(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2014-219809
(P2014-219809A)
(43)公開日　平成26年11月20日(2014.11.20)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 17/27	(2006.01)	G06F 17/27	D	5B091

審査請求 未請求　請求項の数 13　ＯＬ　（全 20 頁）

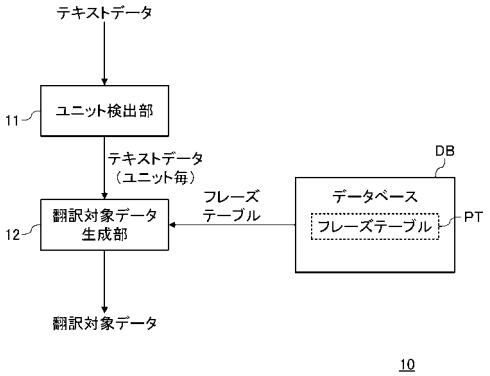
(21)出願番号　特願2013-97857(P2013-97857)
(22)出願日　平成25年5月7日(2013.5.7)

特許法第３０条第２項適用申請有り　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ｈｔｔｐｓ：／／ｉｐｓｊ．ｉｘｓｑ．ｎｉｉ．ａｃ．ｊｐ／ｅｊ／

(71)出願人　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
(74)代理人　100114476
弁理士　政木　良文
(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
(72)発明者　グラム　ニュービッグ
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　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

(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)【要約】
【課題】　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。
【解決手段】　テキストデータ分割装置１０は、原言語フレーズと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ＰＴを記録するデータベースＤＢと、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、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とともに、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し、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１２と、を備える。
【選択図】　図１



【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、
を備えることを特徴とするテキストデータ分割装置。
【請求項２】
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または２に記載のテキストデータ分割装置。
【請求項４】
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１～３のいずれか１項に記載のテキストデータ分割装置。
【請求項５】
請求項１～４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６】
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５に記載の翻訳装置。

【請求項７】
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５または６に記載の翻訳装置。
【請求項８】
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５～７のいずれか１項に記載の翻訳装置。
【請求項９】
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。
【請求項１０】
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ことを特徴とする請求項９に記載のテキストデータ分割方法。
【請求項１１】
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９または１０に記載のテキストデータ分割方法。
【請求項１２】
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９～１１のいずれか１項に記載のテキストデータ分割方法。
【請求項１３】
請求項９～１２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】

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まっ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。
【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。
【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。

【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。

【００２２】

さらに、上記特徴の翻訳装置において、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。

【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。
【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。
【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。
【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】
【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。
【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。
【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。

【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。

【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】

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】

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。

【００４４】

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開始されると、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。

【００４５】

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）。

【００４６】

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（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。

【００４７】

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ＰＴ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ＰＴを参照する（ステップ＃３）。

【００４８】

対象語句が原言語フレーズに該当する場合（ステップ＃４，ＹＥＳ）、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，ＹＥＳ）、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。

【００４９】

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、ＮＯ）、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、上述の保留語句とする。

【００５０】

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，ＹＥＳ）、新たなユニットが入力されない場合（ステップ＃５、ＮＯ）、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）。

【００５１】

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。

【００５２】

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。

【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。

【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。

【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。

【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ため、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】
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）。
【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。

【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】

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。

【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＞

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。

【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。

【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順次行い、その翻訳結果を順次出力する。

【００８４】

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。

【００８５】

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言える。

【００８６】

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。

【００８７】

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。

【００８８】

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。

【００８９】

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。

【００９０】

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。

【００９１】

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。

【００９２】

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。

【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。

【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ＢＬＥＵスコアと、翻訳速度を示す遅延時間と、を用いて翻訳性能を表す。ＢＬＥＵスコアは、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、ＢＬＥＵス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。

【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、英語の旅行対話文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、日本語の旅行対話文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）の旅行対話文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ＢＬＥＵスコアとしている。

【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ＢＬＥＵスコアとしている。

【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】

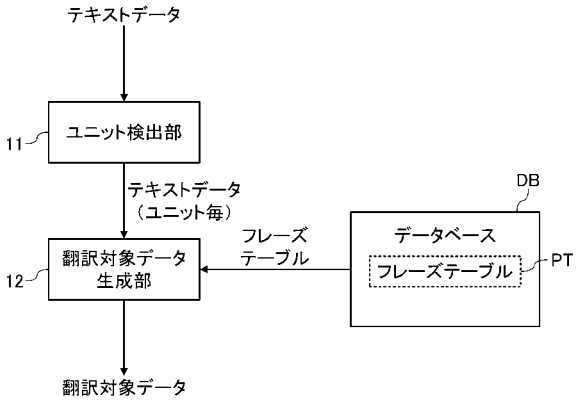
本発明は、テキストデータを分割するテキストデータ分割装置、テキストデータ分割方法及びテキストデータ分割プログラムや、当該テキストデータ分割装置を用いた翻訳装置に利用可能である。特に、本発明は、入力された音声をリアルタイムで翻訳して出力する翻訳装置や、当該翻訳装置に用いられるテキストデータ分割装置、テキストデータ分割方法及びテキストデータ分割プログラムに、好適に利用可能である。

【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
５０　：　翻訳結果出力部
ＤＢ　：　データベース
ＰＴ　：　フレーズテーブル
ＬＭ　：　言語モデル

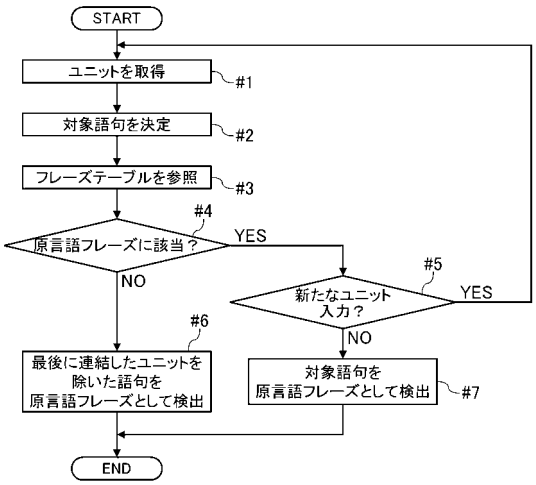
【図１】



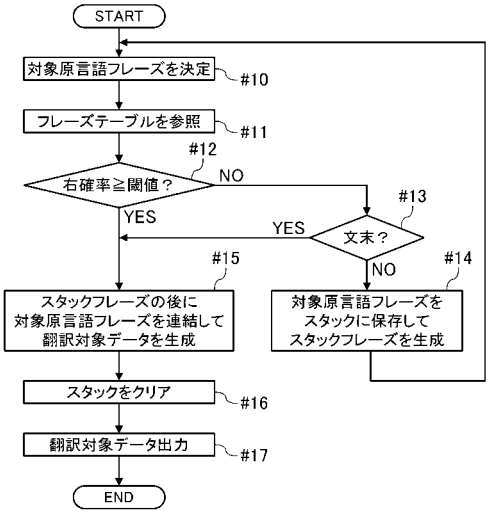
【図２】

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

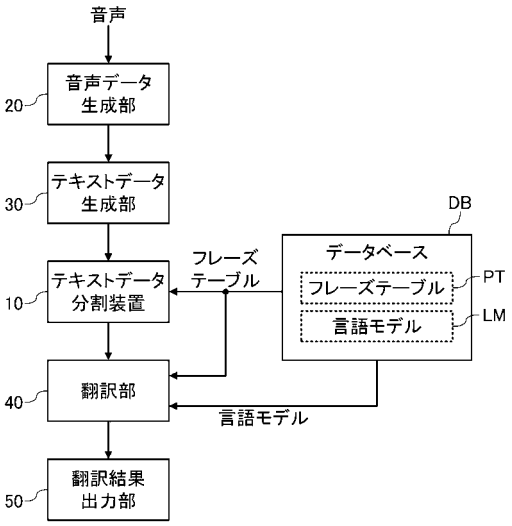
【図３】



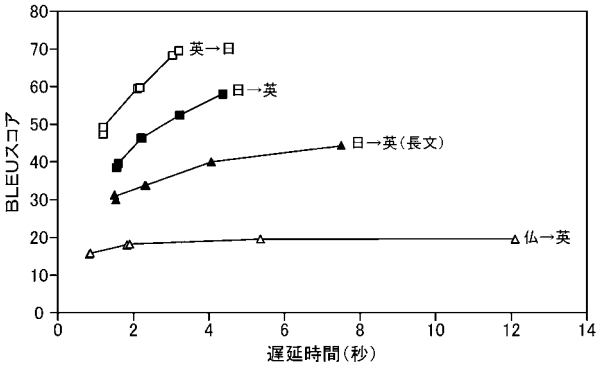
【図４】



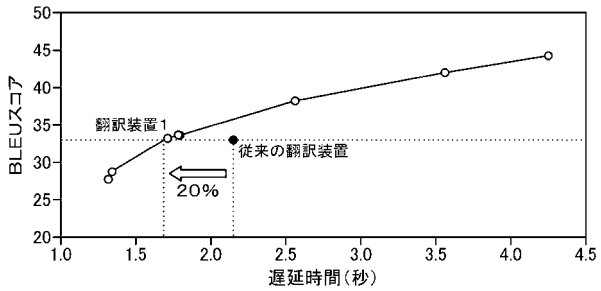
【図５】


音声
音声データ生成部
20
テキストデータ生成部
30
テキストデータ分割装置
10
翻訳部
40
翻訳結果出力部
50
フレーズテーブル
データベース
DB
フレーズテーブル
PT
言語モデル
LM
言語モデル
1

【図６】


BLEUスコア
遅延時間(秒)
英→日
日→英
日→英(長文)
仏→英

【図７】


BLEUスコア
遅延時間(秒)
翻訳装置1
従来の翻訳装置
20%

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

Ｆターム(参考) 5B091 AA03　AB20　BA04　BA12　CA01　CA21　CC15