

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第6265923号
(P6265923)

(45) 発行日 平成30年1月24日 (2018. 1. 24)

(24) 登録日 平成30年1月5日 (2018. 1. 5)

(51) Int. Cl. F 1
G 0 6 F 17/28 (2006.01)
 G 0 6 F 17/28 6 1 8
 G 0 6 F 17/28 6 2 7

請求項の数 5 (全 20 頁)

(21) 出願番号	特願2015-7131 (P2015-7131)	(73) 特許権者	000004226
(22) 出願日	平成27年1月16日 (2015. 1. 16)		日本電信電話株式会社
(65) 公開番号	特開2016-133920 (P2016-133920A)		東京都千代田区大手町一丁目5番1号
(43) 公開日	平成28年7月25日 (2016. 7. 25)	(74) 代理人	110001519
審査請求日	平成28年12月9日 (2016. 12. 9)		特許業務法人太陽国際特許事務所
		(72) 発明者	貞光 九月
			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	松尾 義博
			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	浅野 久子
			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】 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】

入力された、翻訳元言語文と、前記翻訳元言語文に対応する論理形式文である翻訳先言語文とのペアの集合に基づいて、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について、前記翻訳元言語文の単語と前記翻訳先言語文の単語との対応付けである単語アラインメントを推定する単語アラインメント推定部と、

前記単語アラインメント推定部により推定された前記単語アラインメントに基づいて、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について、前記ペアの前記単語アラインメントにおいて、前記翻訳元言語文の1つの単語が、前記翻訳先言語文の複数の部分文字列に対応し、かつ、前記複数の部分文字列が同一又は類似している場合、前記1つの単語を複製した前記翻訳元言語文の単語と前記翻訳先言語文の単語との対応付けである複製済単語アラインメントを生成すると共に、前記翻訳元言語文の単語の各々について複製された単語数に基づいて、前記翻訳元言語文の単語の複製数を推定するための単語複製モデルを学習する単語複製部と、

前記翻訳元言語文及び前記翻訳先言語文のペアの各々について前記単語複製部により生成された前記複製済単語アラインメントに基づいて、前記翻訳元言語文を前記翻訳先言語文に翻訳するための翻訳モデルを学習する翻訳モデル学習部と、

を含む翻訳学習装置。

【請求項 2】

前記翻訳元言語文及び前記翻訳先言語文のペアの各々について、前記翻訳元言語文の部

分文字列と前記翻訳先言語文の部分文字列との対応付けを表す予め定められた部分文字列アラインメントに基づいて、前記ペアについての前記複製済単語アラインメントを補正し、前記ペアの各々について補正した前記複製済単語アラインメントに基づいて、前記翻訳元言語文のフレーズと前記翻訳先言語文のフレーズとのペアを格納したフレーズテーブルを推定する半教師ありフレーズ推定部を更に含み、

前記翻訳モデル学習部は、前記半教師ありフレーズ推定部により生成された前記フレーズテーブルに基づいて、前記翻訳モデルを学習する請求項1に記載の翻訳学習装置。

【請求項3】

フレーズフィルタリング部と、パラフレーズ適用部とを更に含み、

前記フレーズフィルタリング部は、前記半教師ありフレーズ推定部により生成されたフレーズテーブルから、前記翻訳先言語文について予め定められた制約を満たさないフレーズを含むペアを削除し、

前記翻訳モデル学習部は、前記フレーズフィルタリング部によりフレーズが削除された前記フレーズテーブルに基づいて、前記翻訳モデルを学習し、

前記パラフレーズ適用部は、前記翻訳モデル学習部により学習された前記翻訳モデルに含まれる各フレーズについて、言い換え可能なフレーズを格納したパラフレーズ辞書を参照して、言い換え可能なフレーズに拡張した前記翻訳モデルを、パラフレーズ済翻訳モデルとして生成する請求項2に記載の翻訳学習装置。

【請求項4】

単語アラインメント推定部と、単語複製部と、翻訳モデル学習部とを含む翻訳学習装置における翻訳学習方法であって、

前記単語アラインメント推定部が、入力された、翻訳元言語文と、前記翻訳元言語文に対応する論理形式文である翻訳先言語文とのペアの集合に基づいて、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について、前記翻訳元言語文の単語と前記翻訳先言語文の単語との対応付けである単語アラインメントを推定するステップと、

前記単語複製部が、前記単語アラインメント推定部により推定された前記単語アラインメントに基づいて、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について、前記ペアの前記単語アラインメントにおいて、前記翻訳元言語文の1つの単語が、前記翻訳先言語文の複数の部分文字列に対応し、かつ、前記複数の部分文字列が同一又は類似している場合、前記1つの単語を複製した前記翻訳元言語文の単語と前記翻訳先言語文の単語との対応付けである複製済単語アラインメントを生成すると共に、前記翻訳元言語文の単語の各々について複製された単語数に基づいて、前記翻訳元言語文の単語の複製数を推定するための単語複製モデルを学習するステップと、

前記翻訳モデル学習部が、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について前記単語複製部により生成された前記複製済単語アラインメントに基づいて、前記翻訳元言語文を前記翻訳先言語文に翻訳するための翻訳モデルを学習するステップと、

を含む翻訳学習方法。

【請求項5】

コンピュータを、請求項1～請求項3のいずれか1項に記載の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、翻訳学習装置、方法、及びプログラムに係り、特に、翻訳元言語文を翻訳先言語文に翻訳するための翻訳学習装置、方法、及びプログラムに関する。

【背景技術】

【0002】

従来より、構文解析の手法として、自然言語文に対して構造を付与したものや、文脈自由文法（CFG：Context free Grammar）や依存文法が用いられている。文脈自由文法では、{}が1つの部分木を表し、その部分木に対応するラベルを（）として記す。例えば、

10

20

30

40

50

【0003】

{I(NP) { like(V) it(NP)}(VP) }(S)

【0004】

という形式のものである。

【0005】

また、文脈自由文法の規則の適用例は、

【0006】

S->{NP, VP}, NP->{ “ I ” }, VP->{V, NP}, V->{ “ like ” }, N->{ “ it ” }

【0007】

という形式となる。ここで、NP (noun phrase) は名詞フレーズ、VP (verb phrase) は動
詞フレーズ、N (noun) は名詞、V (verb) は動詞、S (string) は文字列を表す。 10

【0008】

また、同期文脈自由文法 (SCFG: Synchronous Context free Grammar) に基づいて翻訳
元言語文 (自然言語文) と翻訳先言語文 (論理形式文) のフレーズアラインメントを獲得
する手法が知られている (非特許文献 1 参照)。

【先行技術文献】

【非特許文献】

【0009】

【非特許文献 1】Y.Wong, R.Mooney, “ Learning for Semantic Parsing with Statistical
Machine Translation ”, The Annual Meeting of the North American Chapter of the
ACL (NAACL2006) 20

【非特許文献 2】Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch
, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran,
Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, “ Mos
es: Open Source Toolkit for Statistical Machine Translation ”, Annual Meeting of
the Association for Computational Linguistics (ACL), demonstration session, Pra
gue, Czech Republic, June 2007.

【発明の概要】

【発明が解決しようとする課題】

【0010】

しかし、第 1 の課題として、非特許文献 1 に記載の論理形式文には形式的な単語 (上記
例の場合、カッコや.,?等) が多く、翻訳の結果生成された S P A R Q L 文 (論理形式文
の一種) が、S P A R Q L 文の体裁をなさず、破たんする可能性が高いという問題があっ
た。このような問題を非特許文献 1 では、別途獲得した論理形式文への変換ルールを併用
することでこの問題を回避しているが、統計的機械翻訳 (以下、S M T と記載する) の本
来得意とする表現の揺れ、曖昧性の解消といった利点を十全に発揮できていない。例えば
、元言語文「価格は？」→論理形式文「 “ ?p=価格 ?o) ” 」という対応があった場合、出
力時にカッコの数が合わなくなる可能性がある。 30

【0011】

また、第 2 の課題として、最も広く使われている S M T ツールは非特許文献 2 に記載の
同期文脈自由文法であるが、t r e e b a s e d S M T を用いる場合であっても、同期
文脈自由文法を基本として設計されている場合、変数の数が自然言語文と論理形式文間で
異なる場合、翻訳が困難となるという問題があった。 40

【0012】

例えば、論理形式文側で単語の意味が一つある場合、「AAAの価格は？」という自然言
語文は、「SELECT ?o; WHERE {?s = AAA, ?p = prop-ja:価格, ?o }」という論理形式文
となる。一方、論理形式文で単語の意味が二つある場合、AAAの価格は？」という自然言
語文は、「SELECT ?o; WHERE {?s = AAA, ?p = prop-ja:価格, ?o ; ?p=prop-ja-nl:価格
?o. }」という論理形式文となる。このように、自然言語文側で 1 度しか出現しない価格
が、論理形式文側では 2 回出現する。一般の S M T 手法では、これらの現象を明に扱わな 50

いため、翻訳が困難となる。

【0013】

また、第3の課題として、自然言語文と論理形式文とのペアを多く入手することは困難なため、それらの学習データから得られる翻訳モデル中のフレーズペアは疎になってしまい、未知の文が入力された場合、処理できない可能性が高いという問題があった。例えば、「AAAの価格はいくら?」についての学習データがあり、それを用いて学習した翻訳モデルを用いれば、「AAAの価格はいくら?」という自然言語文から論理形式文の翻訳を行うことができる。一方、「AAAの値段はいくら?」という自然言語文では、「値段」という単語が未知のため、翻訳できない。

【0014】

また、第4の課題として、自然言語文と論理形式文とのペアの全ての部分文字列に対してアラインメントがあれば、翻訳モデル学習は極めて容易であるが、そのようなアノテーションには大きなコストがかかってしまうという問題がある。一方、共通して用いられる部分的なアラインメントを作ることは比較的成本を低く抑えて作成することが可能である。しかし、このように部分的なアラインメントだけが判明している場合、残る部分文字列のアラインメントを推定したり、翻訳モデルの学習を行う方法は自明ではない。

【0015】

本発明は、上記問題点を解決するために成されたものであり、精度良く、翻訳元言語文を翻訳先言語文に翻訳するための翻訳モデルを学習することができる翻訳学習装置、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】

【0016】

上記目的を達成するために、第1の発明に係る翻訳学習装置は、入力された、翻訳元言語文と、前記翻訳元言語文に対応する翻訳先言語文とのペアの集合に基づいて、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について、前記翻訳元言語文の単語と前記翻訳先言語文の単語との対応付けである単語アラインメントを推定する単語アラインメント推定部と、前記単語アラインメント推定部により推定された前記単語アラインメントに基づいて、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について、前記ペアの前記単語アラインメントにおいて、前記翻訳元言語文の1つの単語が、前記翻訳先言語文の複数の部分文字列に対応し、かつ、前記複数の部分文字列が同一又は類似している場合、前記1つの単語を複製した前記翻訳元言語文の単語と前記翻訳先言語文の単語との対応付けである複製済単語アラインメントを生成すると共に、前記翻訳元言語文の単語の各々について複製された単語数に基づいて、前記翻訳元言語文の単語の複製数を推定するための単語複製モデルを学習する単語複製部と、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について前記単語複製部により生成された前記複製済単語アラインメントに基づいて、前記翻訳元言語文を前記翻訳先言語文に翻訳するための翻訳モデルを学習する翻訳モデル学習部と、を含んで構成されている。

【0017】

また、第1の発明に係る翻訳学習装置において、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について、前記翻訳元言語文の部分文字列と前記翻訳先言語文の部分文字列との対応付けを表す予め定められた部分文字列アラインメントに基づいて、前記ペアについての前記複製済単語アラインメントを補正し、前記ペアの各々について補正した前記複製済単語アラインメントに基づいて、前記翻訳元言語文のフレーズと前記翻訳先言語文のフレーズとのペアを格納したフレーズテーブルを推定する半教師ありフレーズ推定部を更に含み、前記翻訳モデル学習部は、前記半教師ありフレーズ推定部により推定された前記フレーズテーブルに基づいて、前記翻訳モデルを学習してもよい。

【0018】

また、第1の発明に係る翻訳学習装置において、フレーズフィルタリング部と、パラフレーズ適用部とを更に含み、前記フレーズフィルタリング部は、前記半教師ありフレーズ推定部により生成されたフレーズテーブルから、前記翻訳先言語文について予め定め

10

20

30

40

50

られた制約を満たさないフレーズを含むペアを削除し、前記翻訳モデル学習部は、前記フレーズフィルタリング部によりフレーズが削除された前記フレーズテーブルに基づいて、前記翻訳モデルを学習し、前記パラフレーズ適用部は、前記翻訳モデル学習部により学習された前記翻訳モデルに含まれる各フレーズについて、言い換え可能なフレーズを格納したパラフレーズ辞書を参照して、言い換え可能なフレーズに拡張した前記翻訳モデルを、パラフレーズ済翻訳モデルとして生成してもよい。

【0019】

第2の発明に係る翻訳学習装置は、入力された、翻訳元言語文と、前記翻訳元言語文に対応する翻訳先言語文とのペアの集合に基づいて、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について、前記翻訳元言語文の単語と前記翻訳先言語文の単語との対応付けである単語アラインメントを推定する単語アラインメント推定部と、前記翻訳元言語文の部分文字列と前記翻訳先言語文の部分文字列との対応付けを表す予め定められた部分文字列アラインメントに基づいて、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について、前記ペアについての前記単語アラインメントを補正し、前記ペアの各々について補正した前記単語アラインメントに基づいて、前記翻訳元言語文のフレーズと前記翻訳先言語文のフレーズとのペアを格納したフレーズテーブルを推定する半教師ありフレーズ推定部と、前記翻訳元言語文及び前記翻訳先言語文のペアの集合と、前記半教師ありフレーズ推定部により推定された前記フレーズテーブルとに基づいて、前記翻訳元言語文を前記翻訳先言語文に翻訳するための翻訳モデルを学習する翻訳モデル学習部と、を含んで構成されている。

【0020】

第3の発明に係る翻訳学習装置によれば、推定された単語アラインメントに基づいて、翻訳元言語文及び翻訳先言語文のペアの各々について、ペアの単語アラインメントにおいて、翻訳元言語文の1つの単語が、翻訳先言語文の複数の部分文字列に対応し、かつ、複数の部分文字列が同一又は類似している場合、1つの単語を複製した翻訳元言語文の単語と翻訳先言語文の単語との対応付けである単語アラインメントである複製済単語アラインメントを生成すると共に、翻訳元言語文の単語の各々について複製された単語数に基づいて、翻訳元言語文の単語の複製数を推定するための単語複製モデルを学習し、翻訳元言語文及び翻訳先言語文のペアの各々について生成された複製済単語アラインメントに基づいて、翻訳元言語文を翻訳先言語文に翻訳するための翻訳モデルを学習することにより、精度良く、翻訳元言語文を翻訳先言語文に翻訳するための翻訳モデルを学習することができる。

【0021】

第1の発明に係る翻訳学習方法は、単語アラインメント推定部と、半教師ありフレーズ推定部と、翻訳モデル学習部とを含む翻訳学習装置における翻訳学習方法であって、単語アラインメント推定部が、入力された、翻訳元言語文と、前記翻訳元言語文に対応する翻訳先言語文とのペアの集合に基づいて、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について、前記翻訳元言語文の単語と前記翻訳先言語文の単語との対応付けである単語アラインメントを推定するステップと、前記半教師ありフレーズ推定部が、前記翻訳元言語文の単語と前記翻訳先言語文の単語との一部分の対応付けである部分文字列アラインメントを示す予め定められたアノテートデータに基づいて、前記単語アラインメントの全体のアラインメントを補正したフレーズテーブルを推定するステップと、前記翻訳モデル学習部が、前記翻訳元言語文及び前記翻訳先言語文のペアの集合と、前記半教師ありフレーズ推定部により推定された前記フレーズテーブルとに基づいて、前記翻訳元言語文を前記翻訳先言語文に翻訳するための翻訳モデルを学習するステップと、を含んで実行することを特徴とする。

【0022】

第2の発明に係る翻訳学習方法は、単語アラインメント推定部と、単語複製部と、翻訳モデル学習部とを含む翻訳学習装置における翻訳学習方法であって、前記単語アラインメント推定部が、入力された、翻訳元言語文と、前記翻訳元言語文に対応する翻訳先言語文

とのペアの集合に基づいて、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について、前記翻訳元言語文の単語と前記翻訳先言語文の単語との対応付けである単語アラインメントを推定するステップと、前記単語複製部が、前記単語アラインメント推定部により推定された前記単語アラインメントに基づいて、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について、前記ペアの前記単語アラインメントにおいて、前記翻訳元言語文の1つの単語が、前記翻訳先言語文の複数の部分文字列に対応し、かつ、前記複数の部分文字列が同一又は類似している場合、前記1つの単語を複製した前記翻訳元言語文の単語と前記翻訳先言語文の単語との対応付けである複製済単語アラインメントを生成すると共に、前記翻訳元言語文の単語の各々について複製された単語数に基づいて、前記翻訳元言語文の単語の複製数を推定するための単語複製モデルを学習するステップと、前記翻訳モデル学習部が、前記翻訳元言語文及び前記翻訳先言語文のペアの各々について前記単語複製部により生成された前記複製済単語アラインメントに基づいて、前記翻訳元言語文を前記翻訳先言語文に翻訳するための翻訳モデルを学習するステップと、を含んで実行することを特徴とする。

10

【0023】

第4の発明に係るプログラムは、コンピュータを、第1又は第2の発明に係る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である。

【発明の効果】

【0024】

本発明の翻訳学習装置、方法、及びプログラムによれば、精度良く、翻訳元言語文を翻訳先言語文に翻訳するための翻訳モデルを学習することができる、という効果が得られる。

20

【図面の簡単な説明】

【0025】

【図1】第1の実施の形態に係る翻訳学習装置の機能的構成を示すブロック図である。

【図2】自然言語文の単語と論理形式文の単語の対応付けである単語アラインメントを誤って推定する場合の例を示す概念図である。

【図3】部分文字列アラインメントアノテーションデータの一例を示す図である。

【図4】半教師ありフレーズ推定による単語アラインメントの改善例を示す図である。

【図5】第1の実施の形態に係る翻訳装置の機能的構成を示すブロック図である。

30

【図6】第1の実施の形態に係る翻訳学習装置における翻訳学習処理ルーチンを示すフローチャートである。

【図7】第1の実施の形態に係る翻訳装置における翻訳処理ルーチンを示すフローチャートである。

【発明を実施するための形態】

【0026】

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。

【0027】

<本発明の実施の形態に係る概要>

【0028】

40

まず、本発明の実施の形態における概要を説明する。

【0029】

本発明の実施の形態に係る翻訳学習装置及び翻訳装置は、翻訳元言語文（自然言語文）から自動的に翻訳先言語文（論理形式文）へと翻訳することを目的としている。例えば、「AAAの価格は」という自然言語文を、自動的に

【0030】

?s prop-ja:定価 ?o. FILTER(regex(str(?s), “AAA”)).

【0031】

という論理形式文へと翻訳することである。なお、本実施の形態では、翻訳元言語文を自然言語文とし、翻訳先言語文として論理形式文とした場合を例に説明するが、これに限定さ

50

れるものではなく、翻訳先言語文を他の自然言語文としてもよい。

【0032】

ここで、「?s」は求めたい主語、「?o」は求めたい目的語、「prop-ja」は述語が定義されたURIを示し、上記例では条件文として用いられる。「FILTER」も条件文を表し、「regex」は正規表現であることを示す。「str(?s)」は主語を文字列として解釈し、「“AAA”」という文字列を含む場合、条件を満たすことを表す。なお、論理形式文とは、論理式で表される、計算機が理解可能な表現形式であり、入計算や一般的なデータベースへの問い合わせ言語を含むものである。

【0033】

＜本発明の第1の実施の形態に係る翻訳学習装置の構成＞

10

【0034】

次に、本発明の第1の実施の形態に係る翻訳学習装置の構成について説明する。図1に示すように、本発明の第1の実施の形態に係る翻訳学習装置100は、CPUと、RAMと、後述する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出来る。この翻訳学習装置100は、機能的には図1に示すように入力部10と、演算部20とを備えている。

【0035】

入力部10は、自然言語文と、自然言語文に対応する論理形式文とのペアの集合を受け付ける。

【0036】

20

演算部20は、自然言語文間翻訳モデル学習部28と、単語アラインメント推定部30と、単語アラインメント32と、単語複製部34と、複製済単語アラインメント36と、半教師ありフレーズ推定部38と、フレーズテーブル40と、フレーズフィルタリング部42と、フィルタリング済フレーズテーブル44と、翻訳モデル学習部46と、翻訳モデル48と、パラフレーズ適用部50と、パラフレーズ済翻訳モデル52と、自然言語文間翻訳モデル60と、単語複製モデル62とを含んで構成されている。

【0037】

自然言語文間翻訳モデル学習部28は、以下に説明するように、入力部10で受け付けた自然言語文と、論理形式文とのペアの集合に基づいて、自然言語文間翻訳モデル60を学習する。自然言語文間翻訳モデル60の学習には既知の技術、例えば非特許文献2のmo

30

【0038】

自然言語文間翻訳モデル学習部28は、まず、自然言語文及び論理形式文に対応付けられる部分文字列の組み合わせを表す部分文字列アラインメントに抽象化ラベル（例えば「“AAA”」に対して抽象化ラベル「entity」がラベル付けされる）が付与された部分文字列アラインメントアノテートデータ26に基づいて、各ペアの自然言語文及び論理形式文を抽象化する。例えば、自然言語文側を抽象化すると「AAA(=annotated as entity) の価格(=annotated as property)」は、「[entity] の [property] は」となる。論理形式文側を抽象化すると「?s prop-ja: 定価 ?o.(=annotated as property) FILTER(regex(str(?s), “AAA”(=annotated as entity))).」は、「?s [property] ?o. FILTER(regex(str(?s), [entity])).」となる。

40

【0039】

そして、自然言語文間翻訳モデル学習部28は、同じ抽象化された論理形式文と対となる、抽象化された自然言語文の集合を求め、求められた抽象化された自然言語文の集合に含まれる自然言語文を互いに翻訳対であるとみなして翻訳モデルを学習する。例えば「?s [property] ?o. FILTER(regex(str(?s), [entity])).」という抽象化された論理形式文と対となる、抽象化された自然言語文の集合において、{[entity] の [property] は, [entity] の [property] って何?} 等がある場合、この2文を自然言語文のテンプレートとなる翻訳ペアと見なす。翻訳ペアの組み合わせとして、10文ある場合には、全組み合わせである45通りの翻訳ペアを用いて、自然言語文間翻訳モデル60を学習しても良い

50

し、同じ文を使わないように組み合わせた5通りの翻訳ペアを用いて学習するなどしても良い。また、以下のように、機能語である助詞（例えば「の」）等をあらかじめ削除してから翻訳モデルを作成することで学習データの疎性を抑えることができる。

【0040】

「[entity] の [property] は」→「[entity] [property] は」

【0041】

単語アラインメント推定部30は、入力部10で受け付けた自然言語文と、論理形式文とのペアの集合に基づいて、自然言語文及び論理形式文の間の単語の対応付けである単語アラインメントを推定して単語アラインメント32として記憶する。なお、単語アラインメントの推定には既知の技術として、例えば非特許文献2で用いられているGIZA++ (IBM (R) モデル)等を用いればよい。

10

【0042】

単語複製部34は、単語アラインメント推定部30で推定された単語アラインメント32に基づいて、自然言語文及び論理形式文のペアの各々について、当該ペアの単語アラインメントにおいて、自然言語文の1つの単語が、論理形式文の複数の部分文字列に対応し、かつ、複数の部分文字列が同一又は類似している場合、当該1つの単語を複製し、複製した自然言語文の単語と論理形式文の単語との対応付けである単語アラインメントである複製済単語アラインメント36を生成すると共に、自然言語文の単語の各々について複製された単語数に基づいて、自然言語文の単語の複製数を推定するための単語複製モデル62を学習する。

20

【0043】

単語複製部34は、具体的には以下の第1の処理及び第2の処理を行う。

【0044】

まず、単語複製部34は、第1の処理として、単語アラインメント推定部30での推定により得られた単語アラインメント32の結果から、「1:n対応しており、n側の出現形が同一か、又は類似している」という条件を満たす自然言語側単語を複製し、アラインメントを分割して、複製済単語アラインメント36として記憶する。ここでn側の出現形が同一というのは、「論理形式文側で同じ表層として単語が2回以上現れる場合」であり、n側の出現形が類似しているというのは、「論理形式文側で単語の意味がn個ある場合」である。

30

【0045】

第1の処理における「論理形式文側で単語の意味がn個ある場合」に単語を複製する例を説明する。例えば自然言語文「AAAの価格は」についての単語アラインメントの場合であれば、

【0046】

アラインメント1:「価格」<->「prop-ja:価格」

【0047】

アラインメント2:「価格」<->「prop-ja-nl:価格」

【0048】

という対応が、単語アラインメント推定部30による単語アラインメント推定の結果により既知であるとする。この場合に、「prop-ja」や「prop-ja-nl」において、論理形式文のデータベースのエンティティ「AAA」の属性「prop-ja:価格」に入っている値が「1000000円」、エンティティ「AAA」の属性「prop-ja-nl:価格」に入っている値が「軽量化によるコストダウンを実施しました」というように、それぞれ「一問一答の答え」と「説明文」であると想定する。これは「AAAの価格は」と聞かれた場合、答えだけを返すよりも、説明も加えた方が、対話が自然に成り立つ場合があるため、このようなケースに対応する状況を仮定している。そして、単語アラインメントにおいて、n個対象となる単語を複製する。上記例の「AAAの価格は」であれば、「AAAの価格 価格 は」というように4単語目に「価格」を挿入して、次のようにアラインメントする。

40

【0049】

50

アラインメント 1 : 「価格(3 単語目)」<->「prop-ja : 価格」

【0050】

アラインメント 2 : 「価格(4 単語目)」<->「prop-ja-nl : 価格」

【0051】

この場合、論理形式文側の「prop-ja : 価格」と「prop-ja-nl : 価格」とが類似している、と判定する。この時の類似尺度には表記文字列に対する、既存の表記編集距離等と閾値を用いれば良い。

【0052】

一方、複製しない例としては、例えば

【0053】

アラインメント 1 : 「最大」<->「max : 」

【0054】

アラインメント 2 : 「最大」<->「xsd : 」

【0055】

という単語アラインメントであれば、論理形式文側において単語が非類似である。この場合は、後段のフレーズとして獲得可能であるため、そのまま扱えばよい。なお、2 単語以上の単語列内の全単語が同じ 1 : n アラインメントをとる場合には、次のように、A 単語及び B 単語からなる単語列内の順番を保持したまま、単語を複製すると共にアラインメントを含めて分割する。これを形態素として 1 単語として形態素解析器に記憶することで、デコード時にも同様に扱うことが可能である。

【0056】

(“ A B ” -> “ A B A B ”)

【0057】

例えば、自然言語文側の A から論理形式文側の a 及び a ' に、自然言語文側の B から論理形式文側の b 及び b ' のそれぞれに 2 か所、計 4 本のアラインメントがあった場合、自然言語文を A B 及び A B の 2 つに複製した後、アラインメントを A B<->a b (2 本) 、 A B<->a ' b ' (2 本) のそれぞれに分割する。

【0058】

次に、第 1 の処理における「論理形式文側で同じ表層として単語が 2 回以上現れる場合」に単語を複製する例を説明する。この場合は単なる論理形式文側の S P A R Q L 制約上の問題として扱われるものである。例えば「モーターの最大トルクが一番大きな車は」についての単語アラインメントの場合であれば、対応する論理形式文の S P A R Q L は以下のようになる。

【0059】

```
SELECT DISTINCT ?s ?o ?ol WHERE { ?s (prop-ja:最大トルク-モーター) ?o ; (prop-ja-nl:最大トルク-モーター) ?ol . FILTER ( ?o = ?temp ) { SELECT ( ?o AS ?temp ) WHERE { ?s (prop-ja:最大トルク-モーター) ?o . } ORDER BY DESC(xsd:integer(?o)) LIMIT 1 } } }
```

【0060】

上記の例では「prop-ja:最大トルク-モーター」が 2 回出現しているが、これは演算のための一時的な代入を「?temp」という変数に対して行うためであり、単なる S P A R Q L の記述方法の問題にすぎず、本質的な問題ではない。このような場合には、機械的に単語を必要な数分だけ複製すればよい。上記例の場合は 2 個に複製すればよい。

【0061】

単語複製部 3 4 は、第 2 の処理として、第 1 の処理において自然言語文の単語の各々について複製された単語数に基づいて、デコード時に自然言語文の単語の複製数を推定するための識別モデルとして単語複製モデル 6 2 を学習する。単語複製モデル 6 2 の学習は論理形式文側の単語毎 (類似単語の場合はどちらか一方) 、あるいは論理形式文側の単語のクラス、つまり R D F で定義されているフィールド名 (s , p , o 等) 毎に行う。教師信号は複製された単語数とする。例えば、上記第 1 の処理の例では 2 単語に複製されたので、“

10

20

30

40

50

2”とする。学習結果の例としては、

【0062】

$p(2\text{個} | \text{“価格”}, \phi) = 0.8$ $p(1\text{個} | \text{“価格”}, \phi) = 0.2$

【0063】

となる。ここで、 ϕ は当該文における素性集合である。第2の処理での識別モデルとしては、既知のモデル、例えばSVM（サポートベクタマシン）やロジスティック回帰分類器を用いればよく、素性もBOW（Bag of Words）や係り受け素性等、既知の素性を用いればよい。

【0064】

半教師ありフレーズ推定部38は、自然言語文及び論理形式文のペアの各々について、自然言語文の部分文字列と論理形式文の部分文字列との対応付けである予め定められた部分文字列アラインメントに基づいて、当該ペアについての複製済単語アラインメント36を補正し、ペアの各々について補正した複製済単語アラインメント36に基づいて、自然言語文のフレーズと論理形式文のフレーズとのペアを格納したフレーズテーブル40を推定する。

【0065】

具体的には、半教師ありフレーズ推定部38は、まず、部分文字列アラインメントアノテートデータ26を適用して、自然言語文及び論理形式文のペアの各々についての複製済単語アラインメント36に当てはまる部分文字列アラインメントを定める。図2に、単語アラインメント推定部30で自動推定された単語アラインメントの例を示す。図2では、黒で塗りつぶされたマスが単語アラインメントを、点線で囲まれたマスはフレーズアラインメントとなり得る箇所を表している。図3に、部分文字列アラインメントアノテートデータ26における部分文字列アラインメントの例を示す。

【0066】

次に、半教師ありフレーズ推定部38は、図2の単語アラインメントのうち、図3の部分文字列アラインメントアノテートデータと矛盾する単語アラインメントを削除する。矛盾する単語アラインメントを削除した結果を図4に示す。「×」は灰色の部分文字列アラインメントアノテートデータと矛盾するため、削除された単語アラインメントを示す。また、アラインメントが削除された単語については、単語アラインメント推定部30で得られている単語アラインメントの確率値を参照し、矛盾せずに最適となるアラインメントの探索及び変更を行うことにより複製済単語アラインメント36を補正する。例えば、図4において、「×」で示された、アラインメントが削除された箇所に対して、点線で囲まれたフレーズアラインメントとなり得る箇所が変更される。

【0067】

そして、半教師ありフレーズ推定部38は、補正された複製済単語アラインメント36を用いて、対応付けられる自然言語文のフレーズ及び論理形式文のフレーズのペアの各々及びペアの各々の翻訳確率を格納したフレーズテーブル40の作成を行う。ここでは既存のヒューリスティックとして、例えば非特許文献2のgrow-diag等の処理を行い、フレーズテーブル40を得る。

【0068】

フレーズフィルタリング部42は、半教師ありフレーズ推定部38により推定されたフレーズテーブル40から、論理形式文について予め定められた制約を満たさないフレーズを含むペアを削除し、フィルタリング済フレーズテーブル44を得る。本実施の形態では、予め定められた制約として、開き括弧“(”や閉じかっこ“) ”の数が合うことを用い、開き括弧“(”や閉じかっこ“) ”の数が合わないフレーズを含むペアを削除する。

【0069】

翻訳モデル学習部46は、フレーズフィルタリング部42で得られたフィルタリング済フレーズテーブル44に基づいて、自然言語文を論理形式文に翻訳するための翻訳モデル48を学習する。本実施の形態では、後述する翻訳装置における翻訳時（デコード時）に必要となる、フレーズ同士の翻訳確率や、フレーズが前後に移動する際の歪み確率等のス

10

20

30

40

50

コアを記憶した翻訳モデルを学習する。学習方式は既存手法を用いればよく、例えば、非特許文献3（統計的機械翻訳ツールmoses、<http://www.statmt.org/moses/>）にフリーツールとして公開されているmosesを用いることができる。

【0070】

パラフレーズ適用部50は、翻訳モデル学習部46により学習された翻訳モデル48に含まれる各フレーズについて、言い換え可能なフレーズを格納したパラフレーズ辞書（図示省略）を参照して、言い換え可能なフレーズに拡張し、言い換え可能なフレーズに拡張した翻訳モデルを、パラフレーズ済翻訳モデル52として生成する。ここで用いるパラフレーズ辞書は既存のものでよく、例えば非特許文献4（水上 雅博, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲. 「日本語言い換えデータベースの構築と言語的個人性変換への応用」 言語処理学会第20回年次大会 (NLP). 2014/3）等に記載されているものを用いればよい。本実施の形態では、翻訳モデル学習前にパラフレーズを拡張しても、学習データにパラフレーズの例が現れないため、翻訳確率が付与されないため、翻訳モデル学習後に適用することとした。

10

【0071】

<本発明の第1の実施の形態に係る翻訳装置の構成>

【0072】

次に、本発明の第1の実施の形態に係る翻訳装置の構成について説明する。図6に示すように、本発明の第1の実施の形態に係る翻訳装置200は、CPUと、RAMと、後述する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出来る。この翻訳装置200は、機能的には図6に示すように入力部210と、演算部220とを備えている。

20

【0073】

入力部210は、翻訳対象の自然言語文を受け付ける。

【0074】

演算部220は、自然言語文間翻訳部228と、単語複製部230と、デコーディング部232と、パラフレーズ済翻訳モデル252と、自然言語文間翻訳モデル260と、単語複製モデル262とを含んで構成されている。

【0075】

自然言語文間翻訳モデル260は、自然言語文間翻訳モデル60と同じものが記憶されている。

30

【0076】

部分文字列アラインメントアノテートデータ226は、部分文字列アラインメントアノテートデータ26と同じものが記憶されている。

【0077】

単語複製モデル262は、単語複製モデル62と同じものが記憶されている。

【0078】

パラフレーズ済翻訳モデル252は、パラフレーズ済翻訳モデル52と同じものが記憶されている。

【0079】

40

自然言語文間翻訳部228は、まず、入力部10で受け付けた自然言語文と、自然言語文及び論理形式文に対応付けられる部分文字列の組み合わせを表す予め定められた部分文字列アラインメントアノテートデータ226とに基づいて、自然言語文を抽象化する。例えば、自然言語文の部分文字列「AAA」に対し、部分文字列アラインメントアノテートデータ226において抽象化ラベル“entity”が付与されているため、これをそのまま抽象化するための辞書と見なして自然言語文を抽象化する。例えば、自然言語文が「AAA の価格 って どのくらい」であれば、「[entity] の [property] って どのくらい」と抽象化する。

【0080】

次に、自然言語文間翻訳部228は、自然言語文間翻訳モデル260に基づいて、抽象

50

化された自然言語文を自然言語文に翻訳する。ここではn b e s t 解を得る。つまり抽象化された自然言語文集合に拡張する。そして、抽象化された自然言語文の部分文字列を元の抽象化される前の部分文字列に変換し、拡張された自然言語文を得る。例えば、抽象化された自然言語文が「[entity] の [property] って何？」であれば「AAAの価格って何？」と変換する。ここで拡張したn b e s t 文集合の全ての文に対して後段の単語複製部 2 3 0 及びデコーディング部 2 3 2 の処理を行い、パラフレーズ済翻訳モデル 2 5 2 の翻訳確率に基づいて、最終的な論理形式文を生成すればよい。

【0081】

なお、自然言語文間翻訳部 2 2 8 において、自然言語文間翻訳モデル 2 6 0 の学習データの網羅性が高ければ、抽象化された自然言語文集合のうち、入力文からの翻訳確率の最も高いものを選択し、選択された、抽象化自然言語文に対応する抽象化論理形式文へ一意に変換し、さらに抽象化されていた部分文字列に対し、対応する論理形式文側の変換結果を用いて、論理形式文を得てもよい。この場合は、後段の単語複製部 2 3 0 及びデコーディング部 2 3 2 の処理を行う必要はない。

【0082】

単語複製部 2 3 0 は、単語複製モデル 2 6 2 に基づいて、自然言語文間翻訳部 2 2 8 で拡張された自然言語文の各々に対し、当該自然言語文の単語の各々について、当該単語の複製数を推定して、当該単語を複製する。

【0083】

具体的には、単語複製部 2 3 0 は、まず、単語複製モデル 2 6 2 を用いて、自然言語文に含まれる各単語の複製数を推定する。推定結果の例としては、 $p(2\text{個} | \text{“価格”}, \phi) = 0.85$ 、 $p(1\text{個} | \text{“価格”}, \phi) = 0.15$ というものである。ここで「 ϕ 」は当該自然言語文における素性集合を表す。次に、推定結果に基づいて、単語を複製する。この例の場合は“価格”を2個に複製する。なお、複製の際、単語複製モデル 2 6 2 の上位N種類の複製数を全て出力し、単語複製済自然言語文として保持し、複数の単語複製済自然言語文に対してデコーディングを行ってもよい。

【0084】

デコーディング部 2 3 2 は、パラフレーズ済翻訳モデル 2 5 2 に基づいて、単語複製部 2 3 0 で単語が複製された、拡張された自然言語文の各々を、論理形式文に翻訳し、翻訳確率に基づいて選択された論理形式文を出力する。翻訳処理には、既存手法を用いればよく、例えば、上記非特許文献 3 にフリーツールとして公開されている m o s e s を用いることができる。

【0085】

<本発明の第1の実施の形態に係る翻訳学習装置の作用>

【0086】

次に、本発明の実施の形態に係る翻訳学習装置 1 0 0 の作用について説明する。入力部 1 0 において自然言語文と、自然言語文に対応する論理形式文とのペアの集合を受け付けると、翻訳学習装置 1 0 0 は、図 6 に示す翻訳学習処理ルーチンを実行する。

【0087】

まず、ステップ S 1 0 0 では、入力部 1 0 において受け付けた、自然言語文と、自然言語文に対応する論理形式文とのペアの集合を取得する。

【0088】

次に、ステップ S 1 0 2 では、ステップ S 1 0 0 で取得した、自然言語文と論理形式文とのペアの集合、及び部分文字列アライメントアノテートデータ 2 6 に基づいて、自然言語文間翻訳モデル 6 0 を学習する。

【0089】

ステップ S 1 0 4 では、ステップ S 1 0 0 で取得した、自然言語文と論理形式文とのペアの集合に基づいて、自然言語文及び論理形式文の間の単語の対応付けである単語アライメント 3 2 を推定する。

【0090】

ステップS106では、ステップS104で推定された単語アラインメント32に基づいて、自然言語文及び論理形式文のペアの各々について、当該ペアの単語アラインメントにおいて、自然言語文の1つの単語が、論理形式文の複数の部分文字列に対応し、かつ、複数の部分文字列が同一又は類似している場合、当該1つの単語を複製し、複製した自然言語文の単語と論理形式文の単語との対応付けである単語アラインメントである複製済単語アラインメント36を生成する。

【0091】

ステップS108では、ステップS106で自然言語文の単語の各々について複製された単語数に基づいて、自然言語文の単語の複製数を推定するための単語複製モデル62を学習する。

10

【0092】

ステップS110では、ステップS106で生成された複製済単語アラインメント36の自然言語文及び論理形式文のペアの各々について、自然言語文の部分文字列と論理形式文の部分文字列との対応付けである予め定められた部分文字列アラインメントに基づいて、当該ペアについての複製済単語アラインメント36を補正する。

【0093】

ステップS112では、ステップS110でペアの各々について補正した複製済単語アラインメント36に基づいて、自然言語文のフレーズと論理形式文のフレーズとのペアを格納したフレーズテーブル40を推定する。

20

【0094】

ステップS114では、ステップS112で推定されたフレーズテーブル40から、論理形式文について予め定められた制約を満たさないフレーズを含むペアを削除し、フィルタリング済フレーズテーブル44を得る。

【0095】

ステップS116では、ステップS114で得られたフィルタリング済フレーズテーブル44に基づいて、自然言語文を論理形式文に翻訳するための翻訳モデル48を学習する。

【0096】

そして、ステップS118では、ステップS116で学習された翻訳モデル48に含まれる各フレーズについて、パラフレーズ辞書を参照して、言い換え可能なフレーズに拡張し、言い換え可能なフレーズに拡張した翻訳モデルを、パラフレーズ済翻訳モデル52として生成して処理を終了する。

30

【0097】

以上説明したように、第1の実施の形態に係る翻訳学習装置によれば、推定された単語アラインメントに基づいて、自然言語文及び論理形式文のペアの各々について、当該ペアの単語アラインメントにおいて、自然言語文の1つの単語が、論理形式文の複数の部分文字列に対応し、かつ、複数の部分文字列が同一又は類似している場合、1つの単語を複製した自然言語文の単語と論理形式文の単語との対応付けである単語アラインメントである複製済単語アラインメントを生成すると共に、自然言語文の単語の各々について複製された単語数に基づいて、自然言語文の単語の複製数を推定するための単語複製モデルを学習し、自然言語文及び論理形式文のペアの各々について生成された複製済単語アラインメントからフレーズテーブルを推定し、フィルタリング済フレーズテーブルに基づいて、翻訳モデルを学習し、言い換え可能なフレーズに拡張することにより、精度良く、自然言語文を論理形式文に翻訳するための翻訳モデルを学習することができる、という効果が得られる。

40

【0098】

<本発明の第1の実施の形態に係る翻訳装置の作用>

【0099】

次に、本発明の第1の実施の形態に係る翻訳装置200の作用について説明する。入力部210において翻訳対象の自然言語文を受け付けると、翻訳装置200は、図7に示す

50

翻訳処理ルーチンを実行する。

【0100】

まず、ステップS200では、入力部210において受け付けた翻訳対象の自然言語文を取得する。

【0101】

次に、ステップS202では、ステップS200で取得した自然言語文と、部分文字列アラインメントアノテートデータ226とに基づいて、自然言語文を抽象化する。

【0102】

ステップS204では、自然言語文間翻訳モデル260に基づいて、ステップS202で抽象化された自然言語文を、拡張された自然言語文に翻訳する。

10

【0103】

ステップS206では、単語複製モデル262に基づいて、ステップS204で翻訳されて得られた拡張された自然言語文の各々に対し、当該自然言語文の単語の各々について、当該単語の複製数を推定して、当該単語を複製する。

【0104】

ステップS208では、パラフレーズ済翻訳モデル252に基づいて、ステップS206で単語が複製された、拡張された自然言語文の各々を論理形式文に翻訳し、翻訳確率に基づいて選択された論理形式文を出力し、処理を終了する。

【0105】

以上説明したように、第1の実施の形態に係る翻訳装置によれば、拡張された自然言語文の単語の各々について、当該単語の複製数を推定して、当該単語を複製し、単語が複製された、拡張された自然言語文を論理形式文に翻訳することで、精度良く、自然言語文を論理形式文に翻訳することができる。

20

【0106】

<本発明の第2の実施の形態に係る翻訳学習装置の構成>

【0107】

次に、本発明の第2の実施の形態に係る翻訳学習装置の構成について説明する。なお、第1の実施の形態の翻訳学習装置100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。

【0108】

上記図1に示すように、第2の実施の形態に係る翻訳学習装置100は、CPUと、RAMと、後述する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出来る。この翻訳学習装置100は、機能的には図1に示すように入力部10と、演算部20とを備えている。

30

【0109】

第2の実施の形態に係る演算部20は、自然言語文間翻訳モデル学習部28と、単語アラインメント推定部30と、単語アラインメント32と、単語複製部34と、複製済単語アラインメント36と、半教師ありフレーズ推定部38と、フレーズテーブル40と、フレーズフィルタリング部42と、フィルタリング済フレーズテーブル44と、翻訳モデル学習部46と、翻訳モデル48と、パラフレーズ適用部50と、パラフレーズ済翻訳モデル52と、自然言語文間翻訳モデル60と、単語複製モデル62とを含んで構成されている。

40

【0110】

第2の実施の形態に係る自然言語文間翻訳モデル学習部28は、第1の実施の形態と同様に、入力部10で受け付けた自然言語文と、論理形式文とのペアの集合に基づいて、自然言語文間翻訳モデル60を学習する。そして、第1の実施の形態と同様の手法を用いて抽象化した自然言語文と、論理形式文とのペアの集合を単語アラインメント推定部30に出力する。

【0111】

第2の実施の形態に係る単語アラインメント推定部30は、自然言語文間翻訳モデル学

50

習部 28 で得られる、抽象化された自然言語文と論理形式文とのペアの集合に基づいて、自然言語文及び論理形式文の間の単語の対応付けである単語アラインメントを推定して単語アラインメント 32 として記憶する。

【0112】

第 2 の実施の形態に係る単語複製部 34 は、単語アラインメント推定部 30 で推定された単語アラインメント 32 に基づいて、抽象化された自然言語文及び論理形式文のペアの各々について、当該ペアの単語アラインメントにおいて、抽象化された自然言語文の 1 つの単語が、抽象化された論理形式文の複数の部分文字列に対応し、かつ、複数の部分文字列が同一又は類似している場合、当該 1 つの単語を複製し、複製した自然言語文の単語と論理形式文の単語との対応付けである単語アラインメントである複製済単語アラインメント 36 を生成すると共に、抽象化された自然言語文の単語の各々について複製された単語数に基づいて、抽象化された自然言語文の単語の複製数を推定するための単語複製モデル 62 を学習する。

10

【0113】

第 2 の実施の形態に係る半教師ありフレーズ推定部 38 は、抽象化された自然言語文及び論理形式文のペアの各々について、抽象化された自然言語文の部分文字列と抽象化された論理形式文の部分文字列との対応付けである予め定められた部分文字列アラインメントに基づいて、当該ペアについての複製済単語アラインメント 36 を補正し、ペアの各々について補正した複製済単語アラインメント 36 に基づいて、抽象化された自然言語文のフレーズと抽象化された論理形式文のフレーズとのペアを格納したフレーズテーブル 40 を推定する。

20

【0114】

第 2 の実施の形態に係るフレーズフィルタリング部 42 は、半教師ありフレーズ推定部 38 により推定されたフレーズテーブル 40 から、論理形式文について予め定められた制約を満たさないフレーズを含むペアを削除し、フィルタリング済フレーズテーブル 44 を得る。

【0115】

第 2 の実施の形態に係る翻訳モデル学習部 46 は、フレーズフィルタリング部 42 で得られたフィルタリング済フレーズテーブル 44 のフレーズに含まれる抽象化ラベルを抽象化前の単語に変換する。そして、抽象化ラベルが単語に変換されたフィルタリング済フレーズテーブル 44 に基づいて、自然言語文を論理形式文に翻訳するための翻訳モデル 48 を学習する。

30

【0116】

なお、第 2 の実施の形態に係る翻訳学習装置の他の構成及び作用は、第 1 の実施の形態の翻訳学習装置 100 と同様であるため詳細な説明を省略する。

【0117】

<本発明の第 2 の実施の形態に係る翻訳装置の構成>

【0118】

次に、本発明の第 2 の実施の形態に係る翻訳装置の構成について説明する。なお、第 1 の実施の形態の翻訳装置 200 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。

40

【0119】

上記図 6 に示すように、第 2 の実施の形態に係る翻訳装置 200 は、CPU と、RAM と、後述する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出来る。この翻訳装置 200 は、機能的には図 6 に示すように入力部 210 と、演算部 220 とを備えている。

【0120】

第 2 の実施の形態に係る演算部 220 は、自然言語文間翻訳部 228 と、単語複製部 230 と、デコーディング部 232 と、パラフレーズ済翻訳モデル 252 と、自然言語文間翻訳モデル 260 と、単語複製モデル 262 とを含んで構成されている。

50

【0121】

第2の実施の形態に係る自然言語文間翻訳部228は、第1の実施の形態と同様に、入力部10で受け付けた自然言語文と、部分文字列アラインメントアノテートデータ226とに基づいて、自然言語文を抽象化する。

【0122】

第2の実施の形態に係る単語複製部230は、単語複製モデル262に基づいて、自然言語文間翻訳部228で抽象化された自然言語文の単語の各々について、当該単語の複製数を推定して、当該単語を複製する。

【0123】

第2の実施の形態に係るデコーディング部232は、まず、単語複製部230で、単語が複製された、抽象化された自然言語文の部分文字列を、抽象化される前の自然言語文の部分文字列に変換し、拡張された自然言語文を得る。そして、パラフレーズ済翻訳モデル252に基づいて、第1の実施の形態と同様の手法を用いて、拡張された自然言語文の各々を、論理形式文に翻訳し、翻訳確率に基づいて選択された論理形式文を出力する。

【0124】

なお、第2の実施の形態に係る翻訳装置の他の構成及び作用は、第1の実施の形態の翻訳装置200と同様であるため詳細な説明を省略する。

【0125】

以上説明したように、第2の実施の形態に係る翻訳学習装置によれば、抽象化された自然言語文と、論理形式文とのペアの集合に基づいて、単語アラインメントを推定し、推定された単語アラインメントに基づいて、抽象化された自然言語文及び論理形式文のペアの各々について、当該ペアの単語アラインメントにおいて、抽象化された自然言語文の1つの単語が、論理形式文の複数の部分文字列に対応し、かつ、複数の部分文字列が同一又は類似している場合、1つの単語を複製した自然言語文の単語と論理形式文の単語との対応付けである単語アラインメントである複製済単語アラインメントを生成すると共に、抽象化された自然言語文の単語の各々について複製された単語数に基づいて、抽象化された自然言語文の単語の複製数を推定するための単語複製モデルを学習し、抽象化された、自然言語文及び論理形式文のペアの各々について生成された複製済単語アラインメントからフレーズテーブルを推定し、フィルタリング済フレーズテーブルに基づいて、翻訳モデルを学習し、言い換え可能なフレーズに拡張することにより、精度良く、自然言語文を論理形式文に翻訳するための翻訳モデルを学習することができる、という効果が得られる。

【0126】

また、本発明の第2の実施の形態に係る翻訳装置によれば、抽象化された自然言語文の単語の各々について、当該単語の複製数を推定して、当該単語を複製し、単語が複製された抽象化自然言語文の部分文字列を、抽象化前の自然言語文の部分文字列に変換し、変換された翻訳対象の自然言語文を論理形式文に翻訳することで、精度良く、自然言語文を論理形式文に翻訳することができる。

【0127】

なお、本発明は、上述した実施の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。

【0128】

例えば、上述した実施の形態では、フィルタリング済フレーズテーブル44に基づいて、自然言語文を論理形式文に翻訳するための翻訳モデル48を学習していたが、これに限定されるものではなく、単語複製部34により生成された複製済単語アラインメント36、又は半教師ありフレーズ推定部38により推定されたフレーズテーブル40に基づいて、翻訳モデル48を学習してもよい。

【0129】

また、上述した実施の形態では、半教師ありフレーズ推定部38は、複製済単語アラインメント36を補正し、ペアの各々について補正した複製済単語アラインメント36に基づいて、自然言語文のフレーズと論理形式文のフレーズとのペアを格納したフレーズテー

10

20

30

40

50

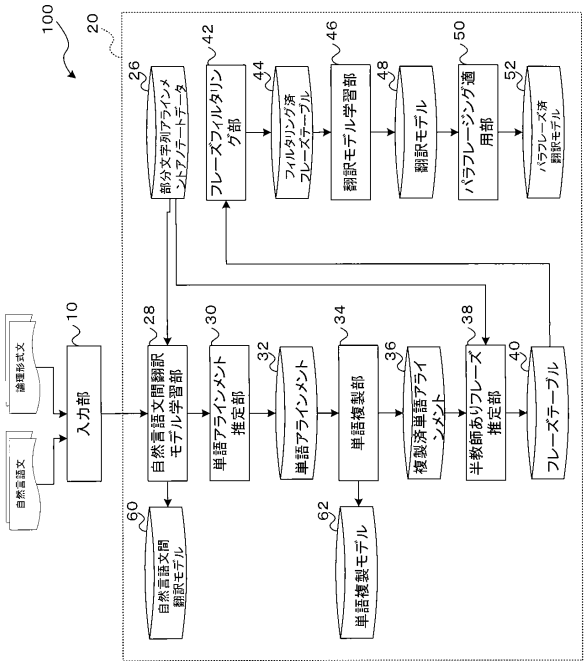
ブル40を推定しているが、これに限定されるものではなく、単語アラインメント32を補正し、ペアの各々について補正した単語アラインメント32に基づいて、自然言語文のフレーズと論理形式文のフレーズとのペアを格納したフレーズテーブル40を推定してもよい。

【符号の説明】

【0130】

10、210	入力部	
20、220	演算部	
26、226	部分文字列アラインメントアノテートデータ	
28	自然言語文間翻訳モデル学習部	10
30	単語アラインメント推定部	
32	単語アラインメント	
34	単語複製部	
36	複製済単語アラインメント	
38	半教師ありフレーズ推定部	
40	フレーズテーブル	
42	フレーズフィルタリング部	
44	フィルタリング済フレーズテーブル	
46	翻訳モデル学習部	
48	翻訳モデル	20
50	パラフレージング適用部	
52、252	パラフレーズ済翻訳モデル	
60、260	自然言語文間翻訳モデル	
62、262	単語複製モデル	
100	翻訳学習装置	
200	翻訳装置	
228	自然言語文間翻訳部	
230	単語複製部	
232	デコーディング部	

【図 1】



【図 2】

	AAA	の	価格	は
?s				
prop-ja:				
定価				
?o				
.				
FILTER(				
regex(				
str(?s)				
AAA				
)				
).				

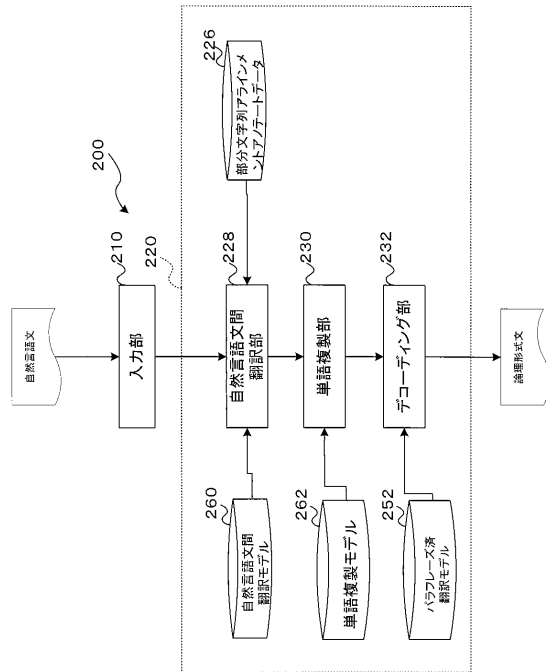
【図 3】

	AAA	の	価格	は
?s				
prop-ja:				
定価				
?o				
.				
FILTER(				
regex(				
str(?s)				
AAA				
)				
).				

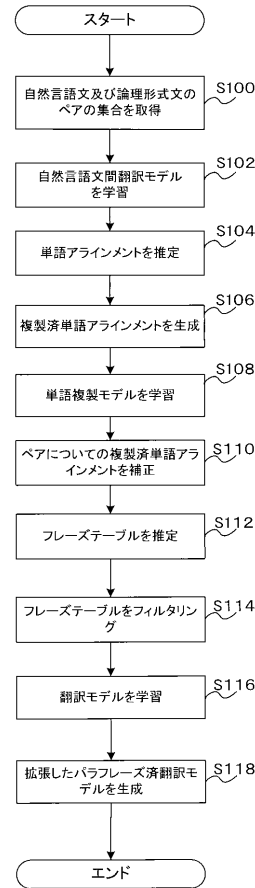
【図 4】

	AAA	の	価格	は
?s				
prop-ja:				
定価				
?o				
.				
FILTER(				
regex(				x
str(?s)				x
AAA				
)		x		
).		x		

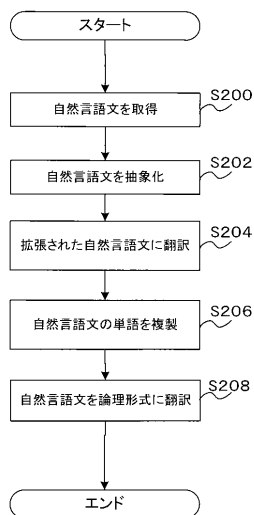
【図 5】



【図 6】



【図 7】



フロントページの続き

- (72)発明者 西川 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 斉藤 いつみ
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

審査官 成瀬 博之

- (56)参考文献 特表2009-543190 (JP, A)
特開2011-180823 (JP, A)
特開2013-120407 (JP, A)
米国特許出願公開第2011/0246177 (US, A1)

- (58)調査した分野(Int.Cl., DB名)
G06F 17/20-17/28