1B3-1

Tweet の reply 対からの返答パターン獲得

Reply Pattern Extraction from Reply Tweets Pairs

横野 光 *1
Hikaru Yokono

*1株式会社 富士通研究所
Fujitsu Laboratories Ltd.

One of the important elements for a dialogue system is the knowledge of how to reply the input utterance. Our goal is to obtain patterns which have a certain amount of context information and which is usable for several sentences. In this paper, we try to extract patterns of reply from tweet reply pairs and discuss how we can obtain good patterns.

1. はじめに

対話システムにおいて必要とされる基本的な要素として，ユーザの発話に対して適切に返答するということが挙げられる．返答を生成するためには，どのような発話に対してどのような返答がなされるのかといったような知識が必要となり，それをどのように獲得するかが重要となる．実際の返答生成では，大量の発話対からなる学習データを用意し，入力された発話に対してどのような返答が生成されるかを直接学習するようなモデルを用いた手法や，データからや人手で返答のためのテンプレートを作成しそれを元に生成するといった手法がある．

事例ベースのように発話の内容とそれをどのように発話するかの情報が統合されている知識は，返答生成に比較的容易に用いることができるが，同じ内容であっても発話の仕方が異なれば違う知識として扱われるため，様々な状況に対応するためには大規模なデータが必要となる．これに対して，内容に該当する部分をスロットとしたテンプレートのような知識は，発話の仕方に焦点を当てたものと見なすことができ，返答の内容とは独立して用いることができるため，様々な状況で用いることが可能である．しかし，一方で，発話内容の選択と発話の仕方の選択という処理を行う必要があるため，返答生成の負荷が大きくなる．

本稿では，ユーザの発話に対して，返答の内容の選択は別処理として行い，その処理によって得られた結果を相手の発話に合わせた形でテンプレートを利用して出力するという返答生成システムの構築に向けて，どのような発話に対してどのように答えればよいかについての知識である返答パターンの獲得について述べる．事例として，Twitter*1 の reply 関係にある tweet 対を発話対と見なし，そこから獲得できた返答パターンについて有用か否か，また，有用なパターンの抽出で考慮すべき要素について議論する．

2. 返答パターン獲得

発話とそれに対する返答からなる発話対データから，どのような発話に対してどのような返答がなされるか，を表したパターン対の抽出を行う．ここでパターンとは直接係り受けの関係にある 2 つの節からなり，各節において名詞，動詞，形容詞に該当する箇所がそれぞれの品詞の語が入るスロットとなっているもののことを指す (例 “[名詞] を-[動詞] たい EOS”)．

テンプレートの獲得ではどのようなテンプレートが有用であるか，を考える必要がある．汎用的なテンプレートであれば少数のもので多くの発話に対応することが可能であるが，そのテンプレートをどのようなときに適用すべきかが問題となり，一方で具体的なテンプレートであれば適用すべき状況は限定されるが，その代わりに実際の運用には数多くのテンプレートが必要となり，それをどう獲得するかが問題となる．

本研究では発話全体の係り受け対ではなく，文末とそれに直接係る節の対と文末の節のみを対象とする．これは文末の節にはその文が質問であるか問いかけであるか，という情報が含まれているため，その節を返答パターンに含めることによって返答パターンがどのような状況で使われうるかを明示できると考えたからである．文末の節に複数の節が係っている場合，それぞれの係り元の節と文末の節の組を考える．

発話対データの各事例に対して，形態素解析，係り受け解析を行い，文末に関わる節を抽出し，名詞，動詞，形容詞のスロット化を行う．そして，発話文から得られた節の組と返答文から得られた節の文とのすべての組み合わせを生成し，返答パターンの候補とする．発話対データから得られた返答パターン候補のうち，品詞のスロットを 1 つも含まないもの，品詞のスロットを含むが，そのスロットに入る要素が 1 種類しかないパターン対は候補から除外する．返答パターン候補獲得の段階では “(ただいま, おかえり)“のような定型のやりとりも獲得できる．これらも対話システムにおいて有用な知識であるが，本研究ではある程度汎用的に使える返答パターンの獲得を目指しているため，そのような事例は対象としない．

対話におけるやりとりでは，相手の発話を受けてそれに関することを返答することが多い．たとえば，「京都行きたいなあ」という発話に対する「京都良いよね」という応答のように相手の発話中の要素をそのまま返答に利用するといったような場合である．そこで本スロット中に出現した要素が発話文と返答文で共通していたもののみに候補を限定し，共通した要素が出現しているスロットに ID を振り，同じ要素が出現しうることを表す．

提案手法の流れと，得られる返答パターン対の例を図 1 に示す *2．

これらの処理によって得られる返答パターンは，どのような

連絡先: 横野光, 株式会社富士通研究所，神奈川県川崎市中原区上小田中 4-1-1, yokono.hikaru@jp.fujitsu.com

∗1 https://twitter.com

∗2 “-” は係り関係を示す

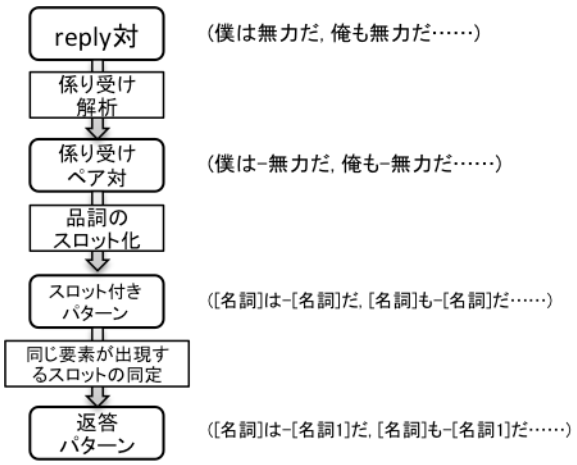

図 1: 返答パターン生成と獲得例

種類の発話に対して関連する話題をどのように返答に用いるか，を表現したものとみることができる．

3. Tweet の reply 対からのパターン獲得と分析

実際にどのようなパターンが獲得できるか，また，よいパターンを獲得するにはどうすればよいかを分析するために，発話対の事例からのパターン獲得を試みた．

分析に利用したデータは Twitter の reply 関係にある tweet 対である．reply 関係にある tweet は必ずしも対話であるというわけではなく，独り言に対しての反応なども含まれているが，相手の発話に対して反応を行うという点では共通であると見なし，tweet 対が実際に対話になっているかどうかを見ることはしない．以下，reply されている側を source，reply している側を reply と呼ぶ．

tweet 集合は Twitter Streaming API を用いて収集した 2016/8/1 から 2016/09/09 までの tweet から “in_reply_to_status_id” が空でない tweet を抽出し，その id を持つ tweet を Twitter API で収集することで構築した．得られた tweet 集合から bot による自動 tweet や解析が困難であると考えられるものなどを以下の条件によって除去し，残りをパターン抽出の元データとした．

- ユーザ名に “bot” を含まない
- URL，ハッシュタグ，記号，括弧，絵文字を含まない
- 1 文のみで構成されている *3

この条件を満たした tweet 対は約 995,000 件であった．

得られた source と reply のそれぞれの tweet に対し，形態素解析と係り受け解析を行い，文末の節とその節に直接係る節の組，文末の節単体に対して，要素中の品詞が名詞，動詞，形容詞である形態素の箇所を品詞のスロットに置き換える．スロットを含まないパターンは候補から削除する．最後にデータ中の tweet 集合を得られたパターンでまとめ上げ，その結果，品詞のスロットに入る事例が 1 種類であったパターンは候補から削除した．形態素解析は形態素辞書として mecab-ipadic-neologd*4 を用いた MeCab を，係り受け解析には CaboCha を利用した *5．

スロットを含まないパターンの多くは “おはよう”-“おはよう” のような挨拶や感動詞のやりとりであり，また品詞のスロットに入る事例が 1 種類のものは定型のやりとりである．これらは発話に対する返答の知識としては有用であるが，前述の通り，本研究ではある程度汎用的に使えるテンプレートを獲得することを目的としているため，これらは除外した．

はじめに，tweet 集合から得られた返答パターンの自己相互情報量 (PMI) を求め，それによってパターンをランク付けした上位の結果を表 1 に示す．

表 1: 得られた返答パターンの例

	source	reply
(1)	ほんとかよ [名詞 1]-EOS	ほんとほんと [名詞 1]-EOS
(2)	[動詞 1] へ-[名詞]EOS	[動詞 1] へ-[名詞]EOS
(3)	[名詞 1] で-[動詞 2] ないんだっけ EOS	[名詞 1] で-[動詞 2] ないんでしたと EOS
(4)	[名詞 1] が-[動詞 2] なかった EOS	[名詞 1] が-[動詞 2] なかった EOS
(5)	[名詞 1] ただいま-EOS	[名詞 1] おかえり-EOS
(6)	お [名詞] で-[動詞 1]EOS	[動詞] で-[動詞 1]EOS
(7)	[動詞 1] か-[動詞] てる EOS	[動詞 1] べき-EOS
(8)	[名詞 1] も-[名詞]EOS	[名詞 1] も-[動詞] き EOS
(9)	んね [名詞 1]-EOS	うんうん [名詞 1]-EOS
(10)	[動詞 1] よ [名詞]-EOS	[動詞 1] んかい [名詞]-EOS

獲得できたすべての返答パターンが有用であるわけではないが，獲得できたパターンの中には実際に利用できるようなものも含まれている．たとえば，(3) は内容の確認に対してそれをそのまま受けるというパターンである．挨拶のやりとりであっても，(5) は「マジただいま」「マジお帰り」のように相手の発話に合わせて挨拶を変えるという返答の仕方があることが知識として獲得できている．

一方で，どのように利用できるか分からない，或いは，有用とは思えないパターンも獲得されている．たとえば，(4) は相手の発話をそのまま返答しているパターンであるが，これがどのように使えるかについてはこのパターンだけからは不明である．また，Twitter という比較的砕けた表現を対象としているため，形態素解析誤りによって得られてしまったパターンも存在した．(2) の元の source 側の tweet は “ねるへぇす” であり，解析誤りによって得られたパターンである．

スロットに入る事例の種類が多いようなパターンは汎用的に用いることができると考えられる．そこで，source に対応するパターンの事例の経験分布のエントロピーによってランク付けを行った．その上位の結果を表 2 に示す．

表 1 と比べると，どのように用いればよいかがパターンからは読み取ることができないものが多く，スロットに入る事例の種類が多ければよいというわけではないということが分かる．

汎用的なパターンは少数で多くの事例をカバーしうるが，それが用いられる状況やスロットに入る要素間の関係がパターン

*3 “。”，“！”，“？” を文境界と見なしている

*4 https://github.com/neologd/mecab-ipadic-neologd

*5 http://taku910.github.io/{mecab, cabocha}/

表 2: エントロピーによるランク付け

	source	reply
(11)	[名詞 1]-EOS	[名詞 1]-EOS
(12)	[動詞 1]-[名詞]EOS	[動詞 1]-[名詞]EOS
(13)	[動詞]-[名詞 1]EOS	[動詞]-[名詞 1]EOS
(14)	[名詞 1]-EOS	[形容詞]-[名詞 1]EOS
(15)	[名詞 1]-EOS	[動詞]-[名詞 1]EOS
(16)	[名詞 1]-EOS	[動詞] た-[名詞 1]EOS
(17)	[動詞]-[名詞 1]EOS	[名詞 1]-EOS
(18)	[名詞 1]-EOS	[名詞 1]-[動詞]EOS
(19)	[名詞 1]-EOS	[動詞 1] た-EOS
(20)	[動詞]-[名詞 1]EOS	[形容詞]-[名詞 1]EOS

自体からは判別できないため，そのようなパターンからの返答生成は困難である．

実際に獲得されたパターンに対して，有用であると考えられるパターンを人手で選択した．その結果の一部を表 3 に示す．

表 3: 人手で選択した有用と考えられるパターン

	source	reply
(21)	[形容詞 1] ですね-[名詞]EOS	[形容詞 1] ですよね-[名詞]EOS
(22)	[形容詞 1] よね-EOS	[形容詞 1] です-[名詞]EOS
(23)	[名詞] が-[動詞 1] たい EOS	[名詞] も-[動詞 1] たい EOS
(24)	[名詞 1] かよ-[名詞]EOS	[名詞 1] だよ-[名詞]EOS

source と reply で同じ語を共有しているものに対象を制限しているため，得られるパターンでその用途が比較的明らかなものの多くは発話に対する同意を表すものであった．しかし，単なる同意であっても，相手の発話に合わせてどのように返答するかは異なるため，発話のスタイルに合わせたパターンが獲得できる可能性があることは有用であると考えられる．

本研究で考える有用なパターンの条件の一つとして，それ単体でその用法が推定できるということが挙げられる．文の用法はモダリティによって表現されることが多く [日本 03]，従ってモダリティに関わる表現を含むパターンは比較的有用であると考えることができるため，例えば，パターンのスコアを定義する際にどの程度スロット化されていない要素を含むか，を考慮することが有効であると考えられる．

また，スロットに対してどのような情報を付与するかという問題もある．本稿では，そのスロットに入りうる品詞と，source と reply でどのスロットが同じ語によって共有されるかをスロットに対する情報とした．しかし，前述の通り同意を表すようなパターンや，相手に合わせた形で挨拶を返すといったものが主に抽出され，他の用法のものは少なかった．

実際の対話では，必ずしも同じ語が返答に用いられるわけではなく，発話中の語と関係する語を用いることもある．例えば，“お願い” に対して “了解” といったやりとりに関する対応や，前の発話の関連語，或いは橋渡し照応などのように様々な関係があり，それらの関係を考慮してスロットに対して情報を付与する必要がある．

4. 関連研究

対話システムにおける発話生成の基本的なアプローチとして，テンプレートによるものと事例に基づくものとがある．発話テンプレートに関する研究として，塚原らは発話文中の固有表現に着目してそれが出現する発話から対話パターンを生成し，応答文生成に用いている [塚原 15b]．

用例ベースの発話生成では対応できる発話は利用する用例コーパスのサイズに影響されるが，Web コーパスのような大規模なコーパスが利用可能となり，これらを用いることで幅広く対応できる用例ベースの対話システムが提案されている．特に Twitter の tweet は非常に数が多く，話し言葉に近いテキストであるため，対話システムの知識として用いられることが多い．例えば，杉山らは係り受けのペアを単位とした話題についてどのように返答すべきかの知識を構築し，それを用いた返答生成を提案している [杉山 15]．木村らは発話された時期を考慮した返答選択の手法を提案している [木村 16]．

また，近年大規模なデータを獲得する手段としてクラウドソーシングによるデータ作成が注目されている．クラウドソーシングでは作業者は非専門家であることが多いため，複雑な作業を行うことは困難であるが比較的容易な作業を安価で行うことができる．クラウドソーシングを用いた対話に関するデータ構築としては，塚原らによる対話コーパスの構築 [塚原 15a] や Mitchell らによる言い換えパターン生成 [Mitchell 14] などがある．

他にも，用例ベースの対話システムに関する研究として，水上らはシステムの応答候補からユーザの選考に基づいて応答を決定する手法を提案している [水上 16]．

5. おわりに

本稿ではユーザ発話に対するシステムの返答を生成するためのパターン獲得について述べた．Twitter の reply 関係にある tweet から内容語に当たる箇所をスロットとしたパターンを抽出し，どのようなパターンが得られるかを調査し，有用なパターン抽出に向けて考慮すべきことについて議論した．

スロットに入る語の制約がほとんどなく，パターン自体を見てそれが使われる状況が想定できないようなパターンは，汎用的であって様々な表現を生成できうるとしても有用とはいえない．また逆に具体的すぎるパターンはそのパターンの用途は明らかであっても生成できる表現の数は少なく，実用においては大量のパターンを用意する必要がある．

そこで本稿では，発話とそれに対する返答においてどのスロットが同じ語を共有しているかという点に着目してパターンを抽出した．これによって相手の発話に対する同意を表す表現を獲得することができた．しかし，実際の対話では相手の発話に出てくる語を受けて話を続けることもある一方で，全く同じ語を使う以外にも対象の語の関連語によって受けるということもある．そのため，今後は発話と返答における語の共有に関して，同一語ではなく関連語まで広げた対応を考慮したパターン獲得を試みる予定である．また，獲得において単に関連しているかどうか，だけではなく関連している語の間の関係にはどのようなものがあり，より有用なパターンにするためにはどのような情報をスロットに付与すれば良いかについて取り組む予定である．

参考文献

[Mitchell 14] Mitchell, M., Bohus, D., and Kamar, E.: Crowdsourcing Language Generation Templates for Dialogue Systems, in *Proceedings of the INLG and SIGDIAL 2014 Joint Session*, pp. 16–24 (2014)

[水上 16] 水上 雅博, Neubig, G., 吉野 幸一郎, Sakti, S., 鈴木 優, 中村 哲：快適度推定に基づく用例ベース対話システム, 言語処理学会第 22 回年次大会, pp. 298–301 (2016)

[杉山 15] 杉山 弘晃, 目黒 豊美, 東中 竜一郎, 南 泰浩：任意の話題を持つユーザ発話に対する係り受けと用例を利用した応答文の生成, 人工知能学会論文誌, Vol. 30, No. 1, pp. 183–194 (2015)

[塚原 15a] 塚原 裕史, 内海 慶：オープンプラットフォームとクラウドソーシングを活用した対話コーパス構築方法, 言語処理学会 第 21 回年次大会, pp. 147–150 (2015)

[塚原 15b] 塚原 裕史, 内海 慶：単語と対話パターンの相関ネットワーク上のラベル伝搬による対話生成, 第 29 回人工知能学会全国大会 (2015)

[日本 03] 日本語記述文法研究会（編）：現代日本語文法 4 第 8 部 モダリティ, くろしお出版 (2003)

[木村 16] 木村 葵, 目良 和也, 黒澤 義明, 竹澤 寿幸：Twitter と word2vec を用いた時期に合った返答発話選択手法, 言語処理学会第 22 回年次大会, pp. 79–82 (2016)