(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2018年5月31日(31.05.2018)

WIPO | PCT

(10) 国際公開番号
WO 2018/097091 A1

(51) 国際特許分類:
G06F 17/30 (2006.01)

(21) 国際出願番号： PCT/JP2017/041630

(22) 国際出願日： 2017年11月20日(20.11.2017)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(30) 優先権データ：
特願 2016-229072 2016年11月25日(25.11.2016) JP

(71) 出願人: 日本電信電話株式会社 (**NIPPON TELEGRAPH AND TELEPHONE CORPORATION**) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目５番１号 Tokyo (JP).

(72) 発明者: 大塚 淳史 (**OTSUKA, Atsushi**); 〒1808585 東京都武蔵野市緑町３丁目９－１１ ＮＴＴ知的財産センタ内 Tokyo (JP). 別所 克人(**BESSHO, Katsuji**); 〒1808585 東京都武蔵野市緑町３丁目９－１１ ＮＴＴ知的財産センタ内 Tokyo (JP). 西田 京介(**NISHIDA, Kyosuke**); 〒1808585 東京都武蔵野市緑町３丁目９－１１ ＮＴＴ知的財産センタ内 Tokyo (JP). 浅野 久子(**ASANO, Hisako**); 〒1808585 東京都武蔵野市緑町３丁目９－１１ ＮＴＴ知的財産センタ内 Tokyo (JP). 松尾 義博(**MATSUO, Yoshihiro**); 〒1808585 東京都武蔵野市緑町３丁目９－１１ ＮＴＴ知的財産センタ内 Tokyo (JP).

(74) 代理人: 伊東 忠重, 外(**ITOH, Tadashige** et al.); 〒1000005 東京都千代田区丸の内二丁目１

(54) **Title:** MODEL CREATION DEVICE, TEXT SEARCH DEVICE, MODEL CREATION METHOD, TEXT SEARCH METHOD, DATA STRUCTURE, AND PROGRAM

(54) 発明の名称： モデル作成装置、テキスト検索装置、モデル作成方法、テキスト検索方法、データ構造、及びプログラム

[図1]

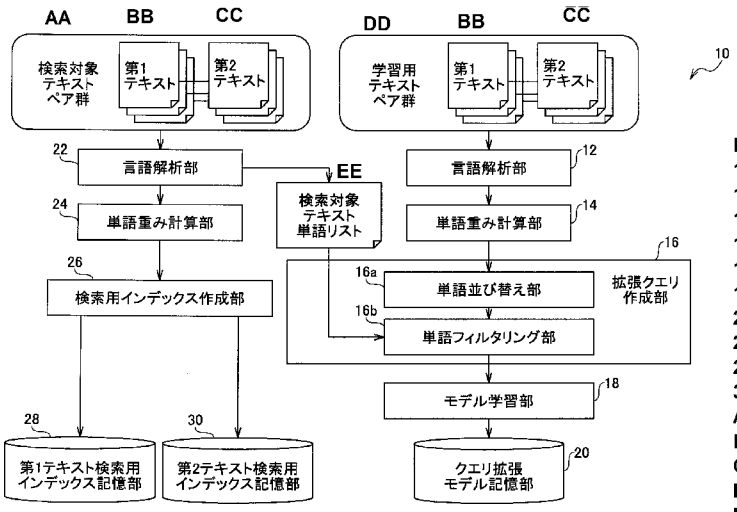


FIG. 1:
12, 22 Language analysis unit
14, 24 Vocabulary weighting calculation unit
16 Extended query creation unit
16a Vocabulary sorting unit
16b Vocabulary filtering unit
18 Model learning unit
20 Query extension model storage unit
26 Search index creation unit
28 First text search index storage unit
30 Second text search index storage unit
AA Search subject text pair group
BB First text
CC Second text
DD Text pair group for learning
EE Search subject text vocabulary list

(57) **Abstract:** With a text pair group for learning being taken as input, said text pair group for learning being formed from a pair of a first text for learning and a second text for learning which is an answer when the first text for learning is treated as a question, a query extension model is learned which creates text which forms an extended query with regard to text which forms a query.

(57) 要約：学習用の第１テキストと、学習用の第１テキストを質問としたときの回答となる学習用の第２テキストとのペアからなる学習用テキストペア群を入力として、クエリとなるテキストに対して、拡張クエリとなるテキストを作成するクエリ拡張モデルを学習する。

WO 2018/097091 A1

［続葉有］

番１号 丸の内 ＭＹ ＰＬＡＺＡ （明治安田生命ビル） １６階 Tokyo (JP).

(81) 指定国（表示のない限り、全ての種類の国内保護が可能）：AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国（表示のない限り、全ての種類の広域保護が可能）：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類：

— 国際調査報告（条約第21条(3)）

明　細　書

発明の名称 ：

モデル作成装置、テキスト検索装置、モデル作成方法、テキスト検索方法、データ構造、及びプログラム

技術分野

[0001] 本発明は、検索用に入力された音声又はテキストを拡張するクエリ拡張モデルを学習するモデル作成装置、入力した音声又はテキストについて検索するテキスト検索装置、検索用に入力された音声又はテキストを拡張するクエリ拡張モデルを学習するモデル作成方法、入力した音声又はテキストについて検索するテキスト検索方法、及び、検索用に入力された音声又はテキストを拡張したり、クエリ拡張モデルを学習するプログラム又は入力した音声又はテキストについて検索したりするプログラムに関する。

背景技術

[0002] 情報検索システムでは、ユーザが入力した検索キーワード集合（クエリ）に対して、キーワードマッチ等の処理によってクエリに適合するテキストを検索している。キーワードマッチ検索の場合はクエリとして入力されたキーワードと、テキスト内のキーワードとが完全一致していなくてはならず、検索の再現率（Recall）が低下してしまうという課題があった。そこで、クエリ内に含まれるキーワードを自動的に増やすことでより幅広い文書にマッチさせる技術にクエリ拡張がある。クエリ拡張では、検索ログから統計処理により拡張するキーワードを決定している。

[0003] 情報検索の応用として、質問応答（ＦＡＱ検索）、対話処理等がある。これらのシステムでは、ユーザの質問や発言に対して、妥当な応答を返すことが目的となる。質問応答システムや対話システムでは予め大量のＦＡＱ、応答候補文等をデータベースに保存しておき、ユーザの入力に対して情報検索のアプローチで最も妥当な候補を選択する問題となる。応答文検索では、データベースに質問と応答とのペアを保持しておき、ユーザの入力と質問とを

比較し、最も類似度が高かった質問の応答文を出力する。このようにすることで、ユーザの質問、発話等に対して適切な応答が可能になる。

先行技術文献

特許文献

[0004] 特許文献1：特開２０１４－９９０６２号公報

特許文献2：特開２０１１－１０３０１８号公報

発明の概要

発明が解決しようとする課題

[0005] ＦＡＱ検索、対話システム等では、Ｑ（質問：Question）及びＡ（答え：Answer）、発話文及びその応答文等の２つのテキストのペアをデータベースに保存しておき、実際の検索では、Ｑ及び発話文といったペアの第１テキストを主に使用して、実際の入力クエリとの比較を行う。その場合、Ａ及び応答文といった第２テキストは、検索で使用しない場合が多い。

[0006] しかしながら、入力クエリ及び第１テキストのみで比較を行っても関連性が判別せず、入力クエリと第２テキストとを比較して初めて関連性が明らかになるケースがある。

[0007] 例えば、ＦＡＱ検索で「動画が重くて見られません」という質問が入力クエリとして入ってきた時、「通信量が多いと帯域制限により通信速度が低下する場合があります」というＡを検索したいとする。これは、動画が見られない原因の一つとして、動画の見過ぎによる帯域制限が考えられるため、関連するＱＡとして妥当なものである可能性が高い。

[0008] しかしながら、実際のＦＡＱでは、「動画が重くて見られません」のようなより具体的な事象での言及で記載されているＱは少なく、「帯域制限について教えて下さい」の様に、一般化された内容で記載されていることが多い。このとき、「動画が重くて見られません」と「帯域制限について教えて下さい」とは通常の検索では類似性が低く、検索できない可能性が高い。

[0009] 入力クエリと第１テキストとで検索を行うことでは十分な検索結果が得ら

れない場合、入力クエリとＦＡＱのＡ等の第２テキストとを検索で使用することは可能である。しかし、入力クエリに含まれるキーワードと第２テキストで使用されているキーワードとが異なっていることから、入力クエリに含まれるキーワードで直接検索しても、適切な検索結果が得られず、十分な検索精度が得られない場合が多い。

[0010] その際、クエリ拡張等の手法を用いて入力クエリに含まれるキーワードを拡張することで対応することも可能であり、これまでは検索ログ等を用いたクエリ拡張の手法が用いられてきたが、この手法ではキーワードの意味的な類似性に基づいてキーワード拡張を行うため、キーワード拡張を行っても、第２テキストを検索するのに適したキーワードが作成できない場合が多い。

[0011] 例えば、上述した例では、「動画」というキーワードを拡張する際には、「動画」と意味的な類似性が高い「ビデオ」、「視聴」等といったキーワードが作成されることが多く、「通信量」、「帯域制限」等のキーワードが作成されることは稀である。

[0012] このように、入力クエリに含まれるキーワードに対してキーワード拡張を行っても、キーワード拡張によって適切なキーワードが作成されないため、入力クエリに対する検索結果を精度良く得ることができなかった。

[0013] 本発明は、以上のような事情に鑑みてなされたものであり、入力されたクエリに対するテキストのペアの検索結果を精度良く得ることができるモデル作成装置、テキスト検索装置、モデル作成方法、テキスト検索方法、データ構造、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

[0014] 上記目的を達成するために、本発明のモデル作成装置は、学習用の第１テキストと、前記学習用の第１テキストを質問としたときの回答となる学習用の第２テキストとのペアからなる学習用テキストペア群を入力として、クエリとなるテキストに対して、拡張クエリとなるテキストを作成するクエリ拡張モデルを学習するモデル学習部、を含む。

[0015] なお、検索対象の第１テキストと、前記検索対象の第１テキストを質問と

したときの回答となる検索対象の第２テキストとのペアからなる検索対象テキストペア群に基づいて、前記検索対象の第１テキストについての検索用インデックス、及び前記検索対象の第２テキストについての検索用インデックスを作成する検索インデックス作成部を更に含むようにしても良い。

[0016] また、検索対象の第１テキストと、前記検索対象の第１テキストを質問としたときの回答となる検索対象の第２テキストとのペアからなる検索対象テキストペア群に含まれる各単語からなる検索対象テキスト単語リストを用いて、前記学習用テキストペア群から、前記検索対象テキスト単語リストに含まれない単語を除去する単語フィルタリング部を更に含み、前記モデル学習部は、前記単語フィルタリング部によって前記検索対象テキスト単語リストに含まれない単語を除去された前記学習用テキストペア群に基づいて、前記クエリ拡張モデルを学習するようにしても良い。

[0017] 上記目的を達成するために、本発明のテキスト検索装置は、検索対象の第１テキストと、前記検索対象の第１テキストを質問としたときの回答となる検索対象の第２テキストとのペアからなる検索対象テキストペア群から、入力クエリに対応する、第１テキスト及び第２テキストのペアを検索するテキスト検索装置であって、前記入力クエリに対して、拡張クエリを作成するためのクエリ拡張モデルに基づいて、前記第２テキストを検索するための拡張クエリを作成する拡張クエリ作成部と、前記入力クエリと前記拡張クエリとに基づいて、前記第１テキスト及び前記第２テキストのペアを検索するテキストペア検索部と、を含む。

[0018] なお、前記テキストペア検索部は、前記第１テキストについての検索用インデックスと、前記入力クエリと、前記第２テキストについての検索用インデックスと、前記拡張クエリとに基づいて、前記第１テキスト及び前記第２テキストのペアを検索するようにしても良い。

[0019] また、前記テキストペア検索部は、前記第１テキストについての検索用インデックスと、前記入力クエリとに基づいて、前記第１テキストの各々について、第１テキスト検索スコアを算出する第１テキスト検索スコア算出部と

、前記第２テキストについての検索用インデックスと、前記拡張クエリとに基づいて、前記第２テキストの各々について、第２テキスト検索スコアを算出する第２テキスト検索スコア算出部と、前記第１テキスト及び前記第２テキストのペアの各々について、前記第１テキスト検索スコアと第２テキスト検索スコアとを統合し、前記第１テキスト及び前記第２テキストのペアを検索する検索スコア統合結果出力部と、を含むようにしても良い。

[0020] 上記目的を達成するために、本発明のデータ構造は、検索対象の第１テキストと、前記検索対象の第１テキストを質問としたときの回答となる検索対象の第２テキストとのペアからなる検索対象テキストペア群から、入力クエリに対応する、第１テキスト及び前記第２テキストのペアを検索するテキスト検索装置で用いるためのデータ構造であって、学習用の第１テキストと、前記学習用の第１テキストを質問としたときの回答となる学習用の第２テキストとのペアからなる学習用テキストペア群、及び前記検索対象テキストペア群を入力として得られる、クエリとなるテキストに対して、拡張クエリとなるテキストを作成するクエリ拡張モデルと、前記検索対象の第１テキスト及び第２テキストについての検索用インデックスと、を含む。

[0021] 上記目的を達成するために、本発明のモデル作成方法は、モデル学習部を含んだモデル作成装置におけるモデル作成方法であって、前記モデル学習部が、学習用の第１テキストと、前記学習用の第１テキストを質問としたときの回答となる学習用の第２テキストとのペアからなる学習用テキストペア群、及び検索対象の第１テキストと、前記検索対象の第１テキストを質問としたときの回答となる検索対象の第２テキストとのペアからなる検索対象テキストペア群を入力として、クエリとなるテキストに対して、拡張クエリとなるテキストを作成するクエリ拡張モデルを学習するステップと、を含む。

[0022] 上記目的を達成するために、本発明のテキスト検索方法は、拡張クエリ作成部、及びテキストペア検索部を含み、第１テキストと、前記第１テキストを質問としたときの回答となる第２テキストとのペアからなる検索対象テキストペア群から、入力クエリに対応する、第１テキスト及び前記第２テキス

トのペアを検索するテキスト検索装置におけるテキスト検索方法であって、前記拡張クエリ作成部が、前記入力クエリに対して、拡張クエリを作成するための予め学習されたクエリ拡張モデルに基づいて、前記第２テキストを検索するための拡張クエリを作成するステップと、前記テキストペア検索部が、前記入力クエリと前記拡張クエリとに基づいて、前記第１テキスト及び前記第２テキストのペアを検索するステップと、を含む。

[0023] 上記目的を達成するために、本発明のプログラムは、コンピュータを、上記モデル作成装置又はテキスト検索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] 本発明によれば、入力されたクエリに対するテキストのペアの検索結果を精度良く得ることが可能となる。

図面の簡単な説明

[0025] [図1]実施形態に係るモデル作成装置の構成を示す機能ブロック図である。

[図2]実施形態に係るテキスト検索装置の構成を示す機能ブロック図である。

[図3]実施形態に係る検索用インデックスの一例を示す模式図である。

[図4]実施形態に係るニューラルネットワークを説明するための説明図である。

[図5]実施形態に係るモデル作成装置により実行されるモデル作成処理の流れを示すフローチャートである。

[図6]実施形態に係るモデル作成装置により実行される検索用インデックス作成処理の流れを示すフローチャートである。

[図7]実施形態に係るテキスト検索装置により実行される検索処理の流れを示すフローチャートである。

[図8]コンピュータのハードウェア構成の一例を示す図である。

発明を実施するための形態

[0026] 以下、本発明の実施形態について図面を用いて説明する。

[0027] 本実施形態に係る検索システムは、検索用のモデルを学習するモデル作成

装置、及び、検索を実行するテキスト検索装置を含んで構成される。

[0028] 図１に、モデル作成装置１０の構成を示すブロック図を示した。また、図２に、テキスト検索装置４０の構成を示すブロック図を示した。まず、モデル作成装置１０について説明し、次に、テキスト検索装置４０について説明する。

[0029] モデル作成装置１０では、学習用の第１テキストと、学習用の第１テキストを質問としたときの回答となる学習用の第２テキストとのペアで構成される学習用テキストペア群を入力として、入力クエリを拡張するためのクエリ拡張モデルを作成する。また、モデル作成装置１０は、検索対象の第１テキストと、検索対象の第１テキストを質問としたときの回答となる検索対象の第２テキストとのペアで構成される検索対象テキストペア群を入力として、第１テキスト検索用インデックス、及び、第２テキスト検索用インデックスを作成する。

[0030] 本実施形態では、学習用テキストペア群及び検索対象テキストペア群が入力されると、学習用テキストペア群及び検索対象テキストペア群について、まず後述する言語解析部１２及び言語解析部２２により形態素解析を行うと共に、入力されたキーワード抽出を行う。次に、学習用テキストペア群及び検索対象テキストペア群に出現する各単語についての重みを計算する。ここから、通常の全文検索エンジンと同様に、上記重みを使用した転置インデックスを作成する。また、ニューラルネットワークを用いたクエリ拡張モデルに入力するために学習用テキストペア群の整形を行う。そして、ニューラルネットワークを用いて学習用の第２テキストを検索するための拡張クエリを作成するためのクエリ拡張モデルを学習する。

[0031] 本実施形態では、学習用テキストペア群として、検索対象テキストペア群よりも大規模な学習用コーパスを使用する。例えば、Ｗｅｂの質問回答サイトのコーパスを用いる。質問回答サイトとは、ユーザが自然文で質問文を投稿すると、他のユーザがその質問文に対する回答文を投稿することができるサービスである。本実施形態では、質問文を学習用の第１テキストとし、回

答文を学習用の第２テキストとして学習用テキストペア群を作成し、学習用コーパスとして用いる。

[0032] また、本実施形態では、クエリ拡張モデルを学習する際、クエリ拡張モデルの出現単語を、検索対象テキストペア群に含まれている単語をリスト化した検索対象テキスト単語リストに含まれる単語に限定することで、クエリ拡張モデルの学習を効率的かつ高速に行うことが可能になる。

[0033] ただし、本実施形態に係るモデル作成装置をテキスト検索装置４０以外の装置で使用する場合、検索対象テキスト単語リストを用いた、クエリ拡張モデルの出現単語の限定処理は不要となる。

[0034] 図１に示すように、モデル作成装置１０は、言語解析部１２、単語重み計算部１４、拡張クエリ作成部１６、モデル学習部１８、クエリ拡張モデル記憶部２０、言語解析部２２、単語重み計算部２４、検索用インデックス作成部２６、第１テキスト検索用インデックス記憶部２８、及び、第２テキスト検索用インデックス記憶部３０を備えている。

[0035] 以降、それぞれの処理部について詳細に説明する。

[0036] 言語解析部１２及び言語解析部２２は、学習用テキストペア群及び検索対象テキストペア群に対して言語処理を適用し、キーワード抽出を行う。この際、各テキストペア群が日本語のように単語区切り無しで記述されている場合には、形態素解析を行うことによりテキストを単語単位に区切り、各テキストペア群が英語のように自明な単語区切りが含まれる言語で記述されている場合には、単語区切りにすることで、文を単語単位に区切る。なお、このとき、言語解析部１２及び言語解析部２２は、単語のステミングも行う。

[0037] 言語解析部１２及び言語解析部２２は、区切られた単語について名詞、動詞といった内容語の単語のみを抽出し、抽出した単語を検索で用いるキーワードとする。この際、連続する複数の単語によって１つの固有名詞が表示されるような場合には、これら複数の単語を接合する等の処理を行っても良い。

[0038] なお、言語解析部２２は、検索対象テキストペア群から抽出したキーワー

ドをリスト化し、検索対象テキスト単語リストとして記憶しておく。この検索対象テキスト単語リストは、単語フィルタリング部１６ｂによって単語をフィルタリングする際に使用される。

[0039] 単語重み計算部１４及び単語重み計算部２４は、後述する検索用インデックス作成部２６による検索用インデックスの作成、及び後述する単語並び替え部１６ａによる単語の並び替えで使用するために、抽出したキーワードの重要度を表す重みを計算する。重み計算には、情報検索でよく使用されるＩＤＦ値を用いる。単語ｗのＩＤＦ値ＩＤＦ（ｗ）は、下記（１）式により計算される。下記（１）におけるｄｆ（ｗ）は、単語ｗが出現するテキストの数であり、Ｎは、テキストの総数である。

[0040] [数1]

$$idf(w) = \log \frac{N}{df(w)} \quad \cdots (1)$$

[0041] なお、ＩＤＦ値と同様の性質を持つ計算式であれば、Ｏｋａｐｉ　ＢＭ２５等の上記以外の変形式を用いても良い。

[0042] 検索用インデックス作成部２６は、検索対象テキストペア群を検索するための検索用インデックスを作成する。検索用インデックスは、図３に示すように、テキストＩＤ、キーワード、及び当該キーワードの重みが、テキストＩＤ及び単語の組み合わせ毎に対応付けられて格納されたデータベースである。この際、重みは単語重み計算部２４により計算されたＩＤＦ値を用いてＴＦ／ＩＤＦで計算する。キーワードとする、テキストｄの単語ｗの重みは

、下記（２）式で表される。下記（２）式におけるｔｆ（ｄ，ｗ）は、テキストｄ中で単語ｗが出現する回数である。

[0043] [数2]

$$weight(d,w) = tf(d,w) \times idf(w) \quad \cdots (2)$$

[0044] なお、重みを計算したＴＦ／ＩＤＦ以外にも、Ｏｋａｐｉ　ＢＭ２５等の他の重み計算手法を用いて、重みを計算しても良い。

[0045] 拡張クエリ作成部１６は、単語並び替え部１６ａ、及び、単語フィルタリング部１６ｂを有している。

[0046] 単語並び替え部１６ａは、重みが計算された各単語を重みに応じて並び替える。ニューラルネットワークを用いたｅｎｃｏｄｅｒ－ｄｅｃｏｄｅｒモデルでは、出力時には単語や文字の系列を出力する。一般的な翻訳モデルでは、文法上正しい順番に文字や単語を出力できるように学習を行う。しかし、本実施形態では、クエリ拡張モデルによって出力された単語を検索クエリとして使用するため、文法上の並びは不必要である。

[0047] そこで、本実施形態では、クエリ拡張モデルによって出力された単語を検索において有効に活用するために、重要な単語順に単語が出力されるようにする。重要な単語順に出力されるようにするには、学習用テキストペア群に出現する単語を、形態素解析、単語区切り等を行った後の出現順の並びから、重要な単語から先頭に並び替えれば良い。

[0048] 単語の重要度は、上述した検索用インデックス作成部２６による処理と同

様に、ＴＦ／ＩＤＦによって計算することができる。例えば、「通信量が多いと帯域制限により通信速度が低下する場合があります」という文を形態素解析して並べると、「通信量　多い　帯域制　限　通信速度　低下　場合　ある」のような単語の並びになるが、これらの単語を単語の重み順で並び替えると、「帯域制限　通信速度　通信量　低下　場合　多い　ある」のような並び順になる。このように並び替えることで、クエリ拡張を行う際に、出力時にはより重要な単語が優先して出力されやすくなる。

[0049] 本実施形態では、ｄｅｃｏｄｅ時には単語の系列データを出力するため、単語並び替え部１６ａは、出力の学習データとなる学習用の第２テキストに対しては上述した単語の並び替えを必ず行う。一方、入力時には、単語の系列データ等における単語の語順を考慮しないｅｎｃｏｄｅを行うため、単語並び替え部１６ａは、学習用の第１テキストに関しては上述した単語の並び替えを必ずしも行わなくても良い。

[0050] しかしながら、学習用の第１テキストに入力する単語数等を制限する場合、先頭からｎ文字目で切ってしまうと、例えば日本語等では、「私」等の主語等、文の先頭に出現しやすい単語は常に学習される一方で、述語等、文の後半に出現しやすい重要な単語は学習され難い状況が生じる。そこで、本実施形態では、より重要な単語を学習で利用するために、学習用の第１テキストについても学習用の第２テキストと同様に単語の並び替えを行う。これにより、学習に有効な単語を学習で常に使用することができる。

[0051] 単語フィルタリング部１６ｂは、例えばｅｎｃｏｄｅｒ－ｄｅｃｏｄｅｒモデルを用いて学習用の第２テキストを検索するための検索クエリを作成する。学習用テキストペア群では、大量のテキストに様々な記載がなされている。学習用テキストペア群をそのまま用いてｅｎｃｏｄｅｒ－ｄｅｃｏｄｅｒモデルで学習した場合、ｄｅｃｏｄｅ時には様々な単語が出力されることになる。しかしながら、出力された単語を検索として使用することを考えると、検索対象テキストペア群に出現しない単語については、どれだけ作成してもヒットなしで使用されることがないため、作成する意味があまり無い。

そこで、ｄｅｃｏｄｅ時に出力される単語の語彙を、検索対象テキストペア群に含まれる語彙の範囲に限定することで、効率的かつ高速に学習を行うことが可能になる。

[0052] 単語フィルタリング部１６ｂでは、言語解析部２２で取得した検索対象テキスト単語リストと、単語並び替え部１６ａにより単語が並び替えられた学習用テキストペア群を照合し、検索対象テキスト単語リストに存在しない単語については、並び替えられた学習用テキストペア群から削除する。これは、出力側である学習用の第２テキストに対してのみ行う処理である。学習用の第１テキストについては、多様な入力を受け付けるため、単語のフィルタリングを行わない。

[0053] なお、クエリ拡張モデルを学習する際に、検索対象テキストペア群が確定していない場合等、検索対象テキストペア群を取得できない場合には、単語フィルタリング部１６ｂによる検索対象テキスト単語リストによる単語のフィルタリングをスキップして、汎用のクエリ拡張モデルを学習しても良い。

[0054] モデル学習部１８は、入力と出力との変換を行うモデルを学習する。本実施形態では、このモデルとして、ｅｎｃｏｄｅｒ－ｄｅｃｏｄｅｒモデルを用いる。ｅｎｃｏｄｅｒ－ｄｅｃｏｄｅｒモデルは、ニューラルネットワークを用いて、入力と出力との変換を学習できるモデルであり、ニューラルネットワークの学習には、学習用テキストペア群に含まれる第１テキスト及び第２テキストのペアをそのまま入力すれば良く、学習パラメータがニューラルネットワークによって自動学習されるという点が特徴となっている。

[0055] 例えば、下記参考文献１に記載されているように、「私　は　テニス　がしたい」という文を入力とし、「I want to play tennis」を出力として学習した場合には、学習用テキストペア群を入力するだけで自動翻訳器を作成することができる。

[0056] ［参考文献１］Ilya Sutskever, Oriol Vinyals, Quoc V. Le. Sequence to Sequence Learning with Neural Networks. 2013.

[0057] ｅｎｃｏｄｅｒ－ｄｅｃｏｄｅｒモデルは、入力の文字列を特徴ベクトル

に変換するＥＮＣＯＤＥ部、及びＥＮＣＯＤＥ部で変換された特徴ベクトルから出力文字列を作成するＤＥＣＯＤＥ部から構成される。

[0058] 一般的なｅｎｃｏｄｅｒ－ｄｅｃｏｄｅｒモデルでは、ＬＳＴＭという系列構造に強い活性関数を用いたニューラルネットワークを用いて、ＥＮＣＯＤＥ部もＤＥＣＯＤＥ部も構成するが、本実施形態では、ＥＮＣＯＤＥ部ではｓｉｇｍｏｉｄ等の通常の活性関数を用いたニューラルネットワークを用いる。これは、検索を想定した場合、入力クエリには、キーワード集合の場合と自然文の場合とのどちらの可能性も想定される。このように入力クエリのフォーマットが不定である場合には、単語の順番に大きく影響されやすいＬＳＴＭ等の系列モデルを用いるのはふさわしくない。そのため、本実施形態では、ｅｎｃｏｄｅｒ－ｄｅｃｏｄｅｒモデルではＥＮＣＯＤＥ部に系列モデルを使用しない。

[0059] 図４に、本実施形態で使用するｅｎｃｏｄｅｒ－ｄｅｃｏｄｅｒモデルを示した。図４に示すように、ＥＮＣＯＤＥ部は、Ｗ＿ｉｎ１～Ｗ＿ｉｎＮまでの単語ベクトル層、ＣＯＮＴＥＸＴ層、ＡＴＴＥＮＴＩＯＮ層から構成される。また、ＤＥＣＯＤＥ部は、単語の意味を表現した特徴ベクトルを占めるＥＭＢＥＤＥＤ層とＬＳＴＭ層から構成される。なお、ＤＥＣＯＤＥ部に入力される＜／Ｓ＞は文頭を意味する。

[0060] 単語ベクトル層では、単語をベクトル表現に変換したベクトルを用いる。このベクトルとして、該当する要素（単語）を１とし、他の要素（単語）を０にする１－ｈｏｔ型のベクトル、Ｗｏｒｄ２ｖｅｃ（登録商標）等により事前に学習した単語ベクトル等を用いても良い。

[0061] ＣＯＮＴＥＸＴ層では、全単語ベクトルの総和ベクトルの総和が入力となる。また、ＡＴＴＥＮＴＩＯＮ層では、下記参考文献２に示すＧｌｏｂａｌ ａｔｔｅｎｔｉｏｎと同様の計算を行うことによって出力を決定するが、本実施形態のＥＮＣＯＤＥ部にはＬＳＴＭのＨＩＤＤＥＮ層が存在しないため、単語ベクトルの出力をＨＩＤＤＥＮ層の出力として扱う。

[0062] ［参考文献２］Minh-Thang Luong, Hieu Pham, Christopher D. Manning.

Effective Approaches to Attention-based Neural Machine Translation. 2015.

[0063] ＤＥＣＯＤＥ部は、上記参考文献１及び２と同様にＬＳＴＭベースの作成モデルとなっている。この際、ニューラルネットワークの各層のユニット数、学習のための最適化手法、及びエラー関数の設定方法については、本実施形態においては特に指定せず、適用する学習用テキストペア群の規模、使用言語等を考慮して適宜設定できるものとする。また、ニューラルネットワークモデルについても本実施形態で示したものは最小構成であるため、ＣＯＮＴＥＸＴ層等のニューラルネットワークの各層を多段化する等の変形を行っても良い。

[0064] モデル学習部１８は、学習したクエリ拡張モデルを、クエリ拡張モデル記憶部２０に記憶させる。

[0065] 検索用インデックス作成部２６は、検索対象テキストペア群を用いて、検索対象の第１テキストを検索するための第１テキスト検索用インデックスと、検索対象の第２テキストを検索するための第２テキスト検索用インデックスと、の２種類のインデックスを作成する。この際、検索用インデックス作成部２６は、検索対象の第１テキストのみから、第１テキスト検索用インデックスを作成する。また、検索用インデックス作成部２６は、第２テキスト検索用インデックスを作成する際には、検索対象の第２テキストのみから第２テキスト検索用インデックスを作成しても良いし、検索対象の第１テキストと検索対象の第２テキストとを結合して１つにまとめたテキストから第２テキスト検索用インデックスを作成しても良い。

[0066] 検索用インデックス作成部２６は、作成した第１テキスト検索用インデックスを第１テキスト検索用インデックス記憶部２８に記憶させる。また、検索用インデックス作成部２６は、作成した第２テキスト検索用インデックスを第２テキスト検索用インデックス記憶部３０に記憶させる。

[0067] このように、検索用インデックス作成部２６により、検索対象テキストペア群から第１テキスト検索用インデックス及び第２テキスト検索用インデッ

クスが作成されると共に、モデル学習部１８により、学習用テキストペア群から第２テキスト検索用のクエリ拡張モデルが学習される。

[0068] ここで、従来の情報検索技術では、第１テキストと第２テキストとを結合し、１つの文書とみなして検索を実行していた。一方、本実施形態に係るテキスト検索装置４０は、第１テキストの検索と第２テキストの検索とで異なるクエリで別々に検索を実行することに特徴がある。

[0069] テキスト検索装置４０は、検索対象の第１テキストと、検索対象の第１テキストを質問としたときの回答となる検索対象の第２テキストとのペアからなる検索対象テキストペア群から、入力クエリに対応する、検索対象の第１テキスト及び検索対象の第２テキストのペアを検索する。この際、入力クエリに対して、拡張クエリを生成するための予め学習されたクエリ拡張モデルに基づいて、検索対象の第２テキストを検索するための拡張クエリを作成する。また、入力クエリと拡張クエリとに基づいて、検索対象の第１テキスト及び検索対象の第２テキストのペアを検索する。

[0070] 図２に示すように、テキスト検索装置４０は、言語解析部４２、テキストペア検索部４３、及び、拡張クエリ作成部４６を含んで構成される。また、テキストペア検索部４３は、第１テキスト検索スコア算出部４４、第２テキスト検索スコア算出部４８、及び、検索スコア統合結果出力部５０を有している。

[0071] 以降、それぞれの処理部について詳細に説明する。

[0072] 言語解析部４２は、モデル作成装置１０の言語解析部１２、２２と同様の処理を行う。

[0073] 第１テキスト検索スコア算出部４４は、入力クエリと検索対象テキストペア群の第１テキストとを比較した比較結果として検索スコアを算出する。第１テキスト検索スコアの算出には、モデル作成装置１０で作成し、第１テキスト検索用インデックス記憶部２８に記憶されている第１テキスト検索用インデックスを用いる。

[0074] 第１テキスト検索スコアは、第１テキスト検索用インデックスに格納され

ている重みの総和で計算できる。第１テキストｄと入力クエリＱとの第１テキスト検索スコアｓｃｏｒｅ１は、下記（３）式で表される。下記（３）式におけるｑは、入力クエリＱ中に含まれるキーワードを示している。ｗｅｉｇｈｔ（ｄ１，ｑ）は、第１テキスト検索用インデックスに格納されている、テキストｄ、単語ｑの重みを表す重み値である。

[0075] [数3]

$$score_1(d_1, Q) = \sum_{q \in Q} weight(d_1, q)$$

…（３）

[0076] 上記（３）式により、入力クエリとより多く、かつ重要なキーワードが一致している第１テキストほど、第１テキスト検索スコアは高い値となる。なお、第１テキスト検索スコアの算出は、一般的なキーワード一致検索を行うものであるため、単語の意味的な類似度を用いたクエリ拡張、単語の意味ベクトルを用いたスコア算出手法等と組み合わせて使用しても良い。

[0077] 拡張クエリ作成部４６は、入力クエリを、クエリ拡張モデル記憶部２０に記憶されているクエリ拡張モデルに入力することにより、キーワード拡張を行い、第２テキスト検索用の拡張クエリを作成する。なお、クエリ拡張モデルには、言語解析部４２で抽出したキーワードをそのまま入力すれば良い。ただし、入力クエリが長文である場合には、モデル作成装置１０の単語並び替え部１６ａと同様に、単語の重みを予め計算しておき、単語の重みに基づいて、抽出されたキーワードから重み値が大きい上位ｎ語の単語のみを抽出してクエリ拡張モデルに入力しても良い。なお、ｎは、ユーザ等によって予

め設定される。ｎとしては、例えば、数語～数十語程度が好ましい。

[0078] クエリ拡張モデルでは、０～Ｎ個のキーワードが動的に出力される。このとき、出力されたＮ個のキーワード全てを拡張クエリに使用してもよいし、出力数が多い場合には任意のｎ番目までに出力されたキーワードを用いても良い。クエリ拡張モデルは、出力時に、検索において重要と思われる順番にキーワードが出力されるように学習されるため、拡張キーワード数を制限する場合には、クエリ拡張モデルが出力した順番に則って使用すれば良い。

[0079] なお、第２テキスト検索スコアを算出する際には、入力クエリのキーワード群に、拡張クエリ作成部４６で出力された拡張キーワード群を追加したキーワード群を拡張クエリとして使用する。

[0080] 第２テキスト検索スコア算出部４８は、拡張クエリ作成部４６により作成された拡張クエリと、モデル作成装置１０の検索用インデックス作成部２６により作成され、第２テキスト検索用インデックス記憶部３０に記憶されている第２テキスト検索用インデックスを用いて、第２テキスト検索スコアを算出する。

[0081] 第２テキスト検索スコアの算出には、第１テキスト検索スコアを算出する際に用いた上記（３）式に加えて、近接重みを考慮する。近接重みとは、拡張クエリ中のキーワードが第２テキスト中のキーワードにヒットした場合、その他のヒットしたキーワードが第２テキスト中のどの程度近くに存在しているかということを考慮した指標である。

[0082] 第２テキストは、第１テキストと比べて長文で記述されている場合が多い。また、ヒットしたキーワードが第２テキストの複数の文に亘って点在している場合よりも、より少ない文中のキーワードに密集してヒットしている方が有用な情報である可能性が高い。そのため、第２テキストの検索では、ヒットしたキーワード間の位置の近さを示す近接重みを導入する。

[0083] 近接重みは、ヒットしたキーワード間の距離の平均によって計算する。第２テキストをｄとした場合に、単語ｑがヒットしたときの近接重みは、下記（４）式に従って計算される。下記（４）式におけるＨは、第２テキストｄ

にヒットした入力クエリのキーワード集合であり、N_Hは、キーワード集合の全キーワード数であり、Ｌは、第２テキストｄの先頭からのキーワードの位置を示している。

[0084] [数4]

$$prox(d_2,q)=\frac{1}{N_H}\sqrt{\sum_{h\in H}(L(d_2,q)-L(d_2,h))^2}$$

…（４）

[0085] 例えば、「通信量多い　帯域制限　通信速度　低下　場合　ある」という第２テキストｄがあるとする。この場合、Ｌ（ｄ２，帯域制限）＝３となり、Ｌ（ｄ２，低下）＝５となる。このように近接重みを用いると、第２テキスト検索スコアは、下記（５）式に従って計算される。

[0086] [数5]

$$score_2(d_2,Q)=\sum_{q\in Q}weight(d_2,q)\times prox'(d_2,q)$$

…（５）

[0087] ここで、上記（５）式におけるｐｒｏｘ’（ｄ２，ｑ）は、０～１に正規化

された近接重み値である。近接重みはキーワード間の距離にもとづいて計算されるため、値の範囲が不定である。そのため、重み係数として使用するために値が０～１の範囲に限定されるように正規化を行う必要がある。正規化手法についてはｓｉｇｍｏｉｄ関数を用いたもの等、任意の正規化手法を使用して良い。また、第２テキストが長文でない場合等には、近接計算を導入した検索スコア計算方法ではなく、第１テキスト検索スコアと同様の計算式を用いても良い。

[0088] 検索スコア統合結果出力部５０は、第１テキスト検索スコアと第２テキスト検索スコアとを統合し、統合スコアを算出し、統合スコアの降順に検索結果を出力する。本実施形態では、統合スコアを、第１テキスト検索スコアと第２テキスト検索スコアとの線形和として計算する。入力クエリＱに対して、第１テキストをd_1とし、第２テキストをd_2とした場合のテキストペア群ＰＤの統合スコアは、下記（６）式に従って計算される。下記（６）式におけるw_1及びw_2は、線形和のための重み係数である。

[0089] [数6]

$$score(PD, Q) = w_1 \times score(d_1, Q) + w_2 \times score(d_2, Q)$$

…（６）

[0090] 検索対象テキストペア群によって第１テキストと第２テキストとの考慮する比率は異なる。例えば、ＦＡＱ検索では、入力クエリＱの部分に質問内容が明確に記載されている場合には、第１テキストであるＱをより考慮した検索を行うべきである。一方、入力クエリＱには「～について」等の簡潔な記

載しかなく、第２テキストであるＡの部分に豊富な記載がある場合には、第２テキストであるＡの第２テキスト検索スコアを優先したほうが良い結果が得られる。また、重み係数に関しては、検索対象テキストペア群の性質を観察し、人手で付与しても良いし、機械学習、統計処理等に基づいて自動的に付与しても良い。

[0091] なお、本実施形態に係るモデル作成装置１０及びテキスト検索装置４０は、例えば、図８に示すようなコンピュータ１００で構成される。図８に示すコンピュータ１００は、入力装置１０１、表示装置１０２、外部Ｉ／Ｆ１０３、ＲＡＭ（Random Access Memory）１０４、ＲＯＭ（Read Only Memory）１０５、ＣＰＵ（Central Processing Unit）１０６、通信Ｉ／Ｆ１０７、及び補助記憶装置１０８を備えている。これらの各ハードウェアはバスＢによって接続されている。なお、コンピュータ１００は、入力装置１０１及び表示装置１０２のうちの少なくとも一方を備えていなくても良い。

[0092] 本実施形態は、ＣＰＵ１０６が、ハードディスク等の補助記憶装置１０８やＲＯＭ１０５に記憶されているプログラムを読み出して実行することにより、上記の各ハードウェア資源とプログラムとが協働し、上述した機能が実現される。なお、当該プログラムは、例えばＣＤ－ＲＯＭ等の記録媒体１０３ａに格納されていても良い。

[0093] 本実施形態に係るモデル作成装置１０によるモデル作成処理の流れを、図５に示すフローチャートを用いて説明する。本実施形態では、モデル作成装置１０に、モデル作成処理の実行を開始するための予め定めた情報が入力されたタイミングでモデル作成処理が開始されるが、モデル作成処理が開始されるタイミングはこれに限らない。

[0094] ステップＳ１０１では、言語解析部１２が、学習用テキストペア群を入力する。

[0095] ステップＳ１０３では、言語解析部１２が、入力した学習用テキストペア群に含まれる各テキストについて形態素分解を行い、単語を抽出する。なお、このとき、言語解析部１２は、抽出した単語のステミングを行う。

[0096] ステップＳ１０５では、単語重み計算部１４が、抽出した各単語について、重みを計算する。

[0097] ステップＳ１０７では、単語並び替え部１６ａが、重みを計算した各単語を、重みに基づいて並び替える。

[0098] ステップＳ１０９では、単語フィルタリング部１６ｂが、並び替えられた各単語のうち、後述する検索用インデックス作成処理で作成される検索対象テキスト単語リストに含まれる単語を削除することにより、単語をフィルタリングする。

[0099] ステップＳ１１１では、モデル学習部１８が、各単語が並び替えられると共に各単語がフィルタリングされた学習用テキストペア群を用いてクエリ拡張モデルを学習する。

[0100] ステップＳ１１３では、学習したクエリ拡張モデルをクエリ拡張モデル記憶部２０に記憶させ、本モデル作成処理のプログラムの実行を終了する。

[0101] 次に、本実施形態に係るモデル作成装置１０による検索用インデックス作成処理の流れを、図６に示すフローチャートを用いて説明する。本実施形態では、モデル作成装置１０に、検索用インデックス作成処理の実行を開始するための予め定めた情報が入力されたタイミングで検索用インデックス作成処理が開始されるが、検索用インデックス作成処理が開始されるタイミングはこれに限らない。

[0102] ステップＳ２０１では、言語解析部２２が、検索対象テキストペア群を入力する。

[0103] ステップＳ２０３では、言語解析部２２が、入力した検索対象テキストペア群に含まれる各テキストについて形態素分解し、単語を抽出する。また、言語解析部２２が、検索対象テキスト単語リストを作成する。なお、このとき、言語解析部２２は、抽出した単語のステミングを行う。

[0104] ステップＳ２０５では、単語重み計算部２４が、抽出された各単語の重みを計算する。

[0105] ステップＳ２０７では、検索用インデックス作成部２６が、抽出された各

単語の重みに基づいて、第１テキスト検索用インデックス及び第２テキスト検索用インデックスを作成する。

[0106] ステップＳ２０９では、検索用インデックス作成部２６が、作成した第１テキスト検索用インデックスを第１テキスト検索用インデックス記憶部２８に記憶させると共に、作成した第２テキスト検索用インデックスを第２テキスト検索用インデックス記憶部３０に記憶させ、本検索用インデックス作成処理のプログラムの実行を終了する。

[0107] 次に本実施形態に係るテキスト検索装置４０による検索処理の流れを、図７に示すフローチャートを用いて説明する。本実施形態では、テキスト検索装置４０に、検索処理の実行を開始するための予め定めた情報が入力されたタイミングで検索処理が開始されるが、検索処理が開始されるタイミングはこれに限らない。

[0108] ステップＳ３０１では、言語解析部４２が、ユーザにより入力された入力クエリを入力する。

[0109] ステップＳ３０３では、言語解析部４２が、入力クエリを形態素分解し、単語を抽出する。

[0110] ステップＳ３０５では、第１テキスト検索スコア算出部４４が、第１テキスト検索用インデックス記憶部２８から第１テキスト検索用インデックスを読み出し、入力クエリと、読み出した第１テキスト検索用インデックスに基づいて、第１テキスト検索スコアを算出する。

[0111] ステップＳ３０７では、拡張クエリ作成部４６が、クエリ拡張モデル記憶部２０に記憶されているクエリ拡張モデルを読出し、入力クエリを、クエリ拡張モデルを用いて拡張し、第２テキスト検索用の拡張クエリを作成する。

[0112] ステップＳ３０９では、第２テキスト検索スコア算出部４８が、第２テキスト検索用インデックス記憶部３０から第２テキスト検索用インデックスを読み出し、拡張クエリと、読み出した第２テキスト検索用インデックスと、に基づいて、第２テキスト検索スコアを算出する。

[0113] ステップＳ３１１では、検索スコア統合結果出力部５０が、検索対象の第

１テキストと第２テキストとのペアの各々について、第１テキスト検索スコアと第２テキスト検索スコアとを統合して統合スコアを算出する。

[0114] ステップＳ３１３では、検索スコア統合結果出力部５０が、算出した統合スコアに基づいて、検索対象の第１テキストと第２テキストとのペアの検索結果を出力し、本検索処理のプログラムの実行を終了する。

[0115] このようにして、本実施形態では、学習用の第１テキストと、学習用の第１テキストを質問としたときの回答となる学習用の第２テキストとのペアからなる学習用テキストペア群、及び検索対象の第１テキストと、検索対象の第１テキストを質問としたときの回答となる検索対象の第２テキストとのペアからなる検索対象テキストペア群を入力として、クエリとなるテキストに対して、拡張クエリとなるテキストを作成するクエリ拡張モデルを学習する。

[0116] また、本実施形態では、第１テキストと、第１テキストを質問としたときの回答となる第２テキストとのペアからなる検索対象テキストペア群から、入力クエリに対応する、第１テキスト及び第２テキストのペアを検索する際、入力クエリに対して、拡張クエリを作成するための予め学習されたクエリ拡張モデルに基づいて、第２テキストを検索するための拡張クエリを作成し、入力クエリと拡張クエリとに基づいて、第１テキスト及び第２テキストのペアを検索する。

[0117] なお、本実施形態では、図１及び図２に示す機能の構成要素の動作をプログラムとして構築し、モデル作成装置１０及びテキスト検索装置４０として利用されるコンピュータにインストールして実行させるが、これに限らず、ネットワークを介して流通させても良い。

[0118] また、構築されたプログラムをハードディスク、ＣＤ－ＲＯＭ等の可搬記憶媒体に格納し、コンピュータにインストールしたり、配布したりしても良い。

[0119] 本願は、日本国に２０１６年１１月２５日に出願された基礎出願２０１６－２２９０７２号に基づくものであり、その全内容はここに参照をもって援

用される。

符号の説明

[0120] １０　モデル作成装置

１２、２２、４２　言語解析部

１４、２４　単語重み計算部

１６　拡張クエリ作成部

１６ａ　単語並び替え部

１６ｂ　単語フィルタリング部

１８　モデル学習部

２０　クエリ拡張モデル記憶部

２６　検索用インデックス作成部

２８　第１テキスト検索用インデックス記憶部

３０　第２テキスト検索用インデックス記憶部

４０　テキスト検索装置

４３　テキストペア検索部

４４　第１テキスト検索スコア算出部

４６　拡張クエリ作成部

４８　第２テキスト検索スコア算出部

５０　検索スコア統合結果出力部

請求の範囲

[請求項1] 学習用の第１テキストと、前記学習用の第１テキストを質問としたときの回答となる学習用の第２テキストとのペアからなる学習用テキストペア群を入力として、クエリとなるテキストに対して、拡張クエリとなるテキストを作成するクエリ拡張モデルを学習するモデル学習部

を含むモデル作成装置。

[請求項2] 検索対象の第１テキストと、前記検索対象の第１テキストを質問としたときの回答となる検索対象の第２テキストとのペアからなる検索対象テキストペア群に基づいて、前記検索対象の第１テキストについての検索用インデックス、及び前記検索対象の第２テキストについての検索用インデックスを作成する検索インデックス作成部

を更に含む請求項１記載のモデル作成装置。

[請求項3] 検索対象の第１テキストと、前記検索対象の第１テキストを質問としたときの回答となる検索対象の第２テキストとのペアからなる検索対象テキストペア群に含まれる各単語からなる検索対象テキスト単語リストを用いて、前記学習用テキストペア群から、前記検索対象テキスト単語リストに含まれない単語を除去する単語フィルタリング部を更に含み、

前記モデル学習部は、前記単語フィルタリング部によって前記検索対象テキスト単語リストに含まれない単語を除去された前記学習用テキストペア群に基づいて、前記クエリ拡張モデルを学習する

請求項１又は２記載のモデル作成装置。

[請求項4] 検索対象の第１テキストと、前記検索対象の第１テキストを質問としたときの回答となる検索対象の第２テキストとのペアからなる検索対象テキストペア群から、入力クエリに対応する、第１テキスト及び第２テキストのペアを検索するテキスト検索装置であって、

前記入力クエリに対して、拡張クエリを作成するためのクエリ拡張

モデルに基づいて、前記第２テキストを検索するための拡張クエリを作成する拡張クエリ作成部と、

前記入力クエリと前記拡張クエリとに基づいて、前記第１テキスト及び前記第２テキストのペアを検索するテキストペア検索部と、

を含むテキスト検索装置。

[請求項5] 前記テキストペア検索部は、前記第１テキストについての検索用インデックスと、前記入力クエリと、前記第２テキストについての検索用インデックスと、前記拡張クエリとに基づいて、前記第１テキスト及び前記第２テキストのペアを検索する

請求項４記載のテキスト検索装置。

[請求項6] 前記テキストペア検索部は、

前記第１テキストについての検索用インデックスと、前記入力クエリとに基づいて、前記第１テキストの各々について、第１テキスト検索スコアを算出する第１テキスト検索スコア算出部と、

前記第２テキストについての検索用インデックスと、前記拡張クエリとに基づいて、前記第２テキストの各々について、第２テキスト検索スコアを算出する第２テキスト検索スコア算出部と、

前記第１テキスト及び前記第２テキストのペアの各々について、前記第１テキスト検索スコアと第２テキスト検索スコアとを統合し、前記第１テキスト及び前記第２テキストのペアを検索する検索スコア統合結果出力部と、

を含む請求項５記載のテキスト検索装置。

[請求項7] 検索対象の第１テキストと、前記検索対象の第１テキストを質問としたときの回答となる検索対象の第２テキストとのペアからなる検索対象テキストペア群から、入力クエリに対応する、第１テキスト及び前記第２テキストのペアを検索するテキスト検索装置で用いるためのデータ構造であって、

学習用の第１テキストと、前記学習用の第１テキストを質問とした

ときの回答となる学習用の第２テキストとのペアからなる学習用テキストペア群、及び前記検索対象テキストペア群を入力として得られる、

クエリとなるテキストに対して、拡張クエリとなるテキストを作成するクエリ拡張モデルと、

前記検索対象の第１テキスト及び第２テキストについての検索用インデックスと、

を含むデータ構造。

[請求項8] モデル学習部を含んだモデル作成装置におけるモデル作成方法であって、

前記モデル学習部が、学習用の第１テキストと、前記学習用の第１テキストを質問としたときの回答となる学習用の第２テキストとのペアからなる学習用テキストペア群、及び検索対象の第１テキストと、前記検索対象の第１テキストを質問としたときの回答となる検索対象の第２テキストとのペアからなる検索対象テキストペア群を入力として、クエリとなるテキストに対して、拡張クエリとなるテキストを作成するクエリ拡張モデルを学習するステップ

を含むモデル作成方法。

[請求項9] 拡張クエリ作成部、及びテキストペア検索部を含み、第１テキストと、前記第１テキストを質問としたときの回答となる第２テキストとのペアからなる検索対象テキストペア群から、入力クエリに対応する、第１テキスト及び前記第２テキストのペアを検索するテキスト検索装置におけるテキスト検索方法であって、

前記拡張クエリ作成部が、前記入力クエリに対して、拡張クエリを作成するための予め学習されたクエリ拡張モデルに基づいて、前記第２テキストを検索するための拡張クエリを作成するステップと、

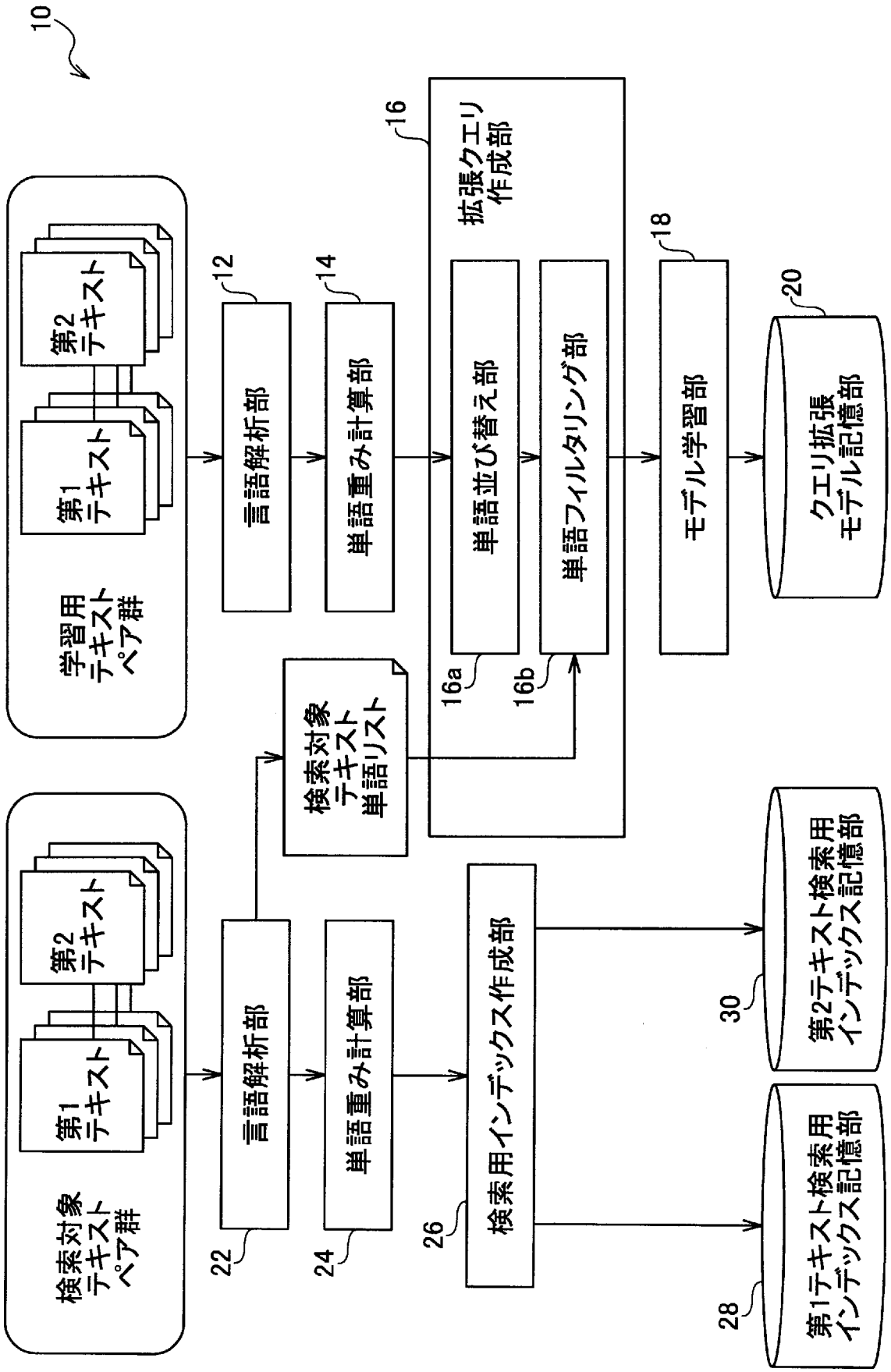
前記テキストペア検索部が、前記入力クエリと前記拡張クエリとに基づいて、前記第１テキスト及び前記第２テキストのペアを検索する

ステップと、

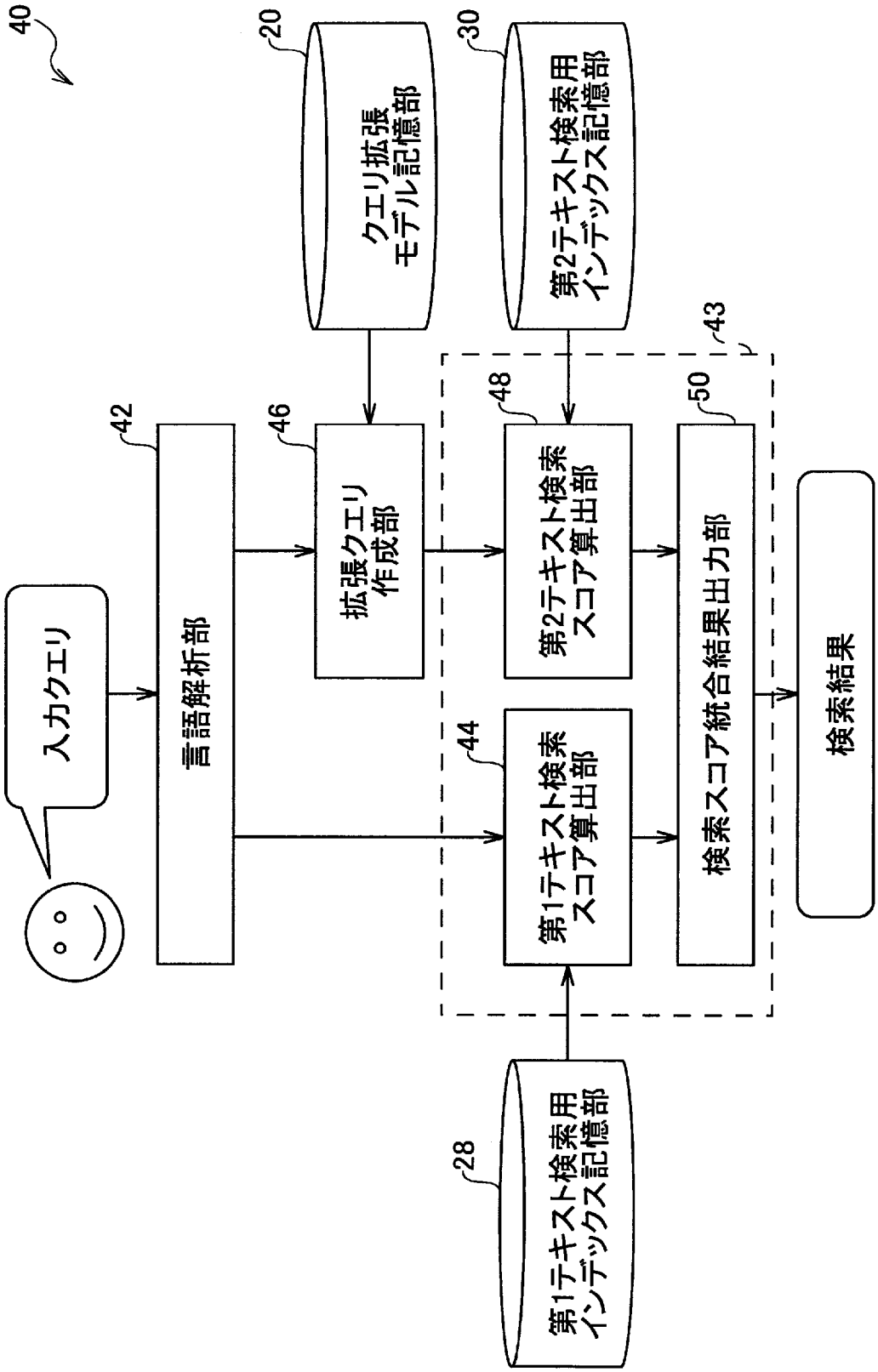
を含むテキスト検索方法。

[請求項10] コンピュータを、請求項１～請求項３の何れか１項記載のモデル作成装置の各部、又は請求項４～請求項６の何れか１項記載のテキスト検索装置の各部として機能させるためのプログラム。

[図1]


10
学習用テキストペア群
第1テキスト
第2テキスト
言語解析部
12
単語重み計算部
14
拡張クエリ作成部
16
単語並び替え部
16a
単語フィルタリング部
16b
モデル学習部
18
クエリ拡張モデル記憶部
20
検索対象テキスト単語リスト
検索対象テキストペア群
第1テキスト
第2テキスト
言語解析部
22
単語重み計算部
24
検索用インデックス作成部
26
第2テキスト検索用インデックス記憶部
30
第1テキスト検索用インデックス記憶部
28

[図2]

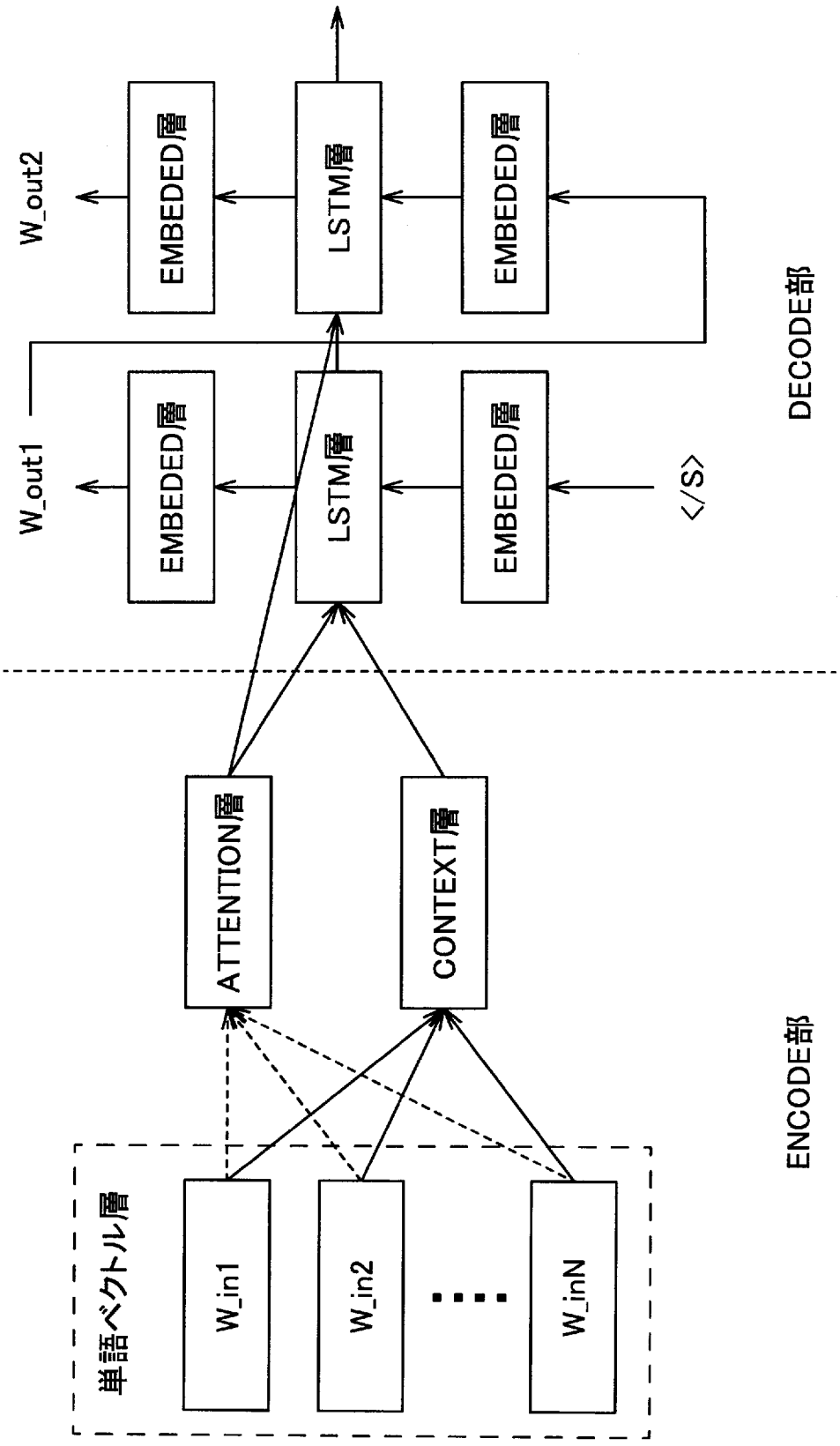

40
20
クエリ拡張
モデル記憶部
30
第2テキスト検索用
インデックス記憶部
43
42
46
48
50
入力クエリ
言語解析部
拡張クエリ
作成部
第2テキスト検索
スコア算出部
検索スコア統合結果出力部
検索結果
44
第1テキスト検索
スコア算出部
28
第1テキスト検索用
インデックス記憶部

[図3]

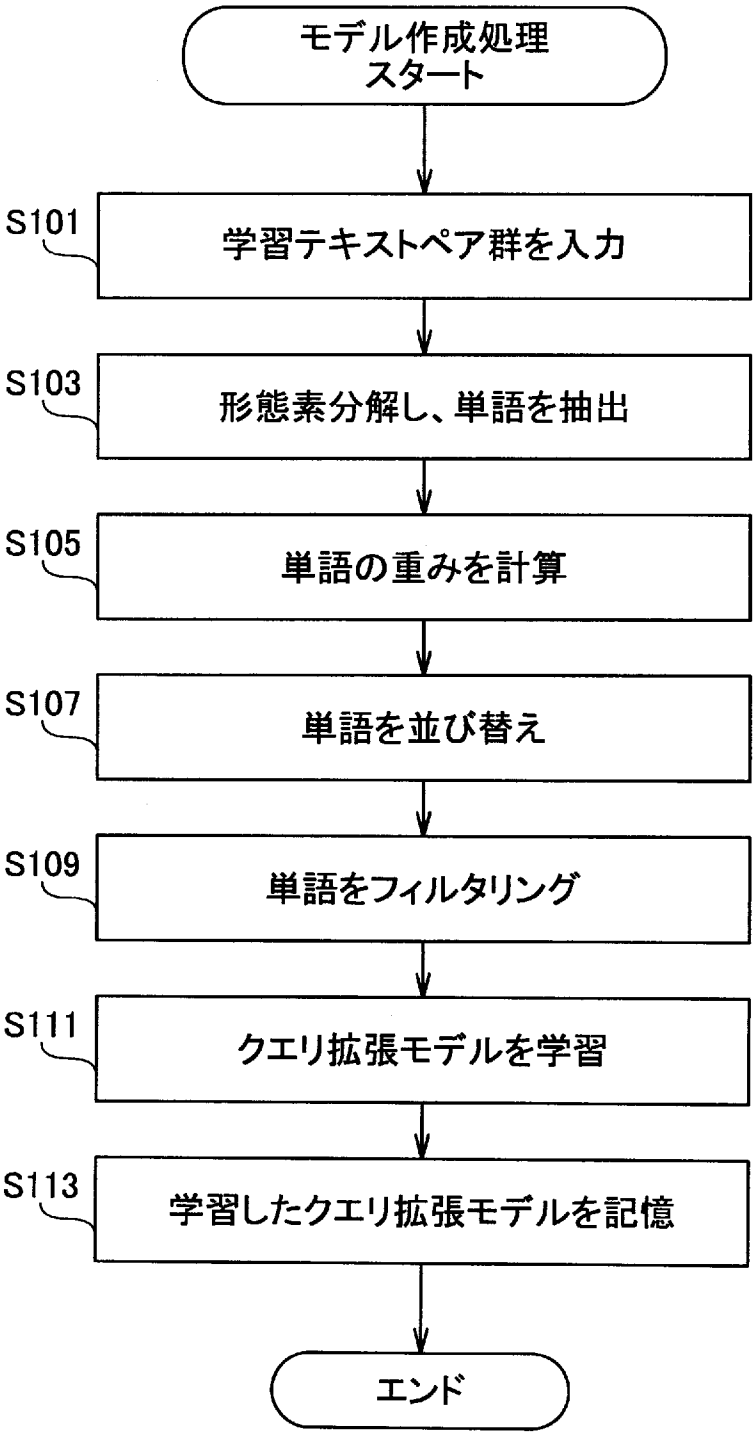
28, 30

テキスト ID	キーワード	重み
01A	帯域制限	2.51
01A	速度	1.22
01A	利用規約	3.43
01B	同意	1.82

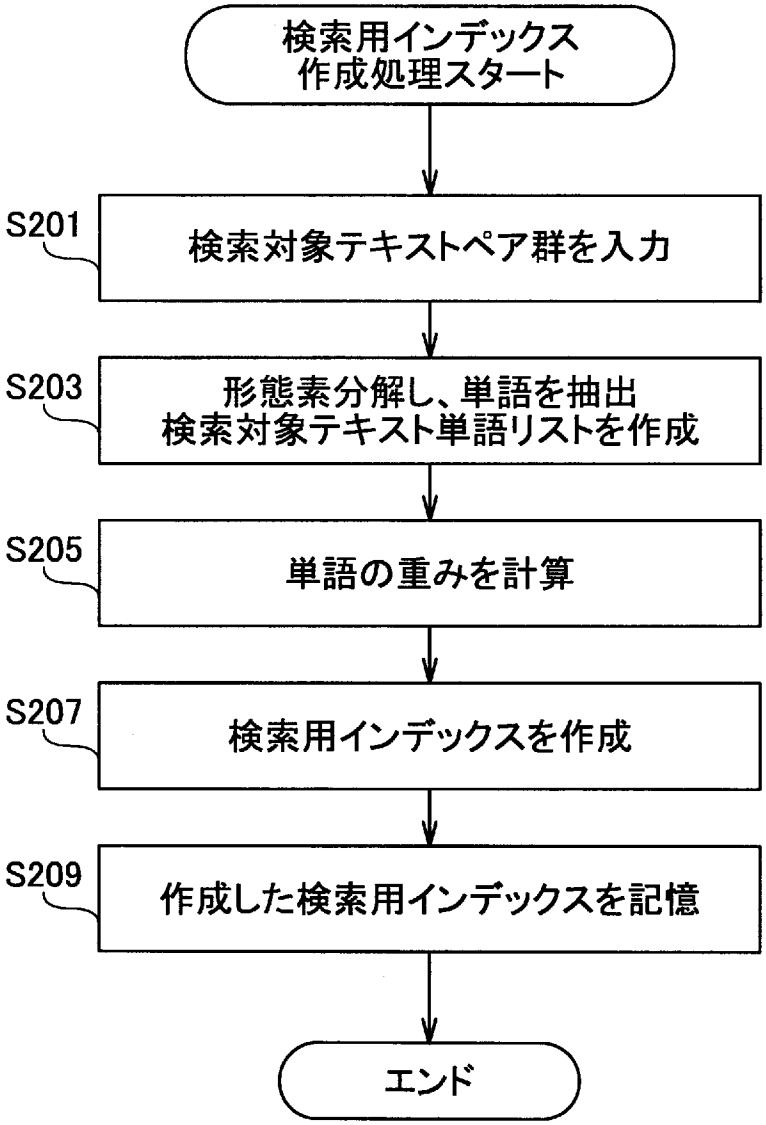
[図4]


W_out2
EMBEDED層
LSTM層
EMBEDED層
W_out1
EMBEDED層
LSTM層
EMBEDED層
</S>
DECODE部
ATTENTION層
CONTEXT層
単語ベクトル層
W_in1
W_in2
W_inN
ENCODE部

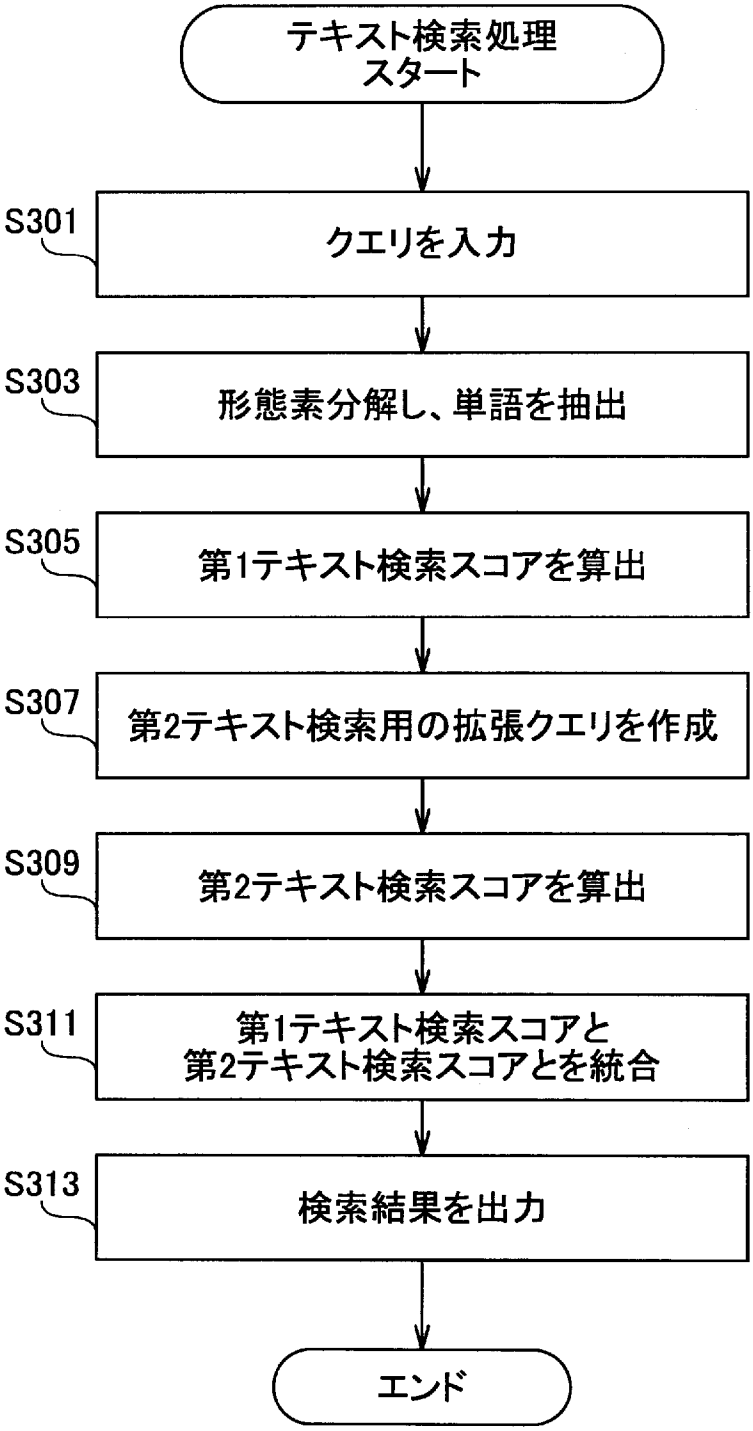
[図5]


モデル作成処理
スタート
S101
学習テキストペア群を入力
S103
形態素分解し、単語を抽出
S105
単語の重みを計算
S107
単語を並び替え
S109
単語をフィルタリング
S111
クエリ拡張モデルを学習
S113
学習したクエリ拡張モデルを記憶
エンド

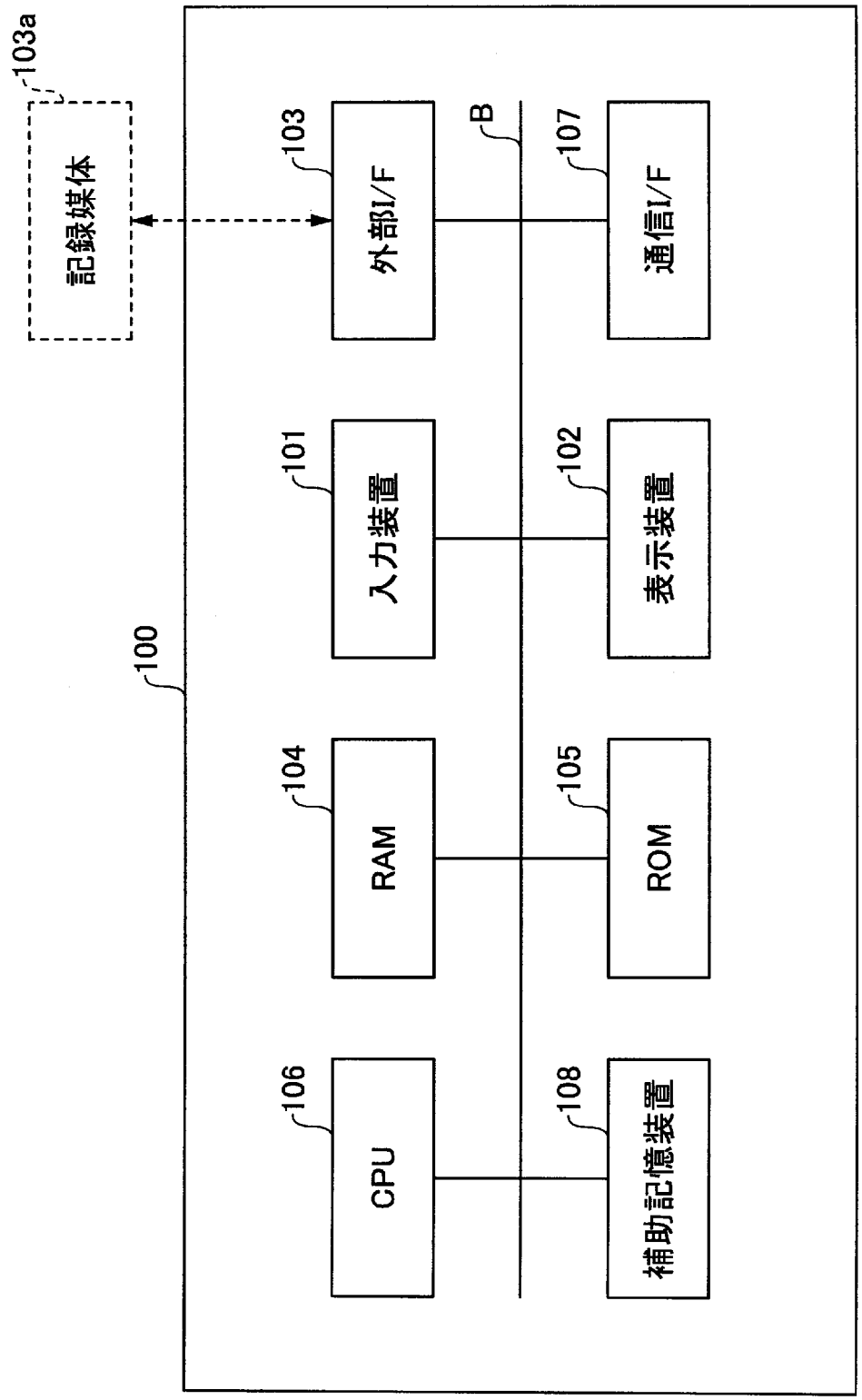
[図6]


検索用インデックス
作成処理スタート
S201
検索対象テキストペア群を入力
S203
形態素分解し、単語を抽出
検索対象テキスト単語リストを作成
S205
単語の重みを計算
S207
検索用インデックスを作成
S209
作成した検索用インデックスを記憶
エンド

[図7]


テキスト検索処理
スタート
S301
クエリを入力
S303
形態素分解し、単語を抽出
S305
第1テキスト検索スコアを算出
S307
第2テキスト検索用の拡張クエリを作成
S309
第2テキスト検索スコアを算出
S311
第1テキスト検索スコアと
第2テキスト検索スコアとを統合
S313
検索結果を出力
エンド

[図8]


103a
記録媒体
100
103
外部I/F
B
107
通信I/F
101
入力装置
102
表示装置
104
RAM
105
ROM
106
CPU
108
補助記憶装置

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2017/041630

A. CLASSIFICATION OF SUBJECT MATTER
Int. Cl. G06F17/30(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int. Cl. G06F 17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Published examined utility model applications of Japan 1922-1996
Published unexamined utility model applications of Japan 1971-2018
Registered utility model specifications of Japan 1996-2018
Published registered utility model applications of Japan 1994-2018

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	JP 2007-304793 A (NIPPON TELEGRAPH AND TELEPHONE CORP.) 22 November 2007, paragraphs [0007], [0013] (Family: none)	1-2, 4-10 3
Y	JP 2000-339314 A (NIPPON TELEGRAPH AND TELEPHONE CORP.) 08 December 2000, paragraphs [0086]-[0088] (Family: none)	1-2, 4-10
Y	JP 2016-66232 A (OKWAVE CO., LTD.) 28 April 2016, paragraphs [0051]-[0052] (Family: none)	2, 5-6

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
Name and mailing address of the ISA/ Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2017/041630

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2016/0147775 A1 (ORACLE INTERNATIONAL CORPORATION) 26 May 2016, paragraph [0038] & WO 2016/081170 A1 & CN 106796608 A	3

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１７／０４１６３０

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１８年
日本国実用新案登録公報	１９９６－２０１８年
日本国登録実用新案公報	１９９４－２０１８年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
Y A	JP 2007-304793 A（日本電信電話株式会社）2007.11.22, 段落[0007]、[0013]（ファミリーなし）	1-2, 4-10 3
Y	JP 2000-339314 A（日本電信電話株式会社）2000.12.08, 段落[0086]-[0088]（ファミリーなし）	1-2, 4-10
Y	JP 2016-66232 A（株式会社オウケイウェイヴ）2016.04.28, 段落[0051]-[0052]（ファミリーなし）	2, 5-6

☑　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー

「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの

「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「Ｏ」口頭による開示、使用、展示等に言及する文献

「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献

「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
０９．０２．２０１８	２０．０２．２０１８

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｍ　５３７７
樋口　龍弥
電話番号　０３－３５８１－１１０１　内線　３５９９

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	US 2016/0147775 A1 (ORACLE INTERNATIONAL CORPORATION) 2016.05.26, 段落[0038] & WO 2016/081170 A1 & CN 106796608 A	3