Vol.2017-SLP-117 No.5
2017/7/28

CNNを用いたレファレンス・インタビューにおける対話行為の推定

河野　誠也[1,a)]　吉野　幸一郎[1,2,b)]　鈴木　優[1,c)]　中村　哲[1,d)]

概要：対話システムによってユーザの情報検索を補助する場合，相手の意図や知識を適切に推定し，その意図や知識に応じて適切な補助を行うために，対話行為を用いた対話管理が重要となる．本稿では，対話システムにおけるユーザの検索意図明確化のための応答戦略を，図書館における書籍検索補助サービスの一部であるレファレンス・インタビューにおける図書館員の振る舞いからモデル化することを志向して，レファレンス・インタビューで行われる対話行為についての検討，及び対話行為自動推定法の検証を行った．具体的には，Byte Pair Encoding に基づいた Tokenizer により，発話を語彙数固定制約のもとでサブワード単位の系列に分割し，Convolutional Neural Network を用いた分類器の素性として用いることで，単語単位で分割した場合と比較した．この結果，提案法では使用する語彙数を抑えつつ対話行為推定の精度が改善されることが確認できた．

キーワード：レファレンス・インタビュー，対話行為自動推定

1. はじめに

情報検索をタスクとする対話システムにおいて，ユーザの質問はしばしば曖昧であり，その質問のみでは応答することが難しい場合がある．このとき，システム側から何らかの聞き返しや情報提示を行い，ユーザの検索意図の明確化を行うことが重要となる [1].

人対人で情報検索をタスクとする対話が行われる場面として，図書館におけるレファレンス・サービスがある．レファレンス・サービスとは，図書館利用者が学習・研究・調査を目的として必要な情報・資料などを求めた際に，図書館員が情報そのもの，あるいはそのために必要とされる資料を検索・提供・回答することによってこれを助ける図書館サービスである [2]．レファレンス・サービスにおいては，図書館利用者の曖昧な情報要求を明確化することを目的としてレファレンス・インタビューというインタビューが行われる．また，このレファレンス・インタビューを事前に行った場合，行わなかった場合と比較して簡易な事実調査に関するレファレンスサービスの回答の正確さが大幅に向上することが知られている [3]. こうした機能を対話システムに導入することができれば，ユーザが必要とする情報にたどり着く可能性を高めることができると期待できる．

そこで本稿では，対話システムにおけるユーザの検索意図明確化のための応答戦略を，レファレンス・インタビューにおける図書館員の振る舞いからモデル化することを志向して，レファレンス・インタビューで行われる対話行為について，Inoue によるタグセット [4] を基に検討を行った．また，この対話行為タグの推定について，Convolutional Neural Network を用いた分類器を構築し，どのような素性が有効かについての検証を行った．その結果，Byte Pair Encoding に基づいた Tokenizer により，発話を語彙数固定制約のもとでサブワード単位の系列に分割し機械学習の素性として用いることで，単語単位で分割した場合と比較して，使用する語彙数を抑えつつ対話行為推定の精度を改善できることを確認した．

2. 図書館におけるレファレンス・インタビュー

レファレンス・サービスにおける図書館員は，レファレンス・インタビューを通じて，利用者の情報要求を明確化したうえで回答を提供する．その際，インタビューによって質問の主題を確定するとともに，要求が生じた状況や背景，情報利用の目的や動機などを利用者が語る内容から明らか

[1] 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
[2] 科学技術振興機構
Japan Science and Technology Agency
[a)] kawano.seiya.kj0@is.naist.jp
[b)] koichiro@is.naist.jp
[c)] ysuzuki@is.naist.jp
[d)] s-nakamura@is.naist.jp

表 1　チャットレファレンス・インタビューの例

No.	Speaker	Utterance
1	Patron	here is a current in a metal wire due to the motion of electrons. sketch a possible path for the motion of a single electron in this wire, the direction of the electric field vector, and the direction of conventional current.
2	Patron	you can just describe what they would look like
3	Librarian	Just a moment, please....
4	Patron	Thanks
5	Librarian	Is this for a school assignment and if so what is your grade level?
6	Patron	I’m a junior in high school... it’s for a physics class... i have a test tomorrow and this stuff and I’m still shakey on it
7	Librarian	What part of your physics books this question comes from: electricity?
8	Patron	ya
9	Librarian	Let me check
10	Librarian	Hold on please
11	Patron	ok
12	Librarian	I am still checking
13	Librarian	Hold on please
14	Librarian	http://www.swansontec.com/set.htm
15	Librarian	The source that I just sent has good graphics that shows the electic currents
16	Librarian	And the graphic is animated so you can see the movement
17	Librarian	Can you see the page?
18	Patron	yes
19	Librarian	Let me check for more hold on please...

にする．また，利用者が対話以前から明確に意識していることだけでなく，図書館員のインタビューによってはじめて意識し思い起こすようなことや，図書館員に問われその時点で考えたことも，適切な文献を推薦する上で重要な手がかりとなる．こうした情報を総合して，ユーザが真に必要とする文献を見つけるためのインタビューを行う [5]．

Radford らは，Online Computer Libaray Center，Inc.（OCLC）が提供するチャットレファレンス・サービス QuestionPoint に記録されたレファレンス・インタビューのチャットログをランダムにサンプリングし，サービスの利用者と被利用者（＝図書館員）を対象とした分析を行った．その結果，チャットレファレンスでは相手との共感を生み出すことが重要であることを指摘しており，そのためには図書館員が自分自身についての情報を提供し，安心を与え，ユーモアのある形式張らない言葉遣いをすることが勧められるとしている．また，レファレンス・サービスにおいては，正しい回答を示すだけでなく，図書館が対話をする能力を発揮することで，利用者のニーズを理解しようとし，信頼関係を築き，利用者に安心感を与えることが重要であり，それが「名前も顔もない白い検索ボックス」と図書館員を分けるものであると述べている [6]．

2.1　レファレンス・インタビューコーパス

本稿では，Radford らが用いた QuestionPoint におけるチャットデータをコーパスとして用いる．本データは，800 セッション，約 12,634 発話から構成される．また，収録されているすべての発話について，参与者，日時（分単位）に関する情報が付与されており，個人情報（利用者の名前やメールアドレスなど）に関わる部分については，匿名化処理が施されている．

レファレンス・インタビューにおける実際の対話例を表 1 に示す．表 1 の例では，利用者の質問に対して，単に回答を提示するのではなく，図書館員がインタビューを通じて，次の日に学校で試験があることなどの背景情報や，利用者の学年，検索履歴などの事項を明らかにしていることがわかる．

2.2　レファレンス・インタビューにおける対話行為

人間同士が行っているような自然な対話を行う対話システムを構築する上で，システムの全ての振る舞いをルールで網羅的に記述する方法は非現実的である．そこで，強化学習などの枠組みを用いてシステムの振る舞いを対話コーパスから統計的に学習する試みがなされている [7]．しかしながら，このようなデータ駆動型のアプローチを用いてモデル化を行う場合，学習に用いることができる対話データの量は限られるため，入力シンボルとなる発話をある程度汎化した状態で，記述し取り扱うことが必要となる [1], [8]．そこで，本稿では発話を汎化し記述するために対話行為タグに着目する．対話行為とは，対話において参与者がどのような意図を持って発話を行ったかの種類を判断するために用いられるタグとして定義される [9]．

レファレンス・インタビューに特化した対話行為タグと

して，Inoue によるタグセットがある [4]．Inoue は，レファレンス・インタビューにおける参与者の行動は，円滑なコミュニケーションの促進と維持のために，1）情報の伝達，2）情報探索活動におけるタスク管理，3）挨拶や感謝などによる社会的関係の維持，4）通信の物理的側面の管理といった 4 つの種類に分類可能であるとしている．具体的には，発話を機能（Dialogue Act Function; DAF）とドメイン（Dialogue Act Domain; DAD）の 2 つの意味的レベルで取り扱い，対話行為タグの定義を行っている（表 2，3）．

表 2　Dialogue Act Function（5 クラス）

No	Dialogue Act Function	Description
1	Information Provision	To provide information
2	Information Request	To request information
3	Task Management	To assign or commit to tasks
4	Social Relationship Management	To manage socio-emotional aspects of communication
5	Communication Management	To manage physical aspects of communication

表 3　Dialogue Act Domain（19 クラス）

No.	Function	Domain
1	Information	Information Problem
2		Search Process
3		Information Object
4		Feedback
5		Other
6	Task Management	Librarian’s Task
7		User’s Task
8		Other
9	Social Relationship Management	Greeting
10		Valediction
11		Exclamation
12		Apology
13		Gratitude
14		Downplay
15		Closing Ritual
16		Rapport Building
17	Communication Management	Channel Checking
19		Pausing
20		Feedback

3. レファレンス・インタビューにおける対話行為自動推定法の改善

本章では，Inoue によるタグセットを用いてレファレンス・インタビューにおける対話行為の自動推定を行った．推定には，Convolutional Neural Network による対話行為推定器を構築し，Byte Pair Encoding により，発話をサブワード単位の系列に分割したものを素性として用いた．

3.1 Byte Pair Encoding

Byte Pair Encoding（以下，BPE）では，出現頻度が高い文字 bigram を使われていない符号に置き換えていくことを繰り返していくことでテキストを圧縮する．Sennrich らは，ニューラル機械翻訳のタスクにおいて，BPE をテキストの単語分割に応用している [10]．具体的には，制約語彙数を目的関数として，テキストを圧縮し，最終的に得られた符号に対応する文字列を利用している．これにより，低頻度語や未知語をサブワードの系列として取り扱うことで，語彙数を抑えながら OOV が削減でき，機械翻訳の評価尺度である BLEU が向上したことを報告している．本稿では，テキスト分類のタスクである対話行為自動推定のタスクにおいても有効であると仮定して，これを素性抽出に用いた．Tokenizer には，BPE の拡張実装である SentencePiece[*1] を用いた．

3.2 Convolutional Neural Network

本稿では，対話行為推定器に Convolutional Neural Network（以下，CNN）を用いた．具体的には，一層の畳み込みと Global Max Pooling，512 × 4 の中間層，出力層で構成されるシンプルなモデルを用いた．畳み込みカーネルの数は 512，フィルターサイズは 3，ストライド幅は 1 を設定し，各層において Dropout（比率 0.5）と Batch Normarizataion を適用した．また，出力層の次元は，対話行為タグのクラス数であり，推定には Softmax 関数を用いた．単語埋め込みベクトルの初期値については，ランダムに設定を行い，次元数は 100 を設定した．

3.3 評価条件

Radford らによるチャットコーパスのうち，Inoue によりアノテーションされた 200 セッション，5,327 件の発話を用いて，10 分割交差検証により評価を行った．SentencePiece によるサブワード分割モデルの学習については，コーパスのうち対話行為タグがアノテーションされていない 600 セッション，7,307 件の発話を用いて学習を行った．また，比較のため，発話を単語単位での分割を行い素性として学習した CNN，従来素性を用いて学習した Multi Layer Perceptron（以下，MLP），Random Forests Classifier（以下，RF）を用意した．従来素性については，先行研究 [11], [12] を参考に以下を用いた．また，すべて実験において，乱数シード値は固定して行った．

- Bag-of-words(BoW): 各要素は単語の出現頻度
- Bag-of-bigram(BoB): 各要素は単語 bigram の出現頻度
- Speaker type: 参与者タイプ
- Text segmentation length: 発話を構成する単語の数
- Message position: 対話における発話の出現位置

[*1] http://github.com/google/sentencepiece

hsize

表 4　各手法における対話行為推定結果の比較

Methods	語彙数	DAD	DAF	平均 OOV	最大単語長	平均単語長	平均発話長
Subword-Level CNN	97	0.8601 *	0.7175	0.0	4	1.5	60.6
Subword-Level CNN	197	**0.8684 *** **	**0.7256 ***	0.0	10	2.5	46.3
Subword-Level CNN	295	0.8622 ***	0.7209 *	0.0	10	2.8	41.1
Subword-Level CNN	395	0.8620 **	0.7130	0.0	12	3.2	37.8
Subword-Level CNN	494	0.8570 **	0.7122	0.0	13	3.4	35.8
Subword-Level CNN	592	0.8585 **	0.7141	0.0	13	3.5	34.3
Subword-Level CNN	686	0.8585	0.7092	0.2	13	3.7	33.2
Subword-Level CNN	784	0.8556	0.7091	0.4	13	3.8	32.3
Subword-Level CNN	881	0.8536	0.7040	0.7	13	3.8	31.5
Subword-Level CNN	977	0.8541	0.7046	1.6	13	4.0	30.9
Word-Level CNN	6091	0.8438	0.6937	333.9	80	6.9	16.5
MLP（BoW + BoB）	6091	0.8498	0.7119	333.9	80	6.9	16.5
MLP（全素性）	6091	0.8515	0.7145	333.9	80	6.9	16.5
RF（BoW + BoB）	6091	0.8292	0.6790	333.9	80	6.9	16.5
RF（全素性）	6091	0.8367	0.7008	333.9	80	6.9	16.5

MLP（全素性）との paired t-test の結果　* 5%水準で有意
** 1%水準で有意
*** 0.1%水準で有意

3.4　評価結果

レファレンス・インタビューにおける対話行為を自動推定した結果を表 4 に示す．表 4 では，100，200，300，400，500，600，700，800，900，1000 の語彙数固定制約のもとで発話をサブワードの系列に分割した場合（以下，Subword-Level CNN）と，単語単位の系列に分割した場合（以下，Word-Level CNN），その他従来手法との比較を行っている．評価の結果，Word-Level CNN（語彙数 6091）と比較して，Subword-Level CNN（語彙数 100 から 1000）を用いた場合のほうが，DAF，DAD のいずれにおいても，精度が高くなることが確認できた．また，Word-Level CNN が話者情報や発話の対話における出現位置といった情報を素性として用いた従来手法の精度より低い結果となった．これに対して，Subword-Level CNN では，DAF において，語彙数を 100，200，300，400，500 とした場合で，有意（$p < 0.05$）に精度が改善されていることが確認できた．さらに，DAD においては，語彙数を 200，300 とした場合において，有意に精度が改善されていることが確認できた．また，単語単位の素性を用いたモデルでは，10 分割交差検証において平均して OOV が 333 程度発生しているのに対して，サブワードを用いたモデルでは，ほとんど OOV が発生していないことが確認できた．

表 5　Dialogue Act Function （19 クラス）の推定結果

DA	Precision	Recall	f1-score	support
Info Provision	0.8883	0.9328	0.9100	2857
Info Request	0.8422	0.8799	0.8606	758
Social Rel Mgmt	0.8735	0.8316	0.8520	689
Comm Mgmt	0.8599	0.7454	0.7986	593
Task Mgmt	0.7644	0.6488	0.7019	430
avg/total	0.8667	0.8684	0.8663	5331

表 6　Dialogue Act Domain （19 クラス）の推定結果

DA	Precision	Recall	f1-score	support
Info:Probrem	0.6582	0.8354	0.7363	1203
Info:Seaech	0.6711	0.5982	0.6326	672
Info:Obj	0.8111	0.8302	0.8205	1184
Info:Feedback	0.2326	0.0901	0.1299	111
Info:Other	0.5688	0.4685	0.5138	393
Task:Librarian	0.5789	0.5238	0.5500	126
Task:User	0.4286	0.2500	0.3158	96
Task:Other	0.0000	0.0000	0.0000	6
Social:Greeting	0.8924	0.9069	0.8996	247
Social:Valediction	0.9444	0.7556	0.8395	45
Social:Exclamation	0.5385	0.3333	0.4118	21
Social:Apology	0.4000	0.0952	0.1538	21
Social:Gratitude	0.8909	0.9456	0.9174	423
Social:Downplay	0.9200	0.7077	0.8000	65
Social:Closing	0.3684	0.2188	0.2745	32
Socail:Rapport	0.4028	0.3537	0.3766	82
Comm:Channel	0.8857	0.4627	0.6078	67
Comm:Feedback	0.7161	0.7229	0.7195	314
Comm:Pausing	0.8981	0.8447	0.8706	219
avg/total	0.7154	0.7256	0.7145	5331

3.5　対話行為の検討

本節では，3.2 節でベストスコアを出した Subword-Level CNN が推定した対話行為についてのエラー分析を行う．各対話行為における推定結果を表 5，表 6 に示す．DAF については，Precision，Recall，F 値において高い値で対話行為を推定できた．一方で，DAD については，学習データの不足からか，各対話行為推定結果における F 値の加重平均は，0.72 程度と低い結果に留まった．特に，Info:Feedback，Task:User，Task:Other，Social:Exclamation，Social:Apology，Social:Clossing，Social:Rapport といった対話行為では、F 値が 0.5 を下回るなどさらなる手法の改善が求められる結果となった．これ

については，例えば，対話履歴の情報を素性として用いるなどして改善が可能であると考えられる．

また，DAD において，対話行為分類の精度が低かった原因としては，対話行為タグのアノテーション作業時における曖昧性も原因であると考えられる．以下の例に示すように，通常，ユーザや図書館員の発話は，複数の対話行為で構成される (カッコ内はアノテーション候補となりうる対話行為)．しかしながら，Inoue のタグセットでは，一つの発話に対して，DAF と DAD をそれぞれ一つアノテーションしているため，アノテーション作業者によってアノテーション結果に齟齬が生じている．このような対話行為タグのアノテーション時に生じる曖昧性については，作業者に対するトレーニングによる熟練度の向上はもちろんのこと，対話行為タグの定義の改良によって解消を行うことが必要となると考えられる．

- thank you so much. this looks great. can you find any reasons why tea would do this? (Social:Gratitude, Info:Probrem)
- ”Name”, welcome to maryland askusnow! i’m looking at your question right now; it will be just a moment. (Social:Greeting, Task:Librarian, Comm:Pausing)

このような問題を防ぐための方法としては，Bunt による対話行為タグの定義 ISO24617-2[13] で与えられるタグの付与方法が有効であると考える．ISO24617-2 では，特定の課題達成に向け対話内容を進行させる一般目的機能 (General-Purpose Function; GPF) に基づくタグと，発話順番の決定や対話を構造化するといった，相互に対話を通じてやり取りを行う上での調整に関する機能を扱う次元特有機能 (Domain-Specific Function; DSF) に基づくタグが区別して用意されている点で特徴的である．Inoue が定義したタグセットをこの ISO24617-2 の定義と照らし合わせ，DSF に相当するものは複数タグの付与を許すなどすることで，こうしたタグ付与上の問題を解決することができる．

4. まとめ

本稿では，レファレンス・インタビューにおける発話の対話行為を自動推定するために，CNN を用いたモデルを構築した．評価実験の結果，従来素性を用いたベースラインと比較して，高い精度で対話行為推定が可能であることを確認することができた．また，本研究で用いたモデルでは，CNN への入力として，発話を単語単位で分割した場合の系列と，サブワード単位で分割した場合の系列を素性として用いてその比較を行った．その結果，目標ドメインにおける対話コーパスで学習したサブワード分割器により，発話をサブワード単位の系列に分割したものを CNN の入力素性として用いることで，単語単位の系列を素性として用いた場合と比較して精度が改善されることが確認できた．一方で，本研究で構築したモデルでは，DAF ついては高い精度で対話行為を推定できたものの，DAD については，比較的低い推定精度に終わった．これはデータの不足と同時に，一貫したアノテーションが困難な場合がある現行アノテーションスキーマの問題でもあると考えられる．今後は，学習に用いるデータを増やすことや，対話履歴の情報を素性として用いるなどしてモデルの改善を行う必要があると考える．また，Inoue による定義を拡張し ISO24617-2 を参考としたアノテーションスキーマの検討を行って行きたい．本研究で取り扱ったレファレンス・インタビューでは，ユーザが求める情報を提供するために必要な情報を提供するために対話を行うという点でタスク志向型対話であるといえる．一方で，自然対話のような目標達成に直接関係がないような発話も多数行われており，非タスク志向型対話の側面もある．これらの点を考慮し，それぞれの観点から情報検索対話に必要となる機能を整理したアノテーションスキーマを検討していく必要がある．

謝辞

本研究は JST さきがけ（JPMJPR165B）および JST CREST（JPMJCR1513）の支援を受けた。

参考文献

[1] Yoshino, K. and Kawahara, T.: Conversational system for information navigation based on POMDP with user focus tracking, *Computer Speech & Language*, Vol. 34, No. 1, pp. 275–291 (2015).
[2] 日本図書館学会. 用語辞典編集委員会：図書館情報学用語辞典，丸善株式会社 (1997).
[3] Radford, M. L., Connaway, L. S., Confer, P. A., Sabolcsi-Boros, S. and Kwon, H.: ” Are we getting warmer?” Query clarification in live chat virtual reference, *Reference & User Services Quarterly*, pp. 259–279 (2011).
[4] Inoue, K.: An Investigation of Digital Reference Interviews: A Dialogue Act Approach, PhD Thesis, Syracuse University (2013).
[5] 泰則斎藤：レファレンス・インタビュ-における利用者モデル，*Library and information science*, No. 27, pp. p69–85（オンライン），入手先 ⟨http://ci.nii.ac.jp/naid/40000019897/en/⟩ (1989).
[6] Radford, M. L. and Connaway, L. S.: Seeking synchronicity: Evaluating virtual reference services from user, non-user, and librarian perspectives, *Proposal for a research project, submitted February*, Vol. 1, p. 2005 (2005).
[7] 目黒豊美，東中竜一郎， 南泰浩，堂坂浩二：POMDP を用いた聞き役対話システムの対話制御，言語処理学会第 17 回年次大会，pp. 912–915 (2011).
[8] 翠輝久，大竹清敬， 堀智織，柏岡秀紀， 中村哲ほか：京都観光案内対話コーパスにおける対話行為タグの設計と分析，情報処理学会研究報告音声言語情報処理 (SLP)，Vol. 2009, No. 10 (2009-SLP-075), pp. 39–44 (2009).
[9] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C. and Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech, *Computational linguistics*, Vol. 26, No. 3, pp. 339–373 (2000).

[10] Sennrich, R., Haddow, B. and Birch, A.: Neural machine translation of rare words with subword units, *arXiv preprint arXiv:1508.07909* (2015).
[11] : Yu Bei., Keisuke Inoue.: “An Investigation of Digital Reference Interviews: Dialogue Act Annotation with the Hidden Markov Support Vector Machine.”, OCLC/ALISE research grant report published electronically by OCLC Research, 2012.，PhD Thesis.
[12] Ivanovic, E.: Dialogue act tagging for instant messaging chat sessions, *Proceedings of the ACL Student Research Workshop*, Association for Computational Linguistics, pp. 79–84 (2005).
[13] Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., Popescu-Belis, A. and Traum, D. R.: ISO 24617-2: A semantically-based standard for dialogue annotation., *LREC*, pp. 430–437 (2012).