

(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
2019年9月12日(12.09.2019)



(10) 国際公開番号

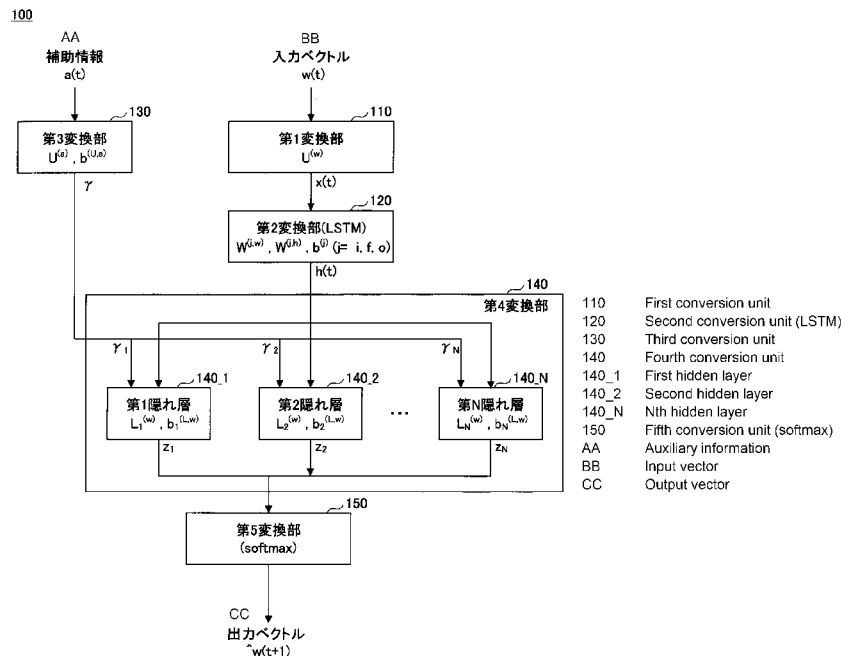
WO 2019/171925 A1

- (51) 国際特許分類:
G06F 17/27 (2006.01) *G06N 3/08* (2006.01)
- (21) 国際出願番号: PCT/JP2019/005871
- (22) 国際出願日: 2019年2月18日(18.02.2019)
- (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 2018-041781 2018年3月8日(08.03.2018) 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: デルクロア マーク(DELICROIX, Marc); 〒1808585 東京都武蔵野市緑町3丁目9-1 NTT知的財産センタ内 Tokyo (JP). 小川 厚徳(OGAWA, Atsunori); 〒1808585 東京都武蔵野市緑町3丁目9-1 NTT知的財産センタ内 Tokyo (JP). 中谷 智広(NAKATANI, Tomohiro); 〒1808585 東京都武蔵野市緑町3丁目9-1 NTT知的財産センタ内 Tokyo (JP). ヘンツェル ミハエル(HENTSCHEL, Michael); 〒6308237 奈良県奈良市中筋町16-2 北羅マンション303 Nara (JP).
- (74) 代理人: 伊東 忠重, 外(ITO, Tadashige et al.); 〒1000005 東京都千代田区丸の内二丁目1

(54) Title: DEVICE, METHOD AND PROGRAM USING LANGUAGE MODEL

(54) 発明の名称: 言語モデルを利用する装置、方法及びプログラム

[図1]



(57) **Abstract:** The model adaptation feature of a language model having higher adaptability is disclosed. One embodiment of the present disclosure is a device having a first neural network unit for converting an inputted symbol and outputting an intermediate state and a second neural network unit for converting inputted auxiliary information and the intermediate state and predicting a symbol that follows the inputted symbol, the second neural network unit having a plurality of hidden layers that accept the intermediate state and the auxiliary information as inputs thereto, the auxiliary information inputted

番 1 号 丸の内 M Y P L A Z A (明治安
田生命ビル) 1 6 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, 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:

- 一 国際調査報告 (条約第21条(3))

to each hidden layer being related to different devices.

(57) 要約: より高い適応能力を有する言語モデルのモデル適応技術が開示される。本開示の一態様は、入力された記号を変換し、中間状態を出力する第1のニューラルネットワーク部と、入力された補助情報と前記中間状態とを変換し、前記入力された記号に後続する記号を予測する第2のニューラルネットワーク部と、を有する装置であって、前記第2のニューラルネットワーク部は、前記中間状態と前記補助情報とを入力とする複数の隠れ層を有し、各隠れ層に入力される補助情報は互いに異なる装置に関する。

明 細 書

発明の名称： 言語モデルを利用する装置、方法及びプログラム 技術分野

[0001] 本発明は、言語モデルのモデル適応技術に関する。

背景技術

[0002] 言語モデルのモデルパラメータを、特定のトピックやジャンルに合わせて適応させる技術は、音声認識等において重要な課題である。一般的な言語モデルと比較して、特定のトピックやジャンルに合わせて適応させた言語モデルは、言語解析や音声認識等のタスクの予測精度の向上が期待される。

[0003] 言語モデルのモデル適応技術として、非特許文献1が知られている。非特許文献1では、リカレントニューラルネットワーク（RNN）で学習された言語モデルのモデル適応技術が開示され、ジャンルに依存しない言語モデルを特定のジャンルに依存するモデルに適応させるため、正解ジャンルのラベルがモデル学習時に追加される。

[0004] また、言語モデルではないが、音響モデルにおけるモデル適応技術として、非特許文献2が知られている。

先行技術文献

非特許文献

[0005] 非特許文献1: Salil Deena, Madina Hasan, Mortaza Doulaty, Oscar Saz, and Thomas Hain, "Combining feature and model-based adaptation of RNN LMs for multi-genre broadcast speech recognition," in INTERSPEECH, 2016, pp. 2343-2347.

非特許文献2: Marc Delcroix, Keisuke Kinoshita, Chengzhu Yu, Atsunori Ogawa, Takuya Yoshioka, and Tomohiro Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation in noisy conditions," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5270-5274.

発明の概要

発明が解決しようとする課題

- [0006] しかしながら、非特許文献 1 に開示されるモデル適応技術では、適応のための補助情報がジャンルのみであり、ジャンルを補助情報とする 1 種類の N N しかないため、適応能力が乏しい。また、正解のジャンルが与えられた状態でなければ適切な学習をすることが難しい。
- [0007] また、非特許文献 2 に開示されるモデル適応技術では、音響モデルにおいて高い適応能力を有することが示されている。しかしながら、言語モデルを対象とする場合に当該モデル適応技術をどのように応用すればよいか、また、言語モデルに応用した場合でも効果があるのか否かは明らかではない。
- [0008] 上述した問題点を鑑み、本発明の課題は、より高い適応能力を有する言語モデルのモデル適応技術を提供することである。

課題を解決するための手段

- [0009] 上記課題を解決するため、本開示の一態様は、入力された記号を変換し、中間状態を出力する第 1 のニューラルネットワーク部と、入力された補助情報と前記中間状態とを変換し、前記入力された記号に後続する記号を予測する第 2 のニューラルネットワーク部と、を有する装置であって、前記第 2 のニューラルネットワーク部は、前記中間状態と前記補助情報とを入力とする複数の隠れ層を有し、各隠れ層に入力される補助情報は互いに異なる装置に関する。

発明の効果

- [0010] 本発明によると、より高い適応能力を有する言語モデルのモデル適応技術を提供することができる。

図面の簡単な説明

- [0011] [図1]図 1 は、本発明の一実施例による推定装置を示すブロック図である。
- [図2]図 2 は、本発明の一実施例による言語モデル学習装置を示すブロック図である。

[図3]図3は、本発明の一実施例による推定装置及び言語モデル学習装置のハードウェア構成を示すブロック図である。

発明を実施するための形態

[0012] 以下の実施例では、学習済みのニューラルネットワーク（言語モデル）を用いて、入力された記号から次の記号を推定する推定装置と、当該言語モデルを学習する言語モデル学習装置とが開示される。ここで、「記号」とは、言語モデルの処理対象であり、例えば、単語、音素、形態素、文字（character）等である。言語モデルは、入力された記号に基づいて、次に出現する可能性の高い記号の予測結果を出力するものである。以下、処理単位が「単語」である場合を例として、処理を説明する。

[0013] 後述される実施例による言語モデルは、単語を所定次元のベクトルで表現した単語ベクトル $w(t)$ を入力とし、補助情報を利用して次の単語の予測結果 $\hat{w}(t+1)$ を出力するものである。例えば、音声認識において、時刻 t に発話された単語 $w(t)$ の次に発話される可能性の高い単語 $\hat{w}(t+1)$ を予測する場合等に用いられる。本発明の言語モデルの主な特徴は、補助情報（トピック、文脈、話者等）に依存するニューラルネットワークが複数の分解された隠れ層から構成されることと、分解された隠れ層の各々において、異なる補助情報（ γ_n ）に基づく変換が行われることである。

〔実施例1：推定装置〕

まず、図1を参照して、本発明の一実施例による言語モデルを利用した推定装置を説明する。図1は、本発明の一実施例による推定装置を示すブロック図である。

[0014] 図1に示されるように、推定装置100は、第1変換部110、第2変換部120、第3変換部130、第4変換部140及び第5変換部150を有し、第4変換部140は、第1隠れ層140__1、第2隠れ層140__2、
・・・、第N隠れ層140__Nを有する。

[0015] 第1変換部110は、入力ベクトル $w(t)$ を重み行列 $U^{(w)}$ により線形変換することによって第1出力状態 $x(t)$ を取得する。例えば、第1変換部110は、

ニューラルネットワークの線形層により実現されうる。具体的には、第1変換部110は、時点tにおいて入力された単語を符号化したベクトル $w(t)$ に対して、

[0016] [数1]

$$x(t) = U^{(w)} w(t)$$

に従って $x(t)$ を計算し、第2変換部120に出力する。

[0017] 第2変換部120は、第1変換部110から出力された $x(t)$ をニューラルネットワークに入力することによって第2出力状態 $h(t)$ を取得する。ニューラルネットワークの構成は特に限定されるものではなく、RNN(Recurrent Neural Network), LSTM(Long Short Term Memory), DNN(Deep Neural Network), CNN(Convolutional Neural Network)等が利用できる。一例として、LSTMで実現する場合には、第2変換部120は、

[0018] [数2]

$$\begin{aligned} i(t) &= \sigma(W^{(i,w)} x(t) + W^{(i,h)} h(t-1) + b^{(i)}) \\ f(t) &= \sigma(W^{(f,w)} x(t) + W^{(f,h)} h(t-1) + b^{(f)}) \\ o(t) &= \sigma(W^{(o,w)} x(t) + W^{(o,h)} h(t-1) + b^{(o)}) \\ g(t) &= \tanh(W^{(g,w)} x(t) + W^{(g,h)} h(t-1) + b^{(g)}) \\ c(t) &= f(t) \odot c(t-1) + i(t) \odot g(t) \\ h(t) &= o(t) \odot \tanh(c(t)) \end{aligned}$$

に従って第2出力状態 $h(t)$ を計算し、第4変換部140に出力する。ここで、 $i(t)$, $f(t)$, $o(t)$ はそれぞれLSTMにおける入力ゲート、忘却ゲート及び出力ゲートであり、 $W^{(i,w)}$, $W^{(i,h)}$ はそれぞれ入力された単語及び前の隠れ層のゲート

トjのウェイト行列である。また、 $\sigma(\cdot)$ はベクトルの要素単位のシグモイドであり、 $\tanh(\cdot)$ はベクトルの要素単位のハイパボリックタンジェントであり、

[0019] [数3]

⊙

はベクトルの要素単位の乗算である。ベクトルの要素単位のシグモイドとは、ベクトルの各要素についてのシグモイド関数値を並べたベクトルを返す関数であり、ベクトルの要素単位のハイパボリックタンジェントとは、ベクトルの各要素についての $\tanh$ 関数値を並べたベクトルを返す関数である。また、 $c(t)$ は隠れ状態である。

[0020] 第3変換部130は、潜在的ディリクレ配分法（LDA）の出力状態である補助情報 $a(t)$ をニューラルネットワークにより線形変換し、変換後の補助情報 γ を取得し、第4変換部140に出力する。具体的には、第3変換部130は、重み行列を $U^{(a)}$ とし、バイアスベクトルを $b^{(U,a)}$ として、

[0021] [数4]

$$\gamma = [\gamma_1, \gamma_2, \dots, \gamma_n, \dots, \gamma_N] = \sigma(U^{(a)}a(t) + b^{(U,a)})$$

に従って線形変換を行う線形層として実現されうる。変換後の補助情報 γ は要素毎に分解され、分解された各要素 γ_n は、第4変換部140の対応する隠れ層140__nに入力される。

[0022] ここで、潜在的ディリクレ配分法（LDA）は、テキスト文書の潜在的なトピック（話題、分野等）を推定するためのトピックモデルの一種であり、テキスト文書を入力として、各潜在的なトピックに属する確率を表す確率分布の

ベクトルを出力する。この出力が第3変換部130の入力として用いられる $a(t)$ である。すなわち、LDAは、テキスト文書の集合におけるトピックを推定するための生成モデルであり、当該集合の各テキスト文書は潜在的なトピックを合成したものとしてモデル化され、これら潜在的なトピックはトピック確率上の合成としてモデル化される。LDAはテキスト文書における単語の順序を無視して、bag-of-wordsモデルでテキスト文書の特徴を表現する。具体的には、テキスト文書中の各文をbag-of-wordsモデルでベクトル表現したものを並べた低ランクの行列表現として、テキスト文書を表現する。LDAでは、各テキスト文書は以下の生成プロセスによって生成される。

- [0023] 1. ポワソン分布からテキスト文書の長さ M をサンプリングする。
- [0024] 2. α で特定されるディリクレ分布からのサンプリングによって、テキスト文書のトピック上の多項分布を選択する。
- [0025] 3. M 個の単語の各単語 q_m に対して、トピックを選択し、当該トピックに関連するユニグラム分布
- [0026] [数5]

$$p(w_m|q_m, \beta)$$

から単語 w_m を選択する。 α, β はLDAにおける主要なパラメータであり、予め学習データに基づいてモデル学習により学習しておく。 α はトピックが抽出される多項分布上のディリクレ分布の形状を決定し、 β は単語確率に直接影響を与えるパラメータである。

- [0027] 第4変換部140は、 N 個の分解された隠れ層140__1, 140__2, ..., 140__ N から構成される。隠れ層の各々は、linear hidden network (LHN)により実現される。 n 番目の隠れ層140__ n のパラメータを $L_n^{(w)}$ 、バイアスベクトルを $b_n^{(L, w)}$ とする。このとき、第3変換部130から出力される変換後の補助情報 γ のうち、 n 番目の隠れ層140__ n へ入力される情報を γ_n と

すると、 n 番目の隠れ層 1 4 0 __ n は、第 2 変換部 1 2 0 から提供される第 2 出力状態 $h(t)$ と第 3 変換部 1 3 0 から提供される γ_n を入力とし、

[0028] [数6]

$$\gamma_n(L_n^{(w)}h(t) + b_n^{(L,w)})$$

である出力状態 z_n を第 5 変換部 1 5 0 に出力する。

[0029] 第 5 変換部 1 5 0 は、第 4 変換部 1 4 0 から出力された出力状態 z_n ($n=1, 2, \dots, N$) に対して、

[0030] [数7]

$$z = \sum_{n=1}^N \underbrace{\gamma_n(L_n^{(w)}h(t) + b_n^{(L,w)})}_{=z_n}$$

$$\hat{w}(t+1) = \text{softmax}(z).$$

に従って softmax 関数値を予測結果 $\hat{w}(t+1)$ として出力する。ここで、 $\text{softmax}(\cdot)$ は softmax 関数であり、 m 次元の実数値からなるベクトルを m 次元のベクトルに変換する関数であり、入力を $e=(e_1, e_2, \dots, e_m)^T$ (T は転置を表す) とし、出力を $\hat{e}=(\hat{e}_1, \hat{e}_2, \dots, \hat{e}_m)^T$ とした場合、

[0031] [数8]

$$\hat{e}_j = \frac{\exp(e_j)}{\sum_{k=1}^m \exp(e_k)}$$

となるように変換する関数である。つまり、出力ベクトルの各要素の総和が 1

になるように正規化する関数である。

[0032] 上述した実施例では、補助情報 $a(t)$ は、潜在的ディリクレ配分法（LDA）により求めたテキスト文書の潜在的なトピック確率を表す確率分布（ベクトル）であるが、これに限定されるものでない。例えば、補助情報 $a(t)$ は、言語モデルの出力となる次の単語の予測結果 $\hat{w}(t+1)$ の選択に影響するような、直前の単語 $w(t)$ のみから生成される情報（以下、「分解された隠れ層への入力情報」という）以外の何れか適切な補助的な情報であってもよい。例えば、補助情報 $a(t)$ として、 $w(t)$ までの過去の文字列の品詞、直前の文の文脈（対話であれば、直前の話者の発話内容に関する情報等）、会話の場の雰囲気（フォーマルなのか、カジュアルなのか）、会話相手との関係（上司と部下、友人、初対面の人等）、話者の性別、年齢、出身地域（言語や方言に影響もの）が利用されてもよい。会話の場の雰囲気や会話相手、話者の性別等の情報は、いわば予測対象の単語 $\hat{w}(t+1)$ よりも時系列的に前の単語列 $w(t_0), \dots, w(t)$ は、当該単語列の獲得環境を表す情報ともいえる。なお、この例では、 t_0 と t は単語列のインデックスであり、 t が小さいほど時系列的に前の単語のインデックスであるとし、 $t_0 < t$ とした。例えば、文の先頭の単語のインデックスを t_0 とすれば、文の先頭から予測対象の単語のインデックス $t+1$ の直前（つまり t ）までの単語列が $w(t_0), \dots, w(t)$ である。

[0033] また、上述した実施例では、補助情報 $a(t)$ は、トピック確率という1種類の情報であり、これを分解された隠れ層の数（ N 個）に分割した補助情報 γ_n （ $n = 1, \dots, N$ ）の何れかを各分解された隠れ層に入力していた。しかし、本発明はこれに限られるものではなく、上述した様々な種類の補助情報の中から、複数の異なる種別の補助情報を生成し、それぞれの分解された隠れ層に異なる種別の補助情報が入力される構成としてもよい。要するに、分解された隠れ層に入力される補助情報が互いに異なるものであれば、種別は同じであっても異なってもよい。

[0034] このようにして、本実施例の第4変換部における分解された隠れ層は、それぞれ、入力された補助情報に応じた適応処理を行った結果を出力すること

になる。そして、第5変換部が、全ての分解された隠れ層の結果を総合的に勘案し、最終的に出力する次の単語の予測結果 $\hat{w}(t+1)$ を出力する役割を担うのである。

〔実施例2：言語モデル学習装置〕

次に、図2を参照して、本発明の一実施例による上述した言語モデルを学習するための言語モデル学習装置を説明する。ここで、言語モデル学習装置200における学習対象となるパラメータは、第1変換部110、第2変換部120、第3変換部130及び第4変換部140において用いられる行列及びバイアスベクトルの要素である。図2は、本発明の一実施例による言語モデル学習装置を示すブロック図である。

[0035] 図2に示されるように、言語モデル学習装置200は、第1変換部210、第2変換部220、第3変換部230、第4変換部240、第5変換部250及びパラメータ更新部260を有する。第1変換部210、第2変換部220、第3変換部230、第4変換部240及び第5変換部250は、推定装置100の第1変換部110、第2変換部120、第3変換部130、第4変換部140及び第5変換部150と基本的に同じであり、本実施例では、説明の簡単化のため相違点のみ説明する。

[0036] 第1変換部210の入力 $w(t)$ は、学習用に用意された単語ベクトルであり、パラメータ更新部260に与えられる正解の出力ベクトル $w(t+1)$ と予め対応付けられているものとする。第1変換部110と同様にして、第1変換部210は、当該入力ベクトル $w(t)$ を重み行列 $U^{(w)}$ により線形変換することによって第1出力状態 $x(t)$ を取得し、第2変換部220に出力する。

[0037] 第2変換部220は、第2変換部120と同様にして、第1変換部210から出力された $x(t)$ をLSTMに入力することによって第2出力状態 $h(t)$ を取得し、第4変換部240に出力する。

[0038] 第3変換部230の入力 $a(t)$ は、学習用に用意された単語ベクトル $w(t)$ に対応する補助情報となる。補助情報 $a(t)$ は、予め学習用データとして $w(t)$ に対応付けて記憶されていてもよいし、第1変換部210に入力された学習用

データの $w(t)$ から上述の潜在的ディリクレ配分法（LDA）等により計算されてもよい。なお、学習時に用いる補助情報の種類と、推定装置100の運用時に用いる補助情報の種類を同一のものとする。第3変換部230は、第3変換部130と同様にして、補助情報 $a(t)$ をニューラルネットワークにより線形変換された補助情報 γ を取得し、第4変換部240に出力する。

[0039] 第4変換部240は、 N 個の分解された隠れ層240__1, 240__2, ..., 240__ N から構成される。 n 番目の隠れ層240__ n は、第4変換部140の n 番目の隠れ層140__ n と同様にして、第2変換部220から提供される第2出力状態 $h(t)$ と第3変換部230から提供される γ_n から出力状態 z_n を取得し、第5変換部250に出力する。

[0040] 第5変換部250は、第5変換部150と同様にして、第4変換部240から出力された出力状態 z_n ($n=1, 2, \dots, N$)を変換して予測出力ベクトル $\hat{w}(t+1)$ を生成し、パラメータ更新部260に出力する。例えば、出力状態 z_n に関するsoftmax関数値を予測出力ベクトル $\hat{w}(t+1)$ とすればよい。なお、ここでは学習用データに含まれる $w(t)$ に対応する正解データ $w(t+1)$ と区別するため、出力された予測出力ベクトルを記号 $\hat{w}(t+1)$ で表記している。

[0041] パラメータ更新部260は、第5変換部250から出力された予測出力ベクトル $\hat{w}(t+1)$ と、学習用データとして予め与えられている正解出力ベクトル $w(t+1)$ とを比較し、当該比較結果に応じて言語モデルの各パラメータを更新する。パラメータの更新は、例えば、誤差伝播法等、ニューラルネットワークにおける何れか適切なパラメータ更新手法に従って実行すればよい。これにより、予測出力ベクトル $\hat{w}(t+1)$ が正解出力ベクトル $w(t+1)$ に近づくように、ニューラルネットワークの各パラメータが繰り返し更新されることとなる。

[0042] そして、パラメータ更新部260は、予測出力ベクトル $\hat{w}(t+1)$ と正解出力ベクトル $w(t+1)$ との比較結果が予め定めた基準に到達したら、ニューラルネットワークの各パラメータの値を出力して、パラメータ更新処理を終了する。予め定めた基準とは、例えば $\hat{w}(t+1)$ と $w(t+1)$ の近さ（距離等）が所定の閾

値未満となる、或いは、繰り返し回数（パラメータ更新部260の更新処理の回数）が所定の回数に到達する、等である。

[0043] なお、学習済みの言語モデルを音声認識に利用する場合には、入力ベクトル $w(t)$ として、時間区間 t の入力音声信号 $s(t)$ を認識して得られる単語列に対応する単語ベクトルが利用されてもよい。これを言語モデルに入力することによって、次に発話される可能性の高い単語候補 $\hat{w}(t+1)$ を得ることができる。この結果に応じて、次の時間区間の音声信号 $s(t+1)$ の音声認識結果である単語列を求める際に、推定装置100は、言語モデルにより予測された $\hat{w}(t+1)$ を参考にして認識結果を求めてもよい。

[0044] ここで、推定装置100及び言語モデル学習装置200は、典型的には、サーバなどの計算装置により実現されてもよく、例えば、図3に示されるように、バスBを介し相互接続されるドライブ装置101、補助記憶装置102、メモリ装置103、プロセッサ104、インタフェース装置105及び通信装置106から構成されてもよい。推定装置100及び言語モデル学習装置200における後述される各種機能及び処理を実現するプログラムを含む各種コンピュータプログラムは、CD-ROM (Compact Disk-Read Only Memory)、DVD (Digital Versatile Disk)、フラッシュメモリなどの記録媒体107によって提供されてもよい。プログラムを記憶した記録媒体107がドライブ装置101にセットされると、プログラムが記録媒体107からドライブ装置101を介して補助記憶装置102にインストールされる。但し、プログラムのインストールは必ずしも記録媒体107により行う必要はなく、ネットワークなどを介し何れかの外部装置からダウンロードするようにしてもよい。補助記憶装置102は、インストールされたプログラムを格納すると共に、必要なファイルやデータなどを格納する。メモリ装置103は、プログラムの起動指示があった場合に、補助記憶装置102からプログラムやデータを読み出して格納する。プロセッサ104は、メモリ装置103に格納されたプログラムやプログラムを実行するのに必要なパラメータなどの各種デー

タに従って、上述した推定装置１００及び言語モデル学習装置２００の各種機能及び処理を実行する。インタフェース装置１０５は、ネットワーク又は外部装置に接続するための通信インタフェースとして用いられる。通信装置１０６は、インターネットなどのネットワークと通信するための各種通信処理を実行する。

[0045] しかしながら、推定装置１００及び言語モデル学習装置２００は、上述したハードウェア構成に限定されるものでなく、他の何れか適切なハードウェア構成により実現されてもよい。

[0046] 以上、本発明の実施例について詳述したが、本発明は上述した特定の実施形態に限定されるものではなく、特許請求の範囲に記載された本発明の要旨の範囲内において、種々の変形・変更が可能である。

符号の説明

- [0047] １００ 推定装置
２００ 言語モデル学習装置
１１０、２１０ 第１変換部
１２０、２２０ 第２変換部
１３０、２３０ 第３変換部
１４０、２４０ 第４変換部
１５０、２５０ 第５変換部
２６０ パラメータ更新部

請求の範囲

- [請求項1] 入力された記号を変換し、中間状態を出力する第1のニューラルネットワーク部と、
- 入力された補助情報と前記中間状態とを変換し、前記入力された記号に後続する記号を予測する第2のニューラルネットワーク部と、
- を有する装置であって、
- 前記第2のニューラルネットワーク部は、前記中間状態と前記補助情報とを入力とする複数の隠れ層を有し、各隠れ層に入力される補助情報は互いに異なる装置。
- [請求項2] 入力された学習用記号を変換し、中間状態を出力する第1のニューラルネットワーク部と、
- 前記学習用記号に対応して入力された補助情報と前記中間状態とを変換し、前記入力された学習用記号に後続する記号を予測する第2のニューラルネットワーク部と、
- 前記入力された学習用記号に対応する正解出力記号と前記予測された後続する記号を比較し、比較結果に応じて前記第1のニューラルネットワーク部と前記第2のニューラルネットワーク部とのパラメータを更新するパラメータ更新部と、
- を有する装置であって、
- 前記第2のニューラルネットワーク部は、前記中間状態と前記補助情報とを入力とする複数の隠れ層を有し、各隠れ層に入力される補助情報は互いに異なる装置。
- [請求項3] 前記補助情報は、前記後続する記号の選択に影響する前記入力された記号以外の情報を含む情報から抽出されたものである、請求項1又は2記載の装置。
- [請求項4] 前記補助情報は、前記入力された記号を含む、該入力された記号よりも時系列的に前の記号列のトピックと、該記号列の文脈と、該記号列の獲得環境とのうち少なくとも1以上の情報を含む、請求項1乃至

3 何れか一項記載の装置。

[請求項5]

プロセッサが、第1のニューラルネットワークモデルを用いて、入力された記号を変換し、中間状態を出力するステップと、

前記プロセッサが、第2のニューラルネットワークモデルを用いて、入力された補助情報と前記中間状態とを変換し、前記入力された記号に後続する記号を予測するステップと、

を有する方法であって、

前記第2のニューラルネットワークモデルは、前記中間状態と前記補助情報とを入力とする複数の隠れ層を有し、各隠れ層に入力される補助情報は互いに異なる方法。

[請求項6]

プロセッサが、第1のニューラルネットワークモデルを用いて、入力された学習用記号を変換し、中間状態を出力するステップと、

前記プロセッサが、第2のニューラルネットワークモデルを用いて、前記学習用記号に対応して入力された補助情報と前記中間状態とを変換し、前記入力された学習用記号に後続する記号を予測するステップと、

前記プロセッサが、前記入力された学習用記号に対応する正解出力記号と前記予測された後続する記号とを比較し、比較結果に応じて前記第1のニューラルネットワークモデルと前記第2のニューラルネットワークモデルとのパラメータを更新するステップと、

を有する方法であって、

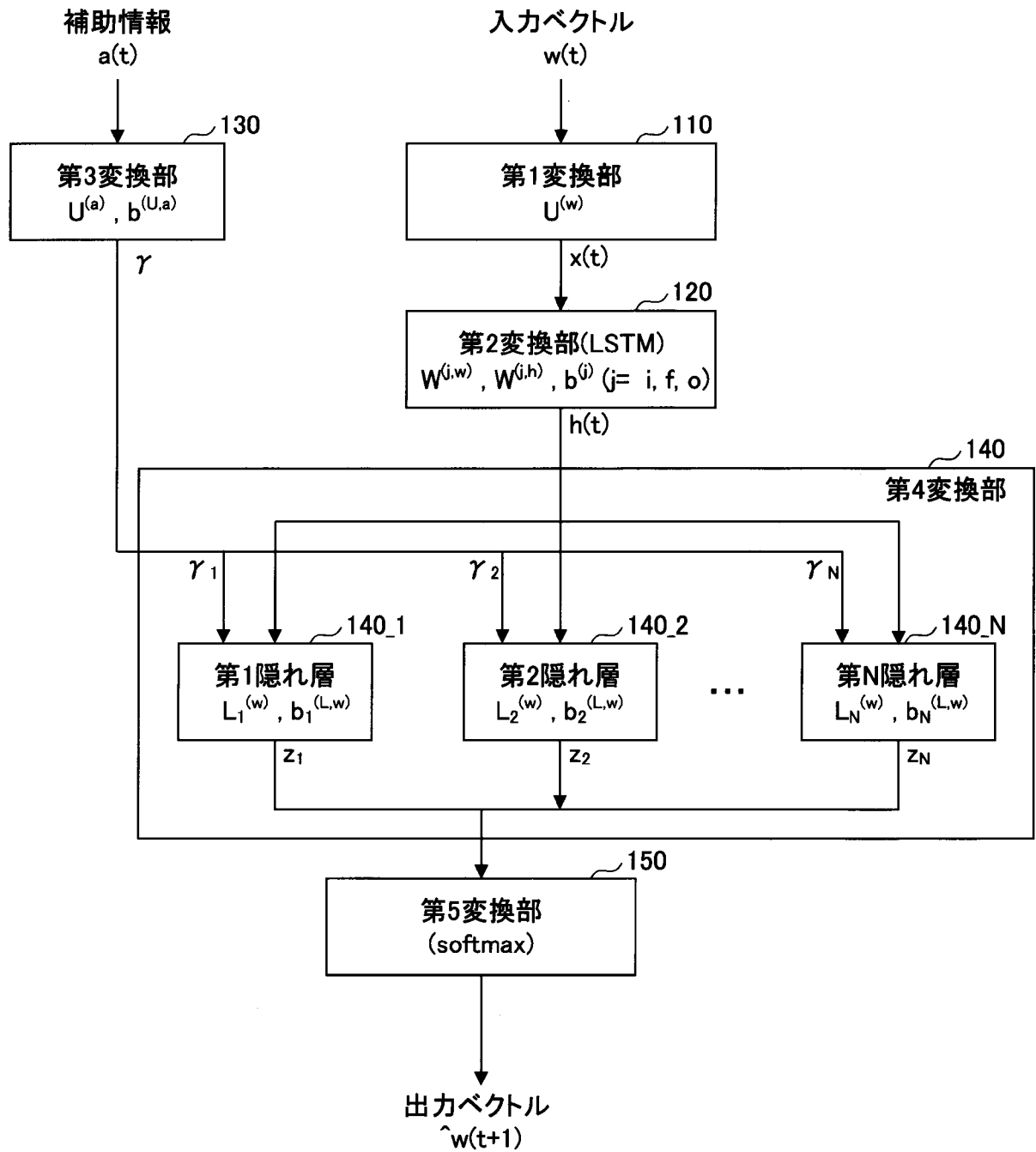
前記第2のニューラルネットワークモデルは、前記中間状態と前記補助情報とを入力とする複数の隠れ層を有し、各隠れ層に入力される補助情報は互いに異なる方法。

[請求項7]

請求項1乃至4何れか一項記載の装置の各部としてプロセッサを機能させるプログラム。

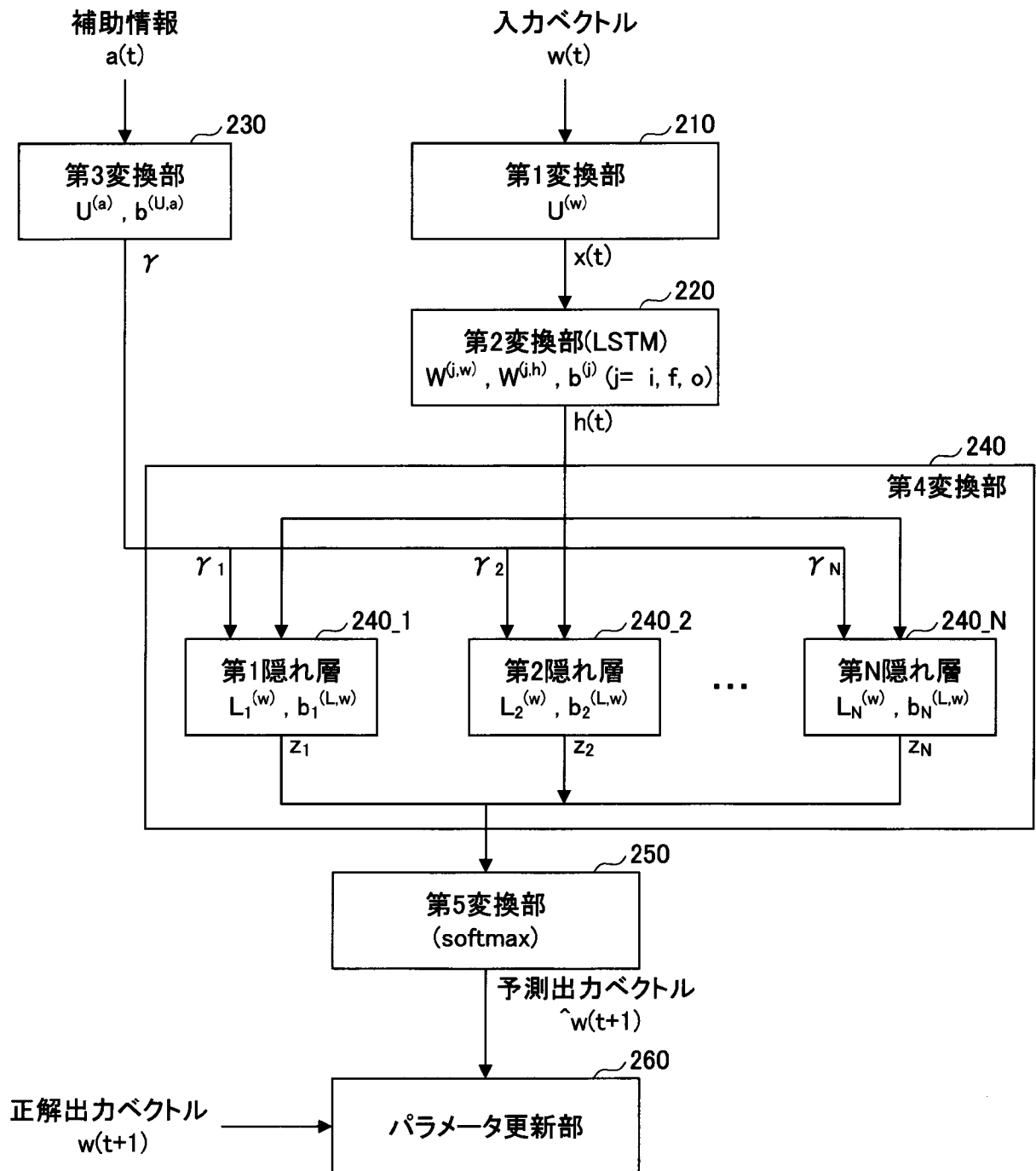
[図1]

100

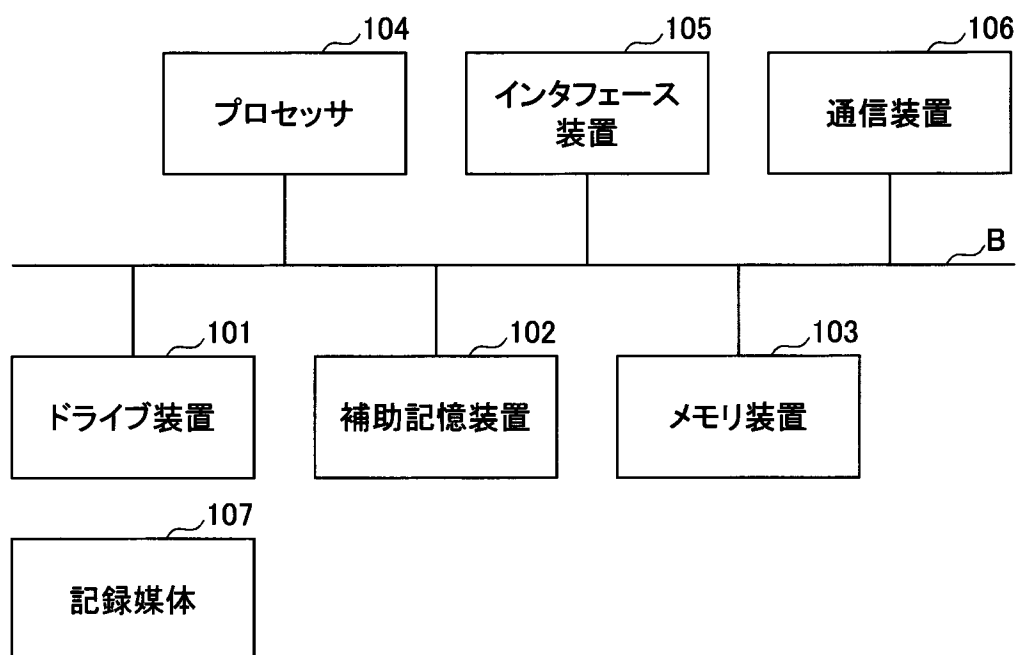


[図2]

200



[図3]

100, 200

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2019/005871

A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl. G06F17/27(2006.01)i, G06N3/08(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. G06F17/20-17/28, G06N3/08

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 1971-2019

Registered utility model specifications of Japan 1996-2019

Published registered utility model applications of Japan 1994-2019

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.
A	JP 2017-16384 A (JAPAN BROADCASTING CORPORATION) 19 January 2017, entire text, all drawings (Family: none)	1-7
A	WO 2017/135148 A1 (NIPPON TELEGRAPH AND TELEPHONE CORP.) 10 August 2017, entire text, all drawings & CN 108701452 A, entire text, all drawings	1-7



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

10 May 2019 (10.05.2019)

Date of mailing of the international search report

21 May 2019 (21.05.2019)

Name and mailing address of the ISA/

Japan Patent Office

3-4-3, Kasumigaseki, Chiyoda-ku,

Tokyo 100-8915, Japan

Authorized officer

Telephone No.

A. 発明の属する分野の分類（国際特許分類（I P C））

Int.Cl. G06F17/27(2006.01)i, G06N3/08(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（I P C））

Int.Cl. G06F17/20-17/28, G06N3/08

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	1971-2019年
日本国実用新案登録公報	1996-2019年
日本国登録実用新案公報	1994-2019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
A	JP 2017-16384 A（日本放送協会）2017.01.19, 全文・全図（ファミリーなし）	1-7
A	WO 2017/135148 A1（日本電信電話株式会社）2017.08.10, 全文・全図 & CN 108701452 A 全文・全図	1-7

☐ 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

10.05.2019

国際調査報告の発送日

21.05.2019

国際調査機関の名称及びあて先

日本国特許庁（I S A/J P）

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）

成瀬 博之

電話番号 03-3581-1101 内線 3586

5 N

9192