データセットの利用経験を通して何を得た？
：楽天データを用いた研究事例から

中山 祐輝
楽天株式会社 楽天技術研究所
2018/11/28(水)

Rakuten

自己紹介：中山 祐輝（なかやま ゆうき）

- 出身：石川県 能美市（松井秀喜さんの出身地, JAISTの所在地）
- 経歴
  - 2008年、石川高専 電子情報工学科 卒業
  - 2010年、金沢大学 卒業、2012年 同大学 大学院修士課程 修了
  - 2018年3月、東京工業大学 大学院博士課程 単位取得満期退学
  - 2018年4月、楽天株式会社 入社
- 受賞歴
  - 2015年度 情報処理学会 山下記念研究賞
  - WebDB Forum 2014 最優秀論文賞
  - IDRユーザフォーラム2017 奨励賞 等
- 研究分野：自然言語処理、意見マイニング、評判分析、金融情報学

著書

目次

- 学生時代における楽天データセットを用いた研究
  - 背景・目的
  - どんなデータセットをどのように用いたか
  - IDRユーザフォーラム2017からの進展
- 楽天との「ヒストーリー」（ヒストリー＋ストーリー）
  - どうして楽天データセット？どうして楽天入った？
  - 入社してから現在までの業務
- 学生さんへのメッセージ
  - データセットの利用経験を通して得たもの

意見マイニングにおける条件付き意見の抽出

(1) 意見抽出

評価の妥当性を限定する

ホテルAの立地は観光目的では素晴らしい。

対象　属性　条件　評価表現

ロビーは時刻表がなくて残念

条件

支配人

ロビー案内所にて掲示してございます。

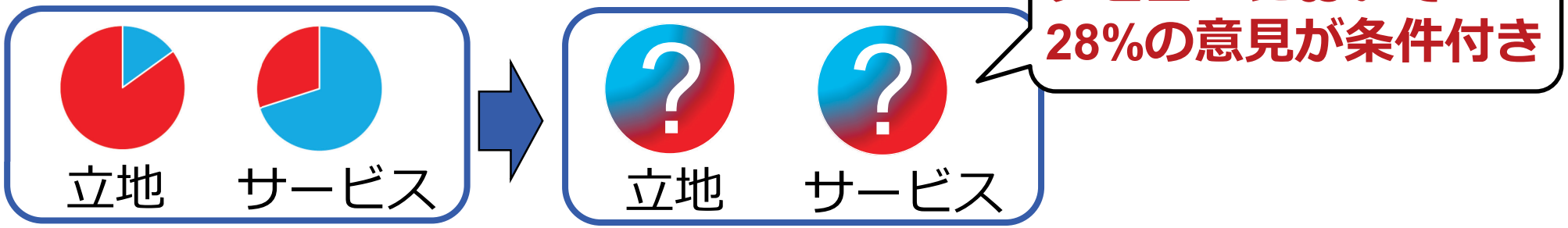
(2) 極性分類

肯定 出張では？

評価

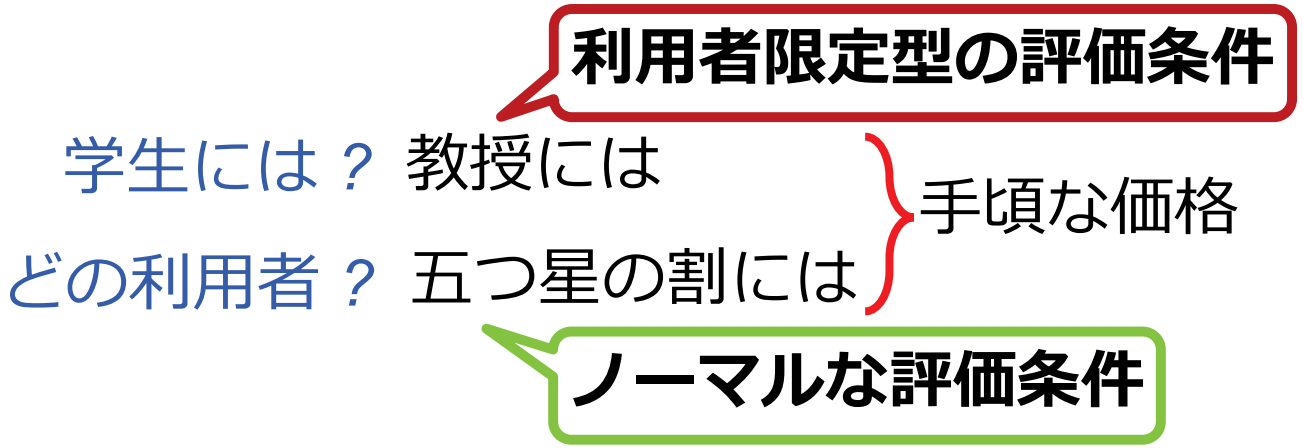
否定 誤解（疑わしい）

(3) 可視化



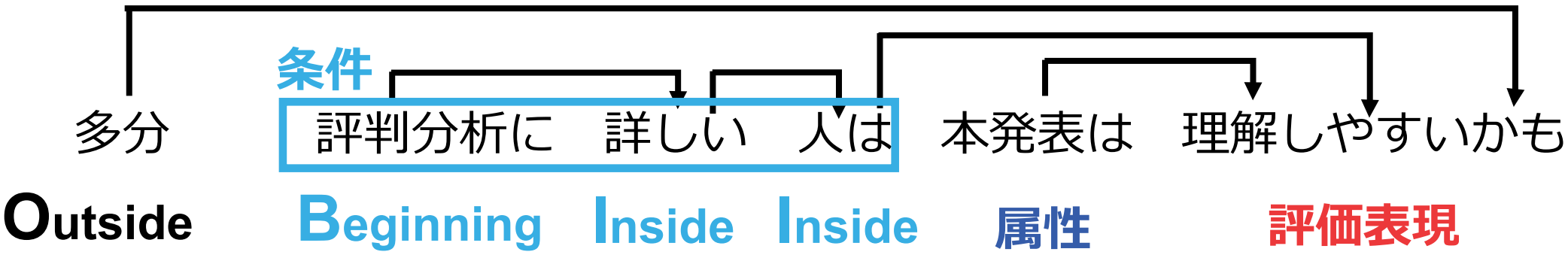
条件 → 評価条件: レビューの場合

- 文法的観点から見た評価条件
  - 節 (出張で利用するなら, 出張で利用したので)、句 (出張には, 出張で)
  - 主語 (意見保持者)
- 意味的観点から見た評価条件
  - 利用者限定型の評価条件: 評価条件の部分集合
    - 利用者の（デモ|サイコ）グラフィック、目的、状況により限定される



研究の目的

1. レビューにおける評価条件の抽出 [Nakayama+ EMNLP2015]



2. 返信文書における評価条件を含む文書の抽出 [IDRユーザフォーラム2017]

どんなデータをどのように使った？IDRユーザフォーラム2017からの進展は？

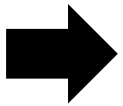
どんなデータセットをどのように用いた？

- 楽天トラベルのレビューデータ約35万件（2010年7月リリース時）
  - 今日現在、約620万件

約17.7倍の増加

- レビュー本文と施設からの回答
  - 人手による学習用＆評価用データの作成（典型的な利用例）

朝食は連泊するときは物足りない



朝食は 連泊する ときは 物足りない

属性　B　I　評価表現

- プランタイトル
- レビューの分類

手法の改善や誤り分析に用いた（以外にも役立った。）

  - 投稿者によって付与されるラベル　感情・情報　or　苦情

どんなデータをどのように用いた？：プランタイトル

【春得】カップル・ファミリーにおすすめ☆お得に札幌ステイ

ユーザ属性 ➡ 手がかりとなる

■ 条件を抽出するための手がかり語として用いた

プラン
タイトル集合 ➡ 手がかり語収集
（手法の詳細は割愛） ➡ 男性 新年会 卒業旅行 春休み 子供連れ 健康志向 …

手法	手がかり語なし	文節内の内容語	手がかり語あり
F値	0.51	0.56	0.58

効果があった

どんなデータをどのように用いた？：レビューの分類

- 「条件を含む返信文書の抽出」における誤り分析の切り口として
  - IDRユーザフォーラム2017からの進展 [中山+ 2018]
  - 背景：レビュー素性の有効性はホテルに依存した →　なぜ？？

レビュー素性追加後に
発生した誤り事例数

施設ID	誤り事例数
a	21
b	11

有効でない

有効

両者の事例には
どんな違いが？

長ったらしく、くどくどと書く苦情が多い

	誤り事例数	
施設ID	苦情レビュー	それ以外
a	10（48%）	11
b	2（18%）	9

[中山+ 2018] 中山 祐輝, 藤井 敦. 宿泊者レビューに対するホテルの返信から何が見えるか?（第12回テキストアナリティクスシンポジウム）

目次

- 学生時代における楽天データセットを用いた研究
  - 背景・目的
  - どんなデータセットをどのように用いたか
  - IDRユーザフォーラム2017からの進展
- 楽天との「ヒストーリー」（ヒストリー＋ストーリー）
  - どうして楽天データセット？どうして楽天入った？
  - 入社してから現在までの業務
- 学生さんへのメッセージ
  - データセットの利用経験を通して得たもの

「楽天」という会社と出会う

2012年9月

中山：条件付き意見の抽出で用いるデータはどうしよう。

F先生：楽天のデータ使ってみたら？多少は信頼性は上がるでしょう。

中山：はい、使ってみます。（へーそんなんあるんや、知らんかった）

某学生が1年前に楽天データを使っていたため、既に研究室にあった。
楽天に入社するなんてこの時は微塵も思っていなかった。

2014年3月

T先生：楽天NYのインターンシップ行ったら？研究発展させられるかもよ。

応募はしたが、色々な理由で辞退した。しかし、東京オフィスの社員さんに研究内容をプレゼンする機会があった。

平手：非常に有用性の高い研究ですね！学会等でお会いすることがあるかもしれませんがよろしくお願いします。

この機会が楽天と深い関係を持つきっかけとなった。

楽天に興味を持ち始める→そうだ楽天へ行こう

8ヶ月後：2014年11月@WebDB Forum

司会：楽天賞は…中山さんです！

平手：弊社が保有しているデータに適用でき、ビジネスへの応用が期待できるという理由で選定しました！

中山：ありがとうございます！（これは楽天に入ってくれという圧力か…）

評価されたことが素直にうれしかった。

楽天に入れば、今までやってきた経験が活かせる。楽天を受けてみよう。

技術やデータの側面だけでなく、英語を話す環境にも興味があった。

学生時代は、留学生に囲まれた環境だったので、英語を話すことには多少は慣れていたのかも。

東工大在籍における研究室のデスク配置

留学生	留学生	日本人	日本人
中山	留学生	日本人	日本人

留学生	留学生	日本人	日本人
留学生	留学生	日本人	日本人

入社からこれまでの業務：朝ごはんフェスティバル

■ 日本一の朝食を提供するホテルとその朝食を決定するイベント

- 楽天トラベルレビューの評判に基づいて決定

評判をスコア化する独自アルゴリズムの開発

■ 入社後にトラベルデータに触れてみて

- 社内データは欠損しているデータがより多い
- データを活かしきれていない
  - ☹アイディアはあるけど、リソース不足
  - ☺学生を巻き込んで研究を行える
- ビジネスに直結する成果を得られる
  - ユーザ（事業側）との関わりを持てる

入社からこれまでの業務：
様々なデータに触れる機会を持てる

Rakuten みん就

https://www.nikki.ne.jp/company/4755/
（閲覧日：2018年11月27日）

Rakuten ブックス

https://books.rakuten.co.jp/rb/15496824/
（閲覧日：2018年11月27日）

目次

- 学生時代における楽天データセットを用いた研究
  - 背景・目的
  - どんなデータセットをどのように用いたか
  - IDRユーザフォーラム2017からの進展
- 楽天との「ヒストーリー」（ヒストリー＋ストーリー）
  - どうして楽天データセット？どうして楽天入った？
  - 入社してから現在までの業務
- 学生さんへのメッセージ
  - データセットの利用経験を通して得たもの

データセットの利用経験を通して得たもの

■ 論文の信頼性向上

- データの出どころがより明確な論文は再現性の観点で優位に立てる
- 難関国際会議に採択された

X社が公開しているyのデータを正解とした。

X社のWebページをクロールし、人手で正解を付与した。

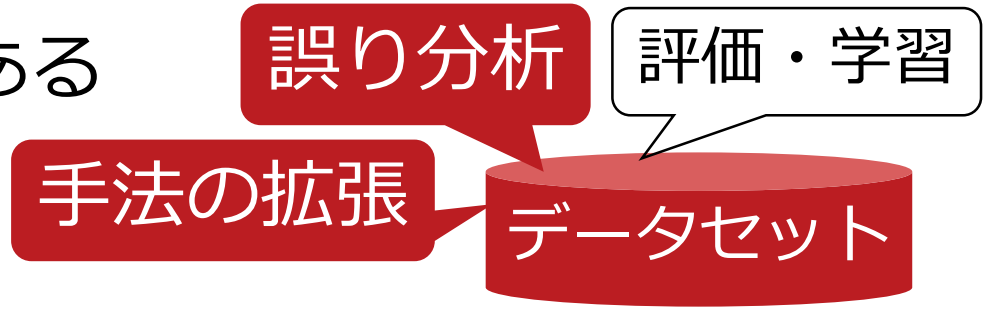
■ 様々な人々と関係を持てる機会が得られた

- 楽天との関わり
- 他の企業さんと議論や会話ができる（e.g., データセットあるある）
  - ホテルによる返信の大部分は 儀礼的な挨拶 or 消極的な対応を主眼とする
- スペインのD学生からデータの提供を依頼された（実際は提供しなかった）
  - 論文引用するからデータ提供してよ

データセットの利用経験を通して得たもの（つづき）

■ 思いもよらない場面で役立つ可能性がある

- 色眼鏡をかけてデータを見ない
- 切り口はいっぱいある



■ 実行可能性の判断や問題設定の手助けとなる

1. 条件付き意見は既存の意見マイニングでは考慮されていない
2. どれほど重要なのか？よく出現するのか？
3. データの調査 → 28%の意見が条件付き → 無視できない

役立った

■ 利用者に適したホテルを推薦できる超能力（できません）


Rakuten

以下からは補足資料

既存の意見マイニングの問題点1：条件付き意見を考慮しない

(1) 意見抽出

ホテルAは良いサービスを提供する。

対象　評価表現　属性

立地は観光目的では素晴らしい。

条件

立地は車で行くと悪いです。

条件

(2) 極性分類

肯定

評価

肯定　出張では？

否定　電車では？

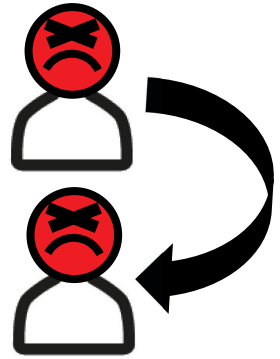
(3) 可視化

既存の意見マイニングの問題点2：レビューに対する返信を考慮しない

ホテルAのロビーは時刻表がなくて残念　否定

支配人：検討させていただきます。

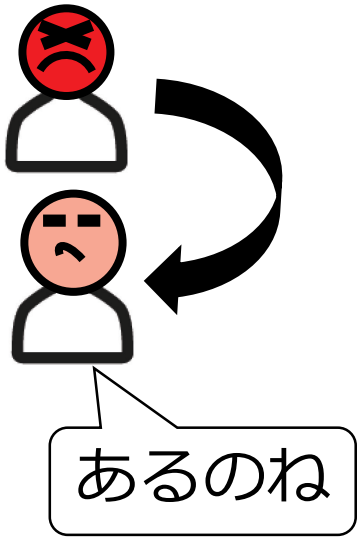
読者の印象



ホテルAのロビーは時刻表がなくて残念　否定　疑わしい

ロビー案内所にて掲示してございます。

評価条件とみなせる



■ 返信が評価の妥当性を限定する場合がある

- 読者の印象が変わる可能性がある

返信における評価条件

誤解の解消

ルームキーが古くてセキュリティ面が心配。

コピーできない特殊なキーですので，ご安心ください

対応できない理由の説明

シャトルバスの案内がわかりにくい。

行政指導により案内表示を行うことができません。

返信における評価条件

解決策の提示

帰宅後、キーを取る時、チェックイン同様
並ばないと行けないのでかなり待たされた。

現時点ではルームキーをお持ちになっての
外出も可能でございますので、次回ご宿泊の
際にご考慮いただければ幸いに存じます。

改善の完了

トイレの鍵が壊れているようでした。

早速改善しました。