宿の推薦根拠説明システムにおける魅力度の考慮と実用を見据えた評価

叶内 晨 [1*]　根石 将人 [2]　林部 祐太 [1]　大内 啓樹 [3]　岡崎 直観 [4]
[1] 株式会社リクルート Megagon Labs　[2] 東京大学　[3] 理化学研究所　[4] 東京工業大学

1 はじめに

近年，機械学習モデルを利用した推薦時において，なぜその候補を推薦するのか，という推薦根拠の説明性の研究に注目が集まっている [1, 2, 3, 4]. Kanouchi ら [5] は抽象的な要求に対する根拠付きの推薦文を提示するために，根拠説明データセットとそのモデルを構築した．しかし、実際にユーザに宿を推薦するシーンを想定すると推薦根拠の候補は無数に考えられ，ユーザにとってより魅力的な根拠を選択して提示する必要があるが，彼らのモデルでは利用する推薦根拠の選定ができていない．

そこで本研究では，推薦根拠説明システムの実サービスへの導入に向け，Kanouchi らの取り組みを基にして根拠がどの程度魅力的なのかを予測するモデルを構築し，その出力の有用性をドメイン知識を有した実務担当者により評価する．評価の結果，根拠判定に加えて魅力度を考慮したモデルが適合率 0.81 で推薦根拠の抽出に成功し，生成した推薦文のうち 55%は実業務で利用可能なことを報告する．

2 先行研究

Kanouchi ら [5] は，宿予約時の推薦根拠説明のために，根拠説明データセット[1)]とそのモデルを構築した．データセット構築では，(1) レビュータイトルと抽象的な要求の類似性を利用して，宿のレビューデータのタイトルの一部を抽象的な要求とみなして収集し，(2) 対応するレビュー本文中で要求に対応する根拠を含む文をアノテーションし，(3) その根拠文を推薦文へ言い換えた．モデル構築では，(1) 要求と各文が与えられた際に根拠文かどうかを予測する BERT モデル [6] と，(2) 要求とその根拠文が与えられた時に推薦文を生成する LSTM ベースのモデル [7] の 2 つのモデルを連結させることで，要求に対する根拠付きの推薦文を生成した．

しかし，推薦文を実際にユーザに提示する際は，抽出される根拠文は複数であるのに対して提示する推薦文は大抵 1 文であり，複数の根拠の中からユーザにとってより魅力的な根拠を選ぶ必要がある．そこで本研究では，魅力度を考慮した推薦文説明システムを構築し，その出力結果を人手評価する．

3 魅力度のアノテーション

複数の根拠の中からユーザにとってより魅力的な根拠を抽出するため，要求に満たす根拠文に対して魅力度をアノテーションした[2)]．クラウドソーシング[3)]を利用し，ワーカーに対して「あなたはとある要求（例えば，絶景の宿）を満たす宿を探しています．次の根拠文が書かれた宿にどれくらい魅力を感じますか」という質問をして，次の 4 段階から 1 つを選択してもらった．

1. **大変魅力的**: このレビュー文が決め手で対象の宿が第一候補になり得る
2. **魅力的**: 第一候補ではないが嬉しく，このレビュー文が理由で予約候補になり得る
3. **普通**: 要求に関する情報だが，特別嬉しくは感じない
4. **魅力的ではない**: その根拠ではこの宿にしたいとは思わない・判断できない

データは，Kanouchi ら [5] が作成した根拠説明データセットに含まれる根拠文からランダムな 8,000 文を対象とし，各文を 5 人でアノテーションした．ただし，宿予約時にユーザがネガティブな要求はしないと仮定し，要求がネガティブな事例 (例えば，接客が微妙) はアノテーションの対象から除いた．

表 1 にアノテーション結果を示す．大変魅力的な根拠が全体の 25.3%，魅力的な根拠が 47.7%なのに対して，魅力的ではない根拠は 4.3%に留まった．

* shin187nlp@megagon.ai
1） https://github.com/megagonlabs/ebe-dataset
2） 根拠文ではなく推薦文に魅力度をアノテーションする方法もあるが，作成される推薦文は手法に大きく依存するため，根拠文に対してアノテーションした．
3） https://crowdworks.jp

表 1　魅力度のアノテーション結果の分布と例

	割合 [%]	例 要求 (タイトル)	根拠文 (レビュー文)
大変魅力的	25.3	アートでおしゃれ	玄関ロビーが，現代アートの美術館になっていました.
魅力的	47.7	ファミリーには最高	子供も遊べるプールもあってよかったです.
普通	22.7	便利でよいホテル	隣のビルの 1 階に居酒屋があり食事をするのに便利でした.
魅力的ではない	4.3	使い勝手の良いホテル	自転車も借りられるので，近隣の移動も便利.

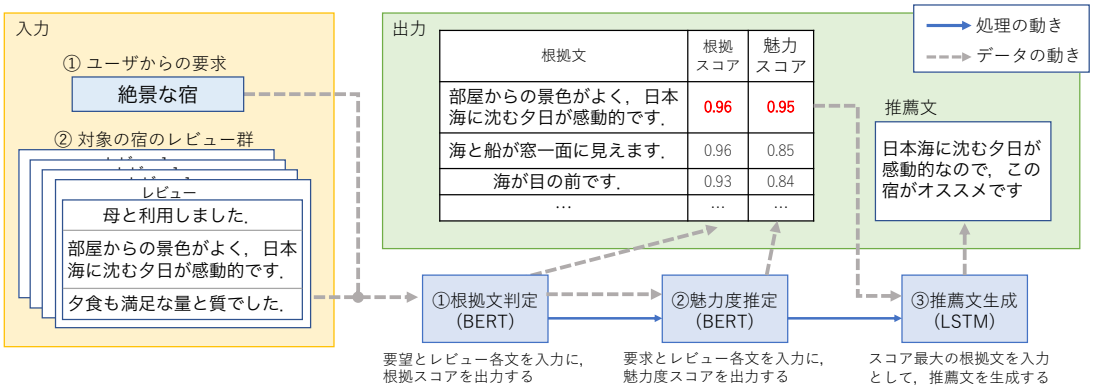

図 1　推薦根拠説明システムの概要図

4　システム構築

本研究で構築した，魅力度も考慮した推薦根拠説明システムの概要とデータの流れを図 1 に示す．Kanouchi ら [5] の根拠文判定モデルと推薦文生成モデルはそのまま利用し，追加で魅力度推定モデルを構築した．魅力度推定のためのデータセットとして，3 節のアノテーションで 5 人中 2 人以上が大変魅力的と答えた文を正例，普通もしくは魅力的ではないと答えた人が 1 人以上で，かつ誰も大変魅力的と答えていない文を負例とした．データ数は正例 2,908 件：負例 2,090 件で，ランダムに 8：1：1 の比率で分割して学習：検証：テストデータとした．

モデルは BERT [6] を用い，その他すべての条件は Kanouchi らの BERT モデルと同様の設定とした．BERT モデルは二値分類問題として finetuning したが，モデル最終層の出力結果を softmax 関数で正規化することで，推論時の確信度を得た．根拠判定モデルに対する確信度として根拠スコア，魅力度推定モデルに対する確信度として魅力スコアを得た．

魅力度も考慮した根拠抽出の有用性確認のため，次の 3 つの手法で根拠文を抽出し，評価した．

手法 1) **根拠スコア最大**: 魅力スコアを利用せずに，根拠スコア最大の根拠文を利用する

手法 2) **魅力スコア最大**: 根拠判定で根拠文と予測された文に対して魅力度推定を行い，魅力スコア最大の根拠文を利用する

表 2　利用可能な根拠文の割合

	利用可能	(強い根拠)
手法 1) 根拠スコア最大	0.79	(0.33)
手法 2) 魅力スコア最大	0.74	(0.32)
手法 3) 根拠 * 魅力最大	**0.81**	(0.32)

表 3　根拠判定モデルで抽出された根拠文数毎の利用可能な根拠文の割合

	12 文以下 (下位 50%)	13 文以上 (上位 50%)
手法 1) 根拠スコア最大	0.75	0.84
手法 2) 魅力スコア最大	0.63	0.86
手法 3) 根拠 * 魅力最大	0.75	**0.88**

手法 3) **根拠スコアと魅力スコアの積が最大**: 根拠判定によって根拠文と予測された文に対して魅力度推定を行い，根拠スコアと魅力スコアの積が最大の根拠文を利用する

5　評価

ドメイン知識を有する実務担当者 3 名により，システムの出力が実サービスで利用可能か評価した[4]．評価対象は 2 種類で，要求と宿が与えられた際の（1）システムが抽出した根拠文と（2）システムが生成した推薦文を評価した．

評価セットは，50 種類の要求毎にそれぞれ宿を 2 つ選定し，合計 100 件の宿とした．クラウドソーシングで「宿を提案してくれるコンシェルジュに宿に関する希望を伝える際，どのようなメッセージを

4）本論文ではサービスでの実運用の可否を調査するため，実務担当者の判断を優先し自動評価は行わない．

表 4　利用不可能な根拠文のエラー分析

エラーの種類	件数	例	
		要求	抽出された根拠文
根拠文が利用水準以下	10	観光に便利な宿	観光地にも近く，大きなホテルよりも車の出入りが簡単で，気軽にあちこち外出できたのが便利でよかったです．
		ゆったりくつろげる宿	初めての利用でしが，滞在中はほとんど貸切状態でゆっくり入浴出来て，研修の疲れも取れました．
根拠文とは言えない	6	繁華街へのアクセスが便利な宿	空港からも近く，夜も静かで目の前が海なので那覇市内というよりも中部方面のホテルに泊まってるような感じでした．
根拠文抽出結果 0 件	3	大人の女性旅に向いた宿	(抽出された根拠文なし)

送りますか」という質問でデータを収集し，実務担当者とは別の関係者が評価セットに用いる要求を 50 種類選んだ．評価に用いる宿は，旅行情報サイトじゃらん net[5] にある宿とし，Li ら [8] の手法を基にした独自のシステムを用いて，要求を入力し宿候補を得た．また根拠説明のみを評価するために，システムが出力した宿候補を関係者が目視で確認し，要求を満たす宿のみを評価対象とした．その際，レビュー数が最低 30 件以上ある宿に限定し，また 100 件以上ある場合は最新の 100 件のみ実験に用いた．

5.1　抽出した根拠文の人手評価

4 節で構築したシステムの手法 3 つを評価した．要求をもつユーザに宿を実際に提案する状況を想定し，その根拠文を使ってユーザにお薦めをできるか，また利用できる根拠文である場合は宿を薦める強い根拠となるかを 3 段階で評価した．

利用可能:

1. 強い根拠なので，積極的にユーザに伝えたい
2. 根拠であり，ユーザに伝える価値がある

利用不可:

3. 根拠ではない，もしくは根拠ではあるが実業務では使えない

全ての事例を実務担当者 3 名で評価し，3 名のアノテーションの一致率（Fleiss の kappa 値）は 0.425 で，中程度の一致となった．3 名の意見が割れた事例については，多数決を採った．

手法ごとの根拠文の評価結果を表 2 に示す．手法 3 が 0.81 の割合で利用可能な根拠を抽出することに成功しており，精度がもっとも良かった．強い根拠の抽出精度は手法毎に大きな差はなかった．

魅力スコアにより結果が改善した例として，「健康的な朝食」という要求に対して，手法 1 は「朝食が品数も多く，朝からワインもあり，地元食材が豊富でした」という文を選択し，利用可能な根拠と判定された一方で，魅力度を考慮した手法 2,3 では，「朝食は北海道の食材満載で，特に野菜のクオリティは非常に高く，サラダ以外に温野菜をチーズフォンデューにしたり，スープカレーに入れたりして食べてました.」と具体的根拠を多く含んだ文の抽出に成功し，強い根拠であると判定された．

5） https://www.jalan.net/

表 5　利用可能な推薦文の割合

そのまま提示可能	0.09
一部修正後に提示可能	0.46
利用不可	0.26
根拠文の時点でのエラー	0.19

根拠判定モデルで抽出された根拠文数毎の評価結果を表 3 に示す．根拠判定モデルによって十分な量の根拠文（13 文以上）が抽出されている場合，手法 3 は，0.88 の精度で利用可能な根拠文を抽出することに成功した．一方で根拠判定モデルによって抽出された根拠文数が少ない場合（12 文以下），すべての手法において精度が低下し，特に手法 2 の魅力スコア最大の根拠文を利用する場合のスコアが 0.63 でもっとも低かった．事例を確認しながら精度低下の原因を調査したところ，BERT の根拠判定モデルによって抽出される根拠文が少ない場合，抽出された根拠文に偽陽性（根拠とは言えない文）が多く含まれることがわかった．そのため，抽出される根拠文が少ない場合は，魅力スコアによる並べ替えをせずに根拠スコアが最大ものをそのまま利用した方が，精度高く根拠文を抽出できたと考えられる．

表 4 に，実務担当者が実業務で利用不可能と答えた手法 3 の出力 19 件に対するエラー分析結果を示す．もっとも多いエラーは 10 件で，根拠文ではあるものの利用できるレベルに達していないと判断された事例であった．例えば実業務では「観光に便利な宿」という要求に回答する場合，近隣の観光地名とそこまでの所要時間の明確な回答が求められる．しかし，抽出された根拠文にその情報はなく，利用水準以下と判断された．他の例には「ゆっくりくつろげる宿」という要求に対して「ほとんど貸切状態

表 6　システムが抽出した根拠文と生成した推薦文の例

例	要求	抽出した根拠文	生成した推薦文	生成評価
1	子供連れにおすすめな宿	食事会場にベビーベットの用意があったり，子供の遊び場スペースや絵本おもちゃの貸し出しもあり，子連れの旅行には良いと思います．	食事会場にベビーベットの用意があったり，子供の遊び場スペースや絵本おもちゃの貸し出しもあり，子連れの旅行には良いので，こちらの宿がオススメです．	全て提示可能
2	自然を楽しめる宿を希望しています	夜は天の川まで見えるほど星がたくさん見え，早朝，部屋と露天風呂からは眼下に雲海を見ることができました．	夜は天の川まで見えるほど星がたくさん見えますので，こちらの宿がオススメです．	全て提示可能
3	料金が格安の宿があれば最高です！	京都駅とホテルの１５分間隔の無料シャトルバス，館内の居酒屋，部屋の携帯充電器等，ホテル横のコンビニ等，安価な宿泊料，サービス，設備等，環境共に満足のいく内容でした．	朝食の携帯充電器等，ホテル横のコンビニ等，安価な宿泊料，サービス，設備等，環境共に満足のいく内容ですので，こちらの宿がオススメです．	一部修正
4	館内だけで満足できる宿	スパ（有料），プール，ジムなどの施設が非常に充実していて，ホテルの中だけでも楽しむことができました．	温泉，プール，ジムなどの施設が非常に充実しているので，こちらの宿がオススメです．	一部修正
5	子どもが喜ぶ要素のあるホテルはありませんか	選べる風呂用の色かご，ひのきや炭の枕，色んな色の浴衣，小学生の娘はもう，楽しくてはしゃいでいました！	普通の風呂用の色かご，ひのきや豊富なので，こちらの宿がオススメです．	利用不可

でゆっくり入浴」というように，根拠が偶発的・限定的すぎるために推薦では利用不可と判断された事例も確認された．これらの文は Kanouchi ら [5] の研究では根拠文としてアノテーションされておりモデルの出力は想定どおりであるため，今後アノテーションの修正かモデルの改善が必要である．

2 番目に多いエラーは根拠文とは言い切れない事例で，例えば「繁華街へのアクセスが便利な宿」という要求に対して「空港からも近く」という文は，アクセスが便利であることを表すが，繁華街へのアクセスの利便性は不明である．このような要求に状況を限定する句が含まれている場合に，モデルがその条件を認識できず，解析に失敗する事例が複数確認された．改善策としては，条件などを含む難しい要求を集めて Kanouchi らの根拠説明データセットを拡張する方法などが考えられる．残りのエラーは 3 件で，根拠を 1 件も抽出できない事例であった．

5.2　生成した推薦文の人手評価

4 節で構築したシステムによって生成された推薦文が提示可能かどうか評価した．評価セットは 5.1 節と同様の設定とし，その中から「利用可能」な根拠文と判定された事例のみを評価対象とした．ただし，生成モデルはすべての手法で同一なため，根拠文抽出の精度がもっとも良かった手法（3）の生成結果にのみ次の 3 択で評価した．

1. **そのままユーザへ提示可能**
2. **一部修正すればユーザへ提示可能**
3. **利用不可，もしくは部分的に使えたとしても自分で書いた方が早い**

評価結果を表 5 に示す．そのまま提示可能な割合が 0.09，一部修正後に提示可能な割合が 0.46 となり，合計で半分以上の推薦文をオペレータ業務で利用可能である．しかし，そのまま提示可能な割合は 0.09 であり，実務担当者がそのまま利用可能な推薦文を生成できる割合は低い．

表 6 に生成した推薦文の例とその評価結果を示す．例 1,2 は生成に成功し，そのまま提示可能な推薦文の例である．一部修正する必要がある例として，例 3 は「格安」をユーザが求めているのに対して推薦文では「安価」であることを主張しているため修正が必要な例で，例 4 は「館内だけで満足できる宿」に対して施設が充実していることを説明しているが，部屋や食事などの重要な要素への言及がなく，実務利用のためには改善の余地がある例である．また利用不可能な事例に関しては，例 5 のように根拠文の時点では利用可能であるが，推薦文生成時に LSTM モデルが誤って出力を短くしすぎる事例が多くみられた．

6　終わりに

本研究では，Kanouchi らの根拠文判定と推薦文言い換えに加えて根拠の魅力度推定を行い，宿推薦時の推薦根拠説明システム全体を人手評価した．評価の結果，抽出した根拠文のうち 81%に利用価値があり，また根拠文が十分に確保できる状況では，さらにその精度が向上することを示した．生成した推薦文は，55%を業務で利用可能であることを示した．今後はこれらのシステムを実サービス上で活用していくための改善と開発をしたい．

参考文献

[1] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web (WWW 2001)*, pp. 285–295, 2001.

[2] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 83–92, 2014.

[3] Guoshuai Zhao, Hao Fu, Ruihua Song, Tetsuya Sakai, Zhongxia Chen, Xing Xie, and Xueming Qian. Personalized reason generation for explainable song recommendation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 10, No. 4, pp. 1–21, 2019.

[4] Yongfeng Zhang, Xu Chen, et al. Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval*, Vol. 14, No. 1, pp. 1–101, 2020.

[5] Shin Kanouchi, Masato Neishi, Yuta Hayashibe, Hiroki Ouchi, and Naoaki Okazaki. You may like this hotel because ...: Identifying evidence for explainable recommendations. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP 2020)*, pp. 890–899. Association for Computational Linguistics, 2020.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pp. 4171–4186, 2019.

[7] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pp. 1412–1421, 2015.

[8] Yuliang Li, Aaron Feng, Jinfeng Li, Saran Mumick, Alon Halevy, Vivian Li, and Wang-Chiew Tan. Subjective databases. *Proceedings of the VLDB Endowment*, Vol. 12, No. 11, pp. 1330–1343, 2019.