(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2008-59440
(P2008-59440A)
(43)公開日　平成20年3月13日(2008.3.13)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 17/28	(2006.01)	G06F 17/28	U	5B091

審査請求 未請求　請求項の数 9　ＯＬ　（全 28 頁）

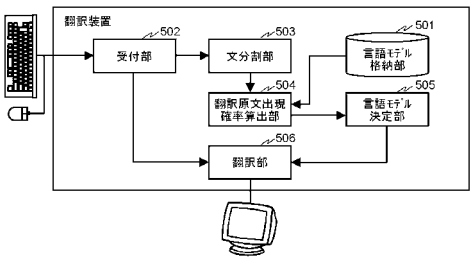
(21)出願番号　特願2006-237639 (P2006-237639)
(22)出願日　平成18年9月1日 (2006.9.1)

(71)出願人　301022471
独立行政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　100115749
弁理士　谷川　英和
(72)発明者　山本　博史
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　隅田　英一郎
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

Ｆターム(参考)　5B091 AA03　BA04　BA11　CA21　EA02

(54)【発明の名称】翻訳装置、クラスタ生成装置、クラスタの製造方法、およびプログラム

(57)【要約】
【課題】従来の翻訳装置においては、精度の高い翻訳ができない、という課題があった。
【解決手段】ｎ種類の区別された言語モデルを格納しており、翻訳対象の第一の言語の文を受け付け、当該文を１以上の用語に分割する文分割部と、各言語モデルを読み出し、当該各言語モデルを用いて、文分割部が取得した１以上の各用語が、各言語モデルが有する１以上の対訳文対中に出現する確率に関する情報である翻訳原文出現確率を、言語モデル毎に算出する翻訳原文出現確率算出部と、言語モデル毎に算出されたｎの翻訳原文出現確率を用いて、最も出現する確率が高い言語モデルを決定する言語モデル決定部と、言語モデル決定部が決定した言語モデルを読み出し、当該読み出した言語モデルを用いて、前記受付部が受け付けた文を第二の言語の文に翻訳する翻訳部を具備する翻訳装置により、精度の高い翻訳ができる。
【選択図】図６



【特許請求の範囲】
【請求項１】
第一の言語の文と、当該文の第二の言語への翻訳文の対の情報である対訳文対を複数格納している対訳文対格納部と、
前記対訳文対格納部から複数の対訳文対を読み出し、当該複数の対訳文対をｎ個のバッファに配置する対訳文対配置部と、
前記バッファ毎に、前記対訳文対配置部が配置する各バッファ中の１以上の対訳文対を１以上の用語に分割し、当該１以上の対訳文対中に用語が出現する確率についての情報である確率情報を取得し、用語と当該用語に対応する確率情報を有する用語出現確率情報を１以上有する情報である言語モデルを取得し、記録媒体上に配置する言語モデル取得部と、
前記言語モデル取得部が取得した１以上の用語出現確率情報が有する１以上の確率情報を用いて、前記ｎ個のバッファ毎に、用語の出現の均一具合についての情報であるｎのエントロピーを算出し、記憶媒体に配置するエントロピー算出部と、
前記ｎのエントロピーを取得し、前記ｎ個のバッファ全体の用語の出現の均一具合についての情報である総エントロピーを算出し、記憶媒体に配置する総エントロピー算出部と、
前記ｎ個のバッファのうちのいずれかのバッファ中のいずれかの対訳文対を読み出し、他の各バッファに移動する対訳文対移動部と、
前記対訳文対移動部が対訳文対を各バッファに移動した後、バッファごとに、前記言語モデル取得部に前記言語モデルを取得し、記録媒体上に配置するように指示し、前記エントロピー算出部に前記ｎのエントロピーを算出し、記憶媒体に配置するように指示し、および前記総エントロピー算出部に対して総エントロピーを算出し、記憶媒体に配置するように指示する第一制御部と、
前記第一制御部の制御に対応して、バッファごとに、得られたｎの総エントロピーを取得し、当該ｎの総エントロピーのうちで最も小さい総エントロピーに対応するバッファに、当該移動対象の対訳文対の移動先のバッファを決定し、当該バッファに前記移動対象の対訳文対を書き込む対訳文対移動先決定部と、
前記対訳文対移動先決定部が、全対訳文対について移動先を決定した後の最近の総エントロピーと、その前のサイクルにおいて、前記対訳文対移動先決定部が全対訳文対について、移動先を決定した後の総エントロピーである直前の総エントロピーを用いて、エントロピーの変化量を算出し、記録媒体に配置する変化量算出部と、
前記変化量算出部が算出した変化量が閾値より小さいか否か、または閾値以下であるか否かを判断する変化判断部と、
前記変化判断部が、変化量が閾値より小さい、または閾値以下であると判断するまで、前記対訳文対移動部、前記第一制御部および前記対訳文対移動先決定部に当該各部の処理を繰り返させる第二制御部と、
前記対訳文対移動先決定部が最後にバッファに対訳文対を書き込んだ後の前記ｎ個のバッファ内の対訳文対のｎ種類の集合を、ｎ種類に区別して蓄積するクラスタ蓄積部を具備するクラスタ生成装置。
【請求項２】
前記言語モデル取得部が取得する確率情報は、
１以上の対訳文対中に一の用語が出現する確率である請求項１記載のクラスタ生成装置。
【請求項３】
ｎ（ｎは２以上の整数）種類の区別された言語モデルであり、用語および当該用語が１以上の対訳文対中に出現する確率についての情報である確率情報を用語毎に有する言語モデルを格納している言語モデル格納部と、
翻訳対象の第一の言語の文を受け付ける受付部と、
前記受付部が受け付けた文を取得し、当該文を１以上の用語に分割し、記憶媒体に配置する文分割部と、
前記言語モデル格納部の各言語モデルを読み出し、当該各言語モデルを用いて、前記文分割部が取得した１以上の各用語が、各言語モデルが有する１以上の対訳文対中に出現する

確率に関する情報である翻訳原文出現確率を、言語モデル毎に算出し、記憶媒体に配置する翻訳原文出現確率算出部と、
前記言語モデル毎に算出されたｎの翻訳原文出現確率を用いて、最も出現する確率が高い言語モデルを決定する言語モデル決定部と、
前記言語モデル決定部が決定した言語モデルを、前記言語モデル格納部から読み出し、当該読み出した言語モデルを用いて、前記前記受付部が受け付けた文を第二の言語の文に翻訳し、当該翻訳結果を出力する翻訳部を具備する翻訳装置。
【請求項４】
前記言語モデル格納部が格納しているｎ種類の区別された各言語モデルは、
請求項１または請求項２記載のクラスタ生成装置が蓄積したｎ種類の各対訳文対の集合から構成された情報であり、ｎ種類の各対訳文対の集合が有する各対訳文対を１以上の用語に分割し、当該１以上の用語が対訳文対の集合中に出現する確率についての情報である確率情報を用語毎に算出されることにより得られた情報である請求項３記載の翻訳装置。
【請求項５】
前記確率情報は、
単語３－ｇｒａｍの確率である請求項３または請求項４記載の翻訳装置。
【請求項６】
前記翻訳部は、
前記言語モデル決定部が決定した言語モデル（Ｔ）を、前記言語モデル格納部から読み出し、当該読み出した言語モデル（Ｔ）において、Ｐ（ｅ｜ｆ，Ｔ）Ｐ（ｆ｜Ｔ）［ｅは入力された翻訳対象の文、ｆは目的言語の文］を最大にする第二の言語の文（ｆ）を、前記言語モデル（Ｔ）が有する１以上の対訳文対が有する第二の言語の翻訳文から選択し、出力する請求項３から請求項５いずれか記載の翻訳装置。
【請求項７】
第一の言語の文と、当該文の第二の言語への翻訳文の対の情報である対訳文対を複数、記憶媒体に格納しており、
コンピュータに、
前記記憶媒体から複数の対訳文対を読み出し、当該複数の対訳文対をｎ個のバッファに配置する対訳文対配置ステップと、
前記バッファ毎に、前記対訳文対配置ステップが配置する各バッファ中の１以上の対訳文対を１以上の用語に分割し、当該１以上の対訳文対中に用語が出現する確率についての情報である確率情報を取得し、用語と当該用語に対応する確率情報を有する用語出現確率情報を１以上有する情報である言語モデルを取得し、記録媒体上に配置する言語モデル取得ステップと、
前記言語モデル取得ステップで取得した１以上の用語出現確率情報が有する１以上の確率情報を用いて、前記ｎ個のバッファ毎に、用語の出現の均一具合についての情報であるｎのエントロピーを算出し、記憶媒体に配置するエントロピー算出ステップと、
前記ｎのエントロピーを取得し、前記ｎ個のバッファ全体の用語の出現の均一具合についての情報である総エントロピーを算出し、記憶媒体に配置する総エントロピー算出ステップと、
前記ｎ個のバッファのうちのいずれかのバッファ中のいずれかの対訳文対を読み出し、他の各バッファに移動する対訳文対移動ステップと、
前記対訳文対移動ステップで対訳文対を各バッファに移動した後、バッファごとに、前記言語モデルを取得し、記録媒体上に配置させ、前記ｎのエントロピーを算出し、記憶媒体に配置させ、および総エントロピーを算出し、記憶媒体に配置させる第一制御ステップと、
前記第一制御ステップにおける処理に対応して、バッファごとに、得られたｎの総エントロピーを取得し、当該ｎの総エントロピーのうちで最も小さい総エントロピーに対応するバッファに、当該移動対象の対訳文対の移動先のバッファを決定し、当該バッファに前記移動対象の対訳文対を書き込む対訳文対移動先決定ステップと、

前記対訳文対移動先決定ステップにおいて、全対訳文対について移動先を決定した後の最近の総エントロピーと、その前のサイクルにおいて、前記対訳文対移動先決定ステップで全対訳文対について、移動先を決定した後の総エントロピーである直前の総エントロピーを用いて、エントロピーの変化量を算出し、記録媒体に配置する変化量算出ステップと、
前記変化量算出ステップで算出した変化量が閾値より小さいか否か、または閾値以下であるか否かを判断する変化判断ステップと、
前記変化判断ステップで、変化量が閾値より小さい、または閾値以下であると判断するまで、前記対訳文対移動ステップにおける処理、前記第一制御ステップにおける処理、および前記対訳文対移動先決定ステップにおける処理を繰り返させ、
前記対訳文対移動先決定ステップにおいて最後にバッファに対訳文対を書き込んだ後の前記ｎ個のバッファ内の対訳文対のｎ種類の集合を、ｎ種類に区別して蓄積するクラスタ蓄積ステップを実行させるためのプログラム。

【請求項８】

ｎ（ｎは２以上の整数）種類の区別された言語モデルであり、用語および当該用語が１以上の対訳文対中に出現する確率についての情報である確率情報を用語毎に有する言語モデルを記録媒体に格納しており、
コンピュータに、
翻訳対象の第一の言語の文を受け付ける受付ステップと、
前記受付ステップで受け付けた文を取得し、当該文を１以上の用語に分割し、記憶媒体に配置する文分割ステップと、
前記記録媒体の各言語モデルを読み出し、当該各言語モデルを用いて、前記文分割ステップで取得した１以上の各用語が、各言語モデルが有する１以上の対訳文対中に出現する確率に関する情報である翻訳原文出現確率を、言語モデル毎に算出し、記憶媒体に配置する翻訳原文出現確率算出ステップと、
前記言語モデル毎に算出されたｎの翻訳原文出現確率を用いて、最も出現する確率が高い言語モデルを決定する言語モデル決定ステップと、
前記言語モデル決定ステップで決定した言語モデルを、前記記録媒体から読み出し、当該読み出した言語モデルを用いて、前記前記受付ステップで受け付けた文を第二の言語の文に翻訳し、当該翻訳結果を出力する翻訳ステップを実行させるためのプログラム。

【請求項９】

ｎ種類の分類された1以上の対訳文対の集合からなるｎのクラスタを製造する方法であって、
第一の言語の文と、当該文の第二の言語への翻訳文の対の情報である対訳文対を複数、記憶媒体に格納しており、
前記記憶媒体から複数の対訳文対を読み出し、当該複数の対訳文対をｎ個のバッファに配置する対訳文対配置ステップと、
前記バッファ毎に、前記対訳文対配置ステップで配置する各バッファ中の１以上の対訳文対を１以上の用語に分割し、当該１以上の対訳文対中に用語が出現する確率についての情報である確率情報を取得し、用語と当該用語に対応する確率情報を有する用語出現確率情報を１以上有する情報である言語モデルを取得し、記録媒体上に配置する言語モデル取得ステップと、
前記言語モデル取得ステップで取得した１以上の用語出現確率情報が有する１以上の確率情報を用いて、前記ｎ個のバッファ毎に、用語の出現の均一具合についての情報であるｎのエントロピーを算出し、記憶媒体に配置するエントロピー算出ステップと、
前記ｎのエントロピーを取得し、前記ｎ個のバッファ全体の用語の出現の均一具合についての情報である総エントロピーを算出し、記憶媒体に配置する総エントロピー算出ステップと、
前記ｎ個のバッファのうちのいずれかのバッファ中のいずれかの対訳文対を読み出し、他の各バッファに移動する対訳文対移動ステップと、
前記対訳文対移動ステップで対訳文対を各バッファに移動した後、バッファごとに、前記

言語モデルを取得し、記録媒体上に配置させ、前記ｎのエントロピーを算出し、記憶媒体に配置させ、および総エントロピーを算出し、記憶媒体に配置させる第一制御ステップと、
前記第一制御ステップにおける処理に対応して、バッファごとに、得られたｎの総エントロピーを取得し、当該ｎの総エントロピーのうちで最も小さい総エントロピーに対応するバッファに、当該移動対象の対訳文対の移動先のバッファを決定し、当該バッファに前記移動対象の対訳文対を書き込む対訳文対移動先決定ステップと、
前記対訳文対移動先決定ステップにおいて、全対訳文対について移動先を決定した後の最近の総エントロピーと、その前のサイクルにおいて、前記対訳文対移動先決定ステップで全対訳文対について、移動先を決定した後の総エントロピーである直前の総エントロピーを用いて、エントロピーの変化量を算出し、記録媒体に配置する変化量算出ステップと、
前記変化量算出ステップで算出した変化量が閾値より小さいか否か、または閾値以下であるか否かを判断する変化判断ステップと、
前記変化判断ステップで、変化量が閾値より小さい、または閾値以下であると判断するまで、前記対訳文対移動ステップにおける処理、前記第一制御ステップにおける処理、および前記対訳文対移動先決定ステップにおける処理を繰り返させ、
前記対訳文対移動先決定ステップにおいて最後にバッファに対訳文対を書き込んだ後の前記ｎ個のバッファ内の対訳文対のｎ種類の集合を、ｎ種類に区別して蓄積するクラスタ蓄積ステップを具備するクラスタの製造方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、自然言語の翻訳を行う翻訳装置等に関するものである。

【背景技術】

【０００２】

近年、Ｎ－ｇｒａｍに代表される統計言語モデルは統計翻訳をはじめとする言語処理において広く用いられている。統計言語モデルはその性格上、学習データと異なるタスクに対しては性能が劣化してしまう。

【０００３】

また、「タスク適応」という考え方がある。「タスク適応」は、特定のタスクに特化したモデルであるタスク依存モデルを作成することが目的である。ここで、タスクとは、例えば、所定の話題や、分類するための指標（例えば、旅行の会話集など）などである。

【０００４】

また、対象のタスクが既知である場合には、あらかじめ「タスク適応」を用いてタスク依存モデルを作成しておき、それを統計翻訳に利用することができる。

【０００５】

従来の翻訳装置において、入力された第１自然言語表現の翻訳として適切な第２自然言語表現を選択することを可能にする翻訳装置があった（例えば、特許文献１参照）。かかる翻訳装置において、第１、第２自然言語共起語収集部は、第１、第２自然言語コーパスを検索してそれぞれ第１、第２自然言語の表現に共起する語および共起語毎の統計情報を取得する。第１、第２自然言語共起情報解析部は、それぞれ受け取った共起語およびその統計情報を用いて各共起語の特徴量を計算し、第１、第２の自然言語共起情報として出力する。共起情報比較部は、第１自然言語共起情報と第２自然言語共起情報と対訳辞書を使用して、第１自然言語表現と第２自然言語表現の全ての組み合わせについてその意味的な類似度を計算し、翻訳候補選択部に送る。翻訳候補選択部は、入力された第１自然言語表現に対して意味的な類似度の最も高い第２自然言語表現を選択して出力する。

【０００６】

また、従来の他の翻訳装置において、対訳フレーズを利用した統計機械翻訳装置において、より高い精度で翻訳を行うことができる装置があった（例えば、特許文献２参照）。かかる統計機械翻訳装置において、日英機械翻訳のデコーダは、日本語フレーズＮグラム

モデル、英語フレーズＮグラムモデル、英語言語モデル、および英語から日本語へのフレーズ翻訳モデルと、日本語の入力文に対し可能な全てのセグメンテーションを行なうセグメンテーション処理部と、得られたセグメンテーションにしたがい、日本語フレーズＮグラムモデル、英語フレーズＮグラムモデル、英語言語モデル、および英語から日本語へのフレーズ翻訳モデルを用い、英語のフレーズを任意の順序で確率付きで並べたフレーズシーケンスを表すラッティスを作成するラッティス作成部と、ラッティス作成部が作成したラッティスのうちで最も確率の高い上位Ｍ個の経路を探索して出力するＡ＊探索処理部とを含む装置である。

【０００７】

なお、本発明に関連する技術として、非特許文献１、非特許文献２に記述された技術がある。

【特許文献１】特開２００２－３５１８７２号公報（第１頁、第１図等）

【特許文献２】特開２００６－０９９２０８号公報（第１頁、第１図等）

【非特許文献１】S. M. Katz, "Estimation of Probabilities from Sparse Data for Language Model Component of a Speech Recognizer,", IEEE Trans. on Acoustics, Speech, and Signal Processing, pp. 400-401, 1987.

【非特許文献２】K. Seymore, R. Rosenfeld,"Using Story Topics for Language Model Adaptation,", Proc. EUROSPEECH, pp. 1987-1990, 1997.

【発明の開示】

【発明が解決しようとする課題】

【０００８】

しかしながら、タスクをあらかじめ想定しておくことが困難な場合も多く、このような場合は通常の「タスク適応」の手法を用いることはできない。

【０００９】

また、従来の翻訳装置においては、用意したモデルのタスクが、翻訳対象の文にマッチしない場合に、翻訳の性能が著しく劣化する、という課題があった。

【課題を解決するための手段】

【００１０】

本第一の発明のクラスタ生成装置は、第一の言語の文と、当該文の第二の言語への翻訳文の対の情報である対訳文対を複数格納している対訳文対格納部と、前記対訳文対格納部から複数の対訳文対を読み出し、当該複数の対訳文対をｎ個のバッファに配置する対訳文対配置部と、前記バッファ毎に、前記対訳文対配置部が配置する各バッファ中の１以上の対訳文対を１以上の用語に分割し、当該１以上の対訳文対中に用語が出現する確率についての情報である確率情報を取得し、用語と当該用語に対応する確率情報を有する用語出現確率情報を１以上有する情報である言語モデルを取得し、記録媒体上に配置する言語モデル取得部と、前記言語モデル取得部が取得した１以上の用語出現確率情報が有する１以上の確率情報を用いて、前記ｎ個のバッファ毎に、用語の出現の均一具合についての情報であるｎのエントロピーを算出し、記憶媒体に配置するエントロピー算出部と、前記ｎのエントロピーを取得し、前記ｎ個のバッファ全体の用語の出現の均一具合についての情報である総エントロピーを算出し、記憶媒体に配置する総エントロピー算出部と、前記ｎ個のバッファのうちのいずれかのバッファ中のいずれかの対訳文対を読み出し、他の各バッファに移動する対訳文対移動部と、前記対訳文対移動部が対訳文対を各バッファに移動した後、バッファごとに、前記言語モデル取得部に前記言語モデルを取得し、記録媒体上に配置するように指示し、前記エントロピー算出部に前記ｎのエントロピーを算出し、記憶媒体に配置するように指示し、および前記総エントロピー算出部に対して総エントロピーを算出し、記憶媒体に配置するように指示する第一制御部と、前記第一制御部の制御に対応して、バッファごとに、得られたｎの総エントロピーを取得し、当該ｎの総エントロピーのうちで最も小さい総エントロピーに対応するバッファに、当該移動対象の対訳文対の移動先のバッファを決定し、当該バッファに前記移動対象の対訳文対を書き込む対訳文対移動先決定部と、前記対訳文対移動先決定部が決定した後の最近の総エントロピーと、その

前に前記対訳文対移動先決定部が決定した後の直前の総エントロピーを用いて、エントロピーの変化量を算出し、記録媒体に配置する変化量算出部と、前記変化量算出部が算出した変化量が閾値より小さいか否か、または閾値以下であるか否かを判断する変化判断部と、前記変化判断部が、変化量が閾値より小さい、または閾値以下であると判断するまで、前記対訳文対移動部、前記第一制御部および前記対訳文対移動先決定部に当該各部の処理を繰り返させる第二制御部と、前記対訳文対移動先決定部が最後にバッファに対訳文対を書き込んだ後の前記ｎ個のバッファ内の対訳文対のｎ種類の集合を、ｎ種類に区別して蓄積するクラスタ蓄積部を具備するクラスタ生成装置である。

【００１１】

かかる構成により、自動的にクラスタを生成できる。ここで、クラスタとは、複数の文を、タスクごとに分類した情報である。また、当該クラスタを用いて、精度の高い機械翻訳が可能となる。

【００１２】

また、本第二の発明のクラスタ生成装置は、第一の発明に対して、前記言語モデル取得部が取得する確率情報は、１以上の対訳文対中に一の用語が出現する確率であるクラスタ生成装置である。

【００１３】

かかる構成により、高速に自動的にクラスタを生成できる。

【００１４】

また、本第三の発明の翻訳装置は、ｎ（ｎは２以上の整数）種類の区別された言語モデルであり、用語および当該用語が１以上の対訳文対中に出現する確率についての情報である確率情報を用語毎に有する言語モデルを格納している言語モデル格納部と、翻訳対象の第一の言語の文を受け付ける受付部と、前記受付部が受け付けた文を取得し、当該文を１以上の用語に分割し、記憶媒体に配置する文分割部と、前記言語モデル格納部の各言語モデルを読み出し、当該各言語モデルを用いて、前記文分割部が取得した１以上の各用語が、各言語モデルが有する１以上の対訳文対中に出現する確率に関する情報である翻訳原文出現確率を、言語モデル毎に算出し、記憶媒体に配置する翻訳原文出現確率算出部と、前記言語モデル毎に算出されたｎの翻訳原文出現確率を用いて、最も出現する確率が高い言語モデルを決定する言語モデル決定部と、前記言語モデル決定部が決定した言語モデルを、前記言語モデル格納部から読み出し、当該読み出した言語モデルを用いて、前記前記受付部が受け付けた文を第二の言語の文に翻訳し、当該翻訳結果を出力する翻訳部を具備する翻訳装置である。

【００１５】

かかる構成により、精度の高い機械翻訳が可能となる。

【００１６】

また、本第四の発明の翻訳装置における前記言語モデル格納部が格納しているｎ種類の区別された各言語モデルは、第一または第二のクラスタ生成装置が蓄積したｎ種類の各対訳文対の集合から構成された情報であり、ｎ種類の各対訳文対の集合が有する各対訳文対を１以上の用語に分割し、当該１以上の用語が対訳文対の集合中に出現する確率についての情報である確率情報を用語毎に算出されることにより得られた情報である。

【００１７】

かかる構成により、効率的に精度の高い機械翻訳が可能となる。

【００１８】

また、本第五の発明の翻訳装置における前記確率情報は、第三、第四いずれかの発明に対して、単語３－ｇｒａｍの確率である翻訳装置である。

【００１９】

かかる構成により、さらに精度の高い機械翻訳が可能となる。

【００２０】

また、本第六の発明の翻訳装置における前記翻訳部は、第三から第五いずれかの発明に対して、前記言語モデル決定部が決定した言語モデル（Ｔ）を、前記言語モデル格納部か

ら読み出し、当該読み出した言語モデル（Ｔ）において、Ｐ（ｅ｜ｆ，Ｔ）Ｐ（ｆ｜Ｔ）［ｅは入力された翻訳対象の文、ｆは目的言語の文］を最大にする第二の言語の文（ｆ）を、前記言語モデル（Ｔ）が有する１以上の対訳文対が有する第二の言語の翻訳文から選択する翻訳装置である。

【００２１】

かかる構成により、精度の高い機械翻訳が可能となる。

【発明の効果】

【００２２】

本発明による翻訳装置によれば、精度の高い機械翻訳が可能となる。

【発明を実施するための最良の形態】

【００２３】

以下、クラスタ生成装置、翻訳装置等の実施形態について図面を参照して説明する。なお、実施の形態において同じ符号を付した構成要素は同様の動作を行うので、再度の説明を省略する場合がある。

（実施の形態１）

【００２４】

図１は、本実施の形態におけるクラスタ生成装置のブロック図である。

【００２５】

クラスタ生成装置は、対訳文対格納部１０１、対訳文対配置部１０２、言語モデル取得部１０３、エントロピー算出部１０４、総エントロピー算出部１０５、対訳文対移動部１０６、第一制御部１０７、対訳文対移動先決定部１０８、変化量算出部１０９、変化判断部１１０、第二制御部１１１、クラスタ蓄積部１１２を具備する。

【００２６】

対訳文対格納部１０１は、第一の言語の文と、当該文の第二の言語への翻訳文の対の情報である対訳文対を複数格納している。第一の言語の文とは、例えば、日本語の文の情報である。また、第二の言語への翻訳文は、例えば、英語の文の情報である。対訳文対は、例えば、日本語の文の情報と英語の文の情報を有する。対訳文対のデータ構造は、問わない。対訳文対は、バッファに連続して、日本語の文の情報と英語の文の情報を有しても良いし、日本語の文の情報と英語の文の情報が異なるバッファに存在し、リンクにより関連付けられていても良い。対訳文対格納部１０１は、通常、大量の対訳文対（例えば、５０万の対の情報）を有する。対訳文対格納部１０１は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。

【００２７】

対訳文対配置部１０２は、対訳文対格納部１０１から複数の対訳文対を読み出し、当該複数の対訳文対をｎ個（ｎは２以上の整数）のバッファに配置する。各バッファは、メインメモリやキャッシュなどに構成されていても良いし、ハードディスクやＤＶＤなどの記録媒体に構成されていても良い。各バッファは、通常、連続する記憶領域で構成されるが、不連続な記憶領域でも良い。対訳文対配置部１０２は、対訳文対格納部１０１から多数の対訳文対を読み出し、例えば、一のバッファに、ｎに対訳文対を分類して（各対訳文対に１からｎの符号を付しても良い）、配置しても良い。かかる場合も、複数の対訳文対をｎ個のバッファに配置したことと同義である、とする。つまり、「複数の対訳文対をｎ個のバッファに配置する」とは、ｎに対訳文対を分類することを示す。対訳文対配置部１０２は、通常、ランダムに、複数の対訳文対をｎ個のバッファに配置する。ただし、対訳文対配置部１０２は、何らかのアルゴリズムに基づいて、複数の対訳文対をｎ個のバッファに配置しても良い。何らかのアルゴリズムとは、例えば、対訳文対を数値に変換し、当該数値をｎで割った余りのＩＤで識別されるバッファに配置する、などである。対訳文対配置部１０２は、通常、ＭＰＵやメモリ等から実現され得る。対訳文対配置部１０２の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００２８】

言語モデル取得部１０３は、バッファ毎に、対訳文対配置部１０２が配置する各バッファ中の１以上の対訳文対を１以上の用語に分割し、当該１以上の対訳文対中に用語が出現する確率についての情報である確率情報を取得し、用語と当該用語に対応する確率情報を有する用語出現確率情報を１以上有する情報である言語モデルを取得し、記録媒体上に配置する。なお、対訳文対は、予めスペース等の区切り文字で区切られていても良く、かかる場合、対訳文対を１以上の用語に分割する処理は、例えば、用語を順次、読み出す処理である。また、対訳文対を１以上の用語に分割する処理は、形態素解析等の言語処理により、単語に分割する処理でも良い。確率情報は、例えば、１以上の対訳文対中に一の用語が出現する確率である。また、確率情報は、例えば、条件付確率でも良い。条件付確率とは、例えば、単語３－ｇｒａｍである。言語モデル取得部１０３は、バッファ毎に、言語モデルを取得し、記録媒体上に配置するので、ｎの言語モデルが構成される。言語モデル取得部１０３は、通常、ＭＰＵやメモリ等から実現され得る。言語モデル取得部１０３の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００２９】

エントロピー算出部１０４は、言語モデル取得部１０３が取得した１以上の用語出現確率情報が有する１以上の確率情報を用いて、ｎ個のバッファ毎に、用語の出現の均一具合についての情報であるｎのエントロピーを算出し、記憶媒体に配置する。出現の均一具合が算出される対象の用語は、各バッファに存在する１以上の対訳文対を構成する複数の用語である。エントロピーは、情報源を観測したときに得られる情報量の期待値のことであり、ここでは、用語の出現の均一具合を示す情報である。エントロピーが小さい値をとるほど、用語が偏って出現していることを示す。エントロピー算出部１０４は、例えば、以下の数式１に示す演算式を用いてエントロピーを算出する。つまり、エントロピー算出部１０４は、例えば、バッファに存在する各用語について、出現確率Ｐを「当該用語が出現する回数／用語の出現回数の和」により算出し、出現確率Ｐから「－ｌｏｇＰ」を演算し、取得する。そして、エントロピー算出部１０４は、全用語の「－ｌｏｇＰ」の和を取得し、エントロピーとする。なお、エントロピー算出部１０４は、「当該用語が出現する回数／用語の出現回数の和」、「－ｌｏｇＰ」などの演算式の情報を予め格納しており、かかる演算式の情報を読み込み、値を代入して、結果を得る。エントロピー算出部１０４は、通常、ＭＰＵやメモリ等から実現され得る。エントロピー算出部の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。また、エントロピーは、通常、情報の均一度合いを示す情報であるが、エントロピー算出部１０４は、その裏返しの情報のばらつき具合を示す情報を算出しても良い。かかる場合も、エントロピー算出部１０４は、エントロピーを算出する、とする。

【００３０】

総エントロピー算出部１０５は、エントロピー算出部１０４が算出し、記憶媒体に配置したｎのエントロピーを取得し、ｎ個のバッファ全体の用語の出現の均一具合を示す情報である総エントロピーを算出し、記憶媒体に配置する。総エントロピー算出部１０５は、通常、ｎのエントロピーの和を算出し、記憶媒体に配置する。総エントロピー算出部１０５は、通常、ＭＰＵやメモリ等から実現され得る。総エントロピー算出部１０５の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００３１】

対訳文対移動部１０６は、ｎ個のバッファのうちのいずれかのバッファ中のいずれかの対訳文対を読み出し、他の各バッファに移動する。対訳文対移動部１０６は、通常、順に、移動対象の対訳文対を選択していく。対訳文対移動部１０６は、通常、ＭＰＵやメモリ等から実現され得る。対訳文対移動部１０６の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００３２】
第一制御部１０７は、対訳文対移動部１０６が対訳文対を各バッファに移動した後、対訳文対が移動したバッファごとに、言語モデル取得部１０３に言語モデルを取得し、記録媒体上に配置するように指示し、エントロピー算出部１０４にｎのエントロピーを算出し、記憶媒体に配置するように指示し、および総エントロピー算出部１０５に対して総エントロピーを算出し、記憶媒体に配置するように指示する。第一制御部１０７は、対訳文対移動部１０６が一の対訳文対を各バッファに移動する毎に、上記の処理（各部に対する指示）を行う。第一制御部１０７は、通常、ＭＰＵやメモリ等から実現され得る。第一制御部１０７の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。なお、クラスタ生成装置がソフトウェアで実現される場合、第一制御部１０７は、単に、言語モデル取得部１０３に対応する関数を呼び出したり、エントロピー算出部１０４に対応する関数を呼び出したり、総エントロピー算出部１０５に対応する関数を呼び出したりするだけでも良い。

【００３３】
対訳文対移動先決定部１０８は、第一制御部１０７の制御に対応して、バッファごとに、得られたｎの総エントロピーを取得し、当該ｎの総エントロピーのうちで最も小さい総エントロピーに対応するバッファ（最も用語の出現の均一具合の小さい場合に、対訳文対が存在するバッファ）に、当該移動対象の対訳文対の移動先のバッファを決定し、当該バッファに前記移動対象の対訳文対を書き込む。ここで、すでに移動対象の対訳文対が、決定されたバッファに書き込まれている場合は、この書き込み処理は省略される。上記の対訳文対移動先決定部１０８の処理は、言い換えれば、全ての対訳文対一つ一つに対して、総エントロピーが最小となるようなバッファへの移動を行う処理である。対訳文対移動先決定部１０８は、通常、ＭＰＵやメモリ等から実現され得る。対訳文対移動先決定部１０８の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００３４】
変化量算出部１０９は、対訳文対移動先決定部１０８が全対訳文対について、移動先を決定した後の最近の総エントロピーと、その前のサイクル（全対訳文対について、移動先を決定する処理を一サイクル、とする。）において、対訳文対移動先決定部１０８が全対訳文対について、移動先を決定した後の総エントロピー（直前の総エントロピー）を用いて、エントロピーの変化量を算出し、記録媒体に配置する。変化量算出部１０９は、通常、「｜最近の総エントロピー　－　直前の総エントロピー｜」により、変化量を算出する。変化量算出部１０９は、通常、ＭＰＵやメモリ等から実現され得る。変化量算出部１０９の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００３５】
変化判断部１１０は、変化量算出部１０９が算出した変化量が閾値より小さいか否か、または閾値以下であるか否かを判断する。なお、変化判断部１１０は、予め閾値を格納している。そして、変化判断部１１０は、閾値を読み出し、変化量算出部１０９が算出した変化量が閾値より小さいか否か、または閾値以下であるか否かを判断し、判断結果を第二制御部１１１に渡す。変化判断部１１０は、通常、ＭＰＵやメモリ等から実現され得る。変化判断部１１０の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００３６】
第二制御部１１１は、変化判断部１１０が、変化量が閾値より小さい、または閾値以下であると判断するまで、対訳文対移動部１０６、第一制御部１０７および対訳文対移動先決定部１０８に当該各部の処理を繰り返させる。各部の処理を繰り返させる処理は、例えば、各部の処理に対応する関数を呼び出す処理である。第二制御部１１１は、通常、ＭＰ

Ｕやメモリ等から実現され得る。第二制御部１１１の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００３７】

クラスタ蓄積部１１２は、対訳文対移動先決定部１０８が最後にバッファに対訳文対を書き込んだ後のｎ個のバッファ内の対訳文対のｎ種類の集合を、ｎ種類に区別して、記録媒体に蓄積する。クラスタ蓄積部１１２は、通常、ｎ種類のバッファに分類された対訳文対の集合を書き込む。ただし、クラスタ蓄積部１１２は、フラグ（例えば、１からｎまでの整数値）が付与された多数の対訳文対を、一のバッファに書き込んでも良い。クラスタとは、ｎ種類に区分されたうちの一の区分の対訳文対の集合をいう。したがって、クラスタ蓄積部１１２は、ｎ個のクラスタを記録媒体に蓄積する処理を行う。クラスタ蓄積部１１２は、通常、ＭＰＵやメモリ等から実現され得る。クラスタ蓄積部１１２の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００３８】

次に、クラスタ生成装置の動作について図２から図５のフローチャートを用いて説明する。

【００３９】

（ステップＳ２０１）対訳文対配置部１０２は、カウンタｉに１を代入する。

【００４０】

（ステップＳ２０２）対訳文対配置部１０２は、ｉ番目の対訳文対が対訳文対格納部１０１に存在するか否かを判断する。ｉ番目の対訳文対が存在すればステップＳ２０３に行き、ｉ番目の対訳文対が存在しなければステップＳ２０７に行く。

【００４１】

（ステップＳ２０３）対訳文対配置部１０２は、ｉ番目の対訳文対を対訳文対格納部１０１から読み出す。

【００４２】

（ステップＳ２０４）対訳文対配置部１０２は、「（ｉ／ｎの余り）＋１」の値を算出し、当該算出した値を、変数ｊに格納する。なお、「ｎ」はバッファ数であり、予め格納されている。対訳文対配置部１０２は、ｎの値および式「（ｉ／ｎの余り）＋１」の情報を読み出し、「ｉ」「ｎ」の値を式に代入し、演算する。

【００４３】

（ステップＳ２０５）対訳文対配置部１０２は、ｊ番目のバッファにｉ番目の対訳文対を移動（または複写）する。

【００４４】

（ステップＳ２０６）対訳文対配置部１０２は、カウンタｉを１、インクリメントする。ステップＳ２０２に戻る。

【００４５】

（ステップＳ２０７）言語モデル取得部１０３は、言語モデルを作成し、記録媒体上に配置する。かかる言語モデル作成の処理について、図３のフローチャートを用いて詳細に説明する。

【００４６】

（ステップＳ２０８）エントロピー算出部１０４は、エントロピー算出処理を行う。エントロピー算出処理について、図４のフローチャートを用いて詳細に説明する。エントロピー算出処理は、各バッファ（言語モデル）に対応して、エントロピーが算出される。つまり、バッファ数（ｎ）に対して、ｎのエントロピーが算出される。

【００４７】

（ステップＳ２０９）総エントロピー算出部１０５は、ステップＳ２０８で算出されたｎのエントロピーの和を算出し、記録媒体に格納する。ｎのエントロピーの和は、総エントロピーである。

【００４８】
（ステップＳ２１０）対訳文対移動先決定部１０８は、カウンタｉに１を代入する。
【００４９】
（ステップＳ２１１）対訳文対移動先決定部１０８は、ｉ番目の対訳文対がｎのバッファ中に存在するか否かを判断する。ｉ番目の対訳文対が存在すればステップＳ２１２に行き、ｉ番目の対訳文対が存在しなければステップＳ２１４に行く。
【００５０】
（ステップＳ２１２）対訳文対移動先決定部１０８は、ｉ番目の対訳文対の移動先を決定する。この移動先決定の処理について、図５のフローチャートを用いて詳細に説明する。
【００５１】
（ステップＳ２１３）対訳文対移動先決定部１０８は、カウンタｉを１、インクリメントする。ステップＳ２１１に戻る。
【００５２】
（ステップＳ２１４）対訳文対移動先決定部１０８は、ステップＳ２１２で決定された総エントロピー（ループの最後に決定された総エントロピー）を一時的にメモリに追記する。
【００５３】
（ステップＳ２１５）変化量算出部１０９は、最新の総エントロピーを取得する。最新の総エントロピーは、最後にステップＳ２１４で追記された総エントロピーである。
【００５４】
（ステップＳ２１６）変化量算出部１０９は、最新の一つ前（直前）の総エントロピーを読み出す。
【００５５】
（ステップＳ２１７）変化量算出部１０９は、ステップＳ２１５で得た総エントロピーと、ステップＳ２１６で得た総エントロピーから、総エントロピーの変化量を算出する。例えば、変化量算出部１０９は、「｜ステップＳ２１５で得た総エントロピー　－　ステップＳ２１６で得た総エントロピー｜」により、総エントロピーの変化量を算出する。
【００５６】
（ステップＳ２１８）変化量算出部１０９は、予め格納している閾値を読み出し、「変化量<閾値」を満たすか否かを判断する。「変化量<閾値」を満たせばステップＳ２１９に行き、「変化量<閾値」を満たさなければステップＳ２１０に戻る。
【００５７】
（ステップＳ２１９）クラスタ蓄積部１１２は、ｎのバッファの対訳文対の集合を、ｎに区別して記録媒体に書き込む。ｎに区別された対訳文対の集合を、それぞれクラスタやタスクなどとも呼ぶ。
【００５８】
次に、ステップＳ２０７の言語モデル作成の処理について、図３のフローチャートを用いて詳細に説明する。
【００５９】
（ステップＳ３０１）言語モデル取得部１０３は、カウンタｉに１を代入する。
【００６０】
（ステップＳ３０２）言語モデル取得部１０３は、「ｉ<＝ｎ」であるか否かを判断する。「ｉ<＝ｎ」であればステップＳ３０３に行き、「ｉ<＝ｎ」でなければ上位関数にリターンする。
【００６１】
（ステップＳ３０３）言語モデル取得部１０３は、ｉ番目のバッファ内の対訳文対の集合を読み出す。
【００６２】
（ステップＳ３０４）言語モデル取得部１０３は、ステップＳ３０３で読み出した対訳

文対の集合を単語に分割し、全単語をメモリ上に配置する。

【００６３】

（ステップＳ３０５）言語モデル取得部１０３は、ステップＳ３０４で分割した単語をソートし、ソートした結果をメモリ上に配置する。

【００６４】

（ステップＳ３０６）言語モデル取得部１０３は、ステップＳ３０４で分割した単語の全数を取得する。

【００６５】

（ステップＳ３０７）言語モデル取得部１０３は、カウンタｊに１を代入する。

【００６６】

（ステップＳ３０８）言語モデル取得部１０３は、ステップＳ３０５でソートした全単語中に、ｊ番目の種類の単語が存在するか否かを判断する。ｊ番目の種類の単語が存在すればステップＳ３０９に行き、ｊ番目の種類の単語が存在しなければステップＳ３１３に行く。

【００６７】

（ステップＳ３０９）言語モデル取得部１０３は、ソートした全単語中における、ｊ番目の種類の単語の出現回数を取得する。

【００６８】

（ステップＳ３１０）言語モデル取得部１０３は、ソートした全単語中における、ｊ番目の種類の単語の出現確率を取得する。言語モデル取得部１０３は、「ステップＳ３０９で取得した出現回数／ステップＳ３０６で算出した単語の全数」により、ｊ番目の種類の単語の出現確率を算出する。

【００６９】

（ステップＳ３１１）言語モデル取得部１０３は、ｊ番目の種類の単語と、ステップＳ３１０で取得した出現確率を対にして、記録媒体に蓄積する。

【００７０】

（ステップＳ３１２）言語モデル取得部１０３は、カウンタｊを１、インクリメントする。ステップＳ３０８に戻る。

【００７１】

（ステップＳ３１３）言語モデル取得部１０３は、カウンタｉを１、インクリメントする。ステップＳ３０２に戻る。

【００７２】

次に、ステップＳ２０８のエントロピー算出処理について図４のフローチャートを用いて説明する。

【００７３】

（ステップＳ４０１）エントロピー算出部１０４は、カウンタｉに１を代入する。

【００７４】

（ステップＳ４０２）エントロピー算出部１０４は、「ｉ<＝ｎ」であるか否かを判断する。「ｉ<＝ｎ」であればステップＳ４０３に行き、「ｉ<＝ｎ」でなければ上位関数にリターンする。

【００７５】

（ステップＳ４０３）エントロピー算出部１０４は、ｉ番目のバッファに対応するｉ番目の言語モデルを読み出す。なお、言語モデルは、図３のフローチャートの処理により、取得されている。

【００７６】

（ステップＳ４０４）エントロピー算出部１０４は、カウンタｊに１を代入し、変数「エントロピー」を０に初期化する。

【００７７】

（ステップＳ４０５）エントロピー算出部１０４は、ステップＳ４０３で読み出した言語モデル中に、ｊ番目の種類の単語が存在するか否かを判断する。ｊ番目の種類の単語が

存在すればステップＳ４０６に行き、ｊ番目の種類の単語が存在しなければステップＳ４１０に行く。

【００７８】

（ステップＳ４０６）エントロピー算出部１０４は、ｊ番目の種類の単語に対応する出現確率Ｐを、ステップＳ４０３で読み出した言語モデルから読み出す。

【００７９】

（ステップＳ４０７）エントロピー算出部１０４は、「－ｌｏｇＰ」を算出し、結果をメモリ上に配置する。

【００８０】

（ステップＳ４０８）エントロピー算出部１０４は、変数「エントロピー」に、「－ｌｏｇＰ」を加算する。

【００８１】

（ステップＳ４０９）エントロピー算出部１０４は、カウンタｊを１、インクリメントする。ステップＳ４０５に戻る。

【００８２】

（ステップＳ４１０）エントロピー算出部１０４は、ｉ番目のバッファに対応するエントロピー（変数「エントロピー」の値）を、記録媒体に蓄積する。

【００８３】

（ステップＳ４１１）エントロピー算出部１０４は、カウンタｉを１、インクリメントする。ステップＳ４０２に戻る。

【００８４】

次に、ステップＳ２１２の移動先決定処理について図５のフローチャートを用いて説明する。

【００８５】

（ステップＳ５０１）対訳文対移動先決定部１０８は、処理対象の対訳文対を読み出す。

【００８６】

（ステップＳ５０２）対訳文対移動先決定部１０８は、カウンタｊに１を代入する。

【００８７】

（ステップＳ５０３）対訳文対移動先決定部１０８は、「ｊ<＝ｎ」であるか否かを判断する。「ｊ<＝ｎ」であればステップＳ５０４に行き、「ｊ<＝ｎ」でなければステップＳ５１０に行く。

【００８８】

（ステップＳ５０４）対訳文対移動先決定部１０８は、ステップＳ５０１で読み出した対訳文対を、ｊ番目のバッファに移動する。

【００８９】

（ステップＳ５０５）言語モデル取得部１０３は、ステップＳ５０４の処理後のｎのバッファの状態で、言語モデルを作成する。

【００９０】

（ステップＳ５０６）エントロピー算出部１０４は、エントロピー算出処理を行う。

【００９１】

（ステップＳ５０７）総エントロピー算出部１０５は、ステップＳ５０６で算出されたｎのエントロピーの和を算出し、記録媒体に格納する。

【００９２】

（ステップＳ５０８）対訳文対移動先決定部１０８は、ｊと、ステップＳ５０７で算出した総エントロピーの組を一時的にメモリに格納する。

【００９３】

（ステップＳ５０９）対訳文対移動先決定部１０８は、カウンタｊを１、インクリメントする。ステップＳ５０３に戻る。

【００９４】

（ステップＳ５１０）対訳文対移動先決定部１０８は、ステップＳ５０８で格納したｎの総エントロピーの中で、最小の総エントロピーを決定する。

【００９５】

（ステップＳ５１１）対訳文対移動先決定部１０８は、ステップＳ５１０で決定した最小の総エントロピーに対応するｊ（バッファの識別子）を取得する。

【００９６】

（ステップＳ５１２）対訳文対移動先決定部１０８は、移動対象の対訳文対をｊ番目のバッファに書き込む。なお、移動対象の対訳文対がｊ番目のバッファに存在する場合、かかる処理は行わない。上位関数にリターンする。

【００９７】

なお、図５のフローチャートにおいて、ステップＳ５０１で読み出した対訳文対が最初に存在するバッファに関して、ステップＳ５０４の処理は省略しても良いことは言うまでもない。

【００９８】

以下、本実施の形態におけるクラスタ生成装置の意義について説明する。クラスタ生成装置で生成し、記録媒体に蓄積されたｎに分類されたクラスタは、例えば、統計翻訳で利用される。

【００９９】

統計翻訳は、以下の数式１に示されるように、与えられた翻訳原言語単語列（ｅ）に対し、確率が最大となる翻訳目的言語単語列（ｆ）を見つける問題である。

【数１】

$$\underset{f}{\operatorname{argmax}}\, P(f|e)$$

【０１００】

この数式１においては、翻訳先目的単語列（ｆ）は翻訳原単語列（ｅ）のみで決るが、実際はトピック等の環境の影響を大きく受ける。ここでは、この環境をタスクとみなし（Ｔ）で表わすこととする。この、タスク（Ｔ）が既知の場合は（Ｔ）を新たな変量として数式１に導入することにより、以下の数式２が得られる。なお、タスク（Ｔ）は、クラスタ（Ｔ）と同義であり、上述した各バッファの対訳文対の集合に対応する。

【数２】

$$\underset{f}{\operatorname{argmax}}\, P(f|e,T)$$

【０１０１】

そして、数式２は、ベイズ則を用いて、数式３のように書き換えることができる。

【数３】

$$\underset{f}{\operatorname{argmax}}\, P(f|T)P(e|f,T)$$

【０１０２】

ここで、Ｐ（ｆ｜ｅ，Ｔ）がタスク依存翻訳モデル、Ｐ（ｆ｜Ｔ）がタスク依存言語モデルである。

【０１０３】

数式２を用いる場合は、あらかじめ適応モデルを構築しておく、すなわちオフライン適応が可能であるが、タスク（Ｔ）は必ずしも既知ではない。この場合は、タスク（Ｔ）と翻訳目的単語列（ｆ）を同時に推定する問題、すなわちオンライン適応として、以下の数式４のように表わされることになる。

【数４】

$$\underset{f,T}{\operatorname{argmax}}\, P(f,T|e)$$
$$= \underset{f,T}{\operatorname{argmax}}\, P(T|e)P(f|e,T)$$

【０１０４】
この数式４と数式２との大きな違いは、数式４においてはタスクを表わす変量（Ｔ）が隠れ変数となっていることである。また数式４の右辺のＰ（Ｔ｜ｅ）はタスク推定、Ｐ（ｆ｜ｅ，Ｔ）はタスク適応を表わしている。
【０１０５】
数式４を満たす翻訳目的単語列（ｆ）を求めるためにはＰ（Ｔ｜ｅ）とＰ（ｆ｜ｅ，Ｔ）を同時に最大化する必要がある。しかしながら、これは困難であるため、本実施の形態、および実施の形態２において、近似としてまずＰ（Ｔ｜ｅ）を最大化し、それによって求まった（Ｔ）を用いてＰ（ｆ｜ｅ，Ｔ）を最大化するという手順をとる。
【０１０６】
オフライン適応の場合、タスクはトピック等の人間の感覚に合ったものとして、あらかじめ規定されていることが多い。しかしながら、オンライン適応の場合は、タスクは隠れ変数として用いられ、外部に出力する必要もない。このため、まずタスクそのものを自由に規定しておくことが可能である。この場合のタスクは、必ずしも人間の感覚に合ったものである必要はないため、タスク（Ｔ）は統計的な観点から、数式４をなるべく大きく、すなわちＰ（Ｔ｜ｅ）とＰ（ｆ｜ｅ，Ｔ）をなるべく大きくできるように規定することが望ましい。
【０１０７】
そこで、この近似として、Ｐ（ｆ｜ｅ，Ｔ）をＰ（ｆ｜Ｔ）で置き換えたＰ（Ｔ｜ｅ）Ｐ（ｆ｜Ｔ）を最大化するような（Ｔ）を規定する。最大化の対象であるＰ（Ｔ｜ｅ）Ｐ（ｆ｜Ｔ）は、ベイズ則を用いて数式５のように書き換えることができる。
【数５】

$$P(f|T)P(e|T)P(T)/P(e)$$

【０１０８】
ここで、Ｐ（ｅ）は（Ｔ）に無関係であり、さらにＰ（Ｔ）を定数とする近似を導入することによって、規定すべきタスク（Ｔ）は、数式６で表わされることになる。
【数６】

$$\operatorname*{argmax}_{T} P(f|T)P(e|T)$$

【０１０９】
この数式６は、（ｆ）と（ｅ）に対して同時に確率を最大化する、すなわち（ｆ）と（ｅ）の尤度の和を最大化するように（Ｔ）を規定することを意味している。これはすなわち、（ｆ）と（ｅ）の対訳文対をエントロピー最小化の基準の元にクラスタリングを行えばよいことを意味している。本クラスタ生成装置は、（ｆ）と（ｅ）の対訳文対をエントロピー最小化の基準の元にクラスタリングを行った結果を得る。この結果は、ｎ個のバッファ内の対訳文対のｎ種類の集合を、ｎ種類に区別して蓄積した情報であり、ｎに分類されたクラスタである。
【０１１０】
以上、本実施の形態によれば、対訳文対の集合に対して、エントロピー最小化の基準に対応した分類が可能になる。したがって、本実施の形態によれば、実施の形態２で述べる機械翻訳に好適な言語モデルを自動的に構築できる。また、本実施の形態によれば、言語モデルの元になるｎ種類に区別して蓄積された対訳文対の集合を自動的に得ることができる。なお、本実施の形態におけるクラスタ生成装置が生成したクラスタから生成された言語モデルを用いれば、精度の高い機械翻訳が可能となる（実施の形態２参照）。
【０１１１】
なお、本実施の形態によれば、言語モデル取得部１０３が取得する確率情報は、１以上の対訳文対中に一の用語が出現する確率（ｕｎｉ－ｇｒａｍの確率）であった。しかし、言語モデル取得部１０３が取得する確率情報は、単語２－ｇｒａｍの確率や、単語３－ｇｒａｍの確率等でも良い。

【０１１２】
また、本実施の形態におけるエントロピーは、本実施の形態で述べた式で算出される値に限られないことは言うまでもない。つまり、エントロピーは、用語の出現の均一具合を示す情報であれば良い。この用語の出現の均一具合は、裏返せば、用語の出現の偏りになり、エントロピーは、この偏り度合いをも含む概念としてとらえる、こととする。
【０１１３】
さらに、本実施の形態における処理は、ソフトウェアで実現しても良い。そして、このソフトウェアをソフトウェアダウンロード等により配布しても良い。また、このソフトウェアをＣＤ－ＲＯＭなどの記録媒体に記録して流布しても良い。なお、このことは、本明細書における他の実施の形態においても該当する。なお、本実施の形態におけるクラスタ生成装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、第一の言語の文と、当該文の第二の言語への翻訳文の対の情報である対訳文対を複数、記憶媒体に格納しており、コンピュータに、前記記憶媒体から複数の対訳文対を読み出し、当該複数の対訳文対をｎ個のバッファに配置する対訳文対配置ステップと、前記バッファ毎に、前記対訳文対配置ステップが配置する各バッファ中の１以上の対訳文対を１以上の用語に分割し、当該１以上の対訳文対中に用語が出現する確率についての情報である確率情報を取得し、用語と当該用語に対応する確率情報を有する用語出現確率情報を１以上有する情報である言語モデルを取得し、記録媒体上に配置する言語モデル取得ステップと、前記言語モデル取得ステップで取得した１以上の用語出現確率情報が有する１以上の確率情報を用いて、前記ｎ個のバッファ毎に、用語の出現の均一具合についての情報であるｎのエントロピーを算出し、記憶媒体に配置するエントロピー算出ステップと、前記ｎのエントロピーを取得し、前記ｎ個のバッファ全体の用語の出現の均一具合についての情報である総エントロピーを算出し、記憶媒体に配置する総エントロピー算出ステップと、前記ｎ個のバッファのうちのいずれかのバッファ中のいずれかの対訳文対を読み出し、他の各バッファに移動する対訳文対移動ステップと、前記対訳文対移動ステップで対訳文対を各バッファに移動した後、バッファごとに、前記言語モデルを取得し、記録媒体上に配置させ、前記ｎのエントロピーを算出し、記憶媒体に配置させ、および総エントロピーを算出し、記憶媒体に配置させる第一制御ステップと、前記第一制御ステップにおける処理に対応して、バッファごとに、得られたｎの総エントロピーを取得し、当該ｎの総エントロピーのうちで最も小さい総エントロピーに対応するバッファに、当該移動対象の対訳文対の移動先のバッファを決定し、当該バッファに前記移動対象の対訳文対を書き込む対訳文対移動先決定ステップと、前記対訳文対移動先決定ステップにおいて、全対訳文対について移動先を決定した後の最近の総エントロピーと、その前のサイクルにおいて、前記対訳文対移動先決定ステップで全対訳文対について、移動先を決定した後の総エントロピーである直前の総エントロピーを用いて、エントロピーの変化量を算出し、記録媒体に配置する変化量算出ステップと、前記変化量算出ステップで算出した変化量が閾値より小さいか否か、または閾値以下であるか否かを判断する変化判断ステップと、前記変化判断ステップで、変化量が閾値より小さい、または閾値以下であると判断するまで、前記対訳文対移動ステップにおける処理、前記第一制御ステップにおける処理、および前記対訳文対移動先決定ステップにおける処理を繰り返させ、前記対訳文対移動先決定ステップにおいて最後にバッファに対訳文対を書き込んだ後の前記ｎ個のバッファ内の対訳文対のｎ種類の集合を、ｎ種類に区別して蓄積するクラスタ蓄積ステップを実行させるためのプログラム、である。
【０１１４】
また、上記プログラムにおいて、前記言語モデル取得部が取得する確率情報は、１以上の対訳文対中に一の用語が出現する確率であることは好適である。
（実施の形態２）
【０１１５】
図６は、本実施の形態における翻訳装置のブロック図である。本翻訳装置は、言語モデル格納部５０１、受付部５０２、文分割部５０３、翻訳原文出現確率算出部５０４、言語

モデル決定部５０５、翻訳部５０６を具備する。

【０１１６】

言語モデル格納部５０１は、ｎ（ｎは２以上の整数）種類の区別された言語モデルを格納している。言語モデルは、用語および当該用語が１以上の対訳文対中に出現する確率についての情報である確率情報を用語毎に有する。言語モデルは、複数の用語および当該複数の連続する用語の組が１以上の対訳文対中に出現する確率についての情報である確率情報を複数の用語毎に有しても良い。ここでの言語モデルは、例えば、日本語と英語の対訳文対の集合から構成される言語モデルであり、日本語、英語ともＧｏｏｄ－Ｔｕｒｉｎｇ（非特許文献１）で平滑化された単語３－ｇｒａｍである。また、入力された日本語文に対し、最大尤度を与えるクラスタを選択する際のモデルは、例えば、クラスタ依存の日本語単語３－ｇｒａｍである。一方、選択されたクラスタに対して、翻訳時に用いられる英語の言語モデルは、クラスタ依存の英語単語３－ｇｒａｍである。また、言語モデル格納部５０１の言語モデルは、実施の形態１で説明したクラスタ生成装置が蓄積したｎ種類の各対訳文対の集合から構成された情報であり、ｎ種類の各対訳文対の集合が有する各対訳文対を１以上の用語に分割し、当該１以上の用語が対訳文対の集合中に出現する確率についての情報である確率情報を１以上の用語毎に算出されることにより得られた情報であることは好適である。ここで、確率情報は、例えば、単語３－ｇｒａｍの確率である。また、確率情報は、例えば、単語ｕｎｉｇｒａｍの確率であっても良い。

【０１１７】

言語モデル格納部５０１は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。

【０１１８】

受付部５０２は、翻訳対象の第一の言語（原言語）の文を受け付ける。受付部５０２は、その他、翻訳処理の開始指示などのユーザからの指示やデータなどを受け付けても良い。第一の言語の文などの入力手段は、テンキーやキーボードやマウスやメニュー画面によるもの等、何でも良い。受付部５０２は、テンキーやキーボード等の入力手段のデバイスドライバーや、メニュー画面の制御ソフトウェア等で実現され得る。

【０１１９】

文分割部５０３は、受付部５０２が受け付けた文を取得し、当該文を１以上の用語に分割し、記憶媒体に配置する。文分割部５０３は、例えば、文に対して形態素解析を行った後、単語に分割しても良い。文を単語に分割する方法は問わない。文を単語に分割する技術は公知技術であるので、詳細な説明を省略する。文分割部５０３は、通常、ＭＰＵやメモリ等から実現され得る。文分割部５０３の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【０１２０】

翻訳原文出現確率算出部５０４は、言語モデル格納部５０１の各言語モデルを読み出し、当該各言語モデルを用いて、文分割部５０３が取得した１以上の各用語が、各言語モデルが有する１以上の対訳文対中に出現する確率に関する情報である翻訳原文出現確率を、言語モデル毎に算出し、記憶媒体に配置する。翻訳原文出現確率は、例えば、受付部５０２が受け付けた文のエントロピーである。翻訳原文出現確率は、例えば、言語モデルにおける、１以上の各用語の出現確率の積でも良い。翻訳原文出現確率算出部５０４は、通常、ＭＰＵやメモリ等から実現され得る。翻訳原文出現確率算出部５０４の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【０１２１】

言語モデル決定部５０５は、言語モデル毎に算出されたｎの翻訳原文出現確率を用いて、最も出現する確率が高い言語モデルを決定する。例えば、翻訳原文出現確率が受付部５０２の受け付けた文（翻訳対象文）のエントロピーである場合、言語モデル決定部５０５は、エントロピーを最小にする言語モデルを選択する。つまり、言語モデル決定部５０５

は、翻訳対象文に対し、最も高い尤度を与える言語モデルを選択する。言語モデル決定部５０５は、通常、ＭＰＵやメモリ等から実現され得る。言語モデル決定部５０５の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【０１２２】

翻訳部５０６は、言語モデル決定部５０５が決定した言語モデルを、言語モデル格納部５０１から読み出し、当該読み出した言語モデルを用いて、受付部５０２が受け付けた文を第二の言語（目的言語）の文に翻訳する。翻訳部５０６は、統計翻訳を行う。翻訳部５０６は、公知技術（例えば、非特許文献２参照）により実現可能であるので、詳細な説明を省略する。翻訳部５０６は、言語モデル決定部５０５が決定した言語モデル（Ｔ）を、言語モデル格納部５０１から読み出し、当該読み出した言語モデル（Ｔ）において、Ｐ（ｅ｜ｆ，Ｔ）Ｐ（ｆ｜Ｔ）［ｅは入力された翻訳対象の文、ｆは目的言語の文］を最大にする第二の言語の文（ｆ）を、言語モデル（Ｔ）が有する１以上の対訳文対が有する第二の言語の翻訳文から選択し、出力する。言語モデル（Ｔ）が決定され、受付部５０２が受け付けた翻訳対象の文（ｅ）が決まっている状況で、翻訳部５０６は、言語モデル（Ｔ）に対応するすべての対訳文対が有する第二の言語の文（ｆ）に対して、「Ｐ（ｅ｜ｆ，Ｔ）Ｐ（ｆ｜Ｔ）」を算出し、最大の「Ｐ（ｅ｜ｆ，Ｔ）Ｐ（ｆ｜Ｔ）」の値を示す（ｆ）を取得することは好適である。出力は、ディスプレイへの表示、プリンタへの印刷、スピーカーへの音出力などである。翻訳部５０６は、通常、ＭＰＵやメモリ等から実現され得る。翻訳部５０６の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【０１２３】

次に、翻訳装置の動作について図７のフローチャートを用いて説明する。

【０１２４】

（ステップＳ７０１）受付部５０２は、翻訳対象の文（ｅ）を受け付けたか否かを判断する。翻訳対象の文（ｅ）を受け付ければステップＳ７０２に行き、翻訳対象の文（ｅ）を受け付けなければステップＳ７０１に戻る。

【０１２５】

（ステップＳ７０２）文分割部５０３は、ステップＳ７０１で受け付けた文（ｅ）をメモリ上に配置する。

【０１２６】

（ステップＳ７０３）文分割部５０３は、ステップＳ７０２でメモリ上に配置した文を1以上の用語に分割し、当該1以上の用語をメモリ上に配置する。

【０１２７】

（ステップＳ７０４）翻訳原文出現確率算出部５０４は、カウンタｉに１を代入する。

【０１２８】

（ステップＳ７０５）翻訳原文出現確率算出部５０４は、「ｉ<＝ｎ」であるか否かを判断する。「ｉ<＝ｎ」であればステップＳ７０６に行き、「ｉ<＝ｎ」でなければステップＳ７１０に行く。なお、ｎは、言語モデルの数であり、クラスタの数であり、タスクの数である、といえる。「ｎ」は予め格納されており、翻訳原文出現確率算出部５０４は、「ｎ」の値を読み出し、「ｉ<＝ｎ」が真であるか、偽であるかを判断する。

【０１２９】

（ステップＳ７０６）翻訳原文出現確率算出部５０４は、ｉ番目の言語モデルを言語モデル格納部５０１から読み出す。

【０１３０】

（ステップＳ７０７）翻訳原文出現確率算出部５０４は、ｉ番目の言語モデルに対する、文分割部５０３が取得した１以上の用語の翻訳原文出現確率を算出し、記憶媒体に配置する。翻訳原文出現確率算出部５０４は、例えば、文（ｅ）に対する尤度（この尤度は、ここでは翻訳原文出現確率）を算出する。翻訳原文出現確率算出部５０４は、例えば、文

分割部５０３が取得した１以上の各用語のｉ番目の言語モデル内の出現確率の積を算出し、翻訳原文出現確率としても良い。

【０１３１】

（ステップＳ７０８）翻訳原文出現確率算出部５０４は、ステップＳ７０７で算出したｉ番目の翻訳原文出現確率をメモリ上に追記する。

【０１３２】

（ステップＳ７０９）翻訳原文出現確率算出部５０４は、カウンタｉを１、インクリメントする。ステップＳ７０５に行く。

【０１３３】

（ステップＳ７１０）言語モデル決定部５０５は、ステップＳ７０８でメモリ上に配置されたｎの翻訳原文出現確率を読み出し、当該ｎの翻訳原文出現確率の中から、最も大きい翻訳原文出現確率に対応する言語モデルを決定する。そして、翻訳部５０６は、言語モデル決定部５０５が決定した言語モデルを、言語モデル格納部５０１から読み出す。

【０１３４】

（ステップＳ７１１）翻訳部５０６は、カウンタｉに１を代入する。

【０１３５】

（ステップＳ７１２）翻訳部５０６は、ｉ番目のフレーズ対が言語モデル格納部５０１に存在するか否かを判断する。ｉ番目のフレーズ対が存在すればステップＳ７１３に行き、ｉ番目のフレーズ対が存在しなければステップＳ７１７に行く。

【０１３６】

（ステップＳ７１３）翻訳部５０６は、ｉ番目のフレーズ対を、言語モデル格納部５０１から読み出す。

【０１３７】

（ステップＳ７１４）翻訳部５０６は、ステップＳ７１３で読み出したフレーズ対と格納している翻訳モデルを用いて、「Ｐ（ｅ｜ｆ，Ｔ）Ｐ（ｆ｜Ｔ）」を算出する。

【０１３８】

（ステップＳ７１５）翻訳部５０６は、ステップＳ７１４で算出した結果をメモリ上に追記する。

【０１３９】

（ステップＳ７１６）翻訳部５０６は、カウンタｉを１、インクリメントする。ステップＳ７１２に行く。

【０１４０】

（ステップＳ７１７）翻訳部５０６は、ステップＳ７１５でメモリ上に配置した値を用いて、最大の値をとる翻訳文（ｆ）を構成する。

【０１４１】

（ステップＳ７１８）翻訳部５０６は、ステップＳ７１７で構成した翻訳文（ｆ）を出力する。処理を終了する。

【０１４２】

以下、本実施の形態における翻訳装置の意義について説明する。

【０１４３】

本翻訳装置において、タスク（Ｔ）を入力された（ｅ）から推定することになる。これはすなわちＰ（Ｔ｜ｅ）を最大化する（Ｔ）を見つけることである。Ｐ（Ｔ｜ｅ）はベイズ則を用いて、Ｐ（ｅ｜Ｔ）Ｐ（Ｔ）／Ｐ（ｅ）と書き換えることができ、クラスタリング時に用いた、Ｐ（Ｔ）を定数とする近似を導入すれば、Ｐ（ｅ｜Ｔ）を最大化すればよいことになる。これはすなわち、（ｅ）に対して、最大尤度を与えるタスク、すなわちクラスタ（Ｔ）を選べばよいことになる。

【０１４４】

次に、推定された（Ｔ）を用いてタスク適応、すなわちＰ（ｆ｜ｅ，Ｔ）の最大化を図る。Ｐ（ｆ｜ｅ，Ｔ）はベイズ則を用いて、下記の数式７に書き換えることができる。

【数７】

$$P(f|e,T)$$
$$=P(e|f,T)P(f,T)/P(e,T)$$
$$=P(e|f,T)P(f|T)P(T)/P(e,T)$$

【０１４５】

ここで、（ｅ）と（Ｔ）は既知であるため、数式７を最大化するためにはＰ（ｅ｜ｆ，Ｔ）Ｐ（ｆ｜Ｔ）を最大化すればよいことになる。この式で、Ｐ（ｅ｜ｆ，Ｔ）がタスク依存（タスク適応後）翻訳モデル、Ｐ（ｆ｜Ｔ）がタスク依存（タスク適応後）言語モデルであり、これらのタスク依存モデルを用いて（ｅ）から（ｆ）を推定することを意味している。従って、上記の図７のフローチャートのステップＳ７１４において、翻訳部５０６は、「Ｐ（ｅ｜ｆ，Ｔ）Ｐ（ｆ｜Ｔ）」を算出した。

【０１４６】

また、本翻訳装置が翻訳処理を行う前処理として、翻訳モデル、言語モデルの学習データである対訳文対に対し、クラスタリングを行い、その後、クラスタごとに翻訳原言語、翻訳目的言語のクラスタ依存言語モデルを作成する。対訳文対に対し、クラスタリングを行う処理は、上記の実施の形態１におけるクラスタ生成装置が行う。また、クラスタごとに翻訳原言語、翻訳目的言語のクラスタ依存言語モデルを作成する処理とは、クラスタごとに、上述の言語モデルを作成する処理であり、公知技術である。

【０１４７】

そして、本翻訳装置において、複数の言語モデル（クラスタ）が用意されており、翻訳原文（ｅ）を受け付け、当該翻訳原文に対し、最も高い尤度を与えるクラスタを選択する。そして、選択されたクラスタの翻訳目的言語のクラスタ依存言語モデルを用いて翻訳を行う。なお、クラスタを選択した後の、クラスタ依存言語モデルを用いた翻訳処理は公知技術であるので、詳細な説明を省略する。

【０１４８】

さらに、以下の本実施の形態で述べた手法の実験結果について述べる。

【０１４９】

本実験において、対象としたドメインは旅行対話で、用いたコーパスは、旅行対話基本表現集（ＡＴＲ旅行対話基本表現集（ＢＴＥＣ））である。翻訳言語対は、日本語から英語であり、学習、および評価コーパスのサイズ等は、図８、図９に示す通りである。

【０１５０】

本実験において、作成した言語モデルは、日本語、英語ともＧｏｏｄ－Ｔｕｒｉｎｇで平滑化された単語３－ｇｒａｍである。入力された日本語文に対し、最大尤度を与えるクラスタを選択する際のモデルとしては、クラスタ依存の日本語単語３－ｇｒａｍをそのまま用いている。

【０１５１】

また、一方、選択されたクラスタに対して、翻訳時に用いられる英語の言語モデルとしては、クラスタ依存の英語単語３－ｇｒａｍと全ての英語学習データを用いて作成したクラスタ非依存単語３－ｇｒａｍを線形補間したものを用いた。

【０１５２】

以上のような状況において、まず、対訳文対に対して行うスラスタリングの際の、クラスタ数を変化させた時の、本明細書における翻訳方法の性能評価を行った。評価基準は翻訳目的言語である英語の評価セットに対するパープレキシティである。この時、対象の英語言語モデルにおける、クラスタ依存モデルとクラスタ非依存モデルの線形補間係数は０．５で固定した。

【０１５３】

変化させたクラスタ数は、５、１０、２０である。その場合の結果を図１０に示す。図１０において左側の軸目盛り"Perplexity"がパープレキシティであり、クラスタ数を変化させた時の値が点線で示されている。クラスタ数が１の場合は適応を行っていない場合、

すなわちベースラインを示している。また、右側の軸目盛り"Reduction Rate"は適応を行うことによってエントロピーが減少した評価セット文の割合をしめしている。すなわち、この値が８９であれば、評価セット文１、５２４文のうち、１、３５７文が適応によってエントロピーが減少し、残りの１６７文では逆に増加したことを示している。

【０１５４】

図１０に示される通り、クラスタ数の増加と共に適応後の言語モデルのパープレキシティは減少している。その値はクラスタ数２０の場合で、ベースラインの２６．９から１８．６に減少しており、割合では約３１％となっている。また、この時エントロピーが減少した評価セット文の割合も８９％と高く、特定の文に限って効果が現れているわけではないことを示している。

【０１５５】

次に、クラスタ依存モデルとクラスタ非依存モデルの線形補間係数を変化させた場合の性能変化を調べた。この時のクラスタ数は前節の評価で最もパープレキシティの低かった２０に固定した。変化させた補間係数は０．０５、および０．１から０．９まで０．１刻みである。このうち、０．５から０．９のあいだの結果を図１１に示す。左右の軸は図１０と同様である。

【０１５６】

図１１に示されるように、補間係数０．７でパープレキシティは最小値１８．１を示しており、これはベースラインの２６．９に対して約３３％の減少となっている。しかしながら、エントロピー減少文の割合は補間係数が大きくなるに従って減少している。この割合は図には示されていないが、補間係数０．１で最小となり、その値は９２％である。またこの時のパープレキシティは２３．１であった。

【０１５７】

以上、本実施の形態によれば、翻訳装置は、入力された翻訳原言語文（ｅ）に対して最も低いパープレキシティを与えるクラスタを選択し、最終的に翻訳目的言語に対する、選択されたクラスタ依存言語モデルを用いることにより、高い性能の翻訳が可能になる。

【０１５８】

なお、本実施の形態において、翻訳装置がクラスタを決定した後の翻訳方法は問わない。

【０１５９】

さらに、本実施の形態における処理は、ソフトウェアで実現しても良い。そして、このソフトウェアをソフトウェアダウンロード等により配布しても良い。また、このソフトウェアをＣＤ－ＲＯＭなどの記録媒体に記録して流布しても良い。なお、このことは、本明細書における他の実施の形態においても該当する。なお、本実施の形態における翻訳装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、ｎ（ｎは２以上の整数）種類の区別された言語モデルであり、用語および当該用語が１以上の対訳文対中に出現する確率についての情報である確率情報を用語毎に有する言語モデルを記録媒体に格納しており、コンピュータに、翻訳対象の第一の言語の文を受け付ける受付ステップと、前記受付ステップで受け付けた文を取得し、当該文を１以上の用語に分割し、記憶媒体に配置する文分割ステップと、前記記録媒体の各言語モデルを読み出し、当該各言語モデルを用いて、前記文分割ステップで取得した１以上の各用語が、各言語モデルが有する１以上の対訳文対中に出現する確率に関する情報である翻訳原文出現確率を、言語モデル毎に算出し、記憶媒体に配置する翻訳原文出現確率算出ステップと、前記言語モデル毎に算出されたｎの翻訳原文出現確率を用いて、最も出現する確率が高い言語モデルを決定する言語モデル決定ステップと、前記言語モデル決定ステップで決定した言語モデルを、前記記録媒体から読み出し、当該読み出した言語モデルを用いて、前記前記受付ステップで受け付けた文を第二の言語の文に翻訳し、当該翻訳結果を出力する翻訳ステップを実行させるためのプログラム、である。

【０１６０】

また、上記プログラムにおける記憶媒体に格納しているｎ種類の区別された各言語モデ

ルは、実施の形態１のクラスタ生成装置が蓄積したｎ種類の各対訳文対の集合から構成された情報であり、ｎ種類の各対訳文対の集合が有する各対訳文対を１以上の用語に分割し、当該１以上の用語が対訳文対の集合中に出現する確率についての情報である確率情報を用語毎に算出されることにより得られた情報である、ことは好適である。

【０１６１】

また、上記プログラムにおける前記確率情報は、単語３－ｇｒａｍの確率である、ことは好適である。

【０１６２】

また、上記プログラムの前記翻訳ステップにおいて、前記言語モデル決定ステップで決定した言語モデル（Ｔ）を、前記記録媒体から読み出し、当該読み出した言語モデル（Ｔ）において、Ｐ（ｅ｜ｆ，Ｔ）Ｐ（ｆ｜Ｔ）［ｅは入力された翻訳対象の文、ｆは目的言語の文］を最大にする第二の言語の文（ｆ）を、前記言語モデル（Ｔ）が有する１以上の対訳文対が有する第二の言語の翻訳文から選択し、出力する、ことは好適である。

【０１６３】

また、上記各実施の形態において、各処理（各機能）は、単一の装置（システム）によって集中処理されることによって実現されてもよく、あるいは、複数の装置によって分散処理されることによって実現されてもよい。

【０１６４】

また、図１２は、本明細書で述べたプログラムを実行して、上述した種々の実施の形態のクラスタ生成装置、または翻訳装置を実現するコンピュータの外観を示す。上述の実施の形態は、コンピュータハードウェア及びその上で実行されるコンピュータプログラムで実現され得る。図１２は、このコンピュータシステム３４０の概観図であり、図１３は、コンピュータシステム３４０のブロック図である。

【０１６５】

図１２において、コンピュータシステム３４０は、ＦＤ（Ｆｌｅｘｉｂｌｅ Ｄｉｓｋ）ドライブ、ＣＤ－ＲＯＭ（Ｃｏｍｐａｃｔ Ｄｉｓｋ Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ）ドライブを含むコンピュータ３４１と、キーボード３４２と、マウス３４３と、モニタ３４４とを含む。

【０１６６】

図１３において、コンピュータ３４１は、ＦＤドライブ３４１１、ＣＤ－ＲＯＭドライブ３４１２に加えて、ＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）３４１３と、ＣＰＵ３４１３、ＣＤ－ＲＯＭドライブ３４１２及びＦＤドライブ３４１１に接続されたバス３４１４と、ブートアッププログラム等のプログラムを記憶するためのＲＯＭ（Ｒｅａｄ－Ｏｎｌｙ Ｍｅｍｏｒｙ）３４１５と、ＣＰＵ３４１３に接続され、アプリケーションプログラムの命令を一時的に記憶するとともに一時記憶空間を提供するためのＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）３４１６と、アプリケーションプログラム、システムプログラム、及びデータを記憶するためのハードディスク３４１７とを含む。ここでは、図示しないが、コンピュータ３４１は、さらに、ＬＡＮへの接続を提供するネットワークカードを含んでも良い。

【０１６７】

コンピュータシステム３４０に、上述した実施の形態のクラスタ生成装置、または翻訳装置の機能を実行させるプログラムは、ＣＤ－ＲＯＭ３５０１、またはＦＤ３５０２に記憶されて、ＣＤ－ＲＯＭドライブ３４１２またはＦＤドライブ３４１１に挿入され、さらにハードディスク３４１７に転送されても良い。これに代えて、プログラムは、図示しないネットワークを介してコンピュータ３４１に送信され、ハードディスク３４１７に記憶されても良い。プログラムは実行の際にＲＡＭ３４１６にロードされる。プログラムは、ＣＤ－ＲＯＭ３５０１、ＦＤ３５０２またはネットワークから直接、ロードされても良い。

【０１６８】

プログラムは、コンピュータ３４１に、上述した実施の形態のクラスタ生成装置、また

は翻訳装置の機能を実行させるオペレーティングシステム（ＯＳ）、またはサードパーティープログラム等は、必ずしも含まなくても良い。プログラムは、制御された態様で適切な機能（モジュール）を呼び出し、所望の結果が得られるようにする命令の部分のみを含んでいれば良い。コンピュータシステム３４０がどのように動作するかは周知であり、詳細な説明は省略する。

【０１６９】

また、上記プログラムを実行するコンピュータは、単数であってもよく、複数であってもよい。すなわち、集中処理を行ってもよく、あるいは分散処理を行ってもよい。

【０１７０】

本発明は、以上の実施の形態に限定されることなく、種々の変更が可能であり、それらも本発明の範囲内に包含されるものであることは言うまでもない。

【産業上の利用可能性】

【０１７１】

以上のように、本発明にかかる翻訳装置は、精度の高い機械翻訳ができる、という効果を有し、翻訳装置等として有用である。

【図面の簡単な説明】

【０１７２】

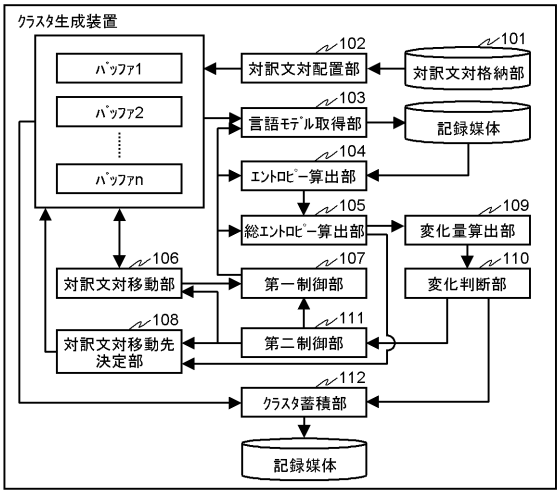
【図１】実施の形態１におけるクラスタ生成装置のブロック図
【図２】同クラスタ生成装置の動作について説明するフローチャート
【図３】同言語モデル作成処理について説明するフローチャート
【図４】同エントロピー算出処理について説明するフローチャート
【図５】同移動先決定処理について説明するフローチャート
【図６】実施の形態２におけるクラスタ生成装置のブロック図
【図７】同翻訳装置の動作について説明するフローチャート
【図８】同評価実験の学習コーパスを示す図
【図９】同評価実験の評価コーパスを示す図
【図１０】同クラスタ数と性能の関係を示す図
【図１１】同補間係数と性能の関係を示す図
【図１２】同クラスタ生成装置等を実現するコンピュータの外観図
【図１３】同クラスタ生成装置等を実現するコンピュータシステムのブロック図

【符号の説明】

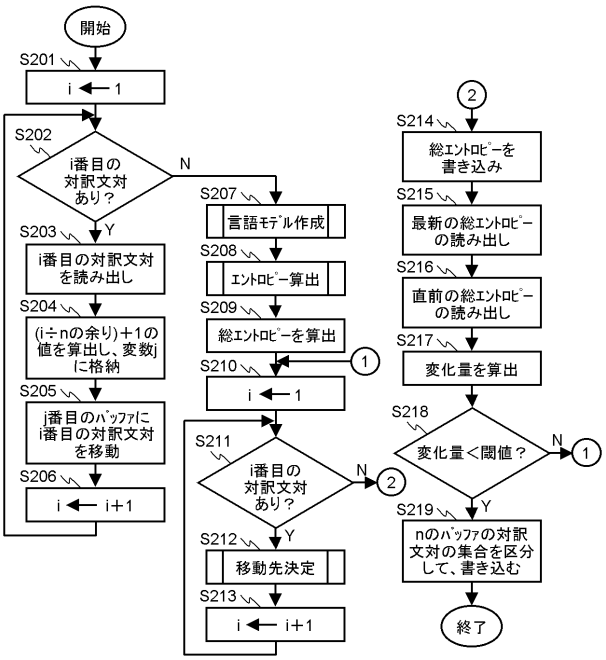
【０１７３】

１０１　対訳文対格納部
１０２　対訳文対配置部
１０３　言語モデル取得部
１０４　エントロピー算出部
１０５　総エントロピー算出部
１０６　対訳文対移動部
１０７　第一制御部
１０８　対訳文対移動先決定部
１０９　変化量算出部
１１０　変化判断部
１１１　第二制御部
１１２　クラスタ蓄積部
５０１　言語モデル格納部
５０２　受付部
５０３　文分割部
５０４　翻訳原文出現確率算出部
５０５　言語モデル決定部
５０６　翻訳部

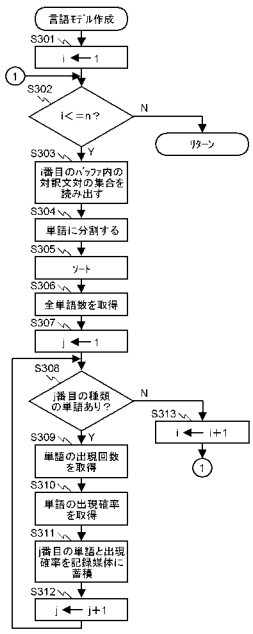
【図１】


クラスタ生成装置
バッファ1
バッファ2
バッファn
対訳文対配置部
102
対訳文対格納部
101
言語モデル取得部
103
記録媒体
エントロピー算出部
104
総エントロピー算出部
105
変化量算出部
109
対訳文対移動部
106
第一制御部
107
変化判断部
110
対訳文対移動先決定部
108
第二制御部
111
クラスタ蓄積部
112
記録媒体

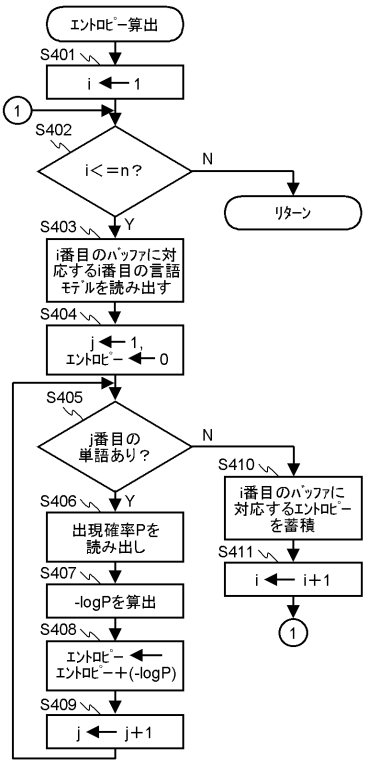
【図２】


開始
S201
i ← 1
S202
i番目の対訳文対あり？
N
Y
S203
i番目の対訳文対を読み出し
S204
(i÷nの余り)＋1の値を算出し、変数jに格納
S205
jバッファにi番目の対訳文対を移動
S206
i ← i+1
S207
言語モデル作成
S208
エントロピー算出
S209
総エントロピーを算出
1
S210
i ← 1
S211
i番目の対訳文対あり？
N
2
Y
S212
移動先決定
S213
i ← i+1
2
S214
総エントロピーを書き込み
S215
最新の総エントロピーの読み出し
S216
直前の総エントロピーの読み出し
S217
変化量を算出
S218
変化量＜閾値？
N
1
Y
S219
nのバッファの対訳文対の集合を区分して、書き込む
終了

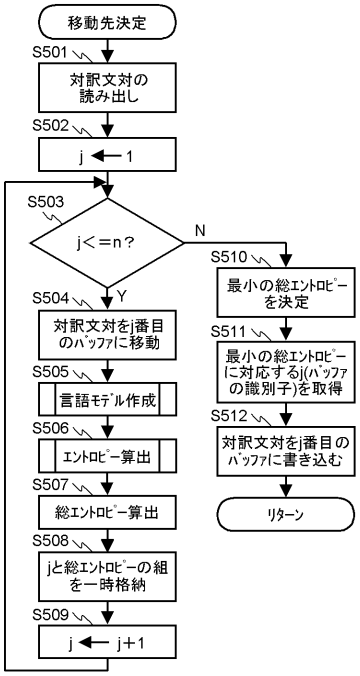
【図３】


言語モデル作成
S301
i ← 1
1
S302
i＜＝n？
N
リターン
Y
S303
iバッファ内の対訳文対の集合を読み出す
S304
単語に分割する
S305
ソート
S306
全単語数を取得
S307
j ← 1
S308
j番目の種類の単語あり？
N
S313
i ← i+1
1
Y
S309
単語の出現回数を取得
S310
単語の出現確率を取得
S311
j番目の単語と出現確率を記録媒体に蓄積
S312
j ← j+1

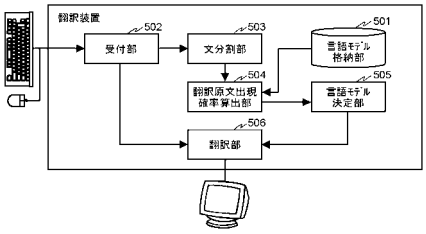
【図４】


エントロピー算出
S401
i ← 1
1
S402
i＜＝n？
N
リターン
Y
S403
iバッファに対応するi番目の言語モデルを読み出す
S404
j ← 1,
エントロピー ← 0
S405
j番目の単語あり？
N
S410
iバッファに対応するエントロピーを蓄積
S411
i ← i+1
1
Y
S406
出現確率Pを読み出し
S407
-logPを算出
S408
エントロピー ← エントロピー＋(-logP)
S409
j ← j+1

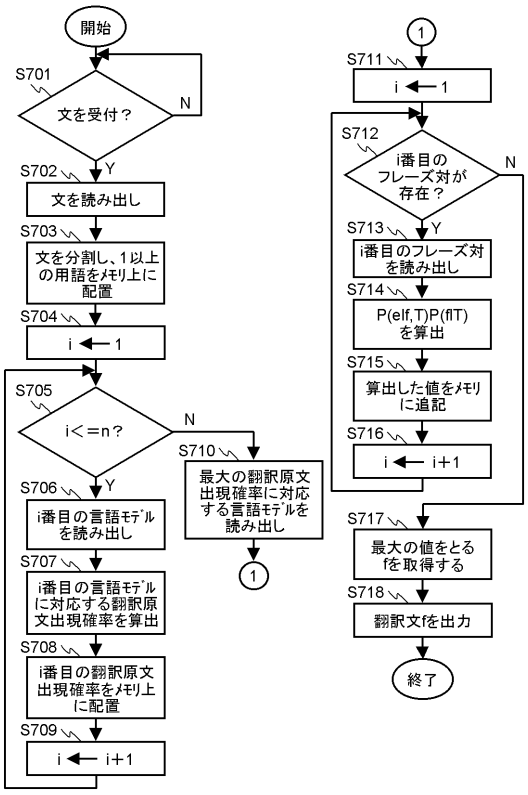
【図５】



【図６】



【図７】



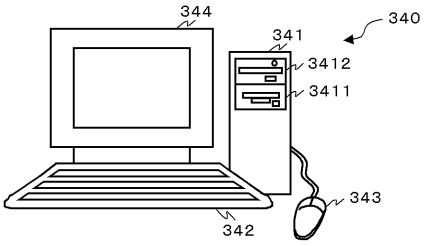
【図８】

言語	文数	総単語数	異なり単語数
日本語	162K	1,449K	18.7K
英語	162K	1,303K	21.0K

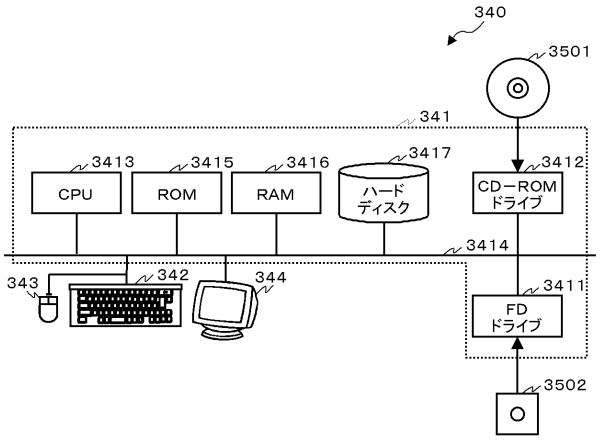
【図９】

言語	文数	総単語数	異なり単語数
日本語	1,524	13,686	2,023
英語	1,524	12,364	1,723

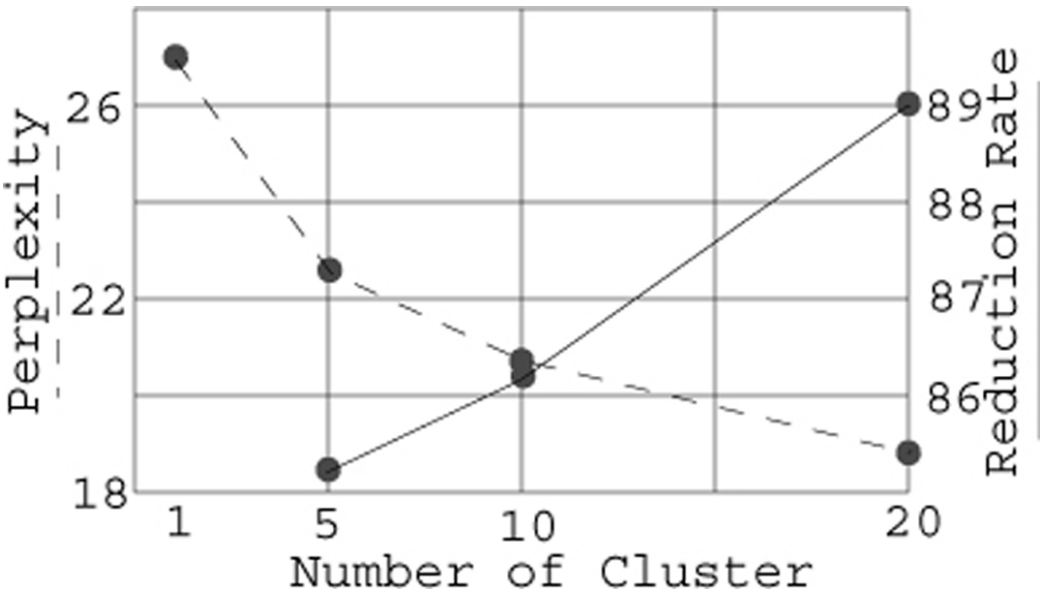
【図１２】



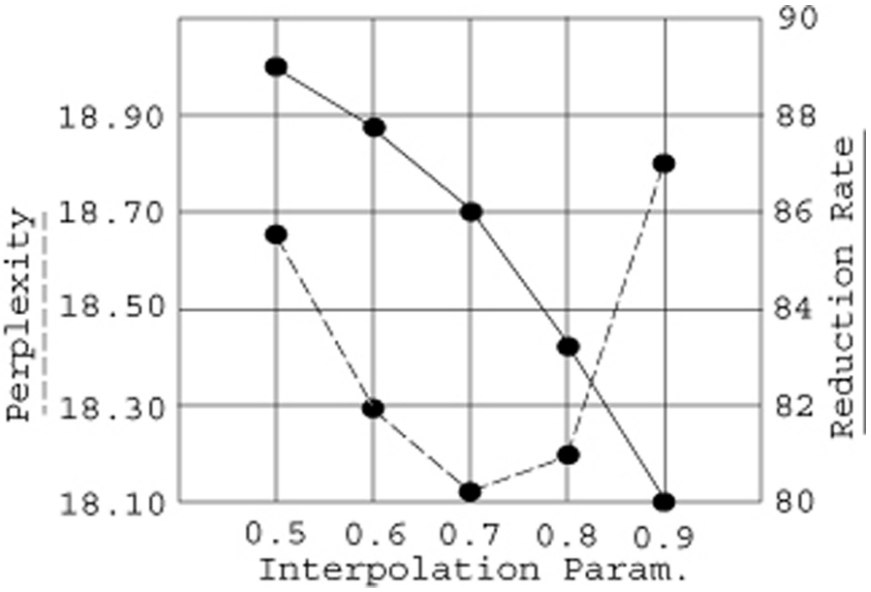
【図１３】


340
3501
341
3413
3415
3416
3417
3412
CPU
ROM
RAM
ハード
ディスク
CD－ROM
ドライブ
3414
343
342
344
3411
FD
ドライブ
3502

【図１０】


Perplexity
26
22
18
1
5
10
20
Number of Cluster
89
88
87
86
Reduction Rate

【図１１】


Perplexity
18.90
18.70
18.50
18.30
18.10
90
88
86
84
82
80
Reduction Rate
0.5 0.6 0.7 0.8 0.9
Interpolation Param.