HOKKAIDO UNIVERSITY

| | |
|---|---|
| Title | Web検索と単語n-gramモデルによる対話処理手法における補佐システムの有効性 |
| Author(s) | 高橋, 瑞希; Rzepka, Rafal; 荒木, 健治 |
| Citation | ことば工学研究会, 35, 17-23 |
| Issue Date | 2010-08-27 |
| Doc URL | http://hdl.handle.net/2115/63986 |
| Type | article |
| File Information | n-gram.pdf |

Instructions for use

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

# Web検索と単語n-gramモデルによる対話処理手法における補佐システムの有効性

高橋瑞希　Rafal Rzepka　荒木健治
北海道大学大学院　情報科学研究科　メディアネットワーク専攻
{ m_taka , kabura , araki }@media.eng.hokudai.ac.jp

## 1. まえがき

ショッピングセンターに設置された案内システム[1]や，インタラクティブな QA システム[2]など，人間と対話を行うシステムの研究および開発が，近年盛んに行われている．これらはタスク指向型対話システムと呼ばれ，「道案内をすること」「ユーザの疑問を解決すること」などといった，達成すべき目的が定まっている対話システムである．しかしながら，これらのシステムはタスクが決まっているがゆえに話題の変遷には対応することができず，それぞれを併用しても補完することのできない需要の隙間が，必然的に生じてしまう．そのため，話題が変化する会話，いわゆる「雑談」を行う非タスク指向型の対話システムに対して，注目が集まっている[3].

しかしながら，文法的な解析が困難な話し言葉で進行し，かつテーマが次々に変遷する雑談において，ユーザに対して意味的にも内容的にもまったく違和感を与えない文を生成することは難しい．また，柔軟な文生成を実現する方法として n-gram モデルが挙げられるが，単純に適用するだけでは，要素間の接続情報しか考慮されない．そのため，文全体をとらえた場合，意味的・文法的に不自然な文が生成されてしまうことがある．こういった違和感を与える応答は，人間との雑談では起こりえないものであり，システムに対する評価の低下に直結する．

本稿では，単語 n-gram モデルに着目した対話システムにおける改良案の一つとして，ユーザとの会話を補佐するシステムの導入を提案する（以後「補佐システム」と呼ぶ）．また，実際に補佐システムを組み込んだ対話処理システムを用い，その有効性，およびユーザに与える印象や影響について検証を行う．

なお本研究における補佐システムの定義は，“メインの対話システムとユーザとの会話に混じって第三者として発言することにより，会話の進行や内容を補佐する役割を担う対話システム”である．すなわち，メイン対話システムに対するサブ対話システムという位置付けである．前提として，補佐システムはあくまでサブの立場であり，実質的な雑談を行うのはユーザとメイン対話システムとなる．本稿では，**2.** で述べる『Maru（マル）』をメイン対話システムとする．また，補佐システムを『Fill（フィル）』と名付ける．Fill に求められる処理，および詳細については，**3.** で記述する．

## 2. メイン対話システム

Maru とは，知識源として WWW を，文生成手法として単語 n-gram モデルを，応答文の選択方法として **2.3** で述べるスコア算出方式を採用した，非タスク指向型の対話システムである．ユーザの発話文解析や文生成の過程において名詞・形容詞を重要な要素としてみなし，また一度の Web 検索によって取得される情報のみを用いるため，効率よく応答文を生成することを可能としている[4].

基本的な処理の流れを，図 1 に示す．以下，各処理について簡単に述べる．

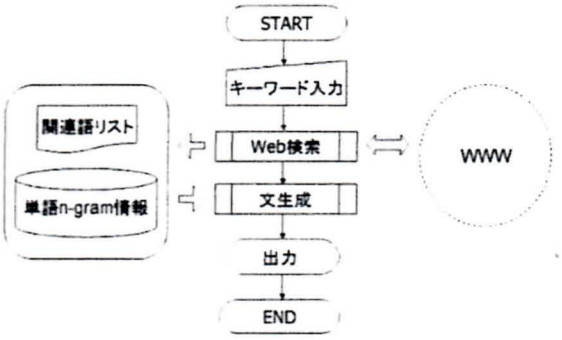


図 1　処理の流れ

### 2.1 Web検索によるWWWの利用

今日 WWW のネットワーク上では，あらゆる話題に関する情報が網羅されている．WWW が対話システムの知識源として有効であることは既に示されており[5]，Web 検索を使用することによって，WWW を一つの巨大な知識データベースとみなすことが可能となる．これにより，知識データベース構築のコストを抑えつつ，常にメンテナンスされたデータを取得することができる．ただし，人手によるデータベースと異なり，Web 検索で得られる情報には必要のないデータが混在しているため，これらを適宜除外する必要がある．

ユーザの入力を受け取った Maru はまず，発話文に含まれるキーワードを抽出し，それらをクエリとして Web 検索を行う．予備実験[4]の結果，有効なデータを効率よく収集することができた，自立名詞および形容詞をキーワードの対象とする．対象形態素解析には MeCab[6] を，検索エンジンには Google[7] を使用する．なお，雑談においては，小説やゲーム，時事問題といった，辞書には含まれない名詞が話題となるケースが数多く想定される．こういった特殊な名詞の解析ミスを事前に防ぐため，MeCab の辞書にはあらかじめ Wikipedia[8] の記事名から収集した名詞を一括して追加登録しておく．

Web 検索によって得られたスニペットには，キーワードに対して関連度の高い語（関連語）が集中している．そこで，スニペットに対して形態素解析を施した後，出現頻度の高い語を関連語として取得し，これらを文生成および応答文選択の際に利用する．

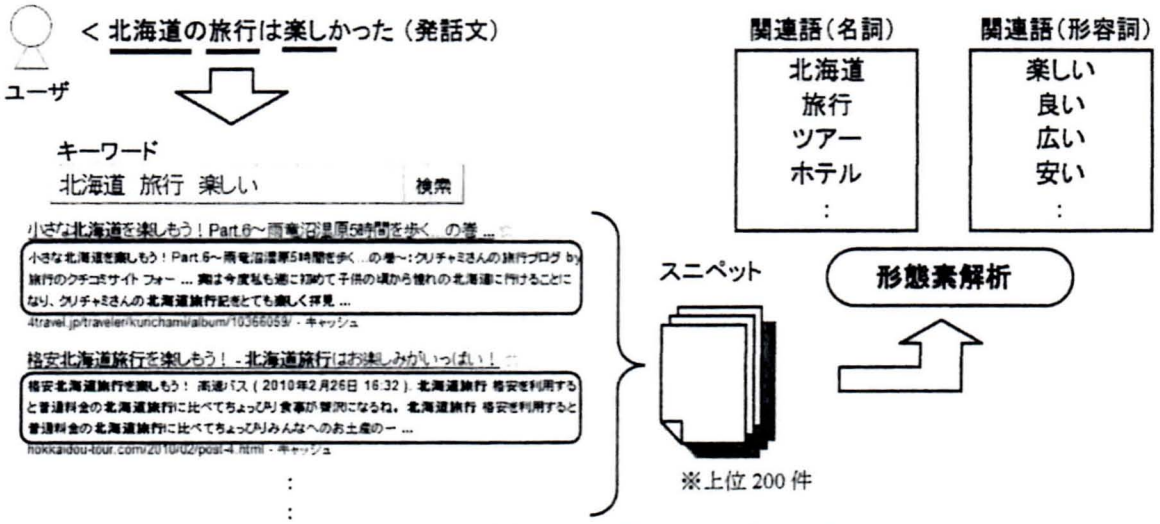


図 2　WWW の利用によって取得されるデータの例

一連の処理によって取得されるキーワード，スニペット，および関連語の例を，図 2 に示す．

### 2.2　候補文生成

Maru では文生成手法として，n-gram モデルを単語単位に拡張した単語 n-gram モデルを採用している．文生成手法には，テンプレートに適宜単語を当てはめる方法，対話ログから文を選択する方法，応答用の辞書を用意する方法など，他にも様々な手法が存在する[9]．しかしながら，これらの手法はいずれも十分な量の対話ログや応答ルールがなければ，画一的な反応に終始することとなる．**2.1** で述べた知識データベースの問題と同様に，対話ログや応答ルールは人手で用意するものなので，応答文のバリエーションを豊かにするためには，必然的に準備段階でコストがかかってしまう．一方，単語 n-gram モデルは単語の接続情報をもとに次々と単語を連鎖させて文を生成するため，型にはまらない柔軟な文生成を行うことができる．また，ランダム性を持つため，同じ応答を繰り返す状態に陥りにくいという利点がある．

ただし，単語 n-gram モデルを単純に適用するだけでは，文脈や文法を考慮することができない．また，ランダム性は上記のような利点を持つが，それと同時に，自然な文が生成されたり不自然な文が生成されたりと，生成結果に不安定さが残る欠点にもなりえる．そのため，何らかの手段を用いて，n-gram モデルの長所を残しつつ，短所を抑制する必要がある．

そこで本システムでは，一度の試行で多くの応答候補文を生成した後，その中から最も自然な応答文を選び出すという処理を行っている．具体的には，以下の手順で候補文群を生成する．

まず，**2.1** で得られたスニペットに単語 n-gram モデル(n=3,4)を適用し，単語 n-gram 情報をまとめたデータを生成する．次に，文の先頭にあたる単語を定め，その単語から単語 n-gram 情報に従って次々に単語を連鎖させる．なお，本システムでは先頭単語として，**2.1** で抽出した関連語，およびユーザの発話から得られたキーワードを用いる．また，単語によっては先頭単語以降の連鎖が途絶えてしまい，文の生成ができなくなる可能性も存在するため，先頭単語は複数個選出する．本システムでは処理時間との兼ね合いから，最大 6 つの単語を選出することとする．

連鎖は条件を変えながら，複数回行う．連鎖方法は大きく分けて，「順生成」と「逆生成」の 2 種類を用いる（図 3）．n-gram モデルの性質上，それぞれ文頭と文末に，違和感の少ない表現が現れやすい．また，両者を複合させた連鎖方法である「双方向生成」も，同時に行う．以上の操作を繰り返すと，多数の出力候補文を得ることができる．

### 2.3　文選択

**2.2** で生成された候補文の中には，ユーザの発話に対して関連性が薄いもの，話し言葉であることを差し引いても文法的に不自然であるもの，冗長であるもの，ノイズが混ざってしまっているものなど，出力すべきではない文が複数含まれている．そこで，各文に対して以下のスコア計算式（式(1)）を適用し，応答文としての確からしさを求める．

$$score(S_i) = sw \times sz\left(sL(S_i) + wC(S_i) + \sum \alpha\right) \quad \cdots(1)$$

$sL(S_i)$：文 $S_i$ の文長スコア
$wC(S_i)$：文 $S_i$ の関連語スコア
$sw$：文 $S_i$ がストップワードを含む時 0，含まない時 1
$sz$：$sL(S_i) = 0$ の時 0，それ以外の時 1
$\alpha$：その他の要素　$-1 < \alpha < 1$

「文長スコア」「関連語スコア」は，「文の長さに関するスコア」および「関連語の有無に関するスコア」で

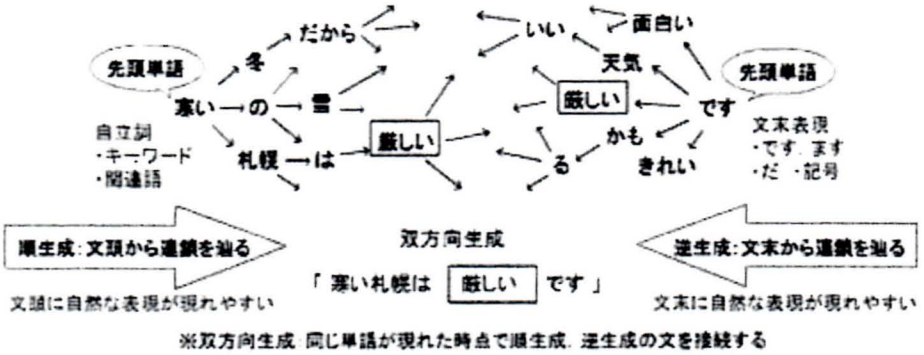


図 3　単語 n-gram モデルを応用した文生成

表 1　スコア計算式の各要素

| | |
|---|---|
| 文長スコア | 極端に長い生成文 → 0 点<br>適切な文長（目安：11～40 文字）→ +11 点<br>その他の文字数は線形補完したスコアを与える<br>（+1～10 点） |
| 関連語スコア | 関連語を含む場合，その数だけ +2 点 |
| 文法 | 話し言葉として見ても不適切な繋がりを含む<br>（例：文頭が助詞）<br>→スコアを 1/2 |
| ストップワード | 不適切な語や，明らかなノイズを含む<br>→無条件に文全体のスコアを 0<br>・"http://～"といった URL の断片<br>・電子掲示板の名前欄に含まれる日付や個人識別用のトリップ<br>・「PDF」「コメント」など Web ページ上にて普遍的に現れる単語<br>・「当店」「情報」など Web 広告や宣伝に現れやすい単語 |
| 特殊なケース | 文長スコアが 0 点<br>→関連語スコアに関わらず全体のスコアを 0 点<br>（極端な長文の中に関連語が繰り返し出現しているケースが多いため） |
| その他 | 助動詞の有無など，1 点刻みでスコアを調整 |

ある．本システムはテキストベースであるため，また簡易的にユーザの発話内容と関連性のある文を求めるため，これら 2 つのスコアの和を基本スコアとした．計算式に含まれる各項の説明を，表 1 に示す．なお，各パラメータの初期値や閾値は，調整を重ねて妥当と判断されたものを採用している[4].

### 2.4　システムの性能

**2.3** までに述べた処理を行う Maru は，単語 n-gram モデルによる文生成をベースとしながら，内部でスコア算出を行うことにより，不適切な文を出力してしまう可能性を低くすることを実現している．同じく WWW を知識源としている対話システム Modalin[10]と印象比較実験を行ったところ，被験者 12 名中 10 名が，Maru により高い評価を与えた[4]．しかしながら，意味的・文法的に不自然な文を完全に除外するまでには至らず，未だ三割以上の確率で，違和感のある文が応答として出力されてしまうという課題が残っている．

この点については，スコアの計算式を変更したり，単語 n-gram 情報の構築方法を工夫したりといった改良案が考えられる．しかしながら，文生成に関する処理の複雑化は応答の遅延に繋がる可能性が高く，システムを利用するユーザのストレスに結びついてしまう．

そこで，本研究では Maru 自身を改良するのではなく，別のアプローチから対話内容を改善させることを考える．以下，このようなアプローチを実現するための「補佐システム」について，詳細を述べていく．

## 3.　補佐システム

### 3.1　補佐システムに求められる要素

メイン対話システム Maru がより良い応答文を生成するためには，ユーザから具体的なキーワードを引き出すことが重要である．なぜなら，このキーワードをもとに Web 検索を行い，情報を収集するためである．例えば，ユーザが「サッカーの○○選手」に興味を抱いていた場合，選手名は当然のこと，「サッカー」という単語や所属チーム名などの付随情報まで含めて揃っていた方が，Web 検索およびその後の処理には都合がよい．

また，Maru の応答がユーザの発話と繋がらない，関連性の低いものであった場合，その突拍子のなさが対話上のエッジとなり，違和感に直接結びついてしまう．しかしながら，両者の間に橋渡しの役目を果たす発話があれば，その発話が違和感を吸収するクッションとなり，会話がその分滑らかに繋がると考えられる．この条件を満たす発話は一種類ではないが，ここでは処理の煩雑化を避けるため，「ユーザの発言を受けて Maru に話題を振り，Maru の発言をうながす」という内容の発話を挟むことにする．上記の仮説が正しければ，例え Maru の応答が違和感あるものだとしても，Fill の発話の有無により，ユーザが受ける印象が変化するはずである．

以上より本研究では，大きく分けて以下の 3 種類の発話文を，補佐システム Fill に出力させる．

・ユーザからより多くの情報を引き出す発話
・ユーザに対して相槌，フィラーを返す発話
・ユーザの発言を受け，Maru の発言をうながす発話

### 3.2 発話文生成処理の実装

補佐システム Fill は，メイン対話システムの Maru と共に，ユーザに対して雑談を提供する，非タスク指向型の役割を担うシステムである．しかしながら，Fill 単体に注目すると，「非タスク指向である Maru を補佐する」という目的を持ったタスク指向の側面も併せ持っている．よって Fill の場合，不確実性のある n-gram モデルより，テンプレートを適用する文生成手法の方が適切であると考えられる．実際，代表的なテンプレート型対話システムである Eliza[11]は，ユーザの発言をうながしたり，相槌を打ったり，ユーザの発話に含まれる名詞や形容詞に反応するなどといった動作を実現している．

本研究では，表 2 に示す発話文テンプレートを Fill に適用する．Eliza にはより多種多様なテンプレートが搭載されているが，Fill は Eliza と異なり単独で対話を行うシステムではないため，**3.1** に挙げた目的を満たすテンプレートを採用するにとどめた．ただし，同じ反応を返すという印象をユーザに与えないため，相槌や定型文は複数用意し，その中からランダムに選択する方式を採用した．

表 2　テンプレートの種類

| |
|---|
| 単語を当てはめるテンプレート<br>・filler1＋（キーワードもしくは関連語）＋ modality<br>・filler2＋（キーワードもしくは関連語）＋ modality<br><br>※filler1：感嘆詞<br>ああ，うん，おお，ええ，うーむ，　…<br>filler2：相手の話に興味を示すフィラー<br>ふむふむ，ほうほう，なるほど，　…<br>modality：名詞もしくは形容詞の後ろに接続する語<br>名詞 ～ ね，かあ，だよね，！（記号），…<br>形容詞 ～ ね，よね，ってことかい，… |
| 定型文テンプレート<br>・Maru の発言をうながす<br>「Maru はどう思う？」「Maru，聞いてる？」…<br>テンプレートの種類が限られているため<br>ランダムで filler と組み合わせる<br>「ふむふむ。」＋「Maru，この話知ってる？」<br><br>・ユーザの発言をうながす<br>「それって何のこと？」「それ，詳しく教えて！」… |
| その他，挨拶への対応<br>こんにちは，こんにちは→「こんにちは！」<br>バイバイ，さよなら→「また来てね」　　など |

### 3.3 発話タイミング

Maru とユーザの対話を補佐する上で，Fill が Maru と共に毎ターン発話を行うことが適切な頻度かどうかを確かめるため，簡易的な実験を行った．被験者は第一著者を含む 3 名で，第一著者以外の 2 名は理系大学院生の 20 代男性，および文系大学生の 20 代女性である．

実験手順
(1)ユーザ，Maru，Fill の三者が発言を行うインターフェースを備えたチャットルーム $\alpha, \beta$ を用意する
(2)部屋 $\alpha$ では，Fill は毎ターン（ユーザと Maru が行う発話の間に）何らかの発話を行う．部屋 $\beta$ では，ユーザの発話にキーワードが 2 つ以上含まれなかった場合のみ発話を行う．
(3)それぞれのルームで雑談を行い，被験者ごとに Fill の発言頻度についての評価を行う．

部屋 $\beta$ での発話タイミングは，Fill が持つ「ユーザからより多くの情報を引き出す」という役割を考慮し，設定した．その結果，部屋 $\alpha$ における Fill の発言頻度は，被験者全員が「過多である」と感じたことが確認された．その後，発話条件を様々に変更して（n 回起きに発話，キーワードが 1 つも得られなかった場合に発話，ランダムに発話，など）同じ実験を繰り返したが，部屋 $\beta$ における設定が最適であるという結果が得られた．よって本研究では，「ユーザの発話にキーワードが 2 つ以上含まれなかった場合」という発話タイミングを，Fill に設定する．

## 4. 評価実験

本章では，補佐システムの有効性，およびユーザに与える影響を検証するための，2 つの実験について述べる．

### 4.1 実験 1：補佐システムの有無による比較

本実験では，Maru とユーザが交互に発話して対話を行うチャットルームをベースラインとする．このベースラインに Fill を加えたチャットルームを用意し，2 つの部屋での雑談を比較することで，補佐システムが存在することによる影響を調査する．なお，2 つの部屋で対話を行った被験者は 10 代から 30 代の男女 10 名で，この中に自然言語処理に関する知識を持つ者は含まれない．また，理系が男性 2 名(大学院生 1, 社会人 1)・女性 2 名(大学生)，文系が男性 1 名(社会人)・女性 5 名(大学生 4, 社会人 1)である．

便宜上，ベースラインのチャットルームを部屋 A, 補佐システムを加えたチャットルームを部屋 B とする．部屋 A・部屋 B ともに，発話の内容や話題には特別な制限を設けなかった．ただし，比較を行うという実験の性質上，それぞれの部屋の最低対話ターン数は 20 ターンとした．

対話後，被験者にアンケートを実施した．アンケートの内容を表 3 に示す．なお，部屋 B で行われる雑談は実質的に三者の会話であるため，文中では対話ではなく，

表3　アンケート内容

アンケート

【1】

それぞれの部屋で行われた会話はいかがでしたでしょうか。次に挙げる各項目について、あなたがより当てはまると　　感じた方の部屋を挙げて下さい。

(1)会話の流れが自然であった
(2)会話の内容が意味的に自然であった
(3)会話の整合性が取れていた
(4)会話のテンポがスムーズであった
(5)会話しやすかった
(6)「会話を続けたい」と感じた

【2】

上記に示した会話のターンの中で、システムの応答(反応)が良かったと感じたものをチェックしてください。(複数可)

【3】

総合的に判断して、より良いと感じた部屋はどちらの部屋ですか？　また、その部屋を選んだ決め手となる理由を教えてください。

【4】

そのほか感じたこと、意見・感想等を自由に記述してください。

※【1】(1)～(4)はシステムと対話する上でユーザが違和感を覚えやすい点に関する評価するため，(5)(6)は対話全体の印象を評価するための項目である．また，【2】は実際の対話ログを表示させ，それぞれのターンを 1 つずつ確認するスペースを設けた．

表4　実験1結果

| 【1】(1)～(6)　部屋AまたはBから選択 | | | | | |
|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) |
| A：5名<br>B：5名 | A：6名<br>B：4名 | A：2名<br>B：8名 | A：7名<br>B：3名 | A：3名<br>B：7名 | A：2名<br>B：8名 |

| 【3】総合的な評価　部屋AまたはBから選択 | |
|---|---|
| A：3名 | B：7名 |
| 主な意見<br>・応答する内容がより良かった<br>・1人を相手に話す方が単純に気楽だった<br>・会話のテンポが良かった | 主な意見<br>・にぎやかだった<br>・雑談をしているという雰囲気が強かった<br>・こちらの話をきちんと聞いていた<br>・反応がより良かった |

表5　実験2結果

| 【1】(1)～(6)　「A」は本文中における部屋A'のこと | | | | | |
|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) |
| A：5名<br>B：4名 | A：4名<br>B：5名 | A：3名<br>B：6名 | A：6名<br>B：3名 | A：2名<br>B：7名 | A：3名<br>B：6名 |

| 【3】総合的な評価 | |
|---|---|
| A：4名 | B：5名 |
| 主な意見<br>・会話のテンポが良かった<br>・不自然な応答がなかった<br>・ちぐはぐな応答が少ないように感じた | 主な意見<br>・人数が多い方が面白い<br>・こちらの話をきちんと聞いていた<br>・違和感が少なかった<br>・すぐに答えてくれたように感じた |

※表4, 表5　部屋A：ベースライン　部屋B：提案システム

「会話」という表現を用いた．アンケートのうち，自由記述の項目を除いた結果をまとめたものを，表 4 に示す．

### 4.2　実験 2：発話量を揃えた場合における比較

実験 1 は，ベースラインとの相違点が，補佐システムが存在するか存在しないかという点だけであった．つまり，他手法との比較実験ではなく，部屋 B は部屋 A の純粋な拡張形といえる．そのため，被験者に対して「部屋 B が本研究の対象である」と予見されなかったとは，残念ながら断言することはできない．そこで，「補佐システムという第三者が加わることによる影響」をよりはっきりと示すために，もう 1 つの実験を行った．

本実験では，2 つのチャットルームにおける発話量・対話量を同じ程度に揃えた上で，補佐システムが与える影響について調査する．具体的には，実験 1 で用意した部屋 A の Maru に Fill の文生成システムを追加し，Maru 自身が持つ文生成システムと Fill の文生成システムの両方を Maru の口から発話させる，という構成に変更する．このチャットルームを「部屋 A'」とし，Fill の処理を追加した Maru の表示名を，部屋 B の Maru と区別するため「Ural」に変更する．ただし，Fill が生成する文のうち，Maru に対して問いかける発話文は，不自然な発話となってしまうため除外する．これにより，部屋 A'と部屋 B の差異は，Fill の存在およびルーム内の発話量から，およそ Fill の存在のみとなる．

その他，アンケートの内容は実験 1 と同様，表 3 に示したものを使用する．被験者は 20 代から 30 代の男女 9 名であり，内訳は理系が男性 4 名(社会人 3, 大学院生 1)，文系が男性 1 名(大学生)・女性 4 名(社会人 2,大学院生 2)である．なお，本稿では便宜上，Ural が存在するチャットルームを部屋 A'と表記しているが，アンケート上で部屋 A'を指す場合は「A'」ではなくそのまま「A」という表記を用いている．これは実験 1 と重複している被験者が存在せず，実験 1 における部屋 A と実験 2 における部

屋 A'を，表記上区別する必要がないためである．また，実験 1 同様，被験者の中には自然言語処理に関する知識を持つ者を含まない．

アンケートのうち，自由記述の項目を除いた結果をまとめたものを，表 5 に示す．

## 5. 考察

本章では，**4.** で得られた実験結果とその検証に加え，対話ログから確認された傾向や，自由記述によるアンケート回答を分析し，考察を行う．

### 5.1 実験 1 の考察

項目【1】における 10 名の結果を総合すると．Fill を加えた部屋 B は，(3)および(5)(6)の評価が高い．つまり，対話における整合性，および対話全体の評価が，ベースラインである部屋 A と比較して高まったことになる．これは，**3.1** で述べた「メイン対話システムの違和感を吸収し，ユーザとの発話を繋げる」という役割を，Fill が補佐システムとして果たしているためと考えられる．一方，(4)の評価，すなわち対話のテンポに関する評価は，部屋 A の方が高いものとなっている．これは単純に，部屋 B では全体の発話量が 1 人分増えたため，対話が進行するスピードが相対的に遅く感じられたためだと考えられる．実際，【3】【4】における自由記述では，「一対一で進行する対話の方がスムーズに感じられる」という意見がいくつか見受けられた．しかしながら，雑談の流れが自然かどうかを問う【1】(1)では，2 つの部屋の差があまり見られず，会話のしやすさを問う(5)では部屋 B の方が高い評価であったため，補佐システムが会話の流れを妨げているとは，必ずしも限らないことがうかがえる．

表 6　評価の高かった対話・会話例

| 部屋 A |
|---|
| ユーザ：最後の夏休みなので海に行こうと思ってます<br>Maru：沖縄までなら 35,000 円ですよ<br><br>ユーザ：小説家では京極夏彦が好きなんだ<br>Maru：斬新な小説『姑獲鳥の夏（うぶめのなつ）』で鮮烈なデビューを果たしたあの<br><br>ユーザ：せっかくホワイト予約したのに当日取りに行けなくてさ<br>Maru：ブラックとホワイトでデザインが違っていますね |
| **部屋 B** |
| ユーザ：夕べは雷雨のせいで眠れなかった…<br>Fill：うーむ…Maru はどう？<br>Maru：心細かった。<br><br>ユーザ：それいいねー。どう？給料弾むよ？<br>Fill：詳しく教えてください。<br>Maru：報酬です<br><br>ユーザ：今大きな悩みがあるんです。聞いてくれますか<br>Fill：なるほど、悩みね。<br>Maru：小さな雲のようそれは風のようなものです。 |

また，【3】の選択式設問より，総合的な判定においても，部屋 B の方が 優れているという結果が得られた．理由としては，「Maru との間を（Fill が）取り持ってくれているように感じた」，「同じテーマを話したが，部屋 B の方が反応が良かったと思う」，「にぎやかだった」など，Fill が加えられたことに対して好意的な印象を持つ回答がほとんどであった．

アンケートの項目【2】では，被験者に対して対話ログを見せ，システムの応答が良かったと感じたターンを選択してもらう，という形式で回答を収集した．なお，システムの応答精度に関する定義は，被験者側の判断に任せた．また，部屋 B ではユーザ，Maru，Fill が発話を行うため，ユーザの発話ごとに区切った発話区間を 1 ターンとした．回答の一部を表 6 に示す．

部屋 A に関しては，意味的・文法的に自然な発話対が選択されており，特に意味的な自然さが重視されている．これは Maru 単体に対して行われた実験[4]と同じ傾向であり，話し言葉が多用される人間同士の雑談を考えても，不自然ではない結果である．一方，部屋 B の場合，意味的・文法的に自然な発話対の他に，不適切と考えられる Maru の発話とユーザの発話との間に Fill が入り込んでいるターンが選択されているケースが散見された．もしこれら対話が部屋 A で行われていた場合，すなわち間に Fill の発話が存在しなかった場合，異なった結果が得られていたと考えられる．【1】【3】の項目同様，この【2】に対する回答からも，補佐システムの有効性を確認することができる．

### 5.2 実験 2 の考察

部屋 A をベースラインとした実験 1 と比較し，実験 2 は，評価がある程度まとまった結果となった．しかしながら，全体を平均すると実験 1 と同様，Fill を用いた部屋 B に関しては【1】(3)(5)(6)の評価が高く，(4)の評価が低い傾向を示している．

その一方で，【3】における選択式設問では，部屋 A' と部屋 B の評価が拮抗しており，実験 1 ほどの差は生じていない．この点について，実際の対話ログを確認したところ，部屋 A'における Ural の発話文，すなわち Maru と Fill を組み合わせた発話文の内容が，不自然な箇所なく噛み合っている，というケースが複数存在した．そして，そういった文を含む対話を行った被験者が，【3】の設問にて部屋 A'を選択していた．つまり，部屋 B からマイナスの印象を受けたために部屋 A'を選択したというよりも，部屋 A'におけるシステムの応答精度が良好だったため，部屋 A'を選択したといえる．この場合，部屋 A' における Ural の発話を Maru と Fill の発話に分解し，部屋 B と同じ形式に変換したとしても，文脈や違和感の有無に変化はないと考えられる．図 7 にこの例を示す．むしろ，【3】の結果に大きな差が見られないにも関わらず

ユーザ：え、富山？あそこ海水浴出来たっけ？
Ural：うん、楽しいよ！広々してたけど錯覚じゃない。はず。

⇩

ユーザ：え、富山？あそこ海水浴出来たっけ？
Fill：うん、楽しいよ！
Maru：広々してたけど錯覚じゃない。はず。

分解しても
会話の流れや整合性に変化がない

図 7　Ural の発話を分解した例

【1】の結果に実験 1 と類似した傾向が見られる点に注目すると，Fill がターン間の違和感を抑えつつ，整合性やターンごとにおける応答文の質を向上させる役目を果たしていると考えることができる．よって本実験からも，補佐システムという第三者を組み込むことに有意性があると考えることができる．

対話ログを直接評価する【2】の設問では，部屋 B に関しては実験 1 と概ね同様の結果が得られた．一方，部屋 A'においては，前述の「不自然な箇所なく噛み合っている発話文」が，高い評価を得ていた．しかしながら，こういった発話文は，部屋 A'でシステムにより生成された応答文全体のおよそ 15%に留まっており，比較的低い確率で偶然生成された結果に過ぎない．また，単語 n-gram モデルとテンプレート型の文生成手法をそのまま接続した結果，被験者の中には「わざとらしさ」，「回りくどさ」を感じた者が存在した．この点については，接続詞を挟むなど，何らかの処理を施した上で文を繋げることにより改善される可能性があるが，本実験の場合は 2 つの文生成手法が共に詳細な文脈情報や意味情報を含まないため，そもそも改善させることが困難であると考えられる．こういった問題点についても，補佐システムという第三者の口から発話を行うという構成が，1 つの解決策であるといえる．

### 5.3　その他の考察

補佐システムが果たすべきものとしてもう 1 つ，ユーザからより多くの情報を引き出させるという役割を，**3.1** にて挙げた．実験 1 の部屋 A・部屋 B の対話ログについて，ユーザの発話にキーワードが 1 つ以下しか含まれなかったケースについて調査すると，部屋 A では 43 ターン，部屋 B では 45 ターンが該当した．さらにその中で，次のターンのユーザの発話に関連語を含むケース（例：「フィギュア」というキーワードしか含まれていない発話があった次のターンの発話に，「スケート」というキーワードが含まれていた）について調査すると，部屋 A では 30 ターン，部屋 B では 42 ターンが該当した．

つまり，部屋 A では 15 回ほど，2 ターン連続で十分な数のキーワードが得られない事態が発生していた一方で，部屋 B ではそのようなケースはわずか 3 回に留まっていたことが明らかとなった．このことから，Fill が補佐システムとして，キーワードを効果的にユーザから引き出していることを確認することができた．

## 5.　まとめ

本研究では，Web 検索と単語 n-gram モデルを使用した対話システムのための補佐システムを考案した．また，条件の異なるチャットルームを用意し，補佐システムの有効性，および会話に与える影響について検証した．その結果，単語 n-gram モデルが持つ文生成の不確実性に対して，違和感を抑制し，あるいは整合性を保つことにより，対話の印象や評価をより良くする効果が確認された．

今後は補佐システムに関して，より適切な発話タイミングや，より効果の高い発話テンプレートのパターンについて，検証を重ねる．特に，評価の低かった雑談のテンポについて，ユーザに与える印象を改善させたい．また，限定されたチャットルームではなく，多くのユーザが存在する Web サービス（Twitter[12]など）を利用して，システムに関する対話データを広く収集し，解析を行う予定である．

## 参考文献


[1] 宮下善太,神田崇行,塩見昌裕,石黒浩,萩田紀博: “来客と顔見知りになる案内ロボット”, インタラクション 2008.

[2] 佐々木智,藤井敦:”取るべき行動を理由と共に答える質問応答システム how 型と why 型の統合”,言語処理学会第 15 回年次大会発表論文集, 2009.

[3] 難波 創, 亀田 弘之, ” 人工無脳における人工意識,” 電子情報通信学会技術研究報告, Vol.107, 2007.

[4] 高橋瑞希,ジェプカ・ラファウ,荒木健治 :“Web 検索と単語 n-gram モデルを用いた文生成手法の性能評価”, 言語処理学会第 15 回年次大会発表論文集, 2010.

[5] 相良春樹,森辰則,中川裕志:”質問応答に対する知識源としての Web 検索エンジンの Snippet の有効性,” 言語処理学会第 12 回年次大会発表論文集, 2006.

[6] MeCab: Yet Another Part-of-Speech and Morpho-logical Analyzer　http://mecab.sourceforge.net/

[7] Google　http://www.google.co.jp/

[8] Wikipwdia　http://ja.wikipedia.org/

[9] 冨坂 亮太, 鈴木 崇史 : “知っておきたいキーワード(第 48 回)人工無脳(会話ボット),” 映像情報メディア学会誌, 64(1) , 2010.

[10] Rafal Rzepka, Shinsuke Higuchi, Michal Ptaszynski, Pawel Dybala and Kenji Araki:”When Your Users Are Not Serious,”人工知能学会論文誌, Vol. 25, 2010.

[11] J.Weizenbaum, “ELIZA—a computer program for the study of natural language communication between man and machine,” Communications of the ACM, 1966

[12] Twitter　http://twitter.com/