(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4876329号
(P4876329)
(45)発行日　平成24年2月15日(2012.2.15)　(24)登録日　平成23年12月9日(2011.12.9)

(51)Int.Cl.　F I
G06F 17/27 (2006.01)　G06F 17/27 Z

請求項の数 9　(全 12 頁)

(21)出願番号　特願2001-144337(P2001-144337)
(22)出願日　平成13年5月15日(2001.5.15)
(65)公開番号　特開2002-342325(P2002-342325A)
(43)公開日　平成14年11月29日(2002.11.29)
審査請求日　平成20年4月15日(2008.4.15)

（出願人による申告）国等の委託研究の成果に係る特許出願（平成１１年度新エネルギー・産業技術総合開発機構「複数言語にまたがる言語知識処理技術の研究（対訳語彙知識抽出技術の研究）」に関する委託研究、産業活力再生特別措置法第３０条の適用を受けるもの）

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100109313
弁理士　机　昌彦
(74)代理人　100124154
弁理士　下坂　直樹
(72)発明者　亀井　真一郎
東京都港区芝五丁目７番１号　日本電気株式会社内
(72)発明者　山端　潔
東京都港区芝五丁目７番１号　日本電気株式会社内

最終頁に続く

(54)【発明の名称】対訳確率付与装置、対訳確率付与方法並びにそのプログラム

(57)【特許請求の範囲】
【請求項１】
第１の言語を第２の言語に翻訳する際に用いられる対訳確率付与装置であって、
第１の言語の文例集と第２の言語の文例集とを有し、
第１の言語の単語に対する第２の言語の訳語候補を単語対応対として格納した第１言語第２言語対訳辞書を有し、
第１の言語の文例集における単語の出現に関する統計量を計算する第１言語統計量計算モジュールを有し、
第２の言語の文例集における単語の出現に関する統計量を計算する第２言語統計量計算モジュールを有し、
前記第２言語統計量計算モジュールによって求められた統計量と、対訳辞書の各単語対応対に付与された対訳確率をパラメータとして、前記第１言語統計量計算モジュールによって求められる統計量から、対訳確率モデル格納部に格納された式に基づいて推定された第２の言語の単語の出現に関する統計量との差を最小にするようにパラメータを調整する対訳確率付与部を有すること
を特徴とする対訳確率付与装置。
【請求項２】
請求項１記載の対訳確率付与装置において、前記対訳確率モデル格納部に格納された式として、第１の言語の単語Ｅ（ｉ）の出現確率ｅ（ｉ）とその単語Ｅ（ｉ）が第２の言語の訳語Ｊ（ｎ）に対応する対応確率Ｓ（ｉ，ｎ）との積を求め、第１の言語の各単語Ｅ（ｉ

）に関して上記の積を可算した和をとることによって第２の言語における単語Ｊ（ｎ）の出現確率ｊ（ｎ）を計算する式を用いることを特徴とする対訳確率付与装置。
【請求項３】
請求項１記載の対訳確率付与装置において、前記対訳確率モデル格納部に格納された式として、第１の言語で一つの文の中に出現する二つの単語のペアの共起確率Ｐ（Ｅ（ｉ）＾Ｅ（ｊ））とその単語ペアを構成する各単語Ｅ（ｉ）および
Ｅ（ｊ）が第２の言語の訳語に対応する対応確率Ｓ（ｉ，ｍ）およびＳ（ｊ，ｎ）との積を求め、第１の言語の各単語ペアＥ（ｉ）およびＥ（ｊ）に関して上記の積を可算した和をとることによって第２の言語で一つの文の中に出現する二つの単語のペアの出現確率Ｐ（Ｊ（ｍ）＾Ｊ（ｎ））を計算する式を用いることを特徴とする対訳確率付与装置。
【請求項４】
請求項１記載の対訳確率付与装置において、前記対訳確率モデル格納部に格納された式として、第１の言語で構文上の係り受け関係にある二つの単語のペアの共起確率Ｐ（Ｅ（ｉ）＾Ｅ（Ｊ））とその単語ペアを構成する各単語Ｅ（ｉ）およびＥ（Ｊ）が第２の言語の訳語に対応する対応確率Ｓ（ｉ，ｍ）およびＳ（Ｊ，ｎ）との積を求め、第１の言語の各単語ペアＥ（ｉ）およびＥ（Ｊ）に関して上記の積を可算した和をとることによって第２の言語で構文上の係り受け関係にある二つの単語のペアの出現確率Ｐ（Ｊ（ｍ）＾Ｊ（ｎ））を計算する式を用いることを特徴とする対訳確率付与装置。
【請求項５】
第１の言語の文例集と第２の言語の文例集とを有し、
第１の言語の単語に対する第２の言語の訳語候補を単語対応対として格納した第１言語第２言語対訳辞書を有し、
（ａ）前記第１の言語の文例集における単語の出現に関する統計量を計算する処理、
（ｂ）前記第２の言語の文例集における単語の出現に関する統計量を計算する処理、
（ｃ）前記第１言語第２言語対訳辞書の各単語対応対に付与された対訳確率をパラメータとして、前記（ａ）によって計算される統計量に基づいて、第２の言語の単語の出現に関する統計量を推定する計算処理、
（ｄ）前記（ｂ）によって計算された統計量と、前記（ｃ）によって推定された統計量との差を最小にするように前記パラメータを調整する処理、
を実行させることを特徴とするプログラム。
【請求項６】
請求項５記載のプログラムにおいて、前記（ｃ）の処理として、第１の言語の単語Ｅ（ｉ）の出現確率Ｅ（ｉ）と、その単語Ｅ（ｉ）が第２の言語の訳語Ｊ（ｎ）に対応する対応確率Ｓ（ｉ，ｎ）との積を求め、第１の言語の各単語Ｅ（ｉ）に関して上記の積を可算した和をとることによって第２の言語における単語Ｊ（ｎ）の出現確率Ｊ（ｎ）を計算する処理を行なうことを特徴とするプログラム。
【請求項７】
請求項５記載のプログラムにおいて、前記（ｃ）の処理として、第１の言語で一つの文の中に出現する二つの単語のペアの共起確率Ｐ（Ｅ（ｉ）＾Ｅ（Ｊ））とその単語ペアを構成する各単語
Ｅ（ｉ）およびＥ（Ｊ）が第２の言語の訳語に対応する対応確率Ｓ（ｉ，ｍ）およびＳ（Ｊ，ｎ）との積を求め、第１の言語の各単語ペアＥ（ｉ）およびＥ（Ｊ）に関して上記の積を可算した和をとることによって第２の言語で一つの文の中に出現する二つの単語のペアの出現確率Ｐ（Ｊ（ｍ）＾Ｊ（ｎ））を計算する処理を行なうことを特徴とするプログラム。
【請求項８】
請求項５記載のプログラムにおいて、前記（ｃ）の処理として、第１の言語で構文上の係り受け関係にある二つの単語のペアの共起確率Ｐ（Ｅ（ｉ）＾Ｅ（Ｊ））とその単語ペアを構成する各単語Ｅ（ｉ）およびＥ（Ｊ）が第２の言語の訳語に対応する対応確率Ｓ（ｉ，ｍ）およびＳ（Ｊ，ｎ）との積を求め、第１の言語の各単語ペアＥ（ｉ）および

Ｅ（Ｊ）に関して上記の積を可算した和をとることによって第２の言語で構文上の係り受け関係にある二つの単語のペアの出現確率Ｐ（Ｊ（ｍ）＾Ｊ（ｎ））を計算する処理を行なうことを特徴とするプログラム。

【請求項９】

第１の言語の文例集と第２の言語の文例集と、

第１の言語の単語に対する第２の言語の訳語候補を単語対応対として格納した第１言語第２言語対訳辞書とを有する対訳確率付与装置が、

（ａ）前記第１の言語の文例集における単語の出現に関する統計量を計算し、

（ｂ）前記第２の言語の文例集における単語の出現に関する統計量を計算し、

（ｃ）前記第１言語第２言語対訳辞書の各単語対応対に付与された対訳確率をパラメータとして、前記（ａ）によって計算される統計量に基づいて、第２の言語の単語の出現に関する統計量を推定する計算処理をし、

（ｄ）前記（ｂ）によって計算された統計量と、前記（ｃ）によって推定された統計量との差を最小にするように前記パラメータを調整すること

を特徴とする対訳確率付与方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

機械翻訳、クロス言語テキスト検索など、異なる言語の間で言葉の対応をとることを課題とする自然言語処理技術に関する。

【０００２】

【従来技術】

機械翻訳、クロス言語テキスト検索など、異なる言語の間で言葉の対応をとることを課題とする自然言語処理技術においては、一方の言語の単語を、もう一方の言語の適切な単語に対応させることは非常に重要な課題であり、訳語選択の問題と呼ばれている。この課題が重要な問題であることは、一般にどのような言語対の場合にも当てはまるが、以下では英語と日本語の場合を取り上げ、具体例を示して説明する。

【０００３】

英語の単語は一般に複数の意味をもち、一般にはそれぞれ異なる日本語の単語に対応する。ところが、自然言語処理分野において元の英単語の使われている状況を正しく判断して適切な日本語の単語を選択することは、一般には非常に困難である。たとえば、英語の単語「ｔｅｒｍ」には「期間」という意味の他に「専門用語」という意味があるが、どのような場合に「期間」という意味となり、どのような場合に「専門用語」という意味になるか、という訳語の選択条件を、あらかじめ明示的に記述することは非常に難しい。

【０００４】

この問題を解決する方法として、言葉が実際に使用された例、すなわち実例文を大量に集めてそれを利用する方法が提案されている。

【０００５】

たとえば「野上宏康、熊野明、田中克己、天野真家『既存目的言語文書からの訳語の自動学習方式』情報処理学会第４２回全国大会（平成３年）」（先行技術文献１）では、以下のような方法が提案されている。

【０００６】

まず、異なる言語（日本語と英語など）で、同じ分野の話題を述べている文例を大量に収集しておく。次に、一方の言語（たとえば英語）の単語が、相手言語（たとえば日本語）の訳語候補のうち、どの訳語に対応するかの確からしさを判定する際に、相手言語の文例集における、各訳語候補の出現確率の高さを用いる。たとえば、今、英語の「ｔｅｒｍ」を「期間」と訳すのが確からしいか「専門用語」と訳すのが確からしいかを判断するのに、同じ分野の話題を述べている日本語の文例集の中に出現する「期間」という単語と「専門用語」という単語の頻度を計測し、その多い方を「ｔｅｒｍ」の訳語とする、という手法である。この手法には、相手言語の文例集のみを分析すればよいという利点がある。

【０００７】
また「中島弘之、梶博行『対訳テキストを利用した訳語選択のための共起関係の自動抽出』情報処理学会第３９回全国大会（平成元年）」（先行技術文献２）では、以下のような方法が提案されている。
【０００８】
まず、異なる言語（日本語と英語など）で、一方が他方の翻訳関係にあるような対訳文例集を用意する。さらに、二つの言語の間の対訳辞書を用意し、第１の言語の例文に含まれる単語に対して対訳辞書を引き、訳語候補を挙げる。その例文と対訳関係にある第２の言語の例文の中に出現する訳語候補の頻度を計測し、最も高頻度で現われる訳語候補を、元の単語に対する訳語とする、という手法である。この手法は、互いに翻訳関係にある対訳例文が利用できる場合には、高い精度で訳語を認定できるという利点がある。
【０００９】
【発明が解決しようとする課題】
しかしながら、先行技術文献１の方法は、相手言語の単語の頻度だけを手がかりにしているため、相手言語で一般的に高頻度で出現する単語が訳語として採用されてしまいやすい、という欠点がある。
【００１０】
たとえば、英語の単語「ｍａｋｅ」には「作る」という訳語の他にも多くの日本語の訳語が相当する。一例として「ｍａｋｅ　ａ　ｃａｌｌ」を「電話をする」と訳すためには「ｃａｌｌ」を「電話」に対応するものとし、「ｍａｋｅ」には「する」という動詞が対応するものとして辞書を構築するのが通常の手法である。このように辞書を作る時「ｍａｋｅ」には少なくとも「作る」と「する」という訳語候補が存在することになる。この場合、先行技術文献１の方法に従って、相手言語、つまり日本語の単語の出現頻度だけを計測すると、訳語「作る」よりも訳語「する」の方が一般に出現頻度が高いので、「ｍａｋｅ」の訳語候補として「する」が最も確からしいものとして選択されてしまう。先行技術文献１には、このように、本来の訳語として適切かどうかとは無関係に、相手言語で出現頻度の高い訳語が選択されやすい、という欠点がある。
【００１１】
また、先行技術文献２の方法は、互いに翻訳関係にある対訳例文が大量に存在する場合に有効な方法であるが、実際には、互いに翻訳関係にある対訳例文の量は極めて限られている。先行技術文献２の方法は対訳例文が大量に存在しない場合には適用できない、という欠点がある。
【００１２】
本願発明の目的は、従来の手法がもつ、上記のような問題点を解決し、より確からしい訳語候補を選択するための方法を提供するところにある。
【００１３】
【課題を解決するための手段】
本発明の対訳確率付与装置は、第１の言語を第２の言語に翻訳する際に用いられる対訳確率付与装置であって、第１の言語の文例集と第２の言語の文例集とを有し、第１の言語の単語に対する第２の言語の訳語候補を単語対応対として格納した第１言語第２言語対訳辞書を有し、第１の言語の文例集における単語の出現に関する統計量を計算する第１言語統計量計算モジュールを有し、第２の言語の文例集における単語の出現に関する統計量を計算する第２言語統計量計算モジュールを有し、対訳辞書の各単語対応対に付与された対訳確率をパラメータとして、第１の言語の文例集から第１言語統計量計算モジュールによって求められる統計量から、第２の言語の単語の出現に関する統計量を推定する対訳確率モデルを有し、第２の言語の文例集から第２言語統計量計算モジュールによって求められた統計量と対訳確率モデルによって第１の言語から推定された第２の言語の統計量との差を最小にするようにパラメータを求める対訳確率付与部を有することを特徴とする。
【００１４】
この場合、第１の言語の単語の出現に関する統計量から第２の言語の単語の出現に関する

統計量を推定する対訳確率モデルとして、第１の言語の単語Ｅ（ｉ）の出現確率Ｅ（ｉ）とその単語Ｅ（ｉ）が第２の言語の訳語Ｊ（ｎ）に対応する対応確率Ｓ（ｉ，ｎ）との積を求め、第１の言語の各単語Ｅ（ｉ）に関して上記の積を可算した和をとることによって第２の言語における単語Ｊ（ｎ）の出現確率Ｊ（ｎ）を計算する対訳確率モデルを用いてもよい。

【００１５】

また、第１の言語の単語の出現に関する統計量から第２の言語の単語の出現に関する統計量を推定する対訳確率モデルとして、第１の言語で一つの文の中に出現する二つの単語のペアの共起確率Ｐ（Ｅ（ｉ）＾Ｅ（Ｊ））とその単語ペアを構成する各単語Ｅ（ｉ）およびＥ（Ｊ）が第２の言語の訳語に対応する対応確率Ｓ（ｉ，ｍ）およびＳ（Ｊ，ｎ）との積を求め、第１の言語の各単語ペアＥ（ｉ）およびＥ（Ｊ）に関して上記の積を可算した和をとることによって第２の言語で一つの文の中に出現する二つの単語のペアの出現確率Ｐ（Ｊ（ｍ）＾Ｊ（ｎ））を計算する対訳確率モデルを用いことにしてもよい。

【００１６】

また、第１の言語の単語の出現に関する統計量から第２の言語の単語の出現に関する統計量を推定する対訳確率モデルとして、第１の言語で構文上の係り受け関係にある二つの単語のペアの共起確率Ｐ（Ｅ（ｉ）＾Ｅ（Ｊ））とその単語ペアを構成する各単語Ｅ（ｉ）およびＥ（Ｊ）が第２の言語の訳語に対応する対応確率Ｓ（ｉ，ｍ）およびＳ（Ｊ，ｎ）との積を求め、第１の言語の各単語ペアＥ（ｉ）およびＥ（Ｊ）に関して上記の積を可算した和をとることによって第２の言語で構文上の係り受け関係にある二つの単語のペアの出現確率Ｐ（Ｊ（ｍ）＾Ｊ（ｎ））を計算する対訳確率モデルを用いることにしてもよい。

【００１７】

【発明の実施の形態】

本発明の実施の形態について図面を参照して説明する。図１は本発明の第１実施の形態の訳語選択システムの構成を示すブロック図である。

【００１８】

本実施の形態は、第１言語の文例集１、第２言語の文例集２、第１の言語の単語に対する第２の言語の訳語候補を単語対応対として格納した第１言語第２言語対訳辞書３、第１言語の文例集における単語の出現に関する統計量を計算する第１言語統計量計算モジュール４、第２言語の文例集における単語の出現に関する統計量を計算する第２言語統計量計算モジュール５、対訳辞書の各単語対応対に付与された対訳確率をパラメータとして、第１の言語の文例集から第１言語統計量計算モジュールによって求められる統計量から、第２の言語の単語の出現に関する統計量を推定する対訳確率モデルを格納した対訳確率モデル格納部６、第２の言語の文例集から第２言語統計量計算モジュールによって求められた統計量と対訳確率モデルによって第１の言語から推定された第２の言語の統計量との差を最小にするようにパラメータを求める対訳確率付与部７とから構成されている。

【００１９】

各ブロックの内容と動作について以下に説明する。第１言語の文例集１には、第１の言語、たとえば英語の実例文が格納されている。第２言語の文例集２には、第２の言語、たとえば日本語の実例文が格納されている。第１言語第２言語対訳辞書３には、第１の言語の各単語に対する第２の言語の訳語候補を単語対応対として格納してある。図２は、第１言語第２言語対訳辞書３の内容の例を示した図である。この図では、第１言語の単語Ｅ（ｉ）に対応する第２言語の訳語候補として、Ｊ（ｋ）、Ｊ（ｍ）、Ｊ（ｎ） が存在する場合を示している。

【００２０】

この図でｅ（ｉ）は、第１言語の単語Ｅ（ｉ）の出現確率、ｊ（ｋ）、ｊ（ｍ）、ｊ（ｎ）はそれぞれ第２言語の単語Ｊ（ｋ）、Ｊ（ｍ）、Ｊ（ｎ）の出現確率を表す。また、Ｓ（ｉ，ｋ）、Ｓ（ｉ，ｍ）、Ｓ（ｉ，ｍ）は、それぞれ、第１言語の単語Ｅ（ｉ）が、第２言語の単語Ｊ（ｋ）、Ｊ（ｍ）、Ｊ（ｎ）に翻訳される確率を表す。

【００２１】
第１言語統計量計算モジュール４は、第１言語の文例集１における単語の出現に関する統計量を計算する。第２言語統計量計算モジュール５は、第２言語の文例集２における単語の出現に関する統計量を計算する。第１言語統計量計算モジュール４および第２言語統計量計算モジュール５は、必要に応じて、第１言語の文例集１および第２言語の文例集２に含まれる文を形態素解析したり構文解析したりして、そこに含まれる単語の出現に関する統計量を計算する。統計量の例としては、各単語の出現確率や二つの単語が同時に出現する共起確率などがある。
【００２２】
対訳確率モデル格納部６には、第１の言語の単語の出現に関する統計量から第２の言語の単語の出現に関する統計量を推定する対訳確率モデルが格納してある。この対訳確率モデルは、対訳辞書の各単語対応対に付与された対訳確率をパラメータとして、第１の言語の文例集から第１言語統計量計算モジュールによって求められる統計量から、第２の言語の単語の出現に関する統計量を推定する。
【００２３】
対訳確率付与部７は、第２の言語の文例集から第２言語統計量計算モジュールによって求められた統計量と対訳確率モデルによって第１の言語から推定された第２の言語の統計量との差を最小にするように、対訳辞書の各単語対応対に付与された対訳確率パラメータを調整する。
【００２４】
図６は、本願発明の第２の実施の形態を説明する図である。
【００２５】
図６において、本願発明の第２の実施の形態は、入力装置１０１と、コンピュータから構成されるデータ処理装置１０２と、出力装置１０３と、記憶装置１０４と、訳語選択プログラムを記録した記憶媒体１０５とを備える。記憶媒体１０５は、磁気ディスク、磁気テープ、光ディスク、半導体メモリその他の記憶媒体よりなる。
【００２６】
訳語選択プログラムは、記憶媒体１０５からデータ処理装置１０２の主記憶装置に読み込まれ、データ処理装置１０２の動作を制御する。データ処理装置１０２は、訳語選択プログラムの制御により以下の処理を行なう。
【００２７】
訳語の選択を行なうべき単語が入力装置１０１から入力されると、第１言語統計量計算モジュール４と第２言語統計量計算モジュール５とが起動される。第１言語統計量計算モジュール４は、第１言語の文例集１における単語の出現に関する統計量を計算する。第２言語統計量計算モジュール５は、第２言語の文例集２における単語の出現に関する統計量を計算する。
【００２８】
次に、対訳確率付与部７が起動される。対訳確率付与部７は、第２の言語の文例集から第２言語統計量計算モジュールによって求められた統計量と対訳確率モデル格納部６に格納された対訳確率モデルによって第１の言語から推定された第２の言語の統計量との差を最小にするように、対訳辞書の各単語対応対に付与された対訳確率パラメータを調整する。
【００２９】
結果として得られた対訳確率パラメータの値にしたがって、訳語が出力装置１０３から出力される。
【００３０】
次に、図１に示した実施の形態における、対訳確率モデル格納部６に格納されている対訳確率モデルの例を用いて、本願発明の動作を説明する。次の式は、対訳確率モデルの一例を表す式である。
【００３１】
【数１】

$$j(m) = \sum_{i} e(i) \times S(i,m)$$

（式１）

この式１において、ｅ（ｉ）は第１言語のｉ番目の単語Ｅ（ｉ）の出現確率を表す。またｊ（ｍ）は第２言語のｍ番目の単語Ｊ（ｍ）の出現確率を表す。Ｓ（ｉ，ｍ）は、第１言語のｉ番目の単語Ｅ（ｉ）が、第２言語のｍ番目の単語Ｊ（ｍ）に翻訳される確率を表す。この式は、第１言語の各単語の出現確率と翻訳確率の積の総和が第２言語の各単語の出現確率を与えるというモデルを表している。

【００３２】

この式１のＳ（ｉ，ｍ）が、この対訳確率モデルにおけるパラメータであり、第１言語の単語Ｅ（ｉ）と第２言語の訳語候補Ｊ（ｍ）との単語対応対に与えられた対訳確率である。このパラメータには、第１言語の単語は第２言語の単語に必ず対応するという仮定の下で、

【００３３】

【数２】

$$\sum_{m} S(i,m) = 1$$

（式２）

という制約がある。

【００３４】

この対訳確率モデルによって各単語の対訳確率を求めるには、第１言語統計量計算モジュール４によって、第１言語の文例集１における単語の出現確率ｅ（ｉ） を計算し、第２言語統計量計算モジュール５によって、第２言語の文例集２における単語の出現確率ｊ（ｍ）を計算し、このようにして求めたｅ（ｉ）およびｊ（ｍ）を上記の対訳確率モデルの式に代入して上記の制約を満たすパラメータＳ（ｉ，ｍ）を定める。

【００３５】

次に、図３、図４、図５を用いて、本願発明と従来方式の差異を説明する。ここでは、例として、英語の単語を日本語の単語に翻訳する場合を考える。

【００３６】

図３は、第１言語第２言語対訳辞書３の中の英単語「ｄｏ」と「ｍａｋｅ」の単語対応対を示している。ここでは簡単のため、英単語「ｄｏ」は日本語の単語「する」１語とだけ訳語候補としての単語対応対をなしており、英単語「ｍａｋｅ」は日本語の単語の「つくる」と「する」の２単語と、訳語候補としての単語対応対をなしている状況を想定する。

【００３７】

図４は、先行技術文献１で示されているような、第２言語の文例集における単語の出現頻度だけを用いて、第１言語の単語の訳語選択を行なう従来方式の動作を、図３で示した単語対応対の構成をもった「ｄｏ」と「ｍａｋｅ」を例に挙げて表した図である。図４は第２言語の文例集における「する」と「つくる」の出現確率が、仮にそれぞれ、０．２０および０．０１である状況を示している。この場合、先行技術文献１の従来方式では、「ｍａｋｅ」の訳語として、出現確率の高い単語「する」が単語「つくる」よりも優先されてしまう。

【００３８】

図５は、本願発明の動作を、図３で示した単語対応対の構成をもった「ｄｏ」と「ｍａｋｅ」を例に挙げて表した図である。図５では、第１言語の文例集における「ｄｏ」と「ｍ

ａｋｅ」の出現確率が、仮にそれぞれ、０．１８および０．０２である状況を示している。第２言語の文例集における「する」と「つくる」の出現確率は、図４の場合と同様に、それぞれ、０．２０および０．０１であるとする。

【００３９】

本願発明では、上述の式で示したような対訳確率モデルを用いて、日本語の同じ単語を訳語としてもつ英単語の影響を考慮した計算を行なう。この方法で、英単語「ｍａｋｅ」が「する」に翻訳される確率および「つくる」に翻訳される確率を計算すると、この例のように「する」の頻度が高くても、その頻度の大部分は英単語「ｄｏ」からの翻訳確率に対応するので、「ｍａｋｅ」から「する」への翻訳確率は低くなる。図５では「ｍａｋｅ」から「つくる」への翻訳確率が０．９、「ｍａｋｅ」から「する」への翻訳確率が０．１という結果が得られた場合を示している。

【００４０】

次に、対訳確率モデル格納部６に格納されている対訳確率モデルの第２の例を用いて、本願発明の動作を説明する。次に挙げる式は、対訳確率モデルの一例を表す式である。

【００４１】

【数３】

$$P(J(m)\hat{}J(n)) = \sum_{i,j} P(E(i)\hat{}E(j)) \times S(i,m) \times S(j,n)$$

（式３）

この式３において、Ｐ（Ｅ（ｉ）＾Ｅ（ｊ））は、第１言語で単語Ｅ（ｉ）と単語Ｅ（ｊ）が同時に出現する共起確率を表す。また、Ｐ（Ｊ（ｍ）＾Ｊ（ｎ）） は、第２言語で単語Ｊ（ｍ）と単語Ｊ（ｎ）が同時に出現する共起確率を表す。この式３は、第１言語における二つの単語の共起確率とそれぞれの単語の対訳確率の積の総和が、第２言語における二つの単語の共起確率を与えるというモデルを表している。

【００４２】

この式のＳ（ｉ，ｍ）およびＳ（ｊ，ｎ）が、この対訳確率モデルにおけるパラメータであり、それぞれ、第１言語の単語 Ｅ（ｉ） と第２言語の訳語候補Ｊ（ｍ）との単語対応対に与えられた対訳確率、第１言語の単語Ｅ（Ｊ）と第２言語の訳語候補Ｊ（ｎ）との単語対応対に与えられた対訳確率である。このパラメータには、第１言語の単語は第２言語の単語に必ず対応するという仮定の下で、

【００４３】

【数４】

$$\sum_{m} S(i, m) = 1$$

（式４）

という制約がある。

【００４４】

この対訳確率モデルを使って各単語の対訳確率を求める場合、二つの単語の共起として、何種類かの共起が考えられる。共起の種類の一つとして、一つの文の中に二つの単語が共に出現する文内共起がある。

【００４５】

この場合、第１言語統計量計算モジュール４によって、第１言語の文例集１における二つの単語の文内共起確率Ｐ（Ｅ（ｉ）＾Ｅ（Ｊ））を計算し、２言語統計量計算モジュール５によって、第２言語の文例集２における二つの単語の文内共起確率Ｐ（Ｊ（ｍ）＾Ｊ（

ｎ））を計算し、このようにして求めたＰ（Ｅ（ｉ）＾Ｅ（Ｊ））およびＰ（Ｊ（ｍ）＾Ｊ（ｎ））を上記の対訳確率モデルの式に代入して、上記の制約を満たすパラメータＳ（ｉ，ｍ）を定める。

【００４６】

もう一つの共起の種類として、二つの単語が、互いに構文的な係り受け関係にある係り受け共起がある。この場合、第１言語統計量計算モジュール４によって、第１言語の文例集１における二つの単語の係り受け共起確率Ｐ（Ｅ（ｉ）＾Ｅ（ｊ））を計算し、第２言語統計量計算モジュール５によって、第２言語の文例集２における二つの単語の係り受け共起確率Ｐ（Ｊ（ｍ）＾Ｊ（ｎ））を計算し、このようにして求めたＰ（Ｅ（ｉ）＾Ｅ（ｊ））およびＰ（Ｊ（ｍ）＾Ｊ（ｎ））を上記の対訳確率モデルの式に代入して、上記の制約を満たすパラメータＳ（ｉ，ｍ）を定める。

【００４７】

【発明の効果】

本願発明によれば、第１言語の単語の訳語を定める際、先行技術文献１とは異なり、第１言語と第２言語の両方の全体の単語の対訳確率を考慮に入れるため、第２言語で出現確率の高い単語が訳語に選ばれやすいという先行技術文献１のもっていた欠点が解消されている。

【００４８】

また、本願発明で用いる第１言語および第２言語の文例集は互いに翻訳関係にあることを仮定していないため、大量に収集することができる。互いに翻訳関係にある文例集が存在しないと適用できないという先行技術文献２のもっていた欠点が解消されている。

【００４９】

さらに、本願発明では文内共起を用いて対訳確率を求めるため、単独の単語の対訳確率だけを用いる場合に比べて、複合語などの場合の翻訳の精度が向上する。

【００５０】

また、本願発明では、係り受け共起を用いて対訳確率を求めるため、動詞とその格要素の名詞が組になって訳語が定まるような場合の翻訳の精度が向上する。

【図面の簡単な説明】

【図１】　本発明の第１の形態をなす訳語選択システムの構成を示すブロック図である。

【図２】　図１に示した実施例における、第１言語第２言語対訳辞書３の内容例を表す図である。

【図３】　従来法と本発明の動作を比較するための、第１言語第２言語対訳辞書３の内容例を示す図である。

【図４】　従来法の動作を説明するための第１言語第２言語対訳辞書３の内容例を示す図である。

【図５】　本発明の動作を説明するための第１言語第２言語対訳辞書３の内容例を示す図である。

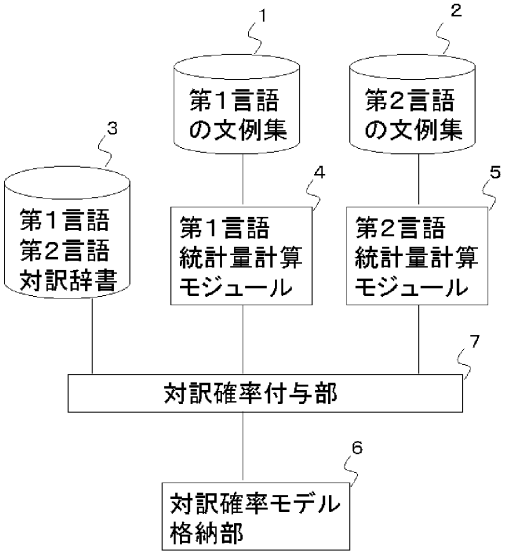
【図６】　本発明の第２の実施の形態をなす訳語選択システムの構成を示すブロック図である。

【符号の説明】

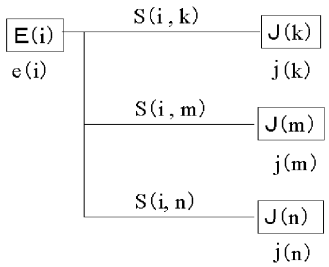
１　　第１言語の文例集
２　　第２言語の文例集
３　　第１言語第２言語対訳辞書
４　　第１言語統計量計算モジュール
５　　第２言語統計量計算モジュール
６　　対訳確率モデル格納部
７　　対訳確率付与部
１０１　入力装置
１０２　データ処理装置
１０３　出力装置

１０４　記憶装置
１０５　記録媒体

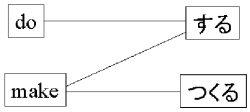
【図１】



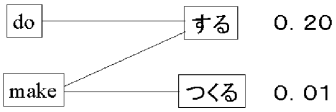
【図２】



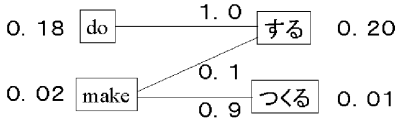
【図３】


do
する
make
つくる

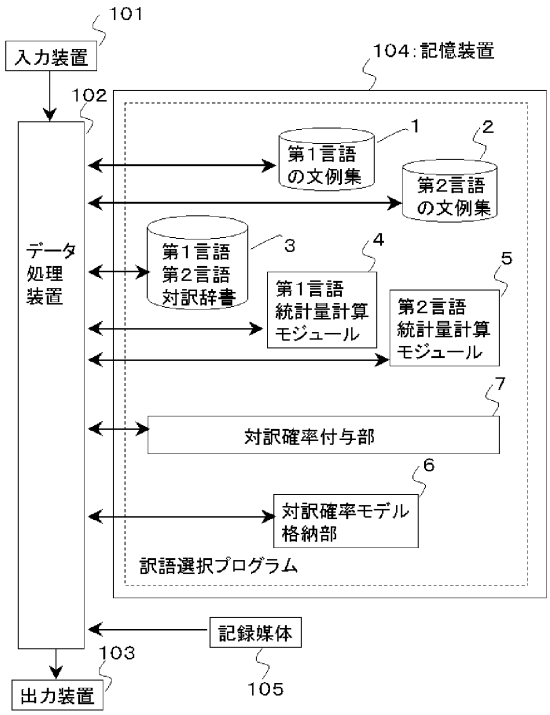
【図４】


do
する
0. 20
make
つくる
0. 01

【図５】


0. 18
do
1. 0
する
0. 20
0. 1
0. 02
make
0. 9
つくる
0. 01

【図６】


101
入力装置
104:記憶装置
102
1
第1言語
の文例集
2
第2言語
の文例集
データ
処理
装置
3
第1言語
第2言語
対訳辞書
4
第1言語
統計量計算
モジュール
5
第2言語
統計量計算
モジュール
7
対訳確率付与部
6
対訳確率モデル
格納部
訳語選択プログラム
記録媒体
103
出力装置
105

フロントページの続き

(72)発明者　長田　誠也
東京都港区芝五丁目７番１号　日本電気株式会社内

審査官　成瀬　博之

(56)参考文献　特開２００２－２２２１８８（ＪＰ，Ａ）

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