雑談システムにおける話題転換

堀内　理沙　　上田　祐大　　原田　孝太　　韓　東力
日本大学文理学部　情報システム解析学科

1. はじめに

近年、対話システムへの期待が高まってきている。その中で、タスク指向対話システムに比べると、雑談システムのような非タスク指向対話システムの研究は会話をいかに飽きることなく続けられるかに重点を置いている。

非タスク指向の既存研究として、樋口らは、人間の発話というものは、命題とモダリティの要素によって構成されるという考えに基づき、「あぁ」や「まぁ」などという多種多様なモダリティを付与することにより、システムの発話がより人間らしくなるような研究を行っている[1]。徳久らは、ユーザの発話の感情がネガティブなもの全てに対して、システムは「それは嫌な気持ちですね」と応答するだけでなく、「それは残念でしたね」、「それは寂しいですね」や「それは心配ですね」というように、ユーザの発話に含まれる感情を正しく判定することにより、システムからの応答文の表現形式を豊かにする試みをしている[2]。斉藤らは、Web 上に存在するニュースや天候などの実世界情報をキャラクタの発話に利用し、事前に整理することで、発話内容を豊富にしようとしている。また、実世界の情報を反映させることで、ユーザには、キャラクタがあたかも現実世界の出来事を把握しているかのような印象を与えている [3]。藤本らは、会話の流れに無関係な話題に遷移することは少ないという考えから、話題・発話間のつながりの自然さに着目し、概念的関連性に基づく話題転換の特徴分析を行っている[4]。

上記に述べた既存研究の[1]と[2]は、システムの応答の仕方に重点を置いており、話題の単調性に触れていない。既存研究の[3]と[4]では、話題提供をしているが、最新の情報をもとに提供していないという問題点がある。そして、上記に述べたすべての既存研究において、時間推移によるユーザ興味度の変化を反映させていないという問題が共通している。

本研究では、以上の問題点をふまえ、

- リアルタイムでの話題生成を行う
- 時間推移を考慮したユーザの興味に基づく話題生成を行う
- 話題転換のタイミング確定手法を考案する

の 3 つを目的とする。また、ユーザに対するシステムからの応答文は、簡単な定型文とテンプレートに Web 検索で得られた話題語を付け加えるだけで生成している。

2. システムの流れ

本研究で提案するシステムの流れを、図 1 に示す。

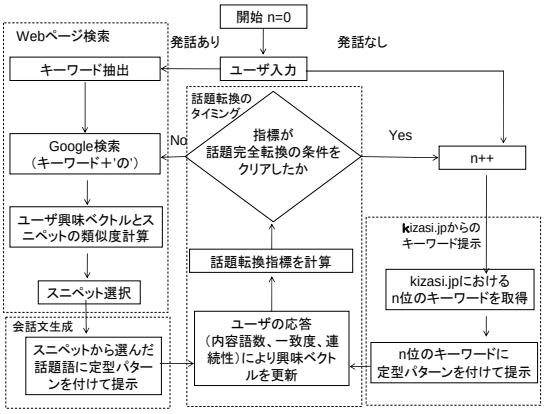
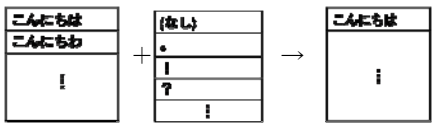


図 1　システムのデータフロー図

図 1 の n は、話題完全転換できる回数を示している。本研究では、n は 10 までとしている。詳しい説明は、次章以降で述べる。

3. 会話文の生成

まず、ユーザが挨拶文を入力してきた場合、システムは挨拶の定型文を返す。ユーザの挨拶文とは、「こんにちは」等のことである。日常生活において、「こんにちは」に対し「こんにちは」と返すのが一般的である。その点を考慮した詳細を、図 2 に示す。



挨拶文のファイル　感嘆符等のファイル　定型文のファイル

図 2　挨拶などの定型文の例

ユーザが、「こんにちは」を「こんにちわ」、「こんにちは。」、「こんにちは！」等の句点や感嘆符を付けて入力した場合も、システムの定型文として「こんにちは」で統一した。

システムがユーザに対し発話する際、テンプレートを用いる。このテンプレートはあらかじめ用意したものであり、相槌、接続詞、語尾の 3 種類である。この 3 種類のテンプレートを組み合わせて利用する。テンプレートの例を、表 1 に示す。

表 1　テンプレートの例

相槌	接続詞	語尾
ええ	ところで	って知ってますか？
そうですね	ではここで	は興味ある？
なるほど～	さて	って説明できますか？
へぇ～	では	について話しましょう

「相槌＋接続詞＋語尾」という生成パターンに基づいて、表 1 に示したテンプレートをランダムに組み合わせて用いる。

4. Web ページ検索

リアルタイムに話題提供を行うために、Google を用いてキーワードを選択している。具体的には、以下の手順に従う。

ステップ 1　ユーザの興味ベクトルを作成
ステップ 2　ユーザ発話を形態素解析器 JUMAN[1]にかけて、名詞を探す
ステップ 3　ステップ 2 で取り出した名詞に「の」をつけて検索キーワードを作成し、検索にかける
ステップ 4　ステップ 3 の検索で出てきた上位 10 件のスニペットを取得する
ステップ 5　10 個のスニペットすべてに対し、スニペットのベクトルを作成
ステップ 6　それぞれのスニペットのベクトルとユーザの興味ベクトルの類似度を計算
ステップ 7　類似度の最も高いスニペットを 1 位として 10 位まで順位付ける
ステップ 8　スニペットの中にクエリ（検索キーワード）があるかどうかを上位から順に調べる
ステップ 9　クエリがあった場合はステップ 10 へ。ない場合は次の順位のスニペットに対してステップ 8 を行う
ステップ 10　クエリがあったスニペットを構文解析器 KNP[2]にかけてクエリの係り受け先の名詞を再帰的に探索する
ステップ 11　クエリと係り受け先の名詞を合わせて、さらにテンプレートをそれに加えてユーザに発話文として返す
ステップ 12　ユーザの発話がなくなるまで、ステップ 2 からステップ 11 を繰り返す

ここでは、名詞が一番ユーザの特徴を捉えることができると考え、名詞を重要視する。また、名詞に「の」を付けて検索するのは、「の」を利用することにより、後ろの文章に関係する名詞が出てくる可能性が高いと考えたからである。

ユーザの興味ベクトルとは、JUMAN の辞書にあるすべての名詞を要素としたベクトルである。ユーザの発話にある名詞が含まれていたらその名詞が示す要素を増やすことでユーザの興味ベクトルを更新する。具体的には、興味認識項と時間減衰項を導入しベクトルの更新を行う。

興味認識項とは、ユーザが多く発話するほど、その発話中にあるキーワードに興味があると考える。よって、ユーザの発話が長い文であればあるほど重みを付ける。また、ユーザの発話に含まれなかったキーワードは興味が小さいと考える。よって重みを減らす。

時間減衰項とは、最後にあったキーワードほどユーザの興味があると考え、過去に出たキーワードすべてに対して重みを減らす。

興味認識項と時間減衰項を使用することでより詳しくユーザの興味を表現する。これは佐竹ら著の[5]を参考にした考えである。本研究で利用するユーザの興味ベクトルの計算式を式 1 に示す。

$$V_i' = \begin{cases} V_i * 0.9 & W_i = 0 \quad and \quad V_i > 0 \ ① \\ V_i - 0.001 & W_i = 0 \quad and \quad V_i \le 0 \ ② \\ V_i + 0.01 * C & W_i > 0 \quad and \quad C \le 100 ③ \\ V_i + 1 & W_i > 0 \quad and \quad C > 100 ④ \end{cases}$$

……式 1

V_i は現在のベクトルの i 番目の要素値で、V'_i は新しいベクトルの i 番目の要素値で、W_i は現在のベクトルの i 番目の単語である。$W_i = 0$ は、現在の発話内に単語 W がなかった時を表し、$W_i > 0$ は、現在の発話内に単語 W が 1 回以上出てきた時を表している。C は発話文の文字数である。また、式 1 中の数値は経験上から導き出した数値である。

式 1 は、条件により①、②、③、④式のどれを使うかが決まる。

①式が時間減衰項の考え方であり、②、③、④式が興味認識項の考え方である。①式も②式もベクトルの要素値を減らす式だが、①式はユーザの発話に少なくとも 1 回は含まれた場合に使用し、②式はユーザの発話に 1 回も含まれなかった場合に使用する。また、②式はベクトルの要素値が-0.5 以下になった場合その処理をしないようにしている。これは、ベクトルの要素値を下げすぎるとユーザの発話に含まれても要素値が 0 より上にならず、結果ユーザの発話に含まれたのに、また②式を使ってしまうので、それを避けるためである。③式は文字数が多ければ多いほど要素値が増える式である。しかし、100 文字以上発話があった時には、④式を使う。これは、いくら興味があるといっても、100 文字は超えることは少ないと考えられることと、もし 100 文字を超えた場合、意味のない記号等が含まれている可能性が高いためである。

ステップ 6 で使用している類似度計算式を式 2 に示す。

[1] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

$$X = \frac{\vec{V}_u \cdot \vec{V}_s}{\left|\vec{V}_u\right| \cdot \left|\vec{V}_s\right|} \cdots\cdots\cdots\text{式 2}$$

X は類似度で、V_u はユーザの興味ベクトルで、V_s はスニペットの全単語ベクトルである。式 2 を用いて、スニペットを類似度の高い順に並べる。

5. 話題転換

本研究では、話題転換のタイミングを決定するうえで、内容語数、一致度、連続性等の指標から閾値を導き、その値に基づき話題転換をするかどうかを判断した。

5.1. 話題転換判断指標

ここからは、話題転換のタイミングを決定するための指標を、詳しく述べていく。

● 内容語数

本研究の内容語数とは、ユーザの発話に含まれる、名詞、動詞、形容詞、形容動詞、連体詞、代名詞、副詞の総数のことである。以下に、内容語数についての例文を示す。

例 1
（システム）北京オリンピックは知っていますか？
（ユーザ）　ええ、知っています。
例 2
（システム）北京オリンピックは知っていますか？
（ユーザ）　ええ、知っています。今回の北京オリンピックは本当に面白かったです。特に陸上の男子 100m は熱狂的でした。

例 1 は、ユーザの発話は単調であるのに対し、例 2 は、「北京オリンピック」というシステムからの話題提供に比較的多く発言しており、発話文中に含まれる単語数も多い。つまり、例 1 において、ユーザは「北京オリンピック」の話題に対しあまり興味が無く、ここで次に話題の転換が起こると考えられる。例 2 においてユーザは、その話題に興味を抱いているので「北京オリンピック」という話題が継続することになる。

● 一致度

この一致度の定義は、システムの発話とユーザの発話がどれだけ一致しているかという事である。一致度の計算式を式 3 に示す。

$$\text{一致度} = \frac{\text{システムからの発話の内容語数と一致した内容語数} * 2}{\text{システムからの発話の内容語数} + \text{ユーザからの発話の内容語数}} \cdots\cdots\text{式 3}$$

以下に、一致度についての例文を示す。

例 3
（システム）WBC って知っていますか？
（ユーザ）　ええ、WBC は知っています。
例 4
（システム）WBC って知っていますか？
（ユーザ）ええ。イチロー選手も参加しますね。

例 3 と例 4 を見ると、システムの質問（発話）に対してユーザの応答（発話）はいずれも長くないが、例 3 では、ユーザは比較的一致した応答（発話）をしている。一方、例 4 においてはユーザの発話文に含まれる内容語はシステムからの発話とまったく異なり、単調な発話ではなく、WBC といった話題に関して比較的深い興味を示していると考えられる。

● 連続性

連続性とはある話題に関してのユーザの反応が、どのぐらい同じ傾向が続くかという事である。以下に連続性についての例文に示す。

例 5
（システム）WBC は興味ありますか？
（ユーザ）　ええ、あります。
（システム）WBC のイチローのこと知っていますか？
（ユーザ）　参加するみたいですね。
（システム）WBC の原監督に興味ありますか？
（ユーザ）　ありません。

初めにシステムがユーザに WBC に興味があるか？と話題を提供している。その返答としてユーザは興味があると答えたため、更にシステムは WBC に関しての話題を深く掘り下げ提供していく。

しかし、システムの更なる発話に対して、「そうみたいですね」、「そのようですね」と単調的で同じ傾向の返答をしている。すなわち、この話題に関して総合的にみればユーザはこの話題に関して、あまり興味が無いのではと考え、このような傾向が続く場合、次の話題に転換することにする。

ここからは、内容語数、一致度と連続性の 3 つの指標がそれぞれどのぐらいの値に達すれば転換が起こるのかということを考える。閾値を決めるための準備実験の結果の一部例を下記の表 2 に示す。

表 2　準備実験の結果(一部)

	組合 1	組合 2	組合 3	組合 4	組合 5	組合 6	組合 7	組合 8	組合 9
内容語数	2 以下	2 以下	2 以下	3	3	3	4 以上	4 以上	4 以上
一致度	0.4 未満	0.4 以上 0.6 未満	0.6 以上	0.4 未満	0.4 以上 0.6 未満	0.6 以上	0.4 未満	0.4 以上 0.6 未満	0.6 以上
連続回数	2	2	2	2	2	2	2	2	2
1 回目	×	×	×	×	×	×	△	×	△
2 回目	△	×	×	×	○	×	×	×	×
⋮					⋮				
10 回目	×	×	×	△	○	×	△	×	×
平均度合	0.36	0.44	0.08	0.04	0.52	0.08	0.42	0.16	0.16

3 つの指標をいろんな数値で 9 つの組み合わせを作成し、1 つの組み合せに対して、話題転換が 10 回起きるよう実験した。ユーザが話題転換をしてほしい時にシステムが話題転換をすれば○。話題転換をしてもしなくてもおかしくない場面であれば△。話題転換をしてはいけない時に話題転換をしたら×をつけた。そして、○が 1、△が 0.8、×が 0 として、話題転換が成立する度合いの平均を出した。

表 2 に示した結果から、我々はユーザが 2 つのパターンに相当する発話をした場合、表 3 に示した値に達した時、話題転換が起こると判断した。

表 3　基準となる閾値

	パターン 1	パターン 2
一致度	0.6 未満	0.4 未満
内容語数	3 以下	4 以上
連続回数	2	2

また、話題が転換する時、すなわち、システムが次の話題提供をする際に必要となるキーワード（話題語）をユーザに提供する必要がある。この話題語を提供する手法は次節で述べる、kizasi.jp[3]というサイトを用いることによって実現する。

5.2. kizasi.jp からのキーワード提示

前述の通り、各指標の閾値に基づき、次の話題に変更すべきと判断した場合、新しい話題をシステムが kizasi.jp から取得し、ユーザに提供する。

この Web サイトを利用するにいたった経緯は、ユーザの興味は日々変化していき、特にその時点で世間が関心にもっている話題に関して興味を示す傾向があるのではないかと判断したからである。また、10 分ごとに話題のランキングが変化していくので、インターネット上で人々が特に関心を抱いていることを話題として提供することが出来る。

また、この kizasi.jp は話題ランキングを順位付けしているだけでなく、ニュアンス（感情）も取り入れている。ニュアンス（感情）とは、その時々でキーワードが悲しいことなのか、楽しい事なのかと表している。ニュアンスは、キーワードの世間の感情を表している。

そのため、我々のシステムでは、ニュアンスが付いているキーワードと、そうでないキーワードを場合分けして、kizasi.jp の注目の話題ランキングのトップ 10 位を、1 位から順に話題語として用いユーザに提供することにした。以下にパターンわけの例を示す。

パターン 1　　キーワードにニュアンスが含まれていた場合

キーワード＋「って」＋ニュアンス＋「って思われているらしいけど、どう思いますか？」

パターン 2　　キーワードにニュアンスが含まれていなかった場合

キーワード＋語尾（ランダムで選択）

以上の 2 パターンを用いて、ユーザに対して話題を提供する。

6.　おわりに

本研究では、Web 検索を利用することにより、リアルタイムに話題を生成することを試みた。また、話題を選択する際にユーザの興味ベクトルを利用することで、ユーザの興味を考慮した話題の提供を実現した。話題転換については、まず、内容語数、一致度と連続性の 3 つの指標を考案し、次に 3 つの指標を組み合わせて準備実験を繰り返すことで、話題転換のタイミングを決定する条件を定めた。最後に、話題転換のために kizasi.jp からキーワードを取得し、ユーザに話題として提供する仕組みも考案した。

しかし、スニペットからキーワードを抽出しているので、バリエーションが豊富でないことや、テンプレートをランダムに利用しているため、システムの発話文がおかしくなるという問題も残っている。また、時間の関係で、この研究に係わっていない人達によるアンケート実験を行っていないため、第三者からみた問題点があると思われる。

今後は、システムとユーザの会話例に対してアンケートによる評価を行い、ユーザの興味ベクトルを導入する効果や話題転換のタイミング決定法などを検証し、手法の改善を行う予定である。

参考文献

[1] 樋口真介,ジェプカラファウ,荒木健治. Web を利用した連想単語及びモダリティ表現による雑談システム，言語処理学会全国大会，PA1-7, pp.175-178. 2008.

[2] 徳久良子,乾健太郎. Web から獲得した感情生起要因コーパスに基づく感情推定，言語処理学会全国大会, A1-8,　pp.33-36. 2008.

[3] 斉藤 哲也,広田 健一,星野 准一.　Web 情報を用いたキャラクタの発話・世間話モデル. 情報処理学会研究報告, NL-181,　pp. 53-58. 2007.

[4] 藤本英輝,高梨克也,河野恭之,木戸出正継. 概念的関連性に基づく雑談の話題転換点分析，人工知能学会全国大会, pp.2G3-01. 2004.

[5] 佐竹聡,川島英之,今井倫太. ニュースコンテンツ提示ロボットにおけるユーザ興味を考慮したコンテンツ選択手法，電子情報通信学会，信学技報, DE2005-50,　pp119－124. 2005.

[3] http://kizasi.jp/