(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2017-84274
(P2017-84274A)
(43)公開日　平成29年5月18日(2017.5.18)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/28	(2006.01)	G06F	17/28	627	5B091
G06N 99/00	(2010.01)	G06F	17/28	618	
		G06N	99/00	150	

審査請求 未請求　請求項の数 8　ＯＬ　（全 16 頁）

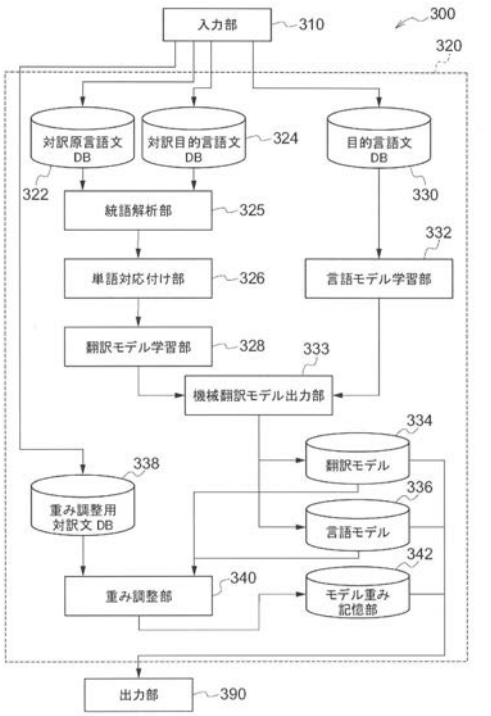
(21)出願番号　特願2015-214659 (P2015-214659)
(22)出願日　平成27年10月30日 (2015.10.30)

(71)出願人　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
(71)出願人　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一丁目５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一丁目５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】単語対応付け装置、機械翻訳学習装置、方法、及びプログラム

(57)【要約】
【課題】対訳関係にある語の自動対応付けを精度良く行う。
【解決手段】単語対応付け部３２６により、対訳となる第１言語文及び第２言語文のペアに含まれる第１言語文について、第２言語の統語要素であって、かつ第１言語に存在しない統語要素に対応する予め定義した仮想単語を、第１言語文に挿入し、仮想単語を挿入した第１言語文に含まれる単語の各々と、第２言語文に含まれる単語の各々との単語の対応関係を推定し、推定された単語の対応関係に基づいて、仮想単語と対応付けられた単語の対応関係を除去し、かつ単語の対応関係に含まれる仮想単語を除去した結果を単語対応付け結果とする。
【選択図】図４



【特許請求の範囲】
【請求項１】
対訳となる第１言語文及び第２言語文のペアに含まれる第１言語文について、
前記第２言語の統語要素であって、かつ前記第１言語に存在しない統語要素に対応する予め定義した仮想単語を、前記第１言語文に挿入し、
前記仮想単語を挿入した前記第１言語文に含まれる単語の各々と、前記第２言語文に含まれる単語の各々との単語の対応関係を推定し、
前記推定された単語の対応関係に基づいて、前記仮想単語と対応付けられた前記単語の対応関係を除去し、かつ前記単語の対応関係に含まれる前記仮想単語を除去した結果を単語対応付け結果とする単語対応付け部
を含む、単語対応付け装置。
【請求項２】
前記第１言語文の統語解析を行う統語解析部を更に含み、
前記単語対応付け部は、前記統語解析部による前記第１言語文の統語解析結果に基づいて、前記第２言語の統語要素であって、かつ前記第１言語に存在しない統語要素に対応する予め定義した仮想単語を、前記第１言語文に挿入する請求項１記載の単語対応付け装置。
【請求項３】
前記第１言語を日本語とし、
前記第２言語を英語とし、
前記単語対応付け部は、前記英語の冠詞に対応する予め定義した仮想単語を、前記日本語文の名詞に係る単語のうち最も左側にある形容詞、前記名詞の直前、又は前記形容詞を修飾している副詞の直前に挿入する請求項１又は２記載の単語対応付け装置。
【請求項４】
請求項１～請求項３の何れか１項記載の単語対応付け装置によって取得した単語対応付け結果に基づいて、前記第１言語の語句が前記第２言語の語句に翻訳される確率を計算したモデルを学習する翻訳モデル学習部
を含む、機械翻訳学習装置。
【請求項５】
単語対応付け部を含む単語対応付け装置における、単語対応付け方法であって、
前記単語対応付け部は、対訳となる第１言語文及び第２言語文のペアに含まれる第１言語文について、
前記第２言語の統語要素であって、かつ前記第１言語に存在しない統語要素に対応する予め定義した仮想単語を、前記第１言語文に挿入し、
前記仮想単語を挿入した前記第１言語文に含まれる単語の各々と、前記第２言語文に含まれる単語の各々との単語の対応関係を推定し、
前記推定された単語の対応関係に基づいて、前記仮想単語と対応付けられた前記単語の対応関係を除去し、かつ前記単語の対応関係に含まれる前記仮想単語を除去した結果を単語対応付け結果とする
単語対応付け方法。
【請求項６】
統語解析部が前記第１言語文の統語解析を行うことを更に含み、
前記単語対応付け部により仮想単語を挿入することは、前記統語解析部による前記第１言語文の統語解析結果に基づいて、前記第２言語の統語要素であって、かつ前記第１言語に存在しない統語要素に対応する予め定義した仮想単語を、前記第１言語文に挿入する請求項５記載の単語対応付け方法。
【請求項７】
翻訳モデル学習部を含む機械翻訳学習装置における、機械翻訳学習方法であって、
前記翻訳モデル学習部は、請求項５又は請求項６記載の単語対応付け方法によって取得した単語対応付け結果に基づいて、前記第１言語の語句が前記第２言語の語句に翻訳され

る確率を計算したモデルを学習する
機械翻訳学習方法。
【請求項８】
コンピュータを、請求項１～請求項３の何れか１項記載の単語対応付け装置、又は請求項４記載の機械翻訳学習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、入力された第１言語と第２言語とにおいて単語の対応関係を取得するための単語対応付け装置、機械翻訳学習装置、方法、及びプログラムに関するものである。
【背景技術】
【０００２】
従来、原言語から目的言語への機械翻訳において、原言語に存在しない統語要素を目的言語への翻訳時に訳出することは一般に容易でなかった。近年用いられている統計的機械翻訳の技術においては、言語に依存しない機械翻訳を実現できる（非特許文献１）。一方で、原言語に存在しない統語要素に対して誤った対訳語句対が獲得され、その結果として翻訳時に訳語の漏れや湧き出しが起こったりするという問題がある。
【０００３】
このような問題への対処方法として、原言語側に存在しないが目的言語側で必要な統語要素を原言語側に補うことによって訳出しやすくする技術がある（非特許文献２、特許文献１）。いずれも韓国語あるいは日本語に存在する主語や目的語を示す助詞相当の統語要素を、英語の統語解析の結果を利用して補うことで、英語から韓国語あるいは日本語への機械翻訳を改善している。
【０００４】
また、翻訳時に英語側の言語モデルや統語構造を考慮することによって、冠詞を後処理として補完する技術も提案されている（非特許文献３）。また、従来の統語構造を利用する方法（非特許文献４）も提案されている。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】特開２０１１－１７５５００号公報
【非特許文献】
【０００６】
【非特許文献１】Phillip Koehn他，”Statistical Phrase-based Translation,” Proc. HLT- NAACL, pp. 263-270,2003.
【非特許文献２】Gumwon Hong他，”Bridging Morpho-Syntactic Gap between Source and Target Sentences for English-Korean Statistical Machine Translation,” Proceeding of the ACL-IJCNLP 2009 Conference Short Papers, pp. 233-236, 2009.
【非特許文献３】Isao Goto他，”Post-ordering by Parsing for Japanese-English Statistical Machine Translation,” Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume2:Short Papers), pp. 311-316, 2012.
【非特許文献４】林克彦他，単語並べ替えと冠詞生成の同時逐次処理：日英機械翻訳への適用，自然言語処理 Vol. 21 No. 5, pp.1037-1057, 2014.
【発明の概要】
【発明が解決しようとする課題】
【０００７】
しかし、上述した従来の方法では、日本語から英語への機械翻訳において補うべき統語要素の一つである冠詞は、抽象名詞や複数形の名詞には不定冠詞が付されないため、上述した単純な方法によるだけでは日本語側に過剰に冠詞相当語が補完されてしまうため、機

械翻訳の結果に不必要な冠詞が多数挿入されてしまうという問題がある。

【０００８】

また、上述した非特許文献３に記載の従来の言語モデルによる方法では原言語側の句とは全く無関係に冠詞が挿入される可能性があるという問題がある。また、上述した非特許文献４記載の方法においては、目的に特化した英語側の構文解析器を要することが問題となる。

【０００９】

本発明では、上記問題点を解決するために成されたものであり、対訳関係にある語の自動対応付けを精度良く行うことができる単語対応付け装置、機械翻訳学習装置、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】

【００１０】

上記目的を達成するために、第１の発明に係る単語対応付け装置は、対訳となる第１言語文及び第２言語文のペアに含まれる第１言語文について、前記第２言語の統語要素であって、かつ前記第１言語に存在しない統語要素に対応する予め定義した仮想単語を、前記第１言語文に挿入し、前記仮想単語を挿入した前記第１言語文に含まれる単語の各々と、前記第２言語文に含まれる単語の各々との単語の対応関係を推定し、前記推定された単語の対応関係に基づいて、前記仮想単語と対応付けられた前記単語の対応関係を除去し、かつ前記単語の対応関係に含まれる前記仮想単語を除去した結果を単語対応付け結果とする単語対応付け部を含んで構成されている。

【００１１】

第２の発明に係る単語対応付け方法は、単語対応付け部を含む単語対応付け装置における、単語対応付け方法であって、前記単語対応付け部は、対訳となる第１言語文及び第２言語文のペアに含まれる第１言語文について、前記第２言語の統語要素であって、かつ前記第１言語に存在しない統語要素に対応する予め定義した仮想単語を、前記第１言語文に挿入し、前記仮想単語を挿入した前記第１言語文に含まれる単語の各々と、前記第２言語文に含まれる単語の各々との単語の対応関係を推定し、前記推定された単語の対応関係に基づいて、前記仮想単語と対応付けられた前記単語の対応関係を除去し、かつ前記単語の対応関係に含まれる前記仮想単語を除去した結果を単語対応付け結果とする。

【００１２】

第１及び第２の発明によれば、単語対応付け部により、対訳となる第１言語文及び第２言語文のペアに含まれる第１言語文について、第２言語の統語要素であって、かつ第１言語に存在しない統語要素に対応する予め定義した仮想単語を、第１言語文に挿入し、仮想単語を挿入した第１言語文に含まれる単語の各々と、第２言語文に含まれる単語の各々との単語の対応関係を推定し、推定された単語の対応関係に基づいて、仮想単語と対応付けられた単語の対応関係を除去し、かつ単語の対応関係に含まれる仮想単語を除去した結果を単語対応付け結果とする。

【００１３】

このように、対訳となる第１言語文及び第２言語文のペアに含まれる第１言語文について、仮想単語を、第１言語文に挿入し、仮想単語を挿入した第１言語文に含まれる単語の各々と、第２言語文に含まれる単語の各々との単語の対応関係を推定し、推定された単語の対応関係に基づいて、仮想単語と対応付けられた単語の対応関係を除去し、かつ単語の対応関係に含まれる仮想単語を除去した結果を単語対応付け結果とすることにより対訳関係にある語の自動対応付けを精度良く行うことができる。

【００１４】

また、第１及び第２の発明において、統語解析部が前記第１言語文の統語解析を行うことを更に含み、前記単語対応付け部により仮想単語を挿入することは、前記統語解析部による前記第１言語文の統語解析結果に基づいて、前記第２言語の統語要素であって、かつ前記第１言語に存在しない統語要素に対応する予め定義した仮想単語を、前記第１言語文に挿入してもよい。

【００１５】
また、第１の発明において、前記第１言語を日本語とし、前記第２言語を英語とし、前記単語対応付け部は、前記英語の冠詞に対応する予め定義した仮想単語を、前記日本語文の名詞に係る単語のうち最も左側にある形容詞、前記名詞の直前、又は前記形容詞を修飾している副詞の直前に挿入してもよい。
【００１６】
また、第３の発明に係る機械翻訳学習装置は、第１の発明に係る単語対応付け装置によって取得した単語対応付け結果に基づいて、前記第１言語の語句が前記第２言語の語句に翻訳される確率を計算したモデルを学習する翻訳モデル学習部、を含んで構成されている。
【００１７】
第４の発明に係る機械翻訳学習方法は、翻訳モデル学習部を含む機械翻訳学習装置における、機械翻訳学習方法であって、前記翻訳モデル学習部は、第２の発明に係る単語対応付け方法によって取得した単語対応付け結果に基づいて、前記第１言語の語句が前記第２言語の語句に翻訳される確率を計算したモデルを学習する。
【００１８】
第３及び第４の発明によれば、翻訳モデル学習部により、第１又は第２の発明によって取得した単語対応付け結果に基づいて、第１言語の語句が第２言語の語句に翻訳される確率を計算したモデルを学習する。
【００１９】
このように、第１又は第２の発明によって取得した単語対応付け結果に基づいて、第１言語の語句が第２言語の語句に翻訳される確率を計算したモデルを学習することによって、精度良くモデルを学習することができる。
【００２０】
また、本発明のプログラムは、コンピュータを、上記の単語対応付け装置、又は機械翻訳学習装置を構成する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】
【００２１】
以上説明したように、本発明の単語対応付け装置、方法、及びプログラムによれば、対訳となる第１言語文及び第２言語文のペアに含まれる第１言語文について、仮想単語を、第１言語文に挿入し、仮想単語を挿入した第１言語文に含まれる単語の各々と、第２言語文に含まれる単語の各々との単語の対応関係を推定し、推定された単語の対応関係に基づいて、仮想単語と対応付けられた単語の対応関係を除去し、かつ単語の対応関係に含まれる仮想単語を除去した結果を単語対応付け結果とすることにより対訳関係にある語の自動対応付けを精度良く行うことができる。
【００２２】
また、本発明の機械翻訳学習装置、方法、及びプログラムによれば、本発明の単語対応位付け装置により取得した単語対応付け結果に基づいて、第１言語の語句が第２言語の語句に翻訳される確率を計算したモデルを学習することによって、精度良くモデルを学習することができる。
【図面の簡単な説明】
【００２３】
【図１】一般的な統計翻訳の一例を示す図である。
【図２】仮想単語の補間（挿入）の一例を示す図である。
【図３】仮想単語の除去の一例を示す図である。
【図４】本発明の実施形態に係る機械翻訳学習装置の機能的構成を示すブロック図である。
【図５】本発明の実施形態に係る機械翻訳学習装置の単語対応付け部の機能的構成を示すブロック図である。
【図６】本発明の実施形態に係る機械翻訳装置の機能的構成を示すブロック図である。

【図７】本発明の実施形態に係る機械翻訳学習装置における機械翻訳学習処理ルーチンのフローチャート図である。
【図８】本発明の実施形態に係る機械翻訳装置における機械翻訳処理ルーチンのフローチャート図である。
【発明を実施するための形態】
【００２４】
以下、図面を参照して本発明の実施形態を詳細に説明する。
【００２５】
＜本発明の実施形態の概要＞
まず、本発明の実施形態の概要について説明する。本実施形態は、非特許文献１記載のような統計的機械翻訳を実現する機械翻訳装置において、自動単語対応付け処理を行う前に、原言語に冠詞相当の仮想単語を補完する処理を行うことと、自動単語対応付け処理完了後に補完した仮想単語と仮想単語に関わる単語対応付けを除去してから対訳語句対の対応付けと対訳語句対集合の獲得、及び翻訳モデルの学習を行うように構成することとを特徴とする。
【００２６】
そのため、仮想単語の補完処理によって、目的言語の冠詞は原言語側に補完された冠詞相当の仮想単語と多く対応付けられ、原言語側のその他の単語と誤って対応付けられることを抑制できる。
【００２７】
そして仮想単語と仮想単語に関わる単語対応付けの除去によって、仮想単語を含まず、かつ冠詞の誤った対応を含まないような原言語と目的言語の対訳語句対が獲得できる。
【００２８】
これによって冠詞は非特許文献１に代表される統計的機械翻訳における「ＮＵＬＬ対応」と呼ばれる、対応相手のいない単語となり、翻訳時に冠詞を付すか否かが翻訳モデル・言語モデルのスコアに応じて自動的に選択されるようになる。
【００２９】
本実施形態においては、自動単語対応付け前に冠詞相当の仮想単語を補完する。具体的には原言語の、別途ルールで定めた名詞句相当箇所の先頭に仮想単語を挿入する処理と、単語対応付け処理完了後に仮想単語と仮想単語を含む単語対応付けを除去してから対訳語句対を抽出する処理とを行う。
【００３０】
ここで、図１に一般的な統計翻訳の例を示す。図１の例においては、「は」が冠詞「ｔｈｅ」に誤って対応付けられている。これは、原言語である日本語側に冠詞相当語がないことによる誤った対応付けを表す。そのため、「は」が「ｔｈｅ」と訳されるような誤った翻訳知識が学習されることになる。また、「流体圧シリンダは」も、必ず「ｔｈｅ」付きで翻訳されることになる。
【００３１】
また、図２に仮想単語の補間（挿入）の例を示す。図２の例においては、冠詞「ｔｈｅ」は、冠詞相当仮想単語「＿ａ」に正しく対応付けられている。また、この場合、仮想単語がないと「ｔｈｅ」を訳出することができない。また、仮想単語が過剰であると、不必要な冠詞が訳出される可能性がある。また、仮想単語が冠詞以外に誤って対応付けられると、図１の例と同様の問題を有することになる。
【００３２】
さらに、図３に仮想単語の除去の例を示す。図３の例においては、冠詞相当仮想単語「＿ａ」と仮想単語の対応付けをすべて除去する。また、仮想単語なしで冠詞の訳出が「選択的に」可能となる。そのため、対応付けのない単語は隣接する語句に連結した対訳語句対の形で学習される。また、仮想単語に誤って対応付けられた単語の影響を排除することができる。
【００３３】

なお、図１から図３の単語間を結ぶ実線が単語対応を示し、右側が単語対応から得られる対訳語句対を示している。図２の仮想単語の挿入及び図３の仮想単語の除去の処理を行うことにより、仮想単語の挿入によって冠詞の誤った単語対応を抑止し、かつ対訳語句対の抽出時には削除することで仮想単語がなくても翻訳ができるような対訳語句対を獲得することを可能にする。また、特許文献１や、非特許文献２は、図２の場合に相当し、翻訳時にも補完をしておかなければならない構成である。また、図１から図３の例においては、原言語を日本語とし、目的言語を英語とした例である。なお、原言語は、日本語に限られず、また、目的言語も英語に限られず、原言語、及び目的言語を任意の言語としてもよい。

【００３４】

＜本発明の実施形態に係る機械翻訳学習装置の構成＞

次に、本発明の実施の形態に係る機械翻訳学習装置の構成について説明する。図４に示すように、本発明の実施の形態に係る機械翻訳学習装置３００は、ＣＰＵと、ＲＡＭと、後述する機械翻訳学習処理ルーチンを実行するためのプログラムや各種データを記憶したＲＯＭと、を含むコンピュータで構成することができる。この機械翻訳学習装置３００は、機能的には図４に示すように入力部３１０と、演算部３２０と、出力部３９０とを備えている。

【００３５】

また、本実施形態においては、例えば、原言語である日本語を第１言語とし、目的言語である英語を第２言語とする。なお、本実施形態においては、第１言語を原言語である日本語とし、第２言語を目的言語である英語として説明するが、第１言語が原言語、又は目的言語であって、かつ第２言語が他方の言語としてもよい。また、第１言語、及び第２言語の組み合わせも日本語と英語との組み合わせに限定されず、他の２言語の組み合わせを用いてもよい。

【００３６】

入力部３１０は、機械翻訳のための学習データとして、対訳文である原言語文と目的言語文とのペアの集合の入力を受け付ける。

【００３７】

また、入力部３１０は、目的言語文の集合の入力を受け付ける。

【００３８】

また、入力部３１０は、モデルの重み調整のための学習データとして、対訳文である原言語文と目的言語文とのペアの集合の入力を受け付ける。

【００３９】

演算部３２０は、対訳原言語文データベース３２２、対訳目的言語文データベース３２４、統語解析部３２５、単語対応付け部３２６、翻訳モデル学習部３２８、目的言語文データベース３３０、言語モデル学習部３３２、機械翻訳モデル出力部３３３、翻訳モデル３３４、言語モデル３３６、重み調整用対訳文データベース３３８、重み調整部３４０、及びモデル重み記憶部３４２を備えている。

【００４０】

対訳原言語文データベース３２２は、入力部３１０により受け付けた対訳文の原言語文の集合を記憶している。

【００４１】

対訳目的言語文データベース３２４は、入力部３１０により受け付けた対訳文の目的言語文の集合を記憶している。

【００４２】

統語解析部３２５は、対訳文である原言語文及び目的言語文のペアの各々について、当該ペアの原言語文及び目的言語文の各々を統語解析し、原言語の構文木を取得する。なお、統語解析には、単語分割や品詞付与の処理を含みえる。また、統語解析の方法は公知の技術、例えば英語についてはBerkeley ParserやEnju等のソフトウェア、日本語についてはHaruniwaやCkylark等のソフトウェアが利用できるが、本実施形態の構成は特定の統語

解析技術に依存しないため、句構造解析に限らず依存構造解析を利用してもよい。また、統語構造を要さず、表層や品詞の情報のみで後述の仮想単語挿入が可能な場合は統語構造の解析を省略してもよい。また、原言語文のみ統語解析してもよい。また、本実施形態においては、後述するように、原言語側に仮想単語を挿入する場合について想定しているため、原言語のみ統語解析してもよいと説明しているが、目的言語側に仮想単語を挿入する場合には、目的言語のみ統語解析してもよい。

【００４３】

単語対応付け部３２６は、対訳文である原言語文及び目的言語文のペアの各々について、当該ペアの原言語文及び目的言語文の間における単語対応付けを行う。また、単語対応付け部３２６は、図５に示すように、仮想単語挿入部３５０、単語対応推定部３５２、及び仮想単語除去部３５４を含む。

【００４４】

図５に示す仮想単語挿入部３５０は、対話文である原言語文及び目的言語文のペアの各々について、統語解析部３２５による当該原言語文の統語解析結果に基づいて、目的言語の統語要素であって、かつ原言語側に存在しないものに相当するものを仮想単語として挿入する。本実施形態においては、補完すべき仮想単語とその挿入位置の決定方法は限定されず、任意に規定してもよい。例えば、日本語の単語単位の依存構造解析結果を利用して英語の冠詞相当の仮想単語を挿入する場合であれば、名詞に係る単語のうち最も左側にある形容詞もしくは名詞の直前、当該形容詞が副詞によって更に修飾されている場合は更に当該副詞の直前に挿入する（非特許文献５：Daniel Flannery他，単語単位の日本語係り受け解析，言語処理学会第18回年次大会発表論文集，pp. 955-958，2012.）等の規則に基づいて決定すればよい。また、非特許文献２あるいは特許文献１に記載の仮想単語挿入方法を利用してもよい。

【００４５】

さらに、例えば、原言語が英語で、目的言語が日本語のように、目的言語側に、原言語の統語要素であって、かつ目的言語側に存在しないものがある場合には、原言語の統語要素であって、かつ目的言語側に存在しないものに相当するものを仮想単語として、目的言語側に挿入してもよい。この場合、第１言語が目的言語である日本語となり、第２言語が原言語である英語となる。

【００４６】

また、仮想単語として利用する文字列は、後述の仮想単語除去ステップでの除去処理を勘案し、原言語または目的言語の文に出現しない固有の文字列とすることが好適である。なお、仮想単語は、予め定義しておくものとし、原言語、又は目的言語を含む任意の言語において、任意の文字列を定義してもよい。

【００４７】

単語対応推定部３５２は、仮想単語挿入部３５０において取得した、対話文である仮想単語が挿入された原言語文及び目的言語文のペアの各々について、当該原言語文に含まれる単語の各々と、当該目的言語文に含まれる単語の各々について単語対応の推定を行い、仮単語対応付け結果を取得する。なお、本実施形態において用いる単語対応の推定方法は。統計的機械翻訳における公知の技術、例えば、ソフトウェアＧＩＺＡ＋＋等を利用する（非特許文献６：Peter F. Brown他，”The Mathematics of Statistical Machine Translation: Parameter Estimation,” Computational Linguistics, pp. 268-311, 1993.）。

【００４８】

仮想単語除去部３５４は、単語対応推定部３５２において取得した仮単語対応付け結果から、仮想単語と、仮想単語との単語対応付けとを除去することにより、単語対応付け結果を取得する。例えば、単語対応が０始まりの原言語、及び目的言語の単語ＩＤの組として与えられた様式で仮単語対応付け結果が取得されている場合には、仮想単語に相当する単語ＩＤを含む単語対応を示す単語ＩＤ組を削除し、仮想単語を除いた０始まりの単語ＩＤに書き換える処理をすればよい。この結果、仮想単語にのみ対応付けられた単語は相手

側言語に対応する単語のない、いわゆるＮＵＬＬ対応となる。

【００４９】

図４の翻訳モデル学習部３２８は、対訳文である原言語文及び目的言語文のペアの各々に対する、単語対応付け部３２６による単語対応付けの結果に基づき、原言語の語句が目的言語の語句に翻訳される確率を計算した翻訳モデルを学習する。モデルの学習は公知の技術、例えば非特許文献１の方法が利用可能である。また、非特許文献１に記載の方法を実装した統計的機械翻訳ソフトウェアＭｏｓｅｓでの対訳語句対応付けアルゴリズムによれば、単語対応が当該対訳語句外への単語対応がない「閉じている対訳語句」を対訳語句候補として抽出するため、ＮＵＬＬ対応の単語は隣接する対訳語句に連結される形の対訳語句として抽出され（図２及び図３の「は」や、図３における「the」）、ＮＵＬＬ対応の単語が後述の翻訳実行部において訳出されるか否かは、翻訳モデル、言語モデル、及びモデル重みに基づいてＮＵＬＬ対応の単語を含めたほうが、確率が高くなるか否かによって決定される。

【００５０】

目的言語文データベース３３０は、入力部３１０により受け付けた目的言語文の集合を記憶している。

【００５１】

言語モデル学習部３３２は、目的言語文データベース３３０に記憶されている目的言語文の集合に基づいて、目的言語の言語モデルを学習する。言語モデルの種類やその学習方法については特に規定しないが、公知の単語Ｎグラム言語モデルや、その種々の学習方法が利用可能である。

【００５２】

機械翻訳モデル出力部３３３は、翻訳モデル学習部３２８において取得された翻訳モデルを、翻訳モデル３３４に出力する。また、機械翻訳モデル出力部３３３は、言語モデル学習部３３２において取得された言語モデルを言語モデル３３６に出力する。

【００５３】

翻訳モデル３３４には、翻訳モデル学習部３２８によって学習された翻訳モデルが記憶されている。

【００５４】

言語モデル３３６には、言語モデル学習部３３２によって学習された言語モデルが記憶されている。

【００５５】

重み調整用対訳文データベース３３８は、入力部３１０により受け付けた、対訳文である原言語文と目的言語文とのペアの集合を記憶している。

【００５６】

重み調整部３４０は、目的言語文データベース３３０に記憶されている目的言語文の集合、翻訳モデル３３４に記憶されている翻訳モデル、及び言語モデル３３６に記憶されている翻訳モデルに基づいて、翻訳モデル及び言語モデルの各々に対する重みを調整する。

【００５７】

複数の統計モデルを利用して機械翻訳を行う場合、それぞれのモデルに適切な重みを設定することで翻訳精度の向上が期待できる。重みの調整には公知の技術、例えば、重み調整用の対訳文を利用して、重み調整用の原言語文を翻訳したときに得られる翻訳結果が、重み調整用の目的言語文に近づくように重みを更新する処理を繰り返し行う方法（非特許文献７：Franz Josef Och, ”Minimum Error Rate Training in Statistical Machine Translation,” Proc. ACL, pp. 160-167, 2003.）が利用可能である。

【００５８】

モデル重み記憶部３４２は、重み調整部３４０によって調整された翻訳モデル及び言語モデルの各々に対する重みを記憶している。

【００５９】

出力部３９０は、翻訳モデル３３４に記憶されている翻訳モデル、及び言語モデル３３

６に記憶されている翻訳モデル、モデル重み記憶部３４２に記憶されている重みを出力する。

【００６０】

＜本発明の実施形態に係る機械翻訳装置の構成＞

次に、本発明の実施の形態に係る機械翻訳装置の構成について説明する。図６に示すように、本発明の実施の形態に係る機械翻訳装置４００は、ＣＰＵと、ＲＡＭと、後述する機械翻訳処理ルーチンを実行するためのプログラムや各種データを記憶したＲＯＭと、を含むコンピュータで構成することができる。この機械翻訳装置４００は、機能的には図６に示すように入力部４１０と、演算部４２０と、出力部４９０とを備えている。

【００６１】

入力部４１０は、翻訳対象となる原言語文の入力を受け付ける。

【００６２】

演算部４２０は、翻訳モデル４２２、言語モデル４２４、モデル重み記憶部４２６、及び翻訳実行部４２８を備えている。

【００６３】

翻訳モデル４２２には、機械翻訳学習装置３００の翻訳モデル３３４と同一の翻訳モデルが記憶されている。

【００６４】

言語モデル４２４には、機械翻訳学習装置３００の言語モデル３３６と同一の言語モデルが記憶されている。

【００６５】

モデル重み記憶部４２６は、機械翻訳学習装置３００のモデル重み記憶部３４２と同一の、翻訳モデル及び言語モデルの各々に対する重みを記憶している。

【００６６】

翻訳実行部４２８は、翻訳モデル４２２に記憶されている翻訳モデル、言語モデル４２４に記憶されている言語モデル、及びモデル重み記憶部４２６に記憶されている重みに基づいて、入力部４１０で受け付けた原言語文を目的言語文へ翻訳する翻訳処理を実行する。翻訳の方法は公知の技術、例えば非特許文献６の技術が利用可能である。

【００６７】

翻訳結果は、出力部４９０を介して、端末または記憶媒体に出力する。

【００６８】

＜本発明の実施形態に係る機械翻訳学習装置の作用＞

次に、本発明の実施の形態に係る機械翻訳学習装置３００の作用について説明する。まず、入力部３１０により、対訳文である原言語文と目的言語文とのペアの集合の入力を受け付け、原言語文の集合が、対訳原言語文データベース３２２に記憶され、目的言語文の集合が、対訳目的言語文データベース３２４に記憶される。

【００６９】

また、入力部３１０により、目的言語文の集合を受け付け、目的言語文データベース３３０に記憶される。また、入力部３１０により、モデルの重み調整のための学習データとして、対訳文である原言語文と目的言語文とのペアの集合の入力を受け付け、重み調整用対訳文データベース３３８に記憶される。

【００７０】

そして、機械翻訳学習装置３００のＲＯＭに記憶されたプログラムを、ＣＰＵが実行することにより、図７に示す機械翻訳学習処理ルーチンが実行される。

【００７１】

まず、ステップＳ３００では、対訳原言語文データベース３２２及び対訳目的言語文データベース３２４に記憶されている、対訳文である原言語文と目的言語文とのペアの集合を読み込む。

【００７２】

次に、ステップＳ３０２では、対訳文のペアの集合に含まれる対訳文のペアの各々につ

いて、仮想単語の挿入、仮単語対応付けの結果の取得、及び仮想単語の除去の処理を行うことにより単語の対応付けを行う。

【００７３】

そして、ステップＳ３０４では、上記ステップＳ３０２による単語の対応付け結果に基づいて、翻訳モデルを学習し、翻訳モデル３３４に記憶して、出力部３９０により出力する。

【００７４】

ステップＳ３０６では、目的言語文データベース３３０に記憶されている目的言語文の集合を読み込む。

【００７５】

そして、ステップＳ３０８では、上記ステップＳ３０６で読み込んだ目的言語文の集合に基づいて、言語モデルを学習し、言語モデル３３６に記憶して、出力部３９０により出力する。

【００７６】

ステップＳ３１０では、重み調整用対訳文データベース３３８に記憶されている対訳文のペアの集合を読み込む。

【００７７】

そして、ステップＳ３１２では、上記ステップＳ３１０で読み込んだ対訳文のペアの集合、翻訳モデル３３４に記憶されている翻訳モデル、及び言語モデル３３６に記憶されている言語モデルに基づいて、各モデルの重みを調整し、モデル重み記憶部３４２に記憶して、出力部３９０により出力し、機械翻訳学習処理ルーチンを終了する。

【００７８】

＜本発明の実施形態に係る機械翻訳装置の作用＞

次に、本発明の実施の形態に係る機械翻訳装置４００の作用について説明する。まず、入力部４１０により、機械翻訳対象の原言語文を受け付けると、機械翻訳装置４００のＲＯＭに記憶されたプログラムを、ＣＰＵが実行することにより、図８に示す機械翻訳処理ルーチンが実行される。

【００７９】

まず、ステップＳ４００では、翻訳モデル４２２に記憶されている翻訳モデル、言語モデル４２４に記憶されている言語モデル、及びモデル重み記憶部４２６に記憶されている各モデルの重みを読み込む。

【００８０】

次に、ステップＳ４０２では、ステップＳ４００において取得した翻訳モデル、言語モデル、及び各モデルの重みに基づいて、入力部４１０において受け付けた原言語文を目的言語文へ翻訳する処理を実行して、翻訳結果を、出力部４９０により出力して、機械翻訳処理ルーチンを終了する。

【００８１】

＜実験例＞

本実施形態に係る機械翻訳学習装置を利用した機械翻訳システムは日本語から英語への翻訳において翻訳評価尺度の一つであるTER（翻訳誤り率）を、Ｍｏｓｅｓを利用した一般的な統計的機械翻訳システムと比較して59.08から58.31に改善することができた。

【００８２】

以上説明したように、本発明の実施形態に係る機械翻訳学習装置によれば、対訳となる第１言語文及び第２言語文のペアに含まれる第１言語文について、仮想単語を、第１言語文に挿入し、仮想単語を挿入した第１言語文に含まれる単語の各々と、第２言語文に含まれる単語の各々との単語の対応関係を推定し、推定された単語の対応関係に基づいて、仮想単語と対応付けられた単語の対応関係を除去し、かつ単語の対応関係に含まれる仮想単語を除去した結果を単語対応付け結果とすることにより対訳関係にある語の自動対応付けを精度良く行うことができる。

【００８３】

また、冠詞の存在しない原言語から冠詞が必要な目的言語への翻訳における、原言語での冠詞の不存在に起因する誤った対訳語句対の獲得を抑制することができる。そのために、原言語の統語解析結果に基づいて原言語の文に冠詞相当の仮想単語を補完した上で対訳関係にある語の自動対応付けを行い、その結果から仮想単語と仮想単語に関わる単語対応付けを除去することによって、目的言語の冠詞が原言語の単語と誤って対応付けられ、その結果として本来冠詞を含んで翻訳すべきでない語句の翻訳において冠詞を伴った翻訳しか許容されないような対訳語句対集合しか獲得できなくなることを防ぐことができる。

【００８４】

また、原言語に対応するもののない目的言語の統語要素が存在するような統計的機械翻訳の学習において誤った対訳語句の対応付けを抑制でき、より精度の高い機械翻訳が可能となる。

【００８５】

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。

【００８６】

例えば、本実施形態においては、機械翻訳学習装置と、機械翻訳装置とを別々の装置として構成する場合について説明したが、これに限定されるものではない。例えば、機械翻訳学習装置と、機械翻訳装置とを１つの装置として構成してもよい。

【００８７】

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し、ネットワークを介して提供することも可能である。

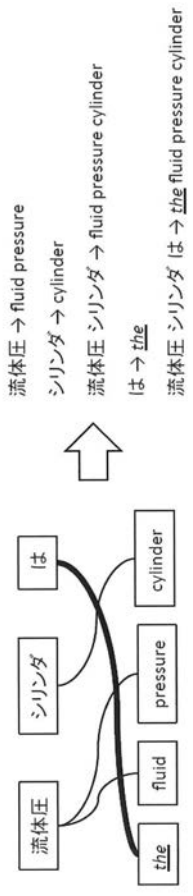
【符号の説明】

【００８８】

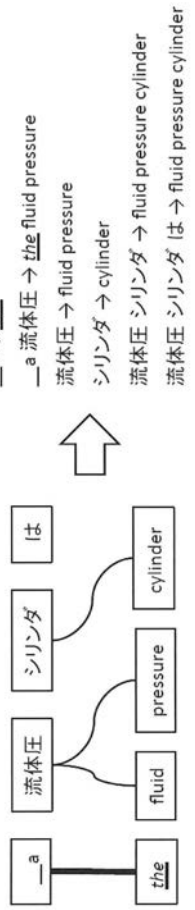
３００ 機械翻訳学習装置
３１０ 入力部
３２０ 演算部
３２２ 対訳原言語文データベース
３２４ 対訳目的言語文データベース
３２５ 統語解析部
３２６ 単語対応付け部
３２８ 翻訳モデル学習部
３３０ 目的言語文データベース
３３２ 言語モデル学習部
３３３ 機械翻訳モデル出力部
３３４ 翻訳モデル
３３６ 言語モデル
３３８ 調整用対訳文データベース
３４０ 重み調整部
３４２ モデル重み記憶部
３５０ 仮想単語挿入部
３５２ 単語対応推定部
３５４ 仮想単語除去部
３９０ 出力部
４００ 機械翻訳装置
４１０ 入力部
４２０ 演算部
４２２ 翻訳モデル
４２４ 言語モデル
４２６ モデル重み記憶部

４２８　翻訳実行部
４９０　出力部

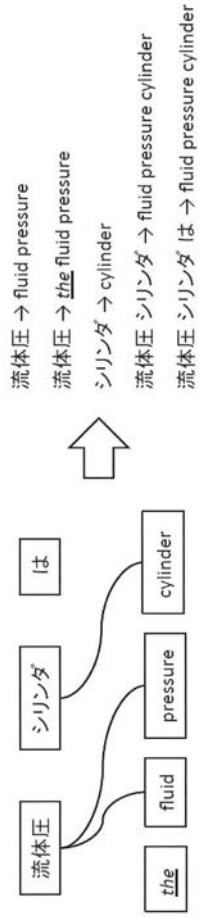
【図１】



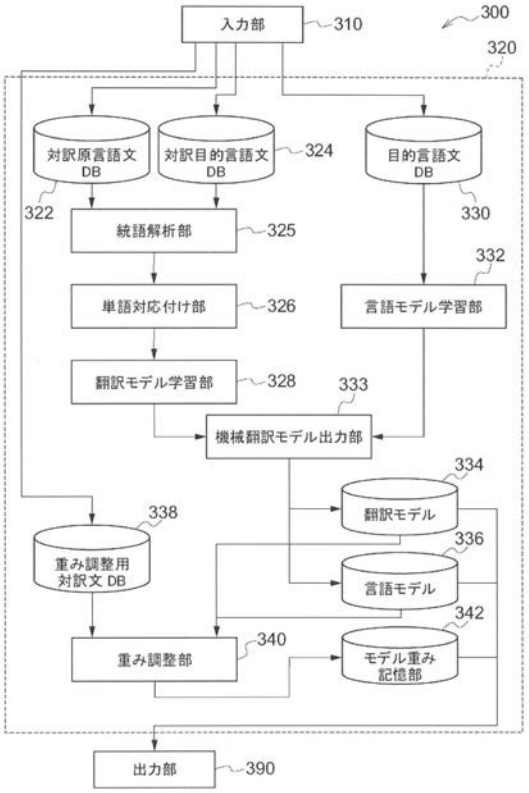
【図２】



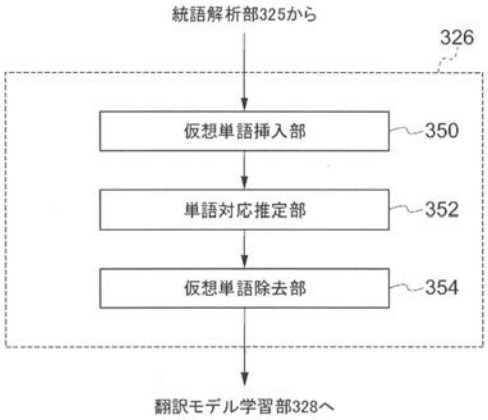
【図３】


流体圧 → fluid pressure
流体圧 → the fluid pressure
シリンダ → cylinder
流体圧 シリンダ → fluid pressure cylinder
流体圧 シリンダ は → fluid pressure cylinder
は
シリンダ
流体圧
cylinder
pressure
fluid
the

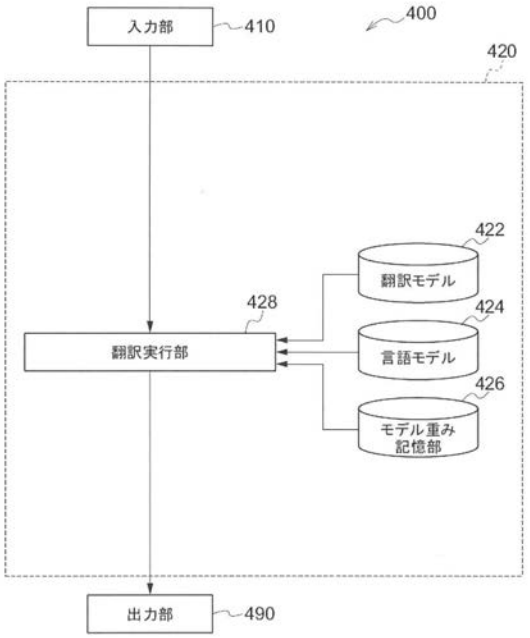
【図４】


入力部
310
300
320
対訳原言語文 DB
対訳目的言語文 DB
324
目的言語文 DB
322
330
統語解析部
325
332
単語対応付け部
326
言語モデル学習部
翻訳モデル学習部
328
333
機械翻訳モデル出力部
334
翻訳モデル
338
336
重み調整用 対訳文 DB
言語モデル
342
重み調整部
340
モデル重み 記憶部
出力部
390

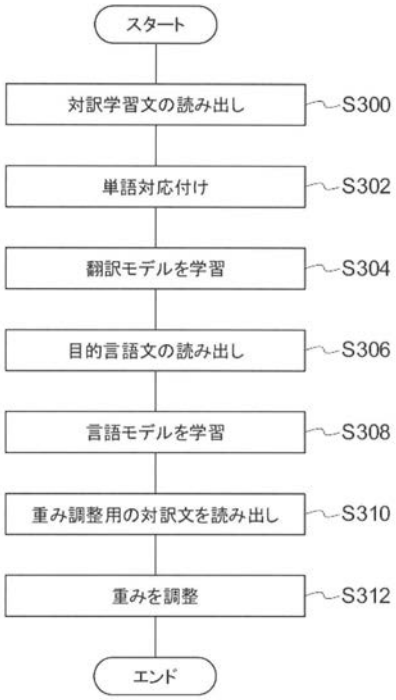
【図５】


統語解析部325から
326
仮想単語挿入部
350
単語対応推定部
352
仮想単語除去部
354
翻訳モデル学習部328へ

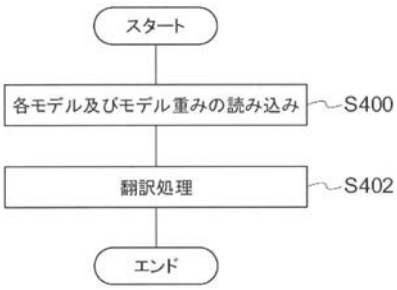
【図６】


入力部
410
400
420
422
翻訳モデル
428
424
翻訳実行部
言語モデル
426
モデル重み 記憶部
出力部
490

【図７】


スタート
対訳学習文の読み出し
S300
単語対応付け
S302
翻訳モデルを学習
S304
目的言語文の読み出し
S306
言語モデルを学習
S308
重み調整用の対訳文を読み出し
S310
重みを調整
S312
エンド

【図８】


スタート
各モデル及びモデル重みの読み込み
S400
翻訳処理
S402
エンド

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
Ｆターム(参考) 5B091 AA03　BA02　EA01