

(19) 日本国特許庁(JP)

(12) 特許公報(B2)

(11) 特許番号

特許第6090785号
(P6090785)

(45) 発行日 平成29年3月8日(2017.3.8)

(24) 登録日 平成29年2月17日(2017.2.17)

(51) Int.Cl. F I
G06F 17/28 (2006.01) G O 6 F 17/28 6 0 9
G06F 17/27 (2006.01) G O 6 F 17/27 6 7 5

請求項の数 11 (全 19 頁)

<p>(21) 出願番号 特願2013-97857 (P2013-97857) (22) 出願日 平成25年5月7日(2013.5.7) (65) 公開番号 特開2014-219809 (P2014-219809A) (43) 公開日 平成26年11月20日(2014.11.20) 審査請求日 平成28年3月14日(2016.3.14)</p> <p>特許法第30条第2項適用 ウェブサイトの掲載日 平成24年11月15日 ウェブサイトのアドレス http://ipsj.ixsq.nii.ac.jp/ej/</p>	<p>(73) 特許権者 504143441 国立大学法人 奈良先端科学技術大学院大学 奈良県生駒市高山町8916-5</p> <p>(74) 代理人 100114476 弁理士 政木 良文</p> <p>(72) 発明者 藤田 朋希 奈良県生駒市高山町8916-5 国立大学法人奈良先端科学技術大学院大学内</p> <p>(72) 発明者 グラム ニュービッド 奈良県生駒市高山町8916-5 国立大学法人奈良先端科学技術大学院大学内</p> <p>(72) 発明者 サクリアニ サクティ 奈良県生駒市高山町8916-5 国立大学法人奈良先端科学技術大学院大学内 最終頁に続く</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

(54) 【発明の名称】 テキストデータ分割装置、テキストデータ分割方法、テキストデータ分割プログラム及び翻訳装置

(57) 【特許請求の範囲】

【請求項1】

原言語の一連の文字列から成るテキストデータを目的言語に翻訳する際に、前記テキストデータを分割して外部に出力するテキストデータ分割装置であって、

前記原言語の少なくとも1つのユニットから成るフレーズである原言語フレーズと、当該原言語フレーズに対応する前記目的言語のフレーズである目的言語フレーズと、から成るフレーズペアを規定するフレーズテーブルを記録するデータベースと、

前記データベースに記録されている前記フレーズテーブルを参照することで、入力される前記テキストデータの先頭から、前記原言語フレーズを順次検出するとともに、検出された少なくとも1つの前記原言語フレーズから成る翻訳対象データを順次生成し、前記翻訳対象データを生成する毎に外部に出力する翻訳対象データ生成部と、

を備え、

前記フレーズテーブルは、前記原言語フレーズの直後に続く前記原言語のフレーズに対応する前記目的言語の目的言語後続フレーズが、前記目的言語フレーズの後方に位置する確率である右確率を、前記原言語フレーズ毎に規定しており、

前記翻訳対象データ生成部は、

前記原言語フレーズの前記右確率が所定の閾値以上であると、当該原言語フレーズで終わる前記翻訳対象データを生成し、

前記原言語フレーズの前記右確率が前記閾値よりも小さいと、当該原言語フレーズの直後に少なくとも1つの前記原言語フレーズが連結された前記翻訳対象データを生成するこ

10

20

とを特徴とするテキストデータ分割装置。

【請求項 2】

前記テキストデータを構成する前記ユニットを検出して、前記翻訳対象データ生成部に対して前記テキストデータを前記ユニット毎に順次出力するユニット検出部を、さらに備え、

前記翻訳対象データ生成部は、前記原言語フレーズに該当しなくなるまで、前記ユニット検出部が出力する順に前記ユニットを連結し、前記原言語フレーズに該当しなくなった時点で、最後に連結した前記ユニットを除いた語句を前記原言語フレーズとして検出することを特徴とする請求項 1 に記載のテキストデータ分割装置。

【請求項 3】

前記翻訳対象データ生成部は、前記テキストデータから前記原言語フレーズを検出する処理と、前記テキストデータから前記翻訳対象データを分割して生成する処理と、を並列的に行うことを特徴とする請求項 1 または 2 に記載のテキストデータ分割装置。

【請求項 4】

請求項 1 ~ 3 のいずれか 1 項に記載のテキストデータ分割装置と、

前記データベースが記録する前記フレーズテーブルを参照して、前記テキストデータ分割装置が順次出力する前記翻訳対象データを順次翻訳して翻訳結果を出力する翻訳部と、を備えることを特徴とする翻訳装置。

【請求項 5】

前記データベースが、前記目的言語の語句の並び方および語句の選択の正しさを示す言語モデルを、さらに記録しており、

前記翻訳部は、前記データベースに記録されている前記言語モデルを参照して、前記翻訳対象データを翻訳するものであり、

前記言語モデルは、前記目的言語の文章を集積して成る目的言語コーパスに対して、前記目的言語の語句の並び方および語句の選択の正しさを示す確率を与える統計的な学習処理を行うことで生成されるものであり、前記学習処理は、前記テキストデータ分割装置と同じ方法で前記目的言語コーパスを分割してから行われていることを特徴とする請求項 4 に記載の翻訳装置。

【請求項 6】

集音した音声を電気信号に変換することで音声データを生成する音声データ生成部と、前記音声データ生成部が生成した前記音声データを変換して前記テキストデータを生成するテキストデータ生成部と、をさらに備えることを特徴とする請求項 4 または 5 に記載の翻訳装置。

【請求項 7】

前記翻訳部が出力する前記翻訳結果を音声合成して出力する翻訳結果出力部を、さらに備えることを特徴とする請求項 4 ~ 6 のいずれか 1 項に記載の翻訳装置。

【請求項 8】

コンピュータによって実行される、原言語の一連の文字列から成るテキストデータを目的言語に翻訳する際における前記テキストデータを分割して出力するテキストデータ分割方法であって、

前記原言語の少なくとも 1 つのユニットから成るフレーズである原言語フレーズと、当該原言語フレーズに対応する前記目的言語のフレーズである目的言語フレーズと、から成るフレーズペアを規定するフレーズテーブルを参照することで、前記テキストデータの先頭から、前記原言語フレーズを順次検出する原言語フレーズ検出ステップと、

前記原言語フレーズ検出ステップから得られる少なくとも 1 つの前記原言語フレーズから成る翻訳対象データを順次生成する翻訳対象データ生成ステップと、

前記翻訳対象データ生成ステップで前記翻訳対象データが生成される毎に、当該翻訳対象データを出力する翻訳対象データ出力ステップと、

を備え、

前記フレーズテーブルは、前記原言語フレーズの直後に続く前記原言語のフレーズに対

10

20

30

40

50

応する前記目的言語の目的言語後続フレーズが、前記目的言語フレーズの後方に位置する確率である右確率を、前記原言語のフレーズ毎に規定しており、

前記翻訳対象データ生成ステップでは、

前記原言語フレーズの前記右確率が所定の閾値以上であると、当該原言語フレーズで終わる前記翻訳対象データを生成し、

前記原言語フレーズの前記右確率が前記閾値よりも小さいと、当該原言語フレーズの直後に少なくとも1つの前記原言語フレーズが連結された前記翻訳対象データを生成することを特徴とするテキストデータ分割方法。

【請求項9】

前記テキストデータの先頭から、前記テキストデータを構成する前記ユニットを順次検出するユニット検出ステップを、さらに備え、

前記原言語フレーズ検出ステップでは、前記原言語フレーズに該当しなくなるまで、前記ユニット検出ステップで検出される順に前記ユニットを連結し、前記原言語フレーズに該当しなくなった時点で、最後に連結した前記ユニットを除いた語句を前記原言語フレーズとして検出することを特徴とする請求項8に記載のテキストデータ分割方法。

【請求項10】

前記原言語フレーズ検出ステップと、前記翻訳対象データ生成ステップと、が並列的に行われることを特徴とする請求項8または9に記載のテキストデータ分割方法。

【請求項11】

請求項8～10のいずれか1項に記載のテキストデータ分割方法における各ステップを、コンピュータ上で実行するプログラムステップを含むことを特徴とするテキストデータ分割プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、原言語を目的言語に翻訳するために原言語のテキストデータを分割するテキストデータ分割装置、テキストデータ分割方法及びテキストデータ分割プログラムと、当該テキストデータ分割装置を用いた翻訳装置と、に関する。

【背景技術】

【0002】

ビジネス、教育、旅行など、様々な分野でグローバル化が進んでおり、外国の言語を見たり聞いたりする機会が増えている。しかし、言語の習得は容易ではなく、外国の言語に触れた時に戸惑ってしまう人は少なくない。

【0003】

そこで、入力されるテキストデータを翻訳して出力する翻訳装置が、広く利用されている。また、近年では、入力された音声を認識してテキストデータを生成し、当該テキストデータを翻訳して合成音声や文字画像として出力することで、入力された音声をリアルタイムで翻訳して出力する翻訳装置が開発されている。

【0004】

このような翻訳装置では、音声の入力から翻訳結果の出力までの時間が短いほど好ましい。また、入力される音声を翻訳する翻訳装置に限らず、入力されるテキストデータを翻訳する翻訳装置であっても、テキストデータの入力から翻訳結果の出力までの時間が短いほど好ましい。

【0005】

しかし、翻訳結果の出力時間を短くするために、入力されたテキストデータを、例えば単語毎に順次翻訳すると、単語の前後関係を無視した翻訳が行われるため、翻訳精度が著しく低下してしまう。反対に、テキストデータの全文が入力された後に翻訳を開始すると、翻訳精度を向上させることはできるが、翻訳結果が出力されるまでに多大な時間を要してしまう。

【0006】

10

20

30

40

50

そこで、音声が入力されないこと（ポーズ）を検出するとともに、入力された音声を認識して得られたテキストデータを、当該ポーズの位置で分割するテキストデータ分割装置が提案されている。このテキストデータ分割装置を用いた翻訳装置であれば、テキストデータをまとめた語句で分割して翻訳することができるため、翻訳精度の低下を抑制しながら迅速に翻訳結果を出力することが可能になる。

【先行技術文献】

【特許文献】

【0007】

【特許文献1】特開2009-58671号公報

【発明の概要】

【発明が解決しようとする課題】

【0008】

しかしながら、特許文献1で提案されているテキストデータ分割装置では、話し方に応じてテキストデータが分割されることから、必ずしも翻訳に適した位置でテキストデータが分割されないため、問題となる。

【0009】

具体的には、例えば、言い淀みが多くポーズが認識され易い話し方では、形態素の間など、分割すると意味が著しく異なってしまう位置でテキストデータが分割されることがあるため、翻訳精度が低下し得る。一方、息継ぎが短くポーズが認識され難い話し方では、テキストデータを十分に分割することができないため、翻訳速度が低下し得る。

【0010】

そこで、本発明は、精度良くかつ迅速に翻訳することができるようにテキストデータを分割するテキストデータ分割装置、テキストデータ分割方法及びテキストデータ分割プログラムと、当該テキストデータ分割装置を備えた翻訳装置と、を提供することを目的とする。

【課題を解決するための手段】

【0011】

上記目的を達成するため、本発明は、原言語の一連の文字列から成るテキストデータを目的言語に翻訳する際に、前記テキストデータを分割して外部に出力するテキストデータ分割装置であって、前記原言語の少なくとも1つのユニットから成るフレーズである原言語フレーズと、当該原言語フレーズに対応する前記目的言語のフレーズである目的言語フレーズと、から成るフレーズペアを規定するフレーズテーブルを記録するデータベースと、前記データベースに記録されている前記フレーズテーブルを参照することで、入力される前記テキストデータの先頭から、前記原言語フレーズを順次検出するとともに、検出された少なくとも1つの前記原言語フレーズから成る翻訳対象データを順次生成し、前記翻訳対象データを生成する毎に外部に出力する翻訳対象データ生成部と、を備えることを特徴とするテキストデータ分割装置を提供する。

【0012】

このテキストデータ分割装置によれば、原言語のテキストデータを、目的言語に翻訳可能な語句である原言語フレーズの単位で分割することで翻訳対象データを生成するとともに、翻訳対象データを生成する都度外部に出力することが可能になる。したがって、原言語のテキストデータを、精度良くかつ迅速に目的言語に翻訳することができるように分割することが可能になる。

【0013】

さらに、上記特徴のテキストデータ分割装置において、前記フレーズテーブルは、前記原言語フレーズの直後に続く前記原言語のフレーズに対応する前記目的言語の目的言語後続フレーズが、前記目的言語フレーズの後方に位置する確率である右確率を、前記原言語フレーズ毎に規定しており、前記翻訳対象データ生成部は、前記原言語フレーズの前記右確率が所定の閾値以上であると、当該原言語フレーズで終わる前記翻訳対象データを生成し、前記原言語フレーズの前記右確率が前記閾値よりも小さいと、当該原言語フレーズの

10

20

30

40

50

直後に少なくとも1つの前記原言語フレーズが連結された前記翻訳対象データを生成すると、好ましい。

【0014】

このテキストデータ分割装置によれば、出力される順に翻訳対象データを目的言語に翻訳すると、正しい語順となる確率が高くなるため、さらに精度良く翻訳することが可能になる。

【0015】

さらに、上記特徴のテキストデータ分割装置において、前記テキストデータを構成する前記ユニットを検出して、前記翻訳対象データ生成部に対して前記テキストデータを前記ユニット毎に順次出力するユニット検出部を、さらに備え、前記翻訳対象データ生成部は、前記原言語フレーズに該当しなくなるまで、前記ユニット検出部が出力する順に前記ユニットを連結し、前記原言語フレーズに該当しなくなった時点で、最後に連結した前記ユニットを除いた語句を前記原言語フレーズとして検出すると、好ましい。

10

【0016】

このテキストデータ分割装置によれば、テキストデータの先頭から順に、できるだけ長い原言語フレーズを検出することが可能である。そのため、この原言語フレーズを用いて構成される翻訳対象データを、精度良く翻訳可能なものとすることができる。

【0017】

さらに、上記特徴のテキストデータ分割装置において、前記翻訳対象データ生成部は、前記テキストデータから前記原言語フレーズを検出する処理と、前記テキストデータから前記翻訳対象データを分割して生成する処理と、を並列的に行うと、好ましい。

20

【0018】

このテキストデータ分割装置によれば、一方の処理の終了を待たずに他方の処理を実行することができるため、効率よく迅速に翻訳対象データを生成することが可能になる。

【0019】

また、本発明は、上記のテキストデータ分割装置と、前記データベースが記録する前記フレーズテーブルを参照して、前記テキストデータ分割装置が順次出力する前記翻訳対象データを順次翻訳して翻訳結果を出力する翻訳部と、を備えることを特徴とする翻訳装置を提供する。

【0020】

さらに、上記特徴の翻訳装置において、前記データベースが、前記目的言語の語句の並び方および語句の選択の正しさを示す言語モデルを、さらに記録しており、前記翻訳部は、前記データベースに記録されている前記言語モデルを参照して、前記翻訳対象データを翻訳するものであり、前記言語モデルは、前記目的言語の文章を集積して成る目的言語コーパスに対して、前記目的言語の語句の並び方および語句の選択の正しさを示す確率を与える統計的な学習処理を行うことで生成されるものであり、前記学習処理は、前記テキストデータ分割装置と同じ方法で前記目的言語コーパスを分割してから行われていると、好ましい。

30

【0021】

この翻訳装置によれば、翻訳対象データの生成時と同じ分割方法で分割された語句から言語モデルが構築されるため、言語モデルを構築した語句の大きさと、翻訳対象データを成す語句の大きさと、を同程度にすることができる。そのため、言語モデルに基づいた翻訳を、精度良く行うことが可能になる。

40

【0022】

さらに、上記特徴の翻訳装置において、集音した音声を変換することで音声データを生成する音声データ生成部と、前記音声データ生成部が生成した前記音声データを変換して前記テキストデータを生成するテキストデータ生成部と、をさらに備えると、好ましい。

【0023】

さらに、上記特徴の翻訳装置において、前記翻訳部が出力する前記翻訳結果を音声合成

50

して出力する翻訳結果出力部を、さらに備えると、好ましい。

【0024】

入力される音声をテキストデータに変換して翻訳したり、翻訳結果を音声合成して出力したりする翻訳装置では、翻訳結果をリアルタイムで生成することが特に強く求められるが、この翻訳装置によれば、上述のようにテキストデータ分割装置が翻訳対象データを順次出力するとともに、翻訳部が翻訳対象データを順次翻訳するため、翻訳結果をリアルタイムで生成することが可能である。

【0025】

また、本発明は、原言語の一連の文字列から成るテキストデータを目的言語に翻訳する際に、前記テキストデータを分割して出力するテキストデータ分割方法であって、前記原言語の少なくとも1つのユニットから成るフレーズである原言語フレーズと、当該原言語フレーズに対応する前記目的言語のフレーズである目的言語フレーズと、から成るフレーズペアを規定するフレーズテーブルを参照することで、前記テキストデータの先頭から、前記原言語フレーズを順次検出する原言語フレーズ検出ステップと、前記原言語フレーズ検出ステップから得られる少なくとも1つの前記原言語フレーズから成る翻訳対象データを順次生成する翻訳対象データ生成ステップと、前記翻訳対象データ生成ステップで前記翻訳対象データが生成される毎に、当該翻訳対象データを出力する翻訳対象データ出力ステップと、を備えることを特徴とするテキストデータ分割方法を提供する。

10

【0026】

さらに、上記特徴のテキストデータ分割方法において、前記フレーズテーブルは、前記原言語フレーズの直後に続く前記原言語のフレーズに対応する前記目的言語の目的言語後続フレーズが、前記目的言語フレーズの後方に位置する確率である右確率を、前記原言語のフレーズ毎に規定しており、前記翻訳対象データ生成ステップでは、前記原言語フレーズの前記右確率が所定の閾値以上であると、当該原言語フレーズで終わる前記翻訳対象データを生成し、前記原言語フレーズの前記右確率が前記閾値よりも小さいと、当該原言語フレーズの直後に少なくとも1つの前記原言語フレーズが連結された前記翻訳対象データを生成すると、好ましい。

20

【0027】

さらに、上記特徴のテキストデータ分割方法において、前記テキストデータの先頭から、前記テキストデータを構成する前記ユニットを順次検出するユニット検出ステップを、さらに備え、前記原言語フレーズ検出ステップでは、前記原言語フレーズに該当しなくなるまで、前記ユニット検出ステップで検出される順に前記ユニットを連結し、前記原言語フレーズに該当しなくなった時点で、最後に連結した前記ユニットを除いた語句を前記原言語フレーズとして検出すると、好ましい。

30

【0028】

さらに、上記特徴のテキストデータ分割方法において、前記原言語フレーズ検出ステップと、前記翻訳対象データ生成ステップと、が並列的に行われると、好ましい。

【0029】

また、本発明は、上記のテキストデータ分割方法における各ステップを、コンピュータ上で実行するプログラムステップを含むことを特徴とするテキストデータ分割プログラムを提供する。

40

【発明の効果】

【0030】

上記特徴のテキストデータ分割装置、翻訳装置、テキストデータ分割方法及びテキストデータ分割プログラムによれば、原言語のテキストデータを、目的言語に翻訳可能な語句である原言語フレーズの単位で分割することで翻訳対象データを生成するとともに、翻訳対象データを生成する都度外部に出力する。したがって、原言語のテキストデータを、精度良くかつ迅速に目的言語に翻訳することができるように分割することが可能になる。

【図面の簡単な説明】

【0031】

50

【図 1】本発明の実施形態に係るテキストデータ分割装置の構成例について示すブロック図。

【図 2】フレーズテーブルの具体例について示す図。

【図 3】翻訳対象データ生成部の具体的な動作例について示したフローチャート。

【図 4】翻訳対象データ生成部の具体的な動作例について示したフローチャート。

【図 5】本発明の実施形態に係る翻訳装置の構成例について示すブロック図。

【図 6】本発明の実施形態に係る翻訳装置における種々のケース毎の翻訳性能を示すグラフ。

【図 7】本発明の実施形態における翻訳装置の翻訳性能と従来の翻訳装置の翻訳性能とを比較して示したグラフ。

10

【発明を実施するための形態】

【0032】

以下、本発明の実施形態に係るテキストデータ分割装置及び翻訳装置について、図面を参照して説明する。なお、テキストデータ分割装置とは、例えば翻訳装置の一部を構成するものであり、原言語（翻訳前の言語、以下同じ）の一連の文字列から成るテキストデータを目的言語（翻訳後の言語、以下同じ）に翻訳する際に、原言語のテキストデータを分割して成る翻訳対象データを出力する装置である。また、以下では、原言語の語句については「」を付して表記し、目的言語の語句については『』を付して表記する。さらに、以下では説明の具体化のため、主として原言語が日本語であり、目的言語が英語である場合について、例示する。

20

【0033】

<テキストデータ分割装置>

最初に、本発明の実施形態に係るテキストデータ分割装置について、図面を参照して説明する。図 1 は、本発明の実施形態に係るテキストデータ分割装置の構成例について示すブロック図である。

【0034】

図 1 に示すように、本発明の実施形態に係るテキストデータ分割装置 10 は、ユニット検出部 11 と、翻訳対象データ生成部 12 と、データベース DB と、を備える。

【0035】

データベース DB は、原言語の少なくとも 1 つのユニットから成るフレーズ（以下、原言語フレーズという）と、当該原言語フレーズの目的言語に対応するフレーズ（以下、目的言語フレーズという）と、から成るフレーズペアを規定するフレーズテーブル PT を記録している。フレーズペアは、例えば対訳コーパス（原言語の文章及び目的言語の文章の対訳を示すデータ）に対して、周知の統計的な学習方法を適用することで抽出することができる。例えば、対訳コーパスに単語アライメントの手法を適用し、その後にフレーズ抽出を行うことで、フレーズペアを生成することができる。なお、フレーズテーブル PT は、テキストデータ分割装置 10 や他の装置によって、データベース DB などに記録されている対訳コーパスが処理されることで生成されたものであってもよいし、予め準備されたものであってもよい。

30

【0036】

ここで、フレーズテーブル PT の具体例について、図面を参照して説明する。図 2 は、フレーズテーブルの具体例について示す図である。

40

【0037】

図 2 に示すように、フレーズテーブル PT では、複数のフレーズペアが規定されている。そして、原言語フレーズ毎に、右確率が規定されている。右確率とは、原言語フレーズの直後に続く原言語のフレーズ（以下、原言語後続フレーズという）に対応する目的言語の目的言語後続フレーズが、目的言語フレーズの後方に位置する確率である。換言すると、原言語フレーズ及び原言語後続フレーズの前後関係と、目的言語フレーズ及び目的言語後続フレーズの前後関係と、が逆順にならない確率である。

【0038】

50

原言語フレーズ及び原言語後続フレーズを翻訳することで得られる、目的言語フレーズ及び目的言語後続フレーズの語順は、全部で以下の4通り存在する。なお、以下では、原言語（日本語）の文や語句に関して、原言語フレーズについては「」を付して表記し、原言語後続フレーズについては《》を付して表記する。例えば、「背の高い 《男》」と表記した場合、「背の高い」が原言語フレーズ、「《男》」が原言語後続フレーズである。また、以下では、目的言語（英語）の文や語句に関して、目的言語フレーズについては『』を付して表記し、目的言語後続フレーズについては《》を付して表記する。例えば、『the tall 《man》』と表記した場合、『the tall』が目的言語フレーズ、『《man》』が目的言語後続フレーズである。

【0039】

(1) 「背の高い 《男》」、『the tall 《man》』のように、目的言語フレーズの直後に目的言語後続フレーズが続く並び方。この並び方を、[連続・同順]という。

(2) 「私は 《太郎を》 訪問した」、『I visited 《Taro》』のように、目的言語フレーズの直後に目的言語後続フレーズが続かないが、目的言語フレーズの後方に目的言語後続フレーズが位置する並び方。この並び方を、[不連続・同順]という。

(3) 「太郎を 《訪問した》」、『《visited》 Taro』のように、目的言語後続フレーズの直後に目的言語フレーズが続く並び方。この並び方を、[連続・逆順]という。

(4) 「背の高い 男を 《訪問した》」、『《visited》 the tall man』のように、目的言語後続フレーズの直後に目的言語フレーズが続かないが、目的言語後続フレーズの後方に目的言語フレーズが位置する並び方。この並び方を、[不連続・逆順]という。

【0040】

右確率とは、ある原言語フレーズ及び原言語後続フレーズを目的言語に翻訳したときに、(1) [連続・同順] 及び (2) [不連続・同順] となる確率である。即ち、上記の例に示すように、右確率とは、原言語後続フレーズを無視して原言語フレーズを翻訳することが可能（順次翻訳が可能）な確率と言うことができる。なお、1つの原言語フレーズに対応する目的言語フレーズが複数ある場合、その1つの原言語フレーズの右確率はそれぞれの目的言語フレーズに応じて複数となる。このような場合、例えば、複数の右確率のうちから最大のものを選択し、当該右確率をその原言語フレーズの右確率として、目的言語フレーズに関わらず、一律に記録することができる。

【0041】

ユニット検出部11は、テキストデータを構成するユニットを検出して、翻訳対象データ生成部12に対してテキストデータをユニット毎に順次出力する。ここで、ユニットとは、文字、形態素、単語のいずれかを意味する。ユニットとして、原言語における文法上の最小単位、または、原言語において意味を有する最小の単位を用いることができる。例えば、ユニットとして、原言語が中国語である場合は文字、日本語である場合は形態素、英語である場合は単語を用いることが好適であるが、これに限らない。ユニット検出部11は、周知の検出方法を用いて、テキストデータを構成するユニットを検出する。例えば、ユニット検出部11は、原言語の文法や単語辞書に基づいて、テキストデータを構成するユニットを検出する。

【0042】

翻訳対象データ生成部12は、データベースDBが記録しているフレーズテーブルPTを参照することで、入力されるテキストデータの先頭から、原言語フレーズを順次検出する。そして、翻訳対象データ生成部12は、検出された少なくとも1つの原言語フレーズから成る翻訳対象データを順次生成する。さらに、翻訳対象データ生成部12は、翻訳対象データを生成する毎に、外部に出力する。

【0043】

10

20

30

40

50

次に、テキストデータ分割装置 10 の具体的な動作例（特に、翻訳対象データ生成部 12 の動作例）について、図面を参照して説明する。図 3 及び図 4 は、翻訳対象データ生成部の具体的な動作例について示したフローチャートである。なお、図 3 は、テキストデータから原言語フレーズを検出する処理を示すものである。また、図 4 は、テキストデータから翻訳対象データを分割して生成する処理を示すものである。また、図 3 及び図 4 に示すフローチャートは、それぞれの処理の 1 サイクル分を示したものであり、これらの処理はそれぞれ繰り返し行われる。

【 0 0 4 4 】

テキストデータ分割装置 10 には、原言語の一連の文字列から成るテキストデータが、先頭から順次入力される。テキストデータ分割装置 10 へのテキストデータの入力が始 10
められると、最初に、ユニット検出部 11 が、テキストデータを構成するユニットを順次検出する。そして、ユニット検出部 11 は、翻訳対象データ生成部 12 に対して、テキストデータをユニット毎に順次出力する。

【 0 0 4 5 】

次に、図 3 に示すように、翻訳対象データ生成部 12 は、ユニット検出部 11 が出力するユニットを取得する（ステップ # 1）。そして、翻訳対象データ生成部 12 は、原言語フレーズに該当するか否かを判断する対象の語句である対象語句を決定する（ステップ # 2）。

【 0 0 4 6 】

翻訳対象データ生成部 12 は、この時点で保留語句（詳細は後述）を有していない場合 20
（例えば、ステップ # 1 において、テキストデータの先頭のユニットを取得した場合）、ステップ # 1 で取得したユニットを、そのまま対象語句とする。一方、翻訳対象データ生成部 12 は、この時点で保留語句を有している場合（例えば、ステップ # 1 において、テキストデータの先頭以外のユニットを取得した場合）、ステップ # 1 で取得したユニットを保留語句の直後に連結して、対象語句とする。

【 0 0 4 7 】

次に、翻訳対象データ生成部 12 は、対象語句がフレーズテーブル P T に規定されている原言語フレーズに該当するか否かを照合するために、フレーズテーブル P T を参照する（ステップ # 3）。

【 0 0 4 8 】

対象語句が原言語フレーズに該当する場合（ステップ # 4 , Y E S）、翻訳対象データ生成部 12 は、新たなユニットが入力されるか否か（直近のステップ # 1 において、テキストデータの末尾ではないユニットが入力されたか否か）を確認する（ステップ # 5）。そして、新たなユニットが入力される場合（ステップ # 5 , Y E S）、翻訳対象データ生成部 12 は、対象語句を上述の保留語句として、ステップ # 1 に戻る。

【 0 0 4 9 】

一方、対象語句が原言語フレーズに該当しない場合（ステップ # 4、N O）、翻訳対象データ生成部 12 は、対象語句から直近のステップ # 2 で連結したユニット（最後に連結したユニット）を除いた語句を、原言語フレーズとして検出する（ステップ # 6）。そして、翻訳対象データ生成部 12 は、対象語句から除かれたユニットを、上述の保留語句と 40
する。

【 0 0 5 0 】

これに対して、対象語句が原言語フレーズに該当する場合であって（ステップ # 4 , Y E S）、新たなユニットが入力されない場合（ステップ # 5、N O）、翻訳対象データ生成部 12 は、対象語句を原言語フレーズとして検出する（ステップ # 7）。

【 0 0 5 1 】

ここで、翻訳対象データ生成部 12 が、図 2 に示したフレーズテーブルを参照して、「私 は 男 です」の日本語のテキストデータに対して図 3 の処理を行った場合について、具体的に例示する。なお、以下に示す例において、ユニットは形態素である。

【 0 0 5 2 】

10

20

30

40

50

まず、翻訳対象データ生成部12は、最初に入力されるテキストデータの先頭のユニット「私」を、そのまま対象語句とする(ステップ#1及びステップ#2)。このとき、翻訳対象データ生成部12は、対象語句「私」が原言語フレーズに該当し(ステップ#4, YES)、新たなユニット「は」が入力されることを確認して(ステップ#5, YES)、「私」を保留語句とする。

【0053】

次に、翻訳対象データ生成部12は、ユニット「は」が入力されると(ステップ#1)、保留語句「私」の直後に連結して「私は」を対象語句とする(ステップ#2)。このとき、翻訳対象データ生成部12は、対象語句「私は」が原言語フレーズに該当し(ステップ#4, YES)、新たなユニット「男」が入力されることを確認して(ステップ#5, YES)、「私は」を保留語句とする。

10

【0054】

次に、翻訳対象データ生成部12は、ユニット「男」が入力されると(ステップ#1)、保留語句「私は」の直後に連結して「私は男」を対象語句とする(ステップ#2)。このとき、翻訳対象データ生成部12は、対象語句「私は男」が原言語フレーズに該当しないことを確認する(ステップ#4, NO)。すると、翻訳対象データ生成部12は、最後に連結したユニット「男」を除いた語句「私は」を、原言語フレーズとして検出する(ステップ#6)。一方、翻訳対象データ生成部12は、対象語句「私は男」から除いたユニット「男」を保留語句とする。

【0055】

これにより、図3に示した1サイクル分の処理が行われたことになる。ただし、上述のように、図3に示す処理は繰り返し行われるため、引き続き原言語フレーズの検出が行われる。

20

【0056】

次に、翻訳対象データ生成部12は、ユニット「です」が入力されると(ステップ#1)、保留語句「男」の直後に連結して「男です」を対象語句とする(ステップ#2)。このとき、翻訳対象データ生成部12は、対象語句「男です」が原言語フレーズに該当するが(ステップ#4, YES)、新たなユニットが入力されないことを確認する(ステップ#5, NO)。すると、翻訳対象データ生成部12は、対象語句「男です」を、原言語フレーズとして検出する。

30

【0057】

このように、翻訳対象データ生成部12は、テキストデータの先頭から順に、できるだけ長い原言語フレーズを検出することが可能である。そのため、この原言語フレーズを用いて構成される翻訳対象データを、精度良く翻訳可能なものとすることができる。

【0058】

次に、図4に示すように、翻訳対象データ生成部12は、図3の処理の繰り返しによって順次検出される原言語フレーズの1つを選択して、処理対象の原言語フレーズである対象原言語フレーズとして決定する(ステップ#10)。このとき、翻訳対象データ生成部12は、図3の処理によって検出された順番(テキストデータの先頭から末尾に向かう順番)で、対象原言語フレーズとするべき原言語フレーズを順次選択する。

40

【0059】

次に、翻訳対象データ生成部12は、フレーズテーブルPTを参照して、対象原言語フレーズの右確率を確認する(ステップ#11)。そして、翻訳対象データ生成部12は、対象原言語フレーズの右確率と所定の閾値とを比較する(ステップ#12)。

【0060】

翻訳対象データ生成部12が、対象原言語フレーズの右確率が所定の閾値よりも小さく(ステップ#12, NO)、当該対象原言語フレーズが文末ではないことを確認すると(ステップ#13, NO)、当該対象原言語フレーズをスタック(メモリ)に保存することで、スタックフレーズを生成する(ステップ#14)。スタックフレーズとは、スタックに保存された対象原言語フレーズを保存された順番に連結したものであり、順番的に後で

50

保存された対象原言語フレーズほどスタックフレーズの後方を成す。また、対象原言語フレーズが文末ではない場合とは、例えば、対象原言語フレーズが、テキストデータの末尾の原言語フレーズではない場合や、フレーズテーブルPTで文末である確率が高いと規定されている特定の原言語フレーズではない場合などである。なお、本発明においては、必ずしも文末であるかどうかを確認するステップ(ステップ#13)を入れる必要はない。つまり、文末に相当する対象原言語フレーズの右確率は比較的大きなものになるため、自ずとステップ#12からステップ#15へ進む確率が高いからである。ただし、ステップ#13を設けることで、文末であり、かつ閾値より右確率の低い対象原言語フレーズから翻訳対象データを生成することが可能となる。

【0061】

翻訳対象データ生成部12は、ステップ#14でスタックフレーズを生成すると、ステップ#10に戻って次の対象原言語フレーズを決定する。そして、翻訳対象データ生成部12は、フレーズテーブルPTを参照して対象原言語フレーズの右確率を確認し(ステップ#11)、対象原言語フレーズの右確率と所定の閾値とを比較する(ステップ#12)。

【0062】

一方、翻訳対象データ生成部12は、対象原言語フレーズの右確率が所定の閾値以上であると(ステップ#12, YES)、スタックフレーズの後に対象原言語フレーズを連結することで翻訳対象データを生成する(ステップ#15)。このとき、スタックにスタックフレーズが無ければ、対象原言語フレーズから成る翻訳対象データを生成する。このようにして生成される翻訳対象データは、右確率が所定の閾値以上である対象原言語フレーズで終わるものとなる。

【0063】

また、翻訳対象データ生成部12は、対象原言語フレーズの右確率が所定の閾値よりも小さく(ステップ#12, NO)、当該対象原言語フレーズが文末であることを確認する場合も(ステップ#13, YES)、上記の場合と同様にスタックフレーズの後に対象原言語フレーズを連結することで翻訳対象データを生成する(ステップ#15)。

【0064】

そして、翻訳対象データ生成部12は、スタックをクリアし(ステップ#16)、生成した翻訳対象データを外部に出力する(ステップ#14)。

【0065】

ここで、翻訳対象データ生成部12が、図2に示したフレーズテーブルを参照して、「何時から プレー できますか」の日本語のテキストデータに対して図4の処理を行った場合について、具体的に例示する。なお、以下の具体例では、ステップ#12の閾値を0.8としている。

【0066】

まず、翻訳対象データ生成部12は、テキストデータの先頭から検出される原言語フレーズ「何時から」を、対象原言語フレーズとする(ステップ#10)。このとき、翻訳対象データ生成部12は、フレーズテーブルPTの原言語フレーズ「何時から」の右確率を参照して、その右確率が0.8333であって閾値0.8以上であることを確認する(ステップ#11及びステップ#12, YES)。すると、翻訳対象データ生成部12は、この時点ではスタックフレーズが無いので、対象原言語フレーズ「何時から」をそのまま翻訳対象データとして生成する(ステップ#15)。そして、翻訳対象データ生成部12は、スタックをクリアするとともに(ステップ#16)、生成した翻訳対象データ「何時から」を外部に出力する(ステップ#17)。

【0067】

これにより、図4に示した1サイクル分の処理が行われたことになる。ただし、上述のように、図4に示す処理は繰り返し行われるため、引き続き翻訳対象データの生成が行われる。

【0068】

次に、翻訳対象データ生成部12は、原言語フレーズ「何時から」の次に検出される原言語フレーズ「プレー」を、対象原言語フレーズとする(ステップ#10)。このとき、翻訳対象データ生成部12は、フレーズテーブルPTの原言語フレーズ「プレー」の右確率を参照して、その右確率が0.25であって閾値0.8よりも小さく(ステップ#11及びステップ#12, NO)、対象原言語フレーズ「プレー」が文末ではないことを確認する(ステップ#13, NO)。すると、翻訳対象データ生成部12は、対象原言語フレーズ「プレー」をスタックに保存することで、スタックフレーズを生成する(ステップ#15)。なお、この時点ではスタックフレーズが無いため、対象原言語フレーズ「プレー」が、そのままスタックフレーズとなってスタックに保存される。

【0069】

10

次に、翻訳対象データ生成部12は、原言語フレーズ「プレー」の次に検出される原言語フレーズ「できますか」を、対象原言語フレーズとする(ステップ#10)。このとき、翻訳対象データ生成部12は、フレーズテーブルPTの原言語フレーズ「できますか」の右確率を参照して、その右確率が0.875であって閾値0.8以上であることを確認する(ステップ#11及びステップ#12, YES)。すると、翻訳対象データ生成部12は、すでにスタックに保存されているスタックフレーズ「プレー」の後に、対象原言語フレーズ「できますか」を連結することで、翻訳対象データ「プレーできますか」を生成する(ステップ#15)。そして、翻訳対象データ生成部12は、スタックをクリアするとともに(ステップ#16)、生成した翻訳対象データ「プレーできますか」を外部に出力する(ステップ#17)。

20

【0070】

この具体例の場合、テキストデータ分割装置10から、まず翻訳対象データ「何時から」が出力され、その次に翻訳対象データ「プレーできますか」が出力される。そして、テキストデータ分割装置10が翻訳対象データを出力する毎に、順次翻訳することによって、『From what time』『can we play?』の翻訳結果が得られる。

【0071】

以上のように、テキストデータ分割装置10は、原言語のテキストデータを、目的言語に翻訳可能な語句である原言語フレーズの単位で分割することで翻訳対象データを生成するとともに、翻訳対象データを生成する都度外部に出力する。したがって、原言語のテキストデータを、精度良くかつ迅速に目的言語に翻訳することができるように分割することが可能になる。

30

【0072】

さらに、テキストデータ分割装置10は、原言語フレーズの右確率が閾値以上であると(後続する原言語フレーズを無視して即座に翻訳することができる確率が高いと)、当該原言語フレーズで終わる翻訳対象データを生成するが、原言語フレーズの右確率が閾値よりも小さいと、当該原言語フレーズの直後に少なくとも1つの原言語フレーズを連結して翻訳対象データを生成する。これにより、テキストデータ分割装置10が出力する順に翻訳対象データを目的言語に翻訳すると、正しい語順となる確率が高くなるため、さらに精度良く翻訳することが可能になる。

【0073】

40

なお、上述した閾値は、0以上1以下の範囲内で、翻訳目的等に応じて任意に設定することが可能である。例えば、翻訳速度よりも翻訳精度が重視される場合や、語順が大きく異なる言語間(例えば、日本語及び英語間)の翻訳を行う場合は、1に近い閾値を設定すると、好ましい。一方、翻訳精度よりも翻訳速度が重視される場合や、語順が同様である言語間(例えば、英語及びフランス語間)の翻訳を行う場合は、0に近い閾値を設定すると、好ましい。

【0074】

また、閾値を0にする場合(即ち、原言語フレーズがそのまま翻訳対象データとなる場合)、図2のフレーズテーブルPTで右確率を規定せず、図4のステップ#11~14, 16を無くしてもよい。あるいは、閾値を1にする場合(即ち、入力されたテキストデー

50

タがそのまま翻訳対象データとなる場合)、図2のフレーズテーブルPTで右確率を規定せず、図4のステップ#11, 12を無くすとともに、常にステップ#13が行われるようにしてもよい。これらの場合、テキストデータ分割装置10の構成及び処理内容を、簡素化することが可能になる。

【0075】

また、テキストデータ分割装置10が、図3に示す処理(テキストデータから原言語フレーズを検出する処理)と、図4に示す処理(テキストデータから翻訳対象データを分割して生成する処理)と、を並列的に行う(例えば、パイプライン処理する)と、一方の処理の終了を待たずに他方の処理を実行することができるため、効率よく迅速に翻訳対象データを生成することが可能になる。

10

【0076】

また、テキストデータ分割装置10が実行する各処理は、少なくとも1つのコンピュータのハードウェア資源(CPU: Central Processing Unit、各種記憶装置など)及びソフトウェア資源(OS: Operating System、各種ドライバなど)を使用した演算処理によって行われる。さらに、かかる演算処理は、CPUによりその実行が制御されるプログラムを実行することによって、ソフトウェア的に実現される。そのため、当該プログラムには、ユニット検出部11及び翻訳対象データ生成部12が行う各処理をコンピュータ上で実行するプログラムステップが含まれる。

【0077】

< 翻訳装置 >

20

次に、上述したテキストデータ分割装置10を備えた翻訳装置について、図面を参照して説明する。図5は、本発明の実施形態に係る翻訳装置の構成例について示すブロック図である。

【0078】

図5に示すように、本発明の実施形態に係る翻訳装置1は、テキストデータ分割装置10と、音声データ生成部20と、テキストデータ生成部30と、翻訳部40と、翻訳結果出力部50と、を備える。なお、図5では、説明の便宜上、データベースDBをテキストデータ分割装置10から分離して図示している。

【0079】

音声データ生成部20は、例えばマイクロフォン等から成り、集音した音声を電気信号に変換することで音声データを生成する。

30

【0080】

テキストデータ生成部30は、周知の音声認識方法を用いて、音声データ生成部20が生成した音声データを文字に変換することで、テキストデータを生成する。例えば、テキストデータ生成部30は、事前に構築した音響モデル(音声と文字との対応を示すデータ)に基づいて、入力される音声データの音声認識を行う。この音響モデルは、例えばデータベースDBに記録される。

【0081】

また、テキストデータ生成部30は、音声データまたはテキストデータの切れ目(データの末尾及び先頭)を検出する周知の検出方法(例えば、音声データから一定時間以上の無音状態(ポーズ)を検出する方法など)を用いて、音声データまたはテキストデータの切れ目を検出する。そして、テキストデータ生成部30は、当該切れ目の前後で別となるテキストデータを生成する。

40

【0082】

テキストデータ分割装置10は、上述のように、データベースDBに記録されているフレーズテーブルPTを参照することで、テキストデータ生成部30が生成したテキストデータから翻訳対象データを順次分割して生成し、順次出力する。

【0083】

翻訳部40は、データベースDBに記録されているフレーズテーブルPTと言語モデルLMとを参照して、テキストデータ分割装置10が順次出力する翻訳対象データの翻訳を

50

順次行い、その翻訳結果を順次出力する。

【 0 0 8 4 】

言語モデルLMとは、目的言語の語句の並び方および語句の選択の正しさ（より具体的には、慣用性、流暢性）を示すものである。例えば、言語モデルLMは、目的言語の文章を集積して成る目的言語コーパス（例えば、上述した対訳コーパスの一部を成す目的言語の文章のデータ）に対して、目的言語の語句の並び方および語句の選択の正しさを示す確率を与える統計的な学習処理を行うことで構築される。

【 0 0 8 5 】

ここで、本発明の実施形態に係る翻訳装置1が、n-gramを利用した言語モデルLMを用いる場合を例示して説明する。この言語モデルLMは、対象となる目的言語の単語が、特定のn-1個の目的言語の単語に後続して使用される条件付き確率を表すものである。即ち、この言語モデルLMは、対象となる目的言語の単語の、目的言語的に正しい用法（より具体的には、慣用的な用法、流暢な用法）を、条件付き確率の高さとして示したものと見える。

【 0 0 8 6 】

具体的に、『I am a man </s>』『I am tired </s>』の2文に基づいて、2-gramの言語モデルLMを構築する場合について例示する。なお、上記例文中の『</s>』は、文末を示す記号である。

【 0 0 8 7 】

まず、それぞれの単語の出現頻度を求める。例えば、『I』の出現頻度C(I)は2、『am』の出現頻度C(am)は2、『a』の出現頻度C(a)は1、『</s>』の出現頻度C(</s>)は2である。同様に、2つの単語の組み合わせの出現頻度を求める。例えば、『I am』の出現頻度C(I am)は2、『am a』の出現頻度C(am a)は1、『am tired』の出現頻度C(am tired)は1、『man </s>』の出現頻度C(man </s>)は1である。

【 0 0 8 8 】

この場合、例えば『am』の後に『a』が用いられる条件付き確率P(a|am)は、 $C(am a) / C(am) = 1 / 2 = 0.5$ となる。また例えば、『I』の後に『am』を用いる条件付き確率P(am|I)は、 $C(I am) / C(I) = 2 / 2 = 1$ となる。また例えば、『man』で文が終わる条件付き確率P(</s>|man)は、 $C(man </s>) / C(man) = 1 / 1 = 1$ となる。

【 0 0 8 9 】

翻訳部40は、フレーズテーブルPTだけでなく言語モデルLMをも参照することによって、フレーズテーブルPTから目的言語的に正しい目的言語フレーズを選択したり、目的言語的に正しい語順や言い回しとなるように目的言語フレーズを並べたり修正したりすることが可能になる。

【 0 0 9 0 】

翻訳結果出力部50は、例えば翻訳結果を文字画像として出力するディスプレイや、翻訳結果を音声合成して出力するスピーカ等から成り、人が知覚可能な態様で翻訳結果を出力する。なお、入力される音声をテキストデータに変換して翻訳したり、翻訳結果を音声合成して出力したりする翻訳装置1では、翻訳結果をリアルタイムで生成することが特に強く求められるが、この翻訳装置1では、上述のようにテキストデータ分割装置10が翻訳対象データを順次出力するとともに、翻訳部40が翻訳対象データを順次翻訳するため、翻訳結果をリアルタイムで生成することが可能である。

【 0 0 9 1 】

なお、上述した言語モデルLMを構築する際に、図3及び図4で述べたテキストデータの分割方法を利用してもよい。この場合、上述の例のような所定の文（例えば、『I am a man </s>』『I am tired </s>』）に対する学習処理によって言語モデルLMが構築されるのではなく、所定の語句（例えば、『I am』『a man </s>』『I am』『tired </s>』）に対する学習処理によって言語モデルLMが構築される。

【 0 0 9 2 】

10

20

30

40

50

上述のように、翻訳対象データは、テキストデータを原言語フレーズの単位で分割したものである。そのため、設定される閾値にも依るが、原則として原言語の文を分割した語句となっている。このような翻訳対象データに対して、文に対する学習処理によって構築した言語モデルLMを用いた翻訳を行うと、翻訳精度が低下することがある。具体的に例えば、翻訳部40が、順次入力される翻訳対象データを、それぞれ一文であると判断して、それぞれの翻訳結果に文末記号</s>を付してしまうことがある。

【0093】

そのため、テキストデータ分割装置10と同じ方法（特に、同じ閾値）で目的言語コーパスを分割することで目的言語の語句を生成して、当該語句に対する学習処理によって言語モデルLMを構築すると、好ましい。この場合、翻訳対象データの生成時と同じ分割方法で分割された語句に対する学習処理によって言語モデルLMが構築されるため、言語モデルLMを構築した語句の大きさと、翻訳対象データを成す語句の大きさと、を同程度にすることができる。したがって、言語モデルに基づいた翻訳を、精度良く行うことが可能になる。

10

【0094】

上記のように言語モデルLMを構築する場合、テキストデータ分割装置10の閾値を変更する毎に、言語モデルLMの再構築が必要になる。しかし、言語モデルLMの再構築は、短時間（例えば、1時間程度）で済ませることが可能である。なお、テキストデータ分割装置10が設定可能な複数の閾値に対応する複数の言語モデルLMを予め構築しておき、それぞれをデータベースDBに記録しておいてもよい。

20

【0095】

また、翻訳装置1の一部を成すテキストデータ生成部30、テキストデータ分割装置10及び翻訳部40のそれぞれが行う各処理は、少なくとも1つのコンピュータのハードウェア資源（CPU、各種記憶装置など）及びソフトウェア資源（OS、各種ドライバなど）を使用した演算処理によって行われる。さらに、かかる演算処理は、CPUによりその実行が制御されるプログラムを実行することによって、ソフトウェア的に実現される。そのため、当該プログラムには、テキストデータ生成部30、テキストデータ分割装置10及び翻訳部40のそれぞれが行う各処理をコンピュータ上で実行するプログラムステップが含まれる。

【0096】

< 翻訳性能 >

上述した翻訳装置1の翻訳性能の一例について、図面を参照して説明する。なお、以下では、翻訳精度を示すBLEUSコアと、翻訳速度を示す遅延時間と、を用いて翻訳性能を表す。BLEUSコアは、例えば人が翻訳した正確な翻訳結果である翻訳モデルに対して、翻訳装置が生成した翻訳結果が類似する程度を、数値化したものである。また、遅延時間は、テキストデータ分割装置10にテキストデータが入力されてから翻訳部40によって翻訳結果が生成されるまでに要した時間である。したがって、BLEUSコアが高いほど翻訳精度が高く、遅延時間が短いほど翻訳速度が速いことになる。

30

【0097】

図6は、本発明の実施形態に係る翻訳装置における種々のケース毎の翻訳性能を示すグラフである。具体的に、図6では、英語の旅行対話を日本語に翻訳したケースにおける翻訳性能を、白塗りの正方形のマーカ（ ）で示している。また、日本語の旅行対話を英語に翻訳したケースにおける翻訳性能を、黒塗りの正方形のマーカ（ ）で示している。また、日本語の長文（11単語以上）の旅行対話を英語に翻訳したケースにおける翻訳性能を、黒塗りの正三角形のマーカ（ ）で示している。また、フランス語のニュース文を英語に翻訳したケースにおける翻訳性能を、白塗りの正三角形のマーカ（ ）で示している。また、図6では、横軸を遅延時間（秒）、縦軸をBLEUSコアとしている。

40

【0098】

また、図6では、それぞれのケースにおいて、閾値を0から1まで0.2ずつ異ならせながら求めた6個の翻訳性能を6個のマーカで示しているが、遅延時間が0に近いものほ

50

ど閾値が小さく、遅延時間が長いものほど閾値が大きくなっている。

【0099】

図6に示すように、全てのケースにおいて、閾値を適宜選択することによって、翻訳精度を維持しながら翻訳速度（遅延時間）を向上することが可能である。即ち、本発明の実施形態における翻訳装置1は、原言語及び目的言語の種類や文の種類（長短、文体）などを問わず、精度良くかつ迅速に翻訳することが可能である。なお、上述のように、英語及びフランス語は語順が同様であるため、閾値を0に近づけても、翻訳精度の低下を抑制することができる。そしてその一方で、閾値を0に近づけると、翻訳速度を格段に向上させることができる。

【0100】

また、図7は、本発明の実施形態における翻訳装置の翻訳性能と従来の翻訳装置の翻訳性能とを比較して示したグラフである。具体的に、図7は、日本語を英語に翻訳するケースにおいて、翻訳装置1の翻訳性能を白塗りの丸のマーカ（○）で示し、例えば特許文献1のようなポーズでテキストデータを分割して翻訳対象データを生成する従来の翻訳装置の翻訳性能を黒塗りの丸のマーカ（●）で示している。また、図6と同様に、横軸を遅延時間（秒）、縦軸をBLEUSコアとしている。

【0101】

また、図6と同様に図7でも、翻訳装置1の翻訳性能について、閾値を0から1まで少しずつ異ならせながら求めた複数の翻訳性能を複数のマーカ（具体的には、0.0、0.2、0.4、0.6、0.7、0.8、0.9、1.0の8個）で示しており、遅延時間が0に近いものほど閾値が小さく、遅延時間が長いものほど閾値が大きくなっている。

【0102】

図7に示すように、従来の翻訳装置の翻訳精度と同様になるように、翻訳装置1の閾値を設定した場合（図中の破線参照）、従来の翻訳装置よりも、遅延時間を20%程度短くすることができる。したがって、本発明の実施形態における翻訳装置1は、従来の翻訳装置と同程度の翻訳精度を維持しながら、従来の翻訳装置よりも翻訳速度を向上させることが可能である。

【0103】

<変形等>

図5において、集音した音声を認識することで生成されたテキストデータを翻訳する翻訳装置1に、本発明の実施形態に係るテキストデータ分割装置10を適用する場合について例示したが、このテキストデータ分割装置10は、外部からテキストデータが入力される翻訳装置にも適用可能である。そして、このような翻訳装置に適用しても、上述の翻訳装置1と同様に、精度良くかつ迅速に目的言語に翻訳する効果を得ることができる。

【産業上の利用可能性】

【0104】

本発明は、テキストデータを分割するテキストデータ分割装置、テキストデータ分割方法及びテキストデータ分割プログラムや、当該テキストデータ分割装置を用いた翻訳装置に利用可能である。特に、本発明は、入力された音声をリアルタイムで翻訳して出力する翻訳装置や、当該翻訳装置に用いられるテキストデータ分割装置、テキストデータ分割方法及びテキストデータ分割プログラムに、好適に利用可能である。

【符号の説明】

【0105】

- 1 : 翻訳装置
- 10 : テキストデータ分割装置
- 11 : ユニット検出部
- 12 : 翻訳対象データ生成部
- 20 : 音声データ生成部
- 30 : テキストデータ生成部
- 40 : 翻訳部

10

20

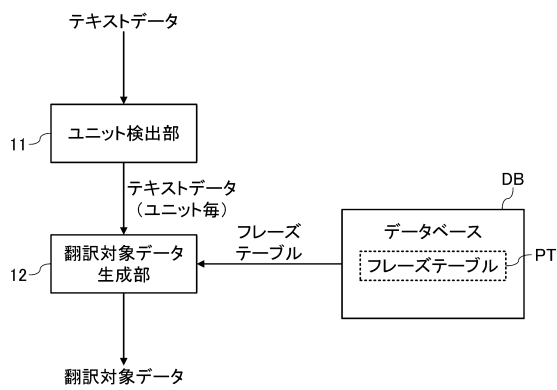
30

40

50

5 0 : 翻訳結果出力部
 D B : データベース
 P T : フレーズテーブル
 L M : 言語モデル

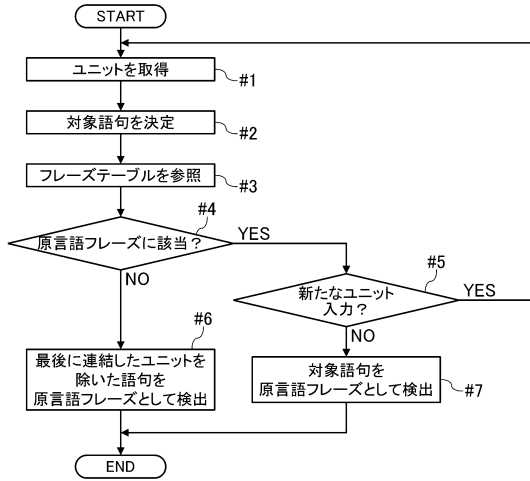
【図1】



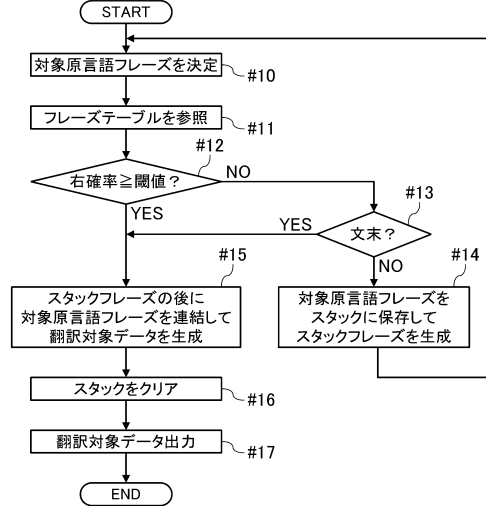
【図2】

原言語フレーズ	目的言語フレーズ	右確率
私	I	0.75
私は	I	0.75
男	man	0.75
男です	am a man	0.75
何	what	0.875
何時	what time	0.8333
何時から	from what time	0.8333
プレー	play	0.25
でき	can	0.9
できますか	?	0.875

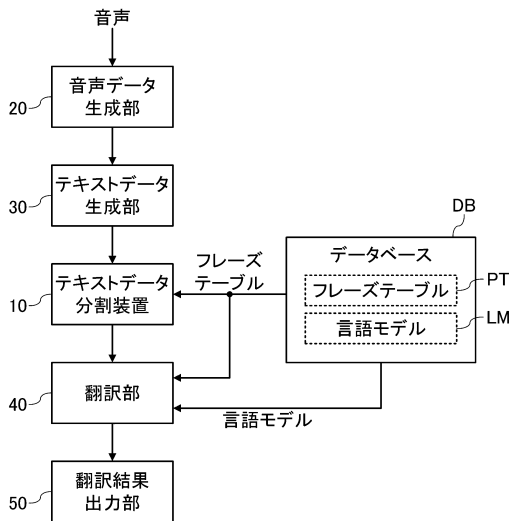
【図3】



【図4】

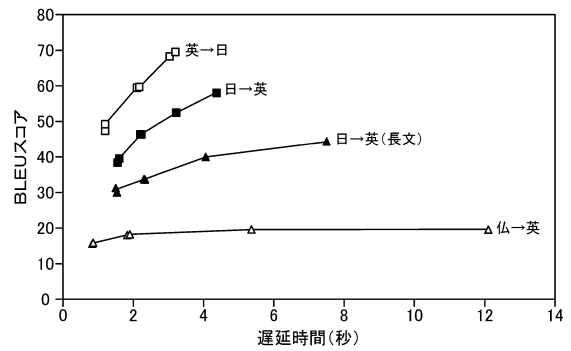


【図5】

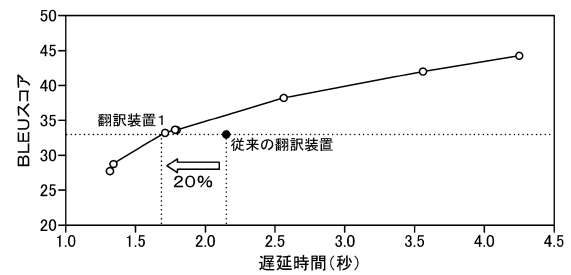


1

【図6】



【図7】



フロントページの続き

(72)発明者 戸田 智基

奈良県生駒市高山町8916-5 国立大学法人奈良先端科学技術大学院大学内

(72)発明者 中村 哲

奈良県生駒市高山町8916-5 国立大学法人奈良先端科学技術大学院大学内

審査官 成瀬 博之

(56)参考文献 特開2001-117920(JP,A)

特開2012-185622(JP,A)

特開2012-073941(JP,A)

清水徹 他2名, 翻訳可能なフレーズを単位とした音声認識結果の信頼度評価, 日本音響学会2005年春季研究発表会講演論文集-I-, 日本, 社団法人日本音響学会, 2005年3月8日, 53-54頁

(58)調査した分野(Int.Cl., DB名)

G06F 17/20-17/28