要約付き宿検索対話コーパス

林部 祐太
株式会社リクルート Megagon Labs, Tokyo, Japan
hayashibe@megagon.ai

1 はじめに

宿を探しているカスタマーへの接客では，オペレータは宿が満たすべき要件を整理して検索し，検索結果を基に適切な宿を提案する．カスタマーが最初から要件を列挙できることはほとんど無く，漠然とした希望しか持っていないこともしばしば有る．そのため，要件の整理にはカスタマーの発話の意図を的確に解釈するだけでなく，質問・補足・提案などしてカスタマーの希望を言語化する手助けが必要である．本研究ではその整理の過程を分析するため，宿を検索する際の要件を列挙した時点までの対話を収集し，次の 3 つのアノテーションを行う．

1 つ目は要約文で，省略されている言葉を補いながら，発話文を簡潔に要約する．例えば，「それがいいです」といった発話文に対して「朝食はバイキングを希望する」といったような要約文を作成する．これは，発話が何を意図しているのかを解釈するためである．

2 つ目は，発話文と要約文の語句の対応付けである．例えば，「それが」と「朝食はバイキングを」，「いいです」と「希望する」をそれぞれ対応付ける．これは，どのような言い換えや補足によって要約が生成されたのかを明らかにするためである．

3 つ目は，要約文とオペレータが列挙した要件の対応付けである．これは，どのようにして要約文が取捨選択され要件としてまとめられるのかを明らかにするためである．

我々は 210 対話を収集し，要約文 3,282 文と 2,134 個の要件，およびそれらの対応付けがアノテーションされているコーパスを構築した．このコーパスは談話理解，要約生成，対話応答生成など幅広い言語処理の研究に役立てられると考えている．

2 宿検索対話の収集

日本語テキスト対話による架空の宿予約サービスを想定し，カスタマーとオペレータを演じる 2 名による対話を，オンライン対話プラットフォーム Slack[1)] で収集した．カスタマーは宿泊先に関する希望をオペレータに伝え，オペレータはそれを具体的な要件まで掘り下げる．宿を大まかに絞り込むのに十分と思われる具体的な要件が挙げられるようになった時点で対話は終了とした．

2.1 参加者

参加者は全員，日本語母語話者かつ Slack 経験者とした．カスタマー役は 35 名に依頼した．オペレータ役は，観光業界での接客経験者と，未経験者の 2 名に依頼した．カスタマーとオペレータのペア 1 組がそれぞれ 6 対話行い，合計 210（= 35 × 6）対話を収集した．このうち，接客経験者がオペレータである対話は 126 対話である．

2.2 前提条件と対話内容

各対話には次の要素と制約からなる「前提条件」をランダムに与え，対話の冒頭でカスタマーとオペレータの双方に表示した．

- 対話が行われる日付：年は無く月と日のみ
- 予約する日付：対話が行われる日付から 3 か月以内の日程で，日付もしくは月の上旬，中旬，下旬のいずれか
- 予約する日数：1 日以上 4 日以下
- エリア：47 都道府県，東北地方，関西地方，四国地方，九州地方の 51 種のうちいずれか
- 人数：大人は 1 人以上，子ども（小学生以下）は 0 人以上で，合計 1 人以上 4 人以下

カスタマーへは，この前提条件にしたがってアドリブを交え，漠然とした希望をオペレータに伝えるよう指示した．オペレータへは，宿を大まかに絞り込むのに十分と思われる具体的な要件が挙げられるようになった時，対話を終了する合図を記入し，それら要件を箇条書きで列挙して対話を終了するよう

1) https://slack.com/

表 1　収集した対話の例．O はオペレータ，C はカスタマー，⏎は改行を示す．表 2 で例示に用いた文は**太字**で示す．

	日付：6 月 6 日⏎予約する日付：7 月 5 日から 4 泊⏎エリア：京都府⏎人数：大人 2 人, 子ども 2 人
O	この度はご利用いただきまして、ありがとうございます。⏎ご宿泊先につきまして、お客様のご希望をお聞かせいただけますでしょうか。
C	夫と幼稚園児の子供 2 名と京都に行きたいとおもっています。
O	さようでございますか。京都の行先は京都市内でしょうか？それとも、丹後地方など別の地域でしょうか？
	(略)
C	**朝食はバイキング希望です。夕食は外で食べる予定です。**
O	承知しました。それでは朝食のみ付いたプランをお調べします。⏎ほかにご希望はございますでしょうか？
C	コンビニが宿の近くにあるといいですね。
O	かしこまりました。それでは、コンビニが近くにあるお宿をお調べします。⏎ご宿泊のお日付と宿泊日数はいかがなさいますか？
C	7 月 5 日から 4 泊です。
O	かしこまりました。4 泊となりますと、観光以外にもお時間があるかと思います。⏎幼稚園児のお子様ですと、7 月は水遊びができると、お喜びいただけるかと思いますが、鴨川沿いのお宿はいかがでしょうか？
C	**それはとてもいいですね。**鴨川で水遊びできるんですね。**知らなかったです。**
	(略)
O	かしこまりました。お子様連れに優しいお宿をお調べします。⏎それでは、以下のご希望条件に合うお宿をお探しいたします。
O	・日付は 7 月 5 日。⏎・宿泊日数は 4 泊。⏎・宿泊地は京都市内。⏎・人数は大人 2 名、子供 2 名。⏎・宿タイプは旅館を希望。⏎・部屋タイプは洋室を希望。⏎・京文化を体験できる宿を希望。⏎**・食事は朝食のみで、バイキングを希望。**⏎・コンビニが近くにある宿を希望。⏎**・鴨川沿いの宿を希望。**⏎・お子様連れに優しい宿を希望。

指示した．また，次の補足事項も双方に案内した．

- 対話の中で宿泊先を決定する必要はない
- 前提条件に記載された情報も漠然としているため，対話の中でこれを掘り下げるのも可
- 前提条件の情報が，対話中に変更されても可
- 旅行プランの計画ではなく，あくまで宿に関する話題をメインとする

2.3　収集結果

1 対話におけるターン[2] は最小 11 回，最大 35 回，平均 19.0 回，合計 3,997 回だった．また，文境界をアノテーションし，1 対話あたりの文数を数えたところ，最小 31 文，最大 78 文，平均 51.5 文，合計 10,814 文だった．

対話の例を表 1 に示す[3]．7 月に子どもと京都旅行という希望から，オペレータは「水遊びができる鴨川沿いの宿」というカスタマーが当初考えていなかった要件を引き出している．そして，対話の最後では要件が箇条書きで 11 件列挙されている．

3　要約アノテーション

表 2　作成した要約文の例

(1) 発話文	朝食はバイキング希望です。
自動要約文	朝食が希望だ。
要約文	朝食はバイキング希望だ。
(2) 発話文	夕食は外で食べる予定です。
自動要約文	[著者] が外で [不特定:人] に夕食を食べる。
要約文	夕食は外で食べる予定だ。
(3) 発話文	それはとてもいいですね。
自動要約文	それが良い。
要約文	水遊びができる鴨川沿いの宿がよい。
(4) 発話文	知らなかったです。
自動要約文	[著者] が良いを知らない。
要約文	鴨川で水遊びできるのを [customer] が知らなかった。

3.1　要約文の作成

カスタマーの各発話に対して，意図の解釈結果として「要約文」をアノテーションした．表 2 に例を示す．

このアノテーションの前処理として，JUMAN++[4] [1] (Revision.1ee40d7) と KNP[5] (Revision.165d699a) [2] で発話文を述語項構造解析した．そして項構造から自然文へ変換する簡易的なルールを作り，各発話の「自動要約文」を作成した．この自動要約文は流暢でなかったり，意味的に誤っていたり，省略の補完が不完全だったりするので，人手で修正し，常体で要約文を作成した．なお，1 つの発話から複数の要

2)　異なる人が入力するまでの，ある人の一連の入力を 1 ターンと数える
3)　各発話の入力時刻も記録しているが例示では省略している
4)　https://github.com/ku-nlp/jumanpp
5)　https://github.com/ku-nlp/knp

表 3 要約文と発話文の対応付けの例．同じ番号を添えた下線で対応を示し，生成に発話文外の情報を必要とする箇所には ⋆ を付与している．

(1) 発話文	朝食は$_1$ バイキング$_2$ 希望です$_3$。
要約文	朝食は$_1$バイキング$_2$希望だ$_3$。
(2) 発話文	夕食は$_4$ 外で食べる$_5$ 予定です$_6$。
要約文	夕食は$_4$ 外で食べる$_5$ 予定だ$_6$。
(3) 発話文	それは$_{7\star}$ とても いいですね$_8$。
要約文	水遊びができる鴨川沿いの宿が$_{7\star}$ よい$_8$。
(4) 発話文	知らなかったです。$_9$
要約文	鴨川で水遊びできるのを$_\star$ [customer] が$_\star$ 知らなかった$_9$。

表 4 要約文と要件の対応付けの例

(1) 要約文	朝食はバイキング希望だ。
要件	食事は朝食のみで、バイキングを希望
(2) 要約文	夕食は外で食べる予定だ。
要件	食事は朝食のみで、バイキングを希望
(3) 要約文	水遊びができる鴨川沿いの宿がよい。
要件	鴨川沿いの宿を希望
(4) 要約文	鴨川で水遊びできるのを [customer] が知らなかった。
要件	-

約文が作られることもありうる．

表 2 において，(1), (2), (4) は，項構造解析が誤っていたため修正した事例である．(3) は，照応表現「それ」を適切な先行詞に修正した事例である．

3.2 要約文と発話文の対応付け

要約文と発話文の間で語句の対応付けを行った．表 3 に例を示す．対応付けの単位は要約文の述語と項を原則とした．例えば (3) では述語とその項の計 2 箇所それぞれに対応付けがなされている．発話文での「いいですね」が要約分では簡潔に「よい」となっていることが分かる．また要約文の「水遊びができる鴨川沿いの宿が」は発話文の「それは」に対応しているが，この生成には発話文外の情報を必要とする．そのような生成に文外情報を必要とする箇所には印を別途付けた．

3.3 要約文と要件の対応付け

オペレータの列挙した要件と，作成した要約文の対応付けを行った．表 4 に例を示す．多くは一対一対応しているが，(1) と (2) のように複数の要約文が同じ要件に対応していることがある．なお，(4) のようにどの要件にも対応していない要約文もあり得る．

表 5 発話文と完全一致しない要約文箇所の例（抜粋）

文外情報	頻度	発話文箇所	要約文箇所
必要	397	-	[customer] が
-	47	いいです	良い
必要	37	お願いします	希望する
-	34	いいですね	良い
-	18	助かります	良い
-	17	希望します	希望する
-	7	ありがたいです	良い
必要	4	はい	良い
必要	3	そうです	観光だ
必要	2	-	朝食は
-	1	母が足が	母の足が
-	1	重視することは見られることです	見たい

表 6 要件との対応が無いまたは複数有る要約文の例

要約文	対応する要件
(5) コネクティングルームが何か知りたい。	-
(6) 食べるのが大好きだ。	-
(7) 高校からの友人 4 人で九州へ旅行へ行きたい。	・人数は大人 4 名 ・宿泊地は長崎か熊本
(8) 露天風呂があって、早朝も入浴ができる旅館がよい。	・宿は旅館 ・露天風呂がある ・早朝に入浴ができる

4 コーパスの分析

4.1 要約文と発話文の対応の分析

発話文と要約文の対応付けは，75 対話 1,268 文の要約文に対して現在アノテーションを終えている[6)]．これら要約文の全 3,724 箇所のうち生成に文外情報を必要としない箇所は 2,812 箇所，必要とする箇所は 912 箇所だった．

前者の 2,812 箇所のうち，発話の文字列と完全一致する箇所は 1,111 箇所，しない箇所は 1,701 箇所であった．発話文箇所の異なり総数は 1,905 種類だったのが，対応する要約文箇所は 1,687 種類であった．完全一致はしない対応の例を表 5 に示す．「重視することは見られることです」という冗長な表現が「見たい」と簡潔に要約されていたり，「いいです」「いいですね」「助かります」など複数の表現が「良い」と要約されていることが分かる．

後者の 912 箇所のうち，694 箇所は発話文に対応する箇所が有った．例えば，ある「はい。」という発話の要約文は「温泉付きの宿で良い。」とアノテーションされているが，これはオペレータの「お宿の

6) 残りの要約文についても付与する予定である．

タイプは温泉付きの旅館でよろしいでしょうか？」という発話に対する返答で，文脈が分からなければこの要約は生成できない．このように「良い」という要約でも発話によっては文外情報が生成に必要な場合があった．

後者の 912 箇所のうち，218 箇所は発話文に対応する箇所が無かった．例えば，「お食事はいかがいたしましょうか？」というオペレータの質問に対するカスタマーの返答「夜はなくて大丈夫で、朝はあってもなくても大丈夫です。何か軽いもので問題ないです。」の第 2 文の要約文は「朝食は軽いものが良い。」とアノテーションされていて，「朝食は」に対応する箇所は発話文内には無い．この「朝食は」の生成には，文脈を理解した上で第 1 文の「朝」が「朝食」を意味しているという推論が必要である．

4.2 要約文と要件の対応の分析

210 対話中カスタマーの発話 3,030 文にアノテーションされている 3,282 の要約文のうち，2,634 文はオペレータが列挙した要件との対応が有った．一方，648 文は対応が無かった．これは，カスタマーの質問や感想であるためであったり，途中でその希望が破棄されたためであったりするなどの理由による．

要件との対応が無いまたは複数有る要約文の例を表 6 に示す．(5), (6) はカスタマーの質問や感想であったりして，要件とは関係ない要約文である．そのため対応する要件が無い．(7), (8) は対応する要件が複数有る要約文である．要約文が対応している要件の数は，619 文が 0 個，2,308 文が 1 個，301 文が 2 個，47 文が 3 個，7 文が 4 個だった．

5 関連研究

5.1 不動産検索対話コーパス

横野らは複数のアノテータが協力して擬似的に対話を収集する手法を提案し，物件を探す客と不動産業者の間での対話からなる，不動産検索対話コーパスを構築した [3, 4]．

福永らは沿線，収納，賃料，間取りといった 38 種類のデータベースフィールドタグを不動産情報サイト SUUMO で不動産を検索する際に指定可能な検索条件をもとに定義し，不動産検索対話コーパスの各発話に対しアノテーションした [5, 6]．彼らはどのデータベースフィールドにも明示的に言及していない「一人暮らしをしたい」といったような情報を「非明示条件」とよび，常識や経験的な知識により「一人暮らしならば一般的に物件の間取りは 1LDK 以下である」ため「間取り ≤ 1LDK」といったような検索条件に変換できるとした．そして，非明示条件を含む発話に対してもタグをアノテーションした．さらに，タグアノテーションの根拠となる単語もアノテーションした．

彼らは検索の要件をタグとフィールド値への変換して表現しようとしているが，我々は自然文への変換（要約）で表現する点が異なる．カスタマーの希望は多種多様であるため，事前定義した有限個の要件に変換するのは情報の欠損が大きいと考える．

5.2 Kyutech Corpus

Kyutech Corpus[7, 8] は 4 名の議論 9 回分の音声書き起こしである．コーパスの各転記単位 [9] には価格，地域といった 28 種のトピックタグが少なくとも 1 つアノテーションされている．また，各対話につき 3 つの要約が 250 字以上 500 文字以内でアノテーションされている．これは，「対話の内容を知らない人が読んだ場合でも，議論の内容を理解できる」ことを趣旨として 3 人のアノテーションによりそれぞれ作成された．

山村らは要約文へトピックタグや各発話と対応の有無のアノテーションを Kyutech Corpus に行った [10, 11]．また，要約と対応付けされている発話文を重要文とし，重要文抽出の実験も行った．彼らは抽象化したトピックタグを介して，内容の一致箇所を分析しているが，我々は具体的な一致語句をアノテーションすることにより分析した．これにより，要約の際にどのように抽象化が行われるかの詳しい分析が可能となった．

6 おわりに

宿検索を目的とした対話においてカスタマーの希望がどのように具体的な要件として整理されていくのかを分析するため，対話データを収集し，発話の要約文・要件との対応関係のアノテーションなどを行った．今後は，構築したコーパスを用いて，要約文や要件の自動生成に取り組む予定である．

謝辞：アノテーションを行っていただき，多くの示唆に富んだご意見をくださった山下華代氏に感謝します．また，有益な助言していただいた荒瀬由紀准教授に感謝します．

参考文献

[1] Arseny Tolmachev, Daisuke Kawahara, et al. Design and Structure of The Juman++ Morphological Analyzer Toolkit. 自然言語処理, Vol. 27, No. 1, pp. 89–132, 2020.

[2] 河原大輔, 黒橋禎夫. 自動構築した格フレーム辞書と先行詞の位置選好順序を用いた省略解析. 自然言語処理, Vol. 11, No. 3, pp. 3–19, 2004.

[3] 横野光, 高橋哲朗. クラウドソーシングを用いた対話コーパス構築. 言語処理学会年次大会, pp. 783–786, 2017.

[4] Tetsuro Takahashi and Hikaru Yokono. Two person dialogue corpus made by multiple crowd-workers. In *Proceedings of the 8th International Workshop on Spoken Dialogue Systems*, 2017.

[5] Shun-ya Fukunaga, Hitoshi Nishikawa, et al. Analysis of Implicit Conditions in Database Search Dialogues. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp. 2741–2745, 2018.

[6] 福永隼也, 西川仁ほか. タスク指向対話におけるユーザ要求の理解とその根拠の抽出. 自然言語処理, Vol. 26, No. 1, pp. 121–154, 2019.

[7] 嶋田和孝, 山村崇ほか. Kyutech コーパス：意思決定タスクを対象とした複数人対話コーパス. 言語処理学会年次大会, pp. 1097–1100, 2016.

[8] Takashi Yamamura, Kazutaka Shimada, et al. The Kyutech corpus and topic segmentation using a combined method. In *Proceedings of the 12th Workshop on Asian Language Resources*, pp. 95–104, 2016.

[9] 小磯花絵, 西川賢哉ほか. 転記テキスト. 国立国語研究所報告 No.124 日本語話し言葉コーパスの構築法, pp. 23–132. 2006.

[10] 山村崇, 嶋田和孝. Kyutech コーパスにおける抜粋要約のアノテーションと分析. 言語処理学会年次大会, pp. 146–149, 2017.

[11] Takashi Yamamura and Kazutaka Shimada. Annotation and Analysis of Extractive Summaries for the Kyutech Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp. 3216–3220, 2018.