

# Web から獲得した知識に基づく雑談対話システム

## Chat-like Conversational System using Large Knowledge Acquired from the Web

江頭 勇佑  
Yuusuke Egashira<sup>†</sup>

柴田 知秀  
Tomohide Shibata<sup>†</sup>

黒橋 禎夫  
Sadao Kurohashi<sup>†</sup>

### 1 はじめに

観光案内や施設案内、切符の予約システムといった、ある特定のタスクを遂行するためにユーザと自然言語を用いて対話を行なうシステムはこれまでも多く研究されており、またそういった課題遂行型の対話システムは我々の身の回りにおいて少しずつ実用化され始めている [1, 2]。その一方で、高齢者や子どもの話し相手となるため [3]、あるいはエンターテインメントとして、話すこと自体を目的とした対話、すなわち雑談を行なうような非課題遂行型の対話システムの必要性も高まっている。Eliza [4] をはじめとした非課題遂行型の対話システムでは、特定の分野について構築された知識データを必要とせず、特定のドメインに限定しない発話に対して応答可能であり、ユーザはシステムと自由な対話を行なうことができる。しかしそれらのシステムではその発話のほとんどが相槌やユーザ発話のオウム返しとなっており、対話によって何らかの情報を得ることができるといった形の意味付けが難しい。

そこで本研究では、ユーザとの自由な対話を行いつつユーザにとって未知の情報を提供する対話を続けることができるシステムの開発を目指し、Web から獲得した知識に基づいてユーザとの雑談対話を遂行する対話システムを構築した。このシステムでは Web から得た情報と、各種の言語処理モジュールを組み合わせることにより、特定のドメインに限定しないような雑談対話を行なう。具体的には、Web から獲得したニュース記事や Wikipedia 等の知識、および質問応答等の自然言語処理の技術を用いたモジュールを組み合わせることにより、ユーザとの雑談対話を実現する。

### 2 関連研究

Web 上の情報を雑談対話に利用する研究として、ウェブニュースを発話の情報源とした対話システムを構築した水野らの研究がある [5]。水野らは雑談において相手を飽きさせない要因として、本研究と同様にシステムがユー

ザにとって未知である情報を提供することを挙げており、次々に更新されるウェブニュースを、ユーザにとって未知である可能性の高い情報として提供できるとしている。このシステムではユーザ発話を検索クエリとしてニュースの記事を検索し、その記事内の文の中でもっとも要約としてふさわしいものをシステム発話として選択する手法をとっているが、対話はユーザ発話と一問一答の形式であり、対話の継続については考慮されていない。

吉野らは知識源からの情報の抽出とシステム発話の手がかりとして述語項構造を利用し、ユーザ発話と意味的に近いシステム発話を生成する手法を提案している [6]。そこでは対話のドメインをプロ野球に関することに制限した上で、Wikipedia やニュースサイトなどの Web 上に存在する情報を用いて、ユーザの質問に回答する形の対話を行なう音声対話システムを構築している。またユーザが発話しない無音時間を検出し、かつ保持している情報の中にそれまでのユーザ発話と関連性のある情報が存在した場合には、その情報の推薦を行なっている。

また鳥澤は、従来のような課題遂行型の対話システムではシステム内に詳細で複雑な知識を保持し、入力される多様な自然言語表現とそれらを結びつける必要性から、回答できる範囲が限られてしまう領域依存性が強いということが問題であるとした [7]。領域依存性を排除するために、ユーザからシステムへと向かうという情報の流れを想定した「ユーザヘインタビューを行なう」というタスクを設定することで、自動獲得された言語知識のみで有意義な対話を実現している。

本研究ではこれらの関連研究をふまえ、Web から獲得した知識をもとにシステム発話を生成する複数のモジュールを組合わせて用いることで、自由な対話を行ないつつシステムからユーザへ新たな情報を提供できる雑談対話を行なうシステムを構築する。

### 3 構築したシステム

構築した対話システムを図 1 に、その動作の概要を図 2 に示す。

システムはあらかじめ取得しておいたウェブニュース、Wikipedia の定義文データベースを保持しており、これ

<sup>†</sup> 京都大学大学院情報学研究科

## 雑談対話システム

S: こんにちは。  
 U: こんにちは。  
 S: 話題を探しています...  
 フェスタに興味はありますか？  
 U: はい  
 S: クラシック音楽が中心のイベント「フェスタサマーミュージック KAWASAKI」が、川崎市で開かれます。  
 U: へえ  
 S: 昨年まで会場だったミュージック川崎シンフォニーホールが東日本大震災で被災しました。  
 U: では今年はどこで開催されるんですか？  
 S: 今年は同市内の別の5会場で分散開催します。  
 U: なるほど  
 S: 新日本フィルハーモニー交響楽団や東京都交響楽団など首都圏で活動する九つのオーケストラが出演しました。  
 U: チケットはいくらですか？  
 S: 気軽に親しんでもらおうと、通常より短めの70分程度のプログラムや、2500円を中心としたチケット価格を設定しました。

U:

[はじめからやり直す](#) [戻る](#)

図1 構築した対話システム

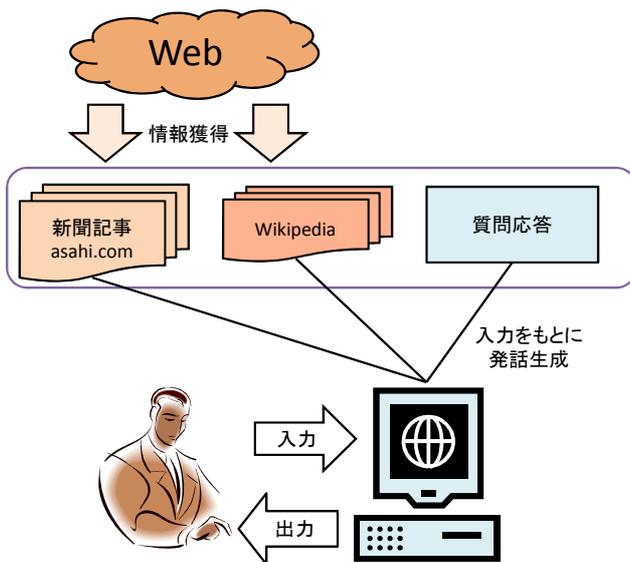


図2 システムの動作概要

らと質問応答処理を用いて雑談対話を行なう。

システムは保持している最新のニュース記事から一つを選択し、ユーザに話題として提案する。続いてユーザの入力をもとに、話題としているニュース記事内の情報、Wikipedia の定義文、関連するニュース記事などを用いて発話を生成する。

ニュース記事は朝日新聞社のニュースサイトである

クラシック音楽が中心のイベント「フェスタサマーミュージック KAWASAKI」が27日～8月14日、川崎市で開かれる。昨年まで会場だったミュージック川崎シンフォニーホール(同市幸区)が東日本大震災で被災。今年は同市内の別の5会場で分散開催する。

新日本フィルハーモニー交響楽団や東京都交響楽団など首都圏で活動する九つのオーケストラが出演。クラシック音楽のほか、映画音楽やジャズなども演奏し、4歳以上の子どもも入場できるファミリーコンサートもある。気軽に親しんでもらおうと、通常より短めの70分程度のプログラムや、2500円を中心としたチケット価格を設定した。

同ホールは震災で天井や照明などが破損し、現在、修理に向けた準備が進む。今年は昭和音楽大学南校舎や洗足学園音楽大学、幸市民会館など市内5会場を使う。

記事本文

キーワード: コンサート フェスタ 東日本大震災

記事に設定されているキーワード

図3 対話に用いる記事の例

asahi.com<sup>\*1</sup>より一定時間ごとに取得し、随時形態素解析、構文解析、固有名詞抽出等の操作を行なう。図1での対話においてシステム発話を生成するために用いたニュース記事の例を図3に示す。

### 3.1 話題の提案

システムは対話のはじめにニュース記事の提案を行なう。提案されるニュース記事は、最新の記事の中からランダムに選択される。更新が頻繁に行われるニュース記事を話題とすることでユーザの興味を引きやすく、対話の自然な導入が可能である。

話題の提案を行なう際にはニュース記事の内容をよく表しているキーワードを用い、「【キーワード】に興味はありますか?」とシステムからの問いかけを行なう。問いかけに対して肯定する回答が得られた場合には、その話題についての対話を開始する。キーワードに関しては出来る限りその記事の内容を表し、かつその記事に特有のものを選択する。記事内に多く、また先頭に近い位置に出現し、その語句が他文書にあまり出現していなければ、その記事に特有な語句であると言える。そこで記事内での出現位置と出現数、取得済のニュース記事全体から算出した IDF 値を用いてキーワードの選択を行なう。ニュース記事のキーワード  $K$  は、記事に付与されている関連記事検索用のキーワードを候補とし、以下の式で算出されたスコアが最も高いものを採用する。

$$Score = \sum_{k \in K} \left(1 - \frac{l(k)}{N}\right) * idf(K)$$

\*1 <http://www.asahi.com/>

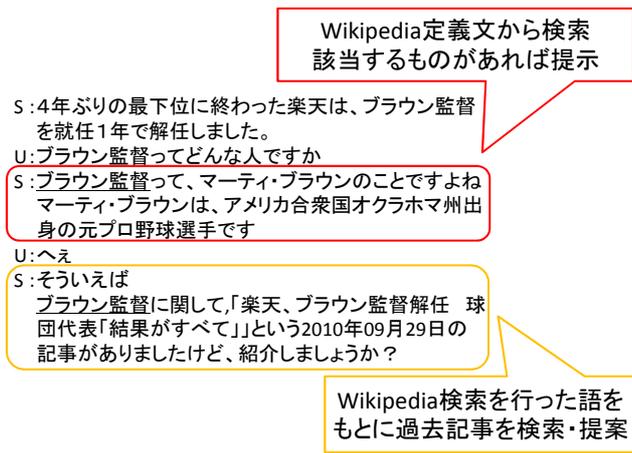


図4 Wikipedia 検索の例

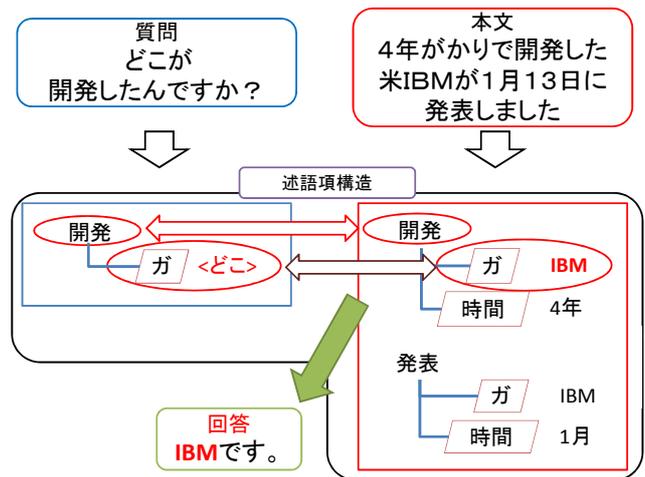


図5 述語項構造マッチングの例

ここで、 $k$  は出現したキーワード候補  $K$  のニュース記事内での各出現、 $l(k)$  は  $k$  が出現した行数であり、 $N$  は記事の文数、 $idf(K)$  は  $K$  の IDF 値である。図3の例ではキーワード候補のうち「フェスタ」が選択され、図1での対話時に話題の提案に用いられている。

### 3.2 話題に関する対話

対話の話題となるニュース記事が決定されると、システムは種々のモジュールを用いてユーザとの対話を行なう。各モジュールはユーザ発話を入力として受け取り、システム発話の候補を生成する。本研究では以下の3種類のモジュールを用いている。

1. Wikipedia 定義文検索
2. ニュース記事を情報源とした質問応答
3. ニュース記事内の情報を提示

システムは Wikipedia 定義文検索、質問応答、記事内情報の提示の順番でシステム発話の生成が可能であることを調べる。

#### 3.2.1 Wikipedia 定義文検索

このモジュールでは例えば「Xってなんですか？」のように、ある語  $X$  の定義を問うような形の文に対して Wikipedia 内の  $X$  の項目を検索し、該当するものがあればその一文目である定義文をシステム発話とする。Wikipedia の定義文はそのままでは話し言葉として適切ではないため、ルールに基づいて「です」「ます」調の話し言葉に変換する。検索を行なう際には Wikipedia に登録されているリダイレクト情報や、曖昧さ回避ページの情報を利用して、表記の揺れや語の曖昧性を吸収する。例を図4に示す。ここではユーザの「ブラウン監督」についての問いに対して、リダイレクト情報を参照し「マーティ・ブラウン」についての定義文をシステム発話としている。

また、Wikipedia に関する発話が終了した際、ユーザの関心が Wikipedia 検索を行なった語「 $X$ 」にあると判断し、「 $X$ 」をもとにして関連記事を検索する。記事の検索には3.1節で採用したキーワードを用い、キーワードとして「 $X$ 」を持つ記事に関連記事とする。該当する記事が存在すれば、その記事についての対話をユーザに提案する。図4の例では、ユーザの問いに回答した後、「ブラウン監督」についての情報に関心があると判断して、「ブラウン監督」をキーワードに持つ記事を新たな話題として提案しなおしている。

#### 3.2.2 ニュース記事を情報源とした質問応答

ユーザ発話が質問文である場合、質問応答を用いて話題としているニュース記事から回答の探索を行ない、システム発話とする。このモジュールではものの名称や数量を問う形の、factoid 型質問文に対して質問応答を行なう。入力されたユーザ発話が質問文であるかどうかは、構文解析器 KNP によって付与される <疑問> タグの有無で判断する。ただし、「そうですか」等の相槌に相当する文に対しても <疑問> タグが付与される場合があるため、一部の文は個別にパターンマッチにより処理から除外する。

回答の探索は図5に示すように述語項構造単位で質問文と記事内の情報をマッチングすることで行なう。質問文と記事内の文はそのままではうまくマッチングがとれないため、質問文中の「誰」や「どこ」といった語句をワイルドカードとして扱い、特定の種類である任意の語とマッチング可能であるとする。マッチング可能である語の種類は、構文解析器 KNP で付与される8種類の固有表現 (ORGANIZATION, PERSON, LOCATION, ARTIFACT, DATE, TIME, MONEY, PERCENT) と数詞である。例えば「どこ」という疑問詞に対しては、固有表現タグ ORGANIZATION あるいは LOCATION を

表 1 対話例の注釈に対応する処理

注釈	対応する処理
話題の提案	話題となるニュース記事の提案 (3.1 節)
Wikipedia 定義文	Wikipedia 定義文検索 (3.2.1 節)
Wikipedia 関連文書	Wikipedia 定義文検索時の関連文書を話題として提案 (3.2.1 節)
質問応答	ユーザの質問に対する質問応答 (3.2.2 節)
類似文選択	ユーザ発話と類似している記事内の未発話文を選択 (3.2.3 節)
未発話文選択	記事内の未発話文のうち先頭に近いものを選択 (3.2.3 節)

持つ任意の語がマッチング可能とする。質問文が入力されると、はじめに小谷らの手法を用いて質問文と記事内の各文を述語項構造に分解し [8]、記事内の各文に対して質問文と構造を比較し、ワイルドカードを含めてマッチングできるかどうかを検査する。マッチングに成功した場合質問文側でのワイルドカードに対応する記事側の語句を回答とする。図 5 の例では「開発」に関する述語項構造において、疑問詞「どこ」に対応している「IBM」が固有表現 ORGANIZATION であるためマッチングが成立し「IBM」が回答として抽出されシステム発話となる。

また述語項構造に関するマッチングによる回答抽出では、本文中に質問に対する回答にふさわしい情報が含まれていても、その述語項構造によっては抽出できない場合がある。そのため述語項構造に関するマッチングに失敗した場合の補助的な質問応答手法として、記事中で前述のワイルドカードに対応する種類の語のうち、その語が出現していた文と質問文間での同一語の出現数と、直前のシステム発話で選択された文からの距離から算出されたスコアが最も高いものを、ユーザの質問に対する回答とする。

### 3.2.3 ニュース記事内の文の提示

このモジュールは話題としているニュース記事内の一文を選択し、3.2.1 節での Wikipedia 定義文と同様に話し言葉変換を行なったうえでシステム発話とする。ニュース記事内にユーザ発話と類似している文が存在し、かつその文が未発話であれば、その文をシステム発話とする。ユーザ発話文と記事内の文が類似しているかどうかは、両文中においてどれだけ同じ名詞が出現しているかで判断する。ユーザ発話と類似していると判断される文が存在しなかった場合には記事の先頭に近い未発話文を選択し、全ての文を発話として提示し終わるとシステムは話題となっている記事の変更を行なう。

## 4 対話例

本システムの対話の例を示す。対話例での各システム発話の右側には、その発話がどのような処理から生成さ

S:富士通東芝モバイルコミュニケーションズに興味はありますか？	話題の提案
U:はい	
S:KDDIは、米マイクロソフトの携帯電話向け基本ソフト「ウィンドウズフォン . 5」を搭載したスマートフォンを9月に発売すると発表しました。	未発話文選択
U:どこが製造するんですか？	
S:モバイルコミュニケーションズです。	質問応答
U:価格はいくらくらいなんですか？	
S:価格は、2年契約などでの割引額を差し引いて3万円台の見込みです。	類似文選択
U:なるほど	
S:MSの最新OSが搭載されたスマートフォンの発表は世界で初めてです。	未発話文選択
U:そうだったんですか？	
S:ウィンドウズ系のOSを搭載したパソコンとの連携のしやすさなどが特徴です。	未発話文選択

図 6 質問に回答することによる対話の例

れたものであるか注釈を付与している。注釈と各処理の対応は表 1 に示す。

ユーザの質問に対して、質問応答の処理を用いて回答することで対話を遂行した例を図 6 に示す。この例ではスマートフォンに関する記事を話題としたうえで、「どこが製造するんですか？」といったユーザの問いに対して質問応答モジュールを用いて回答を行なっている。また、価格についての問いに対しては質問応答モジュールでは回答を発見できなかったものの、記事中の類似した文を発話として選択したことで適切に回答できている。

ユーザの質問に対して Wikipedia 定義文を用いて回答し、関連記事を新たな話題として提案した例を図 7 に示す。対話開始時の話題は天然ガスを燃料とする貨物船についての記事であったが、ユーザが「LNG」、つまり「天然ガス」について聞いたことによりユーザの興味が「天然ガス」に移ったと判断して関連記事を新たな話題として提案しなおしている。

一方で、システムの発話選択が不適切であった例として図 8 を示す。この例において、ユーザは話題としている記事の「コクリコ坂から」の監督が誰であるかと質問を行なったが、システムはそれを「監督」という語句の定義

S: 川崎汽船に興味はありますか?	話題の提案
U: はい	
S: 海運大手の川崎汽船は、液化天然ガスを燃料とする貨物船の開発を始めたと発表しました。	未発話文選択
U: へえ	
S: 重油を燃料とする従来型より二酸化炭素排出量が4割減る見込みです。	未発話文選択
U: それはすごいですね	
S: LNGのみを燃料とする貨物船は世界初といえます。	未発話文選択
U: LNGって何ですか?	
S: LNGって、天然ガスのことですね。天然ガスは、一般に天然に産する化石燃料である炭化水素ガスのことを指します。	Wikipedia 定義文
U: なるほど	
S: そういえば天然ガスに関して、「結晶の中に天然ガス、新鉱物「千葉石」発見 英誌に発表」という2011年02月15日の記事がありましたけど、紹介しましょうか?	Wikipedia 関連文書
U: お願いします	
S: 結晶の中にメタンなどの天然ガスを閉じ込めた新しい鉱物が千葉県南房総市の山中で見つかりました。	未発話文選択

図7 Wikipedia 検索と話題変更が発生した例

S: スタジオジブリに興味はありますか?	話題の提案
U: はい	
S: スタジオジブリのアニメ「コクリコ坂から」が公開中です。	未発話文選択
U: へえ	
S: 東京五輪を翌年に控えた1963年の横浜を舞台に、高校生たちの青春を描きます。	未発話文選択
U: 監督は誰ですか?	
S: 監督とは、多くの事柄や人々・組織など見張ったり、指図をすることで取り締ること、またはそれらを行なう人や組織を言います。	Wikipedia 定義文

図8 システムの発話が不適切であった例

を聞いたものと判断してしまった。これはシステムが発話の生成を Wikipedia 定義文検索モジュールから順に試行しており、さらに Wikipedia 定義文検索でシステム発話が生成されるかどうかを、文脈情報を考慮していない単純なパターンマッチで判断していることから発生した。この問題に対処するためには、それまでの対話履歴を参照し、ユーザ発話で省略されうる要素を各発話生成の際に考慮する必要がある。また、この例で話題としていた記事内には、問いの答えとすることが可能な「ゲド戦記の宮崎吾朗監督が父・宮崎駿の企画を映像化しました」という文が存在しており、質問応答モジュールや記事内の文の提示モジュールが生成したシステム発話では適切に回答できていた。このため、システム発話の生成優先順を場合によって変化させることも有効であると考えられる。

## 5 おわりに

Web から獲得した情報と複数の発話生成モジュールを用いて雑談対話を行なうシステムを提案し、最新のニュース記事を話題として用い、ユーザからの質問に適切に回答することで、ユーザにとって未知の情報を提供しつつ対話を行なうことが可能であることを示した。今後は、より適切なシステム発話を選択するために、各モジュールを制御する枠組みを改善することを考えている。また、システムが行なった対話そのものの評価も行なう予定である。

## 参考文献

- [1] 翠輝久, 河原達也, 正司哲朗, 美濃導彦. 質問応答・情報推薦機能を備えた音声による情報案内システム. 情報処理学会論文誌, Vol. 48, No. 12, pp. 3602-3611, 2007.
- [2] 鹿野清宏, 川波弘道, 西村竜一, 李晃伸. 音声情報案内システム「たけまるくん」および「キタちゃん」の開発. 情報処理学会研究報告, SLP, 音声言語情報処理, Vol. 2006, No. 107, pp. 33-38, 2006.
- [3] 横山祥恵, 山本大介, 古賀敏之, 小林優佳, 土井美和子. 高齢者向け対話インタフェースの開発: 概念辞書を用いた話題展開法. 情報処理学会全国大会講演論文集, Vol. 71, No. 4, 2009.
- [4] J. Weizenbaum. Eliza - a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, Vol. 9, No. 1, pp. 36-45, 1966.
- [5] 水野淳太, 乾健太郎, 松本裕治. ウェブニュースを利用した雑談対話システム. 言語・音声理解と対話処理研究会 55, 1-6, 2009-03-13, 2009.
- [6] Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. Spoken dialogue system based on information extraction using similarity of predicate argument structures. In *Proceedings of the SIG-DIAL 2011 Conference*, pp. 59-66, Portland, Oregon, June 2011.
- [7] 鳥澤健太郎. 一般ユーザーにインタビューする対話エージェント. 信学技報 IEICE Technical Report NLC2007-5(2007-7), 2007.
- [8] 小谷通隆, 柴田知秀, 黒橋禎夫. 言い換え表現の述語項構造への正規化とテキスト含意関係認識での利用. 言語処理学会 第15回年次大会, pp. 260-263, 2009.