主格助詞を考慮した会話文の自動生成

西尾 友佑　　韓 東力
日本大学大学院 総合基礎科学研究科

1. はじめに

近年、非タスク指向型会話システムへの期待が高まっている。しかし、ユーザの興味・関心に対応しつつ、人間らしい柔軟な会話文を提供することは難しい。

これまでに、日本でも非タスク指向型会話システムの研究開発は盛んに行われている。文生成の観点から見てみると、会話文を生成する毎に、必要な単語を含む文章を選んで辞書を生成し、応答文を作成する『人工痴能うずら』[1]が、相手の発言を記憶し学習する『人工無能ロイディ』[2]が開発された。また、高橋ら[3]の研究では、Web 検索と単語 n-gram モデルを組み合わせることにより、柔軟な文生成を行うことに成功している。違う観点から見てみると、吉岡ら[4]の研究では、話者の個性・嗜好情報を考慮した応答文生成が行われている。前田ら[5]の研究においては、ユーザの興味から話題転換する瞬間を調べ、発話文の生成を行った。以上の先行研究において、ある程度の成果は得られたが、大きく二つの課題が残っていた。一つは、実際の人間が話すような口調の文が作られないために、形式がおかしい会話文が生成されてしまうこと。もう一つは、テンプレート化されているために柔軟性の無い会話文が生成されてしまうことである。

2. 先行研究紹介

前述した課題を解決することを目標として、会話の学習に Web を用いた文章生成と、その発話文に一定のキャラクター性を与えるため、『役割語』によるキャラクター性の付与を行うシステムの研究開発が、Han ら[6]によって行われた。

この研究で開発されたシステムは、三つのモジュールによって構成されている。まず、入力文を分析し、Web 検索を用いてキーワードを抽出する Planning Block。次に、渡されたキーワードからマルコフ連鎖を利用して会話文を生成する Generating Block。最後に、会話文に対して役割語変換を適用し、キャラクター性を持たせた上で出力をする Encoding Block。これら三つのモジュールがユーザに対して会話文を提供する。

このシステムはテンプレート化されていない柔軟な会話文の生成には成功した。しかし、会話文間の前後関係を考慮していないため、話題が断ち切られてしまう事例も数多く見られた。また、一つのキーワードから確率のみを指標として次の形態素を選定していたため、日常生活で見られる会話文とは異なる形態素間の繋がりの会話文が生成されてしまう事例が見られた。

これらの事例を踏まえ、Nishio ら[7]は、直近の会話との繋がりを考慮しつつ、意味の通じやすい会話文を生成する会話システムの研究開発を行った。

この研究では、直近のユーザの複数の発話文に対して窓という概念を適用することによって、直近のユーザの複数の発話文中の名詞・動詞・形容詞に重みを付与すること。その後、相互情報量の概念を用いて単語間の窓内における関連度を求めること。この二つの処理を通して、直近の会話の繋がりを考慮した上で、会話文生成に必要な単語対を生成する Extracting Block。生成した単語対のうち、体言を始点に、用言を終点とする断片を含む Tweet を Twitter API[8]を用いて検索し、始点から前へ、終点から後ろへマルコフ連鎖の考えをもとに文生成を行うことによって、意味の通じやすい会話文の生成を行う Multiple Word Generating Module。以上の二つを開発した。このシステムの全体図を図1に示す。

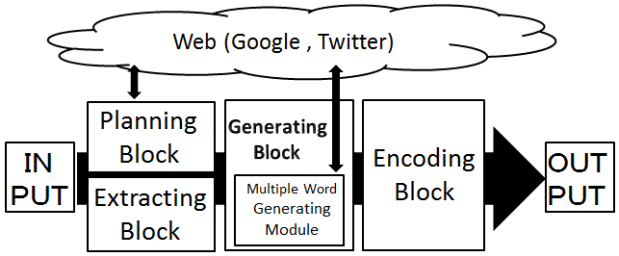

図1　先行研究[7]のシステム全体図

その結果、話題の繋がりを意識した会話文の生成に成功したが、断片からマルコフ連鎖のみを指標としているため、生成した会話文の長さが長くなってしまい、意味の通じにくい会話文が数多く生成されてしまう。といった問題点が見られた。

3. 本研究の概要

前述した問題点を解決することを目指して、本研究では、実行時間の短縮を目的として双方向 Tweet マルコフ連鎖辞書を作成した。また、Multiple Word Generating Module における文生成手法を、主格助詞を考慮した文生成手法に改良することで、より意味の通じやすい会話文生成を行うことを試みた。さらに、評価実験を行い、より意味の通じやすい会話文生成を行うことが出来たかの検証を行った。

なお、改良に合わせて、システムの全体図を図2に示すように変更した。

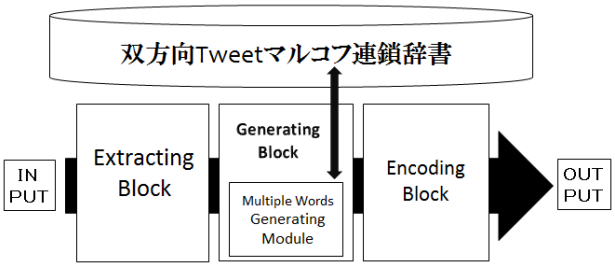

図2　本研究のシステム全体図

4. 双方向 Tweet マルコフ連鎖辞書

応答文生成にあたり、データソースとして用いた双方向 Tweet マルコフ連鎖辞書の生成手法について述べる。双方向 Tweet マルコフ連鎖辞書は、次の二つの辞書によって構成される。

1. 順方向マルコフ連鎖辞書
2. 逆方向マルコフ連鎖辞書

まず、1と2のマルコフ連鎖辞書に関しての説明を行う。1の順方向マルコフ連鎖辞書とは、文頭から文末にかけての形態素の並び方を記録した辞書であり、図3に順方向マルコフ連鎖辞書のデータ構造を示す。

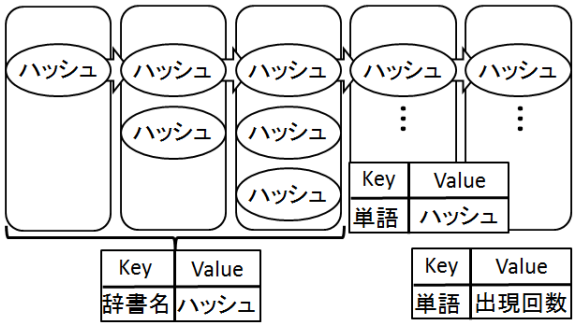

図3　順方向マルコフ連鎖辞書のデータ構造

また、2の逆方向マルコフ連鎖辞書とは、文末から文頭にかけての形態素の並び方を記録した辞書であり、1の順方向マルコフ連鎖辞書とは逆の情報が記録されている辞書である。双方向 Tweet マルコフ連鎖辞書は、1と2を合わせたデータ構造の辞書に対して、大量の Tweet を格納した辞書である。双方向 Tweet マルコフ連鎖辞書は、1と2を合わせたデータ構造の辞書に対して、大量の Tweet を格納した辞書である。

また、辞書に格納する Tweet は Public Timeline 上に一般公開されているものから JUMAN[9]に含まれている単語全てを検索クエリとして、TwitterAPI[8]を用いて抽出した Tweet を用いている。JUMAN に含まれている単語全てを検索クエリとした理由は、基本語彙を網羅していつつも、新聞記事に出てこない難しい単語や固有名詞が含まれていないため、日常会話に出てくる単語を必要最小限に取得することが出来ると考え、JUMAN を採用することとした。

5. 主格助詞に関する文法事項

この章では、文生成手法の改良にあたり採用した、主格助詞に関する文法事項である、埋め込み構造・補文標識・主格助詞保持の原則について紹介する。

まず、埋め込み構造について紹介する。柴田[10]によると、日本語には、「太郎が花子に本を贈った」のように、主語・間接目的語・直接目的語等の要素が一つの名詞から成る名詞句によって構成されている文がある。しかし、このような単純な文は稀で、普段我々が使う文は文の中にもうひとつの文を埋め込んだ文。すなわち、埋め込み構造を持つ文が多く存在する。

続いて、補文標識について紹介する。日本語において、文中に他の文が埋め込まれている際には、一つの文を他の文の補文として埋め込んでいるケースが数多く散見される。その際に、「こと」・「の」・「と」・「ように」の四つの単語が文を他の文に対して埋め込む際に使われる要素である。これらの単語を補文標識という。補文標識は文を名詞化するのとともに、日本語のように主語・目的語・述語の語順を持つ言語では、補文標識は補文の終端の位置を示す機能を有する。

以上より、補文標識を含む文章では一つの文が名詞化され、名詞であるかの様に一つの文に含まれている。そのため、補文標識を含む文は文章の長さが長くなりやすいと言える。

また、柴田[10]は『「文」は少なくとも一つの主格名詞節を含んでいなければならない』という原則を提唱している。これを「主格保持の原則」という。これは文の構造において、主格名詞節を含んでいる文章のみを的確な文とし、それ以外を非文とみなす原則である。

6. 断片選択

以上の各文法事項の紹介を踏まえ、これらの文法事項を適用する理由を以下に述べる。先行研究において、生成した応答文の長さが長くなってしまい、意味の通じにくい応答文が数多く生成されてしまうといった問題点があった。この原因を調査したところ、応答文を構成する断片が複数の文から成り立っているため、生成された会話文も長くなってしまうという事案が数多く見られた。この時、「主格保持の原則」より、文には必ず主格名詞節が一つ以上存在するため、複数の文によって構成された断片には複数の主格名詞節が含まれている場合が数多く存在することが考えられる。

より意味の通じやすい応答文を作るにあたり、断片の長さを短くし、簡潔な応答文を作成する必要がある。そのため、断片の長さを短くするためには主格名詞節が一つだけ含まれている断片を用いて応答文を生成することが望ましいと考えた。また、断片から文頭への文生成の時は、埋め込み構造より、補文標識などの語が連鎖候補とならないようにし、断片から文末への文生成では、マルコフ連鎖のみを指標とするのではなく、文を終わらせる表現が連鎖候補となるようにした。

以上より、文生成アルゴリズムを考案し、その中で応答文生成に必要な断片を選択しようと考え、断片選択アルゴリズム、ならびに断片選択式を考案した。以下に、断片選択アルゴリズム、断片選択式、断片から文頭への文生成、断片から文末への文生成の順に記述する。

まず、Extracting Block で抽出された単語対のうち、体言を W1、用言を W2と定義する。その上で、“W1…W2”という形式になっている断片を含む Tweet を、双方向 Tweet マルコフ連鎖辞書から検索する。この時に抽出した Tweet に対して、補文標識、句読点、括弧、顔文字の一部といった記号が含まれている断片を断片候補から除外する。除外されなかった断片候補を後述する断片選択式に適用し、最もスコアの高い断片を文生成に使用する断片として採用する。

続いて、断片選択式を記述する。

$$Score_i = \begin{cases} 0 & if\ (FL_i = Subst_i + Decl_i\ \&\ N_i = 0) \\ \frac{1}{M_i} & if\ (FL_i > Subst_i + Decl_i\ \&\ N_i = 0) \\ \frac{1}{N_i \cdot M_i} & if\ (FL_i > Subst_i + Decl_i\ \&\ N_i > 0) \end{cases}$$

$Score_i$：各断片に付与するスコア
i：断片の番号
FL_i：断片の文字数
$Subst_i$：断片の先端にある体言の文字数
$Decl_i$：断片の終端にある用言の文字数
N_i：i番目の断片中の主語となりうる助詞の数
M_i：i番目の断片中の形態素数

断片選択式において、$Subst_i$ ならびに $Decl_i$ を考慮しない時、N_i と M_i のみを含んでいる断片が、常に最大値を取ってしまう。そこで、断片の文字数である FL_i を指標のひとつとして採用し、$FL_i > Subst_i + Decl_i$ を満たす断片のみを計算対象として採用するものとする。また、$Score_i$ が同値の断片が存在する時、「主格保持の原則」より、N_i を含む断片を優先するものとする。

次に、断片から文頭への文生成手法を記述する。抽出された断片の先頭に存在する体言からマルコフ連鎖を用いて文生成を行う。また、補文標識、助詞や助動詞などの付属語、句読点、括弧、顔文字の一部といった記号、そして BOS が連鎖候補となった際に文生成を中止し、それ以外の品詞が連鎖候補となる限り、文生成を継続するものとする。

最後に、断片から文末への文生成手法を記述する。抽出された断片の末尾に存在する用言からマルコフ連鎖を用いて文生成を行う。また、句読点や EOS が連鎖候補となった際には文生成を中止し、名詞や動詞などの自立語以外の形態素が連鎖候補となる限り、文生成を継続するものとする。以上が本研究における文生成手法である。また、文生成アルゴリズムについてまとめた図を図4に示す。

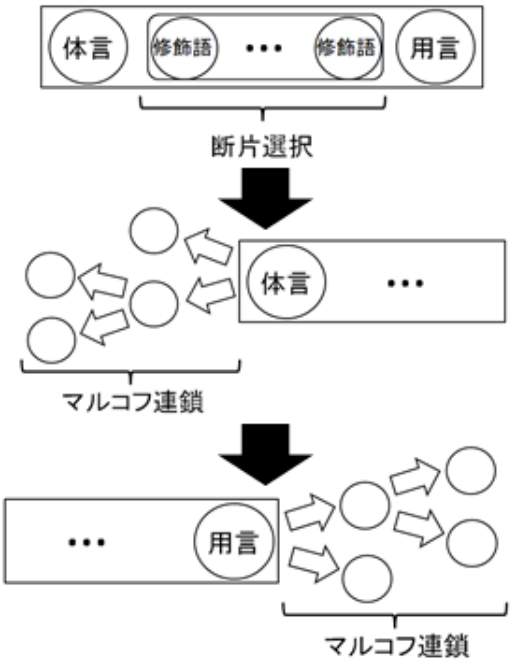

図4 文生成手法アルゴリズム

7. 評価実験

この章では、本研究の内容を実装したシステムを用いて、評価実験を行う。まず、評価実験の目的について記述する。評価実験の目的は、主格助詞を考慮して文生成を行うことに対する効果を検証することである。

次に、実験方法について記述する。実験方法は、先行研究の内容を反映した Ver.1と前述した内容を全て反映させた Ver.2の二つのシステム対話ログを三編ずつ用意し、それぞれのログを読んだ上で、

- 一文単位で意味が分かりやすいか。
- 一文単位で簡潔な形式で書かれているか。
- 会話文全体で意味が分かり易いか。

の三点について五段階評価(1点～5点)で点数を付けてもらうこととする。また、ログを用いて評価実験を行う理由としては、被験者の話す内容により、生成される対話ログの内容に大きなバラつきが発生してしまい、評価に差が出てしまうことを避けるために、各システムが生成したログを用いて評価実験を行うものとする。以下に、実験結果をまとめた表1を記載する。

	Ver.1	Ver.2
意味	2.8	3.7
形式	2.9	3.7
全体	2.8	3.4

表1　評価実験結果

なお、表1の2行目が、一文単位で意味が分かりやすいか。3行目は、一文単位で適切な形式で書かれているか。4行目は会話文全体で意味が分かり易いか。に関する被験者の回答の平均点となっている。

8. 考察と今後の課題

評価実験の結果より、主格助詞を考慮した文生成を行ったところ、先行研究と比較して意味が通じやすく、簡潔な会話文の生成に成功したことが示された。また、一文単位での分かりやすい文生成が可能になった結果、本システムとの会話という行為全体における評価を向上させることに成功したことも示された。さらに、双方向 Tweet マルコフ連鎖辞書のみをデータソースとし、断片を抽出して文生成を行うことで、文法的な崩れも少なく、先行研究で開発された内容を損なうこともなかった。

しかし、断片中に含まれた単語と断片から連鎖することによって応答文に含まれた単語の意味的な対応関係を考慮していなかったため、内容にズレのある会話文が生成されてしまうケースも見られた。

今後は、Tweet マルコフ連鎖辞書のさらなる拡張を行い、日常会話でよく見られる語彙の使われ方の拡張を目指す。また、アルゴリズムの改良を行い、断片とそれ以外の要素との間に意味的な連携を付与し、一文単位での内容を適切なものとすることを目指す。以上を通して、より人間らしい会話文生成を目指していく予定である。

参考文献

1. http://www.din.or.jp/~ohzaki/uzura.htm
2. http://rogiken.org/SSB/reudy.html
3. 高橋瑞希, Rafal Rzepka, 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 16 回年次大会論文集, PA1-32, pp.391-394, 2010.
4. 吉岡 孝治, 吉村 枝里子, 渡部 広一, 河岡 司:“話者の個性・嗜好情報を考慮したコンピュータ会話処理”:FIT2008(第７回情報科学技術フォーラム) 情報科学技術フォーラム講演論文集 7(2), pp.289-290, 2008.
5. 前田 和希, 宋 鑫, 國政 裕友樹, 豊田 博之, 韓 東力:“雑談システムにおける話題展開の性能向上”:言語処理学会第16回年次大会論文集, D2-5, pp.250-253, 2010.
6. Dongli Han, Yasuhiro Kinoshita, Ryu Fukuchi, Tsurugi Kousaki: “Utterance Generation Using Twitter Replying Sentences and Character Assignment.”: International Journal of Digital Content Technology and its Applications, Vol. 5, No. 10, pp.119-126, 2011.
7. Yusuke Nishio, Dongli Han: “Automatic Utterance Generation by Keeping Track of the Conversation's Focus within the Utterance Window”:Lecture Notes in Artificial Intelligence 7614 (Advances in Natural Language Processing), Springer. pp.322-332, 2012.
8. https://dev.twitter.com
9. http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
10. 柴田 方良:“日本語の分析-生成文法の方法-”:大修館書店, 1978.