラベルのバリエーションの強弱を操作するためのパラメータである。温度パラメータτの値を大きく（$\tau \to \infty$）するとノイズラベルの確率分布は一様分布に近づき、小さく（$\tau \to 0$）すると最も高い確率のノイズラベルが選択されるようになる。

【００３４】

例えば温度パラメータτが比較的小さい場合のノイズラベルの決定について説明する。いま、ノイズモデルの出力ベクトルh_t＝（－０．１（０の重みスコア），０．３（＜Ｆ＞の重みスコア），－０．３（＜Ｄ＞の重みスコア））であり、温度パラメータτ＝０．１５であるとする。この場合、h_t/τ＝（－０．６６６６…，２，－２）となる。各ノイズラベルの重みスコアを０以上とすべく指数をとると、$\exp(h_t/\tau)$＝（０．５１，７．３９，０．１３）となる。重みスコアを確率値として扱うべく値域が［０，１］且つ全ての値を足して１になるように正規化すると、確率分布は（０．０６（０がノイズラベルとして選ばれる確率），０．９２（＜Ｆ＞がノイズラベルとして選ばれる確率），０．０２（＜Ｄ＞がノイズラベルとして選ばれる確率））となる。このような確率分布（多項分布）に基づきノイズラベルを１回だけサンプリング（試行）することは、カテゴリカル分布からのサンプリングに相当する。この場合、ノイズラベル＜Ｆ＞の確立が９２％と極めて高く、サンプリング結果として選択される可能性が極めて高い。

【００３５】

例えば温度パラメータτが比較的大きい場合のノイズラベルの決定について説明する。いま、ノイズモデルの出力ベクトルh_t＝（－０．１（０の重みスコア），０．３（＜Ｆ＞の重みスコア），－０．３（＜Ｄ＞の重みスコア））であり、温度パラメータτ＝１．０であるとする。この場合、h_t/τ＝（－０．１，０．３，－０．３）となる。各ノイズラベルの重みスコアを０以上とすべく指数をとると、$\exp(h_t/\tau)$＝（０．９０，１．３５，０．７４）となる。重みスコアを確率値として扱うべく値域が［０，１］且つ全ての値を足して１になるように正規化すると、確率分布は（０．３０（０がノイズラベルとして選ばれる確率），０．４５（＜Ｆ＞がノイズラベルとして選ばれる確率），０．２５（＜Ｄ＞がノイズラベルとして選ばれる確率））となる。このように、温度パラメータτを大きくすると、上述した温度パラメータτ＝０．１５の場合と比較して、ノイズラベル０及びノイズラベル＜Ｄ＞が選択されやすくなっていることがわかる。温度パラメータτが∞に近づくほど、各ノイズラベルの確立は３３．３３３…％に近づき、確率分布が一様分布に近づく。

【００３６】

このように、ノイズ付与部１２は、原言語テキストの各単語の特徴（形態素系列）を入力としてノイズモデルから出力される各ノイズラベルのスコアに基づく確率分布に従ってノイズラベルをサンプリングし、原言語テキストに付与するノイズラベルを決定している。なお、上述した説明においては、ノイズモデルの出力値を基に定義される確率分布が多項分布を表すとして説明したが、これに限定されず、確率分布はポアソン分布又は正規分布等を表すものであってもよい。

【００３７】

ノイズ付与部１２は、つづいて、ノイズモデルを用いて予測したノイズラベル系列を、ノイズを表す単語に置き換える。ノイズ付与部１２は、例えば、各ノイズラベルに対応する語彙集合V_{type}からユニグラム確率に基づきサンプリングを行う。例えば、フィラーのノイズラベル＜Ｆ＞を、フィラーを表す単語へ置換する場合、以下の（２）式に基づきフィラーを表す単語が決定される。

$$w_t' \sim V_{<F>} \cdots (2)$$

上記（２）式において、$V_{<F>}$はノイズラベル＜Ｆ＞の語彙集合、w_t'はタイムステップｔに挿入されるフィラー（ノイズ）を表す単語である。以上によって、原言語テキストの形態素系列ｗ＝（w_0，w_1，…，w_n）からノイズを表す単語を含む系列ｗ´＝（w_0，w_1，w_1'，w_2，w_2'，…，w_n）を得る。

【００３８】
ノイズ付与部１２は、１つの原言語テキストから複数パターンのノイズ付与原言語テキストを得る。ノイズ付与部１２は、例えば、１つのノイズラベルに対して置き換える単語（ノイズを表す単語）を複数パターン導出し、１つの原言語テキストから複数パターンのノイズ付与原言語テキストを得てもよい。また、ノイズ付与部１２は、例えば各形態素に対応するノイズラベルを複数パターン導出し、１つの原言語テキストから複数パターンのノイズ付与原言語テキストを得てもよい。
【００３９】
コーパス構築部１３は、ノイズ付与原言語テキストと、該ノイズ付与原言語テキストのノイズ付与前の原言語テキストに対応する目的言語テキストとを対応付けた疑似対訳コーパスを構築する。図５は、疑似対訳コーパスの構築イメージを示す図である。図５に示される例では、ノイズ付与前の原言語テキストである「主要な高速道路よりも観光ルートの方を走りたいです」についてのノイズ付与原言語テキスト（「えー主要な高速道路よりも観光ルートの方を走りたいです」等の７つのノイズ付与原言語テキスト）と、ノイズ付与前の原言語テキストに対応する目的言語テキストである「I would rather take a scenic route than a main highway.」とが対応付けられた（対訳ペアとした）疑似対訳コーパスが構築されている。
【００４０】
記憶部１４は、コーパス構築部１３によって構築された疑似対訳コーパスを記憶するＤＢである。翻訳モデル学習装置５０は、記憶部１４に記憶されている疑似対訳コーパスを用いて翻訳モデルを学習する。
【００４１】
次に、図６を参照して、翻訳用データ生成システム１が実行する処理を説明する。図６は、翻訳用データ生成システム１が実行する処理を示すフローチャートである。なお、図６に示される処理が実行される前提として、ノイズモデル学習装置４０によってノイズモデルが構築（学習）されているものとする。
【００４２】
図６に示されるように、翻訳用データ生成システム１では、まず、翻訳用データ生成装置１０の解析部１１が対訳コーパスＤＢ２０から原言語テキストを取得する（ステップＳ１）。つづいて、解析部１１は、取得した原言語テキストに対して形態素解析を実行する（ステップＳ２）。
【００４３】
つづいて、翻訳用データ生成装置１０のノイズ付与部１２は、解析部１１が抽出した形態素系列に対してノイズを付与し、ノイズ付与原言語テキストを得る（ステップＳ３）。詳細には、ノイズ付与部１２は、ノイズモデルを用いることにより、入力された原言語テキストの形態素系列に対応するノイズラベル系列を予測し、対応する形態素系列の次にノイズラベルを挿入する。そして、ノイズ付与部１２は、挿入したノイズラベルを、ノイズを表す単語に置換し、最終的な出力であるノイズが付与された原言語テキストであるノイズ付き原言語テキストを得る。
【００４４】
つづいて、翻訳用データ生成装置１０のコーパス構築部１３は、ノイズ付与原言語テキストと、該ノイズ付与原言語テキストのノイズ付与前の原言語テキストに対応する目的言語テキストとを対応付けた疑似対訳コーパスを構築する（ステップＳ４）。
【００４５】
最後に、翻訳モデル学習装置５０は、コーパス構築部１３によって構築された疑似対訳コーパスを用いて翻訳モデルを学習する（ステップＳ５）。以上が、翻訳用データ生成システム１が実行する処理の一例である。
【００４６】
次に、本実施形態の作用効果について説明する。
【００４７】

本実施形態に係る翻訳用データ生成システム１は、原言語テキストにノイズを付与してノイズ付与原言語テキストを得るノイズ付与部１２と、ノイズ付与原言語テキストと、該ノイズ付与原言語テキストのノイズ付与前の原言語テキストに対応する目的言語テキストとを対応付けた疑似対訳コーパスを構築するコーパス構築部１３と、を備える。

【００４８】

本実施形態に係る翻訳用データ生成システム１では、原言語テキストにノイズが付与され、ノイズ付与原言語テキストとノイズ付与前の原言語テキストに対応する目的言語テキストとを対応付けた疑似対訳コーパスが構築される。このように、ノイズ付与原言語テキストがノイズ付与前の原言語テキストに対応する目的言語テキストに対応付けられた対訳コーパスが構築されることにより、このような対訳コーパスを利用して、例えば自然発話入力にフィラー等のノイズが含まれている場合においても、ノイズ付与前の原言語テキストに対応する目的言語テキストを適切に導出することが可能となる。すなわち、本実施形態に係る翻訳用データ生成システム１によれば、ノイズが含まれる自然発話に対して頑健なコーパス（疑似対訳コーパス）を構築することができ、ノイズが含まれる非流暢な自然発話に対しても高精度に翻訳を行うことができる。なお、このような翻訳用データ生成システム１により生成された情報が翻訳に用いられる場合には、利用者の発話内容を修正して翻訳モデルに入力する必要がなく、利用者の発話内容をそのまま翻訳モデルに入力することができる。また、例えば、特開２０１０－０７９６４７号公報及び特開２００７－０５７８４４号公報に記載されたシステムでは、音声認識装置を用いて、逐次利用者の発話を受け取り言い直し判定を行っているが、本実施形態に係る翻訳用データ生成システム１では音声認識装置が不要であり、認識結果のテキスト情報のみが利用できればよい。このように、本実施形態に係る翻訳用データ生成システム１では、発話内容の修正処理や言い直し判定処理が実施されることを抑制できるため、ＣＰＵ等の処理部における処理負荷を軽減するという技術的効果も併せて奏する。

【００４９】

図７は、本実施形態及び比較例の翻訳例を示す表である。図７の上段に示されるように、ノイズが含まれる自然発話入力に対して、比較例では訳抜けが生じている。また、図７の下段に示されるように、ノイズが含まれる自然発話入力に対して、比較例ではノイズを含めた状態で翻訳しており、所望の翻訳を行うことができていない。比較例に示されるように、従来、ノイズが含まれる自然発話に対して高精度に翻訳を行うことは困難であった。この点、図７の上段及び下段に示されるように、本実施形態の翻訳用データ生成システム１によって構築された疑似対訳コーパスが考慮されて翻訳が行われた場合には、ノイズが含まれる自然発話に対しても翻訳誤りが起きにくく、高精度に翻訳を行うことができる。

【００５０】

翻訳用データ生成システム１は、疑似対訳コーパスを用いて翻訳モデルを学習する翻訳モデル学習装置５０を備えている。構築したコーパスに基づいて翻訳モデルが学習されることにより、ノイズが含まれる自然発話に対してより高精度に翻訳を行うことができる。

【００５１】

翻訳用データ生成システム１は、ノイズを含んだ原言語テキスト群である訓練データを用いて、原言語テキストに対するノイズの付与に係るノイズモデルを学習するノイズモデル学習装置４０を備え、ノイズ付与部１２は、ノイズモデルを用いて、原言語テキストにノイズを付与する。予めノイズが含まれている原言語テキスト群に基づきノイズモデルが学習され、該ノイズモデルに基づいてノイズの付与が行われることによって、実際に含まれる可能性が高いノイズが付与され易くなり、翻訳精度をより向上させることができる。

【００５２】

翻訳用データ生成システム１において、ノイズ付与部１２は、原言語テキストの各単語に、ノイズのタイプを示すノイズラベルを付与し、該ノイズラベルを該ノイズラベルに対応する単語へ置き換えることにより、原言語テキストにノイズを付与する。原言語テキストの各単語に応じたノイズラベルが付与された後に該ノイズラベルに応じた単語（ノイズ）が導出されることにより、ノイズ付与の容易性及び妥当性を担保することができる。

【００５３】
翻訳用データ生成システム１において、ノイズ付与部１２は、１つのノイズラベルに対して置き換える単語を複数パターン導出し、１つの原言語テキストから複数パターンのノイズ付与原言語テキストを得る。これにより、１つの原言語テキストから効率的に疑似対訳コーパスを充実させ、翻訳精度をより向上させることができる。
【００５４】
翻訳用データ生成システム１において、ノイズ付与部１２は、各単語に対応するノイズラベルを複数パターン導出し、１つの原言語テキストから複数パターンのノイズ付与原言語テキストを得る。これにより、１つの原言語テキストから効率的に疑似対訳コーパスを充実させ、翻訳精度をより向上させることができる。
【００５５】
翻訳用データ生成システム１において、ノイズ付与部１２は、原言語テキストの各単語の特徴に応じて、ノイズラベルを付与する。これにより、各単語に関連して含まれやすいノイズに係るノイズラベルを、各単語に適切に付与することができる。
【００５６】
翻訳用データ生成システム１において、ノイズ付与部１２は、原言語テキストの各単語の特徴である、形態素、品詞、及び単語の読みの少なくとも一つに応じて、ノイズラベルを付与する。これにより、各単語に関連して含まれやすいノイズに係るノイズラベルを、各単語に適切に付与することができる。
【００５７】
翻訳用データ生成システム１において、ノイズ付与部１２は、原言語テキストの各単語の特徴を入力としてノイズモデルから出力される各ノイズラベルのスコアに基づく各ノイズレベルの確率分布に従ってノイズラベルをサンプリングし、原言語テキストに付与するノイズラベルを決定する。これにより、例えばノイズモデルから出力されたスコアが高いノイズラベルを付与することが可能となり、各単語に関連して含まれやすいノイズに係るノイズラベルを、各単語に適切に付与することができる。
【００５８】
最後に、翻訳用データ生成装置１０のハードウェア構成について、図８を参照して説明する。上述の翻訳用データ生成装置１０は、物理的には、プロセッサ１００１、メモリ１００２、ストレージ１００３、通信装置１００４、入力装置１００５、出力装置１００６、バス１００７などを含むコンピュータ装置として構成されてもよい。
【００５９】
なお、以下の説明では、「装置」という文言は、回路、デバイス、ユニットなどに読み替えることができる。翻訳用データ生成装置１０のハードウェア構成は、図に示した各装置を１つ又は複数含むように構成されてもよいし、一部の装置を含まずに構成されてもよい。
【００６０】
翻訳用データ生成装置１０における各機能は、プロセッサ１００１、メモリ１００２などのハードウェア上に所定のソフトウェア（プログラム）を読み込ませることで、プロセッサ１００１が演算を行い、通信装置１００４による通信や、メモリ１００２及びストレージ１００３におけるデータの読み出し及び／又は書き込みを制御することで実現される。
【００６１】
プロセッサ１００１は、例えば、オペレーティングシステムを動作させてコンピュータ全体を制御する。プロセッサ１００１は、周辺装置とのインターフェース、制御装置、演算装置、レジスタなどを含む中央処理装置（ＣＰＵ：Central Processing Unit）で構成されてもよい。例えば、翻訳用データ生成装置１０のノイズ付与部１２等の制御機能はプロセッサ１００１で実現されてもよい。
【００６２】
また、プロセッサ１００１は、プログラム（プログラムコード）、ソフトウェアモジュールやデータを、ストレージ１００３及び／又は通信装置１００４からメモリ１００２に

読み出し、これらに従って各種の処理を実行する。プログラムとしては、上述の実施の形態で説明した動作の少なくとも一部をコンピュータに実行させるプログラムが用いられる。例えば、翻訳用データ生成装置１０のノイズ付与部１２等の制御機能は、メモリ１００２に格納され、プロセッサ１００１で動作する制御プログラムによって実現されてもよく、他の機能ブロックについても同様に実現されてもよい。上述の各種処理は、１つのプロセッサ１００１で実行される旨を説明してきたが、２以上のプロセッサ１００１により同時又は逐次に実行されてもよい。プロセッサ１００１は、１以上のチップで実装されてもよい。なお、プログラムは、電気通信回線を介してネットワークから送信されても良い。

【００６３】

メモリ１００２は、コンピュータ読み取り可能な記録媒体であり、例えば、ＲＯＭ（Read Only Memory）、ＥＰＲＯＭ（Erasable Programmable ＲＯＭ）、ＥＥＰＲＯＭ（Electrically Erasable Programmable ＲＯＭ）、ＲＡＭ（Random Access Memory）などの少なくとも１つで構成されてもよい。メモリ１００２は、レジスタ、キャッシュ、メインメモリ（主記憶装置）などと呼ばれてもよい。メモリ１００２は、本発明の一実施の形態に係る無線通信方法を実施するために実行可能なプログラム（プログラムコード）、ソフトウェアモジュールなどを保存することができる。

【００６４】

ストレージ１００３は、コンピュータ読み取り可能な記録媒体であり、例えば、ＣＤ－ＲＯＭ（Compact Disc ＲＯＭ）などの光ディスク、ハードディスクドライブ、フレキシブルディスク、光磁気ディスク(例えば、コンパクトディスク、デジタル多用途ディスク、Ｂｌｕ－ｒａｙ（登録商標）ディスク)、スマートカード、フラッシュメモリ(例えば、カード、スティック、キードライブ)、フロッピー（登録商標）ディスク、磁気ストリップなどの少なくとも１つで構成されてもよい。ストレージ１００３は、補助記憶装置と呼ばれてもよい。上述の記憶媒体は、例えば、メモリ１００２及び／又はストレージ１００３を含むデータベース、サーバその他の適切な媒体であってもよい。

【００６５】

通信装置１００４は、有線及び／又は無線ネットワークを介してコンピュータ間の通信を行うためのハードウェア（送受信デバイス）であり、例えばネットワークデバイス、ネットワークコントローラ、ネットワークカード、通信モジュールなどともいう。

【００６６】

入力装置１００５は、外部からの入力を受け付ける入力デバイス（例えば、キーボード、マウス、マイクロフォン、スイッチ、ボタン、センサなど）である。出力装置１００６は、外部への出力を実施する出力デバイス（例えば、ディスプレイ、スピーカー、LEDランプなど）である。なお、入力装置１００５及び出力装置１００６は、一体となった構成（例えば、タッチパネル）であってもよい。

【００６７】

また、プロセッサ１００１やメモリ１００２などの各装置は、情報を通信するためのバス１００７で接続される。バス１００７は、単一のバスで構成されてもよいし、装置間で異なるバスで構成されてもよい。

【００６８】

また、翻訳用データ生成装置１０は、マイクロプロセッサ、デジタル信号プロセッサ（ＤＳＰ：Digital Signal Processor）、ＡＳＩＣ（Application Specific Integrated Circuit）、ＰＬＤ（Programmable Logic Device）、ＦＰＧＡ（Field Programmable Gate Array）などのハードウェアを含んで構成されてもよく、当該ハードウェアにより、各機能ブロックの一部又は全てが実現されてもよい。例えば、プロセッサ１００１は、これらのハードウェアの少なくとも１つで実装されてもよい。

【００６９】

以上、本実施形態について詳細に説明したが、当業者にとっては、本実施形態が本明細書中に説明した実施形態に限定されるものではないということは明らかである。本実施形態は、特許請求の範囲の記載により定まる本発明の趣旨及び範囲を逸脱することなく修正

及び変更態様として実施することができる。したがって、本明細書の記載は、例示説明を目的とするものであり、本実施形態に対して何ら制限的な意味を有するものではない。例えば、本発明の一態様に係る翻訳用データ生成システムは、事前に定義したノイズ単語（フィラー、言い淀み、言い直し等）を原言語テキストのランダムな位置に付与するものであってもよい。ランダムな位置に付与する単語（ノイズ）は、例えばノイズ単語候補からランダムに選択されてもよい。このような構成においては、ノイズモデルの学習データ（ラベル付きデータ）がなくても、ノイズ単語が定義できさえすれば、ノイズをランダムに付与するノイズモデルを構築することができる。

【００７０】

本明細書で説明した各態様／実施形態は、ＬＴＥ（Long Term Evolution）、ＬＴＥ－Ａ（LTE-Advanced）、ＳＵＰＥＲ ３Ｇ、ＩＭＴ－Ａｄｖａｎｃｅｄ、４Ｇ、５Ｇ、ＦＲＡ（Future Radio Access）、Ｗ－ＣＤＭＡ（登録商標）、ＧＳＭ（登録商標）、ＣＤＭＡ２０００、ＵＭＢ（Ultra Mobile Broad-band）、ＩＥＥＥ ８０２．１１（Ｗｉ－Ｆｉ）、ＩＥＥＥ ８０２．１６（ＷｉＭＡＸ）、ＩＥＥＥ ８０２．２０、ＵＷＢ（Ultra-Wide Band）、Ｂｌｕｅｔｏｏｔｈ（登録商標）、その他の適切なシステムを利用するシステム及び／又はこれらに基づいて拡張された次世代システムに適用されてもよい。

【００７１】

本明細書で説明した各態様／実施形態の処理手順、フローチャートなどは、矛盾の無い限り、順序を入れ替えてもよい。例えば、本明細書で説明した方法については、例示的な順序で様々なステップの要素を提示しており、提示した特定の順序に限定されない。

【００７２】

入出力された情報等は特定の場所(例えば、メモリ)に保存されてもよいし、管理テーブルで管理してもよい。入出力される情報等は、上書き、更新、または追記され得る。出力された情報等は削除されてもよい。入力された情報等は他の装置へ送信されてもよい。

【００７３】

判定は、１ビットで表される値（０か１か）によって行われてもよいし、真偽値（Boolean：trueまたはfalse）によって行われてもよいし、数値の比較（例えば、所定の値との比較）によって行われてもよい。

【００７４】

本明細書で説明した各態様／実施形態は単独で用いてもよいし、組み合わせて用いてもよいし、実行に伴って切り替えて用いてもよい。また、所定の情報の通知（例えば、「Ｘであること」の通知）は、明示的に行うものに限られず、暗黙的（例えば、当該所定の情報の通知を行わない）ことによって行われてもよい。

【００７５】

ソフトウェアは、ソフトウェア、ファームウェア、ミドルウェア、マイクロコード、ハードウェア記述言語と呼ばれるか、他の名称で呼ばれるかを問わず、命令、命令セット、コード、コードセグメント、プログラムコード、プログラム、サブプログラム、ソフトウェアモジュール、アプリケーション、ソフトウェアアプリケーション、ソフトウェアパッケージ、ルーチン、サブルーチン、オブジェクト、実行可能ファイル、実行スレッド、手順、機能などを意味するよう広く解釈されるべきである。

【００７６】

また、ソフトウェア、命令などは、伝送媒体を介して送受信されてもよい。例えば、ソフトウェアが、同軸ケーブル、光ファイバケーブル、ツイストペア及びデジタル加入者回線（ＤＳＬ）などの有線技術及び／又は赤外線、無線及びマイクロ波などの無線技術を使用してウェブサイト、サーバ、又は他のリモートソースから送信される場合、これらの有線技術及び／又は無線技術は、伝送媒体の定義内に含まれる。

【００７７】

本明細書で説明した情報、信号などは、様々な異なる技術のいずれかを使用して表されてもよい。例えば、上記の説明全体に渡って言及され得るデータ、命令、コマンド、情報

、信号、ビット、シンボル、チップなどは、電圧、電流、電磁波、磁界若しくは磁性粒子、光場若しくは光子、又はこれらの任意の組み合わせによって表されてもよい。

【００７８】

なお、本明細書で説明した用語及び／又は本明細書の理解に必要な用語については、同一の又は類似する意味を有する用語と置き換えてもよい。

【００７９】

また、本明細書で説明した情報、パラメータなどは、絶対値で表されてもよいし、所定の値からの相対値で表されてもよいし、対応する別の情報で表されてもよい。

【００８０】

ユーザ端末は、当業者によって、移動通信端末、加入者局、モバイルユニット、加入者ユニット、ワイヤレスユニット、リモートユニット、モバイルデバイス、ワイヤレスデバイス、ワイヤレス通信デバイス、リモートデバイス、モバイル加入者局、アクセス端末、モバイル端末、ワイヤレス端末、リモート端末、ハンドセット、ユーザエージェント、モバイルクライアント、クライアント、またはいくつかの他の適切な用語で呼ばれる場合もある。

【００８１】

本明細書で使用する「判断(determining)」、「決定(determining)」という用語は、多種多様な動作を包含する場合がある。「判断」、「決定」は、例えば、計算(calculating)、算出(computing)、処理(processing)、導出(deriving)、調査(investigating)、探索(looking up)（例えば、テーブル、データベースまたは別のデータ構造での探索）、確認(ascertaining)した事を「判断」「決定」したとみなす事などを含み得る。また、「判断」、「決定」は、受信(receiving)（例えば、情報を受信すること）、送信(transmitting)(例えば、情報を送信すること)、入力(input)、出力(output)、アクセス(accessing)（例えば、メモリ中のデータにアクセスすること）した事を「判断」「決定」したとみなす事などを含み得る。また、「判断」、「決定」は、解決(resolving)、選択(selecting)、選定(choosing)、確立(establishing)、比較(comparing)などした事を「判断」「決定」したとみなす事を含み得る。つまり、「判断」「決定」は、何らかの動作を「判断」「決定」したとみなす事を含み得る。

【００８２】

本明細書で使用する「に基づいて」という記載は、別段に明記されていない限り、「のみに基づいて」を意味しない。言い換えれば、「に基づいて」という記載は、「のみに基づいて」と「に少なくとも基づいて」の両方を意味する。

【００８３】

本明細書で「第１の」、「第２の」などの呼称を使用した場合においては、その要素へのいかなる参照も、それらの要素の量または順序を全般的に限定するものではない。これらの呼称は、２つ以上の要素間を区別する便利な方法として本明細書で使用され得る。したがって、第１および第２の要素への参照は、２つの要素のみがそこで採用され得ること、または何らかの形で第１の要素が第２の要素に先行しなければならないことを意味しない。

【００８４】

「含む（include）」、「含んでいる（including）」、およびそれらの変形が、本明細書あるいは特許請求の範囲で使用されている限り、これら用語は、用語「備える(comprising)」と同様に、包括的であることが意図される。さらに、本明細書あるいは特許請求の範囲において使用されている用語「または（or）」は、排他的論理和ではないことが意図される。

【００８５】

本明細書において、文脈または技術的に明らかに1つのみしか存在しない装置である場合以外は、複数の装置をも含むものとする。

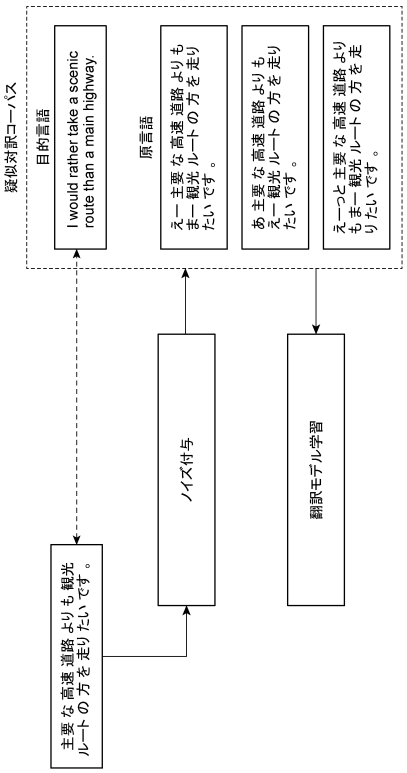
【００８６】

本開示の全体において、文脈から明らかに単数を示したものではなければ、複数のもの

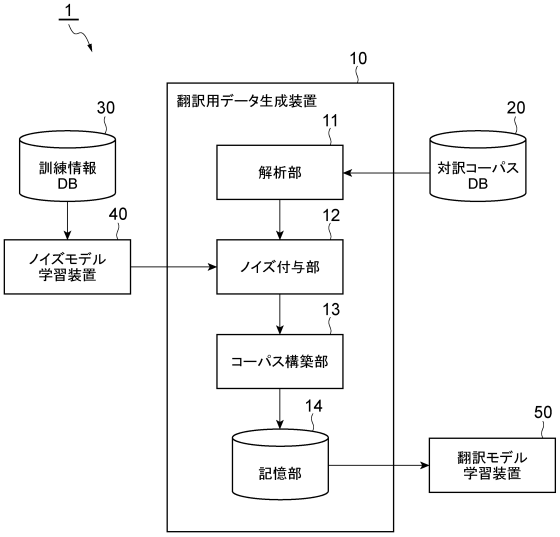
を含むものとする。
【符号の説明】
【００８７】
１…翻訳用データ生成システム、１２…ノイズ付与部、１３…コーパス構築部、４０…ノイズモデル学習装置（ノイズモデル学習部）、５０…翻訳モデル学習装置（翻訳モデル学習部）。

【図面】

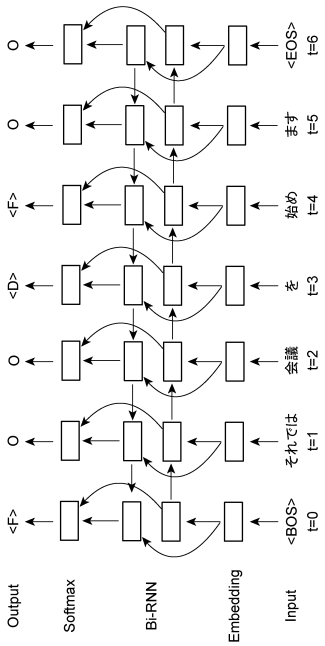
【図１】



【図２】



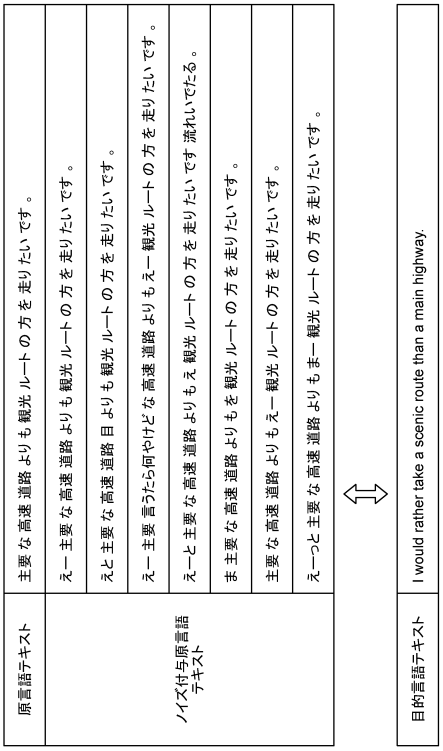
【図３】



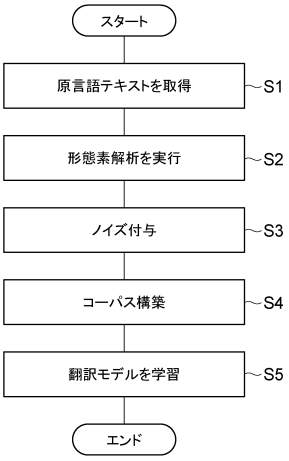
【図４】

ノイズラベル	対象
<F>	フィラー
<D>	言い淀み、言い直し
O	ノイズ無し

【図５】



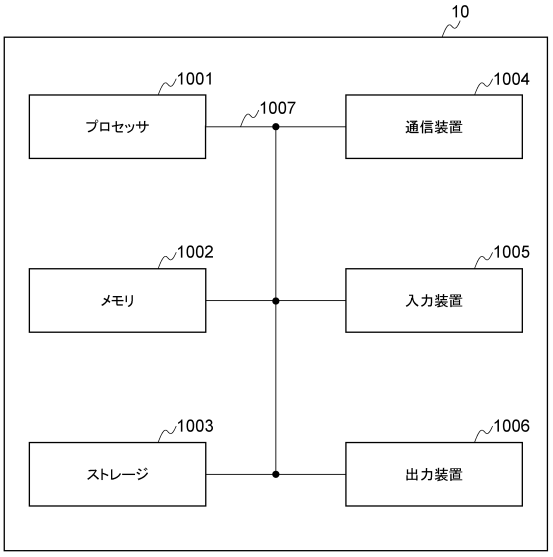
【図６】



【図７】

自然発話(入力)	比較例	本実施形態
えーっとですねーえーちょっとえーキャンセルということでキャンセル料がえー発生する形になるんですがえー本日えー出発の三日前にはいあのー予約のご確認のことはこちらのほうからさせて頂いてませんでしたっでしょうか	Let me see . Well , it seems like cancellation will cause a cancellation fee .	Well , a cancellation will cause a cancellation fee , but you made a reservation three days before your departure , did you confirm the cancellation ?
えっとー大阪にいや東京に行きたいんですがどのように行けばえーいいでしょうか	Well , I do not want to go to Osaka or Tokyo but how would you go if I go out ?	I would like to go to Tokyo , how can I go there ?

【図８】



フロントページの続き

株式会社ＮＴＴドコモ内

審査官　長 由紀子

(56)参考文献　特開２０１８－０５５６７１（ＪＰ，Ａ）
今出 昌宏 外２名，ユーラルネット機械翻訳における自動コーパス生成適用，一般社団法人人工知能学会 第３１回全国大会論文集ＤＶＤ ［ＤＶＤ－ＲＯＭ］，一般社団法人人工知能学会，2017年05月26日，pp.1-4
太田 健吾 外２名，フィラーの書き起こしのないコーパスからのフィラー付き言語モデルの構築，情報処理学会研究報告，日本，社団法人情報処理学会，2007年07月20日，第2007巻 第75号，pp.1-6
増村 亮 外２名，Ｗｅｂ上の言語資源を利用した大規模話し言葉データからの言語モデル作成，日本音響学会 ２０１１年 春季研究発表会講演論文集ＣＤ－ＲＯＭ ［ＣＤ－ＲＯＭ］，社団法人日本音響学会，2011年03月02日，pp.75-78
玉井 孝幸 外２名，音声対話システムにおける発話予測を利用した音声認識，情報処理学会研究報告，日本，社団法人情報処理学会，2002年10月25日，第2002巻 第98号，pp.1-6

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　４０／００－５８
Ｇ０６Ｎ　２０／００