(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7663171号
(P7663171)

(45)発行日　令和7年4月16日(2025.4.16)

(24)登録日　令和7年4月8日(2025.4.8)

(51)国際特許分類
Ｇ０６Ｆ　40/44　(2020.01)

ＦＩ
Ｇ０６Ｆ　40/44

請求項の数　6　（全54頁）

(21)出願番号　特願2020-137323(P2020-137323)
(22)出願日　令和2年8月17日(2020.8.17)
(65)公開番号　特開2022-33437(P2022-33437A)
(43)公開日　令和4年3月2日(2022.3.2)
審査請求日　令和5年7月4日(2023.7.4)
（出願人による申告）令和元年度、総務省、グローバルコミュニケーション計画の推進－多言語音声翻訳技術の研究開発及び社会実証－産業技術力強化法第１７条の適用を受ける特許出願

(73)特許権者　301022471
国立研究開発法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100143498
弁理士　中西 健
(74)代理人　100136319
弁理士　北原 宏修
(74)代理人　100148275
弁理士　山内 聡
(74)代理人　100142745
弁理士　伊藤 世子
(72)発明者　マリ バンジャマン
東京都小金井市貫井北町４－２－１ 国立研究開発法人情報通信研究機構内
(72)発明者　藤田 篤

最終頁に続く

(54)【発明の名称】　疑似対訳データ生成用機械翻訳モデルの学習方法、疑似対訳データ取得方法、および、機械翻訳モデルの学習方法

(57)【特許請求の範囲】

【請求項１】

パラメータを設定することで学習処理を行うことができ、入力データ埋込部と機械学習処理部とを含む疑似対訳データ生成用機械翻訳モデルの学習方法であって、

生成する疑似対訳データの対象とする分野である適応先分野以外の分野の第１言語のデータである他分野第１言語データと、当該他分野第１言語データの第２言語の翻訳データである他分野第２言語データとからなる対訳データを複数含む他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）と、

前記適応先分野以外の分野の第１言語のデータを複数含む他分野単言語データ集合Ｄｓｅｔｍ（Ｌ１）と、

前記適応先分野以外の分野の第２言語のデータを複数含む他分野単言語データ集合Ｄｓｅｔｍ（Ｌ２）と、

前記適応先分野の第１言語のデータを複数含む適応先分野単言語データ集合Ｄｓｅｔｍ（Ｒ１）と、

前記適応先分野の第２言語のデータを複数含む適応先分野単言語データ集合Ｄｓｅｔｍ（Ｒ２）と、

を用いて、パラメータを設定することで学習処理を行うことができ、前記入力データ埋込部とＸＬＭ処理部とを含むＸＬＭモデルに対して、

（Ａ）前記他分野単言語データ集合Ｄｓｅｔｍ（Ｌ１）、前記他分野単言語データ集合Ｄｓｅｔｍ（Ｌ２）、前記適応先分野単言語データ集合Ｄｓｅｔｍ（Ｒ１）、および、前記

適応先分野単言語データ集合Ｄｓｅｔｍ（Ｒ２）に含まれるデータである単言語用入力データの一部をマスクしたマスク化データを入力とし、前記単言語用入力データを正解データとし、当該正解データと前記ＸＬＭモデルの出力との損失が小さくなるように学習する処理であるマスク化処理による学習処理と、
（Ｂ）前記他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）に含まれる対訳データの前記他分野第１言語データおよび前記他分野第２言語データのうちの一方のデータを入力とし、他方のデータを正解データとし、当該正解データと前記ＸＬＭモデルの出力との損失が小さくなるように学習する処理である教師ありデータによる学習処理と
を行うことで、前記ＸＬＭモデルの最適パラメータを取得し、前記最適パラメータが設定された前記ＸＬＭモデルの前記入力データ埋込部に設定されている最適パラメータを、前記疑似対訳データ生成用機械翻訳モデルの前記入力データ埋込部のパラメータの初期パラメータとして設定する初期化ステップと、
前記初期パラメータが設定されている状態の前記入力データ埋込部と、前記機械学習処理部とを含む前記疑似対訳データ生成用機械翻訳モデルに対して、
（１）正解データを入力データと同一にして前記疑似対訳データ生成用機械翻訳モデルの学習処理を行う自己符号化処理、
（２）入力データに対する前記疑似対訳データ生成用機械翻訳モデルの出力データを、再度、前記疑似対訳データ生成用機械翻訳モデルに入力し、その前記疑似対訳データ生成用機械翻訳モデルの出力が前記入力データと同一となるように前記疑似対訳データ生成用機械翻訳モデルの学習処理を行うゼロショット折り返し機械翻訳処理、
（３）他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）に含まれる第１言語のデータおよび第２言語データのいずれか一方を前記疑似対訳データ生成用機械翻訳モデルの入力とし、他方を正解データとして、前記疑似対訳データ生成用機械翻訳モデルの学習処理を行う教師あり機械翻訳処理、
の少なくとも１つを用いて、学習処理を行うことで、前記疑似対訳データ生成用機械翻訳モデルの最適パラメータを取得する最適化ステップと、
を備え、
前記疑似対訳データ生成用機械翻訳モデルは、
制御信号により指定された種別のデータを出力することができ、前記制御信号により指定された、（１）前記他分野第１言語データ、（２）前記他分野第２言語データ、（３）前記適応先分野の第１言語のデータ、および、（４）前記適応先分野の第２言語のデータのいずれかを出力するように設定される、
疑似対訳データ生成用機械翻訳モデルの学習方法。
【請求項２】
請求項１に記載の疑似対訳データ生成用機械翻訳モデルの学習方法により取得された疑似対訳データ生成用機械翻訳モデルを用いて、適応先分野の疑似対訳データを取得する疑似対訳データ取得方法であって、
前記他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）から取得した第１言語のデータに対して、前記制御信号により前記疑似対訳データ生成用機械翻訳モデルの出力が前記適応先分野の第２言語のデータとなるように設定した前記疑似対訳データ生成用機械翻訳モデルを用いた機械翻訳処理を行うことで、前記他分野第１言語データの機械翻訳結果データである前記適応先分野の第２言語疑似翻訳データを取得する第１機械翻訳ステップと、
前記他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）から取得した第２言語のデータに対して、前記制御信号により前記疑似対訳データ生成用機械翻訳モデルの出力が前記適応先分野の第１言語のデータとなるように設定した前記疑似対訳データ生成用機械翻訳モデルを用いた機械翻訳処理を行うことで、前記他分野第２言語データの機械翻訳結果データである前記適応先分野の第１言語疑似翻訳データを取得する第２機械翻訳ステップと、
前記第１機械翻訳ステップで取得された前記適応先分野の第２言語疑似翻訳データと、前記第２機械翻訳ステップで取得された前記適応先分野の第１言語疑似翻訳データと、を対応づけることで、前記適応先分野の疑似対訳データを取得する疑似対訳データ取得ステ

ップと、
を備える疑似対訳データ取得方法。
【請求項３】
請求項１に記載の疑似対訳データ生成用機械翻訳モデルの学習方法により取得された疑似対訳データ生成用機械翻訳モデルを用いて、適応先分野の疑似対訳データを取得する疑似対訳データ取得方法であって、
前記適応先分野単言語データ集合Ｄｓｅｔｍ（Ｒ１）から取得した第１言語のデータまたは前記適応先分野単言語データ集合Ｄｓｅｔｍ（Ｒ２）から取得した第２言語のデータに対して、前記制御信号により前記疑似対訳データ生成用機械翻訳モデルの出力が前記適応先分野の第２言語のデータまたは第１言語のデータとなるように設定した前記疑似対訳データ生成用機械翻訳モデルを用いた機械翻訳処理を行うことで、前記適応先分野の第１言語のデータの機械翻訳結果データである前記適応先分野の第２言語疑似翻訳データまたは前記適応先分野の第２言語のデータの機械翻訳結果データである前記適応先分野の第１言語疑似翻訳データを取得する単言語データ機械翻訳ステップと、
前記単言語データ機械翻訳ステップで、前記疑似対訳データ生成用機械翻訳モデルの入力とした前記適応先分野の第１言語のデータと、前記単言語データ機械翻訳ステップで取得された前記適応先分野の第２言語疑似翻訳データと、を対応づける、または、前記疑似対訳データ生成用機械翻訳モデルの入力とした前記適応先分野の第２言語のデータと、前記単言語データ機械翻訳ステップで取得された前記適応先分野の第１言語疑似翻訳データと、を対応づけることで、前記適応先分野の疑似対訳データを取得する疑似対訳データ取得ステップと、
を備える疑似対訳データ取得方法。
【請求項４】
前記疑似対訳データ取得ステップが取得した前記適応先分野の疑似対訳データの各文対に対して、機械翻訳処理の結果の精度を示す信頼度を取得し、取得した前記信頼度が所定の値以上である文対を含む前記疑似対訳データのみを選択して出力するフィルター処理ステップをさらに備える、
請求項２または３に記載の疑似対訳データ取得方法。
【請求項５】
パラメータを設定することで学習処理を行うことができる機械翻訳モデルであって、適応先分野の第１言語のデータに対して機械翻訳を行い第２言語のデータを取得するための前記機械翻訳モデルの学習方法であって、
請求項２から４のいずれかに記載の疑似対訳データ取得方法により取得された適応先分野の疑似翻訳データと、
前記適応先分野以外の分野の第１言語のデータである他分野第１言語データと、当該他分野第１言語データの第２言語の翻訳データである他分野第２言語データとからなる対訳データを複数含む他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）と、
を用いて、前記機械翻訳モデルに対して、
（Ａ）前記他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）に含まれる対訳データの前記他分野第１言語データを入力とし、前記他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）に含まれる対訳データの前記他分野第２言語データを正解データとし、当該正解データと前記機械翻訳モデルの出力との損失が小さくなるようにする学習処理と、
（Ｂ）前記疑似翻訳データに含まれる前記適応先分野の第１言語のデータを入力とし、前記疑似翻訳データに含まれる前記適応先分野の第２言語のデータを正解データとし、当該正解データと前記機械翻訳モデルの出力との損失が小さくなるようにする学習処理と、
を行うことで、前記機械翻訳モデルの最適パラメータを取得し、当該最適パラメータを機械翻訳モデルに設定することで、学習済み機械翻訳モデルを取得する処理であるの学習処理を行う機械翻訳モデル学習ステップ、
を備える機械翻訳モデルの学習方法。
【請求項６】

請求項１に記載の疑似対訳データ生成用機械翻訳モデルの学習方法により取得された学習済みの機械翻訳モデル、または、請求項５に記載の機械翻訳モデルの学習方法により取得された学習済みの機械翻訳モデルを用いて機械翻訳処理を行う機械翻訳装置。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、ニューラル機械翻訳の技術に関する。

【背景技術】

【０００２】

ニューラル機械翻訳（ＮＭＴ：Ｎｅｕｒａｌ　Ｍａｃｈｉｎｅ　Ｔｒａｎｓｌａｔｉｏｎ）とは、多層ニューラルネットワークを用いた機械翻訳技術である。ニューラル機械翻訳システムは、起点言語の各語および文全体をベクトルまたはテンソルに変換するニューラルネットワーク（エンコーダ）と、得られたベクトルまたはテンソルから目標言語の語の系列を生成するニューラルネットワーク（デコーダ）とで構成され、対訳データに基づいてニューラルネットワークのモデルを学習する（パラメータを最適化する）ものである。ニューラル機械翻訳システムで達成できる翻訳性能は、所与の起点言語、目標言語、対象分野についての対訳データの規模に強く依存する。実用的なニューラル機械翻訳システムの構築には大規模な対訳データが必要である（言語対や分野によるが少なくとも数十万個の対訳データが必要である）。対訳データが小規模にしか存在しない場合、ニューラル機械翻訳システムにおいて、高品質な機械翻訳処理を実現することは困難である。

【０００３】

実現したい機械翻訳の対象とする言語対（対象言語対）と同一言語対の他分野（機械翻訳の対象とする分野以外の分野）の対訳データ、あるいは他の言語対であるが同一分野の対訳データが存在する場合、それらの対訳データを用いた機械翻訳分野適用技術、多言語機械翻訳技術によって対象分野の機械翻訳を実現できる場合がある。

【０００４】

また、対訳データとは異なり、安価かつ大量に入手できる単言語データを活用する技術も生み出されてきている。

【０００５】

また、近年、他分野の対訳データや他の言語対の対訳データを用いることなく、単言語データのみから機械翻訳システムを構築する教師なし機械翻訳技術も提案されている（例えば、非特許文献１を参照）。

【先行技術文献】

【非特許文献】

【０００６】

【文献】Mikel Artetxe, Gorka Labaka, Eneko Agirre (2018). Unsupervised Statistical Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3632-3642.

【発明の概要】

【発明が解決しようとする課題】

【０００７】

しかしながら、上記の技術のいずれにおいても、対象言語対・対象分野における数千文程度の対訳データの存在を前提としており、この前提が成り立たない場合、上記の技術により実現される機械翻訳システムにおいて、高い機械翻訳の性能を達成することはできない。また、上記の技術が必要とする対象言語対・対象分野における数千文程度の対訳データを作成するためには、無視できない規模の人的・金銭的コストが必要となる。

【０００８】

このように、ニューラル機械翻訳システムを新規分野向けに適応させる際には、当該ニューラル機械翻訳システムの機械翻訳の対象とする分野（対象分野）における対訳データが一定量必要であり、そのような対訳データがない場合、機械翻訳の性能を改善すること

は困難である。
【０００９】
そこで、本発明は、上記課題に鑑み、適応先分野（機械翻訳の対象とする分野）の対訳データが一切ない場合であっても、当該適応先分野における機械翻訳を精度良く実行するための疑似対訳データを生成する疑似対訳データ生成装置、および、当該疑似対訳データ生成装置で生成された疑似対訳データを用いて適応先分野における機械翻訳を精度良く実行する機械翻訳システム、および、当該機械翻訳システムで用いられる疑似対訳データ生成用機械翻訳モデルの学習方法、疑似対訳データ取得方法、および、機械翻訳モデルの学習方法を実現することを目的とする。
【課題を解決するための手段】
【００１０】
上記課題を解決するための第１の発明は、疑似対訳データを生成するための疑似対訳データ生成用機械翻訳モデルの学習方法であって、初期化ステップと、最適化ステップと、を備える。
【００１１】
初期化ステップは、
生成する疑似対訳データの対象とする分野である適応先分野以外の分野の第１言語のデータである他分野第１言語データと、当該他分野第１言語データの第２言語の翻訳データである他分野第２言語データとからなる対訳データを複数含む他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）と、
適応先分野以外の分野の第１言語のデータを複数含む他分野単言語データ集合Ｄｓｅｔｍ（Ｌ１）と、
適応先分野以外の分野の第２言語のデータを複数含む他分野単言語データ集合Ｄｓｅｔｍ（Ｌ２）と、
適応先分野の第１言語のデータを複数含む適応先分野単言語データ集合Ｄｓｅｔｍ（Ｒ１）と、
適応先分野の第２言語のデータを複数含む適応先分野単言語データ集合Ｄｓｅｔｍ（Ｒ２）と、
を用いて、疑似対訳データ生成用機械翻訳モデルの学習処理を行い、当該学習処理を実行した後の疑似対訳データ生成用機械翻訳モデルに設定されているパラメータを初期パラメータに設定する。
【００１２】
最適化ステップは、
初期パラメータが設定されている状態の疑似対訳データ生成用機械翻訳モデルに対して、
（１）正解データを入力データと同一にして疑似対訳データ生成用機械翻訳モデルの学習処理を行う自己符号化処理、
（２）入力データに対する疑似対訳データ生成用機械翻訳モデルの出力データ（この「出力データ」は、例えば、入力データとは異なる言語のデータであり、かつ、ゼロショット（学習したことのない分野）のデータである。）を、再度、疑似対訳データ生成用機械翻訳モデルに入力し、その疑似対訳データ生成用機械翻訳モデルの出力（この「出力」は、入力データと同じ言語である。）が入力データと同一となるように疑似対訳データ生成用機械翻訳モデルの学習処理を行うゼロショット折り返し機械翻訳処理、
（３）他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）に含まれる第１言語のデータおよび第２言語データのいずれか一方を疑似対訳データ生成用機械翻訳モデルの入力とし、他方を正解データとして、疑似対訳データ生成用機械翻訳モデルの学習処理を行う教師あり機械翻訳処理、
の少なくとも１つを用いて、学習処理を行うことで、疑似対訳データ生成用機械翻訳モデルの最適パラメータを取得する。
【００１３】
この疑似対訳データ生成用機械翻訳モデルの学習方法では、

（１）精度の高い他分野の対訳データ集合（Ｄｓｅｔｐ（Ｌ１－Ｌ２）（大規模（対訳データ数が多数）であることが好ましい））と、
（２）他分野の第１言語データ集合（Ｄｓｅｔｍ（Ｌ１））と、
（３）他分野の第２言語データ集合（Ｄｓｅｔｍ（Ｌ２））と、
（４）適応先分野の第１言語データ集合（Ｄｓｅｔｍ（Ｒ１））と、
（５）適応先分野の第２言語データ集合（Ｄｓｅｔｍ（Ｒ２））と、
を用いて、疑似対訳データ生成用機械翻訳モデル（例えば、ニューラルネットワークモデル）を事前学習処理（初期化ステップによる処理）により初期化し、さらに、パラメータ最適化処理（最適化ステップによる処理）を行うことで、適応先分野（機械翻訳の対象とする分野）の対訳データが一切ない場合であっても、適応先分野の第１言語および第２言語の疑似対訳データを生成するための疑似対訳データ生成用機械翻訳モデルを学習させることができる。

【００１４】

そして、学習後の（学習済みの）疑似対訳データ生成用機械翻訳モデルにより、適応先分野（対象分野）における対訳データが存在しない場合であっても、適応先分野の疑似対訳データを生成することができる。さらに、生成された適応先分野の疑似対訳データを用いて、機械翻訳モデルを学習させ、学習済みの機械翻訳モデルにより機械翻訳処理を行うことで、適応先分野（対象分野）における対訳データが存在しない場合であっても、適応先分野における機械翻訳を精度良く行うことができる。

【００１５】

なお、事前学習処理により初期化する対象は、入力データから分散表現データを取得する部分のみとしてもよい。例えば、機械翻訳モデルが、入力データから分散表現データを取得する埋込層（例えば、入力データ埋込部により実現）と、機械翻訳用のニューラルネットワークモデル（ＭＴ用ニューラルネットワークモデル）とからなる場合、事前学習処理として、入力データから分散表現データを取得する埋込層（例えば、入力データ埋込部により実現）と、言語横断言語モデル用のニューラルネットワークモデル（ＸＬＭ用ニューラルネットワークモデル、ＸＬＭ：Ｃｒｏｓｓ－ｌｉｎｇｕａｌ ｌａｎｇｕａｇｅ ｍｏｄｅｌ）とからなる言語横断言語モデルを設定し、埋込層および当該言語横断言語モデルのパラメータ最適化処理を行う。そして、機械翻訳モデルを、当該言語横断言語モデルの最適化処理後の埋込層（例えば、入力データ埋込部により実現）と、ＭＴ用ニューラルネットワークモデルとからなるモデルとし、当該言語横断言語モデルの最適化処理後の状態を初期状態として、埋込層および機械翻訳モデルのパラメータ最適化処理を行うようにしてもよい。

【００１６】

第２の発明は、第１の発明である疑似対訳データ生成用機械翻訳モデルの学習方法により取得された疑似対訳データ生成用機械翻訳モデルを用いて、適応先分野の疑似対訳データを取得する疑似対訳データ取得方法であって、第１機械翻訳ステップと、第２機械翻訳ステップと、疑似対訳データ取得ステップと、を備える。

【００１７】

第１機械翻訳ステップは、他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）から取得した第１言語のデータに対して、出力を適応先分野の第２言語に設定して疑似対訳データ生成用機械翻訳モデルを用いた機械翻訳処理を行うことで、他分野の第１言語のデータの機械翻訳結果データである適応先分野の第２言語疑似翻訳データを取得する。

【００１８】

第２機械翻訳ステップは、他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）から取得した第２言語のデータに対して、出力を適応先分野の第１言語に設定して疑似対訳データ生成用機械翻訳モデルを用いた機械翻訳処理を行うことで、他分野の第２言語のデータの機械翻訳結果データである適応先分野の第１言語疑似翻訳データを取得する。

【００１９】

疑似対訳データ取得ステップは、第１機械翻訳ステップで取得された適応先分野の第２

言語疑似翻訳データと、第２機械翻訳ステップで取得された適応先分野の第１言語疑似翻訳データと、を対応づけることで、適応先分野の疑似対訳データを取得する。

【００２０】

これにより、この疑似対訳データ取得方法では、適応先分野（対象分野）における対訳データが存在しない場合であっても、適応先分野における疑似対訳データを取得することができる。

【００２１】

なお、第１機械翻訳ステップおよび第２機械翻訳ステップは、並列に実行されるものであってもよい。

【００２２】

第３の発明は、第１の発明である疑似対訳データ生成用機械翻訳モデルの学習方法により取得された疑似対訳データ生成用機械翻訳モデルを用いて、適応先分野の疑似対訳データを取得する疑似対訳データ取得方法であって、単言語データ機械翻訳ステップと、疑似対訳データ取得ステップと、を備える。

【００２３】

単言語データ機械翻訳ステップは、適応先分野単言語データ集合Ｄｓｅｔｍ（Ｒ１）から取得した第１言語のデータまたは適応先分野単言語データ集合Ｄｓｅｔｍ（Ｒ２）から取得した第２言語のデータに対して、出力を適応先分野の第２言語または第１言語に設定して疑似対訳データ生成用機械翻訳モデルを用いた機械翻訳処理を行うことで、適応先分野の第１言語のデータの機械翻訳結果データである適応先分野の第２言語疑似翻訳データまたは適応先分野の第２言語のデータの機械翻訳結果データである適応先分野の第１言語疑似翻訳データを取得する。

【００２４】

疑似対訳データ取得ステップは、単言語データ機械翻訳ステップにおいて、疑似対訳データ生成用機械翻訳モデルの入力とした適応先分野の第１言語のデータまたは第２言語のデータと、単言語データ機械翻訳ステップで取得された適応先分野の第２言語疑似翻訳データまたは第１言語疑似翻訳データと、を対応づける、または、疑似対訳データ生成用機械翻訳モデルの入力とした適応先分野の第２言語のデータと、単言語データ機械翻訳ステップで取得された適応先分野の第１言語疑似翻訳データと、を対応づけることで、適応先分野の疑似対訳データを取得する。

【００２５】

これにより、この疑似対訳データ取得方法では、適応先分野（対象分野）における対訳データが存在しない場合であっても、適応先分野における疑似対訳データを取得することができる。

【００２６】

第４の発明は、第２または第３の発明であって、疑似対訳データ取得ステップが取得した適応先分野の疑似対訳データの各文対に対して、機械翻訳処理の結果の精度を示す信頼度を取得し、取得した信頼度が所定の値以上である文対を含む疑似対訳データのみを選択して出力するフィルター処理ステップをさらに備える。

【００２７】

これにより、この疑似対訳データ取得方法において、所定の精度以上の疑似対訳データが取得されることが保証される。

【００２８】

なお、「信頼度」は、その値が高い程、信頼度合いが高いことを示す指標である。また、信頼度は、機械翻訳処理により取得されるデータ（処理結果データ）が、単語列やトークンのデータである場合、個々の単語、トークンごとに付与されることが好ましい。

【００２９】

第５の発明は、適応先分野の第１言語のデータに対して機械翻訳を行い第２言語のデータを取得するための機械翻訳モデルの学習方法であって、機械翻訳モデル学習ステップを備える。

【００３０】
機械翻訳モデル学習ステップは、第２から第４のいずれかの発明である疑似対訳データ取得方法により取得された適応先分野の疑似翻訳データと、適応先分野以外の分野の第１言語のデータである他分野第１言語データと、当該他分野第１言語データの第２言語の翻訳データである他分野第２言語データとからなる対訳データを複数含む他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）と、を用いて、機械翻訳モデルの学習を行う。
【００３１】
これにより、この機械翻訳モデルの学習方法では、適応先分野（対象分野）における対訳データが存在しない場合であっても、第２から第５のいずれかの発明である疑似対訳データ取得方法により取得された適応先分野の疑似翻訳データを用いて機械翻訳モデルを学習させることができる。そして、学習済みの機械翻訳モデルにより機械翻訳処理を行うことで、適応先分野（対象分野）における対訳データが存在しない場合であっても、適応先分野における機械翻訳を精度良く行うことができる。
【００３２】
第６の発明は、第１の発明である疑似対訳データ生成用機械翻訳モデルの学習方法により取得された学習済みの機械翻訳モデル、または、第５の発明である機械翻訳モデルの学習方法により取得された学習済みの機械翻訳モデルを用いて機械翻訳処理を行う機械翻訳装置である。
【００３３】
これにより、この機械翻訳装置では、第１の発明である疑似対訳データ生成用機械翻訳モデルの学習方法により取得された学習済みの機械翻訳モデル、または、第５の発明である機械翻訳モデルの学習方法により取得された学習済みの機械翻訳モデルを用いて機械翻訳処理を行うことができる。
【００３４】
なお、第１の発明である疑似対訳データ生成用機械翻訳モデルの学習方法をコンピュータに実行させるためのプログラムも本発明に含まれる。
【００３５】
また、第２から第４のいずれかの発明である疑似対訳データ取得方法をコンピュータに実行させるためのプログラムも本発明に含まれる。
【００３６】
また、第５の発明である機械翻訳モデルの学習方法をコンピュータに実行させるためのプログラムも本発明に含まれる。
【００３７】
さらに、第５の発明である機械翻訳モデルの学習方法により取得された学習済みの機械翻訳モデルを用いて機械翻訳処理を行う機械翻訳方法、当該機械翻訳方法をコンピュータに実行させるためのプログラムも本発明に含まれる。
【発明の効果】
【００３８】
本発明によれば、適応先分野（機械翻訳の対象とする分野）の対訳データが一切ない場合であっても、当該適応先分野における機械翻訳を精度良く実行するための疑似対訳データを生成する疑似対訳データ生成装置、および、当該疑似対訳データ生成装置で生成された疑似対訳データを用いて適応先分野における機械翻訳を精度良く実行する機械翻訳システム、および、当該機械翻訳システムで用いられる疑似対訳データ生成用機械翻訳モデルの学習方法、疑似対訳データ取得方法、および、機械翻訳モデルの学習方法を実現することができる。
【図面の簡単な説明】
【００３９】
【図１】第１実施形態に係る機械翻訳システム１０００の概略構成図。
【図２】第１実施形態に係る疑似対訳データ生成装置１００の概略構成図。
【図３】第１実施形態に係る疑似対訳データ生成装置１００の入力データ取得部１、入力

データ埋込部２、およびＸＬＭ処理部３の概略構成図。
【図４】第１実施形態に係る疑似対訳データ生成装置１００の入力データ取得部１、入力データ埋込部２、および機械翻訳処理部５の概略構成図。
【図５】機械翻訳システム１０００で実行される処理のフローチャート。
【図６】疑似対訳データ生成装置１００で実行される処理のフローチャート。
【図７】疑似対訳データ生成装置１００で実行される処理を説明するための図。
【図８】疑似対訳データ生成装置１００で実行される処理を説明するための図。
【図９】疑似対訳データ生成装置１００で実行される処理を説明するための図。
【図１０】疑似対訳データ生成装置１００で実行される処理を説明するための図。
【図１１】疑似対訳データ生成装置１００で実行される処理を説明するための図。
【図１２】疑似対訳データ生成装置１００で実行される処理を説明するための図。
【図１３】疑似対訳データ生成装置１００で実行される処理を説明するための図。
【図１４】疑似対訳データ生成装置１００で実行される処理を説明するための図。
【図１５】機械翻訳装置ＭＴ１により取得されたデータの一例を示す図。
【図１６】ＣＰＵバス構成を示す図。
【発明を実施するための形態】
【００４０】
［第１実施形態］
第１実施形態について、図面を参照しながら、以下説明する。
【００４１】
＜１．１：機械翻訳システムの構成＞
図１は、第１実施形態に係る機械翻訳システム１０００の概略構成図である。
【００４２】
図２は、第１実施形態に係る疑似対訳データ生成装置１００の概略構成図である。
【００４３】
図３は、第１実施形態に係る疑似対訳データ生成装置１００の入力データ取得部１、入力データ埋込部２、およびＸＬＭ処理部３の概略構成図である。なお、図３では、第１セレクタＳＥＬ１を省略して図示している。
【００４４】
図４は、第１実施形態に係る疑似対訳データ生成装置１００の入力データ取得部１、入力データ埋込部２、および機械翻訳処理部５の概略構成図である。なお、図４では、第１セレクタＳＥＬ１を省略して図示している。
【００４５】
機械翻訳システム１０００は、図１に示すように、対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）と、第１単言語データ記憶部ＤＢｍ（Ｌ１）と、第２単言語データ記憶部ＤＢｍ（Ｌ２）と、第３単言語データ記憶部ＤＢｍ（Ｒ１）と、第４単言語データ記憶部ＤＢｍ（Ｒ２）と、疑似対訳データ生成装置１００と、疑似対訳データ格納部ＤＢ１と、機械翻訳装置ＭＴ１とを備える。
【００４６】
なお、以下では、疑似対訳データ生成装置１００で生成する対訳データの対象とする分野を「適応先分野」といい、当該適応先分野以外の分野を「他分野」という。
【００４７】
対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）は、適応先分野以外の分野（他分野）の第１言語のデータである他分野第１言語データと、当該他分野第１言語データの第２言語の翻訳データである他分野第２言語データとからなる対訳データを複数含む他分野対訳データ集合Ｄｓｅｔｐ（Ｌ１－Ｌ２）を記憶する。そして、対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）は、疑似対訳データ生成装置１００からの読み出し指令に基づいて、対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）に記憶している所定の対訳データＤ０（Ｌ１－Ｌ２）を読み出し、当該対訳データＤ０（Ｌ１－Ｌ２）を疑似対訳データ生成装置１００に出力する。対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）は、例えば、データベースにより実現される。なお、対

訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）に記憶保持されている対訳データの数が多いことが好ましい（対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）が大規模な対訳データを記憶保持していることが好ましい）。

【００４８】

第１単言語データ記憶部ＤＢｍ（Ｌ１）は、適応先分野以外の分野（他分野）の第１言語の単言語データである他分野第１言語データを複数含む他分野第１言語データ集合Ｄｓｅｔｍ（Ｌ１）を記憶する。第１単言語データ記憶部ＤＢｍ（Ｌ１）は、疑似対訳データ生成装置１００からの読み出し指令に基づいて、第１単言語データ記憶部ＤＢｍ（Ｌ１）に記憶している所定の第１言語の単言語データＤ０（Ｌ１）を読み出し、当該単言語データＤ０（Ｌ１）を疑似対訳データ生成装置１００に出力する。第１単言語データ記憶部ＤＢｍ（Ｌ１）は、例えば、データベースにより実現される。

【００４９】

第２単言語データ記憶部ＤＢｍ（Ｌ２）は、適応先分野以外の分野（他分野）の第２言語の単言語データである他分野第２言語データを複数含む他分野第２言語データ集合Ｄｓｅｔｍ（Ｌ２）を記憶する。第２単言語データ記憶部ＤＢｍ（Ｌ２）は、疑似対訳データ生成装置１００からの読み出し指令に基づいて、第２単言語データ記憶部ＤＢｍ（Ｌ２）に記憶している所定の第２言語の単言語データＤ０（Ｌ２）を読み出し、当該単言語データＤ０（Ｌ２）を疑似対訳データ生成装置１００に出力する。第２単言語データ記憶部ＤＢｍ（Ｌ２）は、例えば、データベースにより実現される。

【００５０】

第３単言語データ記憶部ＤＢｍ（Ｒ１）は、適応先分野の第１言語の単言語データである適応先分野第１言語データを複数含む適応先分野第１言語データ集合Ｄｓｅｔｍ（Ｒ１）を記憶する。第３単言語データ記憶部ＤＢｍ（Ｒ１）は、疑似対訳データ生成装置１００からの読み出し指令に基づいて、第３単言語データ記憶部ＤＢｍ（Ｒ１）に記憶している所定の第１言語の単言語データＤ０（Ｒ１）を読み出し、当該単言語データＤ０（Ｒ１）を疑似対訳データ生成装置１００に出力する。第３単言語データ記憶部ＤＢｍ（Ｒ１）は、例えば、データベースにより実現される。

【００５１】

第４単言語データ記憶部ＤＢｍ（Ｒ２）は、適応先分野の第２言語の単言語データである適応先分野第２言語データを複数含む適応先分野第２言語データ集合Ｄｓｅｔｍ（Ｒ２）を記憶する。第４単言語データ記憶部ＤＢｍ（Ｒ２）は、疑似対訳データ生成装置１００からの読み出し指令に基づいて、第４単言語データ記憶部ＤＢｍ（Ｒ２）に記憶している所定の第２言語の単言語データＤ０（Ｒ２）を読み出し、当該単言語データＤ０（Ｒ２）を疑似対訳データ生成装置１００に出力する。第４単言語データ記憶部ＤＢｍ（Ｒ２）は、例えば、データベースにより実現される。

【００５２】

疑似対訳データ生成装置１００は、図２に示すように、データ入力インターフェースＩＦ１と、入力データ取得部１と、第１セレクタＳＥＬ１と、入力データ埋込部２と、ＸＬＭ処理部３と、第１出力データ評価部４と、機械翻訳処理部５と、第２セレクタＳＥＬ２と、第２出力データ評価部６と、第１バッファＢｕｆ１と、第２入力データ取得処理部７と、フィルター処理部８とを備える。

【００５３】

データ入力インターフェースＩＦ１は、疑似対訳データ生成装置１００の各機能部を制御する制御部（不図示）からの制御信号ＣＴＬ１を入力し、当該制御信号ＣＴＬ１に従い、対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）、第１単言語データ記憶部ＤＢｍ（Ｌ１）、第２単言語データ記憶部ＤＢｍ（Ｌ２）、第３単言語データ記憶部ＤＢｍ（Ｒ１）、および、第４単言語データ記憶部ＤＢｍ（Ｒ２）のいずれかから、所定のデータを読み出し、読み出したデータをデータＤ１として入力データ取得部１に出力する。

【００５４】

入力データ取得部１は、図３に示すように、第１入力データ取得処理部１１と、マスク

化処理部１２と、正解データ取得部１３と、入力データ出力部１４と、を備える。入力データ取得部１は、入力されるデータの全部または一部を所定の期間、記憶保持することができるバッファ（不図示）を備えており、処理に応じて、当該バッファに記憶保持されているデータを使用することができる。

【００５５】

第１入力データ取得処理部１１は、データ入力インターフェースＩＦ１から出力されるデータＤ１と、制御部から出力される制御信号ＣＴＬ２とを入力する。第１入力データ取得処理部１１は、制御信号ＣＴＬ２に従い、データＤ１から、マスク化処理部１２および入力データ埋込部２に入力するためのデータを取得（生成）する。具体的には、第１入力データ取得処理部１１は、入力されたデータＤ１（例えば、サブワード列（または単語列）のデータ、または、各サブワード（または各単語）に対応するインデックスの列のデータ）から、（１）トークン（トークンは、文字列を表すものであってもよい）を表すデータであるトークンデータ$xi0_{token}$と、（２）当該トークンの位置を特定するための位置データxi_{pos}と、（３）当該トークンの言語を特定するための言語データxi_{lang}と、を取得する。そして、第１入力データ取得処理部１１は、上記により取得した、トークンデータ$xi0_{token}$をマスク化処理部１２に出力し、位置データxi_{pos}と、言語データxi_{lang}とを第１セレクタＳＥＬ１に出力する。

【００５６】

マスク化処理部１２は、第１入力データ取得処理部１１から出力されるトークンデータ$xi0_{token}$と、制御部から出力される制御信号ＣＴＬ２とを入力する。（１）制御信号ＣＴＬ２がマスク化処理の実行を指示している場合、マスク化処理部１２は、トークンデータ$xi0_{token}$に対してマスク化処理を実行し、マスク化処理後のデータをトークンデータxi_{token}として、第１セレクタＳＥＬ１に出力する。（２）制御信号ＣＴＬ２がマスク化処理の実行を指示していない場合、マスク化処理部１２は、入力したトークンデータ$xi0_{token}$をトークンデータxi_{token}として、第１セレクタＳＥＬ１に出力する。

【００５７】

なお、入力データ取得部１から第１セレクタＳＥＬ１に出力される（１）トークンデータxi_{token}と、（２）位置データxi_{pos}と、（３）言語データxi_{lang}と、を含むデータをデータＤ２ａと表記する。

【００５８】

正解データ取得部１３は、データ入力インターフェースＩＦ１から出力されるデータＤ１と、制御部から出力される制御信号ＣＴＬ２とを入力する。正解データ取得部１３は、制御信号ＣＴＬ２に従い、データＤ１から、言語横断言語モデル（ＸＬＭ）（入力データ埋込部２（埋込層に対応）およびＸＬＭ処理部３のＸＬＭ用ニューラルネットワークモデル３１からなるモデル）、または、機械翻訳モデル（入力データ埋込部２および機械翻訳処理部５により構成されるニューラルネットワークモデル）の学習処理に用いる正解データＤ＿ｃｏｒｒｅｃｔを生成し、当該正解データＤ＿ｃｏｒｒｅｃｔを第１出力データ評価部４および第２出力データ評価部６に出力する。

【００５９】

入力データ出力部１４は、データ入力インターフェースＩＦ１から出力されるデータＤ１と、制御部から出力される制御信号ＣＴＬ２とを入力する。入力データ出力部１４は、制御信号ＣＴＬ２に従い、入力されたデータＤ１を、データＤ１＿ｏｒｇとしてフィルター処理部８に出力する。

【００６０】

第１セレクタＳＥＬ１は、入力データ取得部１の第１入力データ取得処理部１１から出力されるデータＤ２ａと、第２入力データ取得処理部７から出力されるデータＤ２ｂと、制御部から出力される選択信号ｓｅｌ１とを入力する。第１セレクタＳＥＬ１は、選択信号ｓｅｌ１に従い、データＤ２ａ、または、データＤ２ｂを選択して、データＤ３として入力データ埋込部２に出力する。

【００６１】
入力データ埋込部２は、図３、図４に示すように、トークン埋込部２１と、位置埋込部２２と、言語埋込部２３とを備える。
【００６２】
トークン埋込部２１は、データＤ３に含まれるトークンデータxi_{token}を入力し、入力したトークンデータxi_{token}の分散表現データを取得し、取得した分散表現データを分散表現データxi'_{token}として、ＸＬＭ処理部３および機械翻訳処理部５に出力する。なお、トークン埋込部２１は、例えば、トークンデータxi_{token}に対して、分散表現データを取得するための行列による行列演算を行うことで、分散表現データxi'_{token}を取得する。例えば、トークン埋込部２１は、下記の行列演算による処理を行うことで、分散表現データxi'_{token}を取得する。なお、行列W_{token}の各要素（重み付け係数に相当）は、パラメータθ_{emb}の一部である。パラメータθ_{emb}は、ＸＬＭ処理部３または機械翻訳処理部５から入力データ埋込部２に入力されるパラメータ更新データ$update(\theta_{emb})$により更新される。

$$xi'_{token}=xi_{token}\cdot W_{token}$$

xi_{token}：各トークン（入力データ）文字列を表すベクトル（例えば、１×ｎ１の行列（ｎ１次元ベクトル）（ｎ１：自然数））

W_{token}：分散表現データを取得するための行列（例えば、ｎ１×ｍ１の行列（ｎ１，ｍ１：自然数））

xi'_{token}：入力データxi_{token}の分散表現データ（例えば、１×ｍ１の行列（ｍ１次元ベクトル）（ｍ１：自然数））

位置埋込部２２は、データＤ３に含まれる位置データxi_{pos}を入力し、入力した位置データxi_{pos}の分散表現データを取得し、取得した分散表現データを分散表現データxi'_{pos}として、ＸＬＭ処理部３および機械翻訳処理部５に出力する。なお、位置埋込部２２は、例えば、位置データxi_{pos}に対して、分散表現データを取得するための行列による行列演算を行うことで、分散表現データxi'_{pos}を取得する。例えば、位置埋込部２２は、下記の行列演算による処理を行うことで、分散表現データxi'_{pos}を取得する。なお、行列W_{pos}の各要素（重み付け係数に相当）は、パラメータθ_{emb}の一部である。パラメータθ_{emb}は、ＸＬＭ処理部３または機械翻訳処理部５から入力データ埋込部２に入力されるパラメータ更新データ$update(\theta_{emb})$により更新される。

$$xi'_{pos}=xi_{pos}\cdot W_{pos}$$

xi_{pos}：各トークン（入力データ）の位置を表すベクトル（例えば、１×ｎ２の行列（ｎ２次元ベクトル）（ｎ２：自然数））

W_{pos}：分散表現データを取得するための行列（例えば、ｎ２×ｍ２の行列（ｎ２，ｍ２：自然数））

xi'_{pos}：入力データxi_{pos}の分散表現データ（例えば、１×ｍ２の行列（ｍ２次元ベクトル）（ｍ２：自然数））

言語埋込部２３は、データＤ３に含まれる言語データxi_{lang}を入力し、入力した言語データxi_{lang}の分散表現データを取得し、取得した分散表現データを分散表現データxi'_{lang}として、ＸＬＭ処理部３および機械翻訳処理部５に出力する。なお、言語埋込部２３は、例えば、言語データxi_{lang}に対して、分散表現データを取得するための行列による行列演算を行うことで、分散表現データxi'_{lang}を取得する。例えば、言語埋込部２３は、下記の行列演算による処理を行うことで、分散表現データxi'_{lang}を取得する。なお、行列W_{lang}の各要素（重み付け係数に相当）は、パラメータθ_{emb}の一部である。パラメータθ_{emb}は、ＸＬＭ処理部３または機械翻訳処理部５から入力データ埋込部２に入力されるパラメータ更新データ$update(\theta_{emb})$により更新される。

$$xi'_{lang}=xi_{lang}\cdot W_{lang}$$

xi_{lang}：各トークン（入力データ）の言語を表すベクトル（例えば、１×ｎ３の行列（ｎ３次元ベクトル）（ｎ３：自然数））

Ｗｌａｎｇ：分散表現データを取得するための行列（例えば、ｎ３×ｍ３の行列（ｎ３，ｍ３：自然数））
ｘｉ’ｌａｎｇ：入力データｘｉｌａｎｇの分散表現データ（例えば、１×ｍ３の行列（ｍ３次元ベクトル）（ｍ３：自然数））
入力データ埋込部２は、上記により取得された分散表現データをデータＤ４として、ＸＬＭ処理部３および機械翻訳処理部５に出力する。

【００６３】
ＸＬＭ処理部３は、図３に示すように、ＸＬＭ用ニューラルネットワークモデル３１を備える。

【００６４】
ＸＬＭ用ニューラルネットワークモデル３１は、例えば、下記文献Ａに開示されているニューラルネットワークモデルであり、例えば、下記文献Ｂに開示されているトランスフォーマーモデルのアーキテクチャを採用したニューラルネットワークモデルにより実現される。
（文献Ａ）：Alexis Conneau and Guillaume Lample (2019). Cross-Lingual Language Model Pretraining. In Proceedings of the 32nd Neural Information Processing Systems Conference (NeurIPS), pp. 7057-7067.
（文献Ｂ）：Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). Attention is All You Need. In Proceedings of the 30th Neural Information Processing Systems Conference (NeurIPS), pp. 5998-6008.
ＸＬＭ用ニューラルネットワークモデル３１は、入力データ埋込部２から出力される分散表現データＤ４（＝｛ｘｉ’ｔｏｋｅｎ，ｘｉ’ｐｏｓ，ｘｉ’ｌａｎｇ｝）を入力する。

【００６５】
ＸＬＭ用ニューラルネットワークモデル３１は、第１出力データ評価部４から出力されるパラメータ更新データｕｐｄａｔｅ（θ_{XLM}）に基づいて、ＸＬＭ用ニューラルネットワークモデル３１のパラメータを更新する。また、ＸＬＭ用ニューラルネットワークモデル３１は、ＸＬＭ用ニューラルネットワークモデルのパラメータを更新した後、埋込層（入力データ埋込部２のトークン埋込部２１、位置埋込部２２、および、言語埋込部２３に対応）のパラメータを更新するためのパラメータ更新データｕｐｄａｔｅ（θ_{emb}）を生成し、当該パラメータ更新データｕｐｄａｔｅ（θ_{emb}）を入力データ埋込部２に出力する。

【００６６】
また、ＸＬＭ用ニューラルネットワークモデル３１は、入力データ埋込部２から出力されるデータＤ４に対して、ＸＬＭ処理を実行し、データＤ５ｘを取得する。そして、ＸＬＭ処理部３は、上記により取得したデータＤ５ｘを第１出力データ評価部４に出力する。

【００６７】
第１出力データ評価部４は、図２に示すように、ＸＬＭ処理部３から出力されるデータＤ５ｘと、入力データ取得部１から出力される正解データＤ＿ｃｏｒｒｅｃｔとを入力する。第１出力データ評価部４は、損失を評価するために、ＸＬＭ処理部３の出力であるデータＤ５ｘに対する正解データＤ＿ｃｏｒｒｅｃｔから損失評価用のデータＤ＿ｃｏｒｒｅｃｔ’を取得し、データＤ５ｘとデータＤ＿ｃｏｒｒｅｃｔ’とから損失を取得する（詳細については後述）。そして、第１出力データ評価部４は、所定の学習データに対する損失（学習損失）に基づいてＸＬＭ処理部３のＸＬＭ用ニューラルネットワークモデル３１のパラメータθ_{XLM}を更新するためのデータｕｐｄａｔｅ（θ_{XLM}）を生成し、当該データｕｐｄａｔｅ（θ_{XLM}）をＸＬＭ処理部３に出力する。

【００６８】
機械翻訳処理部５は、図４に示すように、ＭＴ用ニューラルネットワークモデル５１を備える。

【００６９】

ＭＴ用ニューラルネットワークモデル５１は、エンコーダ／デコーダ方式のニューラルネットワークモデルであり、例えば、上記文献Ｂに開示されているトランスフォーマーモデルのアーキテクチャによるエンコーダ、デコーダの構成を採用したニューラルネットワークモデルである。

【００７０】

ＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力される分散表現データＤ４（＝｛ｘｉ’ｔｏｋｅｎ，ｘｉ’ｐｏｓ，ｘｉ’ｌａｎｇ｝）と、制御部から出力される制御信号ＣＴＬ３とを入力する。ＭＴ用ニューラルネットワークモデル５１は、制御信号ＣＴＬ３で指示された言語のデータを出力する。

【００７１】

また、ＭＴ用ニューラルネットワークモデル５１は、第２出力データ評価部６から出力されるパラメータ更新データｕｐｄａｔｅ（θ_{MT}）に基づいて、ＭＴ用ニューラルネットワークモデル５１のパラメータを更新する。

【００７２】

また、ＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力されるデータＤ４に対して、機械翻訳処理を実行し、データＤ５を取得する。なお、ＭＴ用ニューラルネットワークモデル５１は、制御信号ＣＴＬ３によって指定された種別のデータがデータＤ５として出力されるように、機械翻訳処理を実行する。なお、制御信号ＣＴＬ３によって指定される種別は、（１）Ｌ１（他分野、第１言語）、（２）Ｌ２（他分野、第２言語）、（３）Ｒ１（適応先分野、第１言語）、および、（４）Ｒ２（適応先分野、第２言語）のいずれかである。

【００７３】

機械翻訳処理部５は、ＭＴ用ニューラルネットワークモデル５１により取得したデータＤ５を第２セレクタＳＥＬ２に出力する。

【００７４】

なお、機械翻訳処理部５は、疑似対訳データを取得する場合、入力データ埋込部２に入力されるデータ（例えば、他分野第１言語のデータＤ３（Ｌ１））に対して機械翻訳処理を行い取得したデータ（例えば、Ｒ２のデータ）と、入力データ埋込部２に入力されるデータ（他分野第１言語のデータＤ３（Ｌ１）の他分野第２言語の対訳文に相当するデータＤ３（Ｌ２））に対して機械翻訳処理を行い取得したデータ（例えば、Ｒ１のデータ）とをペアにして、データＤ５（例えば、データＤ５（Ｒ１－Ｒ２）（適応先分野第１言語の機械翻訳文（入力されたＬ２のデータの機械翻訳結果）と、適応先分野の第２言語の機械翻訳文（入力されたＬ１のデータの機械翻訳結果）とをペアとして構成する対訳データＤ５（Ｒ１－Ｒ２））として、第２セレクタＳＥＬ２に出力する。

【００７５】

第２セレクタＳＥＬ２は、１入力２出力の切替器であり、機械翻訳処理部５から出力されるデータＤ５と、制御部から出力される選択信号ｓｅｌ２とを入力する。第２セレクタＳＥＬ２は、選択信号ｓｅｌ２に従い、データＤ５を、（１）データＤ６ａとして第２出力データ評価部６およびフィルター処理部８に出力する、または、（２）データＤ６ｂとして第１バッファＢｕｆ１に出力する。

【００７６】

第２出力データ評価部６は、図２に示すように、第２セレクタＳＥＬ２から出力されるデータＤ６ａと、入力データ取得部１から出力される正解データＤ＿ｃｏｒｒｅｃｔとを入力する。第２出力データ評価部６は、損失を評価するために、機械翻訳処理部５の出力であるデータＤ６ａの正解データＤ＿ｃｏｒｒｅｃｔから損失評価用のデータＤ＿ｃｏｒｒｅｃｔ’を取得し、データＤ６ａとデータＤ＿ｃｏｒｒｅｃｔ’とから損失を取得する（詳細については後述）。そして、第２出力データ評価部６は、所定の学習データに対する損失（学習損失）に基づいて機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１のパラメータθ_{MT}を更新するためのデータｕｐｄａｔｅ（θ_{MT}）を生成する。

【００７７】

そして、第２出力データ評価部６は、データｕｐｄａｔｅ（θ_{MT}）を機械翻訳処理部５に出力する。

【００７８】

第１バッファＢｕｆ１は、第２セレクタＳＥＬ２から出力されるデータＤ６ｂを記憶保持するためのバッファである。第１バッファＢｕｆ１は、記憶保持しているデータをデータＤ６ｂ’として第２入力データ取得処理部７に出力する。

【００７９】

第２入力データ取得処理部７は、第１入力データ取得処理部１１と同様の機能部であり、第１バッファＢｕｆ１から出力されるデータＤ６ｂ’を入力し、データＤ６ｂ’から入力データ埋込部２に入力するためのデータを取得（生成）する。そして、第２入力データ取得処理部７は、取得したデータをデータＤ２ｂとして、第１セレクタＳＥＬ１に出力する。

【００８０】

フィルター処理部８は、第２セレクタＳＥＬ２から出力されるデータＤ６ａと、入力データ取得部１から出力されるデータＤ１＿ｏｒｇと、制御部から出力される制御信号ＣＴＬ４とを入力する。

【００８１】

また、フィルター処理部８は、（１）データＤ６ａが対訳データ（疑似対訳データ）である場合、各対訳データに信頼度を付与し、当該信頼度によるフィルタリング処理を行い、（２）データＤ６ａが機械翻訳データである場合、当該機械翻訳データとデータＤ１＿ｏｒｇとをペアリングして取得した対訳データ（疑似対訳データ）について、信頼度を付与し、当該信頼度によるフィルタリング処理を行う。

【００８２】

そして、フィルター処理部８は、フィルター処理後の対訳データをデータＤｐｓｄ１として、疑似対訳データ格納部ＤＢ１に出力する。

【００８３】

疑似対訳データ格納部ＤＢ１は、図１に示すように、疑似対訳データ生成装置１００から出力されるデータＤｐｓｄ１を記憶保持するための記憶部である。また、疑似対訳データ格納部ＤＢ１は、機械翻訳装置ＭＴ１から読み出し指令に基づいて、記憶保持しているデータを読み出し、読み出したデータをデータＤｐｓｄ２として機械翻訳装置ＭＴ１に出力する。

【００８４】

機械翻訳装置ＭＴ１は、図１に示すように、対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）および／または疑似対訳データ格納部ＤＢ１から、データを読み出し、読み出した当該データを用いて、機械翻訳モデルの学習処理を行う。そして、機械翻訳装置ＭＴ１は、当該学習処理により取得した機械翻訳モデルの学習済みモデルを用いて、機械翻訳処理を行う。つまり、機械翻訳装置ＭＴ１は、起点言語データＤｉｎ＿ｅを入力し、当該起点言語データＤｉｎ＿ｅに対して機械翻訳処理を行い、目標言語データＤｏｕｔ＿ｊを取得する。

【００８５】

＜１．２：機械翻訳システムの動作＞

以上のように構成された機械翻訳システム１０００の動作について、説明する。

【００８６】

図５は、機械翻訳システム１０００で実行される処理のフローチャートである。

【００８７】

図６は、疑似対訳データ生成装置１００で実行される処理（事前学習処理）のフローチャートである。

【００８８】

図７～図１１は、疑似対訳データ生成装置１００で実行される処理（事前学習処理）を説明するための図である。

【００８９】

図１２～図１４は、疑似対訳データ生成装置１００で実行される処理（疑似対訳データ

生成処理）を説明するための図である。
【００９０】
以下では、図面を参照しながら、機械翻訳システム１０００の動作について、説明する。
【００９１】
（１．２．１：事前学習処理）
まず、機械翻訳システム１０００の疑似対訳データ生成装置１００において実行される事前学習処理（疑似対訳データ生成用ＮＭＴモデル（入力データ埋込部２（埋込層に相当）と機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１とにより実現されるモデル）の事前学習処理）（図５のフローチャートのステップＳ１）について、ＸＬＭの学習処理と、疑似対訳データ生成用ＮＭＴモデルの学習処理とに分けて説明する。
【００９２】
（１．２．１．１：ＸＬＭの学習処理）
まず、ＸＬＭの学習処理について説明する。
【００９３】
ＸＬＭの学習処理は、図６に示すように、
（Ａ）マスク化言語モデル（ＭＬＭ）の処理（ステップＳ１１１）と、
（Ｂ）翻訳言語モデル（ＴＬＭ）の処理（ステップＳ１１２）と、
（Ｃ）損失の計算処理（ステップＳ１１３）と、
（Ｄ）パラメータ（θ_{XLM}，θ_{emb}）の更新処理（ステップＳ１１４）と、
から構成される。
【００９４】
なお、図６のフローチャートに示すように、ステップＳ１１１～ステップＳ１１４の処理は、終了条件が満たされるまで繰り返し実行される（図６のループ処理（ループ１）Ｓ１１０～Ｓ１１５）。
【００９５】
（１．２．１．１Ａ：マスク化言語モデル（ＭＬＭ）の処理）
マスク化言語モデル（ＭＬＭ）の処理において、疑似対訳データ生成装置１００は、図７に示すように、入力データ埋込部２への入力をマスク化データとし、正解データを原データ（マスク化していないデータ）として、学習処理を行う。
【００９６】
≪Ｍ１：ＭＬＭの処理（Ｌ１．ｍａｓｋ→Ｌ１）≫
具体的には、データ入力インターフェースＩＦ１は、第１単言語データ記憶部ＤＢｍ（Ｌ１）から他分野第１言語の単言語データＤ０（Ｌ１）を読み出し、読み出した単言語データをデータＤ１（Ｌ１）として、入力データ取得部１に出力する。
【００９７】
入力データ取得部１の第１入力データ取得処理部１１は、データＤ１（他分野第１言語のデータＤ１（Ｌ１））から、（１）トークンデータ$xi0_{token}$（＝$xi0_{token}$（Ｌ１））と、（２）当該トークンの位置を特定するための位置データxi_{pos}（＝xi_{pos}（Ｌ１））と、（３）当該トークンの言語を特定するための言語データxi_{lang}（＝xi_{lang}（Ｌ１））と、を取得する。
【００９８】
そして、入力データ取得部１は、上記のように取得したトークンデータ$xi0_{token}$をマスク化処理部１２に出力する。
【００９９】
制御部（不図示）は、入力データ取得部１のマスク化処理部１２に対して、マスク化処理を実行することを指示する制御信号ＣＴＬ２を生成し、当該制御信号ＣＴＬ２をマスク化処理部１２に出力する。
【０１００】
マスク化処理部１２は、制御信号ＣＴＬ２に従い、第１入力データ取得処理部１１から入力したトークンデータ$xi0_{token}$の一部のトークンをマスクトークン（例えば、ラ

ベル「[ｍａｓｋ]」を付与したデータ（トークン）に設定する）に置換するマスク化処理を行う。そして、マスク化処理部１２は、当該マスク化処理後のトークンデータをトークンデータｘｉ$_{token}$として、第１セレクタＳＥＬ１に出力する。

【０１０１】

また、第１入力データ取得処理部１１は、位置データｘｉ$_{pos}$（＝ｘｉ$_{pos}$（Ｌ１））を第１セレクタＳＥＬ１に出力し、言語データｘｉ$_{lang}$（＝ｘｉ$_{lang}$（Ｌ１））を第１セレクタＳＥＬ１に出力する。

【０１０２】

つまり、入力データ取得部１は、上記により取得したデータをデータＤ２ａ（＝Ｄ２ａ（Ｌ１．ｍａｓｋ）＝{ｘｉ$_{token}$（Ｌ１），ｘｉ$_{pos}$（Ｌ１），ｘｉ$_{lang}$（Ｌ１）}）として第１セレクタＳＥＬ１に出力する。

【０１０３】

なお、データを示す変数の末尾に括弧書きを追加し、そのデータの種別を表すものとする。例えば、ｘｉ$_{token}$（Ｌ１）は、他分野第１言語（Ｌ１）のデータから導出されたデータであることを示している（以下、同様）。また、上記括弧の中の「．ｍａｓｋ」という表記は、マスク化処理されたトークンデータを含むデータであることを示しており、例えば、「Ｌ１．ｍａｓｋ」は、他分野第１言語（Ｌ１）のデータから導出されたデータにおけるトークンデータに対してマスク化処理を行ったデータを含むデータであることを示している（以下、同様）。

【０１０４】

正解データ取得部１３は、制御信号ＣＴＬ２に従い、データＤ１（＝Ｄ１（Ｌ１））から、入力データ埋込部２およびＸＬＭ処理部３での学習処理（ＸＬＭの学習処理）に用いる正解データＤ＿ｃｏｒｒｅｃｔ（＝Ｄ＿ｃｏｒｒｅｃｔ（Ｌ１））を生成する。そして、正解データ取得部１３は、生成したデータを正解データＤ＿ｃｏｒｒｅｃｔ（＝Ｄ＿ｃｏｒｒｅｃｔ（Ｌ１））として第１出力データ評価部４に出力する。

【０１０５】

制御部は、第１セレクタＳＥＬ１の端子「０」を選択する選択信号ｓｅｌ１を生成し、当該選択信号ｓｅｌ１を第１セレクタＳＥＬ１に出力する。

【０１０６】

第１セレクタＳＥＬ１は、選択信号ｓｅｌ１に従い、端子「０」を選択し、入力データ取得部１から出力されるデータＤ２ａ（Ｌ１．ｍａｓｋ）を、データＤ３（＝Ｄ３（Ｌ１．ｍａｓｋ））として入力データ埋込部２に出力する。

【０１０７】

入力データ埋込部２のトークン埋込部２１は、データＤ３（Ｌ１．ｍａｓｋ）に含まれるトークンデータｘｉ$_{token}$を入力し、入力したトークンデータｘｉ$_{token}$の分散表現データを、例えば、下記数式に相当する処理を実行して取得する。

$$xi'_{token} = xi_{token} \cdot W_{token}$$

ｘｉ$_{token}$：各トークン（入力データ）文字列を表すベクトル（例えば、１×ｎ１の行列（ｎ１次元ベクトル）（ｎ１：自然数））

Ｗ$_{token}$：分散表現データを取得するための行列（例えば、ｎ１×ｍ１の行列（ｎ１，ｍ１：自然数））

ｘｉ'$_{token}$：入力データｘｉ$_{token}$の分散表現データ（例えば、１×ｍ１の行列（ｍ１次元ベクトル）（ｍ１：自然数））

入力データ埋込部２の位置埋込部２２は、データＤ３（Ｌ１．ｍａｓｋ）に含まれる位置データｘｉ$_{pos}$を入力し、入力した位置データｘｉ$_{pos}$の分散表現データを、例えば、下記数式に相当する処理を実行して取得する。

$$xi'_{pos} = xi_{pos} \cdot W_{pos}$$

ｘｉ$_{pos}$：各トークン（入力データ）の位置を表すベクトル（例えば、１×ｎ２の行列（ｎ２次元ベクトル）（ｎ２：自然数））

Ｗ$_{pos}$：分散表現データを取得するための行列（例えば、ｎ２×ｍ２の行列（ｎ２，

ｍ２：自然数））
　ｘｉ’$_{pos}$：入力データｘｉ$_{pos}$の分散表現データ（例えば、１×ｍ２の行列（ｍ２次元ベクトル）（ｍ２：自然数））
　入力データ埋込部２の言語埋込部２３は、データＤ３（Ｌ１．ｍａｓｋ）に含まれる言語データｘｉ$_{lang}$を入力し、入力した言語データｘｉ$_{lang}$の分散表現データを、例えば、下記数式に相当する処理を実行して取得する。
　ｘｉ’$_{lang}$＝ｘｉ$_{lang}$・Ｗ$_{lang}$
　ｘｉ$_{lang}$：各トークン（入力データ）の言語を表すベクトル（例えば、１×ｎ３の行列（ｎ３次元ベクトル）（ｎ３：自然数））
　Ｗ$_{lang}$：分散表現データを取得するための行列（例えば、ｎ３×ｍ３の行列（ｎ３，ｍ３：自然数））
　ｘｉ’$_{lang}$：入力データｘｉ$_{lang}$の分散表現データ（例えば、１×ｍ３の行列（ｍ３次元ベクトル）（ｍ３：自然数））
　入力データ埋込部２は、上記により取得された分散表現データをデータＤ４（Ｌ１．ｍａｓｋ）として、ＸＬＭ処理部３に出力する。なお、上記のトークン文字列や言語識別子の埋め込み方法、位置の埋め込み方法は、上記に限定されることなく、他の方法を用いてもよい。
【０１０８】
　ＸＬＭ処理部３のＸＬＭ用ニューラルネットワークモデル３１は、入力データ埋込部２から出力されるデータＤ４（Ｌ１．ｍａｓｋ）に対して、ＸＬＭ処理を実行し、データＤ５ｘ（＝Ｄ５ｘ（Ｌ１．ｍａｓｋ））を取得する。そして、ＸＬＭ処理部３は、上記により取得したデータＤ５ｘを第１出力データ評価部４に出力する。
【０１０９】
　第１出力データ評価部４は、ＸＬＭ処理部３から出力されるデータＤ５ｘ（Ｌ１．ｍａｓｋ）と、入力データ取得部１から出力される正解データＤ＿ｃｏｒｒｅｃｔ（＝Ｄ＿ｃｏｒｒｅｃｔ（Ｌ１））とを入力する。第１出力データ評価部４は、ＸＬＭ処理部３の出力であるデータＤ５（Ｌ１．ｍａｓｋ）と、その正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）とから損失を取得する。
【０１１０】
　≪Ｍ２：ＭＬＭの処理（Ｌ２．ｍａｓｋ→Ｌ２）≫
　次に、疑似対訳データ生成装置１００のデータ入力インターフェースＩＦ１は、第２単言語データ記憶部ＤＢｍ（Ｌ２）から他分野の第２言語の単言語データＤ０（Ｌ２）を読み出し、読み出した単言語データをデータＤ１（Ｌ２）として、入力データ取得部１に出力する。そして、上記と同様の処理を行う。つまり、入力データ埋込部２への入力データを、データＤ３（Ｌ２．ｍａｓｋ）、すなわち、
　Ｄ３（Ｌ２．ｍａｓｋ）＝｛ｘｉ$_{token}$（Ｌ２），ｘｉ$_{pos}$（Ｌ２），ｘｉ$_{lang}$（Ｌ２）｝
　ｘｉ$_{token}$（Ｌ２）：第２単言語データ記憶部ＤＢｍ（Ｌ２）から読み出した単言語データＤ１（Ｌ２）のトークンデータｘｉ０$_{token}$（Ｌ２）の一部のトークンをマスクトークンに置換するマスク化処理を行うことで取得したマスク化トークンデータ
　ｘｉ$_{pos}$（Ｌ２）：第２単言語データ記憶部ＤＢｍ（Ｌ２）から読み出した単言語データＤ１（Ｌ２）のトークンデータの位置を示すデータ（位置データ）
　ｘｉ$_{lang}$（Ｌ２）：第２単言語データ記憶部ＤＢｍ（Ｌ２）から読み出した単言語データＤ１（Ｌ２）のトークンの言語を示すデータ（言語データ）
とし、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｌ２）（＝原データ（マスク化していないデータ））として、上記と同様の処理を行う。
【０１１１】
　≪Ｍ３：ＭＬＭの処理（Ｒ１．ｍａｓｋ→Ｒ１）≫
　次に、疑似対訳データ生成装置１００のデータ入力インターフェースＩＦ１は、第３単言語データ記憶部ＤＢｍ（Ｒ１）から適応先分野の第１言語の単言語データＤ０（Ｒ１）

を読み出し、読み出した単言語データをデータＤ１（Ｒ１）として、入力データ取得部１に出力する。そして、上記と同様の処理を行う。つまり、入力データ埋込部２への入力データを、データＤ３（Ｒ１．ｍａｓｋ）、すなわち、

Ｄ３（Ｒ１．ｍａｓｋ）＝｛xi_{token}（Ｒ１），xi_{pos}（Ｒ１），xi_{lang}（Ｒ１）｝

xi_{token}（Ｒ１）：第３単言語データ記憶部ＤＢｍ（Ｒ１）から読み出した単言語データＤ１（Ｒ１）のトークンデータ$xi0_{token}$（Ｒ１）の一部のトークンをマスクトークンに置換するマスク化処理を行うことで取得したマスク化トークンデータ

xi_{pos}（Ｒ１）：第３単言語データ記憶部ＤＢｍ（Ｒ１）から読み出した単言語データＤ１（Ｒ１）のトークンデータの位置を示すデータ（位置データ）

xi_{lang}（Ｒ１）：第３単言語データ記憶部ＤＢｍ（Ｒ１）から読み出した単言語データＤ１（Ｒ１）のトークンデータの言語を示すデータ（言語データ）

とし、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ１）（＝原データ（マスク化していないデータ））として、上記と同様の処理を行う。

【０１１２】

≪Ｍ４：ＭＬＭの処理（Ｒ２．ｍａｓｋ→Ｒ２）≫

次に、疑似対訳データ生成装置１００のデータ入力インターフェースＩＦ１は、第４単言語データ記憶部ＤＢｍ（Ｒ２）から適応先分野の第２言語の単言語データＤ０（Ｒ２）を読み出し、読み出した単言語データをデータＤ１として、入力データ取得部１に出力する。そして、上記と同様の処理を行う。つまり、入力データ埋込部２への入力データを、データＤ３（Ｒ２．ｍａｓｋ）、すなわち、

Ｄ３（Ｒ２．ｍａｓｋ）＝｛xi_{token}（Ｒ２），xi_{pos}（Ｒ２），xi_{lang}（Ｒ２）｝

xi_{token}（Ｒ２）：第４単言語データ記憶部ＤＢｍ（Ｒ２）から読み出した単言語データＤ１（Ｒ２）のトークンデータ$xi0_{token}$（Ｒ２）の一部のトークンをマスクトークンに置換するマスク化処理を行うことで取得したマスク化トークンデータ

xi_{pos}（Ｒ２）：第４単言語データ記憶部ＤＢｍ（Ｒ２）から読み出した単言語データＤ１（Ｒ２）のトークンデータの位置を示すデータ（位置データ）

xi_{lang}（Ｒ２）：第４単言語データ記憶部ＤＢｍ（Ｒ２）から読み出した単言語データＤ１（Ｒ２）のトークンデータの言語を示すデータ（言語データ）

とし、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ２）（＝原データ（マスク化していないデータ））として、上記と同様の処理を行う。

【０１１３】

（１．２．１．１Ｂ：翻訳言語モデル（ＴＬＭ）の処理）

次に、翻訳言語モデル（ＴＬＭ）の処理において、疑似対訳データ生成装置１００は、図７に示すように、入力データ埋込部２への入力（対訳データ）をマスク化データとし、正解データを原データ（マスク化していないデータ）として、学習処理を行う。

【０１１４】

≪Ｔ１：ＴＬＭの処理（（Ｌ１－Ｌ２）．ｍａｓｋ→Ｌ１－Ｌ２）≫

具体的には、データ入力インターフェースＩＦ１は、対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）から他分野の対訳データＤ０（Ｌ１－Ｌ２）を読み出し、読み出した対訳データをデータＤ１（Ｌ１－Ｌ２）として、入力データ取得部１に出力する。

【０１１５】

入力データ取得部１の第１入力データ取得処理部１１は、データＤ１（対訳データＤ１（Ｌ１－Ｌ２）（他分野第１言語の文と、当該文の第２言語の対訳文をペアとして構成する対訳データＤ１（Ｌ１－Ｌ２））から、（１）トークンデータ$xi0_{token}$（＝$xi0_{token}$（Ｌ１－Ｌ２））と、（２）当該トークンの位置を特定するための位置データxi_{pos}（＝xi_{pos}（Ｌ１－Ｌ２））と、（３）当該トークンの言語を特定するための言語データxi_{lang}（＝xi_{lang}（Ｌ１－Ｌ２））と、を取得する。

【０１１６】

そして、入力データ取得部１は、上記のように取得したトークンデータｘｉ０$_{token}$（Ｌ１－Ｌ２）をマスク化処理部１２に出力する。

【０１１７】

制御部（不図示）は、入力データ取得部１のマスク化処理部１２に対して、マスク化処理を実行することを指示する制御信号ＣＴＬ２を生成し、当該制御信号ＣＴＬ２をマスク化処理部１２に出力する。

【０１１８】

マスク化処理部１２は、制御信号ＣＴＬ２に従い、第１入力データ取得処理部１１から入力したトークンデータｘｉ０$_{token}$（Ｌ１－Ｌ２）の一部のトークンをマスクトークン（例えば、文字列を人工的なトークン「［ＭＡＳＫ］」としたデータ（トークン）に設定する）に置換するマスク化処理を行う。そして、マスク化処理部１２は、当該マスク化処理後のトークンデータをトークンデータｘｉ$_{token}$（（Ｌ１－Ｌ２）．ｍａｓｋ）として、第１セレクタＳＥＬ１に出力する。なお、対訳データＬ１－Ｌ２（他分野第１言語の文と、当該文の第２言語の対訳文をペアとして構成する対訳データ）に対してマスク化処理を行い取得されるデータを「（Ｌ１－Ｌ２）．ｍａｓｋ」と表記する（以下、同様）。

【０１１９】

また、第１入力データ取得処理部１１は、位置データｘｉ$_{pos}$（＝ｘｉ$_{pos}$）（＝ｘｉ$_{pos}$（Ｌ１－Ｌ２））を第１セレクタＳＥＬ１に出力し、言語データｘｉ$_{lang}$（＝ｘｉ$_{lang}$）（＝ｘｉ$_{lang}$（Ｌ１－Ｌ２））を第１セレクタＳＥＬ１に出力する。

【０１２０】

つまり、入力データ取得部１は、上記により取得したデータをデータＤ２ａ（＝Ｄ２ａ（（Ｌ１－Ｌ２）．ｍａｓｋ）＝｛ｘｉ$_{token}$（（Ｌ１－Ｌ２）．ｍａｓｋ），ｘｉ$_{pos}$（Ｌ１－Ｌ２），ｘｉ$_{lang}$（Ｌ１－Ｌ２）｝）として第１セレクタＳＥＬ１に出力する。

【０１２１】

正解データ取得部１３は、制御信号ＣＴＬ２に従い、データＤ１から、入力データ埋込部２およびＸＬＭ処理部３での学習処理（ＸＬＭの学習処理）に用いる正解データＤ＿ｃｏｒｒｅｃｔ（＝Ｄ＿ｃｏｒｒｅｃｔ（Ｌ１－Ｌ２））を生成する。そして、正解データ取得部１３は、生成したデータを正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１－Ｌ２）として第１出力データ評価部４に出力する。

【０１２２】

制御部は、第１セレクタＳＥＬ１の端子「０」を選択する選択信号ｓｅｌ１を生成し、当該選択信号ｓｅｌ１を第１セレクタＳＥＬ１に出力する。

【０１２３】

第１セレクタＳＥＬ１は、選択信号ｓｅｌ１に従い、端子「０」を選択し、入力データ取得部１から出力されるデータＤ２ａ（（Ｌ１－Ｌ２）．ｍａｓｋ）を、データＤ３（（Ｌ１－Ｌ２）．ｍａｓｋ）として、入力データ埋込部２に出力する。

【０１２４】

入力データ埋込部２のトークン埋込部２１は、データＤ３（（Ｌ１－Ｌ２）．ｍａｓｋ）に含まれるトークンデータｘｉ$_{token}$（＝ｘｉ$_{token}$（（Ｌ１－Ｌ２）．ｍａｓｋ））を入力し、入力したトークンデータｘｉ$_{token}$の分散表現データｘｉ’$_{token}$を、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。

【０１２５】

入力データ埋込部２の位置埋込部２２は、データＤ３（（Ｌ１－Ｌ２）．ｍａｓｋ）に含まれる位置データｘｉ$_{pos}$（＝ｘｉ$_{pos}$（Ｌ１－Ｌ２））を入力し、入力した位置データｘｉ$_{pos}$の分散表現データｘｉ’$_{pos}$を、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。

【０１２６】

入力データ埋込部２の言語埋込部２３は、データＤ３（（Ｌ１－Ｌ２）．ｍａｓｋ）に含まれる言語データｘｉ$_{lang}$（＝ｘｉ$_{lang}$（Ｌ１－Ｌ２））を入力し、入力した

言語データｘｉ$_{lang}$の分散表現データｘｉ’$_{lang}$を、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。

【０１２７】

入力データ埋込部２は、上記により取得された分散表現データをデータＤ４（（Ｌ１－Ｌ２）．ｍａｓｋ）として、ＸＬＭ処理部３に出力する。

【０１２８】

ＸＬＭ処理部３のＸＬＭ用ニューラルネットワークモデル３１は、入力データ埋込部２から出力されるデータＤ４（（Ｌ１－Ｌ２）．ｍａｓｋ）に対して、ＸＬＭ処理を実行し、データＤ５ｘ（＝Ｄ５ｘ（（Ｌ１－Ｌ２）．ｍａｓｋ））を取得する。そして、ＸＬＭ処理部３は、上記により取得したデータＤ５ｘ（（Ｌ１－Ｌ２）．ｍａｓｋ）を第１出力データ評価部４に出力する。

【０１２９】

第１出力データ評価部４は、ＸＬＭ処理部３から出力されるデータＤ５ｘ（（Ｌ１－Ｌ２）．ｍａｓｋ）と、入力データ取得部１から出力される正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１－Ｌ２）とを入力する。第１出力データ評価部４は、ＸＬＭ処理部３の出力であるデータＤ５ｘ（（Ｌ１－Ｌ２）．ｍａｓｋ）と、その正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１－Ｌ２）とから損失を取得する。

【０１３０】

≪Ｔ２：ＴＬＭの処理（（Ｌ２－Ｌ１）．ｍａｓｋ→Ｌ２－Ｌ１）≫

次に、疑似対訳データ生成装置１００のデータ入力インターフェースＩＦ１は、対訳データ記憶部ＤＢｐ（Ｌ１―Ｌ２）から他分野の対訳データＤ０（Ｌ１－Ｌ２）を読み出し、読み出した対訳データの第一言語のデータと第二言語のデータを入れ替えた対訳データをデータＤ１（Ｌ２－Ｌ１）として、入力データ取得部１に出力する。

【０１３１】

入力データ取得部１の第１入力データ取得処理部１１は、データＤ１（対訳データＤ１（Ｌ２－Ｌ１）（他分野第２言語の文と、当該文の第１言語の対訳文をペアとして構成する対訳データＤ１（Ｌ２－Ｌ１））））から、（１）トークンデータｘｉ$_{token}$（＝ｘｉ$_{token}$（Ｌ２－Ｌ１））と、（２）当該トークンの位置を特定するための位置データｘｉ$_{pos}$（＝ｘｉ$_{pos}$（Ｌ２－Ｌ１））と、（３）当該トークンの言語を特定するための言語データｘｉ$_{lang}$（＝ｘｉ$_{lang}$（Ｌ２－Ｌ１））と、を取得する。

【０１３２】

そして、入力データ取得部１は、上記のように取得したトークンデータｘｉ$_{token}$（Ｌ２－Ｌ１）をトークンデータｘｉ０$_{token}$（Ｌ２－Ｌ１）としてマスク化処理部１２に出力する。

【０１３３】

制御部（不図示）は、入力データ取得部１のマスク化処理部１２に対して、マスク化処理を実行することを指示する制御信号ＣＴＬ２を生成し、当該制御信号ＣＴＬ２をマスク化処理部１２に出力する。

【０１３４】

マスク化処理部１２は、制御信号ＣＴＬ２に従い、第１入力データ取得処理部１１から入力したトークンデータｘｉ０$_{token}$（Ｌ２－Ｌ１）の一部のトークンをマスクトークン（例えば、文字列を人工的なトークン「［ＭＡＳＫ］」としたデータ（トークン）に設定する）に置換するマスク化処理を行う。そして、マスク化処理部１２は、当該マスク化処理後のトークンデータをトークンデータｘｉ$_{token}$（（Ｌ２－Ｌ１）．ｍａｓｋ）として、第１セレクタＳＥＬ１に出力する。なお、対訳データＬ２－Ｌ１（他分野第２言語の文と、当該文の第１言語の対訳文をペアとして構成する対訳データ）に対してマスク化処理を行い取得されるデータを「（Ｌ２－Ｌ１）．ｍａｓｋ」と表記する（以下、同様）。

【０１３５】

また、第１入力データ取得処理部１１は、位置データｘｉ$_{pos}$（＝ｘｉ$_{pos}$）（＝ｘｉ$_{pos}$（Ｌ２－Ｌ１））を第１セレクタＳＥＬ１に出力し、言語データｘｉ$_{lang}$（＝

ｘｉｌａｎｇ）（＝ｘｉｌａｎｇ（Ｌ２－Ｌ１））を第１セレクタＳＥＬ１に出力する。

【０１３６】

つまり、入力データ取得部１は、上記により取得したデータをデータＤ２ａ（＝Ｄ２ａ（（Ｌ２－Ｌ１）．ｍａｓｋ）＝｛ｘｉｔｏｋｅｎ（（Ｌ２－Ｌ１）．ｍａｓｋ），ｘｉｐｏｓ（Ｌ２－Ｌ１），ｘｉｌａｎｇ（Ｌ２－Ｌ１）｝）として第１セレクタＳＥＬ１に出力する。

【０１３７】

正解データ取得部１３は、制御信号ＣＴＬ２に従い、データＤ１から、入力データ埋込部２およびＸＬＭ処理部３での学習処理（ＸＬＭの学習処理）に用いる正解データＤ＿ｃｏｒｒｅｃｔ（＝Ｄ＿ｃｏｒｒｅｃｔ（Ｌ２－Ｌ１））を生成する。そして、正解データ取得部１３は、生成したデータを正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ２－Ｌ１）として第１出力データ評価部４に出力する。

【０１３８】

制御部は、第１セレクタＳＥＬ１の端子「０」を選択する選択信号ｓｅｌ１を生成し、当該選択信号ｓｅｌ１を第１セレクタＳＥＬ１に出力する。

【０１３９】

第１セレクタＳＥＬ１は、選択信号ｓｅｌ１に従い、端子「０」を選択し、入力データ取得部１から出力されるデータＤ２ａ（（Ｌ２－Ｌ１）．ｍａｓｋ）を、データＤ３（（Ｌ２－Ｌ１）．ｍａｓｋ）として、入力データ埋込部２に出力する。

【０１４０】

入力データ埋込部２のトークン埋込部２１は、データＤ３（Ｌ２－Ｌ１．ｍａｓｋ）に含まれるトークンデータｘｉｔｏｋｅｎ（＝ｘｉｔｏｋｅｎ（Ｌ２－Ｌ１））を入力し、入力したトークンデータｘｉｔｏｋｅｎの分散表現データｘｉ’ｔｏｋｅｎを、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。

【０１４１】

入力データ埋込部２の位置埋込部２２は、データＤ３（（Ｌ２－Ｌ１）．ｍａｓｋ）に含まれる位置データｘｉｐｏｓ（＝ｘｉｐｏｓ（Ｌ２－Ｌ１））を入力し、入力した位置データｘｉｐｏｓの分散表現データｘｉ’ｐｏｓを、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。

【０１４２】

入力データ埋込部２の言語埋込部２３は、データＤ３（（Ｌ２－Ｌ１）．ｍａｓｋ）に含まれる言語データｘｉｌａｎｇ（＝ｘｉｌａｎｇ（Ｌ２－Ｌ１））を入力し、入力した言語データｘｉｌａｎｇの分散表現データｘｉ’ｌａｎｇを、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。

【０１４３】

入力データ埋込部２は、上記により取得された分散表現データをデータＤ４（（Ｌ２－Ｌ１）．ｍａｓｋ）として、ＸＬＭ処理部３に出力する。

【０１４４】

ＸＬＭ処理部３のＸＬＭ用ニューラルネットワークモデル３１は、入力データ埋込部２から出力されるデータＤ４（（Ｌ２－Ｌ１）．ｍａｓｋ）に対して、ＸＬＭ処理を実行し、データＤ５ｘ（＝Ｄ５ｘ（（Ｌ２－Ｌ１）．ｍａｓｋ））を取得する。そして、ＸＬＭ処理部３は、上記により取得したデータＤ５ｘ（（Ｌ２－Ｌ１）．ｍａｓｋ）を第１出力データ評価部４に出力する。

【０１４５】

第１出力データ評価部４は、ＸＬＭ処理部３から出力されるデータＤ５ｘ（（Ｌ２－Ｌ１）．ｍａｓｋ）と、入力データ取得部１から出力される正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ２－Ｌ１）とを入力する。第１出力データ評価部４は、ＸＬＭ処理部３の出力であるデータＤ５ｘ（（Ｌ２－Ｌ１）．ｍａｓｋ）と、その正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ２－Ｌ１）とから損失を取得する。

【０１４６】

（１．２．１．１Ｃ：損失の計算処理）
上記の処理、すなわち、（Ａ）ＭＬＭの処理（４種類の単言語データを使用）（ステップＳ１１１）と、（Ｂ）ＴＬＭの処理（Ｌ１－Ｌ２の対訳データを使用）（ステップＳ１１２）とを実行した後、第１出力データ評価部４は、損失の計算処理を行う。なお、損失の計算をＭ文（Ｍ個の文、Ｍ：自然数）ごとに行う場合、Ｍ文の中の第ｉ番目の文に含まれるサブワード数をN_i（$1 \leqq i \leqq M$）とし、第ｉ番目の文についてのＸＬＭ処理部３からの出力データＤ５ｘ（X_{in}）（入力データをX_{in}で表す）のｊ番目（ｊ：自然数、$1 \leqq j \leqq N_i$）のサブワードに相当するデータをＤ５ｘ（（$X_{in} \rightarrow X_{out}$），ｉ，ｊ）（「（$X_{in} \rightarrow X_{out}$）」は、入力データが$X_{in}$であり、出力データが$X_{out}$であることを表す）とすると、第１出力データ評価部４は、下記数式のように、ＸＬＭ処理部３から出力されるデータと、正解データとから損失Ｌｏｓｓを取得する。

【数１】

$$Loss = Loss\_MLM + Loss\_TLM$$

【数２】

$$\begin{aligned} Loss\_MLM = &\sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(L1,i,j), D5x((L1.mask \rightarrow L1),i,j)) \\ &+\sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(L2,i,j), D5x((L2.mask \rightarrow L2),i,j)) \\ &+\sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(R1,i,j), D5x((R1.mask \rightarrow R1),i,j)) \\ &+\sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(R2,i,j), D5x((R2.mask \rightarrow R2),i,j)) \end{aligned}$$

【数３】

$$\begin{aligned} Loss\_TLM = &\sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(L1-L2,i,j), D5x(((L1-L2).mask \rightarrow (L1-L2)),i,j)) \\ &+\sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(L2-L1,i,j), D5x(((L2-L1).mask \rightarrow (L2-L1)),i,j)) \end{aligned}$$

【数４】

$$loss(p,q) = -\sum_{k=1}^{V} p(k)\log(q(k))$$

Ｖ：サブワード語彙のサイズ（各トークン（入力データ）文字列を表すベクトルの次元数）
ｐ：確率分布（ｐ（ｋ）は、ｋ番目（第ｋ次元）の要素の確率を示す）
ｑ：確率分布（ｑ（ｋ）は、ｋ番目（第ｋ次元）の要素の確率を示す）
なお、上記数式において、Ｄ＿ｃｏｒｒｅｃｔ’（ｘ，ｉ，ｊ）は、第１出力データ評価部４により正解データＤ＿ｃｏｒｒｅｃｔ（ｘ）の第ｉ番目の文の第ｊ番目のサブワードから取得されるデータ（ベクトル）であり、サブワード語彙のサイズ（＝各トークン（入力データ）文字列を表すベクトルの次元数（これをｎ１とする））と同じ次元数のベクトル（ｎ１次元のベクトル）である。そして、例えば、Ｄ＿ｃｏｒｒｅｃｔ’（ｘ，ｉ，ｊ）は、ｎ１次元のうち１次元のみ値が「１」であり（ｎ１次元ベクトルの要素のうち、当該正解サブワードに対応する要素のみが「１」）、それ以外は値が「０」であるｏｎｅ－ｈｏｔベクトルである。

【０１４７】
また、上記数式において、例えば、Ｄ５ｘ（（Ｌ１．ｍａｓｋ→Ｌ１），ｉ，ｊ）は、ｎ１次元の実数ベクトル（各次元のデータ（ベクトルの要素）が、対応するサブワードである確率を示す実数ベクトル）であり、ｓｏｆｔｍａｘ関数により確率化（実数ベクトルの各要素の総和が「１」となるように正規化）されている。そして、Ｄ５ｘ（（Ｌ１．ｍａｓｋ→Ｌ１），ｉ，ｊ）は、入力データをＬ１．ｍａｓｋとしたときのＸＬＭ処理部３からの出力データであり、第ｉ番目の文の第ｊ番目のサブワードに相当するデータである。

【０１４８】
また、上記数式において、ｌｏｓｓ（ｐ、ｑ）は、交差エントロピー誤差を求める数式であり、これにより、確率分布間（上記数式では、確率分布ｐ、ｑ）の相違を定量化できる。

【０１４９】
このように、第１出力データ評価部４は、上記数式に相当する処理により、損失Ｌｏｓｓを取得する。

【０１５０】
（１．２．１．１Ｄ：パラメータの更新処理）
第１出力データ評価部４は、所定の学習データに対して上記で算出した損失（学習損失）に基づいてＸＬＭ処理部３のＸＬＭ用ニューラルネットワークモデル３１のパラメータθ_{XLM}を更新するためのデータｕｐｄａｔｅ（θ_{XLM}）を生成し、当該データｕｐｄａｔｅ（θ_{XLM}）をＸＬＭ処理部３に出力する。

【０１５１】
ＸＬＭ処理部３のＸＬＭ用ニューラルネットワークモデル３１は、データｕｐｄａｔｅ（θ_{XLM}）に基づいて、パラメータθ_{XLM}を更新する。

【０１５２】
また、ＸＬＭ用ニューラルネットワークモデル３１は、ＸＬＭ用ニューラルネットワークモデル３１のパラメータを更新した後、埋込層（入力データ埋込部２のトークン埋込部２１、位置埋込部２２、および、言語埋込部２３に対応）のパラメータを更新するためのパラメータ更新データｕｐｄａｔｅ（θ_{emb}）を生成し、当該パラメータ更新データｕｐｄａｔｅ（θ_{emb}）を入力データ埋込部２に出力する。

【０１５３】
入力データ埋込部２は、パラメータ更新データｕｐｄａｔｅ（θ_{emb}）に基づいて、埋込層（入力データ埋込部２のトークン埋込部２１、位置埋込部２２、および、言語埋込部２３に対応）のパラメータを更新する。例えば、入力データ埋込部２は、パラメータ更新データｕｐｄａｔｅ（θ_{emb}）に基づいて、変換行列（例えば、行列Ｗｔｏｋｅｎ、Ｗｐｏｓ、Ｗｌａｎｇ）の要素（値）を更新する。

【０１５４】
疑似対訳データ生成装置１００は、所定の終了条件を満たすまで、上記処理（図６のループ１の処理）を繰り返し実行する。その際、例えば、学習に用いるデータとは別の調整

用データに対して第１出力データ評価部４で算出される損失（テスト損失）を評価値として参照する。

【０１５５】

そして、疑似対訳データ生成装置１００は、上記処理（図６のループ１の処理）の終了条件が満たされたときをもって、ＸＬＭの学習処理を終了させる。これにより、入力データ埋込部２の初期化が完了する。

【０１５６】

つまり、ＸＬＭの学習処理が終了した時点において、入力データ埋込部２で設定されている分散表現データを取得するためのパラメータ（例えば、変換行列W_{token}、W_{pos}、W_{lang}）を、疑似対訳データ生成用ＮＭＴモデル（入力データ埋込部２（埋込層に相当）と機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１とにより実現されるモデル）の学習処理の初期値（初期パラメータ）に設定する。

【０１５７】

つまり、疑似対訳データ生成装置１００において、ＸＬＭの学習処理が終了した時点におけるパラメータが設定されている状態（入力データ埋込部２の状態）を、疑似対訳データ生成用ＮＭＴモデルの学習処理の初期状態として、疑似対訳データ生成用ＮＭＴモデルの学習処理を開始させる。

【０１５８】

なお、図６のフローチャートのループ処理（ループ１）の終了条件は、例えば、以下のように設定される。

（１）事前に定めた反復回数だけループ処理（ループ１）が実行された。

（２）言語横断言語モデル（ＸＬＭ）の学習処理において、第１出力データ評価部４における評価値が一定以上（事前に定めた値以上）の変化を示さなかった。

（３）言語横断言語モデル（ＸＬＭ）の学習処理において、第１出力データ評価部４における評価値が事前に定めた値を下回った。

（４）言語横断言語モデル（ＸＬＭ）の学習処理において、第１出力データ評価部４における評価値が事前に定めた回数更新されなかった。

【０１５９】

上記の終了条件を満たす場合、疑似対訳データ生成装置１００は、ＸＬＭの学習処理が完了した判断し、ＸＬＭの学習処理を終了させる。

【０１６０】

そして、疑似対訳データ生成装置１００は、上記のようにして、パラメータが上記初期値に設定された状態で、以下に説明する、疑似対訳データ生成用ＮＭＴモデル（入力データ埋込部２（埋込層に相当）と機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１とにより実現されるモデル）の学習処理を行う。

【０１６１】

（１．２．１．２：疑似対訳データ生成用ＮＭＴモデルの学習処理）

次に、疑似対訳データ生成装置１００では、疑似対訳データ生成用ＮＭＴモデルの学習処理（ステップＳ１２０～Ｓ１２６）が実行される。

【０１６２】

疑似対訳データ生成用ＮＭＴモデルの学習処理は、図６に示すように、

（Ａ）自己符号化処理（ステップＳ１２１）と、

（Ｂ）ゼロショット折り返し機械翻訳処理（ステップＳ１２２）と、

（Ｃ）教師データあり機械翻訳処理（ステップＳ１２３）と、

（Ｄ）損失の計算処理（ステップＳ１２４）と、

（Ｅ）パラメータ（θ_{MT}，θ_{emb}）の更新処理（ステップＳ１２５）と、

から構成される。

【０１６３】

なお、図６のフローチャートに示すように、ステップＳ１２１～Ｓ１２５の処理は、終了条件が満たされるまで繰り返し実行される（図６のループ処理（ループ２）Ｓ１２０～

Ｓ１２６）。
【０１６４】
（１．２．１．２Ａ：自己符号化処理）
自己符号化処理において、疑似対訳データ生成装置１００は、図８に示すように、入力データ埋込部２への入力データと同一の出力データが出力されるように学習処理（疑似対訳データ生成用ＮＭＴモデルの学習処理）を行う。つまり、自己符号化処理において、正解データは、入力データと同一のデータに設定される。すなわち、自己符号化処理において、疑似対訳データ生成装置１００は、
（１）入力データ埋込部２への入力をＤ３（Ｌ１）とし、機械翻訳処理部５の出力データＤ５（Ｌ１）の正解データをＤ＿ｃｏｒｒｅｃｔ（Ｌ１）とする、
（２）入力データ埋込部２への入力をＤ３（Ｌ２）とし、機械翻訳処理部５の出力データＤ５（Ｌ２）の正解データをＤ＿ｃｏｒｒｅｃｔ（Ｌ２）とする、
（３）入力データ埋込部２への入力をＤ３（Ｒ１）とし、機械翻訳処理部５の出力データＤ５（Ｒ１）の正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ１）とする、および、
（４）入力データ埋込部２への入力をＤ３（Ｒ２）とし、機械翻訳処理部５の出力データＤ５（Ｒ２）の正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ２）として、学習処理（自己符号化処理によるパラメータ最適化処理）を実行する。
【０１６５】
≪Ａ１：自己符号化処理（Ｌ１→Ｌ１）≫
入力データ埋込部２への入力をＤ３（Ｌ１）とし、機械翻訳処理部５の出力データＤ５（Ｌ１）の正解データをＤ＿ｃｏｒｒｅｃｔ（Ｌ１）とする場合について、説明する。
【０１６６】
データ入力インターフェースＩＦ１は、第１単言語データ記憶部ＤＢｍ（Ｌ１）から他分野第１言語の単言語データＤ０（Ｌ１）を読み出し、読み出した単言語データをデータＤ１（＝Ｄ１（Ｌ１））として、入力データ取得部１に出力する。
【０１６７】
入力データ取得部１の第１入力データ取得処理部１１は、データＤ１（他分野第１言語のデータＤ１（Ｌ１））（他分野第１言語の文を構成するデータＤ１（Ｌ１））から、（１）トークンデータ$xi0_{token}$と、（２）当該トークンの位置を特定するための位置データxi_{pos}と、（３）当該トークンの言語を特定するための言語データxi_{lang}と、を取得する。
【０１６８】
そして、入力データ取得部１は、上記のように取得したトークンデータ$xi0_{token}$をマスク化処理部１２に出力する。
【０１６９】
制御部（不図示）は、入力データ取得部１のマスク化処理部１２に対して、マスク化処理を実行しないことを指示する制御信号ＣＴＬ２を生成し、当該制御信号ＣＴＬ２をマスク化処理部１２に出力する。
【０１７０】
マスク化処理部１２は、制御信号ＣＴＬ２に従い、第１入力データ取得処理部１１から入力したトークンデータ$xi0_{token}$をトークンデータxi_{token}として、第１セレクタＳＥＬ１に出力する。
【０１７１】
また、第１入力データ取得処理部１１は、位置データxi_{pos}（＝xi_{pos}（Ｌ１））を第１セレクタＳＥＬ１に出力し、言語データxi_{lang}（＝xi_{lang}）（＝xi_{lang}（Ｌ１））を第１セレクタＳＥＬ１に出力する。
【０１７２】
つまり、入力データ取得部１は、上記により取得したデータをデータＤ２ａ（＝Ｄ２ａ（Ｌ１）＝｛xi_{token}（Ｌ１），xi_{pos}（Ｌ１），xi_{lang}（Ｌ１）｝）として第１セレクタＳＥＬ１に出力する。

【０１７３】
正解データ取得部１３は、制御信号ＣＴＬ２に従い、データＤ１（＝Ｄ１（Ｌ１））から、疑似対訳データ生成用ＮＭＴモデル（入力データ埋込部２（埋込層に相当）と機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１とにより実現されるモデル）の学習処理に用いる正解データＤ＿ｃｏｒｒｅｃｔ（＝Ｄ＿ｃｏｒｒｅｃｔ（Ｌ１））を生成する。具体的には、正解データ取得部１３は、第１セレクタＳＥＬ１への入力データＤ２ａ（Ｌ１）と同一のデータを正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）に設定する。そして、正解データ取得部１３は、上記により生成した正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）を第２出力データ評価部６に出力する。

【０１７４】
制御部は、第１セレクタＳＥＬ１の端子「０」を選択する選択信号ｓｅｌ１を生成し、当該選択信号ｓｅｌ１を第１セレクタＳＥＬ１に出力する。

【０１７５】
第１セレクタＳＥＬ１は、選択信号ｓｅｌ１に従い、端子「０」を選択し、入力データ取得部１から出力されるデータＤ２ａ（Ｌ１）を、データＤ３（Ｌ１）として、入力データ埋込部２に出力する。

【０１７６】
入力データ埋込部２のトークン埋込部２１は、データＤ３（Ｌ１）に含まれるトークンデータｘｉ$_{token}$を入力し、入力したトークンデータｘｉ$_{token}$の分散表現データｘｉ’$_{token}$を、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。

【０１７７】
入力データ埋込部２の位置埋込部２２は、データＤ３（Ｌ１）に含まれる位置データｘｉ$_{pos}$を入力し、入力した位置データｘｉ$_{pos}$の分散表現データｘｉ’$_{pos}$を、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。

【０１７８】
入力データ埋込部２の言語埋込部２３は、データＤ３（Ｌ１）に含まれる言語データｘｉ$_{lang}$を入力し、入力した言語データｘｉ$_{lang}$の分散表現データｘｉ’$_{lang}$を、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。

【０１７９】
入力データ埋込部２は、上記により取得された分散表現データをデータＤ４（Ｌ１）として、機械翻訳処理部５に出力する。

【０１８０】
機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力されるデータＤ４（Ｌ１）（分散表現データＤ４（Ｌ１）（＝｛ｘｉ’$_{token}$，ｘｉ’$_{pos}$，ｘｉ’$_{lang}$｝））と、制御部から出力される制御信号ＣＴＬ３とを入力する。

【０１８１】
制御部は、ＭＴ用ニューラルネットワークモデル５１からＬ１のデータを出力することを指示する制御信号ＣＴＬ３を生成し、当該制御信号ＣＴＬ３を機械翻訳処理部５に出力する。

【０１８２】
ＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力されるデータＤ４（Ｌ１）に対して、機械翻訳処理を実行し、データＤ５（Ｌ１）を取得する。なお、ＭＴ用ニューラルネットワークモデル５１は、制御信号ＣＴＬ３に従い、出力データの種別をＬ１（他分野、第１言語）とする。

【０１８３】
ＭＴ用ニューラルネットワークモデル５１は、上記により取得したデータＤ５（Ｌ１）を第２セレクタＳＥＬ２に出力する。

【０１８４】
制御部は、第２セレクタＳＥＬ２の端子「０」を選択する選択信号ｓｅｌ２を生成し、

当該選択信号ｓｅｌ２を第２セレクタＳＥＬ２に出力する。
【０１８５】
第２セレクタＳＥＬ２は、選択信号ｓｅｌ２に従い、端子「０」を選択し、機械翻訳処理部５から出力されるデータＤ５（データＤ５（Ｌ１））を、第２出力データ評価部６にデータＤ６ａ（データＤ６ａ（Ｌ１））として出力する。
【０１８６】
第２出力データ評価部６は、第２セレクタＳＥＬ２を介して機械翻訳処理部５から出力されるデータＤ５（Ｌ１）（＝データＤ６ａ（Ｌ１））と、入力データ取得部１から出力される正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）とを入力する。第２出力データ評価部６は、機械翻訳処理部５の出力であるデータＤ５（Ｌ１）と、その正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）とから損失を取得する。
【０１８７】
≪Ａ２：自己符号化処理（Ｌ２→Ｌ２）≫
入力データ埋込部２への入力をＤ３（Ｌ２）とし、機械翻訳処理部５の出力データＤ５（Ｌ２）の正解データをＤ＿ｃｏｒｒｅｃｔ（Ｌ２）とする場合についても、疑似対訳データ生成装置１００は、上記と同様の処理を実行する。つまり、疑似対訳データ生成装置１００は、上記処理において、Ｄ３（Ｌ１）をＤ３（Ｌ２）に置換し、正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）を正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ２）に置換して、上記処理と同様の処理を行う。
【０１８８】
≪Ａ３：自己符号化処理（Ｒ１→Ｒ１）≫
入力データ埋込部２への入力をＤ３（Ｒ１）とし、機械翻訳処理部５の出力データＤ５（Ｒ１）の正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ１）とする場合についても、疑似対訳データ生成装置１００は、上記と同様の処理を実行する。つまり、疑似対訳データ生成装置１００は、上記処理において、Ｄ３（Ｌ１）をＤ３（Ｒ１）に置換し、正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）を正解データＤ＿ｃｏｒｒｅｃｔ（Ｒ１）に置換して、上記処理と同様の処理を行う。
【０１８９】
≪Ａ４：自己符号化処理（Ｒ２→Ｒ２）≫
入力データ埋込部２への入力をＤ３（Ｒ２）とし、機械翻訳処理部５の出力データＤ５（Ｒ２）の正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ２）とする場合についても、疑似対訳データ生成装置１００は、上記と同様の処理を実行する。つまり、疑似対訳データ生成装置１００は、上記処理において、Ｄ３（Ｌ１）をＤ３（Ｒ２）に置換し、正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）を正解データＤ＿ｃｏｒｒｅｃｔ（Ｒ２）に置換して、上記処理と同様の処理を行う。
【０１９０】
（１．２．１．２Ｂ：ゼロショット折り返し機械翻訳処理）
ゼロショット折り返し機械翻訳処理において、疑似対訳データ生成装置１００は、図９、図１０に示すように、（１）入力データ埋込部２への入力データに対して機械翻訳処理（疑似対訳データ生成用ＮＭＴモデル（入力データ埋込部２（埋込層に相当）と機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１とにより実現されるモデル）による機械翻訳処理）を行い（１回目の機械翻訳処理）、入力データと異なる言語のデータ（ゼロショット機械翻訳のデータ）を出力させ、（２）その出力されたデータに対して機械翻訳処理（疑似対訳データ生成用ＮＭＴモデルにより機械翻訳処理）を行い（２回目の機械翻訳処理）、入力データと同一のデータが出力されるように学習処理を行う。つまり、ゼロショット折り返し機械翻訳処理において、正解データは、入力データと同一のデータに設定される。
【０１９１】
具体的には、ゼロショット折り返し機械翻訳処理において、疑似対訳データ生成装置１００は、以下の（１）～（６）の処理を行う。

（１）Ｌ１→Ｒ２→Ｌ１：
　１回目の機械翻訳処理において、入力データ埋込部２への入力をＤ３（１）（Ｌ１）とし、機械翻訳処理部５からの出力をＤ５（１）（Ｒ２）とし、２回目の機械翻訳処理において、入力データ埋込部２への入力をＤ３（２）（Ｒ２）（＝Ｄ５（１）（Ｒ２））とし、機械翻訳処理部５からの出力をＤ５（２）（Ｌ１）とし、また、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｌ１）とする。
【０１９２】
　なお、１回目の機械翻訳処理の入力データ埋込部２への入力データをＤ３（１）（ｘ）と、入力データ埋込部２からの出力データをＤ４（１）（ｘ）と、２回目の機械翻訳処理の入力データ埋込部２への入力データをＤ３（２）（ｘ）と、入力データ埋込部２からの出力データをＤ４（２）（ｘ）と表記する（以下同様）。
（２）Ｒ１→Ｌ２→Ｒ１：
　１回目の機械翻訳処理において、入力データ埋込部２への入力をＤ３（１）（Ｒ１）とし、機械翻訳処理部５からの出力をＤ５（１）（Ｌ２）とし、２回目の機械翻訳処理において、入力データ埋込部２への入力をＤ３（２）（Ｌ２）（＝Ｄ５（１）（Ｌ２））とし、機械翻訳処理部５からの出力をＤ５（２）（Ｒ１）とし、また、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ１）とする。
（３）Ｒ１→Ｒ２→Ｒ１：
　１回目の機械翻訳処理において、入力データ埋込部２への入力をＤ３（１）（Ｒ１）とし、機械翻訳処理部５からの出力をＤ５（１）（Ｒ２）とし、２回目の機械翻訳処理において、入力データ埋込部２への入力をＤ３（２）（Ｒ２）（＝Ｄ５（１）（Ｒ２））とし、機械翻訳処理部５からの出力をＤ５（２）（Ｒ１）とし、また、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ１）とする。
（４）Ｌ２→Ｒ１→Ｌ２：
　１回目の機械翻訳処理において、入力データ埋込部２への入力をＤ３（１）（Ｌ２）とし、機械翻訳処理部５からの出力をＤ５（１）（Ｒ１）とし、２回目の機械翻訳処理において、入力データ埋込部２への入力をＤ３（２）（Ｒ１）（＝Ｄ５（１）（Ｒ１））とし、機械翻訳処理部５からの出力をＤ５（２）（Ｌ２）とし、また、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｌ２）とする。
（５）Ｒ２→Ｌ１→Ｒ２：
　１回目の機械翻訳処理において、入力データ埋込部２への入力をＤ３（１）（Ｒ２）とし、機械翻訳処理部５からの出力をＤ５（１）（Ｌ１）とし、２回目の機械翻訳処理において、入力データ埋込部２への入力をＤ３（２）（Ｌ１）（＝Ｄ５（１）（Ｌ１））とし、機械翻訳処理部５からの出力をＤ５（２）（Ｒ２）とし、また、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ２）とする。
（６）Ｒ２→Ｒ１→Ｒ２：
　１回目の機械翻訳処理において、入力データ埋込部２への入力をＤ３（１）（Ｒ２）とし、機械翻訳処理部５からの出力をＤ５（１）（Ｒ１）とし、２回目の機械翻訳処理において、入力データ埋込部２への入力をＤ３（２）（Ｒ１）（＝Ｄ５（１）（Ｒ１））とし、機械翻訳処理部５からの出力をＤ５（２）（Ｒ２）とし、また、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ１）とする。
【０１９３】
　≪Ｂ１：ゼロショット折り返し機械翻訳処理（Ｌ１→Ｒ２→Ｌ１）≫
　入力データ埋込部２への１回目の入力をＤ３（１）（Ｌ１）とし、１回目の機械翻訳処理部５からの出力をＤ５（１）（Ｒ２）とし、入力データ埋込部２への２回目の入力をＤ３（２）（Ｒ２）（＝Ｄ５（１）（Ｒ２））とし、２回目の機械翻訳処理部５からの出力をＤ５（２）（Ｌ１）とし、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｌ１）とする場合について、説明する。
【０１９４】
　データ入力インターフェースＩＦ１は、第１単言語データ記憶部ＤＢｍ（Ｌ１）から他

分野の第１言語の単言語データＤ０（Ｌ１）を読み出し、読み出した単言語データをデータＤ１（＝Ｄ１（Ｌ１））として、入力データ取得部１に出力する。

【０１９５】

入力データ取得部１の第１入力データ取得処理部１１は、データＤ１（他分野第１言語のデータＤ１（Ｌ１））（他分野第１言語の文を構成するデータＤ１（Ｌ１））から、（１）トークンデータ$xi0_{token}$と、（２）当該トークンの位置を特定するための位置データxi_{pos}と、（３）当該トークンの言語を特定するための言語データxi_{lang}と、を取得する。

【０１９６】

そして、入力データ取得部１は、上記のように取得したトークンデータ$xi0_{token}$をマスク化処理部１２に出力する。

【０１９７】

制御部（不図示）は、入力データ取得部１のマスク化処理部１２に対して、マスク化処理を実行しないことを指示する制御信号ＣＴＬ２を生成し、当該制御信号ＣＴＬ２をマスク化処理部１２に出力する。

【０１９８】

マスク化処理部１２は、制御信号ＣＴＬ２に従い、第１入力データ取得処理部１１から入力したトークンデータ$xi0_{token}$をトークンデータxi_{token}として、第１セレクタＳＥＬ１に出力する。

【０１９９】

また、第１入力データ取得処理部１１は、位置データxi_{pos}（＝xi_{pos}（Ｌ１））を第１セレクタＳＥＬ１に出力し、言語データxi_{lang}（＝xi_{lang}（Ｌ１））を第１セレクタＳＥＬ１に出力する。

【０２００】

つまり、入力データ取得部１は、上記により取得したデータをデータＤ２ａ（＝Ｄ２ａ（Ｌ１）＝｛xi_{token}（Ｌ１），xi_{pos}（Ｌ１），xi_{lang}（Ｌ１）｝）として第１セレクタＳＥＬ１に出力する。

【０２０１】

正解データ取得部１３は、制御信号ＣＴＬ２に従い、データＤ１（＝Ｄ１（Ｌ１））から、疑似対訳データ生成用ＮＭＴモデル（入力データ埋込部２（埋込層に相当）と機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１とにより実現されるモデル）の学習処理に用いる正解データＤ＿ｃｏｒｒｅｃｔ（＝Ｄ＿ｃｏｒｒｅｃｔ（Ｌ１））を生成する。具体的には、正解データ取得部１３は、第１セレクタＳＥＬ１への入力データＤ２ａ（Ｌ１）と同一のデータを正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）に設定する。そして、正解データ取得部１３は、上記により生成した正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）を第２出力データ評価部６に出力する。

【０２０２】

制御部は、第１セレクタＳＥＬ１の端子「０」を選択する選択信号ｓｅｌ１を生成し、当該選択信号ｓｅｌ１を第１セレクタＳＥＬ１に出力する。

【０２０３】

第１セレクタＳＥＬ１は、選択信号ｓｅｌ１に従い、端子「０」を選択し、入力データ取得部１から出力されるデータＤ２ａ（Ｌ１）を、データＤ３$^{(1)}$（Ｌ１）として、入力データ埋込部２に出力する。

【０２０４】

入力データ埋込部２のトークン埋込部２１は、データＤ３$^{(1)}$（Ｌ１）に含まれるトークンデータxi_{token}を入力し、入力したトークンデータxi_{token}の分散表現データxi'_{token}を、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。

【０２０５】

入力データ埋込部２の位置埋込部２２は、データＤ３$^{(1)}$（Ｌ１）に含まれる位置データxi_{pos}を入力し、入力した位置データxi_{pos}の分散表現データxi'_{pos}を、上

記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。
【０２０６】
入力データ埋込部２の言語埋込部２３は、データＤ３$^{(1)}$（Ｌ１）に含まれる言語データｘｉ$_{lang}$を入力し、入力した言語データｘｉ$_{lang}$の分散表現データｘｉ’$_{lang}$を、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。
【０２０７】
入力データ埋込部２は、上記により取得された分散表現データをデータＤ４$^{(1)}$（Ｌ１）として、機械翻訳処理部５に出力する。
【０２０８】
機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力されるデータＤ４$^{(1)}$（Ｌ１）（分散表現データＤ４$^{(1)}$（Ｌ１）（＝｛ｘｉ’$_{token}$，ｘｉ’$_{pos}$，ｘｉ’$_{lang}$｝））と、制御部から出力される制御信号ＣＴＬ３とを入力する。
【０２０９】
制御部は、ＭＴ用ニューラルネットワークモデル５１からＲ２のデータを出力することを指示する制御信号ＣＴＬ３を生成し、当該制御信号ＣＴＬ３を機械翻訳処理部５に出力する。
【０２１０】
ＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力されるデータＤ４$^{(1)}$（Ｌ１）に対して、機械翻訳処理を実行し、データＤ５$^{(1)}$（Ｒ２）を取得する。なお、ＭＴ用ニューラルネットワークモデル５１は、制御信号ＣＴＬ３に従い、出力データの種別をＲ２（適応先分野、第２言語）とする。
【０２１１】
ＭＴ用ニューラルネットワークモデル５１は、上記により取得したデータＤ５$^{(1)}$（Ｒ２）を第２セレクタＳＥＬ２に出力する。
【０２１２】
制御部は、第２セレクタＳＥＬ２の端子「１」を選択する選択信号ｓｅｌ２を生成し、当該選択信号ｓｅｌ２を第２セレクタＳＥＬ２に出力する。
【０２１３】
第２セレクタＳＥＬ２は、選択信号ｓｅｌ２に従い、端子「１」を選択し、機械翻訳処理部５から出力されるデータＤ５（データＤ５$^{(1)}$（Ｒ２））を、第１バッファＢｕｆ１にデータＤ６ｂ（データＤ６ｂ$^{(1)}$（Ｒ２））として記憶する（図９を参照）。
【０２１４】
次に、第２入力データ取得処理部７は、第１バッファＢｕｆ１から出力されるデータＤ６ｂ’（データＤ６ｂ$^{(1)}$（Ｒ２））を入力し、データＤ６ｂ’（データＤ６ｂ$^{(1)}$（Ｒ２））から入力データ埋込部２に入力するためのデータ（＝｛ｘｉ$_{token}$（Ｒ２），ｘｉ$_{pos}$（Ｒ２），ｘｉ$_{lang}$（Ｒ２）｝）を取得（生成）する。そして、第２入力データ取得処理部７は、取得したデータをデータＤ２ｂ（データＤ２ｂ$^{(2)}$（Ｒ２）（＝｛ｘｉ$_{token}$（Ｒ２），ｘｉ$_{pos}$（Ｒ２），ｘｉ$_{lang}$（Ｒ２）｝））として、第１セレクタＳＥＬ１に出力する。
【０２１５】
制御部は、第１セレクタＳＥＬ１の端子「１」を選択する選択信号ｓｅｌ１を生成し、当該選択信号ｓｅｌ１を第１セレクタＳＥＬ１に出力する。
【０２１６】
第１セレクタＳＥＬ１は、選択信号ｓｅｌ１に従い、端子「１」を選択し、第２入力データ取得処理部７から出力されるデータＤ２ｂ$^{(2)}$（Ｒ２）を、データＤ３$^{(2)}$（Ｒ２）として、入力データ埋込部２に出力する。
【０２１７】
入力データ埋込部２のトークン埋込部２１は、データＤ３$^{(2)}$（Ｒ２）に含まれるトークンデータｘｉ$_{token}$を入力し、入力したトークンデータｘｉ$_{token}$の分散表現デ

ータｘｉ’ｔｏｋｅｎを、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。
【０２１８】
入力データ埋込部２の位置埋込部２２は、データＤ３（２）（Ｒ２）に含まれる位置データｘｉｐｏｓを入力し、入力した位置データｘｉｐｏｓの分散表現データｘｉ’ｐｏｓを、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。
【０２１９】
入力データ埋込部２の言語埋込部２３は、データＤ３（２）（Ｒ２）に含まれる言語データｘｉｌａｎｇを入力し、入力した言語データｘｉｌａｎｇの分散表現データｘｉ’ｌａｎｇを、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。
【０２２０】
入力データ埋込部２は、上記により取得された分散表現データをデータＤ４（２）（Ｒ２）として、機械翻訳処理部５に出力する。
【０２２１】
機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力されるデータＤ４（２）（Ｒ２）（分散表現データＤ４（２）（Ｒ２）（＝｛ｘｉ’ｔｏｋｅｎ，ｘｉ’ｐｏｓ，ｘｉ’ｌａｎｇ｝））と、制御部から出力される制御信号ＣＴＬ３とを入力する。
【０２２２】
制御部は、ＭＴ用ニューラルネットワークモデル５１からＬ１のデータを出力することを指示する制御信号ＣＴＬ３を生成し、当該制御信号ＣＴＬ３を機械翻訳処理部５に出力する。
【０２２３】
ＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力されるデータＤ４（２）（Ｒ２）に対して、機械翻訳処理を実行し、データＤ５（２）（Ｌ１）を取得する。なお、ＭＴ用ニューラルネットワークモデル５１は、制御信号ＣＴＬ３に従い、出力データの種別をＬ１（他分野、第１言語）とする。
【０２２４】
ＭＴ用ニューラルネットワークモデル５１は、上記により取得したデータＤ５（２）（Ｌ１）を第２セレクタＳＥＬ２に出力する。
【０２２５】
制御部は、第２セレクタＳＥＬ２の端子「０」を選択する選択信号ｓｅｌ２を生成し、当該選択信号ｓｅｌ２を第２セレクタＳＥＬ２に出力する。
【０２２６】
第２セレクタＳＥＬ２は、選択信号ｓｅｌ２に従い、端子「０」を選択し、機械翻訳処理部５から出力されるデータＤ５（２）（Ｌ１）を、データＤ６ａ（２）（Ｌ１）として、第２出力データ評価部６に出力する。
【０２２７】
第２出力データ評価部６は、第２セレクタＳＥＬ２から出力されるデータＤ６ａ（２）（Ｌ１）（＝Ｄ５（２）（Ｌ１））と、入力データ取得部１から出力される正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）とを入力する。第２出力データ評価部６は、機械翻訳処理部５の出力であるデータＤ５（２）（Ｌ１）と、その正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）とから損失を取得する。
【０２２８】
≪Ｂ２：ゼロショット折り返し機械翻訳処理（Ｒ１→Ｌ２→Ｒ１）≫
入力データ埋込部２への１回目の入力をＤ３（１）（Ｒ１）とし、１回目の機械翻訳処理部５からの出力をＤ５（１）（Ｌ２）とし、入力データ埋込部２への２回目の入力をＤ３（２）（Ｌ２）（＝Ｄ５（１）（Ｌ２））とし、２回目の機械翻訳処理部５からの出力をＤ５（２）（Ｒ１）とし、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ１）とする場合についても、疑似対訳データ生成装置１００は、上記と同様の処理を実行する。
【０２２９】

つまり、疑似対訳データ生成装置１００は、上記処理において、入力データ埋込部２への１回目の入力Ｄ３$^{(1)}$（Ｌ１）をＤ３$^{(1)}$（Ｒ１）に、機械翻訳処理部５からの出力Ｄ５$^{(1)}$（Ｒ２）をＤ５$^{(1)}$（Ｌ２）に置換し、入力データ埋込部２への２回目の入力Ｄ３$^{(2)}$（Ｒ２）をＤ３$^{(2)}$（Ｌ２）に、機械翻訳処理部５からの出力Ｄ５$^{(2)}$（Ｌ１）をＤ５$^{(2)}$（Ｒ１）に置換し、正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）を正解データＤ＿ｃｏｒｒｅｃｔ（Ｒ１）に置換して、上記処理と同様の処理を行う。

【０２３０】

≪Ｂ３：ゼロショット折り返し機械翻訳処理（Ｒ１→Ｒ２→Ｒ１）≫

入力データ埋込部２への１回目の入力をＤ３$^{(1)}$（Ｒ１）とし、１回目の機械翻訳処理部５からの出力をＤ５$^{(1)}$（Ｒ２）とし、入力データ埋込部２への２回目の入力をＤ３$^{(2)}$（Ｒ２）（＝Ｄ５$^{(1)}$（Ｒ２））とし、２回目の機械翻訳処理部５からの出力をＤ５$^{(2)}$（Ｒ１）とし、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ１）とする場合についても、疑似対訳データ生成装置１００は、上記と同様の処理を実行する。

【０２３１】

つまり、疑似対訳データ生成装置１００は、上記処理において、入力データ埋込部２への１回目の入力Ｄ３$^{(1)}$（Ｌ１）をＤ３$^{(1)}$（Ｒ１）に、機械翻訳処理部５からの出力Ｄ５$^{(1)}$（Ｒ２）を同じくＤ５$^{(1)}$（Ｒ２）とし、入力データ埋込部２への２回目の入力Ｄ３$^{(2)}$（Ｒ２）を同じくＤ３$^{(2)}$（Ｒ２）とし、機械翻訳処理部５からの出力Ｄ５$^{(2)}$（Ｌ１）をＤ５$^{(2)}$（Ｒ１）に置換し、正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）を正解データＤ＿ｃｏｒｒｅｃｔ（Ｒ１）に置換して、上記処理と同様の処理を行う。

【０２３２】

≪Ｂ４：ゼロショット折り返し機械翻訳処理（Ｌ２→Ｒ１→Ｌ２）≫

入力データ埋込部２への１回目の入力をＤ３$^{(1)}$（Ｌ２）とし、１回目の機械翻訳処理部５からの出力をＤ５$^{(1)}$（Ｒ１）とし、入力データ埋込部２への２回目の入力をＤ３$^{(2)}$（Ｒ１）（＝Ｄ５$^{(1)}$（Ｒ１））とし、２回目の機械翻訳処理部５からの出力をＤ５$^{(2)}$（Ｌ２）とし、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｌ２）とする場合についても、疑似対訳データ生成装置１００は、上記と同様の処理を実行する。

【０２３３】

つまり、疑似対訳データ生成装置１００は、上記処理において、入力データ埋込部２への１回目の入力Ｄ３$^{(1)}$（Ｌ１）をＤ３$^{(1)}$（Ｌ２）に、機械翻訳処理部５からの出力Ｄ５$^{(1)}$（Ｒ２）をＤ５$^{(1)}$（Ｒ１）に置換し、入力データ埋込部２への２回目の入力Ｄ３$^{(2)}$（Ｒ２）をＤ３$^{(2)}$（Ｒ１）に、機械翻訳処理部５からの出力Ｄ５$^{(2)}$（Ｌ１）をＤ５$^{(2)}$（Ｌ２）に置換し、正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）を正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ２）に置換して、上記処理と同様の処理を行う。

【０２３４】

≪Ｂ５：ゼロショット折り返し機械翻訳処理（Ｒ２→Ｌ１→Ｒ２）≫

入力データ埋込部２への１回目の入力をＤ３$^{(1)}$（Ｒ２）とし、１回目の機械翻訳処理部５からの出力をＤ５$^{(1)}$（Ｌ１）とし、入力データ埋込部２への２回目の入力をＤ３$^{(2)}$（Ｌ１）（＝Ｄ５$^{(1)}$（Ｌ１））とし、２回目の機械翻訳処理部５からの出力をＤ５$^{(2)}$（Ｒ２）とし、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ２）とする場合についても、疑似対訳データ生成装置１００は、上記と同様の処理を実行する。

【０２３５】

つまり、疑似対訳データ生成装置１００は、上記処理において、入力データ埋込部２への１回目の入力Ｄ３$^{(1)}$（Ｌ１）をＤ３$^{(1)}$（Ｒ２）に、機械翻訳処理部５からの出力Ｄ５$^{(1)}$（Ｒ２）をＤ５$^{(1)}$（Ｌ１）に置換し、入力データ埋込部２への２回目の入力Ｄ３$^{(2)}$（Ｒ２）をＤ３$^{(2)}$（Ｌ１）に、機械翻訳処理部５からの出力Ｄ５$^{(2)}$（Ｌ１）をＤ５$^{(2)}$（Ｒ２）に置換し、正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）を正解データＤ＿ｃｏｒｒｅｃｔ（Ｒ２）に置換して、上記処理と同様の処理を行う。

【０２３６】

≪Ｂ６：ゼロショット折り返し機械翻訳処理（Ｒ２→Ｒ１→Ｒ２）≫

入力データ埋込部２への１回目の入力をＤ３（１）（Ｒ２）とし、１回目の機械翻訳処理部５からの出力をＤ５（１）（Ｒ１）とし、入力データ埋込部２への２回目の入力をＤ３（２）（Ｒ１）（＝Ｄ５（１）（Ｒ１））とし、２回目の機械翻訳処理部５からの出力をＤ５（２）（Ｒ２）とし、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｒ２）とする場合についても、疑似対訳データ生成装置１００は、上記と同様の処理を実行する。

【０２３７】

つまり、疑似対訳データ生成装置１００は、上記処理において、入力データ埋込部２への１回目の入力Ｄ３（１）（Ｌ１）をＤ３（１）（Ｒ２）に、機械翻訳処理部５からの出力Ｄ５（１）（Ｒ２）をＤ５（１）（Ｒ１）に置換し、入力データ埋込部２への２回目の入力Ｄ３（２）（Ｒ２）をＤ３（２）（Ｒ１）に、機械翻訳処理部５からの出力Ｄ５（２）（Ｌ１）をＤ５（２）（Ｒ２）に置換し、正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）を正解データＤ＿ｃｏｒｒｅｃｔ（Ｒ２）に置換して、上記処理と同様の処理を行う。

【０２３８】

以上のようにして、疑似対訳データ生成装置１００では、

（１）Ｌ１→Ｒ２→Ｌ１
（２）Ｒ１→Ｌ２→Ｒ１
（３）Ｒ１→Ｒ２→Ｒ１
（４）Ｌ２→Ｒ１→Ｌ２
（５）Ｒ２→Ｌ１→Ｒ２
（６）Ｒ２→Ｒ１→Ｒ２

の６パターン（６種類の場合）について、ゼロショット折り返し機械翻訳処理が実行される。

【０２３９】

（１．２．１．２Ｃ：教師データあり機械翻訳処理）

教師データあり機械翻訳処理において、疑似対訳データ生成装置１００は、図１１に示すように、入力データ埋込部２への入力データを対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）から取得した対訳データＤ０（Ｌ１－Ｌ２）の一方の言語のデータとし、当該入力データに対応する対訳データが出力されるように学習処理を行う。つまり、教師あり機械翻訳処理において、正解データは、対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）から読み出した対訳データＤ０（Ｌ１－Ｌ２）に基づいて、設定される。

【０２４０】

なお、教師あり機械翻訳処理において、疑似対訳データ生成装置１００は、

（１）入力データ埋込部２への入力をＤ３（Ｌ１）とし、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｌ２）とする、あるいは、
（２）入力データ埋込部２への入力をＤ３（Ｌ２）とし、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｌ１）として、

疑似対訳データ生成用ＮＭＴモデル（入力データ埋込部２（埋込層に相当）と機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１とにより実現されるモデル）の学習処理（パラメータの最適化処理）を行う。

【０２４１】

≪Ｃ１：教師あり機械翻訳処理（Ｌ１→Ｌ２）≫

入力データ埋込部２への入力をＤ３（Ｌ１）とし、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｌ２）とする場合について、説明する。

【０２４２】

データ入力インターフェースＩＦ１は、対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）から他分野の対訳データＤ０（Ｌ１－Ｌ２）を読み出し、読み出した対訳データをデータＤ１（Ｌ１－Ｌ２）として、入力データ取得部１に出力する。

【０２４３】

入力データ取得部１の第１入力データ取得処理部１１は、データＤ１（他分野対訳データＤ１（Ｌ１－Ｌ２））から第１言語のデータをデータＤ１（Ｌ１）として抽出し、当該

第１言語のデータ（言語データ）Ｄ１（Ｌ１）（他分野第１言語の文を構成するデータＤ１（Ｌ１））から、（１）トークンデータｘｉ０$_{token}$と、（２）当該トークンの位置を特定するための位置データｘｉ$_{pos}$と、（３）当該トークンの言語を特定するための言語データｘｉ$_{lang}$と、を取得する。

【０２４４】

そして、入力データ取得部１は、上記のように取得したトークンデータｘｉ０$_{token}$をマスク化処理部１２に出力する。

【０２４５】

制御部（不図示）は、入力データ取得部１のマスク化処理部１２に対して、マスク化処理を実行しないことを指示する制御信号ＣＴＬ２を生成し、当該制御信号ＣＴＬ２をマスク化処理部１２に出力する。

【０２４６】

マスク化処理部１２は、制御信号ＣＴＬ２に従い、第１入力データ取得処理部１１から入力したトークンデータｘｉ０$_{token}$をトークンデータｘｉ$_{token}$として、第１セレクタＳＥＬ１に出力する。

【０２４７】

また、第１入力データ取得処理部１１は、位置データｘｉ$_{pos}$（＝ｘｉ$_{pos}$（Ｌ１））を第１セレクタＳＥＬ１に出力し、言語データｘｉ$_{lang}$（＝ｘｉ$_{lang}$（Ｌ１））を第１セレクタＳＥＬ１に出力する。

【０２４８】

つまり、入力データ取得部１は、上記により取得したデータをデータＤ２ａ（＝Ｄ２ａ（Ｌ１）＝｛ｘｉ$_{token}$（Ｌ１），ｘｉ$_{pos}$（Ｌ１），ｘｉ$_{lang}$（Ｌ１）｝）として第１セレクタＳＥＬ１に出力する。

【０２４９】

正解データ取得部１３は、制御信号ＣＴＬ２に従い、データＤ１（他分野対訳データＤ１（Ｌ１－Ｌ２））から第２言語のデータをＤ１（Ｌ２）（他分野対訳データＤ１（Ｌ１－Ｌ２）から抽出したデータＤ１（Ｌ１）（他分野第１言語のデータ）の対訳データに相当するデータ）として抽出し、抽出したデータＤ１（Ｌ２）から、疑似対訳データ生成用ＮＭＴモデルの学習処理に用いる正解データＤ＿ｃｏｒｒｅｃｔ（＝Ｄ＿ｃｏｒｒｅｃｔ（Ｌ２））を生成する。そして、正解データ取得部１３は、当該正解データＤ＿ｃｏｒｒｅｃｔ（＝Ｄ＿ｃｏｒｒｅｃｔ（Ｌ２））を第２出力データ評価部６に出力する。

【０２５０】

制御部は、第１セレクタＳＥＬ１の端子「０」を選択する選択信号ｓｅｌ１を生成し、当該選択信号ｓｅｌ１を第１セレクタＳＥＬ１に出力する。

【０２５１】

第１セレクタＳＥＬ１は、選択信号ｓｅｌ１に従い、端子「０」を選択し、入力データ取得部１から出力されるデータＤ２ａ（Ｌ１）を、データＤ３（Ｌ１）として、入力データ埋込部２に出力する。

【０２５２】

入力データ埋込部２のトークン埋込部２１は、データＤ３（Ｌ１）に含まれるトークンデータｘｉ$_{token}$を入力し、入力したトークンデータｘｉ$_{token}$の分散表現データｘｉ’$_{token}$を、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。

【０２５３】

入力データ埋込部２の位置埋込部２２は、データＤ３（Ｌ１）に含まれる位置データｘｉ$_{pos}$を入力し、入力した位置データｘｉ$_{pos}$の分散表現データｘｉ’$_{pos}$を、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。

【０２５４】

入力データ埋込部２の言語埋込部２３は、データＤ３（Ｌ１）に含まれる言語データｘｉ$_{lang}$を入力し、入力した言語データｘｉ$_{lang}$の分散表現データｘｉ’$_{lang}$を、上記のＭＬＭの処理で説明したのと同じ処理を実行して取得する。

【０２５５】
入力データ埋込部２は、上記により取得された分散表現データをデータＤ４（Ｌ１）として、機械翻訳処理部５に出力する。
【０２５６】
機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力される分散表現データＤ４（Ｌ１）（＝｛ｘｉ’ｔｏｋｅｎ，ｘｉ’ｐｏｓ，ｘｉ’ｌａｎｇ｝）と、制御部から出力される制御信号ＣＴＬ３とを入力する。
【０２５７】
制御部は、ＭＴ用ニューラルネットワークモデル５１からＬ２のデータを出力することを指示する制御信号ＣＴＬ３を生成し、当該制御信号ＣＴＬ３を機械翻訳処理部５に出力する。
【０２５８】
ＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力されるデータＤ４（Ｌ１）に対して、機械翻訳処理を実行し、データＤ５（Ｌ２）を取得する。なお、ＭＴ用ニューラルネットワークモデル５１は、制御信号ＣＴＬ３に従い、出力データの種別をＬ２（他分野、第２言語）とする。
【０２５９】
ＭＴ用ニューラルネットワークモデル５１は、上記により取得したデータＤ５（Ｌ２）を第２セレクタＳＥＬ２に出力する。
【０２６０】
制御部は、第２セレクタＳＥＬ２の端子「０」を選択する選択信号ｓｅｌ２を生成し、当該選択信号ｓｅｌ２を第２セレクタＳＥＬ２に出力する。
【０２６１】
第２セレクタＳＥＬ２は、選択信号ｓｅｌ２に従い、端子「０」を選択し、機械翻訳処理部５から出力されるデータＤ５（データＤ５（Ｌ２））を、第２出力データ評価部６にデータＤ６ａ（データＤ６ａ（Ｌ２））として出力する。
【０２６２】
第２出力データ評価部６は、第２セレクタＳＥＬ２を介して機械翻訳処理部５から出力されるデータＤ５（Ｌ２）（＝データＤ６ａ（Ｌ２））と、入力データ取得部１から出力される正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ２）とを入力する。第２出力データ評価部６は、機械翻訳処理部５の出力であるデータＤ５（Ｌ２）と、その正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ２）とから損失を取得する。
【０２６３】
≪Ｃ２：教師あり機械翻訳処理（Ｌ２→Ｌ１）≫
入力データ埋込部２への入力をＤ３（Ｌ２）とし、正解データをＤ＿ｃｏｒｒｅｃｔ（Ｌ１）とする場合についても、疑似対訳データ生成装置１００は、上記と同様の処理を実行する。つまり、疑似対訳データ生成装置１００は、上記処理において、Ｄ３（Ｌ１）をＤ３（Ｌ２）に置換し、正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ２）を正解データＤ＿ｃｏｒｒｅｃｔ（Ｌ１）に置換して、上記処理と同様の処理を行う。
【０２６４】
（１．２．１．２Ｄ：損失の計算処理）
上記の処理、すなわち、
（Ａ）自己符号化処理（ステップＳ１２１）、
（Ｂ）ゼロショット折り返し機械翻訳処理（ステップＳ１２２）、および
（Ｃ）教師データあり機械翻訳処理（ステップＳ１２３）
を実行した後、第２出力データ評価部６は、損失の計算処理を行う。なお、損失の計算をＭ文（Ｍ個の文、Ｍ：自然数）ごとに行う場合、Ｍ文の中の第ｉ番目の文に含まれるサブワード数をN_i（１≦ｉ≦Ｍ）とし、第ｉ番目の文についての機械翻訳処理部５からの出力データＤ５（X_{in}）（入力データをX_{in}で表す）のｊ番目（ｊ：自然数、１≦ｊ≦N_i）のサブワードに相当するデータをＤ５（（X_{in}→X_{out}），ｉ，ｊ）（「（X_{in}

→Ｘｏｕｔ）」は、入力データがＸｉｎであり、出力データがＸｏｕｔであることを表す）またはＤ５（（Ｘｉｎ→Ｘｍ→Ｘｏｕｔ），ｉ，ｊ）（「（Ｘｉｎ→Ｘｍ→Ｘｏｕｔ）」は、入力データがＸｉｎであり、１回目の出力データがＸｍであり、２回目の入力がＸｍであり、２回目の出力がＸｏｕｔであることを表す）とすると、第２出力データ評価部６は、下記数式のように、機械翻訳処理部５から出力されるデータと、正解データとから損失Ｌｏｓｓを取得する。

【数５】

$$Loss = Loss\_self + Loss\_zero + Loss\_mt$$

【数６】

$$\begin{aligned} Loss\_self = &\sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(L1,i,j), D5((L1 \to L1),i,j)) \\ &+ \sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(L2,i,j), D5((L2 \to L2),i,j)) \\ &+ \sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(R1,i,j), D5((R1 \to R1),i,j)) \\ &+ \sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(R2,i,j), D5((R2 \to R2),i,j)) \end{aligned}$$

【数７】

$$Loss\_zero = \sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(L1,i,j), D5((L1 \to R2 \to L1),i,j)) + \sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(R1,i,j), D5((R1 \to L2 \to R1),i,j)) + \sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(R1,i,j), D5((R1 \to R2 \to R1),i,j)) + \sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(L2,i,j), D5((L2 \to R1 \to L2),i,j)) + \sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(R2,i,j), D5((R2 \to L1 \to R2),i,j)) + \sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(R2,i,j), D5((R2 \to R1 \to R2),i,j))$$

【数８】

$$Loss\_mt = \sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(L2,i,j), D5((L1 \to L2),i,j)) + \sum_{i=1}^{M}\sum_{j=1}^{N_i} loss(D\_correct'(L1,i,j), D5((L2 \to L1),i,j))$$

【数９】

$$loss(p,q) = -\sum_{k=1}^{V} p(k)\log(q(k))$$

Ｖ：サブワード語彙のサイズ（各トークン（入力データ）文字列を表すベクトルの次元数）

ｐ：確率分布（ｐ（ｋ）は、ｋ番目（第ｋ次元）の要素の確率を示す）

ｑ：確率分布（ｑ（ｋ）は、ｋ番目（第ｋ次元）の要素の確率を示す）

なお、上記数式において、Ｄ＿ｃｏｒｒｅｃｔ’（ｘ，ｉ，ｊ）は、第２出力データ評価部６により正解データＤ＿ｃｏｒｒｅｃｔ（ｘ）の第ｉ番目の文の第ｊ番目のサブワードから取得されるデータ（ベクトル）であり、サブワード語彙のサイズ（＝各トークン（入力データ）文字列を表すベクトルの次元数（これをｎ１とする））と同じ次元数のベクト

ル（ｎ１次元のベクトル）である。そして、例えば、Ｄ＿ｃｏｒｒｅｃｔ’（ｘ，ｉ，ｊ）は、ｎ１次元のうち１次元のみ値が「１」であり（ｎ１次元ベクトルの要素のうち、当該正解サブワードに対応する要素のみが「１」）、それ以外は値が「０」であるｏｎｅ－ｈｏｔベクトルである。

【０２６５】

また、上記数式において、例えば、Ｄ５（（Ｌ１→Ｌ２），ｉ，ｊ）は、ｎ１次元の実数ベクトル（各次元のデータ（ベクトルの要素）が、対応するサブワードである確率を示す実数ベクトル）であり、ｓｏｆｔｍａｘ関数により確率化（実数ベクトルの各要素の総和が「１」となるように正規化）されている。そして、Ｄ５（（Ｌ１→Ｌ２），ｉ，ｊ）は、入力データをＬ１としたときの機械翻訳処理部５からの出力データ種別をＬ２としたときの出力データであり、第ｉ番目の文の第ｊ番目のサブワードに相当するデータである。

【０２６６】

なお、上記数式において、例えば、Ｄ５（（Ｌ１→Ｒ２→Ｌ１），ｉ，ｊ）は、入力データの種別と翻訳処理結果データ（出力データ）の種別を明示した形式で示しており、（Ｌ１→Ｒ２→Ｌ１）は、１回目の処理において種別Ｌ１のデータが種別Ｒ２のデータとして出力され、２回目の処理において、１回目の出力を入力とし、種別Ｌ１のデータが出力されたときのデータであることを示している。

【０２６７】

また、上記数式において、ｌｏｓｓ（ｐ、ｑ）は、交差エントロピー誤差を求める数式であり、これにより、確率分布間（上記数式では、確率分布ｐ、ｑ）の相違を定量化できる。

【０２６８】

このように、第２出力データ評価部６は、上記数式に相当する処理により、損失Ｌｏｓｓを取得する。

【０２６９】

（１．２．１．２Ｅ：パラメータの更新処理）

第２出力データ評価部６は、所定の学習データに対して上記で算出した損失（学習損失）に基づいてＸＬＭ処理部３のＭＴ用ニューラルネットワークモデル５１のパラメータθ_{MT}を更新するためのデータｕｐｄａｔｅ（θ_{MT}）を生成し、当該データｕｐｄａｔｅ（θ_{MT}）を機械翻訳処理部５に出力する。

【０２７０】

機械翻訳処理部５のＭＬ用ニューラルネットワークモデル５１は、データｕｐｄａｔｅ（θ_{MT}）に基づいて、パラメータθ_{MT}を更新する。

【０２７１】

また、ＭＴ用ニューラルネットワークモデル５１は、ＭＴ用ニューラルネットワークモデル５１のパラメータを更新した後、埋込層（入力データ埋込部２のトークン埋込部２１、位置埋込部２２、および、言語埋込部２３に対応）のパラメータを更新するためのパラメータ更新データｕｐｄａｔｅ（θ_{emb}）を生成し、当該パラメータ更新データｕｐｄａｔｅ（θ_{emb}）を入力データ埋込部２に出力する。

【０２７２】

入力データ埋込部２は、パラメータ更新データｕｐｄａｔｅ（θ_{emb}）に基づいて、埋込層（入力データ埋込部２のトークン埋込部２１、位置埋込部２２、および、言語埋込部２３に対応）のパラメータを更新する。例えば、入力データ埋込部２は、パラメータ更新データｕｐｄａｔｅ（θ_{emb}）に基づいて、変換行列（例えば、行列Ｗｔｏｋｅｎ、Ｗｐｏｓ、Ｗｌａｎｇ）の要素（値）を更新する。

【０２７３】

疑似対訳データ生成装置１００は、所定の終了条件を満たすまで、上記処理（図６のループ２の処理）を繰り返し実行する。その際、例えば、学習に用いるデータとは別の調整用データに対して第２出力データ評価部６で算出される損失（テスト損失）を評価値として参照する。

【０２７４】
そして、疑似対訳データ生成装置１００は、上記処理（図６のループ２の処理）の終了条件が満たされたときをもって、疑似対訳データ生成用ＮＭＴモデルの学習処理を終了させる。
【０２７５】
なお、図６のフローチャートのループ処理（ループ２）の終了条件は、例えば、以下のように設定される。
（１）事前に定めた反復回数だけループ処理（ループ２）が実行された。
（２）疑似対訳データ生成用ＮＭＴモデルの学習処理において、第２出力データ評価部６における評価値が一定以上（事前に定めた値以上）の変化を示さなかった。
（３）疑似対訳データ生成用ＮＭＴモデルの学習処理において、第２出力データ評価部６における評価値が事前に定めた値を下回った。
（４）疑似対訳データ生成用ＮＭＴモデルの学習処理において、第２出力データ評価部６における評価値が事前に定めた回数更新されなかった。
【０２７６】
上記の終了条件を満たす場合、疑似対訳データ生成装置１００は、疑似対訳データ生成用ＮＭＴモデルの学習処理が完了した判断し、当該処理を終了させる。
【０２７７】
以上のようにして、疑似対訳データ生成装置１００は、事前学習処理を完了した時点におけるパラメータを疑似対訳データ生成用ＮＭＴモデル（入力データ埋込部２（埋込層に相当）と機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１とにより実現されるモデル）の初期状態（初期パラメータを設定した状態（入力データ埋込部２（埋込層に相当）のパラメータがＸＬＭの学習により最適化された状態））として、上記処理により、疑似対訳データ生成用ＮＭＴモデルの学習処理（パラメータ最適化処理）を行う。
【０２７８】
（１．２．２：疑似対訳データ生成処理）
次に、疑似対訳データ生成処理（図４のステップＳ２）について説明する。
【０２７９】
疑似対訳データ生成装置１００は、上記の事前学習処理によりに取得された疑似対訳データ生成用ＮＭＴモデル（入力データ埋込部２（埋込層に相当）と機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１とにより実現されるモデル）を用いて、適応先分野の疑似対訳データを自動的に生成する。
【０２８０】
具体的には、疑似対訳データ生成装置１００は、以下の（１）、（２）の方法を用いて、疑似対訳データ生成処理を行う。
（１）他分野の対訳データ（Ｌ１－Ｌ２）を適応先分野向けに改変する方法（第１の方法）
（２）適応先分野の単言語データ（Ｒ１またはＲ２）を機械翻訳する方法（第２の方法）
以下、上記２つの方法による疑似対訳データ生成処理について、説明する。
【０２８１】
（１．２．２．１：疑似対訳データ生成処理（第１の方法）（他分野対訳データを利用））
まず、第１の方法（他分野対訳データを利用する方法）について、説明する。
【０２８２】
データ入力インターフェースＩＦ１は、対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）から他分野の対訳データＤ０（Ｌ１－Ｌ２）を読み出し、読み出した対訳データをデータＤ１（Ｌ１－Ｌ２）として、入力データ取得部１に出力する。
【０２８３】
入力データ取得部１の第１入力データ取得処理部１１は、データＤ１（他分野対訳データＤ１（Ｌ１－Ｌ２））から、Ｌ１のデータをデータＤ１（Ｌ１）として抽出し、当該データＤ１（Ｌ１）から、（１）トークンデータ$xi0_{token}$（＝$xi0_{token}$（Ｌ

１））と、（２）当該トークンの位置を特定するための位置データｘｉ$_{pos}$（＝ｘｉ$_{pos}$（Ｌ１））と、（３）当該トークンの言語を特定するための言語データｘｉ$_{lang}$（＝ｘｉ$_{lang}$（Ｌ１））と、を取得する。

【０２８４】

そして、入力データ取得部１は、上記により取得したトークンデータｘｉ０$_{token}$（Ｌ１）をマスク化処理部１２に出力する。

【０２８５】

制御部（不図示）は、入力データ取得部１のマスク化処理部１２に対して、マスク化処理を実行しないことを指示する制御信号ＣＴＬ２を生成し、当該制御信号ＣＴＬ２をマスク化処理部１２に出力する。

【０２８６】

マスク化処理部１２は、制御信号ＣＴＬ２に従い、第１入力データ取得処理部１１から入力したトークンデータｘｉ０$_{token}$（Ｌ１）を、トークンデータｘｉ$_{token}$（Ｌ１）として第１セレクタＳＥＬ１に出力する。

【０２８７】

また、第１入力データ取得処理部１１は、位置データｘｉ$_{pos}$（＝ｘｉ$_{pos}$（Ｌ１））を第１セレクタＳＥＬ１に出力し、言語データｘｉ$_{lang}$（＝ｘｉ$_{lang}$（Ｌ１））を第１セレクタＳＥＬ１に出力する。

【０２８８】

つまり、入力データ取得部１は、上記により取得したデータをデータＤ２ａ（＝Ｄ２ａ（Ｌ１）＝｛ｘｉ$_{token}$（Ｌ１），ｘｉ$_{pos}$（Ｌ１），ｘｉ$_{lang}$（Ｌ１）｝）として第１セレクタＳＥＬ１に出力する。

【０２８９】

制御部は、第１セレクタＳＥＬ１の端子「０」を選択する選択信号ｓｅｌ１を生成し、当該選択信号ｓｅｌ１を第１セレクタＳＥＬ１に出力する。

【０２９０】

第１セレクタＳＥＬ１は、選択信号ｓｅｌ１に従い、端子「０」を選択し、入力データ取得部１から出力されるデータＤ２ａ（Ｌ１）を、データＤ３（Ｌ１）として、入力データ埋込部２に出力する。

【０２９１】

入力データ埋込部２は、データＤ３（Ｌ１）に含まれる（１）トークンデータｘｉ$_{token}$（Ｌ１）、（２）位置データｘｉ$_{pos}$（Ｌ１）、（３）言語データｘｉ$_{lang}$（Ｌ１）から、分散表現データＤ４（Ｌ１）（＝｛ｘｉ’$_{token}$（Ｌ１），ｘｉ’$_{pos}$（Ｌ１），ｘｉ’$_{lang}$（Ｌ１）｝）を取得し、取得した分散表現データＤ４（Ｌ１）を機械翻訳処理部５に出力する。

【０２９２】

機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力されるデータＤ４（Ｌ１）（分散表現データＤ４（Ｌ１）（＝｛ｘｉ’$_{token}$（Ｌ１），ｘｉ’$_{pos}$（Ｌ１），ｘｉ’$_{lang}$（Ｌ１）｝））と、制御部から出力される制御信号ＣＴＬ３とを入力する。

【０２９３】

制御部は、ＭＴ用ニューラルネットワークモデル５１からＲ２のデータを出力することを指示する制御信号ＣＴＬ３を生成し、当該制御信号ＣＴＬ３を機械翻訳処理部５に出力する。

【０２９４】

ＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力されるデータＤ４（Ｌ１）に対して、機械翻訳処理を実行し、データＤ５（Ｒ２）を取得し、データＤ５（Ｒ２）を保持する。なお、ＭＴ用ニューラルネットワークモデル５１は、制御信号ＣＴＬ３に従い、出力データの種別をＲ２（適応先分野、第２言語）とする。

【０２９５】

また、入力データ取得部１の第１入力データ取得処理部１１は、データＤ１（他分野対訳データＤ１（Ｌ１－Ｌ２））から、Ｌ２のデータをデータＤ１（Ｌ２）として抽出し、当該データＤ１（Ｌ２）から、（１）トークンデータ$xi0_{token}$（＝$xi0_{token}$（Ｌ２））と、（２）当該トークンの位置を特定するための位置データxi_{pos}（＝xi_{pos}（Ｌ２））と、（３）当該トークンの言語を特定するための言語データxi_{lang}（＝xi_{lang}（Ｌ２））と、を取得する。

【０２９６】

そして、入力データ取得部１は、上記により取得したトークンデータ$xi0_{token}$（Ｌ２）をマスク化処理部１２に出力する。

【０２９７】

制御部（不図示）は、入力データ取得部１のマスク化処理部１２に対して、マスク化処理を実行しないことを指示する制御信号ＣＴＬ２を生成し、当該制御信号ＣＴＬ２をマスク化処理部１２に出力する。

【０２９８】

マスク化処理部１２は、制御信号ＣＴＬ２に従い、第１入力データ取得処理部１１から入力したトークンデータ$xi0_{token}$（Ｌ２）を、トークンデータxi_{token}（Ｌ２）として第１セレクタＳＥＬ１に出力する。

【０２９９】

また、第１入力データ取得処理部１１は、位置データxi_{pos}（＝xi_{pos}（Ｌ２））を第１セレクタＳＥＬ１に出力し、言語データxi_{lang}（＝xi_{lang}（Ｌ２））を第１セレクタＳＥＬ１に出力する。

【０３００】

つまり、入力データ取得部１は、上記により取得したデータをデータＤ２ａ（＝Ｄ２ａ（Ｌ２）＝｛xi_{token}（Ｌ２），xi_{pos}（Ｌ２），xi_{lang}（Ｌ２）｝）として第１セレクタＳＥＬ１に出力する。

【０３０１】

制御部は、第１セレクタＳＥＬ１の端子「０」を選択する選択信号ｓｅｌ１を生成し、当該選択信号ｓｅｌ１を第１セレクタＳＥＬ１に出力する。

【０３０２】

第１セレクタＳＥＬ１は、選択信号ｓｅｌ１に従い、端子「０」を選択し、入力データ取得部１から出力されるデータＤ２ａ（Ｌ２）を、データＤ３（Ｌ２）として、入力データ埋込部２に出力する。

【０３０３】

入力データ埋込部２は、データＤ３（Ｌ２）に含まれる（１）トークンデータxi_{token}（Ｌ２）、（２）位置データxi_{pos}（Ｌ２）、（３）言語データxi_{lang}（Ｌ２）から、分散表現データＤ４（Ｌ２）（＝｛xi'_{token}（Ｌ２），xi'_{pos}（Ｌ２），xi'_{lang}（Ｌ２）｝）を取得し、取得した分散表現データＤ４（Ｌ２）を機械翻訳処理部５に出力する。

【０３０４】

機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力されるデータＤ４（Ｌ２）（分散表現データＤ４（Ｌ２）（＝｛xi'_{token}（Ｌ２），xi'_{pos}（Ｌ２），xi'_{lang}（Ｌ２）｝））と、制御部から出力される制御信号ＣＴＬ３とを入力する。

【０３０５】

制御部は、ＭＴ用ニューラルネットワークモデル５１からＲ１のデータを出力することを指示する制御信号ＣＴＬ３を生成し、当該制御信号ＣＴＬ３を機械翻訳処理部５に出力する。

【０３０６】

ＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力されるデータＤ４（Ｌ２）に対して、機械翻訳処理を実行し、データＤ５（Ｒ１）を取得する。なお

、ＭＴ用ニューラルネットワークモデル５１は、制御信号ＣＴＬ３に従い、出力データの種別をＲ１（適応先分野、第１言語）とする。

【０３０７】

機械翻訳処理部５は、上記処理により取得したデータＤ５（Ｒ２）とデータＤ５（Ｒ１）とをペアリングして、データＤ５（Ｒ２－Ｒ１）として、第２セレクタＳＥＬ２に出力する。

【０３０８】

制御部は、第２セレクタＳＥＬ２の端子「０」を選択する選択信号ｓｅｌ２を生成し、当該選択信号ｓｅｌ２を第２セレクタＳＥＬ２に出力する。

【０３０９】

第２セレクタＳＥＬ２は、選択信号ｓｅｌ２に従い、端子「０」を選択し、機械翻訳処理部５から出力されるデータＤ５（データＤ５（Ｒ２－Ｒ１））を、データＤ６ａ（Ｒ２－Ｒ１）として、フィルター処理部８に出力する（図１２を参照）。

【０３１０】

制御部は、フィルター処理部８にて、Ｒ２－Ｒ１の疑似対訳データに対してフィルター処理を実行することを指示する制御信号ＣＴＬ４を生成し、当該制御信号をフィルター処理部８に出力する。

【０３１１】

フィルター処理部８は、制御信号ＣＴＬ４に従い、入力されるデータＤ６ａ（Ｒ２－Ｒ１）（Ｒ２およびＲ１の疑似対訳データ）に対してフィルター処理を行う。

【０３１２】

ここで、上記処理により取得された適応先分野の疑似対訳データ（Ｄ６ａ（Ｒ２－Ｒ１））は、教師データがない適応先分野の対訳データを出力するように学習させたニューラルネットワークモデルを用いて処理して取得したデータであるため、対訳データとしての品質が低い可能性がある。

【０３１３】

そこで、フィルター処理部８は、上記処理により取得された適応先分野の疑似対訳データ（Ｄ６ａ（Ｒ２－Ｒ１））に対して、フィルター処理を行う。例えば、フィルター処理部８は、上記処理により取得された適応先分野の疑似対訳データ（Ｄ６ａ（Ｒ２－Ｒ１））の各文対に信頼度を付与し、付与した信頼度に基づいて、フィルタリングを行う。

【０３１４】

例えば、元のＬ１－Ｌ２の対訳データ（対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）から取得した対訳データＤ０（Ｌ１－Ｌ２））を参照することが考えられる。より具体的には、フィルター処理部８は、対訳データＬ１－Ｌ２（Ｄ０（Ｌ１－Ｌ２））における文対ａｌ－ａ２を機械翻訳して取得した、適応先分野Ｒ１－Ｒ２の疑似対訳データの文対ｂｌ－ｂ２の信頼度を、ａｌとｂｌとの類似度、および、ａ２とｂ２の類似度に基づいて算出し、その信頼度が事前に定めた閾値Ｔｈよりも高いか否かを判断する。なお、フィルター処理部８は、対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）から、対訳データＬ１－Ｌ２（Ｄ０（Ｌ１－Ｌ２））における文対ａｌ－ａ２を取得できるものとする（例えば、データ入力インターフェースＩＦ１を介して取得する）。

【０３１５】

そして、フィルター処理部８は、上記により算出した信頼度が所定の閾値Ｔｈよりも高い疑似対訳データ（Ｒ１－Ｒ２）のみをフィルタリングにより取得する。そして、当該フィルタリングにより取得した疑似対訳データを、疑似対訳データＤｐｓｄ１（Ｒ１－Ｒ２，＃１）として、疑似対訳データ格納部ＤＢ１に出力する。

【０３１６】

（１．２．２．２：疑似対訳データ生成処理（第２の方法）（適応先分野単言語データを利用））

次に、第２の方法（適応先分野単言語データを利用する方法）について、説明する。

【０３１７】

図１３に示すように、データ入力インターフェースＩＦ１は、第３単言語データ記憶部ＤＢｍ（Ｒ１）から適応先分野の第１言語の単言語データＤ０（Ｒ１）を読み出し、読み出した第１言語のデータをデータＤ１（Ｒ１）として、入力データ取得部１に出力する。

【０３１８】

入力データ取得部１の第１入力データ取得処理部１１は、データＤ１（適応先分野第１言語データＤ１（Ｒ１））から、当該第１言語のデータ（言語データ）Ｄ１（Ｒ１）（適応先分野第１言語の文を構成するデータＤ１（Ｒ１））から、（１）トークンデータ$xi0_{token}$（$=xi0_{token}$（Ｒ１））と、（２）当該トークンの位置を特定するための位置データxi_{pos}（$=xi_{pos}$（Ｒ１））と、（３）当該トークンの言語を特定するための言語データxi_{lang}（$=xi_{lang}$（Ｒ１））と、を取得する。

【０３１９】

そして、入力データ取得部１は、上記により取得したトークンデータ$xi0_{token}$（Ｒ１）をマスク化処理部１２に出力する。

【０３２０】

制御部（不図示）は、入力データ取得部１のマスク化処理部１２に対して、マスク化処理を実行しないことを指示する制御信号ＣＴＬ２を生成し、当該制御信号ＣＴＬ２をマスク化処理部１２に出力する。

【０３２１】

マスク化処理部１２は、制御信号ＣＴＬ２に従い、第１入力データ取得処理部１１から入力したトークンデータ$xi0_{token}$（Ｒ１）を、トークンデータxi_{token}（Ｒ１）として第１セレクタＳＥＬ１に出力する。

【０３２２】

また、第１入力データ取得処理部１１は、位置データxi_{pos}（$=xi_{pos}$（Ｒ１））を第１セレクタＳＥＬ１に出力し、言語データxi_{lang}（$=xi_{lang}$（Ｒ１））を第１セレクタＳＥＬ１に出力する。

【０３２３】

つまり、入力データ取得部１は、上記により取得したデータをデータＤ２ａ（＝Ｄ２ａ（Ｒ１）＝｛xi_{token}（Ｒ１），xi_{pos}（Ｒ１），xi_{lang}（Ｒ１）｝）として第１セレクタＳＥＬ１に出力する。

【０３２４】

制御部は、第１セレクタＳＥＬ１の端子「０」を選択する選択信号ｓｅｌ１を生成し、当該選択信号ｓｅｌ１を第１セレクタＳＥＬ１に出力する。

【０３２５】

第１セレクタＳＥＬ１は、選択信号ｓｅｌ１に従い、端子「０」を選択し、入力データ取得部１から出力されるデータＤ２ａ（Ｒ１）を、データＤ３（Ｒ１）として、入力データ埋込部２に出力する。

【０３２６】

入力データ埋込部２は、データＤ３（Ｒ１）に含まれる（１）トークンデータxi_{token}、（２）位置データxi_{pos}、（３）言語データxi_{lang}から、分散表現データＤ４（Ｒ１）（＝｛xi'_{token}（Ｒ１），xi'_{pos}（Ｒ１），xi'_{lang}（Ｒ１）｝）を取得し、取得した分散表現データＤ４（Ｒ１）を機械翻訳処理部５に出力する。

【０３２７】

機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力される分散表現データＤ４（Ｒ１）（＝｛xi'_{token}（Ｒ１），xi'_{pos}（Ｒ１），xi'_{lang}（Ｒ１）｝）と、制御部から出力される制御信号ＣＴＬ３とを入力する。

【０３２８】

制御部は、ＭＴ用ニューラルネットワークモデル５１からＲ２のデータを出力することを指示する制御信号ＣＴＬ３を生成し、当該制御信号ＣＴＬ３を機械翻訳処理部５に出力する。

【０３２９】
ＭＴ用ニューラルネットワークモデル５１は、入力データ埋込部２から出力されるデータＤ４（Ｒ１）に対して、機械翻訳処理を実行し、データＤ５（Ｒ２）を取得する。なお、ＭＴ用ニューラルネットワークモデル５１は、制御信号ＣＴＬ３に従い、出力データの種別をＲ２（適応先分野、第２言語）とする。
【０３３０】
ＭＴ用ニューラルネットワークモデル５１は、上記により取得したデータＤ５（Ｒ２）を第２セレクタＳＥＬ２に出力する。
【０３３１】
制御部は、第２セレクタＳＥＬ２の端子「０」を選択する選択信号ｓｅｌ２を生成し、当該選択信号ｓｅｌ２を第２セレクタＳＥＬ２に出力する。
【０３３２】
第２セレクタＳＥＬ２は、選択信号ｓｅｌ２に従い、端子「０」を選択し、機械翻訳処理部５から出力されるデータＤ５（データＤ５（Ｒ２））を、データＤ６ａ（Ｒ２）として、フィルター処理部８に出力する（図１３を参照）。
【０３３３】
制御部は、フィルター処理部８にて、Ｒ１－Ｒ２の疑似対訳データに対してフィルター処理を実行することを指示する制御信号ＣＴＬ４を生成し、当該制御信号をフィルター処理部８に出力する。
【０３３４】
また、入力データ取得部１は、第３単言語データ記憶部ＤＢｍ（Ｒ１）から取得した適応先分野の言語データＤ０（Ｒ１）（機械翻訳処理の入力としたデータ）をデータＤ１＿ｏｒｇ（Ｒ１）として、フィルター処理部８に出力する。
【０３３５】
そして、フィルター処理部８は、入力される、（１）データＤ１＿ｏｒｇ（Ｒ１）（機械翻訳処理の入力としたデータ）と、（２）データＤ６ａ（Ｒ２）（データＤ３（Ｒ１）の機械翻訳データ）とを対応づけることで適応先分野の疑似対訳データ（Ｒ１－Ｒ２）を取得する。そして、フィルター処理部８は、制御信号ＣＴＬ４に従い、疑似対訳データ（Ｒ１－Ｒ２）にフィルター処理を行う。
【０３３６】
ここで、上記処理により取得された適応先分野の疑似対訳データ（Ｒ１－Ｒ２）は、教師データがない適応先分野の単語データを出力するように学習させたニューラルネットワークモデルを用いて処理して取得したデータであるため、対訳データとしての品質が低い可能性がある。
【０３３７】
そこで、フィルター処理部８は、上記処理により取得された適応先分野の疑似対訳データ（Ｒ１－Ｒ２）に対して、フィルター処理を行う。フィルター処理部８は、例えば、下記文献Ｄに開示されている機械翻訳の品質推定に関する技術を用いて、上記処理により取得された適応先分野の疑似対訳データ（Ｒ１－Ｒ２）に対して信頼度を付与する。
（文献Ｄ）：Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold (2018). Quality Estimation for Machine Translation. Morgan & Claypool.
そして、フィルター処理部８は、付与された信頼度に基づいて、フィルタリングを行う。例えば、フィルター処理部８は、上記により算出した信頼度が所定の閾値Ｔｈよりも高い疑似対訳データ（Ｒ１－Ｒ２）のみをフィルタリングにより取得する。そして、フィルター処理部８は、当該フィルタリングにより取得した疑似対訳データを、疑似対訳データＤｐｓｄ１（Ｒ１－Ｒ２，＃２）として、疑似対訳データ格納部ＤＢ１に出力する。
【０３３８】
なお、上記では、データＤ１＿ｏｒｇを適応先分野の第１言語のデータ（Ｒ１のデータ（＝Ｄ１＿ｏｒｇ（Ｒ１））とし、データＤ３（Ｒ１）を機械翻訳処理により取得した翻訳データを適応先分野の第２言語のデータＤ５（Ｒ２）（データＤ６ａ（Ｒ２））とし

たが（図１３の場合）、図１４に示すように、データＤ１＿ｏｒｇを適応先分野の第２言語のデータ（Ｒ２のデータ（＝Ｄ１＿ｏｒｇ（Ｒ２））とし、データＤ３（Ｒ２）を機械翻訳処理により取得した翻訳データを適応先分野の第２言語のデータＤ５（Ｒ１）（データＤ６ａ（Ｒ２））として、上記と同様の処理により、疑似対訳データ（データＤｐｓｄ１（Ｒ１－Ｒ２，＃２））を生成するようにしてもよい。

【０３３９】

≪まとめ≫

以上のように、疑似対訳データ生成装置１００では、

（１）大規模で（対訳データ数が多く）精度の高い他分野の対訳データ（対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）に記憶されている対訳データ）と、

（２）他分野の第１言語データ（第１単言語データ記憶部ＤＢｍ（Ｌ１）に記憶されている単言語データ）と、

（３）他分野の第２言語データ（第２単言語データ記憶部ＤＢｍ（Ｌ２）に記憶されている単言語データ）と、

（４）適応先分野の第１言語データ（第３単言語データ記憶部ＤＢｍ（Ｒ１）に記憶されている単言語データ）と、

（５）適応先分野の第２言語データ（第４単言語データ記憶部ＤＢｍ（Ｒ２）に記憶されている単言語データ）と、

を用いて、言語横断言語モデル（ＸＬＭ）（入力データ埋込部２（埋込層に相当）とＸＬＭ用ニューラルネットワークモデル３１とにより実現されるモデル）を最適化する（事前学習処理による初期化）。そして、疑似対訳データ生成装置１００では、言語横断言語モデル（ＸＬＭ）の最適化処理後（事前学習処理後）の入力データ埋込部２（埋込層に相当）（ＸＬＭの最適化により取得されたパラメータ（例えば、変換行列Ｗｔｏｋｅｎ、Ｗｐｏｓ、Ｗｌａｎｇが設定されている状態の入力データ埋込部２））と、学習前の機械翻訳処理部５のＭＴ用ニューラルネットワークモデル５１とからなる疑似対訳データ生成用ＮＭＴモデルの状態を初期状態として、疑似対訳データ生成用ＮＭＴモデルのパラメータ最適化処理を行う。これにより、疑似対訳データ生成装置１００では、適応先分野（機械翻訳の対象とする分野）の対訳データが一切ない場合であっても、適応先分野の第１言語および第２言語の疑似対訳データ（Ｒ１－Ｒ２）を生成できるモデル（学習済みモデル）を取得できる。

【０３４０】

（１．２．３：翻訳モデル学習処理）

次に、翻訳モデルの学習処理（図５のステップＳ３）について説明する。

【０３４１】

機械翻訳装置ＭＴ１は、ＮＭＴモデルを有しており、当該ＮＭＴモデルにより機械翻訳を行う装置である。機械翻訳装置ＭＴ１は、疑似対訳データ生成装置１００により生成され、疑似対訳データ格納部ＤＢ１に記憶された疑似対訳データを用いて、適応先分野の機械翻訳を行うための機械翻訳モデル（ＮＭＴモデル）の学習処理を行う。

【０３４２】

具体的には、例えば、機械翻訳装置ＭＴ１は、以下の（１）～（３）の方法のいずれかにより、適応先分野の機械翻訳を行うための機械翻訳モデル（ＮＭＴモデル）の学習処理を行う。

（１）微調整法：

機械翻訳装置ＭＴ１は、対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）から、他分野の対訳データＤ０’（Ｌ１－Ｌ２）を取得し、他分野の対訳データＤ０’（Ｌ１－Ｌ２）を用いて、機械翻訳装置ＭＴ１のＮＭＴモデルの学習処理を行う。その後、機械翻訳装置ＭＴ１は、疑似対訳データ生成装置１００により生成された適応先分野の疑似対訳データ（Ｒ１－Ｒ２の対訳データ）を疑似対訳データ格納部ＤＢ１から疑似対訳データＤｐｓｄ２（Ｒ１－Ｒ２）として読み出し、読み出した適応先分野の疑似対訳データＤｐｓｄ２（Ｒ１－Ｒ２）により、機械翻訳装置ＭＴ１のＮＭＴモデルの学習処理（パラメータの微調整）を行う。

（２）データ混合法：
機械翻訳装置ＭＴ１は、対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）から、他分野の対訳データＤ０’（Ｌ１－Ｌ２）を取得するとともに、疑似対訳データ格納部ＤＢ１から、適応分野の疑似対訳データＤｐｓｄ２（Ｒ１－Ｒ２）を取得する。そして、機械翻訳装置ＭＴ１は、他分野（Ｌ１－Ｌ２）の対訳データＤ０’（Ｌ１－Ｌ２）と適応先分野（Ｒ１－Ｒ２）の疑似対訳データＤｐｓｄ２（Ｒ１－Ｒ２）とを混合したデータを生成し、生成した当該データにより、機械翻訳装置ＭＴ１のＮＭＴモデルを学習させる。なお、他分野（Ｌ１－Ｌ２）の対訳データＤ０’（Ｌ１－Ｌ２）と適応先分野（Ｒ１－Ｒ２）の疑似対訳データＤｐｓｄ２（Ｒ１－Ｒ２）とを混合する際に、２種類の対訳データをタグで区別するようにしてもよい。また、２種類の対訳データを混合する前に、一方または両方をオーバーサンプリングまたはアンダーサンプリングすることで対訳データの混合比を変更してもよい。
（３）データ混合微調整法：
機械翻訳装置ＭＴ１は、対訳データ記憶部ＤＢｐ（Ｌ１－Ｌ２）から、他分野の対訳データＤ０’（Ｌ１－Ｌ２）を取得し、他分野の対訳データＤ０’（Ｌ１－Ｌ２）を用いて、機械翻訳装置ＭＴ１のＮＭＴモデルの学習処理を行う。その後、機械翻訳装置ＭＴ１は、機械翻訳装置ＭＴ１のＮＭＴモデルの学習処理に用いた対訳データと同じ対訳データＤ０’（Ｌ１－Ｌ２）と、適応先分野の疑似対訳データ（Ｒ１－Ｒ２）（疑似対訳データ格納部ＤＢ１から読み出した適応先分野の疑似対訳データＤｐｓｄ２（Ｒ１－Ｒ２））とを混合したデータを生成する。そして、機械翻訳装置ＭＴ１は、生成したデータ（２種類の対訳データを混合したデータ）を用いて、機械翻訳装置ＭＴ１のＮＭＴモデルの学習処理（パラメータの微調整）を行う。なお、他分野（Ｌ１－Ｌ２）の対訳データＤ０’（Ｌ１－Ｌ２）と適応先分野（Ｒ１－Ｒ２）の疑似対訳データＤｐｓｄ２（Ｒ１－Ｒ２）とを混合する際に、２種類の対訳データをタグで区別するようにしてもよい。また、２種類の対訳データを混合する前に、一方または両方をオーバーサンプリングまたはアンダーサンプリングすることで対訳データの混合比を変更してもよい。
【０３４３】
機械翻訳装置ＭＴ１は、上記（１）～（３）の方法のいずれかによるＮＭＴモデルの学習処理を行い、ＮＭＴモデルの最適パラメータを取得する。そして、当該最適パラメータを設定したＮＭＴモデル（学習済みモデル）を用いて、適応先分野の機械翻訳を行う。つまり、図１に示すように、機械翻訳装置ＭＴ１は、適応先分野の起点言語データＤｉｎ＿ｅを入力したとき、当該データＤｉｎ＿ｅに対して機械翻訳処理を行い、目標言語のデータＤｏｕｔ＿ｊ（機械翻訳したデータ）を取得する。
【０３４４】
図１５に、機械翻訳装置ＭＴ１により取得されたデータの一例を示す。図１５の上段は、原文（起点言語データ）であり、図１５の下段は、本発明の機械翻訳システム１０００により取得された機械翻訳データである。また、図１５の中段に、比較のため、従来技術の機械翻訳処理を行った結果データ（適応先分野の対訳データによる学習を行うことなく取得したＮＭＴモデルによる機械翻訳の結果データ）を示す。なお、図１５の下段の出力データは、他分野の対訳データとして、日英の対訳データ（適応先分野以外の対訳データ）を用い、適応先分野をＳＮＳ（Ｓｏｃｉａｌ Ｎｅｔｗｏｒｋｉｎｇ Ｓｅｒｖｉｃｅ）分野としたときのデータ（機械翻訳結果データ）である。
【０３４５】
図１５から分かるように、適応先分野なしの機械翻訳（従来の機械翻訳）により翻訳データでは、正しく翻訳できていないが、本発明の機械翻訳システム１０００により取得された機械翻訳データは従来の機械翻訳に比べ良好なものとなっている。
【０３４６】
≪まとめ≫
このように、機械翻訳システム１０００では、適応先分野（対象分野）における対訳データが存在しない場合であっても、疑似対訳データ生成装置１００により、適応先分野の疑似対訳データを生成することができ、また、疑似対訳データ生成装置１００により生成

された疑似対訳データを用いて、機械翻訳装置ＭＴ１のＮＭＴモデルを学習させることができる。さらに、機械翻訳システム１０００では、疑似対訳データ生成装置１００により生成された疑似対訳データにより学習させたＮＭＴモデルを用いて、機械翻訳処理を行うことで、適応先分野（対象分野）における対訳データが存在しない場合であっても、適応先分野における機械翻訳を精度良く行うことができる。

【０３４７】

機械翻訳システム１０００では、ＸＬＭの学習、疑似対訳データ生成用機械翻訳モデルの学習、および、機械翻訳装置ＭＴ１による適応先分野のＮＭＴモデルの学習のいずれにおいても適応先分野の正規の対訳データを一切使用していない。この点において、機械翻訳システム１０００は、従来技術と大きく相違する。なお、機械翻訳システム１０００において、疑似対訳データの中から最も信頼度が高い部分を抽出し、調整用の対訳データとして用いるようにしてもよい。

【０３４８】

また、機械翻訳システム１０００では、適応先分野（対象分野）における対訳データが存在しないが、他分野の対訳データおよび適応先分野の単言語データが大規模に存在する場合に絶大な効果を発揮する。

【０３４９】

［他の実施形態］

上記実施形態で説明した機械翻訳システム１０００、疑似対訳データ生成装置１００、機械翻訳装置ＭＴ１において、各ブロックは、ＬＳＩなどの半導体装置により個別に１チップ化されても良いし、一部または全部を含むように１チップ化されても良い。

【０３５０】

なおここではＬＳＩとしたが、集積度の違いにより、ＩＣ、システムＬＳＩ、スーパーＬＳＩ、ウルトラＬＳＩと呼称されることもある。

【０３５１】

また集積回路化の手法はＬＳＩに限るものではなく、専用回路または汎用プロセサで実現してもよい。ＬＳＩ製造後にプログラムすることが可能なＦＰＧＡ（Ｆｉｅｌｄ Ｐｒｏｇｒａｍｍａｂｌｅ Ｇａｔｅ Ａｒｒａｙ）や、ＬＳＩ内部の回路セルの接続や設定を再構成可能なリコンフィギュラブル・プロセッサーを利用しても良い。

【０３５２】

また上記各実施形態の各機能ブロックの処理の一部または全部は、プログラムにより実現されるものであってもよい。そして上記各実施形態の各機能ブロックの処理の一部または全部は、コンピュータにおいて、中央演算装置（ＣＰＵ）により行われる。また、それぞれの処理を行うためのプログラムは、ハードディスク、ＲＯＭなどの記憶装置に格納されており、ＲＯＭにおいて、あるいはＲＡＭに読み出されて実行される。

【０３５３】

また上記実施形態の各処理をハードウェアにより実現してもよいし、ソフトウェア（ＯＳ（オペレーティングシステム）、ミドルウェア、あるいは所定のライブラリとともに実現される場合を含む。）により実現してもよい。さらにソフトウェアおよびハードウェアの混在処理により実現しても良い。

【０３５４】

例えば上記実施形態の各機能部をソフトウェアにより実現する場合、図１６に示したハードウェア構成（例えばＣＰＵ（ＧＰＵであってもよい）、ＲＯＭ、ＲＡＭ、入力部、出力部、通信部、記憶部（例えば、ＨＤＤ、ＳＳＤ等により実現される記憶部）、外部メディア用ドライブ等をバスＢｕｓにより接続したハードウェア構成）を用いて各機能部をソフトウェア処理により実現するようにしてもよい。

【０３５５】

また上記実施形態の各機能部をソフトウェアにより実現する場合、当該ソフトウェアは、図１６に示したハードウェア構成を有する単独のコンピュータを用いて実現されるものであってもよいし、複数のコンピュータを用いて分散処理により実現されるものであって

もよい。
【０３５６】
　また上記実施形態における処理方法の実行順序は、必ずしも上記実施形態の記載に制限されるものではなく、発明の要旨を逸脱しない範囲で、実行順序を入れ替えることができるものである。
【０３５７】
　前述した方法をコンピュータに実行させるコンピュータプログラム、及びそのプログラムを記録したコンピュータ読み取り可能な記録媒体は、本発明の範囲に含まれる。ここでコンピュータ読み取り可能な記録媒体としては、例えば、フレキシブルディスク、ハードディスク、ＣＤ－ＲＯＭ、ＭＯ、ＤＶＤ、ＤＶＤ－ＲＯＭ、ＤＶＤ－ＲＡＭ、大容量ＤＶＤ、次世代ＤＶＤ、半導体メモリを挙げることができる。
【０３５８】
　上記コンピュータプログラムは、上記記録媒体に記録されたものに限らず、電気通信回線、無線または有線通信回線、インターネットを代表とするネットワーク等を経由して伝送されるものであってもよい。
【０３５９】
　また、本明細書内の記載、特許請求の範囲の記載において、「最適化」とは、最も良い状態にすることをいい、システム（モデル）を「最適化」するパラメータとは、当該システムの目的関数の値が最適値となるときのパラメータのことをいう。「最適値」は、システムの目的関数の値が大きくなるほど、システムが良い状態となる場合は、最大値であり、システムの目的関数の値が小さくなるほど、システムが良い状態となる場合は、最小値である。また、「最適値」は、極値であってもよい。また、「最適値」は、所定の誤差（測定誤差、量子化誤差等）を許容するものであってもよく、所定の範囲（十分収束したとみなすことができる範囲）に含まれる値であってもよい。
【０３６０】
　なお、本発明の具体的な構成は、前述の実施形態に限られるものではなく、発明の要旨を逸脱しない範囲で種々の変更および修正が可能である。
【符号の説明】
【０３６１】
１０００　機械翻訳システム
１００　疑似対訳データ生成装置
１　入力データ取得部
１１　第１入力データ取得処理部
１２　マスク化処理部
２　入力データ埋込部
２１　トークン埋込部
２２　位置埋込部
２３　言語埋込部
３　ＸＬＭ処理部
３１　ＸＬＭ用ニューラルネットワークモデル
５　機械翻訳処理部
５１　ＭＴ用ニューラルネットワークモデル
８　フィルター処理部
ＭＴ１　機械翻訳装置
ＤＢｐ（Ｌ１－Ｌ２）　対訳データ記憶部
ＤＢｍ（Ｌ１）　第１単言語データ記憶部（他分野の単言語データ用（第１言語））
ＤＢｍ（Ｌ２）　第２単言語データ記憶部（他分野の単言語データ用（第２言語））
ＤＢｍ（Ｒ１）　第３単言語データ記憶部（適応先分野の単言語データ用（第１言語））
ＤＢｍ（Ｒ２）　第４単言語データ記憶部（適応先分野の単言語データ用（第２言語））
ＤＢ１　疑似対訳データ格納部

【図面】

【図１】

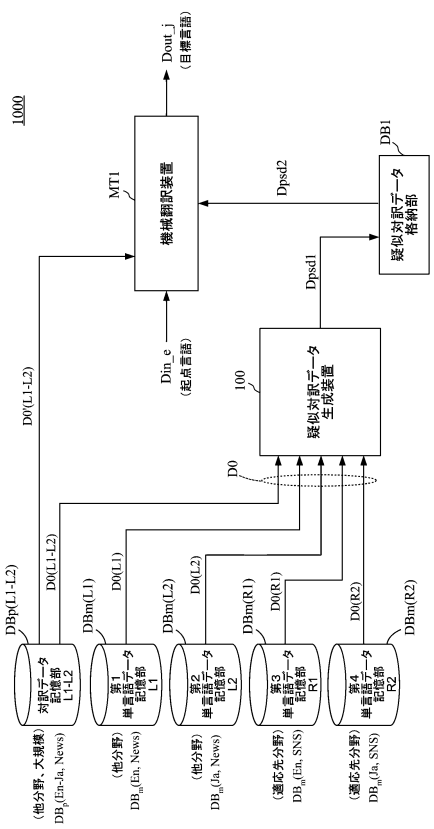

FIG. 1

【図２】

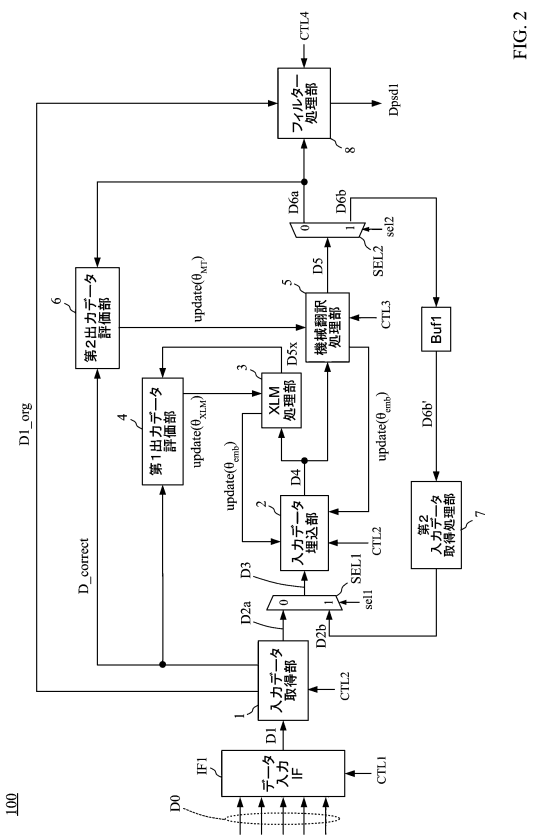

FIG. 2

【図３】

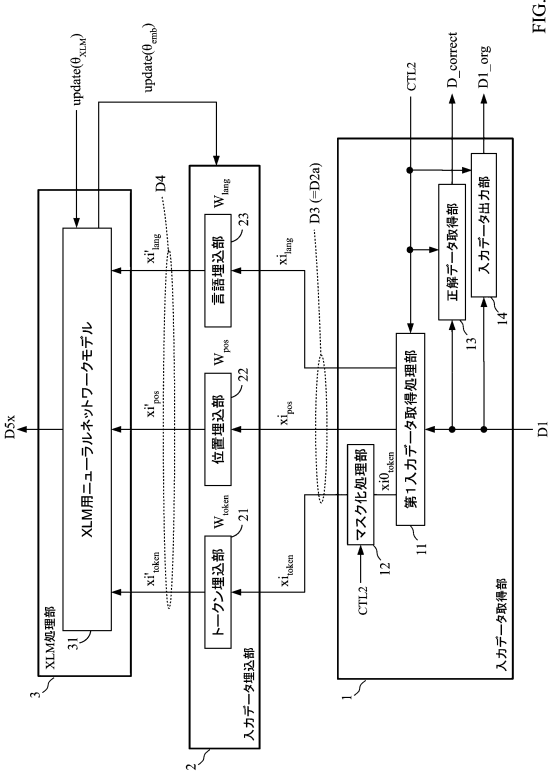

FIG. 3

【図４】

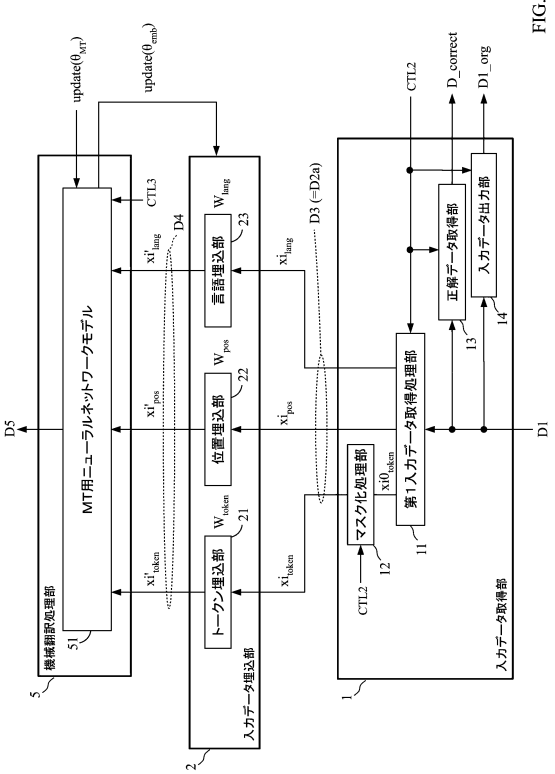

FIG. 4

【図５】

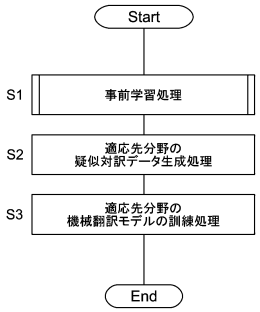

FIG. 5

【図６】

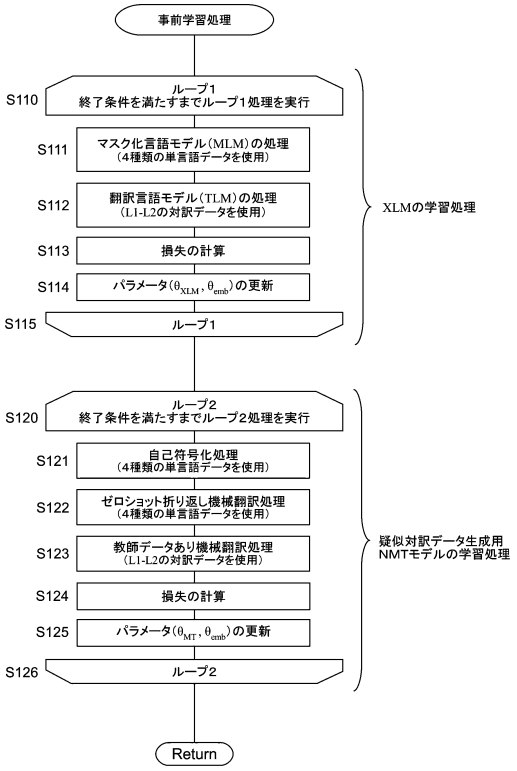

FIG. 6

【図７】

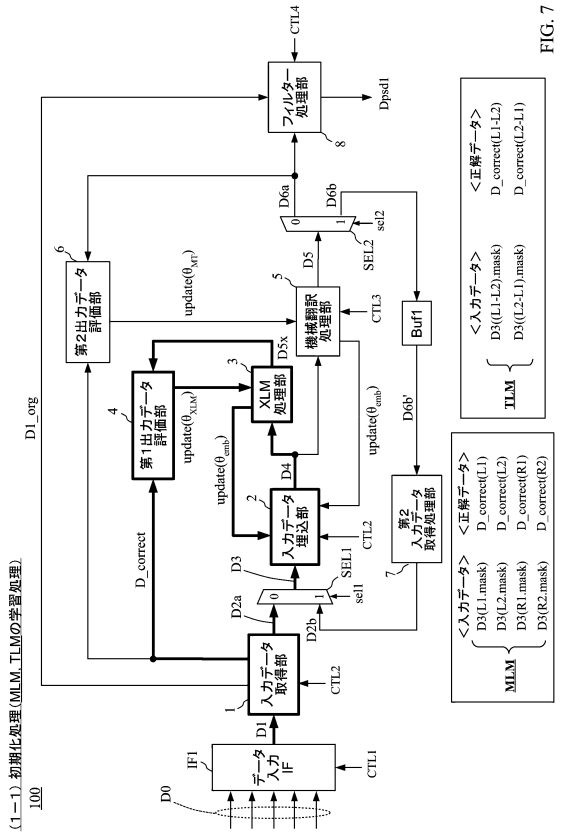

FIG. 7

【図８】

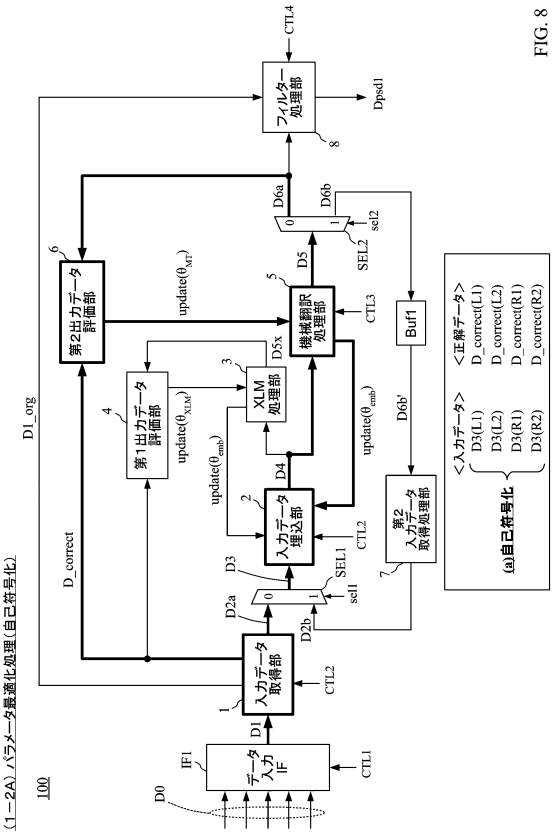

FIG. 8

【図９】

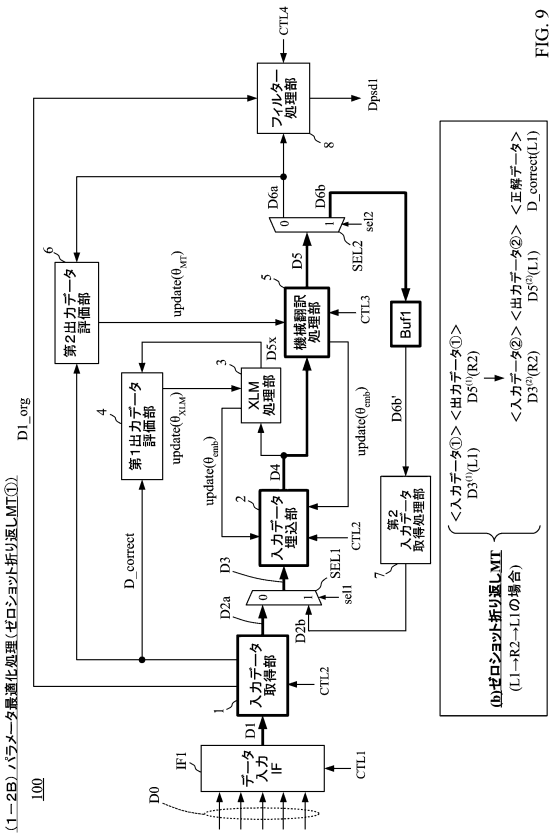

FIG. 9

【図１０】

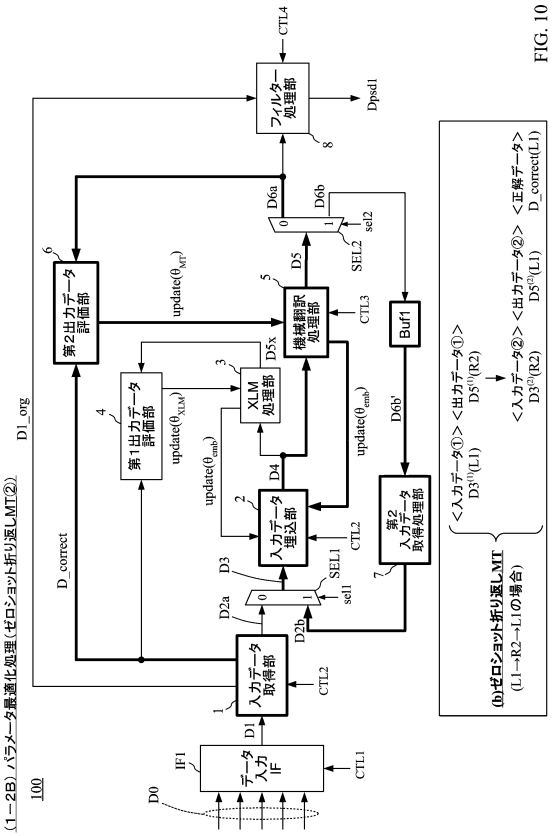

FIG. 10

【図１１】

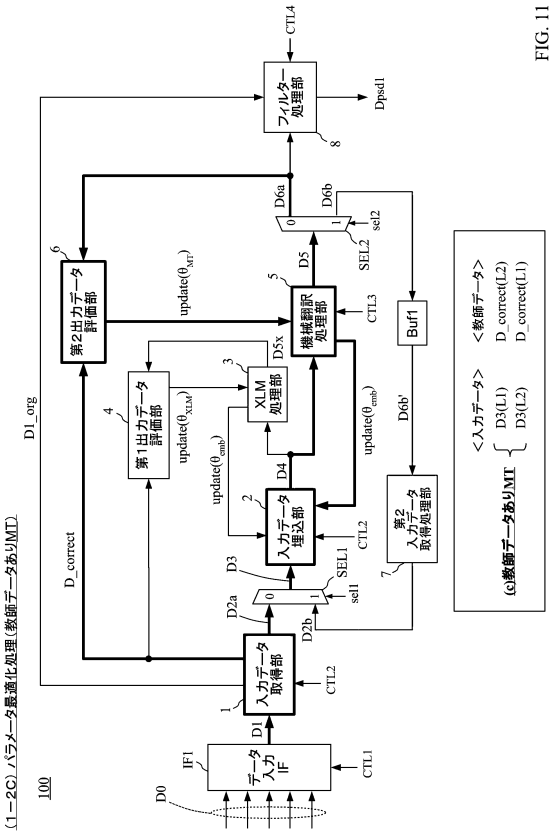

FIG. 11

【図１２】

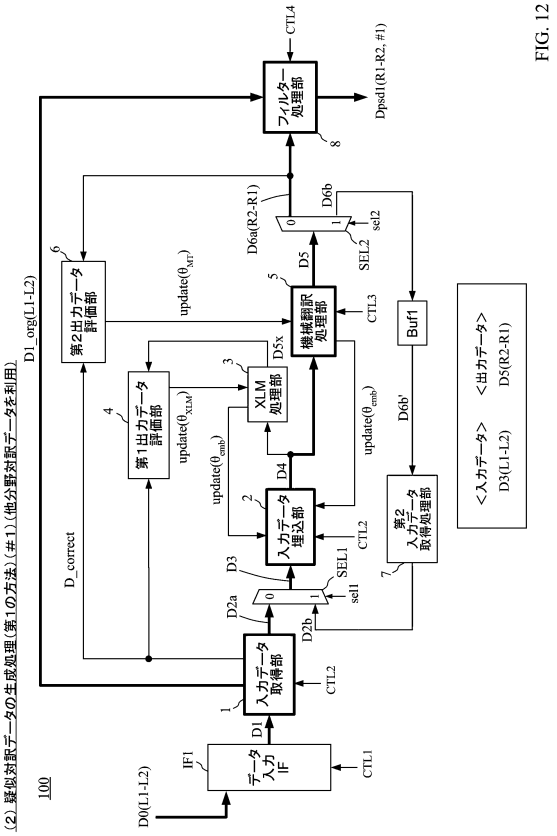

FIG. 12

【図１３】

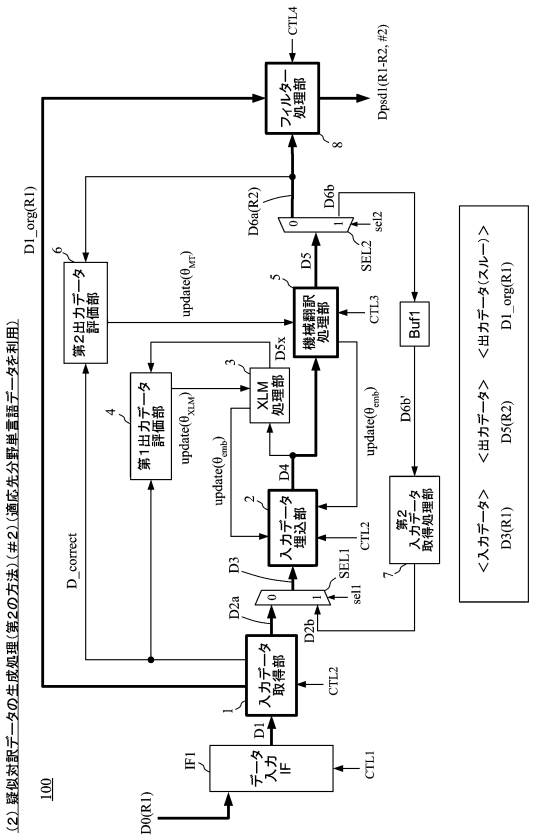

FIG. 13

【図１４】

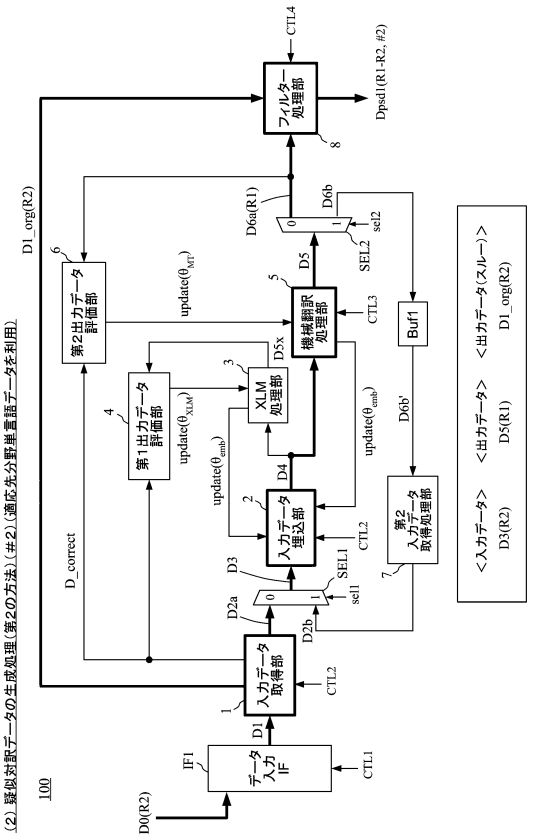

FIG. 14

【図１５】

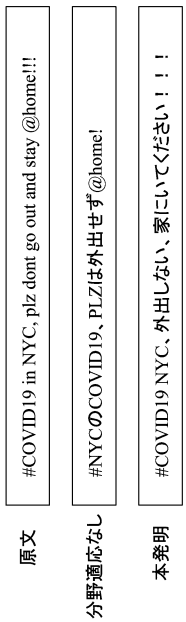

FIG. 15

【図１６】

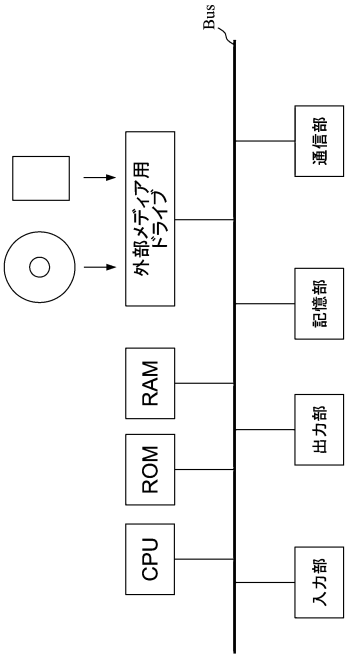

FIG. 16

フロントページの続き

東京都小金井市貫井北町４－２－１ 国立研究開発法人情報通信研究機構内

審査官 成瀬 博之

(56)参考文献 特開２０２０－１１２９１５（ＪＰ，Ａ）
特開２０１８－１１６３２４（ＪＰ，Ａ）
森田知熙 他2名，双方向ニューラル機械翻訳の反復的な教師なし適応の検討，言語処理学会第２５回年次大会 発表論文集，言語処理学会，2019年03月04日，1451-1454頁

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