

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第6558696号  
(P6558696)

(45) 発行日 令和1年8月14日(2019.8.14)

(24) 登録日 令和1年7月26日(2019.7.26)

(51) Int. Cl.		F I			
<b>G06F 17/28</b>	<b>(2006.01)</b>	G06F 17/28	627		
<b>G06N 20/00</b>	<b>(2019.01)</b>	G06F 17/28	618		
		G06N 20/00	130		

請求項の数 8 (全 15 頁)

(21) 出願番号	特願2015-214659 (P2015-214659)	(73) 特許権者	000004226
(22) 出願日	平成27年10月30日(2015.10.30)		日本電信電話株式会社
(65) 公開番号	特開2017-84274 (P2017-84274A)		東京都千代田区大手町一丁目5番1号
(43) 公開日	平成29年5月18日(2017.5.18)	(73) 特許権者	504132272
審査請求日	平成29年12月20日(2017.12.20)		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) 【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

を含み、

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を、前記英  
語の冠詞とし、

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、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】

請求項 1 又は 2 記載の単語対応付け装置と、  
前記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。

10

【請求項 4】

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ことを含み、

20

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が挿入することでは、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  
単語対応付け方法。

【請求項 5】

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 4 記載の単語対応付け方法。

30

【請求項 6】

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、  
請求項 4 又は請求項 5 記載の単語対応付け方法によって単語対応付け結果を取得し、  
前記翻訳モデル学習部は、前記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。

40

【請求項 7】

コンピュータを、請求項 1 又は 2 記載の単語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。

【請求項 8】

コンピュータを、請求項 3 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。

50

## 【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に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。

10

## 【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," Proceedings 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.

20

30

【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。

40

## 【0008】

また、上述した非特許文献3に記載の従来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に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。

50

**【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。

10

**【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。

20

**【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。

30

**【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。

40

**【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。

50

## 【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

10

20

30

40

50

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制し、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】

10

20

30

40

50

< 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。

10

【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。

20

【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。

30

【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。

40

【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

50

び仮想単語除去部 3 5 4 を含む。

【 0 0 4 4 】

図 5 に示す仮想単語挿入部 3 5 0 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 3 2 5 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。

10

【 0 0 4 5 】

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。

【 0 0 4 6 】

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。

20

【 0 0 4 7 】

単語対応推定部 3 5 2 は、仮想単語挿入部 3 5 0 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 G I Z A + + 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.）

30

【 0 0 4 8 】

仮想単語除去部 3 5 4 は、単語対応推定部 3 5 2 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 I D 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 I D を含む単語対応を示す単語 I D 組を削除し、仮想単語を除いた 0 始まりの単語 I D 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 N U L L 対応となる。

40

【 0 0 4 9 】

図 4 の翻訳モデル学習部 3 2 8 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 3 2 6 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 M o s e s 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、N U L L 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」）、N U L L 対応

50



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N U L L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。

【 0 0 5 0 】

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。

【 0 0 5 1 】

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。

10

【 0 0 5 2 】

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。

【 0 0 5 3 】

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。

【 0 0 5 4 】

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。

20

【 0 0 5 5 】

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。

【 0 0 5 6 】

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。

【 0 0 5 7 】

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。

30

【 0 0 5 8 】

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。

【 0 0 5 9 】

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336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。

40

【 0 0 6 0 】

< 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。

【 0 0 6 1 】

50

入力部 4 1 0 は、翻訳対象となる原言語文の入力を受け付ける。

【 0 0 6 2 】

演算部 4 2 0 は、翻訳モデル 4 2 2、言語モデル 4 2 4、モデル重み記憶部 4 2 6、及び翻訳実行部 4 2 8 を備えている。

【 0 0 6 3 】

翻訳モデル 4 2 2 には、機械翻訳学習装置 3 0 0 の翻訳モデル 3 3 4 と同一の翻訳モデルが記憶されている。

【 0 0 6 4 】

言語モデル 4 2 4 には、機械翻訳学習装置 3 0 0 の言語モデル 3 3 6 と同一の言語モデルが記憶されている。

10

【 0 0 6 5 】

モデル重み記憶部 4 2 6 は、機械翻訳学習装置 3 0 0 のモデル重み記憶部 3 4 2 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。

【 0 0 6 6 】

翻訳実行部 4 2 8 は、翻訳モデル 4 2 2 に記憶されている翻訳モデル、言語モデル 4 2 4 に記憶されている言語モデル、及びモデル重み記憶部 4 2 6 に記憶されている重みに基づいて、入力部 4 1 0 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。

【 0 0 6 7 】

翻訳結果は、出力部 4 9 0 を介して、端末または記憶媒体に出力する。

20

【 0 0 6 8 】

< 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 3 0 0 の作用について説明する。まず、入力部 3 1 0 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 3 2 2 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 3 2 4 に記憶される。

【 0 0 6 9 】

また、入力部 3 1 0 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 3 3 0 に記憶される。また、入力部 3 1 0 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 3 3 8 に記憶される。

30

【 0 0 7 0 】

そして、機械翻訳学習装置 3 0 0 の R O M に記憶されたプログラムを、C P U 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。

【 0 0 7 1 】

まず、ステップ S 3 0 0 では、対訳原言語文データベース 3 2 2 及び対訳目的言語文データベース 3 2 4 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。

【 0 0 7 2 】

次に、ステップ S 3 0 2 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。

40

【 0 0 7 3 】

そして、ステップ S 3 0 4 では、上記ステップ S 3 0 2 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 3 3 4 に記憶して、出力部 3 9 0 により出力する。

【 0 0 7 4 】

ステップ S 3 0 6 では、目的言語文データベース 3 3 0 に記憶されている目的言語文の集合を読み込む。

【 0 0 7 5 】

50

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。

10

【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。

20

【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。

30

【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】

40

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が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

50

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。

10

【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

20

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

30

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

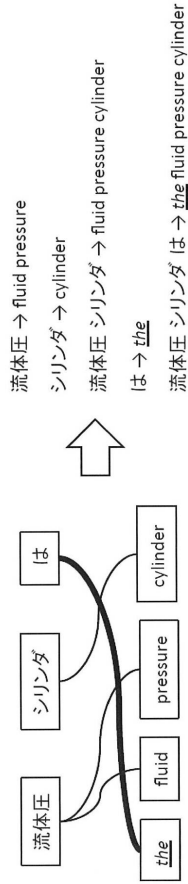
426 モデル重み記憶部

40

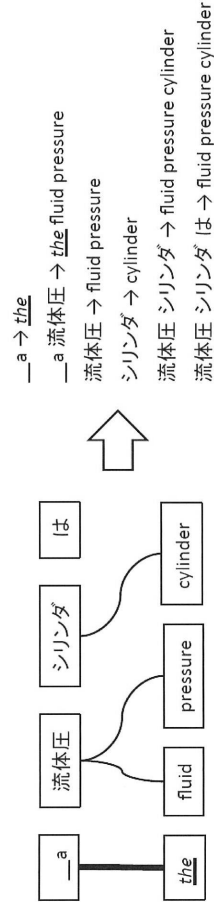
428 翻訳実行部

490 出力部

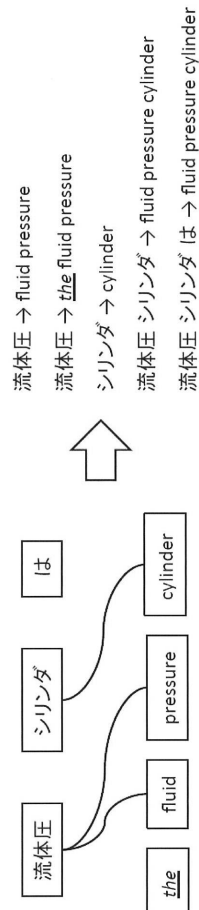
【 図 1 】



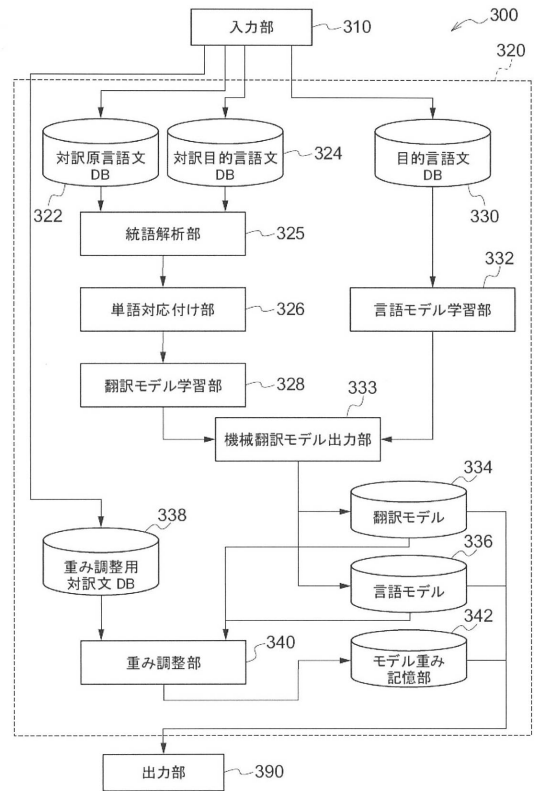
【 図 2 】



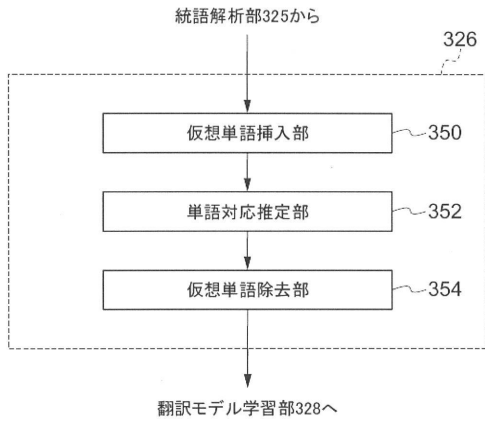
【 図 3 】



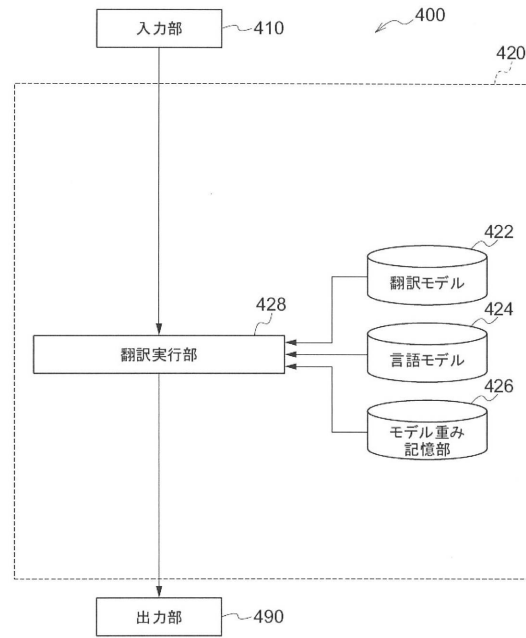
【 図 4 】



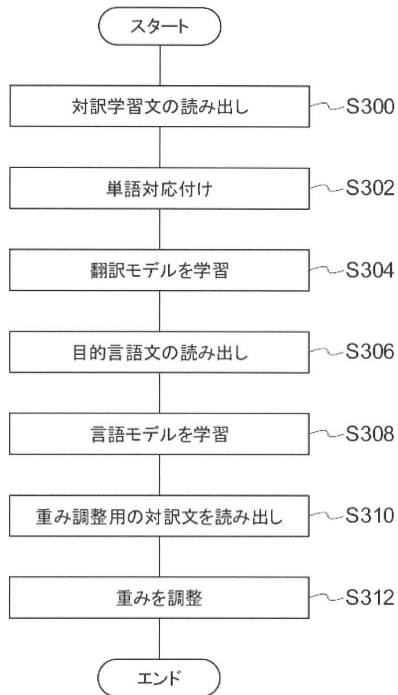
【図5】



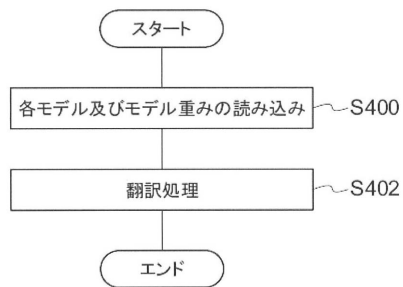
【図6】



【図7】



【図8】



---

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3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2010-198438(JP,A)

特開2013-218524(JP,A)

特開2009-003642(JP,A)

Phillip Koehn他2名, Statistical Phrase-Based Translation, proceedings of HLT-NAACL 2003[online], 2003年, (平成30年11月2日検索), インターネット<URL: <http://www.aclweb.org/anthology/N03-1017>>, URL, <http://www.aclweb.org/anthology/N03-1017>

(58)調査した分野(Int.Cl., DB名)

G06F 17/20 - 17/28

G06N 20/00