Vol.2017-SLP-117 No.4
2017/7/28

再帰型ニューラルネットワークに基づく音素情報を用いた応答選択

牧野 健一郎[1,a)]　李 晃伸[1,b)]

概要：統計的音声対話システムの構築には対話コーパスの収集が必要であり，タスクドメインごとに大規模な対話データを収集するのは困難である．本研究では，少量の対話データからロバストな音声対話システムを構築する手法の 1 つとして，認識結果の音素情報のみから B-LSTM を用いた応答選択手法を検討する．評価実験では音素情報を用いるモデルと単語情報を用いるモデルを比較を行った結果，音声情報案内システム「たけまるくん」のタスクにおいて応答正解率で評価した結果，単語情報を用いるモデルは 77.8 %にであったのに対し，音素情報を用いるモデルは 81.9 %であった．

キーワード：音声対話システム，応答選択，音素情報，B-LSTM

1. はじめに

近年，専門知識を持たなくとも操作が直感的にわかるデバイスとして音声対話システムの普及が注目されている，NTT ドコモの「しゃべってコンシェル」，Apple の “Siri” や Google の “OK Google” 然り，実社会で運用されている．本研究では，ユーザとシステムが QA のやり取りを行うことによってタスクを達成する一問一答形式の音声対話システムを取り扱う．

一問一答の音声対話システムは音声認識の結果から直接適切な応答を選択する．応答選択する手法として対話の事例である質問文と応答文の組からなる用例データを用いて認識結果に近い質問文に対する応答文を選択する用例ベースの応答選択がある．用例ベースの応答選択を統計的に精度よく行うためには用例データが膨大に必要である．しかし，用例データはタスク依存性が高く，タスク設定ごとに実環境を考慮したデータ収集が困難である．また，従来の用例ベースの応答選択を行う手法 [1][2][3][4] は単語レベルのマッチングを行っているが，音声認識には誤認識が不可避の問題であり誤認識した単語によるマッチングではシステムの正しい応答が得ることができない．

そこで本研究では音素情報のみを用いる応答選択を提案する．音素を使う場合，誤認識した単語と正解単語の共通する音素系列をマッチングすることで誤認識にロバストに応答選択ができる．また，音素の種類は単語と比べ少量であり，少ないデータ量でより識別率の高いモデルが作成できる可能性がある．

以下，第 2 節では従来の応答選択手法について，第 3 節では応答選択に音素情報のみを用いる狙いと本研究で用いた B-LSTM (Bidirectional Long Short Term Memory) に基づく応答選択モデルについて述べる．第 4 節では評価実験の結果を示し，第 5 節では本研究のむすびを述べる．

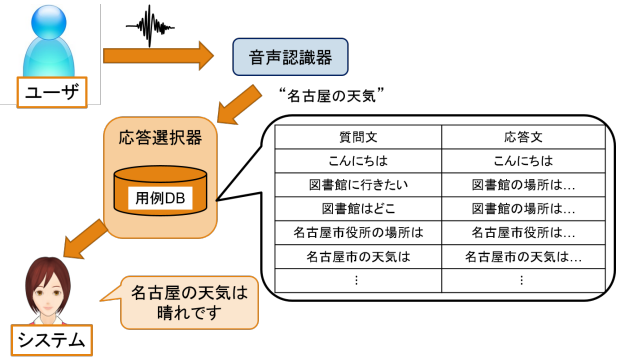

図 1　用例ベースの応答選択の概要図

2. 一問一答の用例ベースの応答選択

用例ベースの音声対話システムの概要を図 1 に示す．用例ベースの音声対話システムは，質問文と応答文の組である用例を持つことでシステムを構築するデータ駆動型の対話システムである．このシステムはユーザの発話の音声認識を行う．その後，認識結果に対し最も類似度の高い用例質問文に対する応答文をシステムの応答として選択する．

[1] 名古屋工業大学大学院 工学研究科
Nagoya Institute of Technology
[a)] makino@slp.nitech.ac.jp
[b)] ri@nitech.ac.jp

本節は用例ベースの応答選択の従来の手法と問題点を述べる．

2.1 単語情報を用いた応答選択

用例質問文と認識結果の類似度の算出する従来の手法は全ての用例質問文に対し認識結果に含まれるキーワード [1] や自立形態素の一致数 [2][3] を求め応答文をスコアリングする手法や認識単語に応答文をラベル付けするモデルを統計学習する手法 [4] など様々な手法が行われてきた．このように従来の手法はいずれも単語情報のマッチングする応答選択を行っていた．

2.2 問題点

応答選択に単語情報を扱う場合，誤認識に頑健ではないという問題がある．つまり，誤認識した単語でマッチングを行うと正しい応答が得られない点である．

音声認識器において誤認識は不可避の問題である．誤認識には音声認識器の性能によるものと音声認識器と応答選択器のタスクのズレによって起こるものがある．後者は音声対話システムのタスク固有の単語や言い回しに使われる語彙が音声認識器の言語モデルにカバーされていない場合起こる．例えば認識辞書にない単語に対して応答選択させたい場合，音声認識部で辞書にある単語で似た読みの単語を認識してしまう．そして応答選択器で誤認識した単語でマッチングを行ってしまう．

また，応答選択を精度良く行うためには用例データベースの構築に大量の対話コーパスが必要である．しかし，音響モデルや言語モデルの作成のために必要な音声コーパスは独話形式の発話が使用できるのに対し，対話システムに必要な対話コーパスはタスク毎に実際の実行環境を意識した発話対が必要であるのでデータ収集のコストが高いといえる．そのため少量のコーパス量で精度よく応答選択器を構築することを考える．

3. 音素情報のみを用いた応答選択

本節では，従来の問題点を解決するために音素情報のみを用いる手法を提案する．音素情報のみを用いる手法に関して詳細に述べるとともに，システムの全体像と実装した B-LSTM に基づく応答選択モデルについて述べる．

3.1 音素情報のみを応答選択に用いる狙い

音素のみの学習を行う場合の第一のメリットとして学習に必要なデータ量の減少が挙げられる．音素は種類数が少ないことから音素系列と応答文との関係を表すための機械学習におけるパラメータ数が少なくなるため，それを学習するためのデータ量は少なくなることが予想される．よって少量のコーパスでの学習には適していると考えられる．

第二のメリットは誤認識に対する頑健性である．誤認識がおきた場合正解単語と誤認識した単語の音素系列は似ていることが多い．例えば「天気」を「定義」と誤認識した場合，音素系列でそれぞれ表すと “t e N k i” と “t e i g i” となり “t e” や “i” など共通する部分がある．このような観点から，音素系列でマッチングを行うと正しく応答選択される確率は高くなる．よって 誤認識による応答選択器への影響の問題が軽減され，認識モデルと応答選択モデルとで音素情報は共通の音素を使えるため可搬性の高いシステムが構築できると考えられる．

一方で，デメリットとして単語の意味情報が失われるので意味情報によるマッチングができなくなることが考えられる．

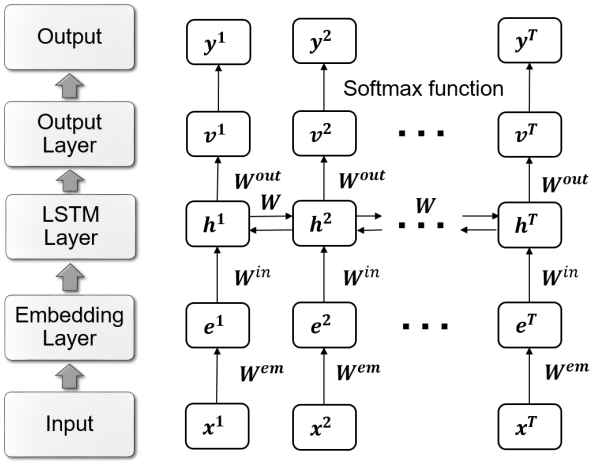

図 2 実装した応答選択モデル

3.2 B-LSTM による応答選択モデル

本研究では音素情報のみを用いた応答選択に B-LSTM (Bidirectional Long Short Term Memory) を適用する．B-LSTM は時系列を扱うニューラルネットワークでより長期の記憶が可能な LSTM を双方向の構造にしたものである．音声対話システムへのユーザからの発話は短い発話が多い．そのため単語系列より系列長が長くとれる音素系列のほうが，時系列データの要素の並び（文脈）を特徴化でき，長距離の文脈の依存関係を捉える B-LSTM によるモデリングに適していると考えられる．

3.3 応答選択モデルの構造

応答選択に適用したニューラルネットワークを図 2 に示す．表 1 のように音素系列と応答文が対応付けられたデータを用意し，対応付けられた応答文を正解ラベルとして応答文分類をするモデルの学習を行う．ニューラルネットワークへの入力は時系列データである文中の音素系列（ID を割り振ったもの）で ID に対応した次元の要素のみ 1 でそれ以外が 0 である one-hot ベクトルである．まず，入力はエンベッディング層で音素を特徴ベクトルで表される．各入力である one-hot ベクトル $\boldsymbol{x}^t$ が重み $\boldsymbol{W}^{em}$ により $\boldsymbol{x}^t$ か

表 1　データセットの例

Input label	Output label
k	こんにちは
o	こんにちは
N	こんにちは
n	こんにちは
i	こんにちは
ch	こんにちは
i	こんにちは
w	こんにちは
a	こんにちは

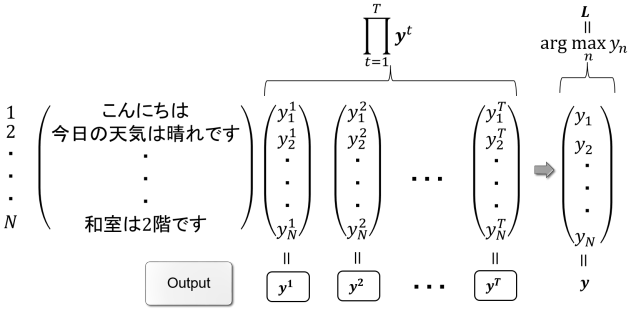

図 3　ニューラルネットの出力から応答選択までの過程

ら連続値ベクトル $\boldsymbol{e}^t(=\boldsymbol{W}^{em}\cdot\boldsymbol{x}^t)$ に変換を行い，LSTM 層の入力とする．LSTM 層で双方向の文脈を考慮した計算の後，出力層ではソフトマックス関数を使い，応答文数を N とすると入力の各音素に対し 1 つの N 次元のベクトル，つまり各応答文が得られる確率が出力される．

ここでユーザの発話に対するシステムの応答文を選択するには，図 3 のように出力の各次元は設計者が設定した各応答文に割り振られているので，文中の各音素から得られた確率の積を質問文に対する各応答文が得られる期待値とみなすことで一番期待値が高い応答文を選択する．ニューラルネットワークの出力から質問文に対する応答文を得るまでを式で表すと以下である．

$$\boldsymbol{y}=(y_1,\cdots,y_N)\approx\prod_{t=1}^{T}\boldsymbol{y}^t \tag{1}$$

$$\hat{L}=\underset{n}{\operatorname{argmax}}\,y_n \tag{2}$$

3.4　システムの全体構成

システムの全体構成を図 4 に示す．まずシステムは設計者が予め設定したキーワード K と応答文の組 A から構成されるデータベースを持つ．質問発話 O は音声波形であるので音声認識部で認識したキーワード K，それ以外の単語 G というテキストが得られる．ここではキーワードを効率よく抽出するために複数のキーワードスポッティング [5] を用いる．そして得られたテキストの音素列 P を用いて，予め学習を行った B-LSTM による識別モデルによって設

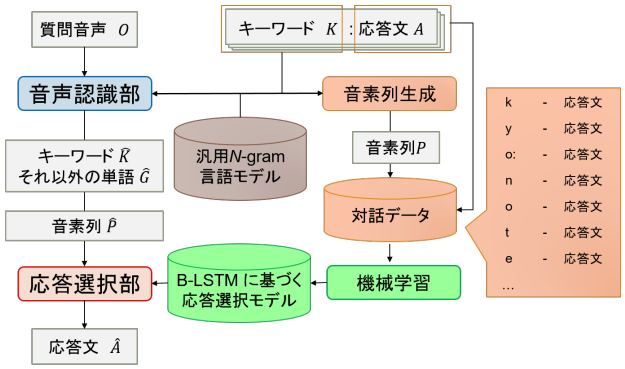

図 4　システムの全体図

計者が設定した応答文 A の中から応答選択を行う．応答選択モデルの学習に用いる対話データは設計者がシステムに与えるキーワードの音素系列，またはキーワードから汎用 N-gram 言語モデルを用いて文生成 [6] を行うことで得られた生成文の音素系列に応答文を対応付けることによって作成する．そして作成した対話データを学習データとすることで B-LSTM に基づいた応答選択モデルを作成する．

4.　評価実験

B-LSTM に基づく音素情報のみを用いた応答選択モデルを比較対象を単語情報のみを用いた応答選択モデルとして評価実験を行った．なお，音素情報のみを用いた応答選択モデルはモノフォンとトライフォンの 2 通り構築した．本節では，実験条件と結果を示す．

4.1　実験条件

4.1.1　使用データベース

実験で使用するデータベースは一問一答形式の音声対話システム「たけまるくん」[2][7] のある期間の対話履歴から作成したものを用いた．データベースの詳細を表 2 に示す．学習データに含まれる文数は文生成時の設定した生成文数によって異なり，キーワードセット種類数 499 に比例する．例えば，生成文数 10 であれば 4990 文である．本実験では生成文数を 1〜10,20,30,40 としてそれぞれ設定し，モデルの構築を行った．テストデータには 2003 年 8 月に収録された有効発話 792 文を用いている．有効発話に想定される音声対話システムの擬似的な対話データベースを構築した．ここで有効発話とは人手によって質問発話と判断された発話と定義する．実験で用いるキーワードの設定は 2002 年 11 月から 2004 年 10 月の期間で出現頻度が 2 回以上ある有効発話の単語列を形態素解析し，その単語列の中から自立語や固有名詞などを人手で設定した．本実験で用いるキーワードと応答文の例を表 3 に示す．表 3 のように，複数の単語をキーワードとみなし，応答文が対応付けられているものも存在する．これをキーワードセットと定義する．学習データは，このキーワードセットをもとに文

表 2 データベースの詳細

	学習データ	テストデータ
文数	499× 生成文数 *1	792
キーワードセット種類	499	173
キーワード種類	378	148
応答種類	168	85

表 3 キーワードセットの例

キーワードセット	応答文
たけまる こんにちは	こんにちは
すごい	すごいでしょ
友達	友達になってください
好き 人	内緒です
のど 渇い	自販機は左奥のトイレの隣にあります

生成を行い，文中の単語や音素に対し応答文を対応付けたものを用いる．

4.1.2 実験環境

文生成と音声認識に用いる汎用言語モデルには CSRC[8] で用いられた Web 言語モデル (WebLM) を使用した．WebLM は，Web からタスクを限定しないで収集したテキストコーパスを学習した単語 3-gram モデルで，語彙数 60,250 語の言語モデルである．音響モデルには CSRC[8] に含まれる全世代話者用音響モデルのうち，総状態数 3000，コードブック数 129，1 コードブックあたり 128 混合分布を持つ PTM モデルを使用した．

音声認識部のワードスポッティングのアルゴリズムは複数キーワードのワードスポッティング [5] に基づく．認識エンジンには大語彙連続音声認識エンジン Julius[9][10] にワードスポッティングを実装したものを使用した．応答選択モデルの学習に用いる文生成は，Julius-4.1.2 のライブラリを使用した．ワードスポッティング時のキーワード遷移確率を −2.3 とした．音声認識時の単語列探索時のビーム幅は 2500，言語重みは 8.0，挿入ペナルティは −2.0 と設定して音声認識を行った．その結果テストデータのキーワード認識率は 81.9 %であった．質問文生成時の最大文長は 10，単語列探索時のビーム幅は 100，単語列結合時のビーム幅は 100 である．文生成の最大文長はキーワード数は 10 である．また，文に対する単語系列を得るために形態素解析器の chasen[11] を用いた．

本実験の応答選択モデルは Python 言語の機械学習用ライブラリである TensorFlow[12] を用い，TensorFlow に用意されている RNN チュートリアル [13] を参考に構築した．実験に用いた計算機の GPU には，GeForce GTX 1080，GPU 演算のためのツールには CUDA-8.0[14] を用いている．

応答選択部への入力である単語の語彙数は WebLM から同一表記で読みが違う単語の統一を行った 50,452 語である．音素の種類数は 41 である．本実験のニューラルネットの中間層には LSTM の層を 2 層重ねた．パラメータ更新手法にバッチサイズを 20 として minibatch 確率的勾配降下法を用いた．また，ニューラルネットの各重みは $[-0.1, 0.1]$ の区間で初期化を行った．学習率は初期学習率と初期学習率学習回数，減衰率を定め，初期学習率である程度学習後，現在の学習率に減衰率を掛けて学習率を小さくして学習を行った．今回は初期学習率を 0.8，初期学習率学習回数を 25，減衰率を 0.8 とした．さらに学習時には各層の入力のドロップアウトを行った．隠れ層のサイズとドロップアウト率は定量的に変化させ，精度が一番高かったもの，また精度が停滞し始めた値を用いた．よって，単語は隠れ層のサイズは 400 ユニット，各層の入力の 50 %をドロップアウトさせた． 音素（モノフォン）は隠れ層のサイズは 200 ユニット，各層の入力の 20 %をドロップアウトさせた． 音素（トライフォン）は隠れ層のサイズは 400 ユニット，各層の入力の 50 %をドロップアウトさせた．

応答選択モデルの評価には以下の応答正解率を用いる．

$$(\text{応答正解率}) = \frac{(\text{正解した評価事例数})}{(\text{評価事例数})}$$

4.2 実験結果

B-LSTM による応答選択モデルの入力に音素系列を用いた場合と単語系列を用いた場合の応答正解率を図 5 に示す．図 5 は学習データ量として 1 キーワードセットあたりの生成文数を 1 ずつ 1〜10 まで，10 からは 40 まで 10 ずつ変えたものである．今回扱った学習データ量では音素を使った場合の精度が単語を使った場合と比べ常に上回っていた．文生成した学習データは少なく，少量の学習データ量では音素系列を学習したほうが有用であると言える．また，入力に単語を使った場合は今回の最小の生成文数 1 のときと最大の生成文数 40 のときの精度の差は大きく，音素（トライフォン）を用いた場合は差が小さく，生成文数 10〜40 の間ではグラフがほぼ横ばいになっているように少量の学習データ量で安定した精度を示した．音素（モノフォン）と音素（トライフォン）を比べると音素（トライフォン）の精度が上回っていることから，音素のコンテキスト情報を含む入力を与えることでより精度が向上したものと考えられる．

今回学習に用いる対話データに生成文ではなくキーワードの音素系列のみを応答文を対応付けることも考えられる．表 4 に対話データとしてキーワードのみを使った場合と文生成を使った場合の応答選択結果を示す．キーワードのみの音素系列の応答正解率は生成文数 1 の時の応答正解率と同等もしくはそれ以下という結果であった．キーワードの数と生成文数 1 のときの学習データの数は同一であるが，生成文を使う場合キーワード以外の発話部分もマッチングに使うのでキーワード以外の発話部分にも応答選択モデルが考慮すべき言い回しが含まれていたと考えられる．

*1 文生成を使った場合の文数

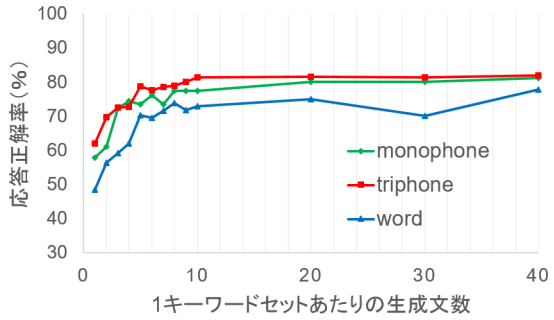

図 5 学習データ量毎の比較

表 4 キーワードまたは生成文を使った場合の応答正解率（%）

対話データ	キーワード	生成文数 1	生成文数 40
単語	49.0	49.0	77.8
音素（モノフォン）	56.9	57.8	81.3
音素（トライフォン）	60.9	62.0	81.9

表 5 書き起こし文での評価（%）

対話データ	認識結果	書き起こし文
単語	77.8	90.5
音素（モノフォン）	81.3	93.2
音素（トライフォン）	81.9	93.6

（生成文数は 40 で学習したモデルを使用）

表 6 双方向及び順方向 LSTM の応答正解率（%）

システム	単語	音素（モノフォン）	音素（トライフォン）
順方向	76.8	75.1	74.2
双方向	77.8	81.3	81.9

（生成文数は 40 で学習したモデルを使用）

また，今回の最大性能である生成文数に 40 を用いた場合のモデルでは単語は 77.8 %，音素（モノフォン）は 81.3 %，音素（トライフォン）は 81.9 %であり，生成文を使う場合文生成による学習データを増やすことで精度を上げることができる点，キーワードを用いた場合学習データ量にデメリットがあるといえる．

さらに今回認識結果だけでなく書き下し文での評価も行った．表 5 に認識結果及び書き起こし文での評価を示す．単語を使った場合と音素（トライフォン）を使った場合を比べると認識結果では 4.1 %差，書き起こし文では 3.1 %差で音素情報を用いたモデルの精度が勝っていた．書き起こし文での差に対し認識結果での差が 1.0 %増加しているので，この増加分が誤認識にロバストな応答選択ができたと言える．

また，B-LSTM だけでなく順方向の LSTM に基づくモデルの実装を行った．生成文数 40 で学習したときの順方向の LSTM と B-LSTM に基づく応答選択モデルの応答正解率を表 6 に示す．順方向 LSTM では単語を用いた場合の精度が高かった．入力した系列毎にどのような応答文が最も尤度が高いかを見ると，文末に近い系列に対しては正しい応答文が選択されているが，文頭に近い系列に対しては間違った応答文が選択されている傾向があった．RNN の構造上文頭に近い系列ほど文脈を考慮した分類がしにくいと考えられる．

一方 B-LSTM では，順方向のみを入力に扱ったときに見られた現象は解決された．その結果，単語では 1.0 %，音素（モノフォン）では 6.2 %，音素（トライフォン）では 7.7 %の精度の向上が見られた．よって，双方向性を取り入れることは有効であるといえる．双方向性を取り入れたことで精度の向上率は音素を用いるシステムが良く，単語を用いるより双方向性の構造が有効であった．

5. むすび

本研究では一問一答の用例ベースの応答選択においてよりコンパクトで識別率の高い手法として音素情報のみを用いた応答選択を提案した．音素情報に適したモデルとして B-LSTM に基づく応答選択を実装し，音素情報のみを用いた応答選択器を単語情報のみを用いた応答選択器を比較対象とした評価実験を行った．

評価実験では，応答正解率として単語情報のみを学習したモデルは 77.8 %，音素情報（モノフォン）のみを学習したモデルは 81.3 %，音素情報（トライフォン）のみを学習したモデルは 81.9 %という結果が得られ，音素情報のみを用いた応答選択モデルは単語情報を用いた応答選択モデルより 4.1 %高かった．また，学習データ量毎の比較を行い，音素情報のみを用いた応答選択モデルが常に精度が単語情報を用いた応答選択モデルよりも上回っていた．今回取り扱った少量の学習データ量では音素情報のみを用いることでよりロバストなモデルができ，B-LSTM は音素情報のみを用いる応答選択モデルに適していた．

今後の課題として，他の機械学習手法の適用やニューラルネットワークの構造をより最適化することにより精度の向上が期待できる．

参考文献

[1] 西村 竜一, 内田 賢志, 李 晃伸, 猿渡 洋, 鹿野 清宏, “Julius を用いた学内案内ロボット用音声対話システムの作成”, 電子情報通信学会技術研究報告 SP2001-99/NLC2001-64, pp. 93-98, 2001

[2] 西村 竜一, 西原 洋平, 鶴見 玲典, 李 晃伸, 猿渡 洋, 鹿野 清宏, “ 実環境研究プラットフォームとしての音声情報案内システムの運用 ” , 電子情報通信学会論文誌, Vol.J87-D2 , No.3 , pp.789-798, 2004

[3] 森本 高弘, 伊藤 仁, 鈴木 基之, 伊藤 彰則, 牧野 正三, “質問応答データベースの自動作成に基づく音声対話システムの評価”,, 情報処理学会研究報告,vol.2008-SLP-74(46),2008-12-10

[4] 平野隆司, 加藤杏樹, 南角吉彦, 李晃伸, 徳田恵一, “登録キーワードと汎用言語モデルを用いた音声認識・応答選択部の密統合に基づく統計的音声対話システム”, 情報処理学会研究報告,vol.2012-SLP-92(3),2012-07-12

[5] 加藤 杏樹, 南角 吉彦, 李 晃伸, 徳田 恵一,“音声対話システムのためのキーワードの共起制約に基づくスポッティング

アルゴリズムの評価”, 信学技報,vol.110,no.357, pp.25-30, Dec.2010.
[6] 吉見 孔孝, 南角 吉彦, 李 晃伸, 徳田 恵一,“音声対話システムのための N-gram に基づくキーワードからの文生成”, 電子情報通信学会技術研究報告,SP2009-83,PP.71-76, Dec.2009.
[7] たけまるくん
http://robotics.naist.jp/research/takemaru/
[8] 連続音声認識コンソーシアム (CSRC)
http://www.lang.astem.or.jp/CSRC/
[9] Akinobu Lee and Tatsuya Kawahara,“Recent Development of Open-Source Speech Recognition Engine Julius”,Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp.131137, Oct. 2009, Sapporo, Japan.
[10] Julius
http://julius.osdn.jp/
[11] chasen
http://chasen.naist.jp/hiki/ChaSen/
[12] TensorFlow
https://www.tensorflow.org/
[13] TensorFlow RNN Tutorial
https://www.tensorflow.org/versions/master/tutorials/recurrent/recurrent-neural-networks
[14] CUDA
https://developer.nvidia.com/cuda-zone