2O5-OS-23b-2

発話状況を意識したオンライン上の対話における応答選択

Domain-aware response selection for online conversation

佐藤 翔悦 *1 Shoetsu Sato
石渡 祥之佑 *1 Shonosuke Ishiwatari
吉永 直樹 *2,*3 Naoki Yoshinaga
豊田 正史 *2 Masashi Toyoda
喜連川 優 *2,*4 Masaru Kitsuregawa

*1東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

*2東京大学 生産技術研究所
Institute of Industrial Science, The University of Tokyo

*3独立行政法人 情報通信研究機構
National Institute of Information and Communications Technology

*4国立情報学研究所
National Institute of Informatics

This paper proposes a domain-aware dialogue system that selects plausible responses for an input utterance from a given pool of response candidates. We employ Long Short-Term Memory-based Recurrent Neural Networks (LSTM-RNNs) to evaluate each response candidate as the suitability as response to the input utterance. In order to capture the diversity of domains (topics, wordings, writing style, etc.) in online conversation, we train multiple LSTM − RNNs from subsets of utterance-response pairs that are obtained by clustering of distributed representations of the utterances, and use the LSTM-RNN that is trained from the utterance-response cluster whose centroid is the closest to the input utterance. Experimental results on Twitter conversation datasets confirmed the effectiveness of multiple LSTM-RNNs.

1. はじめに

コンピュータによる対話システムは，自然言語を用いた情報機器へのアクセス手段，コールセンターなどにおける人間のオペレータの代替手段としての期待から研究が行われてきた．特に近年では，Apple の “Siri” や NTT ドコモの “しゃべってコンシェル” などといった音声対話によるモバイル機器の操作のための知的エージェントが広く普及しており，より洗練された対話システムの重要性が高まっている [Young 13]．

知的エージェントが普及するにつれ，エージェントの言語能力が使用者へ与える，快適さや生産性の向上についても注目されている [Bickmore 05]．そうしたエージェントは特定のタスクにおける補助としての対話機能を持つだけではなく，広範な話題についての雑談が可能であることが望ましい．しかしながら，従来，実用的な対話システムは特定のドメインに対してパターン等を作りこむ事で実現されており，現実には無数に存在する話題や話者の口調といった多様な発話状況（以下ドメインと呼ぶ）に対応した雑談が可能なエージェントの構築は非常に困難である．

そのような背景から，近年では Twitter などのソーシャルメディアから取得可能な幅広い話題を含む大規模な会話データを用いて，統計的手法による学習により，雑談対話を実現しようという研究が盛んである [Ritter 11]．統計的手法においても，ドメインを絞って訓練データを用意して高品質な対話システムを構築する手法 [Hasegawa 13] が研究されつつあるが，膨大なデータから適切なドメインを切り出す難しさから，十分にその有効性が検討されているとは言いがたい．

そこで本研究では，対話データを自動的にサブドメイン（クラスタ）に分割し，各クラスタごとに対話モデルの学習・訓練を行うことで，発話のドメインを意識した応答生成の有効性を探求する．ここで，サブドメインへの分割に教師なしクラスタリングを用いることで，多様なドメインを含む雑談対話から，サブドメインを暗黙に切り出す手法を検討する．また，対話モデルは主に Long-Short Term Memory-based Recurrent Neural Network (LSTM-RNN) [Hochreiter 97] による言語モデルを用いる．

実験では，モデルから生成された応答を定量的に自動評価することは難しいこと，さらに人手による評価はコストが高いことを考慮して，評価が容易でコストが低い応答選択タスクを実験に用いた．具体的には，Twitter から取得したツイートデータとそれに対するリプライデータをある発話に対する適切な応答とみなし，応答候補として与えられた他の候補リプライの中から正しい応答として実際に送られたリプライが選択できているかどうかを検証し，提案手法の有効性を確認した．

2. 関連研究

前節では，対話モデルにおいて，特定ドメインの訓練データを用いることによる，対話システムの性能の向上の可能性について述べた．しかし現実には，対象とするドメインの訓練データが十分に取得出来ない場合が存在する．そのため，ドメイン適応と呼ばれる，異なるドメインのコーパスから得られたモデルをタスクの対象となるドメインに適応させることを目的とした研究や，あるドメインにおけるデータの中でのサブドメインを形成することを目的とした研究が行われている．

Hasegawa らは，発話が聞き手に対し喚起させる感情の種類に着目し，人手で作成した少数の規則によって Twitter から取得した大規模な対話データを怒り，喜び，悲しみなどといった 9 つのカテゴリに分類した感情タグ付き対話コーパスを構築している [Hasegawa 13]．実験では，そのコーパスから対話モデルを学習することで特定の感情を喚起するような応答の生成を試みている．

Higashinaka らは連続した発話の中で，それ以前の文脈を考慮した上で現在の対話行為（例: 挨拶，質問_事実，情報提供など）の種類を決定すべく，高頻度語の bag-of-words 特徴量を用いた無限 HMM によって発話クラスタリングを行って

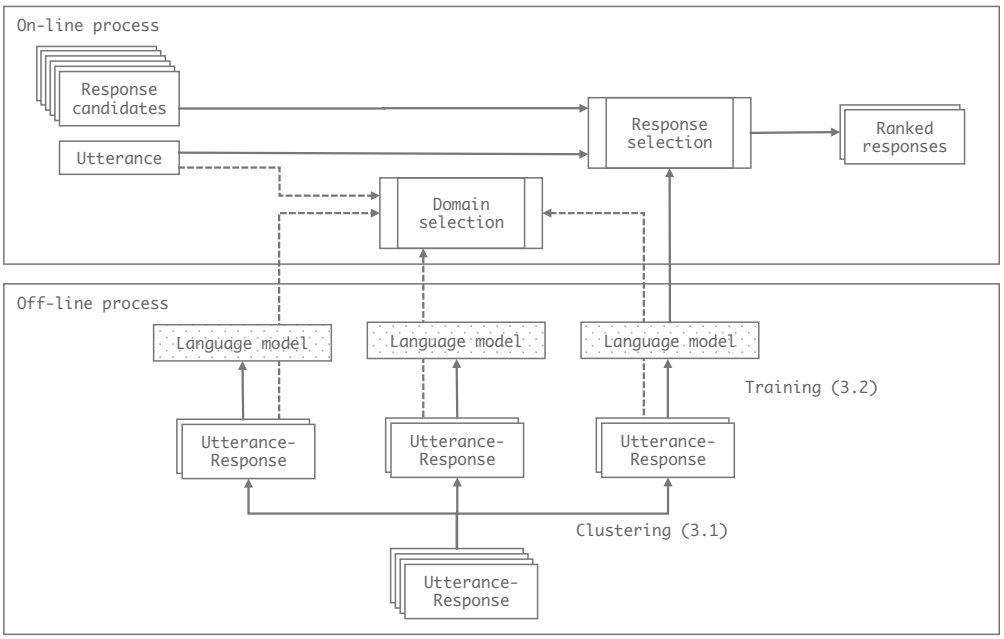

図 1: システムの全体図

いる [Higashinaka 11]．実験では，従来手法である Chinese Restaurant Process (CRP) に対して，系列的な情報を考慮することによってより高精度に発話の対話行為を認識することが出来ることを示した．

また，本研究と基本的な考えを共通するものとして，機械翻訳におけるドメイン適応に関する Yamamoto らの研究がある [Yamamoto 07]．彼らの手法では，各クラスタにおける言語モデルのパープレキシティを最小化するような形で文の組の再配置を繰り返すことで，詳細で一貫性のあるクラスタを得た後，それぞれのクラスタごとに翻訳モデルが構築している．実験では，日英翻訳において BLEU スコアの改善が得られている．

本節ではいくつかの関連研究について触れたが，対話において自動で発話文脈をクラスタリングして応答生成の精度向上の可能性を探求した研究は我々の知る限り数少ない．これらの理由から，我々は対話応答タスクを対象として，Mikolov らが提案した単語の分散表現 [Mikolov 13] に基づくクラスタリングを用いたドメイン適応の手法を提案した．3 節ではその内容について述べる．

3. 提案手法

本研究は，雑談対話（データ）をサブドメインに分割し，得られたサブドメインごとに複数の対話モデルを訓練することで，全データを用いた単一の対話モデルよりも良い性能が得られるのではないかという発想に基づいている．しかし当然ながらデータを分割すればするほど，一つのサブドメインあたりの訓練データは少なくなる．従って，ドメインを考慮することによる性能の向上は訓練データの減少による個々のモデルの性能の低下をどこまで担保出来るのかという点が今回の研究において注目する点である．

図 1 に，我々の提案する対話システムの全体図を示す．以降で，順に各コンポーネントについて述べる．

3.1 分散表現を利用した発話クラスタリング

オンライン対話では一つの発話が短いため，従来の bag-of-words 表現に基づき発話をクラスタリングするとデータの可塑性のため適切なクラスタリングが難しい．そこで，我々は Mikolov ら [Mikolov 13] による単語の分散表現（単語ベクトル）を用いて発話を密ベクトルで表現し，クラスタリングを行う．

クラスタリングの手順としてはまず，対話モデルの訓練データとして，我々が対象とする Twitter から取得した雑談対話（発話-応答ペア）の各発話を形態素解析器 MeCab を用いて各形態素に分割する．この際，今回対象となるデータには新語が多く含まれていると予想されるため，それらに対応した辞書である mecab-ipadic-neologd[*1] を用いた．

このようにして得たそれぞれの発話中の形態素（単語）について，（事前に構築した）単語ベクトルの平均を取ることで発話全体のベクトル表現を構築する．その上で，発話のベクトル表現に対して k-means クラスタリングを適用し，発話状況の共通する発話-応答ペアを収集した．

3.2 LSTM-RNN による応答選択モデル

第 3.1 節で述べた手法に基いて対話データから得られたサブドメインごとの対話データを訓練データとして，LSTM-RNN に基づく言語モデルを用いて応答選択を行う．具体的には，まずテストデータに含まれる各発話に対して，3.1 節と同様にベクトル表現を与える．次に，3.1 節で得られたサブドメインのクラスタのうち，中心がこの発話ベクトルから最も近いものを選択し，発話状況を推定する．最後に，選ばれたサブドメインのクラスタから作られた言語モデルを応答選択に用いる．

今回我々の応答モデルは長期依存が存在する時系列データに対して有効な LSTM-RNN [Hochreiter 97] を用いたシンプルな言語モデルによるものである．会話データにおけるある発話とそれに対する応答を，区切りとなるトークンを挟んで繋げた一文を 1 つの発話－応答として与え，そのパープレキシティによって応答を評価する．

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

4. 実験

対話システムが本来期待されている機能はユーザの発話に対し適切な応答を生成する事であると考えられる．しかし，生成された応答がどの程度適切であるかを客観的な基準に基いて自動評価するのは困難であり，また人手による評価はコストが高い．そのため，本研究では対話応答の中の部分的なタスクである，応答選択タスクによって評価実験を行う．

4.1 設定

実験にあたって，我々は二種類のデータセットを用いた．1つは NTCIR-12 Short Text Conversation タスク [Shang 16] において指定された 500,000 組のツイート・リプライ ID のうち，Twitter API を用いて取得できた 421,050 組からなるデータセットである（以降，NTCIR データセットと呼ぶ）．もう 1 つは 2011 年から 2013 年までに我々が独自に収集したおよそ 230,000,000 組のツイート・リプライからなるデータセットである（以降，UT データセットと呼ぶ）．

我々はまず，分散表現を用いた発話のクラスタリングによって，ツイートデータをそれぞれのサブドメインごとに分割するために，word2vec[*2] を用いた skip-gram による単語ベクトルを UT データセットから学習した．この際，単語ベクトルの次元数は 200，skip-gram における窓幅は 5 に設定している．

NTCIR データセットの中から選択した 100,000 組のツイート・リプライに対し，そのツイート部分に含まれるそれぞれの単語のベクトルを平均する事で，そのツイート全体を表すベクトルとする．このベクトルに対して scikit-learn[*3] で実装されている k-means クラスタリングを適用した．k-means のクラスタ数については 1 から 40 までの範囲を試した．その結果として得られた各クラスタのツイートと，それに伴うリプライを発話に対する適切な応答として考え，TensorFlow[*4] を用いて実装された LSTM-RNN の訓練に用いた．LSTM-RNN のハイパーパラメータについては事前に NTCIR データセットの中の一部を用いてチューニングを行った．

テストデータについては，NTCIR データセットから選んだ 1000 件のツイート・リプライの組を発話に対する正しい応答の組として選択し，その 1000 件のそれぞれに対して UT データセットからランダムに選んだ 19 応答を応答候補として加えた．そのようにして，問題となる 1000 件のツイートそれぞれに対する，正解のリプライを含んだ 20 応答を得た．

実験では，前述の訓練データを用いてクラスタごとに訓練した LSTM-RNN を用いて，その応答候補の応答としての妥当さを順位付けする．1000 件の問題に対し，システムが順位付けした応答の上位 3 位の中に正解の応答が含まれている比率をシステム全体の評価指標 (top-3 精度) として用い，ドメインを考慮した LSTM-RNN の有効性と，クラスタ数の影響について考察する．

4.2 結果と考察

今回の実験では提案手法のクラスタ数 k に対し，全データを単一のモデルの上で訓練したベースライン ($k = 1$) と，k を 10, 20, 40 に設定したものとの比較を行った（表 1）．以降，$k = 1$ のものを single モデル，k=10, 20, 40 のものを k-cluster モデルと呼ぶ．実験結果では，いずれの場合も k-cluster モデルの精度がベースラインとなる single モデルを上回っている．

その中でも，最も良い結果が得られた 20-cluster モデルについてさらに詳細な分析を行う．表 2 に 20-cluster モデルの各クラスタごとの single モデルとの比較結果を示す．「サブドメイン」は，我々が各クラスタに含まれるツイートの内容から判断し，手動でラベル付けしたものである．「要素数」はそれぞれのクラスタで訓練・テストに用いられた発話-応答ペアの数であり，「正解数」では同じ 1000 問のテストケースに対して，ベースラインとなる single モデルと 20-cluster モデルの正解数を比較している．

手法	精度 (top-3)
ランダム選択	15.0%
ベースライン ($k = 1$)	30.8%
提案手法 ($k = 10$)	33.2%
提案手法 ($k = 20$)	35.4%
提案手法 ($k = 40$)	35.0%

表 1: 実験結果: 応答候補の精度 (top-3) (k はクラスタ数)

まず，各クラスタのサイズと精度の向上率に注目すると，比較的小さなクラスタ (〜5000) については軒並み精度が上昇しており，件数で比較した全体の精度の上昇への寄与も主に小さなクラスタによるものである．また，比較的大きなクラスタ (5000〜) だけに注目した全体の精度の変化は合計で-0.38%である．我々が訓練データの量とモデルの精度の関係について検証するために行った予備実験によると，同じテストデータに対して，訓練データ数 10,000 件の single モデルの精度は 26.1%であった．訓練データ 100,000 件の single モデルと比較するとその精度は 4.7%低下している．それに対し single モデルと比較して，k-cluster モデルの大きなクラスタでは 1 つのモデルあたりの訓練データのサイズが 10%〜20% になっている事を考えると，おおむねサブドメインへの分割による効果はデータの減少を十分に補償している．

これらの結果から，サブドメインへの分割によって自身のドメインと関連性の薄い応答が学習データから削減され，それぞれのモデルの詳密化によってより良い学習が行われた，と説明できる．例えば，single モデルにおける応答選択の失敗例としては「おはよう」などのような高頻度な応答が過剰に学習された結果，不適切な状況でもそれらの応答を選択してしまうというものであった．しかし提案手法によって，「おはよう」といった高頻度な発話の一部はサブドメインとして別のクラスタに分割される．その結果，それ以外のクラスタにおける「おはよう」という応答の頻度は下がり，頻出する応答を過剰に選択してしまう問題が緩和されたと考える．一方，小さなクラスタを構成するサブドメイン，誕生日のお祝いや起床・就寝、フォロー・RT に関する挨拶，などはある種典型的な応答が存在するような発話であり，そうした発話に対する典型的な応答がより適切に学習できている，と解釈できる．

また，k-cluster モデルにする事で応答選択に失敗するようになってしまうケースとして，大きな原因として考えられるのが single モデルと比較した場合の未知語の増加である．つまり，訓練データ全体には登場するが，テストケースの発話の分類先クラスタの訓練データには存在しない場合は，本来 single モデルでは既知語であるはずの単語が，k-cluster モデルでは未知語として判定される事になり，結果は single モデルに大きく劣ったものとなってしまう．しかし，それぞれのモデルの小規模化に伴う語彙の減少に関する問題は，訓練データ全体の規模が大きくなるにつれ軽減され，サブドメインへ分割する事の効果の方がより強く現れると考えている．

*2 https://code.google.com/archive/p/word2vec/
*3 http://scikit-learn.org/stable/
*4 https://www.tensorflow.org/

ID	サブドメイン (話題，単語，口調)	#要素数 train	test	#正解数 提案手法	ベースライン	精度向上 $\frac{\Delta\#正解数}{\#要素数\ (test)}$
13	-	11801	108	**38**	25	12.04%
7	-	11524	124	37	**38**	-0.81%
14	政治，経済，社会問題	10294	130	**48**	31	13.08%
3	-	9743	94	**32**	31	1.06%
16	アニメ・漫画	6747	56	**11**	9	3.57%
12	-	6552	66	24	**30**	-9.09%
19	ゲーム	5677	50	**13**	11	4.00%
10	-	5627	45	14	**40**	-57.78%
1	'!' '?' を含む	5190	63	17	**19**	-3.17%
0	眠い，辛い，愚痴	5064	52	17	**20**	-5.77%
15	-	4908	50	**22**	16	12.00%
17	数字を含む	3803	31	5	**6**	-3.23%
6	飲食	2630	16	**6**	4	12.50%
2	フォロー, RT ありがとう (フランク)	2252	33	**29**	11	54.55%
18	'!!!' を含む	1869	17	**8**	4	23.53%
8	フォロー, RT ありがとう (丁寧)	1553	13	**12**	4	61.54%
4	挨拶	1537	21	**7**	2	23.81%
9	'...' を含む	1326	12	**3**	**3**	0.00%
5	おはようございます	1174	13	**9**	3	46.15%
11	叫び、連呼	729	6	**2**	1	16.67%
合計		100000	1000	354	308	4.60%

表 2: 提案手法 ($k = 20$) のそれぞれのクラスタにおけるベースライン ($k = 1$) の結果との比較: #正解数は top-3 の応答候補に正解の応答が含まれていた問題の数を指す.

5. おわりに

本稿では，多様なドメインを持つ対話データに対して，単語の分散表現を元に構成した，発話のベクトル表現を利用することで，人手による明示的なドメインの指定を必要とせずに対話データをサブドメインに分割し，それぞれの発話のドメインを考慮した応答選択をする LSTM-RNN 対話モデルを提案した．その結果，対話データの分割によってそれぞれの対話モデル辺りのデータ量は小規模になるにも関わらず，クラスタによっては単一のモデルで学習した場合を上回る結果が得られ，全体としての応答選択の精度も向上した．

本稿で提案したモデルでは各クラスタを排他的に扱ったが，近接するクラスタ，もしくは階層的なクラスタリングを行った場合における上位・下位クラスタのモデルとのスムージングといったことを行うことによって，ある発話のドメインを複数の観点から捉えるような働きが期待できると考える．今後は本稿で行った分析による結果を元に，こうした可能性について更なる検討を行う予定である．

謝辞

本研究の一部は JSPS 科研費 25280111 の助成を受けたものです.

参考文献

[Bickmore 05] Bickmore, T. W. and Picard, R. W.: Establishing and maintaining long-term human-computer relationships, *ACM Transactions on Computer-Human Interaction (TOCHI)*, Vol. 12, No. 2, pp. 293–327 (2005)

[Hasegawa 13] Hasegawa, T., Kaji, N., Yoshinaga, N., and Toyoda, M.: Predicting and Eliciting Addressee's Emotion in Online Dialogue., in *Proceedings of ACL*, pp. 964–972 (2013)

[Higashinaka 11] Higashinaka, R., Kawamae, N., Sadamitsu, K., Minami, Y., Meguro, T., Dohsaka, K., and Inagaki, H.: Unsupervised Clustering of Utterances Using Non-Parametric Bayesian Methods., in *Proceedings of INTERSPEECH*, pp. 2081–2084 (2011)

[Hochreiter 97] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural computation*, Vol. 9, No. 8, pp. 1735–1780 (1997)

[Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in *Advances in NIPS*, pp. 3111–3119 (2013)

[Ritter 11] Ritter, A., Cherry, C., and Dolan, W. B.: Data-driven response generation in social media, in *Proceedings of EMNLP*, pp. 583–593 (2011)

[Shang 16] Shang, L., Sakai, T., Lu, Z., Li, H., Higashinaka, R., and Miyao, Y.: Overview of the NTCIR-12 Short Text Conversation Task, in *Proceedings of NTCIR-12* (2016)

[Yamamoto 07] Yamamoto, H. and Sumita, E.: Bilingual Cluster Based Models for Statistical Machine Translation, in *Proceedings of EMNLP-CoNLL*, pp. 514–523 (2007)

[Young 13] Young, S., Gasic, M., Thomson, B., and Williams, J. D.: POMDP-based statistical spoken dialog systems: A review, *Proceedings of IEEE*, Vol. 101, No. 5, pp. 1160–1179 (2013)