

(19) 日本国特許庁(JP)

(12) 特許公報(B2)

(11) 特許番号

特許第5067777号
(P5067777)

(45) 発行日 平成24年11月7日(2012.11.7)

(24) 登録日 平成24年8月24日(2012.8.24)

(51) Int.Cl. F 1
G 0 6 F 17/28 (2006.01) G 0 6 F 17/28 U

請求項の数 9 (全 29 頁)

<p>(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) (65) 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) 審査請求日 平成21年8月28日(2009.8.28)</p>	<p>(73) 特許権者 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 審査官 梅本 達雄</p>
---	---

最終頁に続く

(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) 【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

10

20

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、

10

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、

20

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、

30

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。

40

【請求項 5】

前記確率情報は、

単語 3 - g r a m 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

50

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、

10

20

30

40

50

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】

対訳文対配置部、言語モデル取得部、エントロピー算出部、総エントロピー算出部、対訳文対移動部、第一制御部、対訳文対移動先決定部、変化量算出部、変化判断部、第二制御部、およびクラスタ蓄積部により実現され、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、

前記対訳文対配置部が、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、

前記言語モデル取得部が、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、

前記エントロピー算出部が、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、

前記総エントロピー算出部が、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、

前記対訳文対移動部が、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、

前記第一制御部が、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、

前記対訳文対移動先決定部が、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、

前記変化量算出部が、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、

前記変化判断部が、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、

前記第二制御部が、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

10

20

30

40

50

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、前記クラスタ蓄積部が、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 - g r a m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ティスを作成するラティス作成部と、ラティス作成部が作成したラ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, Speec

10

20

30

40

50

h, 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

10

20

30

40

50

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。

10

【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。

20

【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。

30

【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。

40

【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)

50

【 0 0 2 4 】

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。

【 0 0 2 5 】

クラスタ生成装置は、対訳文対格納部 1 0 1、対訳文対配置部 1 0 2、言語モデル取得部 1 0 3、エントローピー算出部 1 0 4、総エントローピー算出部 1 0 5、対訳文対移動部 1 0 6、第一制御部 1 0 7、対訳文対移動先決定部 1 0 8、変化量算出部 1 0 9、変化判断部 1 1 0、第二制御部 1 1 1、クラスタ蓄積部 1 1 2 を具備する。

【 0 0 2 6 】

対訳文対格納部 1 0 1 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 1 0 1 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 1 0 1 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。

【 0 0 2 7 】

対訳文対配置部 1 0 2 は、対訳文対格納部 1 0 1 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 1 0 2 は、対訳文対格納部 1 0 1 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 1 0 2 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 1 0 2 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 1 0 2 は、通常、MPU やメモリ等から実現され得る。対訳文対配置部 1 0 2 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。

【 0 0 2 8 】

言語モデル取得部 1 0 3 は、バッファ毎に、対訳文対配置部 1 0 2 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 - g r a m である。言語モデル取得部 1 0 3 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 1 0 3 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 1 0 3 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。

【 0 0 2 9 】

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「 $-\log P$ 」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「 $-\log P$ 」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」、「 $-\log P$ 」などの演算式の情報予め格納しており、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。

10

20

【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。

30

【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。

40

【0033】

50

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。

10

【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。

20

【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。

30

【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、MPU 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。

40

【0038】

50

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に行く、*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。 10

【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。 20

【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。 30

【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に行く、*i*番目の対訳文対が存在しなければステップS214に行く。 40

【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 50

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。

10

【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。

20

【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、「 $i \leq n$ 」であるか否かを判断する。「 $i \leq n$ 」であればステップS303に行き、「 $i \leq n$ 」でなければ上位関数にリターンする。

30

【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。

40

【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】

50

(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。

10

【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】

20

(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、「 $i \leq n$ 」であるか否かを判断する。「 $i \leq n$ 」であればステップS403に行き、「 $i \leq n$ 」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】

30

(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。

40

【0079】

(ステップS407) エントロピー算出部104は、「 $- \log P$ 」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、「 $- \log P$ 」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】

50

(ステップS 4 1 0) エントロピー算出部 1 0 4 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S 4 1 1) エントロピー算出部 1 0 4 は、カウンタ i を 1、インクリメントする。ステップS 4 0 2 に戻る。

【0084】

次に、ステップS 2 1 2 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S 5 0 1) 対訳文対移動先決定部 1 0 8 は、処理対象の対訳文対を読み出す。 10

【0086】

(ステップS 5 0 2) 対訳文対移動先決定部 1 0 8 は、カウンタ j に 1 を代入する。

【0087】

(ステップS 5 0 3) 対訳文対移動先決定部 1 0 8 は、「 $j \leq n$ 」であるか否かを判断する。「 $j \leq n$ 」であればステップS 5 0 4 に行き、「 $j \leq n$ 」でなければステップS 5 1 0 に行く。

【0088】

(ステップS 5 0 4) 対訳文対移動先決定部 1 0 8 は、ステップS 5 0 1 で読み出した対訳文対を、 j 番目のバッファに移動する。 20

【0089】

(ステップS 5 0 5) 言語モデル取得部 1 0 3 は、ステップS 5 0 4 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S 5 0 6) エントロピー算出部 1 0 4 は、エントロピー算出処理を行う。

【0091】

(ステップS 5 0 7) 総エントロピー算出部 1 0 5 は、ステップS 5 0 6 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S 5 0 8) 対訳文対移動先決定部 1 0 8 は、 j と、ステップS 5 0 7 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。 30

【0093】

(ステップS 5 0 9) 対訳文対移動先決定部 1 0 8 は、カウンタ j を 1、インクリメントする。ステップS 5 0 3 に戻る。

【0094】

(ステップS 5 1 0) 対訳文対移動先決定部 1 0 8 は、ステップS 5 0 8 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S 5 1 1) 対訳文対移動先決定部 1 0 8 は、ステップS 5 1 0 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。 40

【0096】

(ステップS 5 1 2) 対訳文対移動先決定部 1 0 8 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S 5 0 1 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S 5 0 4 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 50

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。

【 0 0 9 9 】

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】

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

【 0 1 0 0 】

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) を新たな変数として数式 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、上述した各パツファの対訳文対の集合に対応する。

【数 2】

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

【 0 1 0 1 】

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】

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

【 0 1 0 2 】

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。

【 0 1 0 3 】

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】

$$\begin{aligned} & \operatorname{argmax}_{f, T} P(f, T|e) \\ & = \operatorname{argmax}_{f, T} P(T|e)P(f|e, T) \end{aligned}$$

【 0 1 0 4 】

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数 (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。

【 0 1 0 5 】

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、それによって求めた (T) を用いて $P(f|e, T)$ を最大化するという手順をとる。

【 0 1 0 6 】

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、

10

20

30

40

50

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

10

20

30

40

50

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 - g r a mの確率である。また、確率情報は、例えば、単語u n i g r a m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や、メニュー画面の制御ソフトウェア等で実現され得る。

10

【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。

20

【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以上の対訳文対中出现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。

30

【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）

40

50

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。

【 0 1 2 3 】

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。

【 0 1 2 4 】

(ステップ S 7 0 1) 受付部 5 0 2 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 S 7 0 2 に行き、翻訳対象の文 (e) を受け付けなければステップ S 7 0 1 に戻る。

【 0 1 2 5 】

(ステップ S 7 0 2) 文分割部 5 0 3 は、ステップ S 7 0 1 で受け付けた文 (e) をメモリ上に配置する。

【 0 1 2 6 】

(ステップ S 7 0 3) 文分割部 5 0 3 は、ステップ S 7 0 2 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。

【 0 1 2 7 】

(ステップ S 7 0 4) 翻訳原文出現確率算出部 5 0 4 は、カウンタ i に 1 を代入する。

【 0 1 2 8 】

(ステップ S 7 0 5) 翻訳原文出現確率算出部 5 0 4 は、「 $i \leq n$ 」であるか否かを判断する。「 $i \leq n$ 」であればステップ S 7 0 6 に行き、「 $i \leq n$ 」でなければステップ S 7 1 0 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 5 0 4 は、「n」の値を読み出し、「 $i \leq n$ 」が真であるか、偽であるかを判断する。

【 0 1 2 9 】

(ステップ S 7 0 6) 翻訳原文出現確率算出部 5 0 4 は、i 番目の言語モデルを言語モデル格納部 5 0 1 から読み出す。

【 0 1 3 0 】

(ステップ S 7 0 7) 翻訳原文出現確率算出部 5 0 4 は、i 番目の言語モデルに対する、文分割部 5 0 3 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 5 0 4 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 5 0 4 は、例えば、文分割部 5 0 3 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。

【 0 1 3 1 】

(ステップ S 7 0 8) 翻訳原文出現確率算出部 5 0 4 は、ステップ S 7 0 7 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。

【 0 1 3 2 】

(ステップ S 7 0 9) 翻訳原文出現確率算出部 5 0 4 は、カウンタ i を 1、インクリメントする。ステップ S 7 0 5 に行く。

【 0 1 3 3 】

(ステップ S 7 1 0) 言語モデル決定部 5 0 5 は、ステップ S 7 0 8 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 5 0 6 は、言語モデル決定部 5 0 5 が決定した言語モデルを、言語モデル格納部 5 0 1 から読み出す。

【 0 1 3 4 】

(ステップ S 7 1 1) 翻訳部 5 0 6 は、カウンタ i に 1 を代入する。

10

20

30

40

50

【 0 1 3 5 】

(ステップ S 7 1 2) 翻訳部 5 0 6 は、i 番目のフレーズ対が言語モデル格納部 5 0 1 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 S 7 1 3 に行き、i 番目のフレーズ対が存在しなければステップ S 7 1 7 に行く。

【 0 1 3 6 】

(ステップ S 7 1 3) 翻訳部 5 0 6 は、i 番目のフレーズ対を、言語モデル格納部 5 0 1 から読み出す。

【 0 1 3 7 】

(ステップ S 7 1 4) 翻訳部 5 0 6 は、ステップ S 7 1 3 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。

10

【 0 1 3 8 】

(ステップ S 7 1 5) 翻訳部 5 0 6 は、ステップ S 7 1 4 で算出した結果をメモリ上に追記する。

【 0 1 3 9 】

(ステップ S 7 1 6) 翻訳部 5 0 6 は、カウンタ i を 1、インクリメントする。ステップ S 7 1 2 に行く。

【 0 1 4 0 】

(ステップ S 7 1 7) 翻訳部 5 0 6 は、ステップ S 7 1 5 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。

【 0 1 4 1 】

(ステップ S 7 1 8) 翻訳部 5 0 6 は、ステップ S 7 1 7 で構成した翻訳文 (f) を出力する。処理を終了する。

20

【 0 1 4 2 】

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。

【 0 1 4 3 】

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。

30

【 0 1 4 4 】

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】

$$\begin{aligned} P(f|e, T) &= P(e|f, T)P(f, T)/P(e, T) \\ &= P(e|f, T)P(f|T)P(T)/P(e, T) \end{aligned}$$

【 0 1 4 5 】

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 S 7 1 4 において、翻訳部 5 0 6 は、「 $P(e|f, T)P(f|T)$ 」を算出した。

40

【 0 1 4 6 】

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

50

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トリングの際の、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。

10

20

30

40

50

【 0 1 5 6 】

図 1 1 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 1 8 . 1 を示しており、これはベースラインの 2 6 . 9 に対して約 3 3 % 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 9 2 % である。またこの時のパープレキシティは 2 3 . 1 であった。

【 0 1 5 7 】

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。

10

【 0 1 5 8 】

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。

【 0 1 5 9 】

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。

20

30

【 0 1 6 0 】

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。

【 0 1 6 1 】

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

40

【 0 1 6 2 】

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。

【 0 1 6 3 】

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

50

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。

10

【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。

20

【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。

30

【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。

40

【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】

50

【 0 1 7 2 】

【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

10

【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】

【 0 1 7 3 】

1 0 1 対訳文対格納部

1 0 2 対訳文対配置部

1 0 3 言語モデル取得部

1 0 4 エントロピー算出部

20

1 0 5 総エントロピー算出部

1 0 6 対訳文対移動部

1 0 7 第一制御部

1 0 8 対訳文対移動先決定部

1 0 9 変化量算出部

1 1 0 変化判断部

1 1 1 第二制御部

1 1 2 クラスタ蓄積部

5 0 1 言語モデル格納部

5 0 2 受付部

30

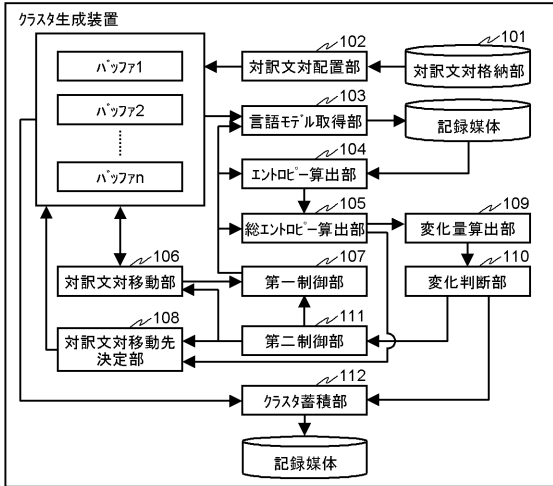
5 0 3 文分割部

5 0 4 翻訳原文出現確率算出部

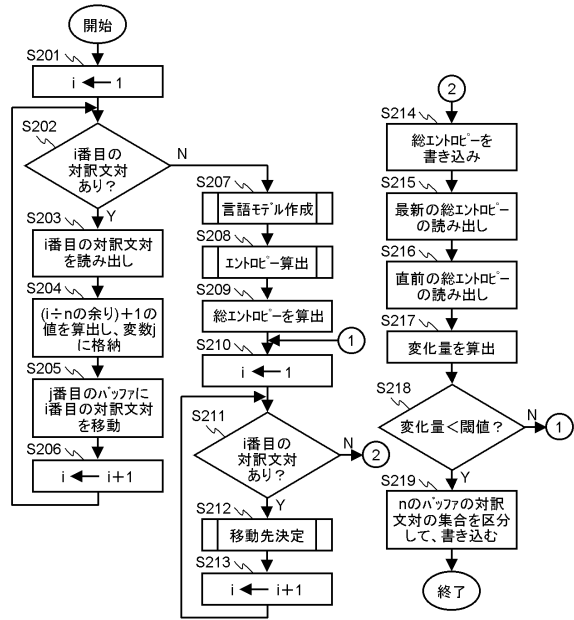
5 0 5 言語モデル決定部

5 0 6 翻訳部

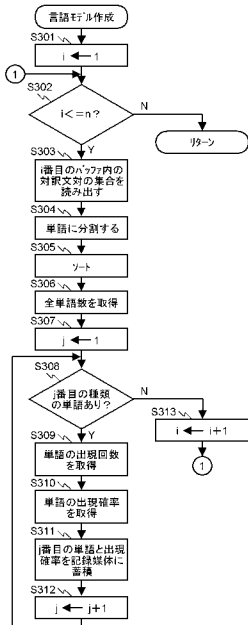
【図1】



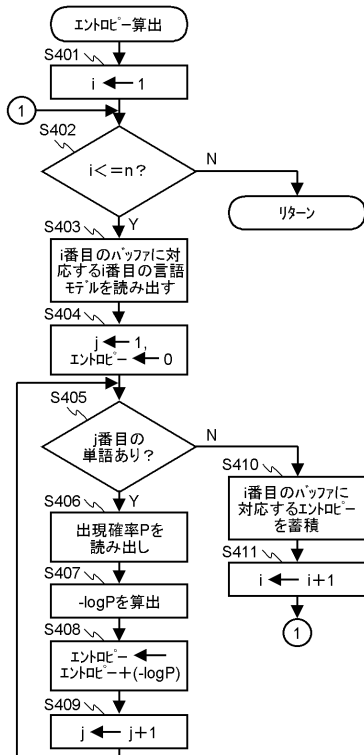
【図2】



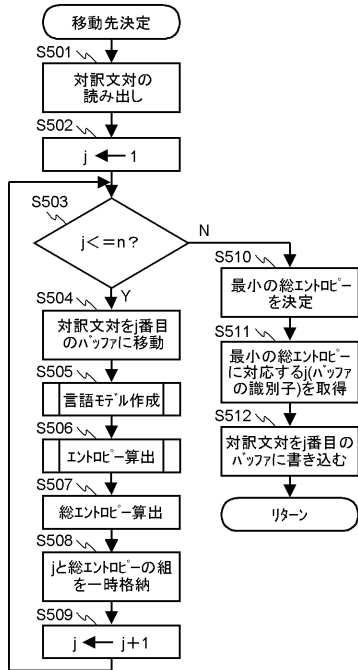
【図3】



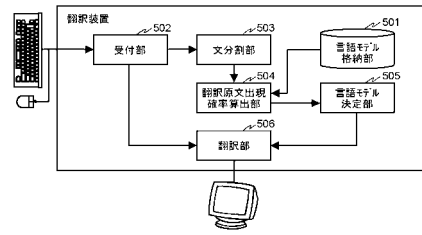
【図4】



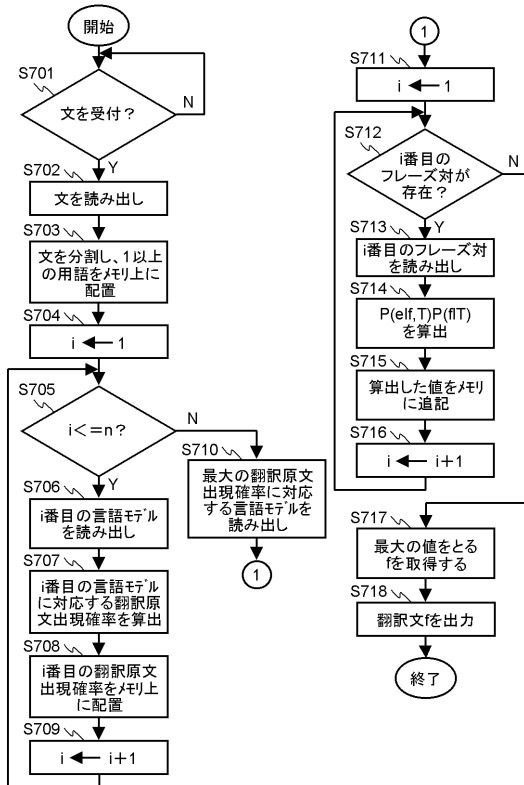
【図5】



【図6】



【図7】



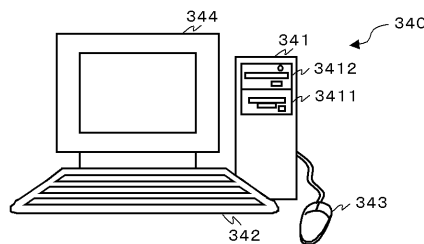
【図8】

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

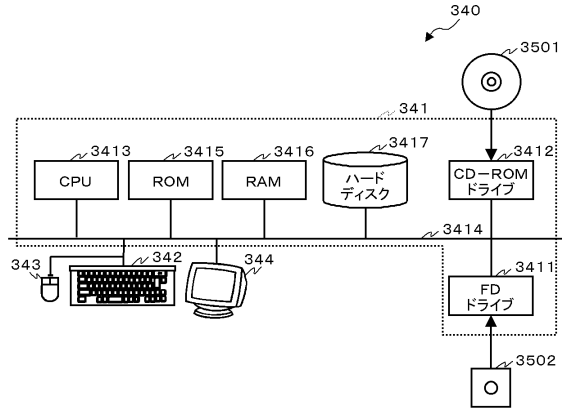
【図9】

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

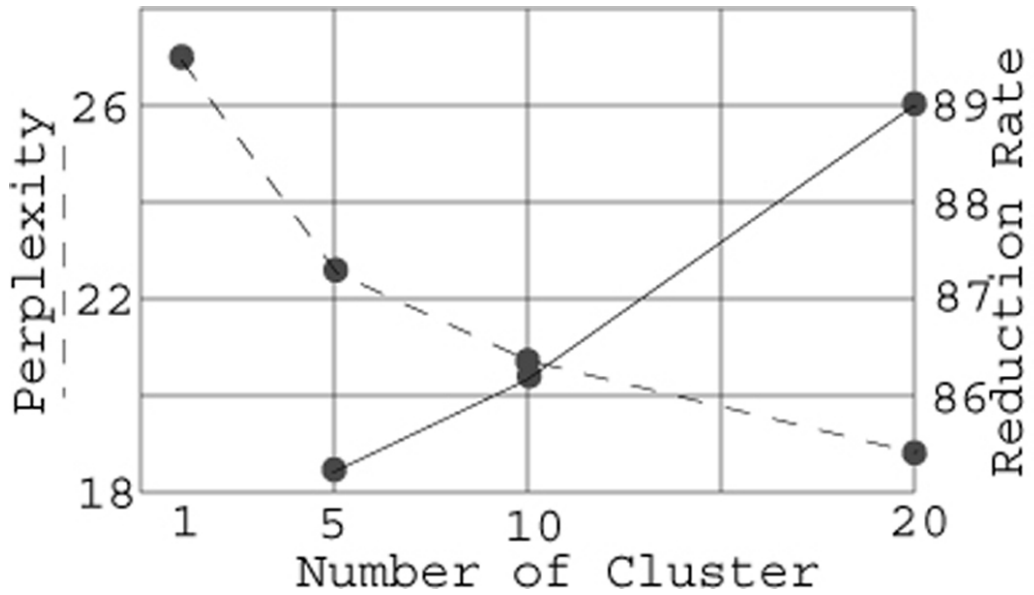
【図12】



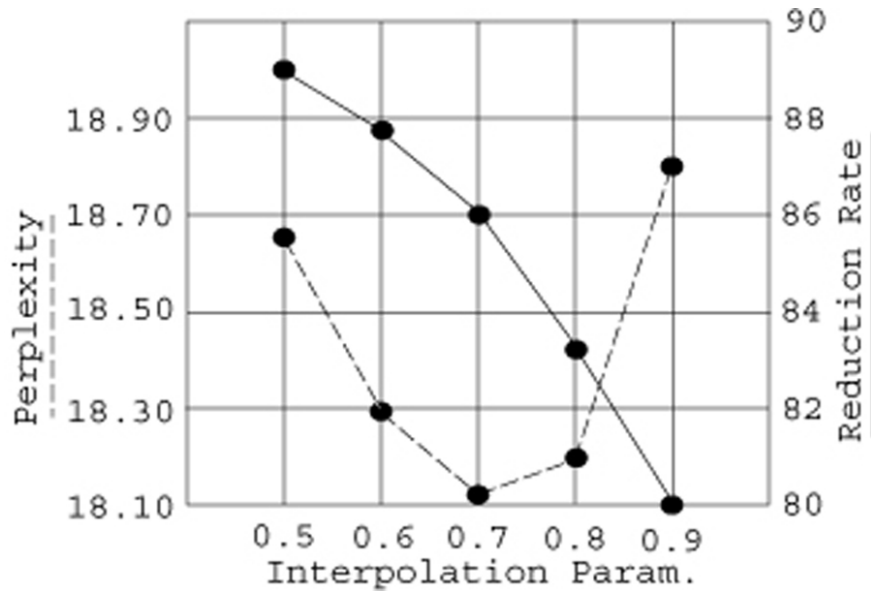
【図13】



【図10】



【図 11】



フロントページの続き

(56)参考文献 特開2004-038406(JP,A)

(58)調査した分野(Int.Cl., DB名)

G06F 17/20 - 17/28