宿レビューからの肯定的事実と推薦対象の抽出

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

オンラインでの宿予約では，フォームに条件を入力して検索し，検索結果から選んで決めるという流れが一般的である．フォームには，日付・エリア・人数・予算などの基本的な条件のほか，食事の有無，部屋のタイプ，喫煙の可否，大浴場の有無などの「こだわり条件」が設定できることがある．しかし，それらの「こだわり条件」は操作性やスペースの制約のため，多くの人が気にすると思われる一般的な条件の中からしか選べるようになっていない．そのため,「自動販売機のビールが安い」「キッズスペースがある」といった，よりきめ細やかな条件では宿を探せない．

そこで我々は，ユーザの旅行の状況を聞き出し，その状況に合わせた宿を提案する対話システムの構築を目指している．本研究では，的確な推薦のために，「肯定的事実」と「推薦対象」という言語知識を宿のレビューテキストから抽出する手法を提案する．例えば,「自販機のビールがかなり安いので，酒飲みには嬉しい」や「子どもたちには、キッズスペースや図書館など楽しかったようです」というレビュー文から「自販機のビールがかなり安い」ことは「酒飲み」に肯定的であることや,「キッズスペースや図書館」が「子どもたち」に肯定的である，という知識を抽出する．

提案手法を用いて，旅行情報サイト「じゃらん net」[1] に投稿された宿のレビューから肯定的事実と推薦対象を 7,701 組抽出した．そして，そのうち 2,439 組に対して肯定的事実のアスペクトのアノテーションと，肯定的事実を推薦対象にアピールポイントとして用いることが妥当であるかの評価を，クラウドソーシングを用いて実施した．

2 関連研究

2.1 レビューから得た知識に基づく推薦

Reschke らはレストランレビューサイト Yelp から得た知識をもとに，レストラン推薦対話システムを提案した [1].

まず，和食，イタリアンなどといったレストランのカテゴリのレビューごとに，Latent Dirichlet Allocation (LDA)[2] を適用し，単語クラスタリングを行った．得られたクラスタをサブカテゴリとみなし，人手でクラスタラベルをアノテーションした．また，人手で定義した構文パターンと評価極性辞書を用いて，カテゴリごとに肯定的に用いられている名詞句をアスペクトとして抽出した．

(1) This place has some really great yogurt and toppings.

(2) Do you want a place with a good yogurt?

例えば，例 (1) からは “yogurt” と “toppings” を抽出する．これは，レビューで好評なアスペクトをもつレストランを例 (2) のように推薦するためである．

対話システムは，まずカテゴリに関して質問する．次にユーザの回答に基づき，サブカテゴリに関する質問する．その後は，アスペクトに関する質問をテンプレートで生成し，繰り返す．

本研究では彼らと同様に，構文パターンと評価極性を表す語を用いて肯定的な事実をレビューから抽出するが，抽出単位は名詞句に限定せず，句とする．

2.2 構文パターンに基づく評価文抽出

鍛治らはウェブコーパスに含まれる定型文，箇条書き，表という 3 つの記述形式から評価文を自動で抽出する手法を提案した [3]．レビュー文には箇条書きや表はほとんど無いので，ここでは定型文に基づく手法について述べる．

彼らは,「良いところは計算が速いことです」のように「良い」,「素晴らしい」のような「好評手がかり句」が,「〜ところは〜ことだ」のような特定の構文パターンで用いられている文を抽出する方法を提案した．約 10 億件の HTML テキストに対して実施したところ，約 54 万文が得られ，Precision は 80%から 90%程度と報告している．

[1] https://www.jalan.net/

予備実験として，同じ手法をレビューコーパスに適用した所，非常に低い Recall となった．これは，彼らのパターンは Precision を重視しているためだと考えられる．そこで，本研究ではより多くの文を抽出できるパターンを提案する．また彼らの研究では抽出していない推薦対象も抽出する．

2.3 レビュー文内の句に対する評価極性アノテーション

中澤らは宿のレビュー文の句に対して 5 段階の評価極性をアノテーションした [4]．このデータは首都大学東京日本語評価極性タグ付きコーパスとして公開されている[2]．

まず彼らは，楽天トラベルのレビュー文に評価極性を人手でアノテーションされている TSUKUBA コーパス[3] に対して単語分割と句構造解析を行い，59,758 種類の句を自動抽出した．このうち 10,000 句に 3 人で，49,758 句に 2 人で，評価極性をアノテーションした．アノテーション時には句のみを与え，前後文脈は与えなかった．

本研究では彼らと同じドメインである宿のレビュー文を扱うが，肯定的な文に限定して自動的に抽出する．また，推薦対象の抽出とアスペクトのアノテーションも行う．

2.4 レビューのアスペクト分類

SemEval-2016 Task 5 では宿のレビューの感情分析において 7 つの Entity Label (hotel, rooms, room amenities, facilities, service, location, food&drinks) を定義し，アノテーションしたデータが使われた [5]．

Fukumoto らは楽天トラベルのレビューのアスペクトを，レビュー投稿フォームのスコア付けの 7 つ基準（サービス，風呂，部屋，食事，立地，設備・アメニティ，総合）で分類した [6]．

安藤らは楽天トラベルのレビューを「商品の絶対的事実」「売り手の要望」「買い手の購入理由」など 23 項目で分類した [7]．また，そのホテルに行きたいと思わせる文か，行きたくないと思わせる文かの判断を「インパクト」とよび，10 名でアノテーションした．彼女らは商品に関する絶対的事実がインパクトに影響することを示した．

[2] https://github.com/tmu-nlp/sentiment-treebank

[3] http://www.nlp.mibel.cs.tsukuba.ac.jp/~inui/SA/corpus/

表 1: 抽出される文が満たすべき条件

1. 主節[4]が次のいずれかの述語である
   楽しい，好きだ，最高だ，素晴らしい，大好きだ，便利だ，満足だ，面白い，良い，良好だ，優れる，素敵だ，嬉しい，助かる，完璧だ，抜群だ，優秀だ，最強だ，絶妙だ，良質だ，有り難い，十二分だ，ユニークだ，文句無しだ，打って付けだ，パーフェクトだ
2. 主節の feature に “否定表現” と “準否定表現” が無い
3. 主節に条件節[5]が係らない
4. 主節のガ格項が「方（ほう）」または「以外」でない
5. 次の制約を満たす節が存在する
   (a) 意味マーカ “人” をもつ（「敵」[6]を除く）
   (b) feature に “一人称” がある場合と，次のいずれかを含む場合は，係元が存在する
   <数量>+人，私，個人的，自分
   (c) どちらかの制約を満たす
      i. 「には」を含み，主節に係る
      ii. 「に」で終わり，直後の節が「とって」で始まり主節に係る

3 肯定的事実と推薦対象の自動抽出

宿を推薦するための言語知識として，肯定的な評価を受けている特徴を含む**肯定的事実**と，その特徴が誰に対して推薦できるかを表す**推薦対象**を考える．本研究ではそれらをレビュー文から自動抽出する．

(3) フロント横にあるコーヒーマシンがいつでも無料で利用でき、コーヒー好きの私には嬉しかったです。

例えば例 (3) というレビュー文から，肯定的事実「フロント横にあるコーヒーマシンがいつでも無料で利用でき、嬉しかったです」と，推薦対象「コーヒー好きの人」を抽出する．

3.1 前処理

はじめに，NFKC 正規化や半角文字の全角化，ハイフンや長音記号の正規化[7]を行う．次に，句点を含まなかったり特殊文字が入ったりしているレビューは除外する．そして，句点，感嘆符，疑問符，音符や星記号など文末に置かれやすい記号を手がかりに文分割する．

[4] ここでの “節” は，一般的な「節」ではなく，1 つの自立語と後続する付属語から構成される KNP++の係り受け解析や格解析の基本単位である “phrase” をさす

[5] 「自分てきには」の「てき」が「敵」と誤解析され，意味マーカ “人” が付与されるような場合があったため

[6] “節機能-条件”, “T 条件節候補”, “節機能疑-条件” のいずれかの feature をもつ節を条件節とみなした

[7] https://github.com/neologd/mecab-ipadic-neologd/wiki/Regexp.ja での処理を参考に実施

表 2: 抽出した肯定的事実と推薦対象の組の例

肯定的事実	推薦対象
朝食のサラダバーは良かった。	野菜が大好きな人
お部屋は広く、洗面ボウルが二つあり、嬉しい！	女性
連泊者は、無料の洗濯サービスがあり、助かります。	出張者
お風呂は畳で足が滑らず、とても良かったです。	足の悪い母

表 3: 推薦対象の例

頻度	推薦対象	頻度	推薦対象
843	女性	5	喉の弱い人
154	子供	5	妊娠中の人
82	女子	5	小さな子供
79	出張者	5	愛煙家の人
73	ビジネスマン	3	タバコが苦手な人
58	コーヒー好きの人	3	コーヒー党の人
54	子連れの人	2	低学年の子供
50	旅行者	1	朝からガッツリ食べる人

最後に，JUMAN++ (v2.0.0-rc2)[8]で形態素解析，KNP++ (0.9-21cc58c)[9]で構文・格解析を行う．

3.2 ルールに基づく抽出

KNP++の解析結果を用いて，表 1 に示した条件をすべて満たす文をコーパスから抽出する．これにより，肯定的な評価が述べられていて（条件 1, 2），願望や仮定といった要望表現ではなく（条件 3, 4），推薦対象が明示されている（条件 5）文が抽出される．feature の詳細はウェブサイト[10] を参照されたい．

条件 5 を満たす節と，その係元の子孫すべてからなる部分木より「には」または「にとって」を除いたものを推薦対象とする．なお，その部分木の親が制約 5b に該当する場合は「人」に置き換える．また，文から部分木を除いた文を肯定的事実とする．

4 抽出した知識の分析

旅行情報サイト「じゃらん net」に投稿された宿のレビューに 3 節の手法を適用し，肯定的事実と推薦対象を 7,701 組得た．抽出した例を表 2 に示す．

4.1 推薦対象の分析

推薦対象の例を表 3 に示す．異なり数は 4,454 だった．「出張者」や「子連れの人」など，さまざまな推薦対象が抽出できている．今後の課題は次のとおりである．

表 4: 肯定的事実のアスペクト分類の結果

アスペクト	件数	平均選択率
接客・サービス	374	70.9
客室の様子，設備・アメニティ	558	80.8
ホテルの様子，設備・アメニティ	543	77.7
食事	561	91.1
立地	196	80.1
以上の複数に該当	99	66.7
その他（ホテルに関係，周辺情報）	64	63.4
その他（ホテルと無関係，選択不能）	44	70.0
全体	2,439	79.6

表 5: 妥当性判定の結果

判定	件数	平均選択率
はい	2,229	83.5
どちらともいえない	94	49.4
いいえ	78	50.8
選択不能	38	49.5
全体	2,439	80.6

表 6: 妥当性判定における組ごとの「はい」の回答数の合計

0	1	2	3	4	5	6	7	8	9	10	計
16	44	55	85	88	142	166	223	310	528	782	2,439

- 「女性」と「女子」や,「コーヒー好きの人」と「コーヒー党の人」などの類義表現の正規化
- 「足の悪い母」や「足の悪い父」を「足の悪い人」とするなど，本質的な表現に着目して汎化
- 構文解析のエラーによる不適切な推薦対象を除外

  (4) 料理は、船盛のお刺身が、とても新鮮で美味しく食べきれない程の量で、お刺身が大好きな人には最高です。

  という文では下線部が推薦対象とされてしまった
- フィルタリングの改善

  (5) 他はフロントスタッフの親切な対応や、バイキングには満足でした。

  という文から，意味マーカ “人” をもつ「バイキング」[11]を抽出してしまっていた．また,「我ら年寄り」「我々年配者」「私のような人」のように，一人称が修飾句に含まれる場合も除去すべきである．

4.2 肯定的事実のアスペクト分類

肯定的事実として表されている内容の傾向を把握するために，2.4 節で述べた関連研究を参考に，表 4 で

[8] https://github.com/ku-nlp/jumanpp
[9] 著者より直接入手
[10] http://nlp.ist.i.kyoto-u.ac.jp/?plugin=attach&refer=KNP&openfile=knp_feature.pdf
[11] 海賊の語義もあるため

示すように肯定的事実を 8 種類に分類した．

そして，得られた肯定的事実 7,701 個のうち 2,439 個を対象に，クラウドソーシングでアノテーションした．アノテーションの質を確保するため，アノテーション対象の 10 問にあらかじめ正解を人手でアノテーションしたチェック問題を 1 問加えて，11 問を 1 タスクとした．そして，チェック問題に正解した場合のみタスクの回答を受理し，ワーカーに報酬を与えた．肯定的事実 1 個につき 5 人がアノテーションする．

アノテーション結果を表 4 に示す．肯定的事実ごとに投票数が最多のアスペクトを選択し，同数の投票が存在した場合，リストで上位のものを選択した．また，その選択肢が選ばれている割合を「選択率」として求め，アスペクトごとの平均を表に示した．選択率が高いほどアノテーションの信頼度が高いといえる．全体での選択率の平均は 79.6%であった．

この表から，食事に関する肯定的事実が最も多く，アノテーション信頼度も最も高いことが分かる．一方，「接客・サービス」は平均選択率が比較的低い．例えば，

(6)　温度調節もしてもらえるので良かったです。

のアスペクトとして 3 人が「接客・サービス」を，2 人が「客室の様子，設備・アメニティ」を選択した．

(7)　あのバリエーションはうれしかったです。

のように，省略や照応があり，文脈が無いと判断できない場合は「その他」が選択されていた．

4.3　肯定的事実と推薦対象の組の妥当性

自動抽出した肯定的事実と推薦対象の組はレビュー記入者の主観に基づいており，一般的な知見とはいえない可能性がある．そのため，肯定的事実を推薦対象にアピールポイントとして用いることが妥当であるか，の判定を行った．判定は，「はい」，「どちらともいえない」，「いいえ」，「選択不能」の 4 択で行った．4.2 節同様にクラウドソーシングを用い，1 組につき 10 人がアノテーションした．推薦システムに用いる際は，多くのワーカーが「はい」と答えた組のみを用いることと考えている．

アノテーション結果を表 5 に示す．また，組ごとに「はい」の回答数を集計し，回答数ごとにまとめた統計を表 6 に示す．この表によれば全員が「はい」と答えた組は全部で 782 組，全員が「はい」と答えなかった組は全部で 16 組あった．7 割以上「はい」と答えた組は全体で 75.6%だった．例 (7) のように文脈がないと判断できない肯定的事実に対しては「はい」が少なかった．判断が分かれた例を次に示す．

(8)　部屋が自動ロックでないのは、有難かったです。

という肯定的事実と推薦対象「年寄り」の組に対しては「はい」に 5 人，「どちらともいえない」に 5 人がアノテーションした．

5　おわりに

本研究では肯定的事実と推薦対象の自動抽出手法を提案し，宿のレビューテキストに対し適用した．そして，得られた組の一部に対して，アスペクトと妥当性をアノテーションした．

今後は，さらにデータを洗練・拡充させ，宿を推薦する対話システムに用いる予定である．また，推薦対象が明示されていない文に対する推薦対象のアノテーションや，抽出源をウェブテキストとした一般的な知識の抽出なども考えている．

参考文献

[1] Kevin Reschke, Adam Vogel, et al. Generating Recommendation Dialogs by Extracting Information from User Reviews. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 499–504, 2013.

[2] David M Blei, Andrew Y Ng, et al. Latent Dirichlet Allocation. *The Journal of Machine Learning Research,*, Vol. 3, pp. 993–1022, 2003.

[3] 鍛治伸裕, 喜連川優. HTML 文書集合からの評価文の自動収集. 自然言語処理, Vol. 15, No. 3, pp. 77–90, 2008.

[4] 中澤真人, 池田可奈子ほか. リビュー文書を対象とした句単位の日本語評価極性タグ付きコーパス. 言語処理学会 第 24 回年次大会 発表論文集, pp. 781–784, 2018.

[5] Maria Pontiki, Dimitris Galanis, et al. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pp. 19–30, 2016.

[6] Fumiyo Fukumoto, Hiroki Sugiyama, et al. Incorporating Guest Preferences into Collaborative Filtering for Hotel Recommendation. In *Proceedings of 6th International Joint Conference on Knowledge Discovery, Knowledge Enginnering and Knowledge Management*, pp. 22–30, 2014.

[7] 安藤まや, 関根聡. レビューには何が書かれていて、読み手は何を読んでいるのか? 言語処理学会 第 20 回年次大会 発表論文集, pp. 884–887, 2014.

[8] Arseny Tolmachev, Daisuke Kawahara, et al. Juman++: A Morphological Analysis Toolkit for Scriptio Continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 54–59, 2018.

[9] Daisuke Kawahara, Yuta Hayashibe, et al. Automatically Acquired Lexical Knowledge Improves Japanese Joint Morphological and Dependency Analysis. In *Proceedings of the 15th International Conference on Parsing Technologies*, pp. 1–10, 2017.