(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6165657号
(P6165657)

(45)発行日　平成29年7月19日(2017.7.19)　(24)登録日　平成29年6月30日(2017.6.30)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 170A
G06F 17/30 350C
G06F 17/30 210A

請求項の数 10　(全 23 頁)

(21)出願番号　特願2014-58246(P2014-58246)
(22)出願日　平成26年3月20日(2014.3.20)
(65)公開番号　特開2015-184749(P2015-184749A)
(43)公開日　平成27年10月22日(2015.10.22)
審査請求日　平成28年8月30日(2016.8.30)

(73)特許権者　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(74)代理人　110002147
特許業務法人酒井国際特許事務所
(72)発明者　中田　康太
東京都港区芝浦一丁目１番１号　株式会社東芝内
(72)発明者　蟻生　政秀
東京都港区芝浦一丁目１番１号　株式会社東芝内

審査官　樋口　龍弥

最終頁に続く

(54)【発明の名称】情報処理装置、情報処理方法およびプログラム

(57)【特許請求の範囲】
【請求項１】
複数の候補文書から言語モデルの学習に用いる文書を選択する情報処理装置であって、
前記言語モデルが利用される目的に合致した目的文書について、それぞれのトピックに対する文書の関連の強さを表すトピック特徴量を算出する第１特徴量算出部と、
前記複数の候補文書のそれぞれについて、前記トピック特徴量を算出する第２特徴量算出部と、
前記複数の候補文書のそれぞれの前記トピック特徴量について、前記目的文書の前記トピック特徴量との類似度を算出する類似度算出部と、
前記類似度が基準値より大きい候補文書を、前記言語モデルの学習に用いる文書として選択する選択部と、
を備える情報処理装置。
【請求項２】
トピック毎に、単語と、前記単語のトピックとの関連の強さを表すスコアとのペアの集合を含むトピック情報を取得するトピック情報取得部をさらに備え、
前記第１特徴量算出部および前記第２特徴量算出部は、前記トピック情報に基づき、前記トピック特徴量を算出する
請求項１に記載の情報処理装置。
【請求項３】
前記第１特徴量算出部および前記第２特徴量算出部は、トピック毎に、対象の文書に含

まれる単語のスコアを累積して、前記トピック特徴量を算出する
請求項２に記載の情報処理装置。
【請求項４】
選択された前記候補文書に基づき、前記言語モデルを学習する学習部
をさらに備える請求項１に記載の情報処理装置。
【請求項５】
前記トピック情報取得部は、前記複数の候補文書を用いて前記トピック情報を生成する
請求項２に記載の情報処理装置。
【請求項６】
前記トピック情報取得部は、異なるトピック数の複数の前記トピック情報を生成し、生成した複数の前記トピック情報に基づき、前記目的文書の複数の前記トピック特徴量を算出し、算出した複数の前記トピック特徴量に基づき、生成した複数の前記トピック情報のうちの１つの前記トピック情報を選択する
請求項５に記載の情報処理装置。
【請求項７】
前記トピック情報取得部は、品詞群毎に前記トピック情報を生成し、
前記第１特徴量算出部および前記第２特徴量算出部は、前記品詞群毎の前記トピック情報に基づき、前記品詞群毎の前記トピック特徴量を算出する
請求項５に記載の情報処理装置。
【請求項８】
前記目的文書と内容が異なり前記言語モデルの学習の基準となる
学習対象の言語モデルと類似した用途で用いられる言語モデルを学習するための類似目的文書に対する、品詞群毎の前記トピック特徴量を算出する第３特徴量算出部をさらに備え、
前記類似度算出部は、
前記複数の候補文書のそれぞれの第１の品詞群に関する前記トピック特徴量に対して、前記目的文書の前記第１の品詞群に関する前記トピック特徴量との第１の類似度を算出し、
前記複数の候補文書のそれぞれの第２の品詞群に関する前記トピック特徴量に対して、前記類似目的文書の前記第２の品詞群に関する前記トピック特徴量との第２の類似度を算出し、
前記選択部は、前記第１の類似度が第１の基準値より大きく、且つ、前記第２の類似度が第２の基準値より大きい候補文書を、前記言語モデルの学習に用いる文書として選択する
請求項７に記載の情報処理装置。
【請求項９】
複数の候補文書から言語モデルの学習に用いる文書を選択する情報処理方法であって、
前記言語モデルが利用される目的に合致した目的文書について、それぞれのトピックに対する文書の関連の強さを表すトピック特徴量を算出する第１特徴量算出ステップと、
前記複数の候補文書のそれぞれについて、前記トピック特徴量を算出する第２特徴量算出ステップと、
前記複数の候補文書のそれぞれの前記トピック特徴量について、前記目的文書の前記トピック特徴量との類似度を算出する類似度算出ステップと、
前記類似度が基準値より大きい候補文書を、前記言語モデルの学習に用いる文書として選択する選択ステップと、
を実行する情報処理方法。
【請求項１０】
コンピュータを、複数の候補文書から言語モデルの学習に用いる文書を選択する情報処理装置として機能させるためのプログラムであって、
前記情報処理装置は、

前記言語モデルが利用される目的に合致した目的文書について、それぞれのトピックに対する文書の関連の強さを表すトピック特徴量を算出する第１特徴量算出部と、
前記複数の候補文書のそれぞれについて、前記トピック特徴量を算出する第２特徴量算出部と、
前記複数の候補文書のそれぞれの前記トピック特徴量について、前記目的文書の前記トピック特徴量との類似度を算出する類似度算出部と、
前記類似度が基準値より大きい候補文書を、前記言語モデルの学習に用いる文書として選択する選択部と、
を備えるプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明の実施形態は、情報処理装置、情報処理方法およびプログラムに関する。
【背景技術】
【０００２】
コンピュータおよびインターネット環境の普及により、大量の文書が電子化され蓄積されている。このような電子化された大量の文書を用いて、音声認識等の技術に利用される言語モデルを学習することができる。例えばウェッブ上で公開されている大量の文書を用いて、一般的な用途に利用される言語モデルを学習することにより、その言語モデルの性能を向上させることができる。しかし、ある特定の目的に利用される言語モデルをウェッブ上で公開されている大量の文書を用いて学習しても、特定の目的以外に関する文書が多量に含まれるので、性能を大幅に向上させることはできない。
【０００３】
ある特定の目的に利用される言語モデルの性能を向上させるには、特定の目的に関する文書（目的文書）のみを用いて言語モデルを学習すればよい。例えば、特定の目的がコールセンターにおける音声認識である場合、コールセンターにおけるオペレータのやり取りの音声を書き起こした文書を用いて言語モデルを学習すれば、その特定の目的に利用される言語モデルの性能を向上させることができる。
【０００４】
ところで、このような方法は、十分な量の目的文書を用いて学習しなければ、多様な表現に対応した言語モデルとすることができない。しかし、特定の目的に関する文書を数多く収集することは困難である。例えば、音声を書き起こして文書化する作業は、経済的および時間的なコストが大きく、十分な量の目的文書を得ることは困難である。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】特開２００９－２３８２３５号公報
【発明の概要】
【発明が解決しようとする課題】
【０００６】
発明が解決しようとする課題は、言語モデルを学習するために用いられる文書を、目的外の文書を大量に含む複数の候補文書から適切に選択することにある。
【課題を解決するための手段】
【０００７】
実施形態の情報処理装置は、複数の候補文書から言語モデルの学習に用いる文書を選択する。前記情報処理装置は、第１特徴量算出部と、第２特徴量算出部と、類似度算出部と、選択部と、を備える。前記第１特徴量算出部は、前記言語モデルが利用される目的に合致した目的文書について、それぞれのトピックに対する文書の関連の強さを表すトピック特徴量を算出する。前記第２特徴量算出部は、前記複数の候補文書のそれぞれについて、前記トピック特徴量を算出する。前記類似度算出部は、前記複数の候補文書のそれぞれの

前記トピック特徴量について、前記目的文書の前記トピック特徴量との類似度を算出する。前記選択部は、前記類似度が基準値より大きい候補文書を、前記言語モデルの学習に用いる文書として選択する。

【図面の簡単な説明】

【０００８】

【図１】第１実施形態に係る情報処理装置の構成を示す図。
【図２】トピック数が５０個のトピック情報の一例を示す図。
【図３】第１実施形態に係る情報処理装置の処理フローを示す図。
【図４】目的文書の第１例を示す図。
【図５】候補文書の第１例を示す図。
【図６】候補文書の第２例を示す図。
【図７】候補文書の第３例を示す図。
【図８】トピック特徴量の算出フローを示す図。
【図９】単語の一致度の高い文書の一例を示す図。
【図１０】トピック数が１０個のトピック情報の一例を示す図。
【図１１】トピック数が２００個のトピック情報の一例を示す図。
【図１２】トピック情報を選択するための処理フローを示す図。
【図１３】第２変形例に係るトピック情報の一例を示す図。
【図１４】第２実施形態に係る情報処理装置の構成を示す図。
【図１５】第２実施形態に係る情報処理装置の処理フローを示す図。
【図１６】目的文書の第２例を示す図。
【図１７】類似目的文書の一例を示す図。
【図１８】第１の品詞群のトピック情報の一例を示す図。
【図１９】第２の品詞群のトピック情報の一例を示す図。
【図２０】情報処理装置のハードウェア構成を示す図。

【発明を実施するための形態】

【０００９】

（第１の実施形態）

図１は、第１実施形態に係る情報処理装置１０の構成を示す図である。図２は、トピック数が５０個のトピック情報の一例を示す図である。

【００１０】

情報処理装置１０は、ウェッブ上等の複数の候補文書から言語モデルの学習に用いる文書を選択し、選択した候補文書を用いて言語モデルを学習する。情報処理装置１０は、目的文書格納部２１と、候補コーパス格納部２２と、トピック情報取得部２３と、第１特徴量算出部２４と、第２特徴量算出部２５と、類似度算出部２６と、選択部２７と、学習部２８とを備える。

【００１１】

目的文書格納部２１は、学習対象の言語モデルが利用される目的に合致した文書（目的文書）を格納する。目的文書は、一例として、ユーザにより手動で選択される。学習対象の言語モデルがコールセンターにおける音声認識に利用される場合には、目的文書は、一例として、コールセンターにおけるオペレータの音声を書き起こしたテキストである。

【００１２】

候補コーパス格納部２２は、言語モデルの学習に用いる文書の候補となる複数の文書（候補文書）を格納する。複数の候補文書は、一例として、ウェッブから収集した大量のテキストである。複数の候補文書には、例えば、ニュースサイトの記事、および、掲示板上に書き込まれたコメント等の、多様な目的で用いられる文書が含まれ、言語モデルが利用される目的以外で用いられる文書も含まれる。候補コーパス格納部２２は、情報処理装置１０内に設けられるのではなく、ネットワーク上のサーバに設けられていてもよいし、複数のサーバに分散して設けられていてもよい。

【００１３】

トピック情報取得部２３は、トピック情報を取得する。トピック情報は、図２に示すような、トピック毎に、単語とスコアとのペアの集合を含む。

【００１４】

トピックとは、文書で述べられている中心的な対象（テーマ）およびその文書の発話のスタイル等の特徴をいう。１つの文書に複数のトピックが含まれていてもよい。例えば、図２のトピック番号＃１は、デジタル家庭電化製品のトピックを表す。また、図２のトピック番号＃２は、食品に関するトピックを表す。さらに、トピック情報は、例えば、丁寧な発話スタイルを表すトピック、および、書き言葉のスタイル（書く場合に用いるスタイル）を表すトピックを含んでもよい。

【００１５】

トピック情報におけるそれぞれのトピックに属する単語は、そのトピックに関連する単語であって、そのトピックに関する文書に含まれる可能性がある。また、トピック情報に含まれるそれぞれの単語は、スコアとペアとなっている。スコアは、その単語が属するトピックとの関連の強さを表す。本実施形態においては、スコアは、大きいほど、対するトピックとの関連が強いことを表す。

【００１６】

なお、トピック情報は、１つの単語が、複数のトピックに属していてもよい。また、トピック情報に含まれるトピックの数は、何個であってもよい。

【００１７】

トピック情報は、一例として、ユーザが複数のトピックを設定し、ユーザがそれぞれのトピックに関する単語を収集することにより、生成される。また、トピック情報は、一例として、ユーザが複数のトピックを設定し、ユーザがトピック毎に関連する文書とを準備し、コンピュータが準備した複数の文書内の単語の頻度を算出することにより、生成される。

【００１８】

また、トピック情報取得部２３は、例えば、下記の文献に記載されているような教師無しトピック分析技術により、トピック情報を自動で生成してもよい。

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): P. 993-1022.

【００１９】

この方法では、まず、ユーザがトピック数Ｎを設定する。そして、トピック情報取得部２３は、大量で多様な文書を解析して、Ｎ個のトピックに分割されたトピック情報を生成する。この方法によれば、トピック情報取得部２３は、トピックに関する事前の知識を用いずにトピック情報を生成することができる。

【００２０】

第１特徴量算出部２４は、トピック情報に基づいて、目的文書格納部２１に格納された目的文書に対するトピック特徴量を算出する。トピック特徴量は、それぞれのトピックに対する、その文書の関連の強さを表す。本実施形態では、トピック特徴量は、下記の数１に示されるような、ベクトル（配列）により表される。

【数１】

$$\vec{T}(t) = (T_1, T_2, ..., T_{49}, T_{50}) = (0.74, 0.03, ..., 0.06, 0.65)$$

【００２１】

ベクトルで表されたトピック特徴量は、トピック情報に含まれるトピックの数分の成分（例えば、T_1，T_2，…，T_{49}，T_{50}）を含む。トピック特徴量に含まれるそれぞれの成分は、トピック情報に含まれるそれぞれのトピックに一対一で対応する。それぞれの成分は、対応するトピックに対する、その文書の関連の強さを表す。例えば、数１の成分T_1は、図２に示すトピック情報におけるトピック番号＃１のトピックに対する、文書の関連の強さを表す。

【００２２】
このようなトピック特徴量は、その文書のトピック毎の関連の強さの分布を表している。なお、トピック特徴量のより詳細な算出方法については、後述の図８を参照して説明する。
【００２３】
第２特徴量算出部２５は、トピック情報に基づいて、候補コーパス格納部２２に格納されたそれぞれの候補文書に対するトピック特徴量を算出する。なお、候補文書に対するトピック特徴量は、目的文書に対するトピック特徴量と、同一の形式であり、同一の算出方法で算出される。
【００２４】
類似度算出部２６は、複数の候補文書のそれぞれのトピック特徴量に対する、目的文書のトピック特徴量との類似度を算出する。すなわち、類似度算出部２６は、複数の候補文書のそれぞれのトピック毎の関連の強さの分布が、目的文書のトピック毎の関連の強さの分布とどれだけ類似しているかを算出する。
【００２５】
本実施形態においては、類似度算出部２６は、ベクトルで表されたトピック特徴量の内積を演算することにより、類似度を算出する。すなわち、類似度算出部２６は、候補文書に対するトピック特徴量に含まれるそれぞれの成分と、目的文書に対するトピック特徴量の対応する成分とを乗算し、乗算結果を全て加算した値を類似度として算出する。
【００２６】
選択部２７は、複数の候補文書のうち、類似度が基準値より大きい候補文書を、言語モデルの学習に用いる文書として選択する。ここで、基準値は、ユーザが設定した値であってもよい。また、基準値は、複数の候補文書の類似度に基づき算出された値であってもよい。例えば、基準値は、複数の候補文書の類似度の平均値、または、複数の候補文書の類似度の最大値から一定量小さい値等であってもよい。
【００２７】
学習部２８は、選択部２７により選択された候補文書に基づき、言語モデルを学習する。学習部２８は、一例として、ｎグラム言語モデルを一般的な公知技術を用いて学習する。
【００２８】
図３は、第１実施形態に係る情報処理装置１０の処理フローを示す図である。なお、本フローでは、家庭電化製品のメーカーのコールセンターの音声認識に用いられる言語モデルを学習する例を説明する。また、本フローでは、図２で示したトピック情報を用いる例を説明する。
【００２９】
処理に先立って、予めユーザにより目的文書が目的文書格納部２１に格納される。目的文書格納部２１は、一例として、図４に示されるような、テレビジョン受像機（テレビとも称する。）のリモートコントローラ（リモコンとも称する。）についての問い合わせに対する応答音声を書き起こしたテキストを、目的文書として格納する。
【００３０】
また、処理に先立って、情報処理装置１０は、ウェッブ等から複数の候補文書を取得し、候補コーパス格納部２２に格納する。候補コーパス格納部２２は、一例として、図５、図６および図７に示されるような、候補文書を格納する。なお、図５に示される候補文書Ｃ＿｛ｎ１｝は、家庭電化製品のメーカーのコールセンターに対する、ＤＶＤレコーダーについての問い合わせ音声を書き起こしたテキストである。図６に示される候補文書Ｃ＿｛ｎ２｝は、テレビの調子がおかしいとのウェッブ上での書き込みのテキストである。図７に示される候補文書Ｃ＿｛ｎ３｝は、食品メーカーのコールセンターに対する、アレルギー源に対する問い合わせの音声を書き起こしたテキストである。
【００３１】
まず、ステップＳ１１において、トピック情報取得部２３は、トピック情報を生成する

。トピック情報取得部２３は、予め保存されているトピック情報を取得してもよい。

【００３２】

続いて、ステップＳ１２において、第１特徴量算出部２４は、トピック毎に、目的文書に含まれる単語のスコアを累積して、目的文書のトピック特徴量を算出する。具体的には、第１特徴量算出部２４は、図８のステップＳ２１からステップＳ２９に示す手順で、目的文書のトピック特徴量を算出する。

【００３３】

図８のステップＳ２１において、第１特徴量算出部２４は、トピック特徴量を初期化する。本例においては、下記の数２に示すように、トピック特徴量に含まれる全ての成分を０．０に初期化する。

【数２】

$$\vec{T}(t)=(T_1,T_2,...,T_{49},T_{50})=(0.0,0.0,...,0.0,0.0)$$

【００３４】

続いて、第１特徴量算出部２４は、対象の文書に含まれる全ての単語毎に、ステップＳ２３からステップＳ２７までの処理を繰り返して実行する（ステップＳ２２とステップＳ２８との間のループ処理）。第１特徴量算出部２４は、一例として、対象の文書の先頭の単語から最後の単語まで、１つずつ単語を選択して、ステップＳ２３からステップＳ２７の処理を実行する。

【００３５】

単語毎のループ処理において、第１特徴量算出部２４は、さらに、トピック情報に示されたトピック毎に、ステップＳ２４からステップＳ２６の処理を繰り返して実行する（ステップＳ２３とステップＳ２７との間のループ処理）。第１特徴量算出部２４は、一例として、トピック情報のトピック番号＃１からトピック番号＃５０まで順次にトピックを選択して、ステップＳ２４からステップＳ２６の処理を実行する。

【００３６】

トピック毎のループ処理において、まず、ステップＳ２４において、第１特徴量算出部２４は、選択した単語が、トピック情報における対象のトピックの単語の集合に含まれるか否かを判断する。含まれない場合には（ステップＳ２４のＮｏ）、第１特徴量算出部２４は、処理をステップＳ２７に進める。含まれる場合には（ステップＳ２４のＹｅｓ）、第１特徴量算出部２４は、処理をステップＳ２５に遷移させる。

【００３７】

ステップＳ２５において、第１特徴量算出部２４は、トピック情報における対象のトピックの単語の集合から、選択した単語に対応する（ペアとなる）スコアを取得する。続いて、ステップＳ２６において、第１特徴量算出部２４は、取得したスコアにより、トピック特徴量の対応する成分を更新する。第１特徴量算出部２４は、一例として、トピック特徴量の対応する成分に、取得したスコアを加算する。

【００３８】

例えば、ループ処理の対象の単語が「テレビ」であり、ループ処理の対象のトピックがトピック番号＃１であるとする。この場合、トピック番号＃１の単語の集合の中の「テレビ」が存在する。従って、第１特徴量算出部２４は、トピック番号＃１の「テレビ」に対応するスコア（０．１１）を、トピック特徴量の１番目の成分T_1に加算する。下記の数３は、初期化したトピック特徴量に対して、「テレビ」に対応するスコア（０．１１）を加算した後のトピック特徴量である。

【数３】

$$\vec{T}(t)=(T_1,T_2,...,T_{49},T_{50})=(0.11,0.0,...,0.0,0.0)$$

【００３９】

第１特徴量算出部２４は、ステップＳ２６の処理が終了すると、処理をステップＳ２７に進める。ステップＳ２７において、全てのトピックについて、まだ、ステップＳ２４からステップＳ２６の処理を終了していない場合には、第１特徴量算出部２４は、処理をステップＳ２３に戻して次のトピックについて処理を繰り返す。終了した場合には、第１特徴量算出部２４は、処理をステップＳ２８に進める。

【００４０】

ステップＳ２８において、全ての単語について、まだ、ステップＳ２３からステップＳ２７の処理を終了していない場合には、第１特徴量算出部２４は、処理をステップＳ２２に戻して、次の単語について処理を繰り返す。終了した場合には、第１特徴量算出部２４は、処理をステップＳ２９に進める。

【００４１】

下記の数４は、全ての単語についての更新処理が完了した後のトピック特徴量である。本例では、目的文書にトピック番号＃１に属する単語が多く含まれているので、T_1の値が他の成分より大きくなっている。

【数４】

$$\vec{T}(t)=(T_1,T_2,...,T_{49},T_{50})=(2.5,0.1,...,0.2,2.2)$$

【００４２】

ステップＳ２９において、第１特徴量算出部２４は、トピック特徴量を正規化する。本例では、下記の数５に示される演算によりトピック特徴量を正規化する。すなわち、第１特徴量算出部２４は、それぞれの成分T_iを、全ての成分の二乗平均で除算することにより、トピック特徴量を正規化する。

【数５】

$$T_i=\frac{T_i}{\sqrt{\sum_{i=1}^{50}T_i^2}}$$

【００４３】

下記の数６は、目的文書に対する正規化後のトピック特徴量を示す。

【００４４】

【数６】

$$\vec{T}(t)=(T_1,T_2,...,T_{49},T_{50})=(0.74,0.03,...,0.06,0.65)$$

本例において、正規化後のトピック特徴量は、各成分の二乗和が１となる。このように正規化することにより、トピック特徴量は、対象の文書が何れのトピックと関連性が強いのかを表すことができる。なお、数６のトピック特徴量は、成分T_3～T_{48}までは０．０である。従って、本実施形態において、目的文書は、トピック番号＃１とトピック番号＃５０のトピックとの関連性が強い。

【００４５】

第１特徴量算出部２４は、以上のように目的文書に対するトピック特徴量を算出する。

【００４６】

図３に戻る。続いて、情報処理装置１０は、候補コーパス格納部２２に格納されている候補文書毎に、ステップＳ１４からステップＳ１７の処理を繰り返して実行する（ステップＳ１３とステップＳ１８との間のループ処理）。

【００４７】

候補文書毎のループ処理では、まず、ステップＳ１４において、第２特徴量算出部２５は、トピック毎に、対象の文書に含まれる単語のスコアを累積して、候補文書のトピック特徴量を算出する。具体的には、第２特徴量算出部２５は、図８のステップＳ２１からス

テップＳ２９に示した手順で、候補文書のトピック特徴量を算出する。

【００４８】

下記の数７は、候補文書Ｃ＿｛ｎ１｝、候補文書Ｃ＿｛ｎ２｝および候補文書Ｃ＿｛ｎ３｝に対するトピック特徴量を示す。

【数７】

$$\vec{T}(c_{n1}) = (0.70, 0.01, ..., 0.04, 0.70)$$

$$\vec{T}(c_{n2}) = (0.71, 0.02, ..., 0.69, 0.02)$$

$$\vec{T}(c_{n3}) = (0.01, 0.68, ..., 0.09, 0.68)$$

【００４９】

なお、数７に示すトピック特徴量は、成分Ｔ$_{3}$～Ｔ$_{48}$までは０．０である。候補文書Ｃ＿｛ｎ１｝は、トピック番号＃１およびトピック番号＃５０のトピックとの関連性が強い。候補文書Ｃ＿｛ｎ２｝は、トピック番号＃１およびトピック番号＃４９のトピックとの関連性が強い。候補文書Ｃ＿｛ｎ３｝は、トピック番号＃２およびトピック番号＃５０のトピックとの関連性が強い。

【００５０】

続いて、ステップＳ１５において、類似度算出部２６は、目的文書のトピック特徴量と候補文書のトピック特徴量との類似度を算出する。本実施形態においては、類似度算出部２６は、下記の数８に示されるように、目的文書のトピック特徴量と、候補文書のトピック特徴量との内積を演算する。

【数８】

$$sim(t, c_j) = \vec{T}(t) \cdot \vec{T}(c_j)$$

【００５１】

下記の数９は、候補文書Ｃ＿｛ｎ１｝、候補文書Ｃ＿｛ｎ２｝および候補文書Ｃ＿｛ｎ３｝に対する類似度を示す。

【数９】

$$sim(t, c_{n1}) = 0.74*0.70 + 0.03*0.01 + 0.06*0.04 + 0.65*0.70 = 0.98$$

$$sim(t, c_{n2}) = 0.74*0.71 + 0.03*0.02 + 0.06*0.69 + 0.65*0.02 = 0.58$$

$$sim(t, c_{n3}) = 0.74*0.01 + 0.03*0.68 + 0.06*0.09 + 0.65*0.68 = 0.48$$

【００５２】

候補文書Ｃ＿｛ｎ１｝の類似度は、０．９８となる。候補文書Ｃ＿｛ｎ２｝の類似度は、０．５８となる。候補文書Ｃ＿｛ｎ３｝の類似度は、０．４８となる。目的文書および候補文書Ｃ＿｛ｎ１｝は、共に、トピック番号＃１およびトピック番号＃５０のトピックとの関連性が強いので、類似度が他よりも高くなっている。

【００５３】

続いて、ステップＳ１６において、選択部２７は、類似度が基準値より大きいか否かを判断する。類似度が基準値以下である場合には（ステップＳ１６のＮｏ）、選択部２７は、処理をステップＳ１８に進める。類似度が基準値より大きい場合には（ステップＳ１６のＹｅｓ）、選択部２７は、処理をステップＳ１７に進める。

【００５４】

ステップＳ１７において、選択部２７は、対象の候補文書を、言語モデルの学習に用いる文書として選択する。本例においては、選択部２７は、基準値が０．７０に設定されており、類似度が０．７０より大きい候補文書Ｃ＿｛ｎ１｝を選択する。そして、選択部２

７は、処理をステップＳ１８に進める。

【００５５】

ステップＳ１８において、全ての候補文書について、まだステップＳ１４からステップＳ１７の処理を終了していない場合には、選択部２７は、処理をステップＳ１３に戻して、次の候補文書について処理を繰り返す。終了した場合には、選択部２７は、処理をステップＳ１９に進める。

【００５６】

ステップＳ１９において、学習部２８は、選択された候補文書を用いて、言語モデルを学習する。そして、ステップＳ１９の処理を終えると、情報処理装置１０は、本フローを終了する。

【００５７】

以上のように、本実施形態に係る情報処理装置１０によれば、目的外の文書を大量に含む複数の候補文書から、言語モデルを学習するために適切な文書を効率良く選択することができる。特に、情報処理装置１０によれば、目的文書に含まれる単語と一致する単語が比較的に少ない候補文書であっても、トピックの分布が類似していれば、言語モデルを学習するために用いる文書として選択することができる。

【００５８】

例えば、図４に示す目的文書と、図５に示す候補文書Ｃ＿｛ｎ１｝とを比較すると、含まれる単語の多くが異なっており、単語毎の一致度は低い。しかし、例えば、図４に示す目的文書の「テレビ」と図５に示す候補文書Ｃ＿｛ｎ１｝の「ＤＶＤ」とは、両者ともデジタル家庭電化製品に関連する単語として認識されるので、人間の感覚では類似すると判断される。情報処理装置１０は、このような候補文書Ｃ＿｛ｎ１｝を選択する。

【００５９】

また、単語の一致度の高い文書は、ほとんどが同一の単語を用いたテキストで構成される可能性がある。例えば、図９は、図４に示す目的文書と、単語の一致度の高い候補文書の一例を示す図である。図９の候補文書は、目的文書とほぼ同様の表現で構成された文書となっている。従って、図９に示すような候補文書を用いて言語モデルを学習したとしても、多様な表現に対して脆弱な言語モデルとなってしまう。

【００６０】

情報処理装置１０は、目的文書および候補文書のトピック特徴量を比較して類似度を判断する。従って、情報処理装置１０は、目的文書と単語の一致度が低くても、同一のトピックに属する単語が含まれる候補文書を選択することができる。例えば、図５に示す候補文書Ｃ＿｛ｎ１｝は、図４に示す目的文書と同様に、トピック番号＃１およびトピック番号＃５０のトピックの成分が大きいので、言語モデルを学習するための文書として選択される。従って、情報処理装置１０では、人間の感覚では目的文書と類似すると判断される候補文書を適切に選択することができる。これにより、情報処理装置１０によれば、目的に関する多様な表現な含む文書により言語モデルを学習することができるので、多様な表現に対して頑健な言語モデルを生成することができる。

【００６１】

（第１変形例）

つぎに、第１実施形態の第１変形例に係る情報処理装置１０について説明する。

【００６２】

図１０は、トピック数が１０個のトピック情報の一例を示す図である。図１１は、トピック数が２００個のトピック情報の一例を示す図である。

【００６３】

トピック数が少ない場合、１つのトピックには、広い範囲に関連する単語が含まれる。例えば、図１０に示されるように、トピック数が１０個のトピック情報には、トピック番号＃１のトピックに「テレビ」「ＤＶＤ」等のデジタル家庭電化製品に関連する単語に加えて、「番組」「年末」等のテレビジョン番組に関連する単語が含まれてしまう。

【００６４】

トピック数が多い場合、１つのトピックには、狭い範囲に関連する単語が含まれる。例えば、図１１に示されるように、トピック数が２００個のトピック情報には、トピック番号＃１のトピックとトピック番号＃２のトピックとに、「テレビ」と「ＤＶＤ」とが分かれて属してしまう。そして、トピック番号＃１には「テレビ」に関連する単語が含まれ、トピック番号＃２には「ＤＶＤ」に関連する単語が含まれてしまう。

【００６５】

そこで、第１変形例に係るトピック情報取得部２３は、複数のトピック数Ｎに対してトピック情報を生成し、生成されたトピック情報の中から最も適切なトピック情報を選択する。

【００６６】

図１２は、適切なトピック数のトピック情報を選択するための処理フローを示す図である。

【００６７】

まず、ステップＳ３１において、トピック情報取得部２３は、トピック数が異なる複数のトピック情報を生成する。本例においては、トピック情報取得部２３は、トピック数Ｎ＝１０、Ｎ＝５０、Ｎ＝２００のトピック情報を生成する。

【００６８】

続いて、ステップＳ３２において、トピック情報取得部２３は、トピック数が異なる複数のトピック情報のそれぞれに基づいて、目的文書のトピック特徴量を算出する。下記の数１０は、トピック数Ｎ＝１０、Ｎ＝５０、Ｎ＝２００の場合のトピック情報を示す。なお、数１０に示すトピック特徴量は、T_3以降の成分の値は０．０である。

【数１０】

$$\vec{T}_{10}(t)=(T_1,T_2,...)=(0.80,0.04,...)$$

$$\vec{T}_{50}(t)=(T_1,T_2,...)=(0.74,0.03,...)$$

$$\vec{T}_{200}(t)=(T_1,T_2,...)=(0.54,0.50,...)$$

【００６９】

トピック数Ｎ＝１０およびトピック数Ｎ＝５０のトピック情報は、「テレビ」および「リモコン」がトピック番号＃１のトピックに属する。従って、トピック数Ｎ＝１０およびトピック数Ｎ＝５０のトピック情報に基づく、トピック特徴量は、トピック番号＃１の成分T_1の値が大きい。

【００７０】

トピック数Ｎ＝２００のトピック情報は、「テレビ」がトピック番号＃１のトピックに属し、「リモコン」がトピック番号＃２のトピックに属する。従って、トピック数Ｎ＝２００のトピック情報に基づく、トピック特徴量は、トピック番号＃１の成分T_1とトピック番号＃２の成分T_2がほぼ同等となっている。

【００７１】

続いて、ステップＳ３３において、トピック情報取得部２３は、生成した複数のトピック情報のうち、含まれる最大の成分の値が、閾値以上であるトピック情報を抽出する。本例の場合、トピック数Ｎ＝１０のトピック情報に基づくトピック特徴量の最大の成分の値は、０．８０である。また、トピック数Ｎ＝５０のトピック情報に基づくトピック特徴量の最大の成分の値は、０．７４である。また、トピック数Ｎ＝２００のトピック情報に基づくトピック特徴量の最大の成分の値は、０．５４である。そして、閾値を０．７とした場合、トピック情報取得部２３は、閾値以上であるトピック情報として、トピック数Ｎ＝１０のトピック情報、および、トピック数Ｎ＝５０のトピック情報を抽出する。

【００７２】

続いて、ステップＳ３４において、トピック情報取得部２３は、抽出したトピック情報

のうち、トピック数が最大となるトピック情報を選択する。本例の場合、トピック情報取得部２３は、トピック数Ｎ＝５０のトピック情報を選択する。

【００７３】

第１変形例に係る情報処理装置１０は、このように適切な数のトピック数に設定されたトピック情報を用いて、言語モデルを学習するための候補文書を選択する。これにより、本変形例に係る情報処理装置１０によれば、より性能の良い言語モデルを学習することができる。

【００７４】

（第２変形例）

つぎに、第１実施形態の第２変形例に係る情報処理装置１０について説明する。図１３は、第２変形例に係るトピック情報の一例を示す図である。

【００７５】

第２変形例に係るトピック情報は、文章および発話のスタイルを表すトピックの単語の集合を含む。例えば、図１３に示すトピック情報におけるトピック番号＃４９のトピックは、親しい友人との会話で使用されるような通常の発話スタイルで用いられる単語の集合を含む。また、図１３に示すトピック情報におけるトピック番号＃５０のトピックは、接客等で用いられるような丁寧な発話スタイルで用いられる単語の集合を含む。

【００７６】

例えば、コールセンターのオペレータは、通常、丁寧な発話スタイルの音声を発生する。従って、デジタル家庭電化製品に属する単語が含まれている文書であって、且つ、日本語において文章の語尾に用いられる「です」または「ます」等の丁寧な発話スタイルに用いる単語を含む文書を選択することにより、コールセンターのオペレータの音声認識に用いられる言語モデルを、効率良く学習することができる。

【００７７】

従って、第２変形例に係る情報処理装置１０によれば、トピック情報が発話スタイルを表すトピックの単語の集合を含むことにより、特定の用途の言語モデルを学習するために、より適切な候補文書を選択することができる。

【００７８】

（第２実施形態）

つぎに、第２実施形態に係る情報処理装置１０について説明する。なお、第２実施形態に係る情報処理装置１０は、第１実施形態に係る情報処理装置１０と略同一の機能および構成を有する。従って、略同一の機能および構成を有する要素には同一の符号を付けて、相違点を除き詳細な説明を省略する。

【００７９】

図１４は、第２実施形態に係る情報処理装置１０の構成を示す図である。第２変形例に係る情報処理装置１０は、類似目的文書格納部６１と、第３特徴量算出部６２とをさらに備える。

【００８０】

類似目的文書格納部６１は、学習対象の言語モデルと類似した用途で用いられる言語モデルを学習するための文書（類似目的文書）を格納する。例えば、学習対象の言語モデルが、デジタル家庭電化製品のメーカーのコールセンターの音声認識に用いられる場合であれば、類似目的文書により学習する言語モデルは、異なる商品のメーカーのコールセンターの音声認識に用いられる。

【００８１】

トピック情報取得部２３は、含まれる単語が品詞群毎に分割されたトピック情報を取得する。トピック情報取得部２３は、一例として、名詞（第１の品詞群）を含むトピック情報と、名詞以外の単語（例えば、助詞、助動詞、動詞および代名詞等の第２の品詞群）を含むトピック情報とを生成する。

【００８２】

第１特徴量算出部２４は、品詞群毎のトピック情報に基づき、目的文書に対する品詞群

毎のトピック特徴量を算出する。第１特徴量算出部２４は、一例として、目的文書に対する、名詞（第１の品詞群）に関するトピック特徴量および名詞以外の単語（第２の品詞群）に関するトピック特徴量を算出する。

【００８３】

第２特徴量算出部２５は、品詞群毎に分割されたトピック情報に基づき、それぞれの候補文書に対する品詞群毎のトピック特徴量を算出する。第２特徴量算出部２５は、一例として、候補文書に対する、名詞（第１の品詞群）に関するトピック特徴量および名詞以外の単語（第２の品詞群）に関するトピック特徴量を算出する。

【００８４】

第３特徴量算出部６２は、品詞群毎に分割されたトピック情報に基づき、類似目的文書に対する品詞群毎のトピック特徴量を算出する。第３特徴量算出部６２は、一例として、類似目的文書に対する、名詞（第１の品詞群）に関するトピック特徴量および名詞以外の単語（第２の品詞群）に関するトピック特徴量を算出する。

【００８５】

類似度算出部２６は、第１算出部７１と、第２算出部７２とを有する。第１算出部７１は、目的文書に対する品詞群毎のトピック特徴量、および、それぞれの候補文書に対する品詞群毎のトピック特徴量を入力する。また、第１算出部７１は、第１の品詞群の指定を入力する。そして、第１算出部７１は、複数の候補文書のそれぞれの第１の品詞群に関するトピック特徴量に対して、目的文書の第１の品詞群に関するトピック特徴量との第１の類似度を算出する。第１算出部７１は、一例として、それぞれの候補文書の名詞（第１の品詞群）に関するトピック特徴量に対して、目的文書の名詞（第１の品詞群）に関するトピック特徴量の類似度（第１の類似度）を算出する。

【００８６】

第２算出部７２は、類似目的文書に対する品詞群毎のトピック特徴量、および、それぞれの候補文書に対する品詞群毎のトピック特徴量を入力する。また、第２算出部７２は、第２の品詞群の指定を入力する。そして、第２算出部７２は、複数の候補文書のそれぞれの第２の品詞群に関するトピック特徴量に対して、類似目的文書の第２の品詞群に関するトピック特徴量との第２の類似度を算出する。第２算出部７２は、一例として、それぞれの候補文書の名詞以外の品詞（第２の品詞群）に関するトピック特徴量に対して、類似目的文書の名詞以外の品詞（第２の品詞群）に関するトピック特徴量の類似度（第２の類似度）を算出する。

【００８７】

選択部２７は、複数の候補文書のうち、第１の類似度が第１の基準値より大きく、且つ、第２の類似度が第２の基準値より大きい候補文書を、言語モデルの学習に用いる文書として選択する。

【００８８】

ここで、第１の基準値および第２の基準値は、ユーザが設定した値であってもよい。また、第１の基準値は、複数の候補文書の第１の類似度に基づき算出された値（平均値または最大値に基づく値等）であってもよい。また、第２の基準値は、複数の候補文書の第２の類似度に基づき算出された値（平均値または最大値に基づく等）であってもよい。

【００８９】

図１５は、第２実施形態に係る情報処理装置１０の処理フローを示す図である。なお、本フローでは、家庭電化製品のメーカーのコールセンターの音声認識に用いられる言語モデルを学習する例を説明する。

【００９０】

処理に先立って、予めユーザにより目的文書が目的文書格納部２１に格納される。目的文書格納部２１は、一例として、図１６に示されるような、家庭電化製品のメーカーのコールセンターのオペレータにより作成された、対話内容をまとめたレポート等のテキストを、目的文書として格納する。

【００９１】

また、処理に先立って、情報処理装置１０は、ウェッブ等から複数の候補文書を取得して、候補コーパス格納部２２に格納する。候補コーパス格納部２２は、一例として、第１実施形態と同様の、図５、図６および図７に示されるような、候補文書を格納する。

【００９２】

また、処理に先立って、予めユーザにより類似目的文書が類似目的文書格納部６１に格納される。類似目的文書格納部６１は、一例として、図１７に示されるようなテキストを類似目的文書として格納する。図１７のテキストは、家庭電化製品とは異なる製品（食品）のメーカーのコールセンターの音声認識に用いられる言語モデルの学習に利用される文書である。

【００９３】

まず、ステップＳ４１において、トピック情報取得部２３は、品詞群毎に、トピック情報を生成する。下記の数１１は、本実施形態の品詞群の集合の一例を示す式である。

【数１１】

$$PoS=(A,B)=([\text{名詞}],[\text{助詞},\text{助動詞},\text{動詞},\text{代名詞}])$$

【００９４】

数１１の式では、第１の品詞群Ａは、名詞であること示し、第２の品詞群Ｂは、助詞、助動詞、動詞および代名詞であることを示す。なお、トピック情報取得部２３は、３以上の品詞群に分割したトピック情報を生成してもよい。

【００９５】

トピック情報取得部２３は、一例として、第１の品詞群Ａのトピック情報として、図１８に示すようなトピック情報を生成する。また、トピック情報取得部２３は、一例として、第２の品詞群Ｂのトピック情報として、図１９に示すようなトピック情報を生成する。

【００９６】

このように品詞群毎にトピック情報を生成することにより、例えば、名詞のトピック情報は、「デジタル家庭電化製品」（トピック番号＃Ａ＿１）または「食品」（トピック番号＃Ａ＿２）等のトピック毎に、名詞である単語を分類することができる。また、助詞、助動詞、動詞および代名詞のトピック情報は、「書く場合に用いるスタイル」（トピック番号＃Ｂ＿１）または「丁寧な発話のスタイル」（トピック番号＃Ｂ＿２）等の文章または発話のスタイル毎に単語を分類することができる。なお、第１の品詞群のトピック情報と第２の品詞群のトピック情報とは、トピック数が異なっていてよい。

【００９７】

続いて、ステップＳ４２において、第１特徴量算出部２４は、品詞群毎のトピック情報に基づき、目的文書に対する品詞群毎のトピック特徴量を算出する。下記の数１２は、目的文書に対する第１の品詞群Ａに関するトピック特徴量、および、目的文書に対する第２の品詞群Ｂに関するトピック特徴量を示す。

【数１２】

$$\vec{T}_A(t)=(T_{A1},T_{A2},\ldots)=(0.74,0.03,\ldots)$$

$$\vec{T}_B(t)=(T_{B1},T_{B2},\ldots)=(0.81,0.09,\ldots)$$

【００９８】

数１２に示されるように、目的文書は、トピック番号＃Ａ＿１およびトピック番号＃Ｂ＿１の値が大きいので、「デジタル家庭電化製品」および「書く場合に用いるスタイル」との関連性が高いことがわかる。

【００９９】

続いて、ステップＳ４３において、第３特徴量算出部６２は、品詞群毎のトピック情報に基づき、類似目的文書に対する品詞群毎のトピック特徴量を算出する。下記の数１３は、類似目的文書に対する第１の品詞群Ａに関するトピック特徴量、および、類似目的文書

に対する第２の品詞群Ｂに関するトピック特徴量を示す。

【０１００】

【数１３】

$$\vec{T}_A(t') = (0.01, 0.85, \ldots)$$

$$\vec{T}_B(t') = (0.10, 0.80, \ldots)$$

数１３に示されるように、類似目的文書は、トピック番号＃Ａ＿２およびトピック番号＃Ｂ＿２の値が大きいので、「食品」および「丁寧な発話スタイル」との関連性が高いことがわかる。

【０１０１】

続いて、情報処理装置１０は、候補コーパス格納部２２に格納されている候補文書毎に、ステップＳ４５からステップＳ４９の処理を繰り返して実行する（ステップＳ４４とステップＳ５０との間のループ処理）。

【０１０２】

候補文書毎のループ処理では、まず、ステップＳ４５において、第２特徴量算出部２５は、候補文書に対する品詞群毎のトピック特徴量を算出する。下記の数１４は、候補文書Ｃ＿｛ｎ１｝、候補文書Ｃ＿｛ｎ２｝および候補文書Ｃ＿｛ｎ３｝に対する、第１の品詞群Ａおよび第２の品詞群Ｂに関するトピック特徴量を示す。

【数１４】

$$\begin{cases} \vec{T}_A(c_{n1}) = (0.79, 0.01, \ldots) \\ \vec{T}_B(c_{n1}) = (0.10, 0.80, \ldots) \end{cases}$$

$$\begin{cases} \vec{T}_A(c_{n2}) = (0.76, 0.06, \ldots) \\ \vec{T}_B(c_{n2}) = (0.75, 0.10, \ldots) \end{cases}$$

$$\begin{cases} \vec{T}_A(c_{n3}) = (0.03, 0.84, \ldots) \\ \vec{T}_B(c_{n3}) = (0.06, 0.79, \ldots) \end{cases}$$

【０１０３】

数１４に示すように、候補文書Ｃ＿｛ｎ１｝は、トピック番号＃Ａ＿１およびトピック番号＃Ｂ＿２の値が大きいので、「デジタル家庭電化製品」および「丁寧な発話スタイル」との関連性が高いことがわかる。また、候補文書Ｃ＿｛ｎ２｝は、トピック番号＃Ａ＿１およびトピック番号＃Ｂ＿１の値が大きいので、「デジタル家庭電化製品」および「書く場合に用いるスタイル」との関連性が高いことがわかる。また、候補文書Ｃ＿｛ｎ３｝は、トピック番号＃Ａ＿２およびトピック番号＃Ｂ＿２の値が大きいので、「食品」および「丁寧な発話スタイル」との関連性が高いことがわかる。

【０１０４】

続いて、ステップＳ４６において、類似度算出部２６の第１算出部７１は、品詞群毎に、目的文書のトピック特徴量と候補文書のトピック特徴量との類似度（第１の類似度）を算出する。本実施形態においては、第１算出部７１は、下記の数１５に示されるように、第１の品詞群Ａおよび第２の品詞群Ｂのそれぞれについて、目的文書のトピック特徴量と、候補文書のトピック特徴量との内積を演算する。

【数１５】

$$sim_A(t,c_j)=\vec{T}_A(t)\cdot\vec{T}_A(c_j)$$

$$sim_B(t,c_j)=\vec{T}_B(t)\cdot\vec{T}_B(c_j)$$

【０１０５】

続いて、ステップＳ４７において、類似度算出部２６の第２算出部７２は、品詞群毎に、類似目的文書のトピック特徴量と候補文書のトピック特徴量との類似度（第２の類似度）を算出する。本実施形態においては、第１算出部７１は、下記の数１６に示されるように、第１の品詞群Ａおよび第２の品詞群Ｂのそれぞれについて、類似目的文書のトピック特徴量と、候補文書のトピック特徴量との内積を演算する。

【数１６】

$$sim_A(t',c_j)=\vec{T}_A(t')\cdot\vec{T}_A(c_j)$$

$$sim_B(t',c_j)=\vec{T}_B(t')\cdot\vec{T}_B(c_j)$$

【０１０６】

続いて、ステップＳ４８において、選択部２７は、第１の類似度が第１の基準値（th_A）より大きく、且つ、第２の類似度が第２の基準値（th_B）より大きいか否かを判断する。下記の数１７は、選択部２７による判断条件を示す式である。

【数１７】

$$sim_A(t,c_n)>th_A \text{ かつ } sim_B(t',c_n)>th_B$$

【０１０７】

条件を満たさない場合には（ステップＳ４８のＮｏ）、選択部２７は、処理をステップＳ５０に進める。条件を満たす場合には（ステップＳ４８のＹｅｓ）、選択部２７は、処理をステップＳ４９に進める。

【０１０８】

ステップＳ４９において、選択部２７は、対象の候補文書を、言語モデルの学習に用いる文書として選択する。本例においては、選択部２７は、第１の基準値および第２の基準値が０．５０に設定されており、第１の類似度および第２の類似度が共に０．５０より大きい候補文書Ｃ＿｛ｎ１｝を選択する。そして、選択部２７は、処理をステップＳ５０に進める。

【０１０９】

ステップＳ５０において、全ての候補文書について、まだステップＳ４５からステップＳ４９の処理を終了していない場合には、選択部２７は、処理をステップＳ４４に戻して、次の候補文書について処理を繰り返す。終了した場合には、選択部２７は、処理をステップＳ５１に進める。

【０１１０】

ステップＳ５１において、学習部２８は、選択された候補文書を用いて、言語モデルを学習する。そして、ステップＳ５１の処理を終えると、情報処理装置１０は、本フローを終了する。

【０１１１】

ここで、第２実施形態においては、候補文書Ｃ＿｛ｎ１｝についての数１７の条件式は、下記の通りとなる。

ｓｉｍ＿Ａ（ｔ，Ｃ＿｛ｎ１｝）＝０．７４＊０．７９+０．１１＊０．０３＝０．５９、かつ、ｓｉｍ＿Ｂ（ｔ´，Ｃ＿｛ｎ１｝）＝０．１０＊０．１０+０．８＊０．８＝０．６５

【０１１２】
従って、候補文書Ｃ＿｛ｎ１｝は、第１の品詞群Ａおよび第２の品詞群Ｂの両方で条件を満たすので、学習用の文書として抽出される。候補文書Ｃ＿｛ｎ１｝は、デジタル家庭電化製品についての丁寧な発話スタイルの文書であり、コールセンターで発話される内容と一致する。従って、情報処理装置１０は、このような文書を用いて学習を行うことで、性能の高い言語モデルを生成することができる。
【０１１３】
もし、第１の品詞群および第２の品詞群の両方に対して、目的文書との類似度を用いた場合、候補文書Ｃ＿｛ｎ１｝についての、第２の品詞群Ｂに関する数１７の条件式は、ｓｉｍ＿Ｂ（ｔ，Ｃ＿｛ｎ１｝）＝０．１５となる。従って、この場合、候補文書Ｃ＿｛ｎ１｝は、条件を満たさず、学習用の文書として選択されない。一方で、候補文書Ｃ＿｛ｎ２｝についての数１７の条件式は、ｓｉｍ＿Ａ（ｔ，Ｃ＿｛ｎ２｝）＝０．５６、ｓｉｍ＿Ｂ（ｔ，Ｃ＿｛ｎ２｝）＝０．６５となる。従って、この場合、候補文書Ｃ＿｛ｎ２｝が学習用の文書として選択され、コールセンターで実際には発話されないような、書く場合に用いるスタイルの単語を含んだ文書が、学習用の文書として選択されてしまう。
【０１１４】
また、もし、第１の品詞群および第２の品詞群の両方に対して、類似目的文書との類似度を用いた場合には、候補文書Ｃ＿｛ｎ１｝についての、第１の品詞群Ａに関する数１７の条件式は、ｓｉｍ＿Ａ（ｔ´，Ｃ＿｛ｎ１｝）＝０．１１となる。従って、この場合、候補文書Ｃ＿｛ｎ１｝は、条件を満たさず、学習用の文書として選択されない。
【０１１５】
一方で、候補文書Ｃ＿｛ｎ３｝についての数１７の条件式は、ｓｉｍ＿Ａ（ｔ´，Ｃ＿｛ｎ３｝）＝０．７１、ｓｉｍ＿Ｂ（ｔ，Ｃ＿｛ｎ３｝）＝０．６４となる。従って、この場合、候補文書Ｃ＿｛ｎ３｝が学習用の文書として選択され、異なる話題のコールセンターの発話と類似した文書が、学習用の文書として選択されてしまう。
【０１１６】
このように第２実施形態に係る情報処理装置１０によれば、目的文書の主要なテーマと、類似目的文書の発話スタイルが予め分かっている場合に、両文書の特徴を組み合わせて、目的に合った学習用の文書を選択することができる。
【０１１７】
（ハードウェア構成）
図２０は、実施形態に係る情報処理装置１０のハードウェア構成の一例を示す図である。実施形態に係る情報処理装置１０は、ＣＰＵ１０１（Central Processing Unit）等の制御装置と、ＲＯＭ１０２（Read Only Memory）およびＲＡＭ１０３（Random Access Memory）等の記憶装置と、ネットワークに接続して通信を行う通信Ｉ／Ｆ１０４と、各部を接続するバスとを備えている。
【０１１８】
実施形態に係る情報処理装置１０で実行されるプログラムは、ＲＯＭ１０２等に予め組み込まれて提供される。また、実施形態に係る情報処理装置１０で実行されるプログラムは、インストール可能な形式または実行可能な形式のファイルでＣＤ－ＲＯＭ（Compact Disk Read Only Memory）、フレキシブルディスク（ＦＤ）、ＣＤ－Ｒ（Compact Disk Recordable）、ＤＶＤ（Digital Versatile Disk）等のコンピュータで読み取り可能な記録媒体に記録してコンピュータプログラムプロダクトとして提供されてもよい。
【０１１９】
さらに、実施形態に係る情報処理装置１０で実行されるプログラムは、インターネット等のネットワークに接続されたコンピュータ上に格納され、情報処理装置１０がネットワーク経由でダウンロードすることにより提供されてもよい。また、実施形態に係る情報処理装置１０で実行されるプログラムは、インターネット等のネットワーク経由で提供または配布されてもよい。
【０１２０】

実施形態に係る情報処理装置１０で実行されるプログラムは、トピック情報取得モジュール、第１特徴量算出モジュール、第２特徴量算出モジュール、第３特徴量算出モジュール、類似度算出モジュール、選択モジュールおよび学習モジュールを含む構成となっており、コンピュータを上述した情報処理装置１０の各部（トピック情報取得部２３、第１特徴量算出部２４、第２特徴量算出部２５、類似度算出部２６、第３特徴量算出部６２、選択部２７および学習部２８）として機能させうる。このコンピュータは、ＣＰＵ１０１がコンピュータ読取可能な記憶媒体からこのプログラムを主記憶装置上に読み出して実行することができる。なお、トピック情報取得部２３、第１特徴量算出部２４、第２特徴量算出部２５、類似度算出部２６、第３特徴量算出部６２、選択部２７および学習部２８は、一部または全部がハードウェアにより構成されていてもよい。

【０１２１】

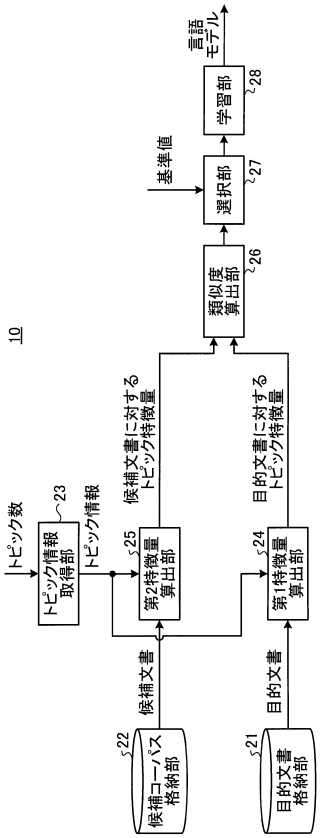
本発明のいくつかの実施形態を説明したが、これらの実施形態は、例として提示したものであり、発明の範囲を限定することは意図していない。これら新規な実施形態は、その他の様々な形態で実施されることが可能であり、発明の要旨を逸脱しない範囲で、種々の省略、置き換え、変更を行うことができる。これら実施形態やその変形は、発明の範囲や要旨に含まれるとともに、請求の範囲に記載された発明とその均等の範囲に含まれる。

【符号の説明】

【０１２２】

１０　情報処理装置
２１　目的文書格納部
２２　候補コーパス格納部
２３　トピック情報取得部
２４　第１特徴量算出部
２５　第２特徴量算出部
２６　類似度算出部
２７　選択部
２８　学習部
６１　類似目的文書格納部
６２　第３特徴量算出部
７１　第１算出部
７２　第２算出部
１０１　ＣＰＵ
１０２　ＲＯＭ
１０３　ＲＡＭ
１０４　通信Ｉ／Ｆ

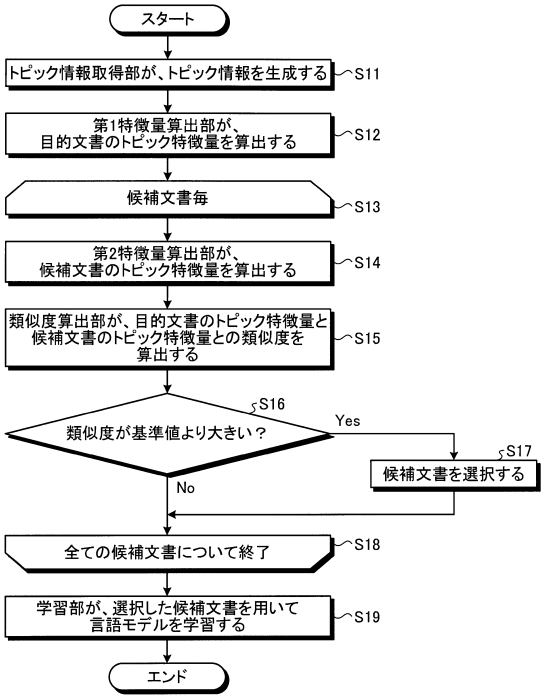
【図１】



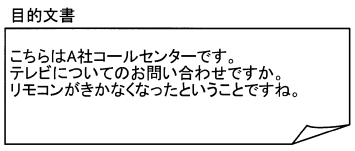
【図２】

トピック番号#1		トピック番号#2		…	トピック番号#49		トピック番号#50	
単語	スコア	単語	スコア		単語	スコア	単語	スコア
テレビ	0.11	食事	0.11		xxx	0.11	xxx	0.11
DVD	0.10	食品	0.10		xxx	0.10	xxx	0.10
リモコン	0.10	アレルギー	0.10		xxx	0.10	xxx	0.10
画面	0.09	成分	0.09		xxx	0.09	xxx	0.09
録画	0.09	ミルク	0.09		xxx	0.09	xxx	0.09
機種	0.08	添加	0.09		xxx	0.09	xxx	0.09
…	…	…	…		…	…	…	…

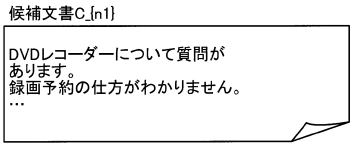
【図３】



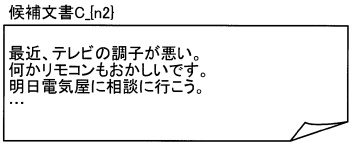
【図４】



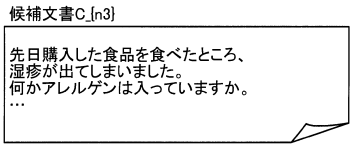
【図５】



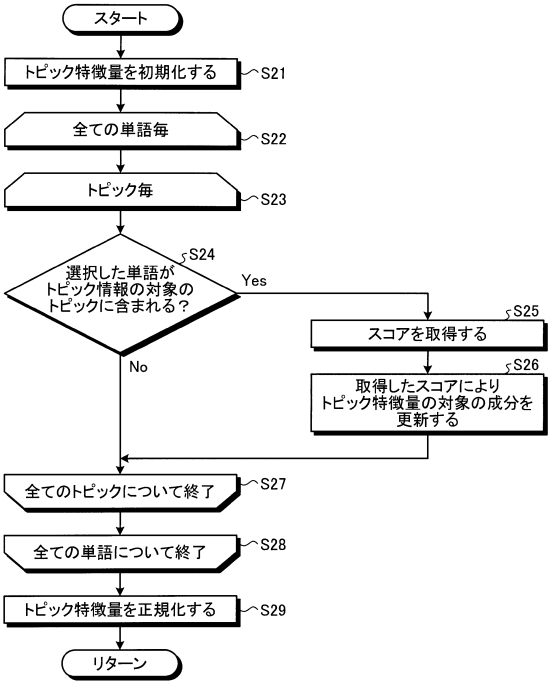
【図６】



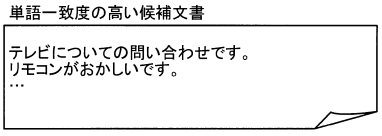
【図７】



【図８】



【図９】



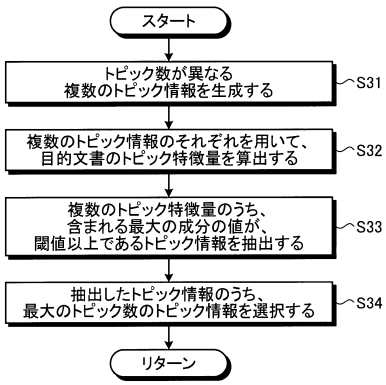
【図１０】

トピック番号#1		…
単語	スコア	
テレビ	0.11	
番組	0.10	
冷蔵庫	0.10	
電気	0.09	
年末	0.09	
DVD	0.08	
…	…	

【図１１】

トピック番号#1		トピック番号#2		…
単語	スコア	単語	スコア	
テレビ	0.11	DVD	0.11	
画面	0.10	リモコン	0.10	
チャンネル	0.10	容量	0.10	
ケーブル	0.09	予約	0.09	
ベゼル	0.09	編集	0.09	
色	0.08	起動	0.09	
…	…	…	…	

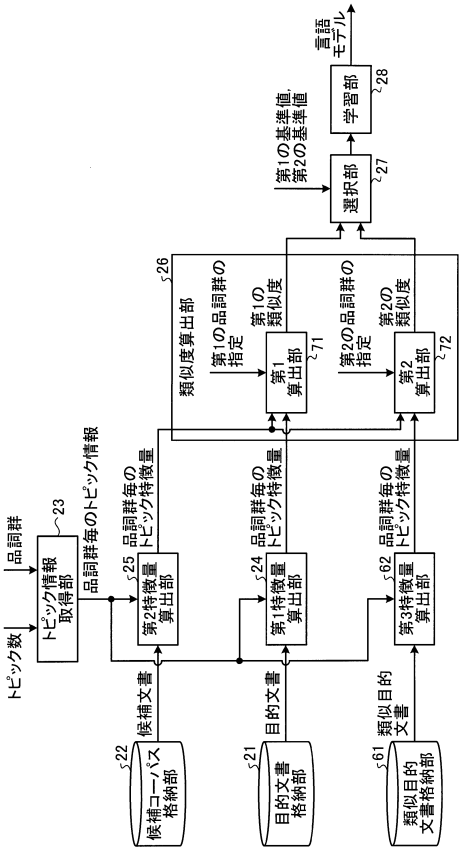
【図１２】



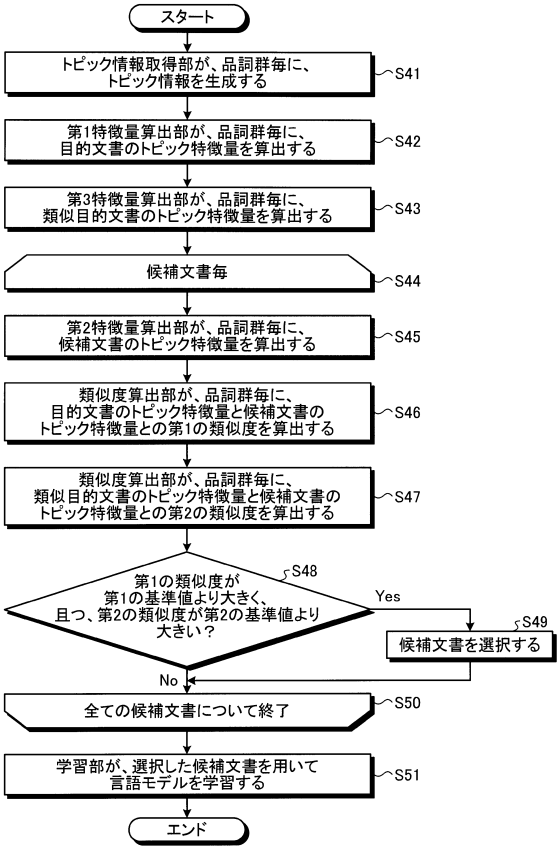
【図１３】

トピック番号#1		トピック番号#2		…	トピック番号#49		トピック番号#50	
単語	スコア	単語	スコア		単語	スコア	単語	スコア
テレビ	0.11	食事	0.11		って	0.11	です	0.11
DVD	0.10	食品	0.10		じゃ	0.10	ます	0.10
リモコン	0.10	アレルギー	0.10		から	0.10	私	0.10
画面	0.09	成分	0.09		けど	0.09	ましょう	0.09
録画	0.09	ミルク	0.09		それ	0.09	申し	0.09
機種	0.08	添加	0.09		みたい	0.09	先日	0.09
…	…	…	…		…	…	…	…

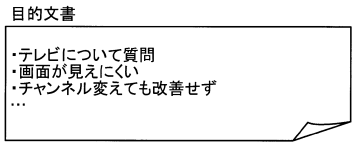
【図１４】



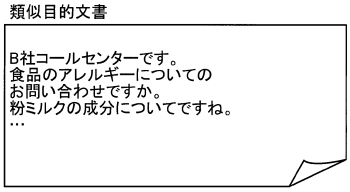
【図１５】



【図１６】



【図１７】



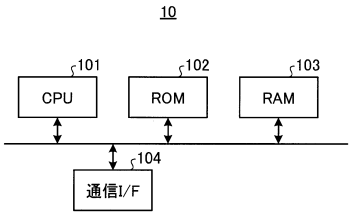
【図１８】

トピック番号#A_1		トピック番号#A_2		…
単語	スコア	単語	スコア	
テレビ	0.11	食事	0.11	
DVD	0.10	食品	0.10	
予約	0.10	アレルギー	0.10	
画面	0.09	成分	0.09	
録画	0.09	ミルク	0.09	
機種	0.08	添加	0.09	
…	…	…	…	

【図１９】

トピック番号#B_1		トピック番号#B_2		…
単語	スコア	単語	スコア	
ない	0.11	です	0.11	
だった	0.10	ます	0.10	
した	0.10	ましょう	0.09	
なった	0.09	申し	0.09	
する	0.09	ませ	0.09	
有り	0.09	…	…	
…	…	…	…	

【図２０】



フロントページの続き

(56)参考文献 特開２０１０－９７３１８（ＪＰ，Ａ）
特開平４－３１４１７１（ＪＰ，Ａ）
特開２００８－１０２７９０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０