# 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

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

近年，機械と人が対話を行うという話題性やエンターテイメント性から，非タスク指向型の対話システムに注目が集まっている[1]．しかしながら，非タスク指向型の対話システムは，タスク指向型と異なり，無数の話題に対応しなければならない．そのため，自由度の高い発話文生成手法や，どのような話題にも対応するための知識が必要となる．本稿では，柔軟な文生成を可能とする単語 n-gram モデルに着目し，入力文に対応した応答文を生成する手法を提案する．また，本手法では生成した複数の文に対して評価を行い，最終的に出力する文を決定するが，その方法の妥当性について，比較実験を行って検証する．

## 2. 提案手法の概要

本稿で提案する手法を組み込んだ対話システムが行う処理の流れを，図 1 に示す．以下，各処理について述べる．

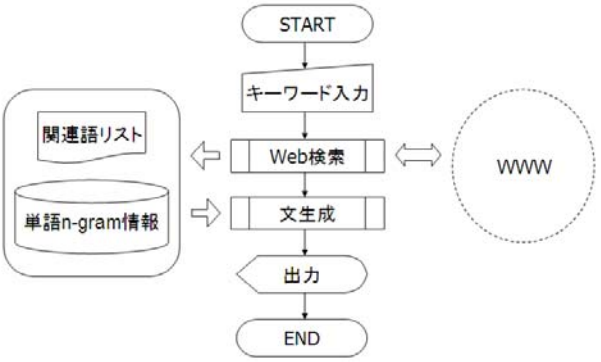


図 1　処理の流れ

### 2.1 Web 検索の利用

本手法では，WWW を一つの巨大な知識データベースとして扱う．WWW を知識源とした対話システムが有効であることは，既に確認されている[2]．まず，ユーザの入力文に含まれる自立名詞および形容詞を，キーワードとして抽出する．そのキーワードをクエリとして Web 検索を行い，スニペット（検索語を含む箇所を抽出した，テキストの断片）を取得する．

さらに，得られたスニペットに形態素解析を施した後，出現頻度の高い語（関連語）を抽出し，これらを文生成の際に利用する．本システムでは，形態素の解析に MeCab[3]を使用する．なお，予備実験[4]の結果に基づき，名詞は出現頻度上位 100 個，形容詞は上位 10 個を，関連語として取り出す．

例えば，「北海道の旅行で楽しいことがあった」という入力文からは，「北海道，旅行，楽しい」といったキーワードを抽出することが可能である．「こと」は非自立名詞なので，キーワードには該当しない．これらのキーワードをクエリとして Web 検索を行い，前述の条件に適合する単語を抽出すると，「札幌，体験，面白い，すごい」といった関連語を取得することができる．

また，以上の処理を行う際，スニペットから単語 n-gram 情報を同時に取得する．詳細は **2.2** で述べる．

### 2.2 文生成

文生成には，チャットの対話ログから適切な応答を学習して文を抽出する方法，あらかじめテンプレートとして用意された応答ルールに適宜単語を代入する方法，n-gram モデル（マルコフモデル）を用いて文を作り出す方法など，様々な手法が存在する[5]．これら 3 つの文生成手法をログ型，テンプレート型，n-gram モデル型と分類した場合，それぞれ表 1 のような特徴を持つ．

本手法では，自由度の高い文を作成することが可能である n-gram モデルを，単語単位に拡張（単語 n-gram モデル）した上で，文の生成法として採用する．上述した他の 2 つの手法は，その仕組み上，十分な対話ログや応答ルールがなければ，画一的な反応を繰り返すことになる．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，準備段階でコストがかかってしまう．

ただし，単語 n-gram モデルでは文全体の整合性や文の意味などは考慮されないため，単純にこのモデル適用するだけでは，応答として適切な文を作ることは困難である．

そこで，本システムでは，以下の手順で応答文を生成する．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，

表 1　文生成手法の種類とその特徴

| | |
|---|---|
| ログ型 | 対話ログの中から適切な応答文を探す<br>〇文脈に沿った発話が可能<br>×膨大な対話ログという知識源が必要 |
| テンプレート型 | 応答ルールに従う<br>〇文法に違和感のない文を出力可能<br>×人手でルールを構築する必要がある |
| n-gram モデル型 | n-gram モデルをコーパスに適用<br>〇型にはまらない柔軟な文生成が可能<br>×単体では文脈や文法を考慮できない |

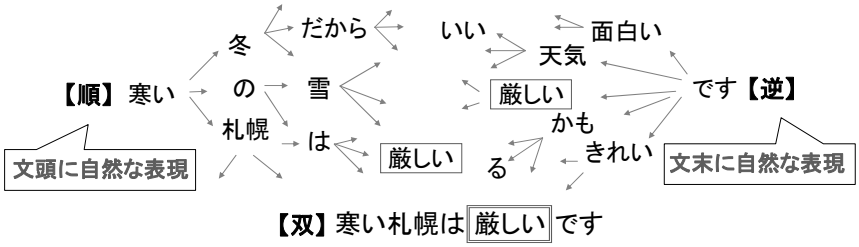


図 2　n-gram モデルを用いた文生成

その単語から単語 n-gram 情報に従って次々に単語を繋げていく．これを連鎖と呼ぶ．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．

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．順生成は文頭から連鎖を辿っていく方法，逆生成は文末から連鎖を辿っていく方法である．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，両者を組み合わせた「双方向生成」も，同時に行う．以上の操作(図 2)を繰り返すと，多数の出力候補文が生成される．

単語 n-gram モデルは，その性質上，自然な文が生成されたり不自然な文が生成されたりと，結果が安定しない．しかしながら，一回の試行で多くの文を生成した後，何らかの方法で最も自然な応答文を選び出すことができれば，応答精度の変動を抑制することが可能であると考えられる．

### 2.3　文生成

**2.2** で得られた複数の候補文に対し，妥当性を評価するためにスコア付与を行う．このスコアが最も高いものを，最終的な応答文として出力する．

各文のスコアは，文の長さが適切かどうかを判定する「文長スコア」，ユーザの発話内容に関連している語の有無を反映する「関連語スコア」，及び文法や不適切な語句などの要素を考慮して算出される．具体的な計算式は式(1)の通りである．また，スコアの分類を表 2 に示す．なお，各スコアの値や閾値は，試行を繰り返して調整を行った結果，妥当と判断されたものである．

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

$sL(S_i)$ ; 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　$-1 < \alpha < 1$

表 2　生成された文のスコアを算出するための要素

| | |
|---|---|
| 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（+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>→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 |

## 4.　性能評価実験

本章では，2 つのシステムを用いた比較実験について述べる．

### 4.1　文生成処理について

本実験では，**3.** で述べた一連のスコア算出処理が，単語 n-gram モデルによる文生成の結果を改善しているかどうかを調べる．実験は，ある入力文に対してスコアを基に応答文を選択した結果(A)と，生成された文の中からランダムに選択した結果(B)を比較する形式で行う．被験者は 10 代から 40 代の男女 15 名，使用した入力文は予備実験[4]時に収集したサンプル 50 文である．また，評価は 5 段階で，内訳は「A の方が優れている・どちらかといえば A が優れている・同じくらい優れている（劣っている）・どちらかといえば B が優れている・B の方が優れている」とした．

実験結果を表 3 に示す．なお，数値は被験者が選んだ文数の平均値である．スコア計算を行わなかった B と比較し，スコア計算を行った A の方が，応答として約 5 ポイント適切な文を出力していることがわかる．

表 3　文生成処理　実験結果

| | |
|---|---|
| A の方が優れている | 11.0 |
| どちらかといえば A が優れている | 11.2 |
| どちらも同じくらい優れている(劣っている) | 15.2 |
| どちらかといえば B が優れている | 5.4 |
| B の方が優れている | 7.2 |

### 4.2　他システムとの比較

本節では，ベースラインに樋口ら[2]が提案した雑談システム”Modalin”を用いた印象評価実験を行う．これは，Modalin が本システムと同様に Web 上のデータを知識源とし，関連語を抽出している点と，文生成法にテンプレートを使用している点を考慮したためである．これにより，知識源の違いによる出力の差異を小さくすることができ，なおかつ単語 n-gram モデルを使用した生成法の有効性を確認することが可能となる．

実験は，「1 つの発話文を入力すると，本システムと Modalin からそれぞれ応答文が出力される」というインターフェイスを作成して行った．被験者は，10 代から 40 代の男女 12 名である（自然言語処理についての知識を持つ者を 1 名含む）．「発話文入力→応答文出力」を 1 つの試行とし，1 名につき最低 10 回の対話を行った．その後，被験者に以下のアンケートを実施した．なお便宜上，本システムを「システム A」，Modalin を「システム B」という名称で，被験者に伝えた．

アンケート
(1)個別の試行結果についてシステム A とシステム B それぞれの出力を比較し，五段階評価を行う
（A がよい(2)・どちらかといえば A がよい(1)
同程度(0)・どちらかといえば B がよい(-1)
B がよい(-2)　　( )内は計算用の数値）

(2)各試行結果を総合的に評価した場合について以下の項目に答える
・システム A（システム B）は文法的に自然か
・システム A（システム B）は意味/内容的に自然か
・より知識があると感じたシステムはどちらか
・より興味深い応答を行ったシステムはどちらか
※上 2 問は五段階評価（1～5）とする

実験結果を表 4 に示す．なお，アンケート(2)の結果における数値は，全回答の平均値である．

(1)の結果について，五段階評価は-2 から 2 までの配点で行われるため，ある被験者が付けた評価の平均値が 0 を超えていれば，その被験者はシステム A（本システム）をシステム B（Modalin）より高く評価したとみなすことができる．アンケートの結果，12 名中 10 名が，個別の試行結果を通じてシステム A をより高く評価した．

また，(2)の結果について，システム A とシステム B の間に文法面での評価差はほとんど見られず，0.1 ポイントであった．一方，意味/内容面の評価および二択式の

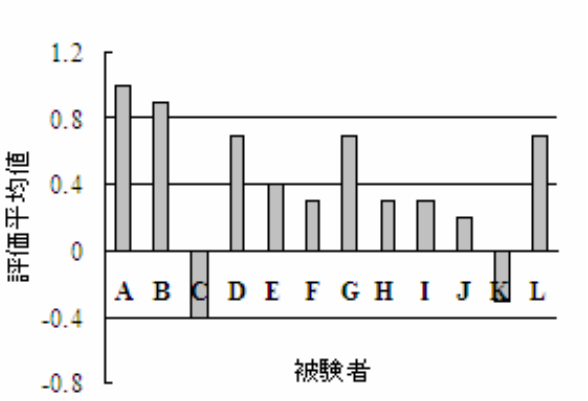


図 3　他システムとの比較実験結果(1)

表 4　他システムとの比較実験結果(2)

| システム A | | システム B | |
|---|---|---|---|
| 文法 | 意味/内容 | 文法 | 意味/内容 |
| 2.9 | 3.3 | 3.0 | 2.3 |

| より知識ある応答 | より興味深い応答 |
|---|---|
| A：9 名 / B：3 名 | A：11 名/ B：1 名 |

印象評価では，システム A の方がシステム B より高い評価を得ている．**2.2** で挙げたように，単語 n-gram モデルはテンプレートを使用した場合と比較し，文法的・意味的に自然な文を生成しにくい．このことから，本システムの手法には，単語 n-gram モデルの欠点を補いつつ，非タスク型の対話に適した応答を行う効果があるといえる．

## 5.　考察

本手法における文生成法は，ヒューリスティックな工夫を施した部分も多い．スコア計算の有効性については，**4.1** および **4.2** で行った実験において確認されたが，それぞれの要素が具体的にどのように働いているかについて，実際の対話データを基に調査を行った．

例えば，図 4 の応答について考える．「入力」は被験者が入力した文，「出力候補」はスコア計算前に生成された候補文および基本スコア（文長スコア＋関連語スコア）の一部，「出力」はシステムが最終的に選択した文，そして「除外」はスコアが 0 となった場合におけるその理由である．

(2)の文は文長スコアが(4)と等しく，また関連語スコアが(5)と等しい．また，(1)は文が冗長であるため文長スコアは 0 だが，その中に関連語（今回抽出した語は「審査，演技，テレビ，面白い」）を 4 つ含んでいるため，関連語スコアが高くなっている．もし，基本スコアとして文の長さ，あるいは関連語の有無のいずれかしかチェックを行わなかった場合，(2)以外の，(1)や(4)といった文が出力される可能性が生じる．本システムは意味解析を行っていないため，関連語の有無というシンプルな要素でも，ユーザの発話に関係があると思われる文を選択する手掛かりとなる．当然ながら，意味解析処理を搭載すること

図 4 入力に対する処理と応答（抜粋）

<入力>
今年のフィギュアは面白い！誰が勝つかドキドキ
<生成>
(1)自分は細かい審査基準があって、それに沿って演技してるのを見るのが好きでテレビでやっているとほぼ毎回見ていて、今も面白いブログです。(0+8)
(2)細かい審査基準があったので余計に面白かったのです。(11+2)
(3)情報局～面白い事探し～アニメ・ゲーム・PC・映画関係の話から色々と http://www... (11+4)
(4)格闘技しか思えないんですけど(11+0)
(5)嬉しいな(1+2)
<出力>
細かい審査基準があったので余計に面白かったのです。
<除外>
(1)文長スコアが 0 点
(3)ストップワード（URL の断片,“情報”）

により，さらに細かい判別が可能となると考えられる．しかし観点を変えれば，複雑な意味解析を使わなくとも，関連語をチェックすることで簡易的に意味・話題の整合性を保つことが可能，とも考えられる．このケースでは，(4)ではなく(2)の文が応答文として選択されたという部分が，具体例として該当する．

一方，基本スコアに注目すると，(3)の文が 15 と最も高い値を持つ．しかし，(3)はノイズを含んでいる上，内容も応答文として相応しくない．本手法ではあらかじめストップワードを設けることにより，こういった「基本スコアが高くなってしまう不適切な表現を含む文」を候補から除外している． ストップワードの設定は人手によるが，元となるコーパスが Web テキストであるため，ノイズの傾向をつかむことが可能である．これは，Web 情報を知識源として利用する際の，利点の 1 つであるといえる．なお，ストップワードの具体的な例は，表 2 に示した通りである．

同様に，1 対話を 1 ターンとして 50 の対話例について調査したところ，スコア付与処理について表 5 に示した効果を確認することができた．

個々の処理別に見た場合，文長スコアの算出が有効に働いたケースは 36 件，関連語スコアの算出が有効に働いたケースは 29 件，ストップワードや冗長な文に注目した処理が有効に働いたケースは 41 件存在した．つまり，基本スコアを計算した後に施す処理の効果が最も大きい． 1 回の試行で生成される候補約 16 文のうち，ストップワードを持つ文は 3 文から 6 文ほど存在したため，これらを除外することができる効果は大きい．しかしながら，2 つ以上の要素が効果的に働いたケースも，全体の 4 割から 5 割を占めている．これは，基本スコアを算出する処理とその後の処理が，互いに補完していることを表している．ストップワードの有無や文の長さだけで，出力候補から文を除外するかどうかを決定した場合，50 の入力文のうち 36 文の試行において， 応答文として不自然な

表 5 対話例の考察

| 該当するケース | [件] |
|---|---|
| A | 36 |
| B | 29 |
| C | 41 |
| A, B | 25 |
| B, C | 25 |
| A, C | 23 |
| A,B,C | 19 |

A：文長スコアを考慮した結果
不適切な文の出力を抑制することができたケース
B：関連語スコアを考慮しなければ
不適切な文の出力を抑制することができたケース
C：基本スコアに関わらず候補から除外する処理により
不適切な文を除外することができたケース

文が出力されてしまった．

## 6. まとめ

Web 検索と単語 n-gram モデルを組み合わせることにより，テンプレートに依存しない，柔軟な文生成を実現する手法の提案を行った．また，複数の角度から文を生成した後に各文を評価するという処理に，不適切な文や違和感のある文の出力を抑制する効果があることが確認された．さらに，スコアを計算する処理について，考察を行った．

今後は，より自然な文を選出することが可能なスコア付与方法について調査する予定である．また，ヒューリスティックに値を定めた部分（文長スコアの閾値など）について，統計的な分析を行い，有効性を追究する．その上で，独立した試行で継続する対話ではなく，現在の話題を考慮し，専門的なテーマに遷移しても対話を続けることができる雑談システムの作成を目指したい．

## 参考


[1]「酢鶏」作者が語る「一家に一台、人工無脳」の未来像 http://ascii.jp/elem/000/000/417/417954/

[2] 樋口真介,ジェプカ・ラファウ, 荒木健治:“Web による単語共起頻度及びモダリティを用いた雑談システム”,平成 19 年度　電気・情報関係学会北海道支部連合大会講演論文集, pp.148,2007.

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

[4] 高橋瑞希,ジェプカ・ラファウ,荒木健治 :“単語 n-gram モデルを用いた文生成手法の改善案”, 平成 21 年度電気・情報関係学会北海道支部連合大会

[5] 森部 敦, 毛利 公美, 森井 昌克:“自動会話システム(人工無能)の開発とその応用：Web テキストからの会話文生成と会話形成に関する研究,” 電子情報通信学会技術研究報告, Vol.105, No.283(20050908)