

叶内 晨 (Megagon Labs)

根石 将人 (東京大学)

林部 祐太 (Megagon Labs)

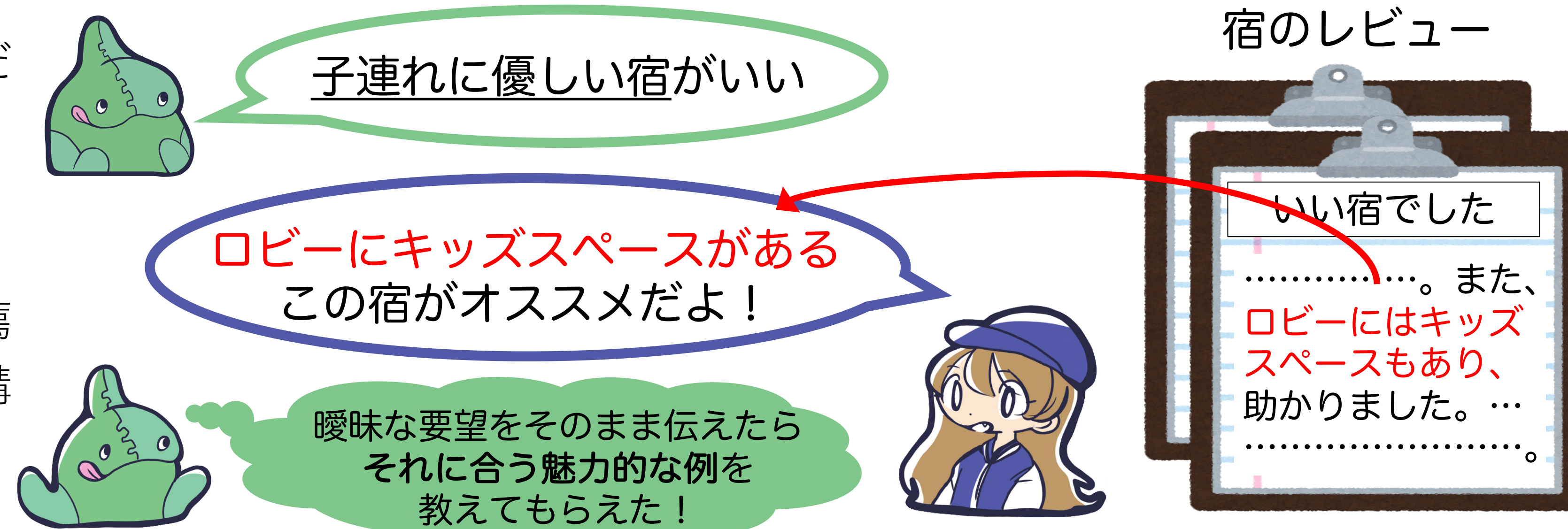
岡崎 直観 (東京工業大学)

概要

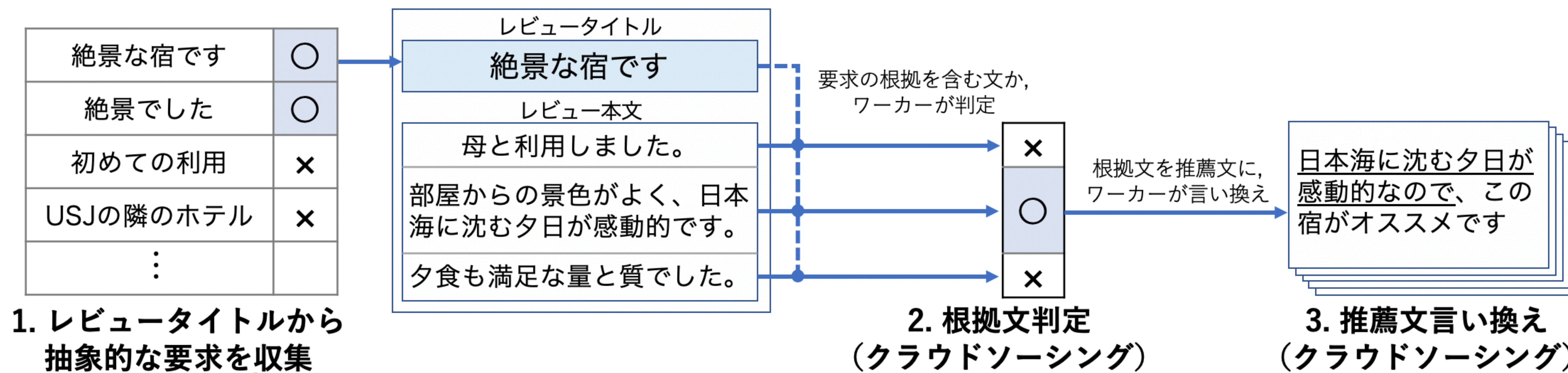
本研究では、抽象的な要求に対して根拠を含んだ推薦文を提示する対話システムの実現を目指す。

具体的には、抽象的な要求と推薦対象の宿のレビュー*が与えられた状況下で、レビュー本文から推薦根拠を含む文を抽出し、得られた根拠文を推薦文に言い換える。そのために必要なデータセット構築後、推薦文生成システムを構築・評価した。

*本研究では旅行情報サイトじゃらんnetのレビューデータを利用した



根拠説明データセットの構築



- 宿に対する要求抽出のため、宿を表す表現を含むレビュータイトルを抽出 (絶景な宿)
- 固有名詞を含まないなどのルールにより、ノイズをフィルタリング
- 収集した要求から「宿」を表す表現を除き、それと同じ内容語を含むタイトルも収集 (絶景な宿 -> 絶景 -> 絶景でした)
- Sudachipyの代表表記で表記揺れを吸収後、頻度上位のタイトルから人手でカテゴリ分けし、頻度を考慮して利用データを決定

表1: 収集した抽象的な要求の例と収集数・利用数

カテゴリ	例		収集数		利用数
	拡張前	拡張後	拡張前	拡張後	
綺麗なホテル	綺麗なホテル	綺麗な宿	15k	71k	4.6k
ゆったり	寛げる旅館	寛げました	8k	80k	4.6k
接客	親切な宿	対応が親切	10k	143k	4.6k
便利	便利なホテル	観光に便利	4k	113k	3.6k
子連れ	子供に優しい宿	子供に優しい	3k	81k	3.6k
景色	絶景のホテル	絶景でした	1k	34k	3.1k
食事	美味しい宿	美味しい朝食	1k	145k	3.1k
コスト	コストの良い宿	コストが最高	5k	89k	3.1k
グッド	最高の宿	最高でした	33k	278k	3.6k
その他	昭和な宿	昭和の雰囲気	19k	297k	16.0k

- ワーカーに要求(タイトル)と本文を提示し次の判定作業をお願い
 - 各文は要求と関係性がある文か
 - (a)の時、要求に対する根拠を含む文か (なぜなら~だから、と言えるような事物・サービスなどの具体例を含むか)
- 収集した5万レビューを対象として、各データを5人で作業
- チェック問題を作成し、それに正解したワーカーのデータのみ採用

表2: 根拠文の数とレビュー数

根拠文の数	レビュー数	割合
なし	24,034	(48.1%)
1文	21,675	(43.4%)
2文	3,464	(7.0%)
3文以上	827	(1.7%)

- ワーカーに要求と根拠文を提示し次の作文をお願い

「要求の宿を探している人に対し、~なのでこの宿がオススメですと根拠付きの説明をしてください」
- 可能な限り元の表現の利用をお願い
- 要求の感情表現が否定的で推薦できない場合と、根拠がない場合は別途報告をお願い
- 根拠文判定で5人中3人が根拠文と答えた31,351文を対象に、各文を5人で作文

表3: ワーカー間の推薦文の一致数

一致数	文の数	割合
3人以上一致	7,646	(24.4%)
2人一致	13,850	(44.2%)
一致なし	6,438	(20.5%)
根拠なし報告	1,155	(3.7%)
ネガティブ報告	2,262	(7.2%)

実験

① 根拠文判定のための二値分類タスク

- 根拠文判定で3人以上が根拠文と答えた文を正例、それ以外を負例
- Train : Dev : Test = 40k件(202k文) : 5k件(25k文) : 5k件(25k文)
- モデル
 - ロジスティック回帰
 - 特徴量は要求・本文それぞれのTFiDF、w2vの平均ベクトル、差分ベクトルなど
 - BERT[1]
 - 入力順: [CLS]、要求文、[SEP]、本文1文

手法	F1
ロジスティック回帰	47.92
BERT	74.07

② 要求と根拠文からの推薦文言い換えタスク

- 推薦文言い換えで3件以上推薦文が付与された文のみ利用
- 推薦文の共通文末である「この宿がオススメです」は削除し学習
- モデル
 - LuongらのLSTM[2]
 - VaswaniらのTransformer[3]
 - 入力順: 要求、[SEP]、根拠文1文

手法	BLEU
言い換えなし	49.17
LSTM	67.10
Transformer	59.50

③ 要求とレビュー本文からの推薦文生成タスク

- 要求に対する根拠がない場合は[根拠なし]トークンを出力
- 段階的モデル (実験①のBERT → 実験②のLSTM)
 - BERTで本文の各文が要求に対する根拠文か予測
 - 要求とスコア最大の1文を入力とし、LSTMで推薦文を生成
 - BERTで根拠文と予測された文がない場合は[根拠なし]
- 一括モデル (LSTM)
 - 入力順: 要求、[SEP]、レビュー本文 (複数文)
 - 推薦文がない場合は [根拠なし] を単体で出力するよう学習

手法	BLEU	F1*
段階的モデル (BERT→LSTM)	41.54	62.21
一括モデル (LSTM)	18.20	46.74

* SQuAD[4]と同様のMacro F1値。参照文(推薦文)がある場合は、参照文と生成文のbag-of-tokenの一致率をF1で計算。ない場合はないと答えたとスコア1で、それ以外はスコア0。

表7: レビュー入力時の推薦文生成の例

タイトル	本文	正解の推薦文 (参照文)	段階的モデル	一括モデル
見晴らしが最高でした	初日はチェックイン前にプレミアム観光券を購入してから、東山動物園にイケメンゴリラのシャバーニに会いに行き、翌日はリニア・鉄道館に行って満喫しました。お部屋は名古屋城が綺麗に見えるところで、子供達も大喜びでした。...	[お部屋は名古屋城が綺麗に見えるところなので][お部屋は名古屋城が綺麗に見えるところで、子供達も大喜びでしたので]	名古屋城が綺麗に見えるところで、子供達も大喜びでしたので	お部屋は名古屋城が綺麗に見えるところで
わんこ連れで満足!	夫婦2人と父、わんこ2匹でお世話になりました。... お部屋はゆったりしていて、大きな窓からは海が一望でき、とっても満足できました。... ベット可能ホテルはまだまだがっかりする所が多い中、こちらのホテルはお薦めですので[ベットと楽しめるので]	[ベット可能ホテルはまだまだがっかりする所が多い中、こちらのホテルはお薦めですので][ベットと楽しめるので]	ベット可能ホテルはとっても便利なので	お部屋はゆったりしていて、大きな窓からは海が一望でき、とっても満足できましたので
子連れにオススメの旅館	一歳四か月の子供がいるので、夕食は部屋出しなのと貸切温泉があるプランに惹かれて宿泊しました。... 敷き布団が年季が入っている感じで薄かったのもっと厚いと良かったかな。温泉や接客は良く、子供にも優しく接して貰えて良かったです。	[子供にも優しく接してくれるので][温泉や接客は良く、子供にも優しく接して貰えて良かったです]	温泉や接客は良く、子供にも優しく接して貰えたので	プールや接客は良く、子供にも優しいので

今後

- 根拠文の選び方の工夫・複数根拠文からの説明文生成
- 実システムへの導入

参考文献

[1] Jacob Devlin, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT, 2019.
 [2] Thang Luong, et al. Effective approaches to attention-based neural machine translation. In EMNLP, 2015.
 [3] Ashish Vaswani, et al. Attention is all you need. In NIPS, 2017.
 [4] Pranav Rajpurkar, et al. Squad: 100,000+ questions for machine comprehension of text. In EMNLP, 2016.