Vol.2017-DBS-164 No.10
2017/1/17

トピックモデルによる話題知識を考慮したテンプレート穴埋め型発話生成

久保田 豊久[1,a)]　若林 啓[1,b)]

概要：近年，携帯端末やロボット技術の発達によって，対話システム高度化への期待が高まっている．多様な状況での対話システム活用のため，雑談に対応する機能の実現が求められるようになると言われている．現在，ルールベースや web 上テキスト資源から対話文を選択するといった手法が提案されている．しかし，ルールベースは多様な話題に合わせた構築が難しいという問題，テキスト選択手法は資源量に合わせて返答が限られてしまうという問題が指摘されている．これらの問題から，発話生成型手法の実現が期待されるが，これまでに提案されているテンプレート穴埋めベースの手法は，テンプレートと置換語句の連接関係を考慮出来ていない．そこで，本研究はトピックモデルを併せて使用することで話題知識を考慮したテンプレート穴埋めベースの発話生成手法を提案し，既存手法との比較を行うことで，提案手法の有用性を検討する．

キーワード：雑談対話，発話テンプレート，発話生成

Kubota Toyohisa[1,a)]　Wakabayashi Kei[1,b)]

1. はじめに

近年，ロボット技術や携帯端末の発達とともに，対話システムの実装が人々の身近に現れたことで，その高度化に注目が集まっている．自然言語対話によって様々な形式の情報を伝えることができるが，その中でも特定の目的を持たない対話である雑談は，人間同士の対話全体の 6 割を占めることが報告されている．[1] 人間が自然に対話システムを利用する上で，システムが雑談に対応する機能をもつことは，システムに対する信頼感の向上や，話者の潜在的な情報要求の発見において重要な役割を果たすと考えられる．このことから，特定の話題に限定されないオープンドメインな雑談対話に対応できるシステムへの期待は，関連の産業において高まっているといえる．

オープンドメインな対話システムとして，これまでには，web 上のテキストから関連度の高い文章を選択する手法 [2][3] や，人手で作成したルールに基づいた返答を行う対話システムの構築手法 [4] が提案されている．ルールベースの手法は，設定したルールで想定された範囲内の形式や話題の対話を行う上では自然な対話を行うことができるが，対応可能な話題を増やすとルールが難解になっていくという問題を抱えている．一方，web テキストの選択に基づく対話手法は多くの話題に対応できるが，web テキストの集合は有限であるのに対して，自然言語対話は状況ごとに無限に多様であるため，状況と web テキストとの間の「ずれ」が本質的に解消できないという問題を持つ．

本研究では，発話の形式と話題を分離し，それらを組み合わせて状況に合わせた発話を生成する方式の手法を検討する．安藤ら [5] は，生成方式の対話システムの実現を目指し，テンプレートを用いた発話生成手法を提案している．これは，過去の会話履歴データにある発話文の一部を空欄に置き換えることで作成した発話テンプレートを用いて，発話生成時に空欄に適切な語を当てはめることで，文構造を保ちつつ，様々な話題に適応可能な手法である．しかし，この提案手法では，テンプレートを選択する基準がなくランダムに選ばれることや，空欄に当てはめる単語の選択方法において前後の単語や文脈との連接関係を考慮していないといった問題がある．本論文では，テンプレートを用いた発話文生成手法において，Support Vector Machine

[1] 筑波大学
University of Tsukuba，Ibaraki 305-8550，Japan
[a)] s1521613@u.tsukuba.ac.jp
[b)] kwakaba@slis.tsukuba.ac.jp

(SVM) を用いた発話タイプ推定に基づいたテンプレート選択と，N-gram モデルによる当てはめ単語の連接関係の考慮，トピックモデルを用いた話題を考慮した単語選択により，より尤もらしいオープンドメイン発話文の生成を行う手法を提案する．実験により，提案手法と従来手法の比較を行い，有用性を検証する．

2. 先行研究

安藤ら [5] は，特定のチャットルームでの会話ログを用例文集合として用いて，テンプレートに基づく発言モデルによりテキストコミュニケーションを行う対話システムの構築する手法を提案している．発言モデルは，用例文集合から構築した空欄を含む発話テンプレートの集合から 1 つを選び，別途用例文集合から抽出した語彙用例集から選んだ語彙列を空欄に当てはめることで発話文生成を行う．テンプレートの空欄は，「名詞句」と「述部」の 2 種類とする．テンプレートの構築では，まず用例文に対して形態素解析を行い，名詞が連続している部分単語列を「名詞句」，最初に出現した動詞から文末までの部分単語列を「述部」と置き換えて，テンプレートとする．テンプレートの例を図 1 に示す．語彙用例集は，用例文集合から抽出した連接辞書であり，それぞれの単語について，次の位置に出現したことのある単語のリストが登録される．発話文の生成では，ランダムに選択されたテンプレートについて，それぞれの空欄に置換語句を当てはめる．置換語句の生成は，以下の手順で行う．

- 空欄が名詞句ならば名詞を，述部ならば動詞を，語彙集合からランダムに選択し，1 番目の単語とする．
- $n = 2, 3, 4, ...$ について，以下を繰り返す．
  - 語彙用例集から， $n-1$ 番目の単語の次に出現しうる単語のリストを参照し，その中からランダムに単語を 1 つ選ぶ．選んだ単語を n 番目の単語とする．
  - 空欄が名詞句ならば名詞，述部ならば句読点が，終了語彙である．もし，選んだ n 番目の単語が，終了語彙ならば，ループを抜けて終了する．

この手法は，基本的に名詞句と述部のランダムな置き換えに基づいており，意味の通らない文が生成されることが多いが，安藤らは改善を施す余地が十分にあると述べている．この手法の利点は，文構造を保ちつつ，豊富な種類の文生成が行える点にあるが，置換語句選択する際に，テンプレートとの連接関係の考慮や話題の考慮が出来ていないという問題がある．提案手法は，N-gram モデルとトピックモデルを用いることで，連接関係と話題を考慮した置換語句の選択を行うことで，これらの課題を解決することを検討する．

3. 生成手法

先行研究では，名詞句と述部という空欄を設定することでテンプレート生成を行っていたが，本研究では単純化のために名詞のみを空欄に設定した．文の生成に際しては，テンプレート選択方法として Support Vector Machine (SVM)[6][7] による発話タイプ推定を用い，N-gram モデル，トピックモデル，両モデルを組み合わせた提案手法によって文生成を行う．

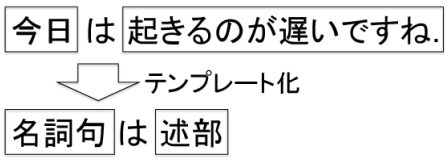

図 1 テンプレート化

3.1 提案手法の概要

テンプレート穴埋め発話生成は，テンプレートを選択するステップと，空欄を穴埋めする置換語句を選択するステップに分けられる．処理の概略を図 2 に示す．

テンプレート選択のステップでは，テンプレート同士の隣接関係を考慮した方法と，発話タイプを考慮した方法を適用する．コーパス全体をテンプレート化する際に，テンプレートの隣接関係を記録しておき，入力文テンプレートと全く同一のテンプレートがコーパス中に存在する場合，コーパス中でその次の返答に用いられているテンプレートを選択する．入力文のテンプレートがコーパス中に存在しない場合，入力文の発話タイプの推定を行い，その返答の発話タイプにふさわしいと考えられる発話タイプをもつテンプレートをランダムに選択する．発話タイプの推定に基づくテンプレート選択については 3.2 節で述べる．

テンプレート選択後の置換語句選択ステップにおいては，トピックモデルのみを用いた手法，N-gram のみを用いた手法，およびトピックモデルと N-gram の両方を考慮した提案手法のそれぞれで置換する語を選択する．トピックモデルの学習については 3.3 節で，それぞれの置換語句の選択手法については 3.4 節で述べる．

表 1 発話タイプ遷移表

入力発話タイプ	出力発話タイプ
sv	sv,qw,fo
qw	no,fo
no	sv,qw,fo
fo	sv,qw,fo
%	%

3.2 発話タイプの推定

テンプレート選択のステップでは，大まかな発話の意図が決定すると考えられることから，入力文の発話タイプの推定に基づいてテンプレートの選択を行う．発話タイプは SWBD-DAMSL タグ [8] を独自に統合し，表 2 に示

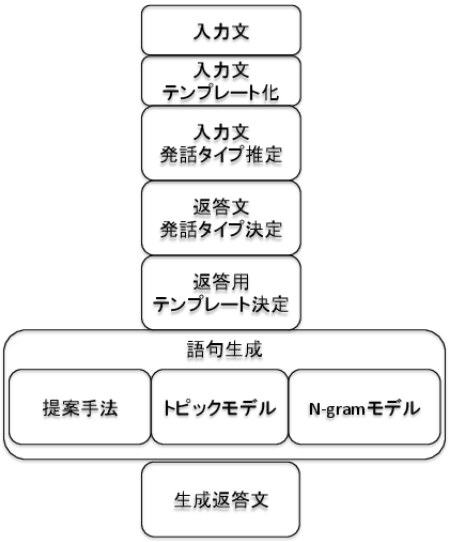

図 2 生成手順概略

すように，Statement-opinion (sv)，Wh-Question (qw)，Other answers (no)，Other-forward-function (fo)，Uninterpretable (%) の５つを用いる．本研究では，Twitter から抽出した会話データに含まれる 800 会話について，人手で上記の 5 種類の発話タイプのいずれかを付与するアノテーションを行った．

この訓練データを用いて，Support Vector Machine (SVM)[6][7] により発話タイプの分類器を構築する．アノテーションされたそれぞれの発話データを，まずテンプレートに変換する．その上で，テンプレートに含まれる全ての単語を用いて単語頻度ベクトルに変換する．この単語頻度ベクトルを特徴ベクトルとして，SVM を学習する．

学習した SVM を用いて，コーパス全体から構築したテンプレートそれぞれについて発話タイプを推定し，テンプレート集合全体を 5 種類のそれぞれの発話タイプに従って分類する．発話生成を行う際には，入力文の発話タイプを推定し，あらかじめ表 1 のように定めた発話タイプ遷移表に従って，入力文の発話タイプの返答としてふさわしいと考えられる発話タイプをランダムに一つ選択する．選択した発話タイプをもつテンプレートを一つランダムに選択し，返答文のテンプレートとする．

SVM の実装には，scikit-learn の LinearSVC を用いた．

表 2 SWBD-DAMSL 統合タグ

SWBD-DAMSL タグ	内容
Statement-opinion (sv)	意見
Wh-Question (qw)	質問
Other answers (no)	回答
Other-forward-function (fo)	挨拶等
Uninterpretable (%)	解釈不能

3.3 トピックモデルの学習

本研究では，トピックモデルとして Latent Dirichlet Allocation (LDA) を用いる [9]．LDA は，Bag-of-words で表現された文書の集合 $X = \{X_1, ..., X_D\}$ ， $X_d = \{x_{d1}, ..., x_{dN_d}\}$ に関する確率的生成モデルである．文書 d に含まれる個々の単語 x_{di} は，トピックと呼ばれる潜在変数 z_{di} に依存して生成される．各文書は個別のトピック分布 θ_d を潜在的にもつ．また，トピック k に対応する単語の確率分布を ϕ_k とする．LDA では，これらの変数の同時分布について，以下の条件付き独立性を仮定する．

$$p(X, Z, \theta, \phi|\alpha, \beta) = \prod_{d=1}^{D}\prod_{i=1}^{N_d} p(x_{di}|\phi_{z_{di}})p(z_{di}|\theta_d) \times \prod_{d=1}^{D} p(\theta_d|\alpha)\prod_{k=1}^{K} p(\phi_k|\beta_k) \quad (1)$$

ここで，K はトピックの数である．$p(\theta_d|\alpha)$ および $p(\phi_k|\beta_k)$ は，それぞれパラメータ α ，β_k のディリクレ分布である．$\beta = \{\beta_1, ...\beta_K\}$ とする．

LDA の学習とは，所与の文書集合 X に対して，尤もらしい Z, θ, ϕ の値を推定することである．本研究では，LDA の学習アルゴリズムとして Mimno ら [10] が提案した確率的変分ベイズ法を用いる．この手法は確率的最適化に基づいており，反復最適化の 1 回のイテレーションごとに文書集合 X から B 件の文書をランダムサンプリングして得られたミニバッチを用いて学習を行う．ここで， B はバッチサイズと呼び，本研究では $B = 4000$ とした．この手法により，大規模な文書集合の学習を行う場合でもスケーラブルに学習を行うことが可能である．

このアルゴリズムにより，対話コーパスを 1 対話 1 文書とみなした文書集合を用いて LDA の学習を行う．学習により，各トピックの単語生成確率分布 ϕ が得られる．また，これを用いることで，未知の入力文章に対して，トピック分布 $\bar{\theta}$ を推論することができる．

3.4 置換語句の選択

本節では，選択されたテンプレートにおける置換語句を選択する手法として，トピックモデルのみを用いた手法，マルコフ連鎖のみを用いた手法およびトピックモデルとマルコフ連鎖を組み合わせた手法についてそれぞれ述べる．トピックモデルを用いる手法の場合には，まず，ユーザの入力文を単語に分割し，上記で学習した LDA を用いて入力文のトピック分布 $\bar{\theta}$ を求める．

トピックモデルのみを用いた手法では，前後の連接関係を無視して，置換語句 x の確率分布を以下のように展開する．

$$p(x|\bar{\theta}) = \sum_{k=1}^{K} p(x|z=k)p(z=k|\bar{\theta})$$

ここで， $p(z=k|\bar{\theta})$ は $\bar{\theta}$ をパラメータとする離散分布であ

るから，$p(z=k|\bar{\theta})=\bar{\theta}_k$ である．また，$p(x|z=k)$ は，トピック k が単語 x を生成する確率であるから，コーパスで学習されたパラメータを用いて $p(x|z=k)=\phi_k$ である．

マルコフ連鎖のみを用いた手法では，コーパスから学習した 2 次のマルコフ連鎖を用いて，置換語句 x の前 2 単語 x_{-2}, x_{-1} および後 2 単語 x_{+1}, x_{+2} を用いて以下のように x のスコアを求める．

$$p(x|x_{-2}, x_{-1}, x_{+1}, x_{+2}) \propto p(x|x_{-2}, x_{-1})p(x_{+1}|x_{-1}, x)p(x_{+2}|x, x_{+1})$$

ここで，$p(x|x_{-2}, x_{-1})$, $p(x_{+1}|x_{-1}, x)$, $p(x_{+2}|x, x_{+1})$ はそれぞれコーパスから最尤推定した 2 次のマルコフモデルで推定される遷移確率である．

提案手法である，トピックモデルとマルコフ連鎖を組み合わせた手法では，置換語句 x 自体はトピックモデルから生成され，続く後 2 単語 x_{+1}, x_{+2} が 2 次のマルコフモデルによって生成されると仮定する．置換語句 x のスコアは以下のように求める．

$$p(x|x_{-1}, x_{+1}, x_{+2}) \propto \sum_{k=1}^{K} p(x|z=k)p(z=k|\bar{\theta})p(x_{+1}|x_{-1}, x)p(x_{+2}|x, x_{+1})$$

それぞれの手法で，全ての語彙についてスコアを求め，その中で最もスコアの大きい語彙を置換語句として選択する．

4. 実験

4.1 実験方法

N-gram のみを用いた文生成，トピックモデルのみを用いた文生成，提案手法による文生成をそれぞれ行い，実験結果に対してアンケート調査を実施することで，評価を行う．実験では 2013 年投稿において，リプライが連鎖しているようなツイートを会話として抽出した対話コーパス (約 150MB) から返答文生成を行う．また，Twitter 日本語対話文 800 対話に発話タイプアノテーションを施した，発話タイプ推定のための学習コーパスを用いて入力文及び出力文の発話タイプ推定を行う．入力文，入力文発話タイプ推定，出力文発話タイプ推定，発話テンプレート選択は共通とし，アンケート調査はクラウドソーシングサービス「Lancers」を通じて行う．アンケートは表 3 のように入力文と選択肢が与えられ，最も相応しいと思われる回答を選択する．返答文 A は提案手法によって生成された文，返答文 B はトピックモデルのみを用いて生成された文，返答文 C は N-gram のみを用いて生成された文である．返答文 D は回答者がどれも適切な文でないと判断した場合に選択する．入力文として，10 種類の文章を作成した．それぞれの入力文について，8 回ずつそれぞれ手法で返答文を生成し，80 種類のアンケートを作成した．各アンケートにつき，30 人の回答者に回答してもらった．

表 3 アンケート例

入力文	今日も暇だな
返答文 A	今日遊ぼうか
返答文 B	こと遊ぼうか
返答文 C	こんど遊ぼうか
返答文 D	どれも適切でない

4.2 実験結果

個々の生成結果を見ると，表 4 の生成例では入力はダイエットに関する文であり，返答用テンプレートは疑問形の一言となっている．それぞれの手法の生成文は，提案手法と N-gram による生成では「誰得？」，「何故？」といった前後の連接を考慮し，自然な疑問文が生成されている．一方でトピックモデルのみによる生成は「www？」となっており，長音記号もしくは笑いを意味する「w」の連続が選択され，疑問形の文としては何が疑問となっているのか不明な文となっている．表 5 の生成例では入力文は呟きや，ぼやきのような入力文である．トピックモデル，N-gram によって生成された文はそれぞれ，意味の通らない文となっている．提案手法のみ「俺だって」という同意を示す文を生成している．

次にアンケートの結果である図 6 を見ると，「どれも適切でない」が全体の約 84 %を占め，N-gram のみを用いた生成手法が約 8%，提案手法が約 5%，トピックモデルのみを用いた生成手法が約 2%という結果となった．まず，多くの回答が「どれも適切でない」を選択したことから，生成手法には多くの改善余地が残されている．「どれも適切でない」を除いた場合の割合としては，N-gram のみを用いた生成手法が約 54%，提案手法が約 31%，トピックモデルのみを用いた生成手法が約 15%という割合となる．回答者が入力文に対して適切な生成文であると回答した設問の内，半数が N-gram のみを用いた生成手法を支持したことから，前後の連接関係を考慮することが今回の生成条件に於いては重要な要素であったと考えられる．

表 4 各手法生成例 1

入力文	最近、ダイエット始めたんだ
選択テンプレート	[ ]？
提案手法生成文	誰得？
トピックモデル手法生成文	www？
N-gram 手法生成文	何故？

表 5 各手法生成例 2

入力文	毎度の如く月曜日がつらい
選択テンプレート	[ ] だって
提案手法生成文	俺だって
トピックモデル手法生成文	*だって
N-gram 手法生成文	東京だって

表 6 アンケート回答結果

手法	回答数（全 24000 回答）	全体から占める割合
提案手法	1173	0.0489
トピックモデル	555	0.0231
N-gram モデル	2024	0.0843
どれも適切でない	20248	0.843

4.3 実験考察

実験結果より，前後の連接関係考慮が生成結果に大きく貢献していたと考えられる．N-gram のみを用いた生成では前後 2 形態素の連接関係を考慮したために最も支持され，提案手法は前後 1 形態素のみの連接関係考慮となったために十分に文脈を考慮することが出来なかったと考えられる．連接関係を考慮しない場合には，トピック成分が最も大きな単語を選出していたために意味の通らない語が多く選出され，今回のような低い選択率となったと考えられる．しかし個別に実験結果を見ていくと，提案手法は話題を一致させつつ意味の通る文を生成できている場合もあることから，より多くの前後連接関係を考慮することで今回設定した実験方法での支持率は向上可能であると考えられる．また本実験は文全体におけるテンプレート文字量の割合が大きかったこともあり，テンプレート選択時点で不適な文とならないためには連接関係が非常に重要であったと考えられる．

5. おわりに

本研究はテンプレートを用いた発話生成手法に対して，Support Vector Machine (SVM) を用いた発話タイプ推定に基づくテンプレート選択，N-gram モデルによる単語連接関係考慮，トピックモデルを加えることで，文構造を維持しつつ話題考慮が可能であると考え，従来手法と提案手法の比較実験・評価による検討を行った．実験では，より多くの連接関係を考慮することが文生成の尤もらしさに対して，有効である可能性が示唆された．これは今回の用いた発話テンプレートは，空白部分が名詞のみ，かつ 1 語のみの置換であったことから，テンプレート選択時点で文における話題がほぼ固定されてしまったことが原因として考えられる．今後の展望としては，名詞に限定せず，動詞や形容詞といった品詞をテンプレート空白に置換し，より生成領域を大きくすることで結果にどのような変化が生じるかを検証すると共に，どの程度の連接関係を考慮することでより尤もらしい文生成が可能となるのかを検証していく．また，話題を考慮するための手法についても引き続き模索を続ける．

謝辞

本研究の一部は，JSPS 科研費 (課題番号 16H02904) の助成によって行われた．

参考文献

[1] 小磯 花絵，石本 祐一，菊池 英明，：大規模日常会話コーパスの構築に向けた取り組み:会話収録法を中心に，言語・音声理解と対話処理研究会，(2015)
[2] 柴田 雅博，冨浦 洋一，西口 友美：雑談自由対話を実現するための WWW 上の文章からの妥当な候補文選択手法，人工知能学会論文誌，(2009)
[3] 稲葉 通将，神園 彩香，高橋 健一，：Twitter を用いた非タスク指向型対話システムのための発話候補文獲得，人工知能学会論文誌，(2014)
[4] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo,：*Towards an open-domain conversational system fully based on natural language processing*, In Proc. COLING，(2014)
[5] 安藤 秀哲，高橋 勇，黒岩 丈介，小高 知宏，小倉 久和，：チャットにおける会話の特徴と会話エージェントの検討，福井大学工学部研究報告，(2002)
[6] Suykens, Johan AK, and Joos Vandewalle. ：*Least squares support vector machine classifiers*, Neural processing letters (1999)
[7] Furey, Terrence S., et al.：*Support vector machine classification and validation of cancer tissue samples using microarray expression data*, Bioinformatics (2000)
[8] Jurafsky, Dan, Elizabeth Shriberg, and Debra Biasca. *Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual*, Institute of Cognitive Science Technical Report (1997)
[9] Blei, D, Ng, A, Jordan, M, :*Latent dirichlet allocation*, Journal of Machine Learning Research，(2003)
[10] Mimno, David, Matt Hoffman, and David Blei.：*Sparse stochastic inference for latent Dirichlet allocation*, arXiv preprint arXiv:1206.6425 (2012)