不足情報を自律的に問う対話エージェントの実現に向けた聞き返しの必要性検知

大原　康平† 佐藤　翔悦† 吉永　直樹†† 豊田　正史†† 喜連川　優††,†††

† 東京大学大学院 情報理工学系研究科　〒113–8654 東京都文京区本郷 7–3–1
†† 東京大学 生産技術研究所　〒153–8505 東京都目黒区駒場 4–6–1
††† 国立情報学研究所　〒101–0003 東京都千代田区一ツ橋 2–1–2
E-mail: †{ohara,shoetsu,ynaga,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　人間は対話を行う際，相手の発言に対して応答するために十分な情報が無いと感じた場合には，不足している情報を補完するよう聞き返しを行う．しかし既存の対話システムではそのような状況を把握することができず，不自然な応答をしてしまうことが多い．そこで本稿では，不足情報を自律的に獲得する対話エージェントの実現の為に，与えられた発話に対し聞き返しの必要性の有無を判別するタスクを提案する．具体的な手法としては，大規模なマイクロブログ対話データを用いた教師あり学習による分類器を用い，2 段階で分類する．一段階目では，全ての発話から応答に疑問符と疑問詞（だれ，なに，いつ等）が含まれるものを分類する．2 段階目では，疑問符と疑問詞を含む応答を持つ発話から，情報を補うための聞き返しが必要な発話を分類する．2 段階目の分類で扱う聞き返しを必要とする発話とは，実際の応答に疑問符が含まれる対話のうち，対話を行う二者が共有している文脈を持たない第三者が，聞き返しをせずに応答できない発話とする．実験ではマイクロブログから学習・評価用データを構築し，データの一部をテストとすることで提案手法の有効性を検証する．

キーワード　自然言語処理，対話，情報補完，マイクロブログ，聞き返し

1. はじめに

近年，マイクロソフトが開発するりんな [6] のような雑談対話システムに加え，Apple の Siri(注1) のようにタスク指向型対話システムに対してもユーザとの雑談対話が発生する場合が増えており，雑談対話システムの必要性が高まっている．このような中 Twitter をはじめとするマイクロブログの出現により大規模な対話データが得られるようになったことを背景として，統計的機械翻訳やニューラルネットワークに基づく対話モデルによって雑談対話を実現しようとする研究 [4, 5] が盛んに行われている．

現状，単独の発話だけを入力として応答を出力する対話システムは，ユーザの発話に対し文脈に沿わない応答や論理的に不自然な応答をする場合がある．これには様々な理由が考えられるが，その原因の一つに，対話が発話者と応答者の間である程度閉じた世界で行われるため，互いが共有している情報を可能な限り省略して情報のやり取りを行うことから，ユーザの単独の発話には自然な（人間に近い）応答をするための情報が不足していることが考えられる．我々人間が対話を行う際は，相手の発言を理解する上で情報が不足している場合や更に会話を続ける上で知っておきたいことがある場合に聞き返しを行うことで会話の文脈を明らかにする．例えば,「お昼食べた？」と尋ねて「食べたよ」と返された場合，人であれば「何を食べたの？」等と聞き返しを行う．しかしながら，既存の対話システムは情報が不足している状況を把握できず，不自然な応答をしてしまうことが多い．先の例では,「食べたよ」と話しかけると対話システムでは「美味しいよね」等と不自然な応答を返すことがある．

本稿では，与えられた発話に対して聞き返しの必要性の有無を判別するタスクを提案し，分類器を用いてこれを解く手法を提案する．ある発話が聞き返しが必要であると識別可能であれば,「詳しく」などと聞き返しを行うことで情報を補完し，より自然な応答を返すことが可能である．また，既存の対話モデルに関しても聞き返しが必要な発話を学習データから除くことで、本来情報不足から回答不能であるような発話・応答が対話モデル構築時にノイズとして働くことを避けられると考えられる．

学習データを作成する際に，発話に疑問を返している応答の中には発話に不足している情報を聞き返している場合もあるがそうでないものも多く,「聞き返しを必要とする発話」データを集めるのは困難である．そこで本稿では 2 段階の分類器を用いた聞き返しの必要性予測手法を提案する．具体的には，全発話から応答に疑問符と 5w1h（When，Where，Who，What，Why，How）のパターン（何，いつ等）が含まれる発話を予測する分類器と，疑問符と聞き返しに頻出するパターンが応答に含まれるような発話のうち応答が聞き返しとなるような発話を判別する分類器を用い，段階的に聞き返しを予測する手法である．この 2 段階の分類器を用いることで，学習用データ作成の困難さが緩和される．緩和される理由は，一段階目は簡単な文字列のパターン一致で学習データを収集することができ，2 段階目は疑問符と 5w1h のパターンが応答に含まれる対話の中か

(注1)：http://www.apple.com/jp/ios/siri/

ら聞き返しが必要である対話をラベル付けすれば良いためである．聞き返しが発生する原因としては発話が情報不足であることや急に話題が転換すること，応答者にとって未知の単語が出現すること等が影響すると考えられるため，そういった情報を捉えられるような特徴量を設計した．

学習・評価用データとしては，研究室で収集している Twitter データセットにおけるメンションのやり取りを対話として用いた．本稿では 2 段階の分類器のそれぞれの性能を適合率・再現率・F 値によって評価し，どの特徴量が有効であるか評価するための ablation test を行った．分類器は共にランダムに予測する方法より性能が良い結果となった．

2. 関 連 研 究

不足情報を問うための聞き返しの必要性の有無を判別する本研究と同一のタスクに取り組む研究は著者の知る限り存在しない．しかし雑談対話を続けるために適切な質問や応答を返すという点で関連がある対話システムの研究について述べる．

近年，音声対話システムの分野において，相手の話す内容に合わせ，適切な応答を行うことで相手の話したい内容を引き出す傾聴対話についての研究が行われている [2, 8, 10, 13]．Han ら [2] は発話を意図に応じてラベル付けし，それぞれの意図について応答を返すシステムを提案した．主観的評価実験を行い，従来システムより満足度が高い結果を得た．石田ら [13] はテキスト上の雑談対話において，発話の焦点となる単語と疑問詞との n-gram 確率を計算，その結果を元に質問や応答を生成する手法を提案し，雑談対話コーパス [14] を使用して評価を行った．下岡ら [8] は，音声認識の信頼度が高い場合は繰り返し/問い返し応答と共感応答を生成し，そのいずれかから選択して応答を返す傾聴システムを提案した．評価では 110 名の一般被験者に対する実験を行い，77%のユーザ発話に対して対話を破綻させることなく応答を生成した．これらの研究は，発話のみの情報から傾聴対話のラベルに分類または応答を返しているのに対して，本研究では発話に至るまでの対話の情報も用いている点で異なる．

傾聴対話に関しては，会話を円滑に進める上で重要な相槌に関する研究も行われている [9, 12]．山口ら [12] は発話の区切りに出現する相槌を対象に，先行の発話と相槌の関係の分析を行った．それに加え，先行発話の統語構造や韻律的特徴から相槌の形態の予測と生成を行い，その有効性を示した [9]．これらの研究は，発話に対してどのように相槌を打つことが適切かに着目した研究であるが，本タスクは不足情報を補うための聞き返しの必要性の有無を判別するものである．

さらに，傾聴対話に関する研究ではないが，Li ら [3] はユーザの質問と知識ベースとの齟齬を聞き返しによって解消する対話システムを提案した．評価ではユーザの発話に対して聞き返しを行うシステムと聞き返しを行わないシステムの性能を比較し，聞き返しによって質問応答システムの性能が向上することを示した．ユーザの発話からだけでは適切な応答ができない際に聞き返しを行うことでシステムの応答性能を向上させるという目的は本研究と同じであるが，この研究では発話に対する聞き返しの必要性検知には取り組んでおらず，本研究の取り組みとは異なる．

これらの関連研究の多くは小規模な対話データセットによる実験を行っている．一方で，本研究では分類を 2 段階とすることで部分的に大規模なデータを用いた教師あり学習を行い発話の分類を行った．また，注目する発話だけでなく発話に至るまでの対話の情報も利用した実験を行った．

3. 提 案 手 法

本節では，提案手法について述べる．概要を図 1 に示した．提案手法は 2 段階の分類器から構成される．学習データを作成する困難さのため 2 段階とした．具体的には，全対話データの中で発話に不足している情報を聞き返しているものは少数であり，聞き返しが必要な発話データを収集するのは困難である．そこで，聞き返しの必要条件を満たす応答を自動で集めて分類器を構築してフィルタとして利用し，その結果を実際の聞き返し対話に分類する分類器を人手で構築したデータから学習する．

1 段階目の分類器は全ての発話から応答に疑問符と 5w1h を表すパターンを含むものを判別するモデルである．パターンとは「いつ，どこ，だれ，誰，なんだ，なに，何，なぜ，なんで，どうして，どうなん，どんな，どのよう，どっち」である．ただし，このパターンを応答に含む対話のうち疑問文で用いられる代名詞でない「あいつ，誰かと，だれかと，なにも」はパターンから除いた．2 段階目の分類器は上記の応答を持つ発話のうち，応答が発話で不足している情報を補うための聞き返しを行っているかを判別するモデルである．以下では 1 段階目の分類器を「疑問符+5w1h 応答予測器」と、2 段階目の分類器を「聞き返しの必要性予測器」と呼ぶ．

聞き返しの必要性予測器を学習させる際，4.1 節で述べるように，人手によるアノテーションを行ったため，学習用データセットは小さく，それ自体から特徴を学習することは難しい．そこで，発話や着目する発話までの対話の中で出現した単語から抽出する特徴量の一部を別の大規模データを利用することで得た．利用する特徴量については 3.1 節で述べる．この特徴量を使い 2 つの分類器を学習した．

3.1 特 徴 量

聞き返しの必要性の有無を分類する際，発話が情報不足であることや話題転換，応答者にとって未知の単語が出現すること等の情報が有効であると考えられる．そこで，そのような情報を捉えるために，以下の聞き返しが生じる場面を想定した特徴量を用いた．なお，特徴量を抽出するため，必要に応じて NEologd v0.0.5 [11](注2) を辞書に用いた形態素解析器 MeCab により形態素解析を，JDepP を用いて係り受け解析をそれぞれ行った．

3.1.1 情報の欠損を捉える特徴量

名詞数　発話の長さがある一定以下である，つまり発話が短いことが聞き返しに影響すると仮定し，名詞の数という形で特徴

(注2)：https://github.com/neologd/mecab-ipadic-neologd/releases/tag/v0.0.5

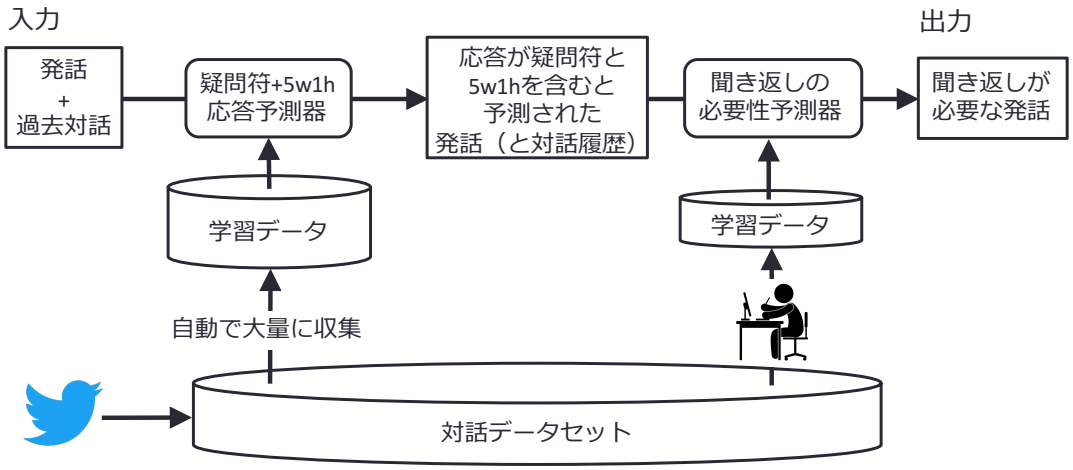

図 1　提案手法概要

量とした．本稿においては実験的に，名詞が 3 回以上出現するか否かを特徴とした．

代名詞の有無　発話における代名詞が何を指すのか応答者が理解できない場合がある．そこで，代名詞の有無を特徴とした．

動詞の必須格の有無　発話中に存在する動詞に係るべき助詞が不足している場合に聞き返しが必要になる場合が多い．具体的には，以下のような対話が考えられる．

> A：食べたよ
> B：何を食べたの？

動詞「食べる」はヲ格で食べ物を示すことが多いが，この発話では欠落している．そこで，4. 節で述べたデータセットの中で一日分の発話について動詞に係る文節を収集し，その中で助詞の頻度を数えて，その中でもっとも頻度が高い助詞で示される格を必須格とみなす．その上で，発話中の各動詞が必須格を伴う場合には必須格が有りとし，それ以外の場合は必須助詞が無しとした．

3.1.2　発話者のみが知りうる話題を捉える特徴量

固有名詞の有無　発話における名詞の中でも固有名詞に関しては，応答者が知らない場合それが何（誰，いつ，どこで）なのか聞き返す場合が多いと考えられる．そこで，発話における固有名詞の有無を特徴量とした．

形態素の出現頻度順位　発話において応答者が知らない単語または形態素が含まれる場合に聞き返しが発生しやすいと考えられる．そこで，4.1 節で述べる方法で学習した形態素の出現頻度を使用し，発話に含まれる形態素の出現頻度を特徴量とする．出現頻度の低い形態素は，未知語として，聞き返しの必要性を検知する際に有効な特徴量であると考えられる．

本稿では，形態素を出現頻度により，4 つの区分に分け特徴とした．具体的には，2013 年の研究室のツイートデータを全ての月のデータが含まれるよう 100 分の 1 にサンプリングしたデータにおける出現回数 100 未満，100 以上 1,000 未満，1,000 以上 10,000 未満，1,000 以上，の 4 つである．

3.1.3　話題の新規性を捉える特徴量

発話以前の対話に出現した名詞の有無　発話において，それまでの対話で出てこなかった名詞が現れる場合，応答者が，その名詞がなぜ出現したのかまたは何なのか理解できない場合に聞き返しが発生すると考えられる．

発話以前の対話に出現した動詞の有無　聞き返しが発生する一つの場合として，発話者の発言がそれまでの対話の流れに沿わないものである時がある．これは，発話以前の対話に出現した名詞の有無でも同様であるが，そういった，急な話題転換を特徴として捉えようとした特徴量である．

3.1.4　n-gram

1,2-gram　特定の語彙が分類に影響することを考え，発話の表層の n-gram と，発話に至るまでの対話で出現した発話者による投稿に含まれる表層と応答者による投稿に含まれる表層の n-gram を特徴量とした．n について，疑問符+5w1h 応答予測器には 1 と 2 を，聞き返しの必要性予測器では 1 のみを用いた．聞き返しの必要性予測器で 2-gram を用いないのは，学習用データが小さいために学習できず，ノイズとて働き得ることが想像されるためである．

発話から特徴量を抽出する際に使用したデータセットについて詳細を述べる．まず，形態素の出現頻度区分の特徴を抽出する際に使用したデータセットは，4. 節で述べた研究室のデータセットのうち，2013 年の全ての期間のツイートを 100 分の 1 にサンプリングしたものを使用した．合計のツイート数は，約 2,900 万件であった．

次に，動詞における必須助詞の有無の特徴量を抽出する際に使用したデータセットは，4. 節のうち，一般的な日（元旦やクリスマス等でない）である 2013 年 11 月 6 日を選択し，この日のツイートをデータとして用いた．

4.　聞き返し対話データセット

本節では，発話に対する聞き返しの必要性の有無を分類する際に学習と評価で使用したデータセットについて述べる．本研究では，モデルの学習と評価用の対話データとして，マイクロブログサービスの 1 つである Twitter から収集したデータを使

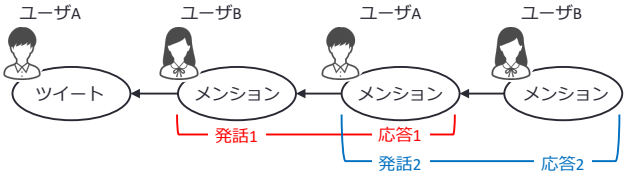

図 2　Twitter データと対話データ．赤字と青字の発話と応答のペアがそれぞれ 1 件の対話．

用する．本稿で用いる Twitter データセットは，著者らの研究室において 2011 年 3 月より継続して蓄積しているものであり，約 150 万の公開ユーザからタイムラインを継続的に収集したものである．収集対象のユーザについては，2011 年 3 月に約 30 名の著名な日本人ユーザを選び，それらのユーザがメンションやリツイートを行ったユーザを更に収集対象として拡大している．2017 年の 1 月時点では，累積約 380 億のツイートが保存されている．この中から 2013 年と 2014 年のデータの一部を実験と評価に用いた．

4.1　学習用データセット

Twitter データセットのうち聞き返しの必要性予測器には，一般的な日（元旦やクリスマス等でない）である 2014 年 6 月 18 日に投稿されたツイートをデータとして用いた．疑問符+5w1h 応答予測器には 2014 年 6 月 18 日を除く 2014 年のツイートをデータとして用いた．

対話としては，お互いが相手と話していることを認識している状態での発話と応答のペアを得ることを考える．具体的にはあるユーザ A のツイートに対して別のユーザ B がメンションをし，更にユーザ A がユーザ B へメンションを返しているもののみを対話とし，この時の，ユーザ B のメンションを発話，ユーザ A のユーザ B へのメンションを応答とする．この発話と応答の 1 ペアを 1 件の対話とする．図 2 における Twitter のツイート・メンションのデータからは，2 件の対話が得られる．2014 年では約 1487 万件の対話が存在し，2014 年 6 月 18 日の対話は約 65000 件存在した．更にこの対話データを以下のようにクリーニングを行う．

- Twitter 特有の表現を削除．引用を示す記号である RT や QT，相手への返信を表す記号の *@user_name* は削除した．
- スクリーンネームに「bot」（大文字小文字区別なし）が含まれるユーザは削除した．ボット（人ではなく，自動的に投稿を行うプログラム）が含まれる対話は，人同士の対話を扱う本研究の目的には沿わないためである．
- 2 名のユーザが交互にメンションをしているやりとりのみ対象にし，その他のデータは除去した．
- 発話が極端に短い発話や長い発話は削除．発話の文字数が 2~24 字以外の対話は除去した．

以下では 2 つの予測器で学習・評価に用いた対話データについてそれぞれ詳細を述べる．

疑問符+5w1h 応答予測　クリーニング後のデータから正例として，応答に疑問符（半角全角）と 3. 節で述べた 5w1h のパターンを含む対話を集めることを考える．簡単なルールで収集

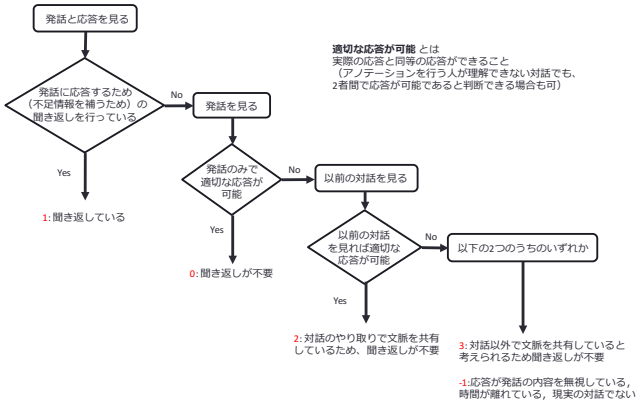

図 3　アノテーションガイドライン

表 1　各ラベルの詳細と件数

ラベル	発話の詳細	件数
0（負例）	聞き返す必要がない	72
1（正例）	会話を継続するため，不足情報を補うために実際に聞き返している	125
2（負例）	聞き返しておらず，対話のやり取りから文脈を共有しているため聞き返しが不要となった	237
3（負例）	聞き返しておらず，対話以外で 2 者間で共有している文脈があるために聞き返しが不要となった	37
-1（負例）	応答が発話の内容を無視している，時間が離れている，現実の対話でない	53

可能であるため，2014 年全体の約 1487 万件の対話を対象とし，197,072 件の正例を得た．

聞き返しの必要性予測　クリーニング後のデータから正例として聞き返しが実際に行われている発話を集めることを考える．聞き返しが発生する場合，応答には疑問符と 5w1h のパターンを含まれると考え，応答に疑問符と 5w1h のパターンが含まれる対話について人手によるアノテーションをすることで，実際に聞き返しが行われている対話のデータを得た．

対象とした 2014 年 6 月 18 日の対話データのうち応答に疑問符と 5w1h のパターンが含まれる対話 524 件存在した．アノテーションは著者が行った．具体的には，ある対話に対して図 3 のようなフローチャートに従い発話を 5 つのラベルへ分類した．各ラベルの件数としては表 1 のようになった．各ラベルの実例を以下に示す．下線が聞き返しの有無を判定する発話を示す．

ラベル 0　「聞き返す必要がない」を表す．

> A：夏休みだよ！
> B：うらやましいいつまで？

ラベル 1　「会話を継続するため，不足情報を補うために実際に聞き返している」を表す．

> A：絶対体に悪いっすよね…
> B：なにが？

ラベル 2　「聞き返しておらず，対話のやり取りから文脈を共

有しているため聞き返しが不要となった」を表す.

B：北海道の七夕、八月七日なのよくわからないな。
A：な。よくわからないな。
B：っしょ？何の関係があんだよな。

発話だけでは，何が「よくわからないな」なのか不明なので，人間同士であれば「何が？」といった応答が自然と出てくると考えられるが，応答では聞き返していない．そこで以前の発話を見ると,「北海道の七夕」に関する話題であることがわかるため，このラベルをつけた.

ラベル 3　「聞き返しておらず，対話以外で 2 者間で共有している文脈があるために聞き返しが不要となった」を表す.

B：攻撃体力タイプ…だと…
A：回復力マイナスの未来しかみえない
B：この属性だとディノライダーわんちゃんか？なにげに 2 枚おるんだが w

発話から，ゲームの話題だと考えられるが，何のゲームかは不明である．そして応答では聞き返しが発生しておらず，直前の発話を見ても，何のゲームかは不明である．よって，この対話以外での文脈共有が行われていると考え，ラベル 3 をつけた.

ラベル-1　「応答が発話の内容を無視している，時間が離れている，現実の対話でない」を表す．このラベルは，発話に対して応答が時間的・意味的ずれが発生しているものを指す.「現実の対話でない」とは，ユーザがアニメ等のキャラクターを演じて対話を行っているものを指す.

A：おはよー
B：おはよー
A：おはよおおお
B：わーい！
A：うぇーい！
B：いつからから凸復帰するん？

これは，応答が発話を無視しているため，-1 とした.

以上のアノテーションを行ったデータにおいてラベル 1 の「会話を継続するため，不足情報を補うために実際に聞き返している」をつけたもののみを聞き返しが必要とし，この 125 件を正例とした．他のラベルの 399 件を負例とした.

5. 実験・評価

本稿では，2 段階の分類器（疑問符+5w1h 応答予測器と聞き返しの必要性予測器）についてそれぞれ独立に性能の評価を行った.

5.1 疑問符+5w1h 応答予測器

4.1 節で述べた疑問符+5w1h 応答予測器のためのデータセットから学習データ，ハイパーパラメタ調整用の開発データ，テスト用の評価データを構築した．正例と負例の数に偏りがあるため，学習データは正例と負例の割合が同じになるようにし，開発データと評価データは実データでの正例負例の割合と同じになるように負例をダウンサンプリングした．学習データは全体で 319,241 件，そのうち正例が 159,629 件となった．開発データと評価データはそれぞれ 19,972 件で，そのうち正例は 265 件となった．疑問符+5w1h 応答予測器の評価には，正例を判別する性能を見るために適合率，再現率，F 値を使用する.

表 2　疑問符+5w1h 応答予測器とベースラインとの性能比較

	適合率	再現率	F 値
全て正例と予測	0.0133	**1.000**	0.0262
ランダム	0.0132	0.499	0.0258
提案手法	**0.0228**	0.472	**0.0435**

表 3　疑問符+5w1h 応答予測器の ablation test．太字は各列での最小値.

特徴量	適合率	再現率	F 値
全て	0.0228	0.472	0.0435
– 情報の欠損を捉える	0.0230	0.468	0.0439
– 発話者のみが知りうる話題を捉える	0.0225	0.460	0.0429
– 話題の新規性を捉える	0.0230	0.475	0.0439
– 発話の 1,2-gram	0.0219	**0.411**	0.0416
– 発話に至るまでの対話中の 1,2-gram	**0.0194**	0.555	**0.0374**

分類器としては，オンライン学習器である Passive Aggressive アルゴリズム [1,7] の実装である opal(注3) を用いた．本実験では線形カーネルを使用し，c パラメータを開発データを用いて調整した．c パラメータは $2^{-10} \sim 2^{2}$ まで 2 の乗数で変化させ，もっとも F 値が高かったモデルを評価データに対して適用した性能を見る．比較のためのベースラインとして，全て正例と予測する手法，完全にランダムにラベルを予測する手法の 2 つを使用した．ランダムの手法においては，100 回ランダムに予測（50%の確率で正例と予測）して算出した値の平均をとったものとした．結果を表 2 に示す．適合率と F 値は 2 つのベースラインを上回り，再現率はランダムと近い値をとった.

次に，有効な特徴量を評価するために ablation test を行う．各特徴量をいくつかにまとめて除去した時の性能を見る．具体的には，以下のように 3.1 節で述べた特徴量をまとめ除去する．この ablation test においても，開発データで c パラメータを調整し，もっとも性能が良いモデルを評価データに適用した際の性能を見る．結果を表 3 に示す．1,2-gram 以外の特徴量を除去した場合の F 値は，全ての特徴量を使用した場合と比較して，0.0006 までの差しかなく，全ての特徴量を使用した時の結果と各特徴量を除去した時の結果の p 値を計算したところ 5%よりも大きく統計的に有意でなかった．発話の 1,2-gram を除去すると F 値は小さくなるが，p 値は 5%より大きく統計的に有意でなかった．一方，発話に至るまでの対話における 1,2-gram 特徴量のみを除去した場合は，全ての特徴量を使用した場合と比較して F 値は小さくなっており，p 値の計算によっても統計的な有意性が示された．この結果から発話以前の対話における情報が有効であることが示された．1,2-gram 以外の特徴量が

(注3)：http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/opal/

表 4　5-分割交差検定による聞き返しの必要性予測器とベースラインとの性能比較

	適合率	再現率	F 値
全て正例と予測	0.238	**1.000**	0.385
ランダム	0.239	0.503	0.323
提案手法	**0.324**	0.664	**0.435**

表 5　聞き返しの必要性予測器の ablation test．太字は各列での最小値．

特徴量	適合率	再現率	F 値
全て	0.324	0.664	0.435
– 情報の欠損を捉える	0.338	0.672	0.449
– 発話者のみが知りうる話題を捉える	0.311	0.656	0.422
– 話題の新規性を捉える	0.322	0.648	0.429
– 発話の 1-gram	**0.278**	**0.616**	**0.382**
– 発話に至るまでの対話中の 1-gram	0.298	**0.616**	0.400

有効でない原因として，特徴量の取り方が良くなかったことや，データセットで用いた Twitter の対話は知人同士のやり取りであり，文脈を共有していることが多く疑問が発生しづらいことが考えられる．

5.2　聞き返しの必要性予測器

実験の評価には，4.1 節で述べた聞き返しの必要性予測器のためのデータセットを正例負例の割合が偏らないように 5 分割し，1 つを評価用，残りの 4 つを学習用とした交差検定を行った．学習用のデータでは正例と負例の偏りをなくすため，以下の処理を行った．データセット 524 件のうち正例が 125 件であったため，1 度の検定の学習用データには 100 件の正例が含まれる．そこで学習用の負例の数も 100 件にダウンサンプリングした．

評価には，適合率，再現率，F 値を使用し，5 回のテストにおける平均値を算出した．比較のためのベースラインとして，全て正例と予測する手法，完全にランダムにラベルを予測する手法の 2 つを使用した．ランダムの手法においては，5 分割したデータの 5 回のテストそれぞれに対し 100 回ランダムに予測（50%の確率で正例と予測）して算出した値を，5 回の結果で平均をとったものとした．結果を表 4 に示す．適合率と F 値は 2 つのベースラインを上回り，再現率はランダムの手法よりも高い値をとった．

次に，聞き返しの必要性予測器に関する ablation test を行う．全ての特徴量から，特徴量を疑問符+5w1h 応答予測器の ablation test 同様にまとめて除去した時の性能を見る．結果を表 5 に示す．情報の欠損を捉える特徴量を除去すると F 値は大きくなってしまい，発話者のみが知りうる話題を捉える特徴量と発話者のみが知りうる話題を捉える特徴量をそれぞれ除去した場合は F 値が小さくなる結果となった．しかし，いずれも p 値を計算すると 5%より大きく，統計的に有意でなかった．一方で発話の 1-gram を除去した場合は F 値が小さくなる結果であり，p 値の計算によっても統計的な有意性が示された．また，発話に至るまでの対話中の 1-gram を除去した場合も，全ての特徴量を使用した場合と比較して F 値は小さくなるが，p 値の計算から統計的な優位性は示せなかった．聞き返しが生じる場面を想定した特徴量が有効でない原因としては，特徴量の取り方やデータセットが知人同士の対話である点，実験に使用できたデータが少なかったことが考えられる．

6.　まとめと今後の課題

本研究では，発話とそれ以前の対話を入力として，発話に対する聞き返しの必要性の有無を判別するタスクに取り組んだ．

聞き返しが必要な発話を検知するために，オンライン学習器を使い 2 種類の分類器を提案した．第一段階では，全ての発話から応答に疑問符と疑問詞（だれ，なに，いつ等）が含まれるものを予測する疑問符+5w1h 応答予測器を用いて応答をフィルタし，第 2 段階では，疑問符と疑問詞を含む応答を持つ発話のうち情報を補うための聞き返しが必要な発話を判別する聞き返しの必要性予測器を用いる．発話が情報不足であることや急に話題が転換すること応答者にとって未知の単語が出現すること等が影響すると考えられるため，そういった情報を捉えられるような特徴量を設計した．また，1,2-gram の特徴量も入力した．2 種類の分類器は共にランダムな手法よりも良い性能が得られた．一方で ablation test をしたものの，聞き返しが生じる場面を考慮した特徴量は有意に有効ではなかった．学習・評価用データとしてはマイクロブログの一つである Twitter の大規模データを使用し，対話データセットと人手でアノテーションすることでラベル付き聞き返し対話データセットを作成した．

本稿では，2 種類の分類器を使いそれぞれの性能の実験と評価を行ったが，実際には 2 種類の分類器を合わせて使用することを想定している．そこで，発話全体から聞き返しが必要な発話を検知する 2 種類の分類器を連結して用いた場合の性能の評価を行うことが必要である．また，聞き返しの必要性予測器は筆者がアノテーションを行った非常に小規模なデータによって訓練されている．今後クラウドソーシングなどを用いたより大きなデータセットの作成とそのデータセットを用いた実験と評価を検討している．さらに，発話に対して聞き返すかどうかは応答者や応答者と発話者の関係に依存するという課題がある．本稿で作成したデータでは聞き返しが発生していないが，人によっては聞き返す発話も存在しているということである．対話システムが聞き返すべき発話はどのようなものであるかということに関係する大きな課題であるが，少なくとも人が実際に聞き返している発話に関しては聞き返すべきだと考え，今後は本稿での評価で用いた再現率を高める特徴量やモデルを考えたい．また，本研究の手法を用いて適切な聞き返しを行う対話システムの作成と評価も検討している．

謝　　辞

本研究の一部は JSPS 科研費 16K16109 と 16H02905 の助成を受けたものです．

文　　献

[1] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algo-

rithms. *Journal of Machine Learning Research*, Vol. 7, No. Mar, pp. 551–585, 2006.

[2] Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 129–133, September 2015.

[3] Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. Learning through dialogue interactions by asking questions. *To appear in ICLR 2017*, 2017.

[4] Alan Ritter, Colin Cherry, and William B Dolan. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 583–593. Association for Computational Linguistics, 2011.

[5] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[6] Xianchao Wu, Kazushige Ito, Katsuya Iida, Kazuna Tsuboi, and Momo Klyen. りんな: 女子高生人工知能. 言語処理学会第 22 回年次大会 発表論文集, 2016.

[7] Naoki Yoshinaga and Masaru Kitsuregawa. Kernel slicing: Scalable online training with conjunctive features. In *Proc. of the 23rd COLING*, pp. 1245–1253, 2010.

[8] 下岡和也, 徳久良子, 吉村貴克, 星野博之, 渡部生聖. 音声対話ロボットのための傾聴システムの開発. 自然言語処理= Journal of natural language processing, Vol. 24, No. 1, pp. 3–47, 2017.

[9] 山口貴史, 井上昂治, 吉野幸一郎, 高梨克也, Nigel G. Ward, 河原達也. 傾聴対話システムのための言語情報と韻律情報に基づく多様な形態の相槌の生成. 人工知能学会論文誌, Vol. 31, pp. C–G31_1–10, 2016.

[10] 杉山享志朗, 吉野幸一郎, 田中宏季, 近藤真人, ニュービッグ グラム中村哲. 傾聴対話コーパスの作成と知識獲得行為の分析. 人工知能学会全国大会, 2016.

[11] 佐藤敏紀, 橋本泰一, 奥村学ほか. 単語分かち書き用辞書生成システム neologd の運用-文書分類を例にして. 研究報告自然言語処理 (NL), Vol. 2016, No. 15, pp. 1–14, 2016.

[12] 貴史山口, 幸一郎吉野, 克也高梨, 達也河原. 多様な形態の相槌をうつ音声対話システムのための傾聴対話の分析. 第 77 回全国大会講演論文集, 第 2015 巻, pp. 145–146, 2015.

[13] 石田真也, 井上昂治, 中村静, 高梨克也, 河原達也. 傾聴対話システムのための多様な聞き手応答の生成. 第 78 回情報処理学会全国大会, 2016.

[14] 東中竜一郎, 船越孝太郎ほか. Project next nlp 対話タスクにおける雑談対話データの収集と対話破綻アノテーション. *SIG-SLUD= SIG-SLUD*, Vol. 4, No. 02, pp. 45–50, 2014.