(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6044996号
(P6044996)
(45)発行日　平成28年12月14日(2016.12.14)　(24)登録日　平成28年11月25日(2016.11.25)

(51)Int.Cl.　F I
G06F 17/28 (2006.01)　G06F 17/28 627

請求項の数 5　(全 26 頁)

(21)出願番号　特願2013-149869(P2013-149869)
(22)出願日　平成25年7月18日(2013.7.18)
(65)公開番号　特開2015-22508(P2015-22508A)
(43)公開日　平成27年2月2日(2015.2.2)
審査請求日　平成27年8月27日(2015.8.27)

(73)特許権者　000004226
日本電信電話株式会社
東京都千代田区大手町一丁目５番１号
(73)特許権者　504132272
国立大学法人京都大学
京都府京都市左京区吉田本町３６番地１
(74)代理人　110001519
特許業務法人太陽国際特許事務所
(72)発明者　須藤　克仁
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　永田　昌明
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

最終頁に続く

(54)【発明の名称】文字列対応付け装置、方法、及びプログラム

(57)【特許請求の範囲】
【請求項１】
異なる第１の言語及び第２の言語にそれぞれ属する同じ意味の文字列の組み合わせである文字列組について、前記第１の言語の文字列と、前記第２の言語の文字列との間で文字の対応付けを行う文字列対応付け装置であって、
前記文字列組を複数組記憶した文字列組データベースに記憶された前記文字列組の各々に対して、前記文字列組の各文字列を、前記文字列の先頭から順番に、他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す前置非翻字セグメントと、前記他方の言語の部分文字列と翻字関係にある０文字以上の部分文字列を示す翻字セグメントと、前記他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す後置非翻字セグメントとで構成したときに、前記第１の言語の部分文字列が、前記第２の言語の部分文字列と翻字関係にない非翻字部分である確率を表す非翻字モデル選択確率と、前記第２の言語の部分文字列が、前記第１の言語の部分文字列と翻字関係にない非翻字部分である確率を表す非翻字モデル選択確率と、前記第１の言語の部分文字列が、前記第２の言語の部分文字列と翻字関係にある翻字部分であり、かつ前記第２の言語の部分文字列が、前記第１の言語の部分文字列と翻字関係にある翻字部分である確率を表す翻字モデル選択確率と、前記第１の言語の文字列のうちの前記前置非翻字セグメントの部分文字列、及び前記後置非翻字セグメントの部分文字列の各々に対する前記第１の言語における生成確率を表す非翻字モデル生成確率と、前記第２の言語の文字列のうちの前記前置非翻字セグメントの部分文字列、及び前記後置非翻字セグメントの部分文字列の各々に対する前記第２の

言語における生成確率を表す非翻字モデル生成確率と、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間の部分文字列の各ペアに対する同時生成確率を表す翻字モデル生成確率と、に基づいて尤もらしくなるように、前記文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行う対応付け計算部
を含み、
前記対応付け計算部は、
前記第１の言語の前記非翻字モデル選択確率と、前記第２の言語の前記非翻字モデル選択確率と、前記第２の言語の各部分文字列に対する前記翻字モデル選択確率と、前記第１の言語の各部分文字列に対する前記非翻字モデル生成確率と、前記第２の言語の各部分文字列に対する前記非翻字モデル生成確率と、前記第１の言語の部分文字列と前記第２の言語の部分文字列との間の部分文字列の各ペアに対する前記翻字モデル生成確率と、に対して初期値を各々設定する初期値設定部と、
前記初期値設定部によって設定され、又は前回更新された、前記非翻字モデル選択確率、前記翻字モデル選択確率、前記非翻字モデル生成確率、及び前記翻字モデル生成確率に基づいて、前記文字列組の各々に対して、前記第１の言語の文字列のうちの部分文字列と、前記第２の言語の文字列のうちの部分文字列との間の部分文字列の各ペアについて、前記ペアが翻訳関係にある期待値を計算し、前記第１の言語の文字列のうちの各部分文字列について、前記部分文字列が非翻字部分である期待値を計算し、前記第２の言語の文字列のうちの各部分文字列について、前記部分文字列が非翻字部分である期待値を計算する期待値計算部と、
前記文字列組の各々に対して前記期待値計算部によって計算された各ペアに対する前記翻訳関係にある期待値、前記第１の言語の各部分文字列についての前記非翻字部分である期待値、及び前記第２の言語の各部分文字列についての前記非翻字部分である期待値に基づいて、前記非翻字モデル選択確率、前記翻字モデル選択確率、前記非翻字モデル生成確率、及び前記翻字モデル生成確率を更新するパラメータ更新部と、
予め定められた停止条件が満たされたか否かを判定し、前記停止条件が満たされるまで、前記期待値計算部による計算、及び前記パラメータ更新部による更新を繰り返す停止条件判定部と、を含み
前記文字列組の各々に対して、前記非翻字モデル選択確率、前記翻字モデル選択確率、前記非翻字モデル生成確率、及び前記翻字モデル生成確率の各々に基づいて、前記文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行う文字列対応付け装置。

【請求項２】

異なる第１の言語及び第２の言語にそれぞれ属する同じ意味の文字列の組み合わせである文字列組について、前記第１の言語の文字列と、前記第２の言語の文字列との間で文字の対応付けを行う文字列対応付け装置であって、
前記文字列組を複数組記憶した文字列組データベースに記憶された前記文字列組の各々に対して、前記文字列組の各文字列を、前記文字列の先頭から順番に、他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す前置非翻字セグメントと、前記他方の言語の部分文字列と翻字関係にある０文字以上の部分文字列を示す翻字セグメントと、前記他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す後置非翻字セグメントとで構成したときに、前記第１の言語の部分文字列が、前記第２の言語の部分文字列と翻字関係にない非翻字部分である確率を表す非翻字モデル選択確率と、前記第２の言語の部分文字列が、前記第１の言語の部分文字列と翻字関係にない非翻字部分である確率を表す非翻字モデル選択確率と、前記第１の言語の部分文字列が、前記第２の言語

の部分文字列と翻字関係にある翻字部分であり、かつ前記第２の言語の部分文字列が、前記第１の言語の部分文字列と翻字関係にある翻字部分である確率を表す翻字モデル選択確率と、前記第１の言語の文字列のうちの前記前置非翻字セグメントの部分文字列、及び前記後置非翻字セグメントの部分文字列の各々に対する前記第１の言語における生成確率を表す非翻字モデル生成確率と、前記第２の言語の文字列のうちの前記前置非翻字セグメントの部分文字列、及び前記後置非翻字セグメントの部分文字列の各々に対する前記第２の言語における生成確率を表す非翻字モデル生成確率と、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間の部分文字列の各ペアに対する同時生成確率を表す翻字モデル生成確率と、に基づいて尤もらしくなるように、前記文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行う対応付け計算部

を含み、

前記対応付け計算部は、

前記文字列組の各々に対して、前記文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行って、対応付けの初期設定を行う初期対応設定部と、

前記初期対応設定部によって設定され、又は前回更新された、前記複数組の文字列組のうちの処理対象の文字列組以外の文字列組の各々についての前記対応付けに基づいて、前記第１の言語の前記非翻字モデル選択確率と、前記第２の言語の前記非翻字モデル選択確率と、前記翻字モデル選択確率と、前記処理対象の文字列組の前記第１の言語の文字列のうちの各部分文字列に対する前記非翻字モデル生成確率と、前記処理対象の文字列組の前記第２の言語の文字列のうちの各部分文字列に対する前記非翻字モデル生成確率と、前記処理対象の文字列組の前記第１の言語の文字列のうちの部分文字列と、前記第２の言語の文字列のうちの部分文字列との間の部分文字列の各ペアに対する前記翻字モデル生成確率と、を計算するフィルタリング部と、

前記フィルタリング部によって計算された前記非翻字モデル選択確率、前記翻字モデル選択確率、前記非翻字モデル生成確率、及び前記翻字モデル生成確率に基づいて、前記処理対象の文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行って、前記処理対象の文字列組に対する前記対応付けを更新するサンプリング部と、

予め定められた停止条件が満たされたか否かを判定し、前記停止条件が満たされるまで、各文字列組を処理対象とした前記フィルタリング部による計算及び前記サンプリング部による更新を繰り返す停止条件判定部と、を含む文字列対応付け装置。

【請求項３】

対応付け計算部を含み、異なる第１の言語及び第２の言語にそれぞれ属する同じ意味の文字列の組み合わせである文字列組について、前記第１の言語の文字列と、前記第２の言語の文字列との間で文字の対応付けを行う文字列対応付け装置における文字列対応付け方法であって、

対応付け計算部によって、前記文字列組を複数組記憶した文字列組データベースに記憶された前記文字列組の各々に対して、前記文字列組の各文字列を、前記文字列の先頭から順番に、他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す前置非翻字セグメントと、前記他方の言語の部分文字列と翻字関係にある０文字以上の部分文字列を示す翻字セグメントと、前記他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す後置非翻字セグメントとで構成したときに、前記第１の言語の部分文字

列が、前記第２の言語の部分文字列と翻字関係にない非翻字部分である確率を表す非翻字モデル選択確率と、前記第２の言語の部分文字列が、前記第１の言語の部分文字列と翻字関係にない非翻字部分である確率を表す非翻字モデル選択確率と、前記第１の言語の部分文字列が、前記第２の言語の部分文字列と翻字関係にある翻字部分であり、かつ前記第２の言語の部分文字列が、前記第１の言語の部分文字列と翻字関係にある翻字部分である確率を表す翻字モデル選択確率と、前記第１の言語の文字列のうちの前記前置非翻字セグメントの部分文字列、及び前記後置非翻字セグメントの部分文字列の各々に対する前記第１の言語における生成確率を表す非翻字モデル生成確率と、前記第２の言語の文字列のうちの前記前置非翻字セグメントの部分文字列、及び前記後置非翻字セグメントの部分文字列の各々に対する前記第２の言語における生成確率を表す非翻字モデル生成確率と、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間の部分文字列の各ペアに対する同時生成確率を表す翻字モデル生成確率と、に基づいて尤もらしくなるように、前記文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行い、

前記対応付け計算部は、初期値設定部と、期待値計算部と、パラメータ更新部と、停止条件判定部と、を含み

前記初期値設定部によって、前記第１の言語の前記非翻字モデル選択確率と、前記第２の言語の前記非翻字モデル選択確率と、前記第２の言語の各部分文字列に対する前記翻字モデル選択確率と、前記第１の言語の各部分文字列に対する前記非翻字モデル生成確率と、前記第２の言語の各部分文字列に対する前記非翻字モデル生成確率と、前記第１の言語の部分文字列と前記第２の言語の部分文字列との間の部分文字列の各ペアに対する前記翻字モデル生成確率と、に対して初期値を各々設定するステップと、

前記期待値計算部によって、前記初期値設定部によって設定され、又は前回更新された、前記非翻字モデル選択確率、前記翻字モデル選択確率、前記非翻字モデル生成確率、及び前記翻字モデル生成確率に基づいて、前記文字列組の各々に対して、前記第１の言語の文字列のうちの部分文字列と、前記第２の言語の文字列のうちの部分文字列との間の部分文字列の各ペアについて、前記ペアが翻訳関係にある期待値を計算し、前記第１の言語の文字列のうちの各部分文字列について、前記部分文字列が非翻字部分である期待値を計算し、前記第２の言語の文字列のうちの各部分文字列について、前記部分文字列が非翻字部分である期待値を計算するステップと、

前記パラメータ更新部によって、前記文字列組の各々に対して前記期待値計算部によって計算された各ペアに対する前記翻訳関係にある期待値、前記第１の言語の各部分文字列についての前記非翻字部分である期待値、及び前記第２の言語の各部分文字列についての前記非翻字部分である期待値に基づいて、前記非翻字モデル選択確率、前記翻字モデル選択確率、前記非翻字モデル生成確率、及び前記翻字モデル生成確率を更新するステップと、

停止条件判定部によって、予め定められた停止条件が満たされたか否かを判定し、前記停止条件が満たされるまで、前記期待値計算部による計算、及び前記パラメータ更新部による更新を繰り返すステップと、を含み

前記文字列組の各々に対して、前記非翻字モデル選択確率、前記翻字モデル選択確率、前記非翻字モデル生成確率、及び前記翻字モデル生成確率の各々に基づいて、前記文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行う文字列対応付け方法。

【請求項４】

対応付け計算部を含み、異なる第１の言語及び第２の言語にそれぞれ属する同じ意味の

文字列の組み合わせである文字列組について、前記第１の言語の文字列と、前記第２の言語の文字列との間で文字の対応付けを行う文字列対応付け装置における文字列対応付け方法であって、

対応付け計算部によって、前記文字列組を複数組記憶した文字列組データベースに記憶された前記文字列組の各々に対して、前記文字列組の各文字列を、前記文字列の先頭から順番に、他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す前置非翻字セグメントと、前記他方の言語の部分文字列と翻字関係にある０文字以上の部分文字列を示す翻字セグメントと、前記他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す後置非翻字セグメントとで構成したときに、前記第１の言語の部分文字列が、前記第２の言語の部分文字列と翻字関係にない非翻字部分である確率を表す非翻字モデル選択確率と、前記第２の言語の部分文字列が、前記第１の言語の部分文字列と翻字関係にない非翻字部分である確率を表す非翻字モデル選択確率と、前記第１の言語の部分文字列が、前記第２の言語の部分文字列と翻字関係にある翻字部分であり、かつ前記第２の言語の部分文字列が、前記第１の言語の部分文字列と翻字関係にある翻字部分である確率を表す翻字モデル選択確率と、前記第１の言語の文字列のうちの前記前置非翻字セグメントの部分文字列、及び前記後置非翻字セグメントの部分文字列の各々に対する前記第１の言語における生成確率を表す非翻字モデル生成確率と、前記第２の言語の文字列のうちの前記前置非翻字セグメントの部分文字列、及び前記後置非翻字セグメントの部分文字列の各々に対する前記第２の言語における生成確率を表す非翻字モデル生成確率と、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間の部分文字列の各ペアに対する同時生成確率を表す翻字モデル生成確率と、に基づいて尤もらしくなるように、前記文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行い、

前記対応付け計算部は、初期対応設定部と、フィルタリング部と、サンプリング部と、停止条件判定部と、を含み、

前記初期対応設定部によって、前記文字列組の各々に対して、前記文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行って、対応付けの初期設定を行うステップと、

前記フィルタリング部によって、前記初期対応設定部によって設定され、又は前回更新された、前記複数組の文字列組のうちの処理対象の文字列組以外の文字列組の各々についての前記対応付けに基づいて、前記第１の言語の前記非翻字モデル選択確率と、前記第２の言語の前記非翻字モデル選択確率と、前記翻字モデル選択確率と、前記処理対象の文字列組の前記第１の言語の文字列のうちの各部分文字列に対する前記非翻字モデル生成確率と、前記処理対象の文字列組の前記第２の言語の文字列のうちの各部分文字列に対する前記非翻字モデル生成確率と、前記処理対象の文字列組の前記第１の言語の文字列のうちの部分文字列と、前記第２の言語の文字列のうちの部分文字列との間の部分文字列の各ペアに対する前記翻字モデル生成確率と、を計算するステップと、

前記サンプリング部によって、前記フィルタリング部によって計算された前記非翻字モデル選択確率、前記翻字モデル選択確率、前記非翻字モデル生成確率、及び前記翻字モデル生成確率に基づいて、前記処理対象の文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行って、前記処理対象の文字列組に対する前記対応付けを更新するステップと、

前記停止条件判定部によって、予め定められた停止条件が満たされたか否かを判定し、

前記停止条件が満たされるまで、各文字列組を処理対象とした前記フィルタリング部による計算及び前記サンプリング部による更新を繰り返すステップと、を含む文字列対応付け方法。

【請求項５】

コンピュータを、請求項１又は請求項２に記載の文字列対応付け装置の各部として機能させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、文字列対応付け装置、方法、及びプログラムに係り、特に、異なる言語の文字列の組における文字の対応付けを行う文字列対応付け装置、方法、及びプログラムに関する。

【背景技術】

【０００２】

ある言語の音韻体系で表記された語句を別の言語の音韻体系での表記に変換する機械翻字を、統計モデルとして表現するために、互いが対応する単語の組を統計モデルの学習のためのデータとして利用して、単語を構成する文字同士の対応関係を推定することが広く行われている。例えば、非特許文献１では、英語の音韻表現と日本語におけるカタカナ語のローマ字化された表記との間での音韻記号-ローマ字間の１対多の対応付け方法について記している。さらに、非特許文献２では、英語の文字と音韻表記との多対多の対応付けについて記している。非特許文献３では、記号の多対多の対応関係を自動的に行うコンピュータプログラムについて記している。

【０００３】

一方で、統計モデルの学習に利用する単語の組を大量に収集しようとすると、ある程度の誤りの混入は避けられない。いわゆる「カタカナ語」と英語の対応で言えば、日英対訳辞書の項目において日本語側がカタカナで表記されているものでも、「コンピュータ」と”computer”のように翻字関係となっているものもあれば、「カブトムシ」と”beetle”のように、カタカナで表記されるが翻字関係とはなっていないものもある。こうした誤った単語対応を統計モデルの学習に利用することでノイズが混入し、統計モデルの質を低下することは避けるべきである。この問題に対して、翻字関係となっている文字間対応の統計モデルと、翻字関係となっておらず２言語間で独立な文字列の統計モデルを利用した翻字対応付け方法が提案されており、有効に働くことが示されている（例えば、非特許文献４）。対訳辞書の存在を仮定しない非特許文献５のような「統計的機械翻訳」と呼ばれる技術分野においては、対訳文中の共起関係等を用いて自動的に単語対応を得ている。この自動的に得られた単語対応から翻字対応関係を得ようとすると単語対応に誤りが含まれる可能性も高くなるが、非特許文献４の方法により、１対１の単語対応組を、翻字となっている単語組と翻字となっていない単語組を自動的に分類し、翻字となっている単語組からのみ翻字対応の統計モデルを学習することが可能となる。また、非特許文献６には、上記非特許文献３の技術を、翻字でない文字列に対応させた場合について記載されている。

【先行技術文献】

【非特許文献】

【０００４】

【非特許文献１】Kevin Knight and Jonathan Graehl、「Machine Transliteration」、Computational Linguistics、1998、Volume 24、Number 4、p.599-612

【非特許文献２】Sittichai Jiampojamarn他２名、「Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion」、Proceedings of NAACLHLT、2007、p.372-379

【非特許文献３】Andrew Finch and Eiichiro Sumita、「A Bayesian Model of Bilingual Segmentation for Transliteration」、Proceedings of International Workshop on Spoken Language Translation、2010

【非特許文献４】Hassan Sajjad他２名、「A Statistical Model for Unsupervised and Semisupervised Transliteration Mining」、Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics、2012、p.469-477
【非特許文献５】Philipp Koehn他２名、「Statistical Phrase-Based Translation」、Proceedings of HLT-NAACL、2003、p.48-54
【非特許文献６】Ohnmar Htun他３名、「Improving Transliteration Mining by Integrating Expert Knowledge with Statistical Approaches」、International Journal of Computer Applications、2012、58(17)、p.12-22

【発明の概要】
【発明が解決しようとする課題】
【０００５】
非特許文献４の方法は単語対応が「翻字である」か「翻字でない」かを識別するため、複合語等、複数の単語の間の対応関係において「部分的には翻字であるが、その他の部分は翻字になっていない」ような場合を識別することができない。非特許文献５のような句に基づく統計的機械翻訳においては、２言語間の句と句の対訳関係を自動的に推定するため、部分的に訳語でないものが含まれることがあり、非特許文献４の方法では十分な識別を行うことが期待できず、そこから得られる翻字対応および翻字モデルの正確性に問題が生ずる。例えば、「コンピュータ」に対して”the computer”という句が対応しているという状況においては、”the”に対応する文字列がカタカナ語側にはないため、「翻字である」か「翻字でない」かの２値分類は適さない。
【０００６】
本発明は、上記の事情を鑑みてなされたもので、異なる言語の文字列の組における文字の対応付けを精度よく行うことができる文字列対応付け装置、方法、及びプログラムを提供することを目的とする。
【課題を解決するための手段】
【０００７】
上記の目的を達成するために本発明に係る文字列対応付け装置は、異なる第１の言語及び第２の言語にそれぞれ属する同じ意味の文字列の組み合わせである文字列組について、前記第１の言語の文字列と、前記第２の言語の文字列との間で文字の対応付けを行う文字列対応付け装置であって、前記文字列組を複数組記憶した文字列組データベースに記憶された前記文字列組の各々に対して、前記文字列組の各文字列を、前記文字列の先頭から順番に、他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す前置非翻字セグメントと、前記他方の言語の部分文字列と翻字関係にある０文字以上の部分文字列を示す翻字セグメントと、前記他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す後置非翻字セグメントとで構成したときに、前記第１の言語の部分文字列が、前記第２の言語の部分文字列と翻字関係にない非翻字部分である確率を表す非翻字モデル選択確率と、前記第２の言語の部分文字列が、前記第１の言語の部分文字列と翻字関係にない非翻字部分である確率を表す非翻字モデル選択確率と、前記第１の言語の部分文字列が、前記第２の言語の部分文字列と翻字関係にある翻字部分であり、かつ前記第２の言語の部分文字列が、前記第１の言語の部分文字列と翻字関係にある翻字部分である確率を表す翻字モデル選択確率と、前記第１の言語の文字列のうちの前記前置非翻字セグメントの部分文字列、及び前記後置非翻字セグメントの部分文字列の各々に対する前記第１の言語における生成確率を表す非翻字モデル生成確率と、前記第２の言語の文字列のうちの前記前置非翻字セグメントの部分文字列、及び前記後置非翻字セグメントの部分文字列の各々に対する前記第２の言語における生成確率を表す非翻字モデル生成確率と、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間の部分文字列の各ペアに対する同時生成確率を表す翻字モデル生成確率と、に基づいて尤もらしくなるように、前記文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記

第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行う対応付け計算部を含んで構成されている。

【０００８】

本発明に係る文字列対応付け方法は、対応付け計算部を含み、異なる第１の言語及び第２の言語にそれぞれ属する同じ意味の文字列の組み合わせである文字列組について、前記第１の言語の文字列と、前記第２の言語の文字列との間で文字の対応付けを行う文字列対応付け装置における文字列対応付け方法であって、対応付け計算部によって、前記文字列組を複数組記憶した文字列組データベースに記憶された前記文字列組の各々に対して、前記文字列組の各文字列を、前記文字列の先頭から順番に、他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す前置非翻字セグメントと、前記他方の言語の部分文字列と翻字関係にある０文字以上の部分文字列を示す翻字セグメントと、前記他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す後置非翻字セグメントとで構成したときに、前記第１の言語の部分文字列が、前記第２の言語の部分文字列と翻字関係にない非翻字部分である確率を表す非翻字モデル選択確率と、前記第２の言語の部分文字列が、前記第１の言語の部分文字列と翻字関係にない非翻字部分である確率を表す非翻字モデル選択確率と、前記第１の言語の部分文字列が、前記第２の言語の部分文字列と翻字関係にある翻字部分であり、かつ前記第２の言語の部分文字列が、前記第１の言語の部分文字列と翻字関係にある翻字部分である確率を表す翻字モデル選択確率と、前記第１の言語の文字列のうちの前記前置非翻字セグメントの部分文字列、及び前記後置非翻字セグメントの部分文字列の各々に対する前記第１の言語における生成確率を表す非翻字モデル生成確率と、前記第２の言語の文字列のうちの前記前置非翻字セグメントの部分文字列、及び前記後置非翻字セグメントの部分文字列の各々に対する前記第２の言語における生成確率を表す非翻字モデル生成確率と、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間の部分文字列の各ペアに対する同時生成確率を表す翻字モデル生成確率と、に基づいて尤もらしくなるように、前記文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行うステップを含む。

【０００９】

また、本発明に係る前記対応付け計算部は、前記第１の言語の前記非翻字モデル選択確率と、前記第２の言語の前記非翻字モデル選択確率と、前記第２の言語の各部分文字列に対する前記翻字モデル選択確率と、前記第１の言語の各部分文字列に対する前記非翻字モデル生成確率と、前記第２の言語の各部分文字列に対する前記非翻字モデル生成確率と、前記第１の言語の部分文字列と前記第２の言語の部分文字列との間の部分文字列の各ペアに対する前記翻字モデル生成確率と、に対して初期値を各々設定する初期値設定部と、前記初期値設定部によって設定され、又は前回更新された、前記非翻字モデル選択確率、前記翻字モデル選択確率、前記非翻字モデル生成確率、及び前記翻字モデル生成確率に基づいて、前記文字列組の各々に対して、前記第１の言語の文字列のうちの部分文字列と、前記第２の言語の文字列のうちの部分文字列との間の部分文字列の各ペアについて、前記ペアが翻訳関係にある期待値を計算し、前記第１の言語の文字列のうちの各部分文字列について、前記部分文字列が非翻字部分である期待値を計算し、前記第２の言語の文字列のうちの各部分文字列について、前記部分文字列が非翻字部分である期待値を計算する期待値計算部と、前記文字列組の各々に対して前記期待値計算部によって計算された各ペアに対する前記翻訳関係にある期待値、前記第１の言語の各部分文字列についての前記非翻字部分である期待値、及び前記第２の言語の各部分文字列についての前記非翻字部分である期待値に基づいて、前記非翻字モデル選択確率、前記翻字モデル選択確率、前記非翻字モデル生成確率、及び前記翻字モデル生成確率を更新するパラメータ更新部と、予め定められた停止条件が満たされたか否かを判定し、前記停止条件が満たされるまで、前記期待値計算部による計算、及び前記パラメータ更新部による更新を繰り返す停止条件判定部と、を

含み前記文字列組の各々に対して、前記非翻字モデル選択確率、前記翻字モデル選択確率、前記非翻字モデル生成確率、及び前記翻字モデル生成確率の各々に基づいて、前記文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行うようにすることができる。

【００１０】

また、本発明に係る前記対応付け計算部は、前記文字列組の各々に対して、前記文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行って、対応付けの初期設定を行う初期対応設定部と、前記初期対応設定部によって設定され、又は前回更新された、前記複数組の文字列組のうちの処理対象の文字列組以外の文字列組の各々についての前記対応付けに基づいて、前記第１の言語の前記非翻字モデル選択確率と、前記第２の言語の前記非翻字モデル選択確率と、前記翻字モデル選択確率と、前記処理対象の文字列組の前記第１の言語の文字列のうちの各部分文字列に対する前記非翻字モデル生成確率と、前記処理対象の文字列組の前記第２の言語の文字列のうちの各部分文字列に対する前記非翻字モデル生成確率と、前記処理対象の文字列組の前記第１の言語の文字列のうちの部分文字列と、前記第２の言語の文字列のうちの部分文字列との間の部分文字列の各ペアに対する前記翻字モデル生成確率と、を計算するフィルタリング部と、前記フィルタリング部によって計算された前記非翻字モデル選択確率、前記翻字モデル選択確率、前記非翻字モデル生成確率、及び前記翻字モデル生成確率に基づいて、前記処理対象の文字列組の各文字列を前記前置非翻字セグメント、前記翻字セグメント、及び後置非翻字セグメントで構成し、かつ、前記第１の言語の文字列のうちの前記翻字セグメントの部分文字列と、前記第２の言語の文字列のうちの前記翻字セグメントの部分文字列との間で文字の対応付けを行って、前記処理対象の文字列組に対する前記対応付けを更新するサンプリング部と、予め定められた停止条件が満たされたか否かを判定し、前記停止条件が満たされるまで、各文字列組を処理対象とした前記フィルタリング部による計算及び前記サンプリング部による更新を繰り返す停止条件判定部と、を含むようにすることができる。

【００１１】

本発明に係るプログラムは、コンピュータを、本発明に係る文字列対応付け装置の各部として機能させるためのプログラムである。

【発明の効果】

【００１２】

以上説明したように、本発明の文字列対応付け装置、方法、及びプログラムによれば、異なる第１の言語及び第２の言語にそれぞれ属する同じ意味の文字列の組み合わせである文字列組について、文字列組の各文字列を、文字列の先頭から順番に、前置非翻字セグメントと、翻字セグメントと、後置非翻字セグメントとで構成したときに、翻字モデル選択確率と、非翻字モデル選択確率と、非翻字モデル生成確率と、翻字モデル生成確率と、に基づいて尤もらしくなるように、文字列組の各文字列を前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントで構成し、かつ、第１言語の文字列のうちの翻字セグメントの部分文字列と、第２言語の文字列のうちの翻字セグメントの部分文字列との間の文字を対応付けることにより、異なる言語の文字列組における文字の対応付けを精度よく行うことができる、という効果が得られる。

【図面の簡単な説明】

【００１３】

【図１】本発明の第１の実施の形態に係る文字列対応付け装置の構成を示す概略図である。

【図２】本発明の第１の実施の形態に係る文字列対応付け装置における文字列対応付け処

理ルーチンの内容を示すフローチャートである。
【図３】本発明の第２の実施の形態に係る文字列対応付け装置の構成を示す概略図である。
【図４】本発明の第２の実施の形態に係る文字列対応付け装置における文字列対応付け処理ルーチンの内容を示すフローチャートである。
【図５】対応付けの一例として、本実施の形態に係る文字列対応付け装置に入力された英語とカタカナとを示す図である。
【図６】英語とカタカナの対応付け結果の例を示す図である。
【発明を実施するための形態】
【００１４】
以下、図面を参照して本発明の実施の形態を詳細に説明する。
【００１５】
＜発明の概要＞
本発明の実施の形態は、対応付けの対象となる文字列組について、「翻字である」か「翻字でない」かの区別を部分文字列の単位で識別する。また、単純に部分文字列単位での識別を行うと文字列組の中で「翻字である」部分と「翻字でない」部分が頻繁に入れ替わってしまうことがあり文字列対応方法として適さないため、「翻字である」部分は文字列組において高々１箇所であるという制約を加える。具体的には、上記非特許文献１記載の多対多の文字列対応付け方法において、文字列の生成モデルとして「翻字モデル」「非翻字モデル」の２つを同時に学習することで、文字列組の翻字となっている部分を識別し、非翻字部分が混在したデータにおける適切な文字列間の対応付けを実現する。
【００１６】
本発明の実施の形態では、上記非特許文献３や上記非特許文献４と同様の「翻字である」文字列の確率モデル（以下、「翻字モデル」と称する）と、「翻字でない」文字列の確率モデル（以下、「非翻字モデル」と称する）の２種類のモデルの存在を仮定する。翻字モデルは原言語と目的言語の部分文字列の同時確率Ｐ（¯ｓ，¯ｔ）（¯ｓは原言語の部分文字列、¯ｔは目的言語の部分文字列）のモデルであり、非翻字モデルは「原言語の部分文字列の確率Ｐ（¯ｓ）のモデル」と「目的言語の部分文字列の確率Ｐ（¯ｔ）のモデル」との独立した２つの確率モデルを含んで構成される。
【００１７】
本発明の実施の形態では、上記非特許文献４や、上記非特許文献６のように、学習に用いられる原言語と目的言語の文字列組が、文字列全体として翻字であるか、翻字でないかを区別するのではなく、文字列のどの部分が翻字であり、どの部分が翻字でないかを区別するように確率モデルを学習する。また、翻字である部分は高々１箇所であると仮定する。つまり、原言語と目的言語の文字列の各々は、
【００１８】
・「０文字以上の翻字でない部分」（以下、前置非翻字セグメントと称する。）
・「０文字以上の翻字である部分」（以下、翻字セグメントと称する。）
・「０文字以上の翻字でない部分」（以下、後置非翻字セグメントと称する。）
【００１９】
の順で構成されると仮定する。すなわち、文字列組の各文字列を、文字列の先頭から順番に、他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す前置非翻字セグメントと、他方の言語の部分文字列と翻字関係にある０文字以上の部分文字列を示す翻字セグメントと、他方の言語の部分文字列と翻字関係にない０文字以上の部分文字列を示す後置非翻字セグメントとで構成した場合を想定する。
なお、文字列全体が翻字でない場合は、すべてが前置非翻字セグメントに属するものとする。本実施の形態における文字列の対応付けは、原言語と目的言語の文字列組を構成する文字列を上記セグメントに分割し、非翻字セグメントにおいては原言語と目的言語で独立な非翻字モデルに基づいて部分文字列が生成され、翻字セグメントにおいては原言語と目的言語との翻字モデルに基づいて部分文字列の組が生成される場合の尤度が最大となる

ような対応付けを求める過程である。翻字モデルや非翻字モデル自体の構成及びモデル最適化アルゴリズムは特に規定しないが、翻字モデルは、原言語の０文字以上の文字列と目的言語の０文字以上の文字列の組の同時生成確率（ただし双方とも０文字となる場合は除く）モデルであり、非翻字モデルは原言語と目的言語で独立な、１文字以上の文字列の生成確率モデルであるとする。

【００２０】

また、ある文字列が翻字モデルと非翻字モデルとのどちらから生成されるかを表す確率変数（以下、「翻字モデル選択確率」、及び「非翻字モデル選択確率」と称する）も同時に考慮する。従って、上記の前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントで文字列が構成されることを前提とすると、原言語の非翻字モデル選択確率は、原言語の部分文字列が、目的言語の部分文字列と翻字関係にない非翻字部分である確率を表しており、目的言語の非翻字モデル選択確率は、目的言語の部分文字列が、原言語の部分文字列と翻字関係にない非翻字部分である確率を表している。

また、翻字モデル選択確率は、原言語の部分文字列が、目的言語の部分文字列と翻字関係にある翻字部分であり、かつ目的言語の部分文字列が、原言語の部分文字列と翻字関係にある翻字部分である確率を表している。

【００２１】

モデルの最適化においては、上記非特許文献４で用いられている前向き後向き（forward-backward）アルゴリズムに基づく期待値最大化（ＥＭ）アルゴリズムや、その変種といえる、上記特許文献１で用いられている、ギブス（Ｇｉｂｂｓ）サンプリングに基づく前向きフィルタリング・後ろ向きサンプリングアルゴリズムにおいて、 上記の３セグメントでの構成を考慮した上で、非翻字セグメントでは非翻字モデル生成確率、翻字セグメントでは翻字モデル生成確率を利用して期待値を計算し、その結果を利用してモデルの更新を行う過程を繰り返し行えばよい。

ここで、原言語の部分文字列の非翻字モデル生成確率は、原言語の前置非翻字セグメント又は後置非翻字セグメントにおける部分文字列の、原言語における生成確率を表し、目的言語の部分文字列の非翻字モデル生成確率は、目的言語の前置非翻字セグメント又は後置非翻字セグメントにおける部分文字列の、目的言語における生成確率を表す。

また、翻字モデル生成確率は、原言語の文字列のうちの翻字セグメントの部分文字列と、目的言語の文字列のうちの翻字セグメントの部分文字列との間の部分文字列の各ペアに対する同時生成確率を表している。

【００２２】

〔第１の実施の形態〕

＜システム構成＞

本発明の第１の実施の形態に係る文字列対応付け装置１００は、原言語（第１の言語）の文字列（単語）と目的言語（第２の言語）の文字列(単語)との対訳である複数組の文字列組を入力とし、文字列組の各々について、文字列の対応付けを行う。この文字列対応付け装置１００は、ＣＰＵと、ＲＡＭと、後述する文字列対応付け処理ルーチンを実行するためのプログラムを記憶したＲＯＭとを備えたコンピュータで構成され、機能的には次に示すように構成されている。図１に示すように、文字列対応付け装置１００は、入力部１０と、演算部２０と、出力部３０とを備えている。

【００２３】

入力部１０は、文字列の対応付けを行う対象である複数組の文字列組を受け付ける。具体的には、入力部１０は、翻字又は対訳になっていることが期待され、かつ空白文字等も含んだ文字列組を、入力装置、記憶媒体もしくはネットワークを通じて複数組読み込む。

【００２４】

演算部２０は、文字列組データベース２１、文字列組データ読み込み部２２、及び対応付け計算部２３を備えている。

【００２５】

文字列組データベース２１には、入力部１０により受け付けた複数組の文字列組が格納

される。
【００２６】
文字列組データ読み込み部２２は、文字列組データベース２１から全ての文字列組を読み込む。
【００２７】
対応付け計算部２３は、文字列組毎に、原言語の文字列と、目的言語の文字列との間で部分文字列同士の対応付けを行う。
【００２８】
対応付け計算部２３は、初期値設定部２３１、期待値計算部２３２、パラメータ更新部２３３、停止判定部２３４、及び文字対応付け処理部２３５を備えている。
【００２９】
初期値設定部２３１は、原言語の部分文字列に対する非翻字モデル選択確率と、目的言語の部分文字列に対する非翻字モデル選択確率と、原言語の部分文字列と目的言語の部分文字列とのペアに対する翻字モデル選択確率と、原言語の文字列のうちの各部分文字列に対する非翻字モデル生成確率と、目的言語の文字列のうちの各部分文字列に対する非翻字モデル生成確率と、原言語の文字列のうちの部分文字列と、目的言語の文字列のうちの部分文字列との間の部分文字列の各ペアに対する翻字モデル生成確率と、に対して初期値を設定する。ここで、初期値は、出現頻度や共起頻度の比を用いて設定することが広く行われているが、一様分布としてもよい。
なお、対応付けの単位として上記非特許文献４のように１文字単位のものしか考慮しない場合は、非翻字モデル生成確率は単純な文字ユニグラム確率となるため、平滑化を考慮しなければ単純な文字の出現頻度に基づく出現確率分布に従って求められ、以後の処理において確率を更新せず、固定するようにしてもよい。
また、翻字モデル選択確率・非翻字モデル選択確率については、上記非特許文献４ではそれぞれ０．５を初期値としているが、０より大きく、和が１となるような任意の初期値を設定してもよい。
【００３０】
期待値計算部２３２は、初期値設定部２３１によって設定され、又は前回更新された、非翻字モデル選択確率、翻字モデル選択確率、非翻字モデル生成確率、及び翻字モデル生成確率に基づいて、文字列組の各々に対して、原言語の文字列のうちの部分文字列と、目的言語の文字列のうちの部分文字列との間の部分文字列の各ペアについて、当該ペアが翻訳関係にある期待値を計算し、原言語の文字列のうちの各部分文字列について、当該部分文字列が非翻字部分である期待値を計算し、目的言語の文字列のうちの各部分文字列について、当該部分文字列が非翻字部分である期待値を計算する。期待値の計算方法は上記非特許文献４に記載の方法に類似するが、本実施の形態では、文字列に、前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントの順で、各セグメントが割り当てられて構成されるため、翻字セグメントのみが存在することを仮定している上記非特許文献４の方法を拡張した方法を用いる。
【００３１】
上記非特許文献４に記載の方法で用いられている前向き後ろ向きアルゴリズムでは、動的計画法を用いて、文字列組の先頭から対応付けの確率を記録する前向き確率表と、逆に文字列組の末尾から対応付けの確率を記録する後向き確率表とに、計算した確率を順次記録するが、本実施の形態では、前置非翻字セグメント、翻字セグメント、後置非翻字セグメントのそれぞれに対応する前向き確率表と後向き確率表の、合計６個の確率表を利用する。前向き確率及び後向き確率の計算と記録とに際しては、前置非翻字セグメント、翻字セグメント、後置非翻字セグメントの順を考慮する。前置非翻字セグメントの前向き確率表に記録される確率の計算にあたっては、そこより前の位置における前置非翻字セグメントの前向き確率のみを参照し、翻字セグメントの前向き確率表に記録される確率の計算にあたっては、そこより前の位置における前置非翻字セグメントの前向き確率及び翻字セグメントの前向き確率を参照し、後置非翻字セグメントの前向き確率表に記録される確率の

計算にあたっては、そこより前の位置における翻字セグメントの前向き確率及び後置非翻字セグメントの前向き確率を参照する。

【００３２】

後向き確率表についてはこれと逆の順序となるため、後置非翻字セグメントの後向き確率表に記録される確率の計算にあたっては、 そこより後の位置における後置非翻字セグメントの後向き確率のみを参照し、翻字セグメントの後向き確率表に記録される確率の計算にあたっては、そこより後の位置における後置非翻字セグメントの後向き確率及び翻字セグメントの後向き確率を参照し、前置非翻字セグメントの後向き確率表に記録される確率の計算にあたっては、そこより後の位置における翻字セグメントの後向き確率及び前置非翻字セグメントの後向き確率を参照する。また、確率表の初期値については、前置非翻字セグメントの前向き確率表の先頭と、後置非翻字セグメントの後向き確率表の末尾を１とし、それ以外を０とする。

【００３３】

期待値計算部２３２は、初期値設定部２３１によって設定され、又は前回更新された、非翻字モデル選択確率、翻字モデル選択確率、非翻字モデル生成確率、及び翻字モデル生成確率に基づいて、上記のように、前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントのそれぞれに対応する前向き確率表の各前向き確率を計算して記録すると共に、前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントのそれぞれに対応する後ろ向き確率表の各後ろ向き確率を計算して記録する。なお、前向き及び後ろ向き確率表は、文字列組ごとに計算される。

期待値計算部２３２は、これらの前向き確率表と後向き確率表を用いて、文字列組中の任意の文字対応付けについて翻訳関係にある期待値を計算する。原言語側の文字位置ｊからｊ’－１までの文字列と目的言語側の文字位置ｉからｉ’－１までの文字列が翻字となる期待値Ｅｔ（ｊ→ｊ’，ｉ→ｉ’）は、以下のように計算できる（なお、文字列位置は０から始まるものとする）。

【００３４】

【数１】

$$E_t\left(s_{j\to j'},t_{i\to i'}\right)=\frac{\left(P_{f1}(j,i)+P_{f2}(j,i)\right)\times p_t\times P_t\left(s_{j\to j'},t_{i\to i'}\right)\times\left(P_{b2}(j',i')+P_{b3}(j',i')\right)}{P_{f1}(J,I)+P_{f2}(J,I)+P_{f3}(J,I)} \quad \cdots(1)$$

【００３５】

ここで、$P_t(s_{j\to j'},t_{i\to i'})$は原言語側の文字位置ｊからｊ’－１までの文字列と目的言語側の文字位置ｉからｉ’－１までの文字列の同時生成確率（翻字モデルの確率）、p_tは翻字モデル選択確率、P_{f1}（ｊ，ｉ）、P_{f2}（ｊ，ｉ）、P_{f3}（ｊ，ｉ）はそれぞれ原言語側の文字位置ｊ、目的言語側の文字位置ｉまでの前置非翻字セグメント、翻字セグメント、後置非翻字セグメントに対する前向き確率、P_{b1}（ｊ，ｉ）、P_{b2}（ｊ，ｉ）、P_{b3}（ｊ，ｉ）はそれぞれ原言語側の文字位置ｊ、目的言語側の文字位置ｉより後の前置非翻字セグメント、翻字セグメント、後置非翻字セグメントに対する後向き確率、Ｊは原言語側の文字列長、Ｉは目的言語側の文字列長である。

期待値計算部２３２は、全ての文字列組ｄ毎に、文字列組ｄの部分文字列の各ペアについて、上記（１）式の期待値を計算する。

【００３６】

期待値計算部２３２は、前向き確率表と後向き確率表を用いて、非翻字部分の期待値も

同様に計算する。具体的には、原言語側の文字位置ｊからｊ’－１までの部分文字列が、目的言語側の文字位置ｉの後で目的言語側から（翻字ではなく）独立に生成される期待値$E_{src}(s_{j\to j'},i)$を、以下の式のように前置非翻字セグメントで現れる場合と後置非翻字セグメントで現れる場合の期待値の和として計算する。期待値計算部２３２は、前向き確率表と後向き確率表を用いて、目的言語側についても同様に、目的言語側の文字位置ｉからｉ’－１までの部分文字列が、原言語側の文字位置ｊの後で原言語側から（翻字ではなく）独立に生成される期待値$E_{trg}(s_{i\to i'},j)$を計算する。

【００３７】

【数２】

$$E_{src}(s_{j\to j'},i)=\frac{P_{f1}(j,i)\times p_{src}\times P_{src}(s_{j\to j'})\times(P_{b1}(j',i')+P_{b2}(j',i'))}{P_{f1}(J,I)+P_{f2}(J,I)+P_{f3}(J,I)}+\frac{P_{f3}(j,i)\times p_{src}\times P_{src}(s_{j\to j'})\times P_{b3}(j',i')}{P_{f1}(J,I)+P_{f2}(J,I)+P_{f3}(J,I)}\quad\cdots(2)$$

【００３８】

ここで、$P_{src}(s_{j\to j'})$は原言語側の文字位置ｊからｊ’－１までの部分文字列の生成確率（原言語側の非翻字モデルの確率）、p_{src}は原言語側の非翻字モデル選択確率である。

【００３９】

期待値計算部２３２は、全ての文字列組ｄ毎に、文字列組ｄの原言語側の部分文字列の各々について、上記（２）式の期待値を計算する。また、期待値計算部２３２は、全ての文字列組ｄ毎に、文字列組ｄの目的言語側の部分文字列の各々について、上記（２）式と同様の期待値を計算する。

期待値計算部２３２は、全ての文字列組ｄに対する計算結果に基づいて、原言語及び目的言語の部分文字列のペア（ｓ、ｔ）の各々について、当該ペア（ｓ、ｔ）に関する期待値の計算結果を集計して当該ペア（ｓ、ｔ）に対する翻字モデルからの同時生成の期待値$E_t(s,t)$を計算する。期待値計算部２３２は、原言語の各部分文字列ｓについて、当該部分文字列ｓに関する期待値の計算結果を集計して当該部分文字列ｓに対する非翻字モデルからの生成の期待値$E_{src}(s)$を計算する。また、期待値計算部２３２は、目的言語の各部分文字列ｔについて、当該部分文字列ｔに関する期待値の計算結果を集計して当該部分文字列ｔの非翻字モデルからの生成の期待値$E_{trg}(t)$を計算する。

また、上記の確率表および期待値の計算において、前置非翻字セグメント、後置非翻字セグメント内では、必ず原言語側が先に生成されるものとし、同じ文字列の生成に際して複数回の数え上げが起こらないようにする。

【００４０】

パラメータ更新部２３３は、文字列組の各々に対して、期待値計算部２３２によって計算された各ペアに対する翻訳関係にある期待値、原言語の各部分文字列についての非翻字部分である期待値、及び目的言語の各部分文字列についての非翻字部分である期待値に基づいて、非翻字モデル選択確率、翻字モデル選択確率、非翻字モデル生成確率、及び翻字モデル生成確率を更新する。具体的には、パラメータ更新部２３３は、期待値計算部２３２において計算された期待値をもとに、翻字モデルの同時生成確率、非翻字モデルの生成確率、及び翻字モデル選択確率・非翻字モデル選択確率を更新する。翻字モデルの同時生成確率及び非翻字モデルの生成確率については、期待値が０より大きい文字列組もしくは文字列に対して、期待値の総和との比をとって更新後の同時生成確率又は生成確率とする

。つまり、翻字モデルの同時生成確率については、原言語の部分文字列ｓと目的言語の部分文字列ｔの同時生成確率Ｐｔ（￣ｓ，￣ｔ）を以下のように更新する。

【００４１】

【数３】

$$P_t(\bar{s},\bar{t}) \leftarrow \frac{E_t(\bar{s},\bar{t})}{\sum_{s,t} E_t(s,t)} \qquad \cdots(3)$$

【００４２】

非翻字モデルの生成確率については、原言語側の部分文字列ｓの生成確率Ｐｓｒｃ（￣ｓ）を以下のように更新する。目的言語側の部分文字列ｔの生成確率Ｐｔｒｇ（￣ｔ）も同様に更新する。

【００４３】

【数４】

$$P_{src}(\bar{s}) \leftarrow \frac{E_{src}(\bar{s})}{\sum_{s} E_{src}(s)} \qquad \cdots(4)$$

【００４４】

翻字モデル選択確率・非翻字モデル選択確率については、原言語と目的言語でそれぞれ翻字セグメントに属する文字数・非翻字セグメントに属する文字数の割合に基づくと考え、原言語の非翻字モデル選択確率は非翻字セグメントで現れる原言語の文字数の割合、目的言語の非翻字モデル選択確率は非翻字セグメントで現れる目的言語の文字数の割合とする。つまり、部分文字列の期待値を利用して計算すると、原言語の非翻字モデル選択確率ｐｓｒｃは以下のように更新される。｜ｓ｜は部分文字列ｓの長さ（文字数）を表す。目的言語の非翻字モデル選択確率ｐｔｒｇについても同様に更新される。

【００４５】

【数５】

$$p_{src} \leftarrow \frac{\sum_s |s| + E_{src}(s)}{\sum_s |s| \times E_{src}(s) + \sum_{s,t} |s| \times E_t(s,t)} \qquad \cdots(5)$$

【００４６】

そして、翻字モデル選択確率ｐｔは、以下の式に示すように、（１－原言語の非翻字モデル選択確率）×（１－目的言語の非翻字モデル選択確率）で更新される。

【００４７】

【数６】

$$p_t \leftarrow (1 - p_{src}) \times (1 - p_{trg}) \qquad \cdots(6)$$

【００４８】
停止判定部２３４は、予め定められた停止条件が満たされたか否かを判定し、当該停止条件が満たされるまで、期待値計算部２３２による計算、及びパラメータ更新部２３３による更新を繰り返す。停止条件としては、文字列組データの尤度が一定以上の数値になった、尤度の変動幅が一定以下の数値になった、パラメータ更新の繰り返し回数が一定の回数を超えた、などが考えられる。本実施の形態では、文字列組データの尤度が一定以上の数値になることを、停止条件とする。
【００４９】
文字対応付け処理部２３５は、文字列組の各々に対して、パラメータ更新部２３３により最終的に更新された非翻字モデル選択確率、翻字モデル選択確率、非翻字モデル生成確率、及び翻字モデル生成確率の各々に基づいて、当該文字列組の各文字列を、前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントで構成し、かつ、原言語の文字列のうちの翻字セグメントの部分文字列と、目的言語の文字列のうちの翻字セグメントの部分文字列との間で文字の対応付けを行う。具体的には、文字対応付け処理部２３５は、得られた最終的な非翻字モデル選択確率、翻字モデル選択確率、非翻字モデル生成確率、及び翻字モデル生成確率を用い、当該非翻字モデル選択確率、翻字モデル選択確率、非翻字モデル生成確率、及び翻字モデル生成確率の下で最尤な文字対応付けをビタビアルゴリズムによって求める。ビタビアルゴリズムは当該分野で広く知られた、動的計画法によって最尤な系列を求めるアルゴリズムである。具体的には、期待値計算部２３２における前向き確率表の確率計算とは異なり、前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントの各々の前向き確率表に対して、最も確率の高い経路の確率およびその経路の情報を保持する形で末尾まで計算を行い、その結果得られた確率表を末尾から先頭に向かって辿っていくことで最尤経路を得ることができる。経路が翻字セグメントの前向き確率表内にあるものは該当する部分が翻字の関係にあり、非翻字セグメントの前向き確率表内にあるものは翻字の関係にないことを表す。
【００５０】
出力部３０は、文字対応付け処理部２３５で対応付けられた文字列組の各々の対応付けの情報を出力する。具体的には，出力部３０は、対応付けの情報を付与した文字列組を端末に表示、もしくは記憶媒体やネットワークを通じて書き出す。
【００５１】
＜文字列対応付け装置の作用＞
次に、第１の実施の形態に係る文字列対応付け装置１００の作用について説明する。まず、対訳となっている第１の言語体系の文字列及び第２の言語体系の文字列の組である文字列組が、文字列対応付け装置１００に複数入力されると、文字列対応付け装置１００によって、入力された複数の文字列組が、文字列組データベース２１に格納される。そして、文字列対応付け装置１００によって、図２に示す文字列対応付け処理ルーチンが実行される。
【００５２】
まず、ステップＳ１００において、文字列組データ読み込み部２２によって、文字列組データベース２１から、全ての文字列組を読み込む。
【００５３】
ステップＳ１０２において、初期値設定部２３１によって、原言語及び目的言語の各々の非翻字モデル選択確率と、翻字モデル選択確率と、原言語及び目的言語の各々の各部分文字列に対する非翻字モデル生成確率と、原言語及び目的言語の部分文字列の各ペアに対する翻字モデル生成確率と、に対して初期値を設定する。
【００５４】
次に、ステップＳ１０４において、期待値計算部２３２によって、全ての文字列組の各々について、周知の前向き後ろ向きアルゴリズムを用いて、上記ステップＳ１０２で設定され、又は後述するステップＳ１１０で前回更新された、非翻字モデル選択確率、翻字モ

デル選択確率、非翻字モデル生成確率、及び翻字モデル生成確率に基づいて、前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントのそれぞれに対応する前向き確率表の各前向き確率を計算して前向き確率表に記録すると共に、前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントのそれぞれに対応する後ろ向き確率表の各後ろ向き確率を計算して後ろ向き確率表に記録する。

【００５５】

ステップＳ１０６において、期待値計算部２３２によって、全ての文字列組の各々について、上記ステップＳ１０４で計算された当該文字列組の前向き確率表と後向き確率表とを用いて、当該文字列組の部分文字列の各ペアに対し、翻字関係にある期待値を計算する。

【００５６】

ステップＳ１０８において、期待値計算部２３２によって、全ての文字列組の各々について、上記ステップＳ１０４で計算された当該文字列組の前向き確率表と後向き確率表を用いて、当該文字列組の原言語側の各部分文字列に対し、非翻字部分となる期待値を計算すると共に、当該文字列組の目的言語側の各部分文字列に対し、非翻字部分となる期待値を計算する。

【００５７】

ステップＳ１１０において、パラメータ更新部２３３は、上記ステップＳ１０６で計算された翻訳関係にある期待値、上記ステップＳ１０８で計算された原言語の各部分文字列についての非翻字部分である期待値、及び目的言語の各部分文字列についての非翻字部分である期待値に基づいて、原言語及び目的言語の各々の非翻字モデル選択確率、翻字モデル選択確率、原言語の各部分文字列に対する非翻字モデル生成確率、目的言語の各部分文字列に対する非翻字モデル生成確率、及び原言語及び目的言語の部分文字列の各ペアに対する翻字モデル生成確率を更新する。

【００５８】

ステップＳ１１４において、停止判定部２３４によって、予め定められた停止条件が満たされたか否かを判定する。そして、停止条件が満たされていない場合には、上記ステップＳ１０４へ戻り、上記ステップＳ１０４～ステップＳ１１０の処理を実行する。一方、停止条件が満たされている場合には、ステップＳ１１６へ進む。

【００５９】

ステップＳ１１６において、文字対応付け処理部２３５によって、文字列組の各々に対して、上記ステップＳ１１０で最終的に更新された非翻字モデル選択確率、翻字モデル選択確率、非翻字モデル生成確率、及び翻字モデル生成確率の各々に基づいて、当該文字列組の各文字列を、前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントで構成し、かつ、原言語の文字列のうちの翻字セグメントの部分文字列と、目的言語の文字列のうちの翻字セグメントの部分文字列との間で文字の対応付けを行う。

【００６０】

ステップＳ１１８において、出力部３０によって、上記ステップＳ１１６で対応付けた結果を出力して、文字列対応付け処理ルーチンを終了する。

【００６１】

以上説明したように、本発明の第１の実施の形態に係る文字列対応付け装置によれば、原言語及び目的言語において対訳となる文字列の組み合わせである文字列組について、文字列組の各文字列を、文字列の先頭から順番に、前置非翻字セグメントと、翻字セグメントと、後置非翻字セグメントとで構成したときに、翻字モデル選択確率と、非翻字モデル選択確率と、非翻字モデル生成確率と、翻字モデル生成確率と、に基づいて尤もらしくなるように、文字列組の各文字列を前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントで構成し、かつ、原言語の文字列のうちの翻字セグメントの部分文字列と、原言語の文字列のうちの翻字セグメントの部分文字列との間で文字の対応付けを行うことにより、異なる言語の文字列組における文字の対応付けを精度よく行うことができる。

【００６２】

また、文字列の生成モデルとして「翻字モデル」「非翻字モデル」の２つを同時に学習することで、文字列の組の翻字となっている部分を識別し、非翻字部分が混在したデータにおける適切な文字列間の対応付けを実現することができる。

【００６３】

また、統計的機械翻訳で用いられる自動的に抽出された対訳語句対のような、必ずしも翻字の組となっておらず、部分的に翻字となっているような文字列組データに対しても、翻字となっている文字列を選択的に対応づけることが可能となる。

【００６４】

〔第２の実施の形態〕

＜システム構成＞

次に、第２の実施の形態について説明する。なお、第１の実施の形態と同様の構成となる部分については、同一符号を付して説明を省略する。

【００６５】

第２の実施の形態では、ギブスサンプリングを用いて、文字列組における文字の対応付けを求めている点が、第１の実施の形態と異なっている。

【００６６】

本実施の形態で用いるギブスサンプリングは、上記非特許文献３や上記非特許文献６で用いられており、当該分野では広く知られる方法である。

上記非特許文献３及び非特許文献６に記載の方法のように、参考文献７（Daichi Mochihashi他２名、「Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling」、Proc. of ACL-IJCNLP、2009）に記載の方法に類似したアルゴリズムによって対応付けを繰り返し更新する。

【００６７】

この方法では、文字列組の集合Ｄから１つの文字列組ｄを選び、ｄを除く文字列組の集合Ｄ－ｄにおける対応付けから推定される事後確率分布に基づいてｄ上の対応付け結果をサンプリングする、という過程を、ｄを入れ替えながら繰り返し行う。サンプリングの結果はｄに対する一意の対応付け結果であるため、ＥＭアルゴリズムを利用する第１の実施の形態の構成とは異なり、停止判定部によって学習が終了したと判定された時点ですべてのｄに対する対応付け結果が得られるため、文字対応付け処理部２３５は必要ない。ただし、停止判定後に１度フィルタリングステップ・サンプリングステップを繰り返すことによって対応付けを再度更新してもよい。

【００６８】

図３に示すように、第２の実施の形態に係る文字列対応付け装置２００の演算部２２０は、文字列組データベース２１、文字列組データ読み込み部２２、及び対応付け計算部２４を備えている。

【００６９】

対応付け計算部２４は、ギブスサンプリングを用いて、複数の文字列組の各々に対して対応付けを行う。対応付け計算部２４は、初期値対応部２４１、フィルタリング部２４２、サンプリング部２４３、及び停止判定部２４４を備えている。

【００７０】

初期値対応部２４１は、文字列組の各々に対して、文字列組の各文字列を前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントで構成し、かつ、原言語の文字列のうちの翻字セグメントの部分文字列と、目的言語の文字列のうちの翻字セグメントの部分文字列との間で文字の対応付けを行って、対応付けの初期設定を行う。ここで、ギブスサンプリングを用いる場合は、確率分布の初期値でなく、適当な初期対応付けを与える必要がある。従って、初期対応付の決定方法としては、何かの規則により対応付けを行う、ランダムに対応付けを行う、別途簡便なモデルでの対応付け結果を利用するなどの方法が考えられる。本実施の形態では、ランダムに対応付けを行う。

【００７１】

フィルタリング部２４２は、フィルタリング部２４２によって前回計算された、原言語

の非翻字モデル選択確率と、目的言語の非翻字モデル選択確率と、翻字モデル選択確率と、処理対象の文字列組の原言語の文字列のうちの各部分文字列に対する非翻字モデル生成確率と、処理対象の文字列組の目的言語の文字列のうちの各部分文字列に対する非翻字モデル生成確率と、処理対象の文字列組の原言語の文字列のうちの部分文字列と、目的言語の文字列のうちの部分文字列との間の部分文字列の各ペアに対する翻字モデル生成確率とに基づいて、上記第１の実施の形態における前向き確率の計算の場合と同様の方法で、処理対象の文字列組ｄに対して、文字列組の先頭から前向き確率を計算し、前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントのそれぞれに対応する前向き確率表に、計算した各前向き確率を記録する。ここで、確率が非常に小さい経路を大量に保持すると計算量が大きくなる場合があるため、他の対立する対応付け経路と比較して非常に確率が小さいような場合はその経路を無視してもよい。

また、フィルタリング部２４２は、初期対応設定部２４１によって設定され、又は後述するサンプリング部２４３によって前回更新された、複数組の文字列組のうちの処理対象の文字列組以外の文字列組の各々についての対応付けに基づいて、原言語の非翻字モデル選択確率と、目的言語の非翻字モデル選択確率と、翻字モデル選択確率と、処理対象の文字列組の原言語の文字列のうちの各部分文字列に対する非翻字モデル生成確率と、処理対象の文字列組の目的言語の文字列のうちの各部分文字列に対する非翻字モデル生成確率と、処理対象の文字列組の原言語の文字列のうちの部分文字列と、目的言語の文字列のうちの部分文字列との間の部分文字列の各ペアに対する翻字モデル生成確率と、を計算する。翻字モデル生成確率P_{gt}（ＥＭアルゴリズムによる構成の場合のP_tに相当）は、上記非特許文献３や上記非特許文献６に記載の方法を拡張し、以下のように定義できる。

【００７２】

【数７】

$$P_{gt}\left(s_{j\to j'},t_{i\to i'}\right)=p_t\times\frac{c\left(s_{j\to j'},t_{i\to i'}\right)+\alpha BM\left(s_{j\to j'},t_{i\to i'}\right)}{C+\alpha}\qquad\cdots(7)$$

【００７３】

ここで、ｃ（$s_{j\to j'}$，$t_{i\to i'}$）は，原言語側の文字列$s_{j\to j'}$と目的言語側の文字列$t_{i\to i'}$がｄを除く文字組列の集合で対応付けられている回数、ＢＭ（$s_{j\to j'}$，$t_{i\to i'}$）は基底測度、αはハイパーパラメータ、Ｃはｄを除く文字列組データでの翻字となっているすべての対応付けの数である。非翻字モデルについても同様であり、原言語・目的言語の非翻字モデル生成確率も、同様に以下のように定義できる（ＥＭアルゴリズムによる構成の場合のP_{src}、P_{trg}に相当）。

【００７４】

【数８】

$$P_{gsrc}\left(s_{j\to j'}\right)=p_{src}\times\frac{c_{src}\left(s_{j\to j'}\right)+\alpha BM\left(s_{j\to j'}\right)}{C_{src}+\alpha}\qquad\cdots(8)$$

$$P_{gtrg}\left(t_{i\to i'}\right)=p_{trg}\times\frac{c_{trg}\left(s_{i\to i'}\right)+\alpha BM\left(t_{i\to i'}\right)}{C_{trg}+\alpha}\qquad\cdots(9)$$

【００７５】

翻字モデル選択確率p_t、非翻字モデル選択確率p_{src}、p_{trg}は前記ＥＭアルゴ

リズムの場合と同様、翻字になっている文字数と非翻字になっている文字数の割合で与えることができる。

【００７６】

サンプリング部２４３は、フィルタリング部２４２によって計算された非翻字モデル選択確率、翻字モデル選択確率、非翻字モデル生成確率、及び翻字モデル生成確率に基づいて、処理対象の文字列組の各文字列を前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントで構成し、かつ、原言語の文字列のうちの翻字セグメントの部分文字列と、目的言語の文字列のうちの翻字セグメントの部分文字列との間で文字の対応付けを行って、処理対象の文字列組に対する対応付けを更新する。具体的には、サンプリング部２４３は、フィルタリング部２４２で計算された前向き確率表を末尾から順に辿っていき、前向き確率に基づいて対応付けを決定していき、ｄに対する対応付けを更新する。より具体的には、文字列組データ中のある位置（ｊ，ｉ）に対して、その位置に至る１つ前の位置（ｊ＿，ｉ＿）を選択する際に、（ｊ＿，ｉ＿）から（ｊ，ｉ）に至る確率に基づく重み付きサンプリングを行う（確率が高い経路ほど選ばれやすいようにする）。（ｊ＿，ｉ＿）から（ｊ，ｉ）に至る確率は、（ｊ＿，ｉ＿）における前向き確率と（ｊ＿，ｉ＿）から（ｊ，ｉ）への経路の確率（$s_{j-\rightarrow j'}$，$t_{i-\rightarrow i'}$の翻字あるいは非翻字確率）である。こうして決定された対応付けをｄの新しい対応付けとする。

【００７７】

全ての文字列組の各々を、処理対象の文字列組として、フィルタリング部２４２及びサンプリング部２４３による処理を繰り返し行う。

停止判定部２４４は、予め定められた停止条件が満たされたか否かを判定し、停止条件が満たされるまで、各文字列組を処理対象とした、フィルタリング部２４２による計算及びサンプリング部２４３による更新を繰り返す。本実施の形態では、ギブスサンプリングを利用しており、文字列組ｄを一つ選択するごとにフィルタリング部２４２によるフィルタリング処理と、サンプリング部２４３によるサンプリング処理とを行うので、文字列組の集合Ｄの全てに対してフィルタリング処理とサンプリング処理とを適用した後に停止判定を行う。ただし、途中で停止してもよい。停止条件はＥＭアルゴリズムを利用する第１の実施の形態の場合と同様に設定が可能である。

【００７８】

＜文字列対応付け装置の作用＞

次に、第２の実施の形態に係る文字列対応付け装置２００の作用について説明する。なお、第１の実施の形態と同様の処理については、同一符号を付して説明を省略する。

【００７９】

まず、対訳となっている原言語の文字列及び目的言語の文字列の組である文字列組が、文字列対応付け装置２００に複数入力されると、文字列対応付け装置２００によって、入力された複数の文字列組が、文字列組データベース２１に格納される。そして、文字列対応付け装置２００によって、図４に示す文字列対応付け処理ルーチンが実行される。

【００８０】

ステップＳ１００において、文字列組データベース２１から、全ての文字列組を読み込む。

【００８１】

そして、ステップＳ２０２において、初期値対応部２４１によって、文字列組の各々に対して、文字列組の各文字列を前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントで構成し、かつ、原言語の文字列のうちの翻字セグメントの部分文字列と、目的言語の文字列のうちの翻字セグメントの部分文字列との間で文字の対応付けを行って、対応付けの初期設定を行う。

【００８２】

次にステップＳ２０４において、フィルタリング部２４２によって、処理対象として、１つの文字組のデータｄを設定する。

【００８３】

ステップＳ２０６において、フィルタリング部２４２によって、上記ステップＳ２０４で設定された処理対象の文字列組のデータｄについて、本ステップで前回計算された、原言語の非翻字モデル選択確率と、目的言語の非翻字モデル選択確率と、翻字モデル選択確率と、処理対象の文字列組の原言語の文字列のうちの各部分文字列に対する非翻字モデル生成確率と、処理対象の文字列組の目的言語の文字列のうちの各部分文字列に対する非翻字モデル生成確率と、処理対象の文字列組の原言語の文字列のうちの部分文字列と、目的言語の文字列のうちの部分文字列との間の部分文字列の各ペアに対する翻字モデル生成確率とに基づいて、上記第１の実施の形態における前向き確率の計算の場合と同様の方法で、処理対象の文字列組ｄに対して、文字列組の先頭から前向き確率を計算し、前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントのそれぞれに対応する前向き確率表に、計算した各前向き確率を記録する。

【００８４】

ステップＳ２０８において、フィルタリング部２４２によって、上記ステップＳ２０２で設定され、又は後述するステップＳ２１０で前回更新された、複数組の文字列組のうちの処理対象の文字列組以外の文字列組の各々についての対応付けに基づいて、原言語及び目的言語の各々の非翻字モデル選択確率と、翻字モデル選択確率と、処理対象の文字列組の原言語及び目的言語の各々の、文字列のうちの各部分文字列に対する非翻字モデル生成確率と、処理対象の文字列組の目的言語の文字列のうちの各部分文字列に対する非翻字モデル生成確率と、翻字モデル生成確率と、を計算する。

【００８５】

ステップＳ２１０において、サンプリング部２４３によって、上記ステップＳ２０８で計算された非翻字モデル選択確率、翻字モデル選択確率、非翻字モデル生成確率、及び翻字モデル生成確率に基づいて、処理対象の文字列組の各文字列を前置非翻字セグメント、翻字セグメント、及び後置非翻字セグメントで構成し、かつ、原言語の文字列のうちの翻字セグメントの部分文字列と、目的言語の文字列のうちの翻字セグメントの部分文字列との間で文字の対応付けを行って、処理対象の文字列組に対する対応付けを更新する。

【００８６】

ステップＳ２１２において、文字列組の集合Ｄのうちの全ての文字列組ｄについて、上記ステップＳ２０４～ステップＳ２１０の処理を実行したか否かを判定する。そして、全ての文字列組ｄについて、上記ステップＳ２０４～ステップＳ２１０の処理を実行していない場合には、上記ステップＳ２０４へ戻り、新たな文字列組ｄを処理対象として設定する。一方、全ての文字列組ｄについて、上記ステップＳ２０４～ステップＳ２１０の処理を実行した場合には、ステップＳ２１４へ進む。

【００８７】

ステップＳ２１４において、停止判定部２４４によって、予め定められた停止条件が満たされたか否かを判定する。そして、停止条件が満たされていない場合には、上記ステップＳ２０４へ戻り、再び、全ての文字列組ｄについて、上記ステップＳ２０４～ステップＳ２１２の処理を実行する。一方、停止条件が満たされている場合には、ステップＳ２１６へ進む。

【００８８】

ステップＳ２１６において、出力部３０によって、上記ステップＳ２１０で最終的に対応付けた結果を出力して、文字列対応付け処理ルーチンを終了する。

【００８９】

なお、第２の実施の形態に係る文字列対応付け装置２００の他の構成及び作用については、第１の実施の形態と同様であるため、説明を省略する。

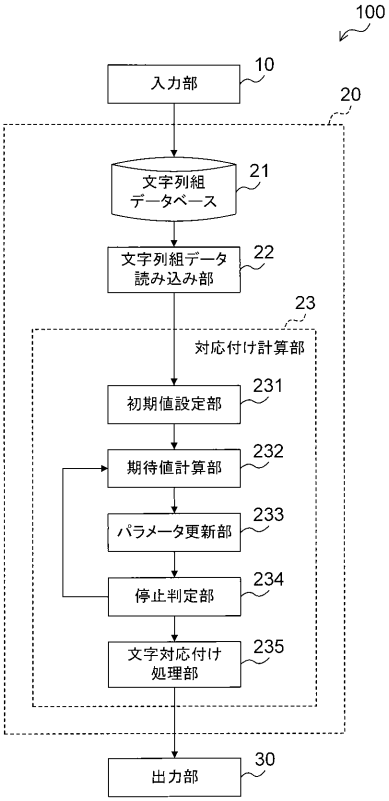
【００９０】

以上説明したように、第２の実施の形態に係る文字列対応付け装置によれば、ギブスサンプリングにより、非翻字モデル選択確率と、翻字モデル選択確率と、非翻字モデル生成確率と、翻字モデル生成確率とを計算し、処理対象の文字列組に対する対応付けを更新することにより、異なる言語の文字列組における文字の対応付けを精度よく行うことができ

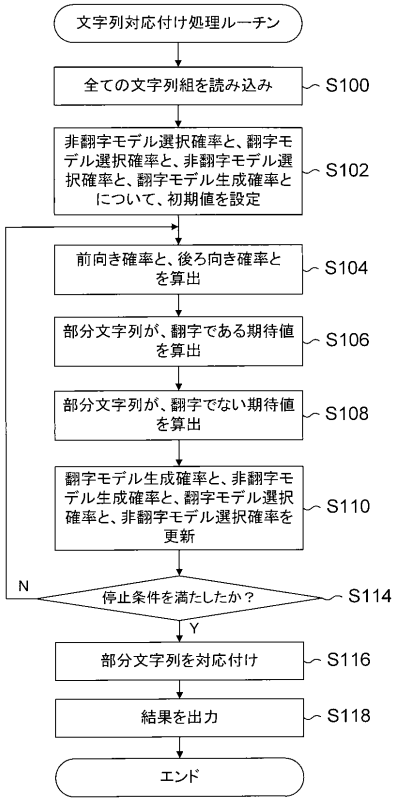
る。
【００９１】
＜実施例＞
次に本発明に係る実施の形態を実施した例について示す。実施例では原言語として日本語（カタカナ語）、目的言語として英語を利用した。なお、本実施例ではギブスサンプリングを利用した第２の実施の形態によって、対応付けの計算を行う。
【００９２】
図５はそれぞれ本実施例で用いた日本語と英語との文字列組データ（およそ５万語対）を抜粋したものである。同じ行に記された文字列組が対応していることを示す。各記号（カタカナ及びアルファベット）は１文字ずつ区切り文字（空白文字）によって分割されている。空白文字自体を対応付けの対象文字として扱いたい場合には、区切り文字として別の文字（例えば”：”）を利用するか、空白文字を別の記号（例えば”[ｓｐ]”）に置換すればよい。
【００９３】
第２の実施の形態によって、上記図５に示した文字列組データに対して対応付けを行った。まず、上記文字列組データを記録した電子ファイルを文字列組データ読み込み部２２により読み込んだ。また、カタカナとアルファベットの任意長さの文字列間の対応を考慮すると計算量が大きくなるため、本実施例においては、カタカナ（日本語）側は最大２文字、アルファベット（英語）側は最大３文字までの対応に限定することとした。
【００９４】
本実施例における初期値対応部２４１では、ランダムな対応付けを与えるために、フィルタリング部２４２によるフィルタリング処理における前向き確率表の前向き確率の値をすべて１と仮定した上で（ただし、カタカナ側３文字以上、アルファベット側４文字以上となる対応付けは許さないようにする）、サンプリング部２４３によるサンプリング処理で利用されるサンプリングアルゴリズムを適用した。これにより、ランダムな初期対応付けを得ることができる。
【００９５】
続いて、文字列組データの各文字列組に対して、フィルタリング処理とサンプリング処理の順に繰り返し計算を行う。各文字列組に対するフィルタリング処理・サンプリング処理を文字列組データのデータ数分行う処理を１ラウンドとして、本実施例では３０ラウンドの処理を行った。また、英語の記号”‐”は通常カタカナに対応するものではないため、常に日本語の長さ０の部分文字列に対応するものとして扱った。なお、ギブスサンプリングにおいてはデータを一個ずつ処理して対応付けとモデルの確率が更新されるため、データの処理順序が学習に与える影響が大きいことが広く知られている。したがって、文字列組データ中のデータの処理順序は毎ラウンド開始時にランダムに入れ替えることとした。
【００９６】
３０ラウンドの繰り返しの後、ビタビアルゴリズムによって再度文字列対応付けを行った。その結果を図６に示す。図６は日本語と英語の文字列がタブ文字で区切られて表記されており、対応付けられる部分文字列が記号”：”で区切られている。つまり、”：”の数は日本語側と英語側で同一となっており、１番目の要素同士が対応する部分文字列組であることを示す。また、”＜ｎｏｉｓｅ＞”となっているのは、対応する部分文字列がない（ノイズである）ことを示している。
【００９７】
図６の結果から、”コンピュータ”と”ｃｏｍｐｕｔｅｒｓ”の組では、英語側末尾の”ｓ”がノイズであること、”バーン”と”ｂｕｒｎ‐ｉｎ”の組では、英語側の”ｉｎ”がノイズであること、”シンメトリー”と”ａｓｙｍｍｅｔｒｙ”との組では、英語側先頭の”ａ”がノイズであること、”サイド”と”ｓｉｄｅ‐ｅｆｆｅｃｔ”の組では、英語側の”ｅｆｆｅｃｔ”がノイズであること、などが見て取れる。また、その他の文字列組についても、妥当な文字列対応付け結果が得られた。

【００９８】
なお、同じデータを上記非特許文献４の方法を実現するコンピュータプログラムによって対応付けした場合、及び上記非特許文献６の方法を実現するコンピュータプログラムによって対応付けした場合は、本実施例において部分的にノイズであると判定された文字列組は、すべてノイズであると判定されたため、本実施例は部分的にノイズが含まれるような文字列組の対応付けに適した方式であることが認められた。
【００９９】
なお、本発明は、上述した実施形態に限定されるものではなく、この発明の要旨を逸脱しない範囲内で様々な変形や応用が可能である。
【０１００】
例えば、文字列組データベース２１は、外部に設けられ、文字列対応付け装置とネットワークで接続されていてもよい。
【０１０１】
上述の文字列対応付け装置は、内部にコンピュータシステムを有しているが、「コンピュータシステム」は、ＷＷＷシステムを利用している場合であれば、ホームページ提供環境（あるいは表示環境）も含むものとする。
【０１０２】
また、本願明細書中において、プログラムが予めインストールされている実施形態として説明したが、当該プログラムを、コンピュータ読み取り可能な記録媒体に格納して提供することも可能である。
【符号の説明】
【０１０３】
１０　入力部
２０、２２０　演算部
２１　文字列組データベース
２２　文字列組データ読み込み部
２３、２４　対応付け計算部
３０　出力部
１００、２００　文字列対応付け装置
２３１　初期値設定部
２３２　期待値計算部
２３３　パラメータ更新部
２３４　停止判定部
２３５　文字対応付け処理部
２４１　初期値対応部
２４２　フィルタリング部
２４３　サンプリング部
２４４　停止判定部

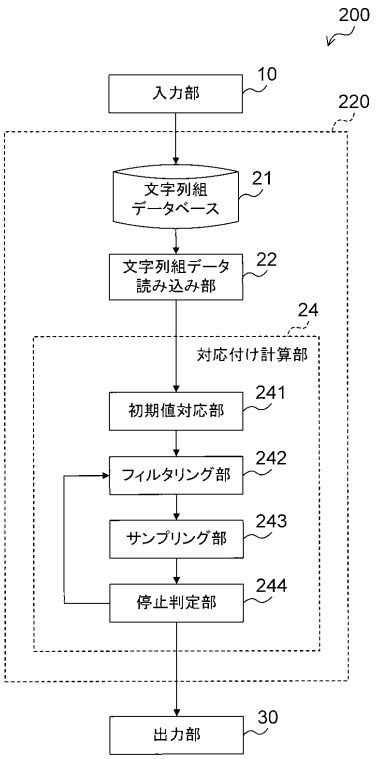
【図１】


100
入力部
10
20
文字列組
データベース
21
文字列組データ
読み込み部
22
23
対応付け計算部
初期値設定部
231
期待値計算部
232
パラメータ更新部
233
停止判定部
234
文字対応付け
処理部
235
出力部
30

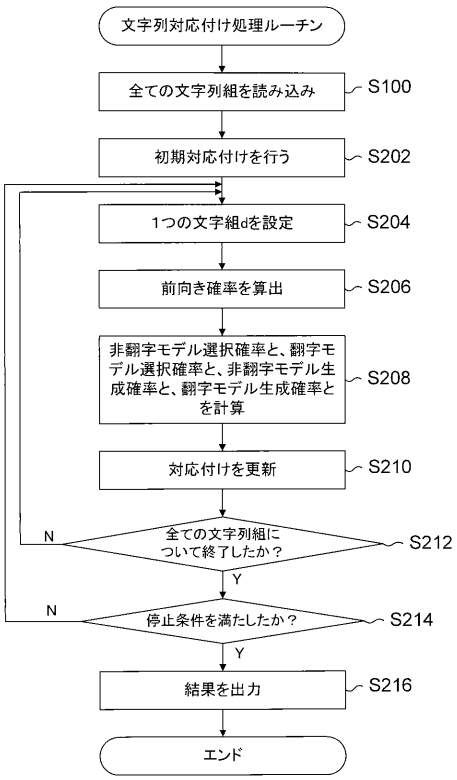
【図２】


文字列対応付け処理ルーチン
全ての文字列組を読み込み
S100
非翻字モデル選択確率と、翻字モデル選択確率と、非翻字モデル選択確率と、翻字モデル生成確率とについて、初期値を設定
S102
前向き確率と、後ろ向き確率とを算出
S104
部分文字列が、翻字である期待値を算出
S106
部分文字列が、翻字でない期待値を算出
S108
翻字モデル生成確率と、非翻字モデル生成確率と、翻字モデル選択確率と、非翻字モデル選択確率を更新
S110
停止条件を満たしたか？
S114
N
Y
部分文字列を対応付け
S116
結果を出力
S118
エンド

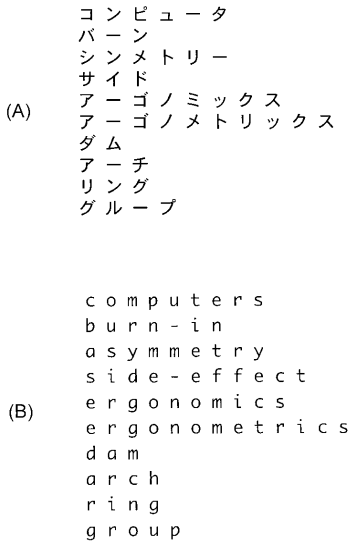
【図３】


200
入力部
10
220
文字列組
データベース
21
文字列組データ
読み込み部
22
24
対応付け計算部
初期値対応部
241
フィルタリング部
242
サンプリング部
243
停止判定部
244
出力部
30

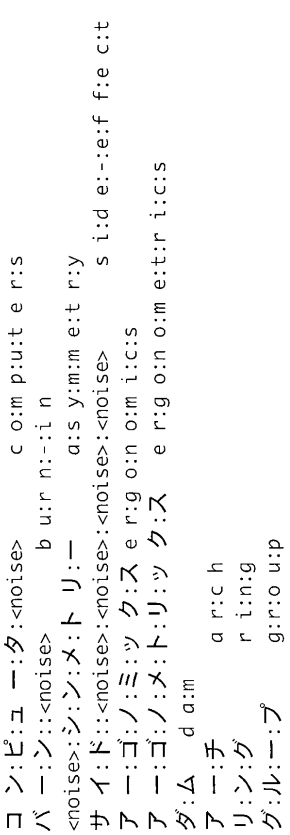
【図４】


文字列対応付け処理ルーチン
全ての文字列組を読み込み
S100
初期対応付けを行う
S202
1つの文字組dを設定
S204
前向き確率を算出
S206
非翻字モデル選択確率と、翻字モデル選択確率と、非翻字モデル生成確率と、翻字モデル生成確率とを計算
S208
対応付けを更新
S210
全ての文字列組について終了したか？
S212
N
Y
停止条件を満たしたか？
S214
N
Y
結果を出力
S216
エンド

【図５】


(A)
コンピューター
バーン
シンメトリー
サイド
アーゴノミックス
アーゴノメトリックス
ダム
アーチ
リング
グループ
(B)
computers
burn-in
asymmetry
side-effect
ergonomics
ergonometrics
dam
arch
ring
group

【図６】


コンピューター:<noise> c o:m p:u:t e r:s
バーン::<noise> b u:r n:-:i n
<noise>:シンメトリー a:s y:m:m e:t r:y
サイド::<noise>:<noise>:<noise> s i:d e:-:e:f f:e c:t
アーゴノミックス e r:g o:n o:m i:c:s
アーゴノメトリックス e r:g o:n o:m e:t:r i:c:s
ダム d a:m
アーチ a r:c h
リング r i:n:g
グループ g:r:o u:p

フロントページの続き

(72)発明者　森　信介
京都府京都市左京区吉田本町３６番地１　国立大学法人京都大学内

審査官　成瀬　博之

(56)参考文献　特開２００１－１４２８７７（ＪＰ，Ａ）
特開２００７－１５６５４５（ＪＰ，Ａ）
特開２０１２－１８５６７９（ＪＰ，Ａ）
特開２００３－２６３４３２（ＪＰ，Ａ）
特開２００５－０９２６８２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／２７－１７／２８