積極傾聴におけるあいづちのタイミングの予測

福野将人　　狩野芳伸
静岡大学総合科学技術研究科情報学専攻
mfukuno@kabolab.net　　kano@inf.shizuoka.ac.jp

1 はじめに

ユーザと対話することを目的としたシステムを対話システムといい、その中でもユーザの話に適切にあいづちを打つなどして話を聞くことに特化した対話システムを傾聴システムという。

本研究ではユーザに楽しんで話してもらうことのできる積極傾聴システムを作ることを目指して、まずはあいづちのタイミングを予測することのできる分類器の制作を試みる。またその過程でその学習・評価に適したコーパスも制作する。

2 背景

傾聴システムには入院患者や高齢者の話し相手となること[1]や、ユーザの話したい、話を聞いてもらいたいといった欲求を満たすこと[2]が期待されている。

このようなシステムでは、話し手であるユーザの発話を聞き手であるシステムが積極的に引き出し、ユーザにできるだけ楽しんで話してもらう必要があると考えた。本研究ではそのような積極的な傾聴のことを積極傾聴と呼びたい。

2.1 先行研究

我々は SVM を用いて以前にも同様の研究に取り組んだが、この際には積極傾聴ではない普通の対話を収録したコーパスである千葉大学 3 人会話コーパス[3]を用いてデータセットを制作し、システムの学習・評価を行っていた[4]。しかし本来積極傾聴を行うシステムを作るなら、その学習には積極傾聴を収録したコーパスを用いることが望ましい。

類似の先行研究としては遠隔操作のアンドロイドを用いて積極傾聴を行わせ、その対話音声を元に間休止単位ごとであいづちを行うかどうか予測したものがある[5]。本研究では生身の人間同士の対話を収録したことに加えて、音響的な特徴量に加えて言語的な特徴量も用いることでさらに自然な予測を行うことを目指した。

3 手法

まず積極傾聴の対話を収録したコーパスを制作し、それを元にあいづちのタイミングを予測することのできる分類器の制作を試みる。

3.1 コーパスの制作手法

対話音声の収録　収録のために 2 人の実験協力者を得た。2 人には対話収録のために設置した防音室に入って対面に着席し、それぞれに指向性のヘッドセットマイクを着用して対話を行ってもらった。

2 人のうちの片方には相手にできるだけ楽しんで話してもらうような積極傾聴を行うように指示した。

1 対話あたりの収録時間は当初は自然に対話が終了するまでとしていたが、この手法では実験協力者からどこで話を終えれば良いのか分からないという声が出た。そこで途中からは収録時間が 10 分を超えたところで側にいる収録担当者(音声をモニタする役目で対話には参加しない)が静かに立ち上がり、対話の終了を促すことにした。

対話は合計で 30 分 38 秒収録した。

収録音声の書き起こし・アノテーション　収録した音声のうち 18 分 44 秒についてアノテーションツール ELAN[6]を用いて書き起こし、アノテーションを行った。

付与するタグは千葉大学 3 人会話コーパスを元に先行研究[7][8]の内容を取り込んだもので、その一覧は付録に掲載した。

形態素ごとの Forced Alignment　音声と書き起こしテキストからだけでは、発話途中の任意の時刻にどの形態素まで発話されているかが分からない。あいづちは対話相手の発話の途中で起こることもあるから、後の学習のために音声認識エンジン Julius[9][10]を用いて Forced Alignment を行い、発話中の各形態素の開始・終了時刻を別途求めることにした。

Forced Alignment は対象の発話ごとにそれ専用の言語モデルを作ることで行った。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

1. 対話音声を各発話(各アノテーションの区間)で分割する
2. アノテーションから転記タグを除去し、kuromoji で形態素単位に分割する
3. よみがなタグの内容を各形態素と対応させ、各形態素のよみがなを調べる
4. Julius を用いて各形態素のよみがなを音素列に変換する
5. 各形態素と音素列を元に Julius の言語モデルを生成する
6. 1 の手順で分割した音声ファイルに対し、5 の手順で生成した言語モデルを用いて Julius で Forced Alignment を実行する

Forced Alignment の精度は、千葉大学 3 人会話コーパスの音声の 1 つ(chiba0132-A.wav)に含まれる 2 語以上の発話を用いて評価したところ RMSE で 0.105 秒だった。

3.2 対話システムの制作手法

まずコーパスを元にデータセットの生成を行い、次にそのデータセットを用いてあいづちタイミング予測モデルの学習・評価を行った。

データセットの作成　本研究で収録したコーパスと、それに加えて事前学習用に千葉大学 3 人会話コーパスを用いてデータセットを作成した。

- 特徴量の計算手法

  コーパスから 1 話者の音声を抽出し、2048 フレームごとの窓区間に分割する

  A) 図 1 に表すように、窓区間ごとの音声波形にプリエンファシスフィルタ・ハミング窓を適用後、17 次元の MFCC(メル周波数ケプストラム係数)を計算し、それぞれ直前の 128 区間ぶんの MFCC を**音響的特徴量**とする

  B) 図 2 に表すように、形態素ごとの Forced Alignment 結果をもとにその区間までに発話された形態素を 10 個求め、その単語埋め込み(100 次元)を**言語的特徴量**とする

- ラベルの計算手法

  特徴量の計算に使った話者の話し相手[i]の音声を抽出し、2048 フレームごとの窓区間に分割する。

  分割後の各区間とその前後の 100 ミリ秒以内にあいづちを含むあらゆる発話の開始があれば正、それ以外に負のラベルを付与する。

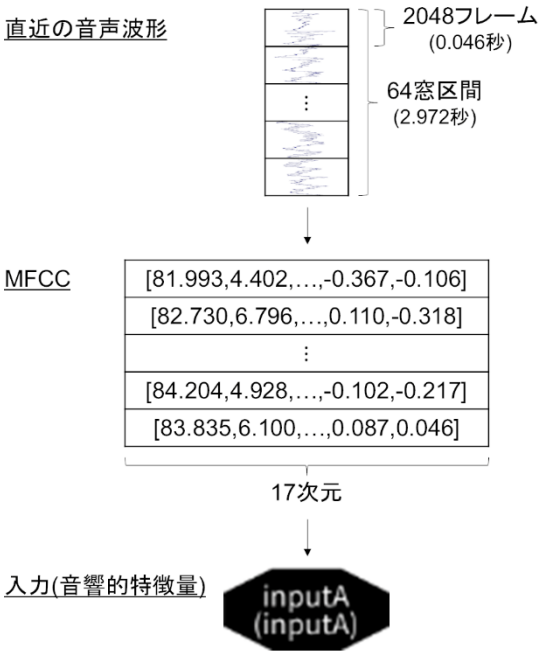

図 1　音響的特徴量の計算手法

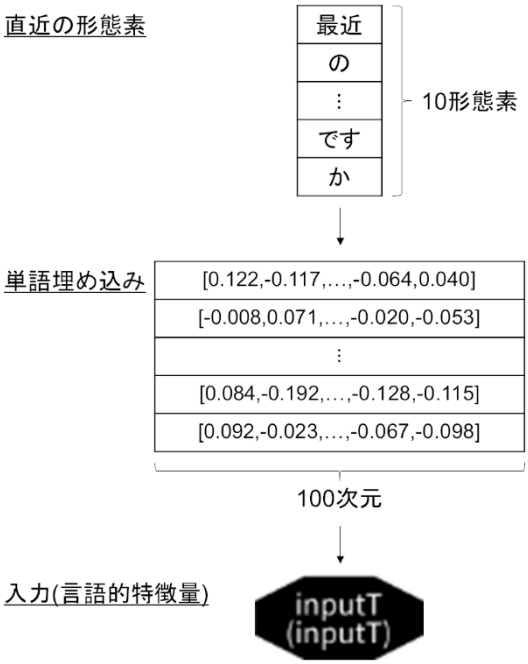

図 2　言語的特徴量の計算手法

[i] 千葉大学 3 人会話コーパスの場合は相手が 2 人いるのでそれぞれに計算する

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

MFCC の計算には内製の音響解析器 nyankyo を用いた。単語埋め込みについては学習済み Word2Vec モデルである日本語 Wikipedia エンティティベクトル[11]を用いて取得した。

またデータセットの作成後、全ての特徴量について標準正規分布に従うよう標準化を行い、その後データセットのおよそ 9 割(バッチサイズで割り切れる数に微調整した量)を学習用、残りを評価用に分割した。さらに学習用については正例を 21 倍にオーバーサンプリングし、正例と負例がおおよそ同数になるようにした。

モデルの作成　今回用いたモデルを図 3 に示す。LSTM と MLP から成る多層構造のニューラルネットワークであり、その構造は以下のようなものである。

1. 【inputA/inputT】事前に計算した音響的特徴量ならびに言語的特徴量
2. 【lstmA/lstmT】1 の出力を many to one 型の LSTM に入力する
3. 【merge】音響的特徴量・言語的特徴量それぞれで計算した 2 の出力を結合する
4. 【dense1】3 の出力を全結合層に入力する
5. 【dense2】4 の出力を全結合層に入力する
6. 【output】出力層

また各層の設定は表 1 に示す通りである。

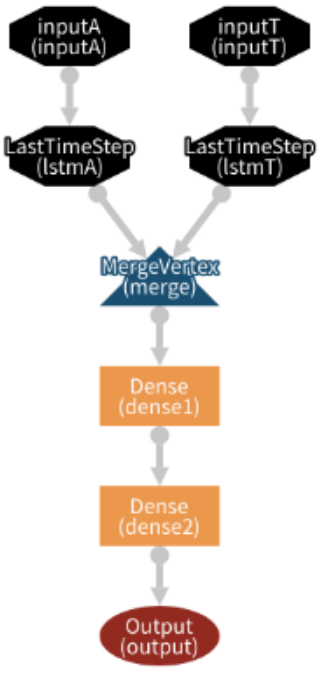

図 3　LSTM+MLP モデル

表 1　モデル各層の設定

lstmA/ lstmT	Layer Type	LastTimeStep
	Input Size	64/ 10
	Layer Size	100
	Weight Init.	xavier
	Updater	Sgd
	Activation Function	tanh
dense1	Layer Type	Dense
	Input Size	200
	Layer Size	300
	Weight Init.	relu
	Updater	Sgd
	Activation Function	relu
dense2	Layer Type	Dense
	Input Size	300
	Layer Size	100
	Weight Init.	relu
	Updater	Sgd
	Activation Function	relu
output	Layer Type	Output
	Input Size	100
	Layer Size	1
	Weight Init.	xavier
	Updater	Sgd
	Activation Function	sigmoid

学習にあたっては、まず千葉大学 3 人会話コーパスを用いて事前学習を行い、次に本研究で収録したコーパスを用いて本学習を行った。モデルの作成や学習には Deeplearning4J を用いた。

またコーパスとは別に、本研究で収録した音声の 1 つについて 4 人の協力者からアンケートを取り、あいづちを打つべきタイミング全てに印をつけてもらった。この結果平均しておよそ 13 秒に 1 回のペースであいづちが打たれることがわかったので、それを参考に手作業でモデルのチューニングを行った。

4 評価

4.1 統計的評価

データセットから評価用に分割したおよそ 1 割を用いて、Accuracy、F1 などを求めた。千葉大学 3 人会話コーパスを用いた評価結果を表 2 に、本研究で収録したコーパスを用いた評価結果を表 3 に示す。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

本学習の効果を検討するため、評価は本学習に本研究で収録したコーパスを用いた場合と用いなかった場合とでそれぞれ別々に行った。また入力に言語的特徴量を用いた効果を検討するため、モデルから言語的特徴量の部分を除いて音響的特徴量のみを用いた場合の結果も求めた。

表 2　千葉大学 3 人会話コーパスを用いた統計的評価の結果

	事前学習+本学習	事前学習のみ	事前学習+本学習(音のみ)
Accuracy	0.6032	0.0892	0.1144
F1	0.1528	0.1637	0.1639
Precision	0.0944	0.0892	0.0895
Recall	0.4014	1.0000	0.9732

表 3　本研究で収録したコーパスを用いた統計的評価の結果

	事前学習+本学習	事前学習のみ	事前学習+本学習(音のみ)
Accuracy	0.6760	0.0390	0.0666
F1	0.0663	0.0751	0.0759
Precision	0.0373	0.0390	0.0395
Recall	0.2949	1.0000	0.9824

ただ、これらの評価値は 2048 フレーム(0.046 秒)ごとに正しく予測できたかで評価されているが、あいづちを打つタイミングには前後に多少のズレが許容できると考えられる。また 3.2 項でチューニングに使ったアンケートでは回答に相当の個人差があった。このことからあいづちには実際打たれるタイミングよりも、打てるけれども打たないタイミングが多くあり、その選択に明確な正解というものは無いとも考えられる。

そこで別途アンケートで人的評価を行うことにした。

4.2 人的評価

本研究で収録した対話音声のうち、音声には問題のない軽微な収録ミス(映像の録画トラブル)が原因で学習にも評価にも用いなかった 1 分 42 秒ぶんを用いてあいづちのタイミング予測を行い、その自然さをアンケートで 5 人の協力者に評価してもらった。結果を表 4 に示す。

表 4　人的評価の結果

	事前学習+本学習	事前学習のみ	事前学習+本学習(音のみ)
平均評価(1~5)	2.80	1.60	1.80

アンケートの手法としては、対象となる音声の波形上に予測されたあいづちのタイミングが表示されるようなアンケート専用の Web サイトを作成し、それらのタイミングについて 1(全く自然でない)から 5(とても自然)の間で 5 段階評価してもらった。

なお 3.2 項で書いたように今回制作したモデルは 13 秒に 1 回ほどあいづちを打つようにチューニングしたが、アンケートへの回答後、協力者全員があいづちの頻度が低くもっと打つべきだとコメントした。

したがって頻度をより高く調整することで評価もより高いものにできる可能性がある。

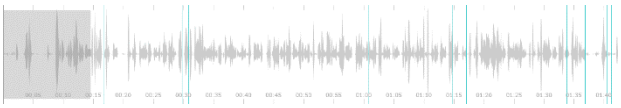

図 4　人的評価に用いた予測結果

人的評価に用いた音声の波形とあいづちタイミングの予測結果を図 4 に示す。青い縦線の瞬間が予測されたタイミングを表しており、主に各発話の間の短い無音区間があいづちのタイミングと予測されていることがわかる。

5 おわりに

本研究では積極傾聴システム制作のための第一歩としてあいづちのタイミングを予測することのできる分類器の制作を試みた。したがって入力特徴量としては既に録音・書き起こし済みのテキストを用いたが、今後はリアルタイムの実音声を入力として、音声認識と特徴量の漸次計算を行いながらあいづちのタイミングの予測を行うような積極傾聴システムに発展させたいと考えている。

またあいづちはそのタイミングに加えて、感動系応答詞、オウム返し、共同補完などの種類の使い分けも重要である。今回収録したコーパスではそれらの種類についてもアノテーションを行っているので、これを元にあいづちのタイミング予測に加えて種類の分類も行えるようにしたい。コーパス自体をさらに拡充させ公開することも予定している。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

1. 北野浩章. 応答やあいづちに用いられる照応的な「そう」について:談話データにみる自然な対話の特徴. 京都大学言語学研究. 2000, no. 19, p. 79–94. https://ci.nii.ac.jp/naid/120001712102, (参照 2019-01-31).
2. 目黒豊美, 東中竜一郎, 堂坂浩二, 南泰浩. 聞き役対話の分析および分析に基づいた対話制御部の構築. 情報処理学会論文誌. 2012, vol. 53, no. 12, p. 2787–2801. http://ci.nii.ac.jp/naid/110009493426/ja/, (参照 2020-03-20).
3. DEN, Y. A scientific approach to conversational informatics : Description, analysis, and modeling of human conversation. Conversational Informatics : An Engineering Approach. 2007. https://ci.nii.ac.jp/naid/10020493675, (参照 2019-02-06).
4. 福野将人, 狩野芳伸. 音響的特徴を用いた応答の使分け・挿入行う 傾聴対話システムの試作. SIG-SLUD. 2018, vol. B5, no. 02, p. 92–93.
5. 原康平, 井上昂治, 高梨克也, 河原達也. “相槌・フィラー予測とのマルチタスク学習による円滑なターンテイキング”. 第80回全国大会講演論文集. 2018, p. 409–410.
6. Han Sloetjes, Peter Wittenburg. Annotation by Category: ELAN and ISO DCR. 2008. http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf, (参照 2021-01-13).
7. Sakishita Masahito, Kishimoto Taishiro, Takinami Akiho, Eguchi Yoko, Kano Yoshinobu. Large-Scale Dialog Corpus Towards Automatic Mental Disease Diagnosis . Springer Verlag , 2020. http://www.scopus.com/inward/record.url?scp=85070566443&partnerID=8YFLogxK.
8. Den Yasuharu, Koiso Hanae, Takanashi Katsuya, Yoshida Nao. “Annotation of response tokens and their triggering expressions in Japanese multi-party conversations.” LREC. 2012, p. 1332–1337.
9. Lee Akinobu, Kawahara Tatsuya, Shikano Kiyohiro. Julius---an open source real-time large vocabulary recognition engine. 2001.
10. 李晃伸. Julius を用いた音声認識インタフェースの作成. ヒューマンインタフェース学会基礎講座 第三回, 2009. 2009, p. 2–8. https://ci.nii.ac.jp/naid/20000503286/, (参照 2019-01-31).
11. 鈴木正敏 , 松田耕史 , 関根聡 , 岡崎直観乾健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与. 言語処理学会 第22回年次大会 (NLP 2016). 2016. https://www.anlp.jp/proceedings/annual_meeting/2016/pdf_dir/A5-2.pdf.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A 付録

記号	種類	説明
(.)	休止	注釈内での休止
:	引き伸ばし	標準的な発音に含まれない語中、語末の音の引き伸ばし
%	詰まり	非語彙的な音の詰まり。休止との違いは非語彙的であるかどうか。ちょうど文節で区切れている場合は休止記号を付与する。
-	語の中断	言いかけた語を途中で止めている場合その中断箇所で使用する。言い直す場合が多いので言いさし記号と併用することが多い。ただし名詞の中で言いかけは言いさし記号と併用するが、連用形と動詞の言いかけは言いさし記号とは併用しない。
(F_word)	フィラー	「えっと」などのフィラー
?	上昇調	質問など上昇調に使用
(B_word)	応答形感動詞	「うん」「はい」などの応答系感動詞
(E_word)	感情表出系感動詞	「あっ」「へえ」などの感情表出系感動詞
(L_word)	語彙的応答	「なるほど」「確かに」などの慣習化された短い表現
(A_word)	評価応答	「すごい」「こわ」などの短い表現で評価する語
(P_word)	繰り返し	直前の相手発話を繰り返したもの
(C_word)	共同補完	相手発話に後続するであろう言葉を補ったもの
(T_phon|word)	言いさし (意図された語が同定可能)	言いさしたもののうち同定可能なもの。記号中の phon は実際の音列、word は意図された語を表す。
(D_phon)	言いさし (意図された語が不明)	言いさしたもののうち不明なもの
(W_phon|word)	いい誤り	いい誤りや発音の怠けなど
(K_kana|kanji)	漢字表記できなくなった文字	語中での引き伸ばしやつまりによって漢字表記できなくなった文字
(R_*)	仮名	個人情報保護のために別の名称に置き換える
(歌_*)	歌	歌いながらの発話
<声>	聞き取れない言語音	全く聞き取れないか、言語音と見なせない音声
<笑>	笑い声	発話を伴わない笑い。どれだけ長く笑っていても単に<笑> と書かれる
<息>	呼吸	ため息などの呼吸。
(Y_TEXT)	読み仮名	注釈終了時にその注釈の読み仮名をカタカナで追加
[	重複開始位置	複数の話者間で発話が重複している場合は、重複開始位置を'[' で記す。
(笑_)	笑い	笑いながら話している音声
<続>	続き	1 つ前の発話区間にまたがった短単位の後半部分
(迷:)		書き起こせない何かしらの理由

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).