楽天グループにおける研究事例の紹介と公開データセットの改良

ぜひご活用ください！

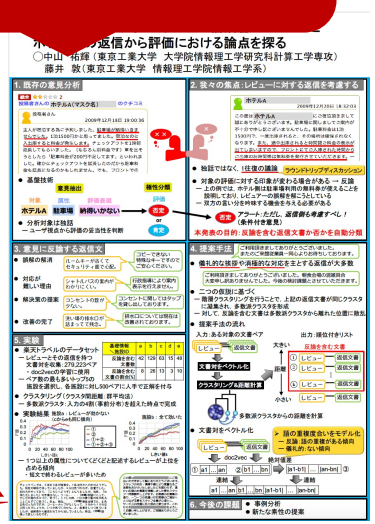
中山 祐輝
楽天グループ株式会社
楽天技術研究所
2023/12/11(月)

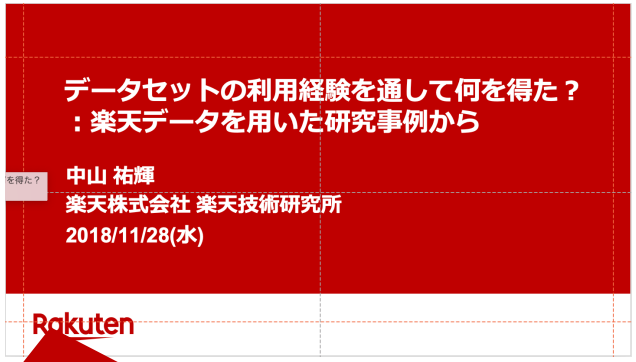
Rakuten

自己紹介：中山 祐輝（なかやま ゆうき）石川県出身

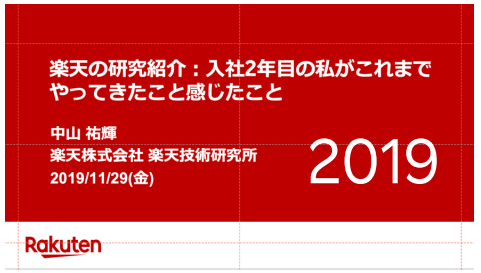
- 自然言語処理のリサーチサイエンティストです（入社6年目）
  - 何でもトライ！（今回は意見分析、曖昧性解消、有害テキスト検出を紹介）
  - YANS2020, YANS2021の運営委員（主にハッカソンを担当）
  - NLP2023, NLP2024のプログラム委員（**NLP2024は第30回記念大会@神戸**）
- 私とIDRユーザフォーラムとの関わり

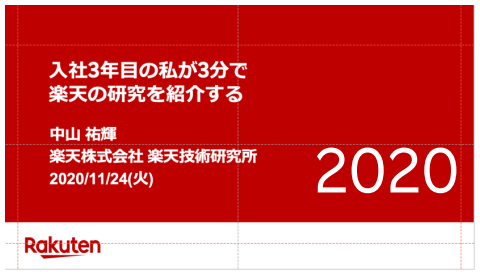
奨励賞受賞（2017）

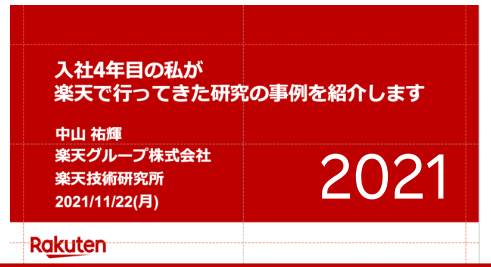


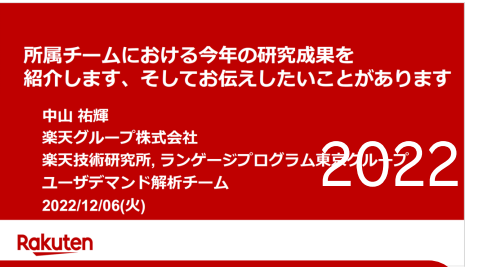


口頭発表セッションで
その後を報告（2018）









「楽天の研究紹介・データセット」と
「働いてみての感想」を発表（2019-）

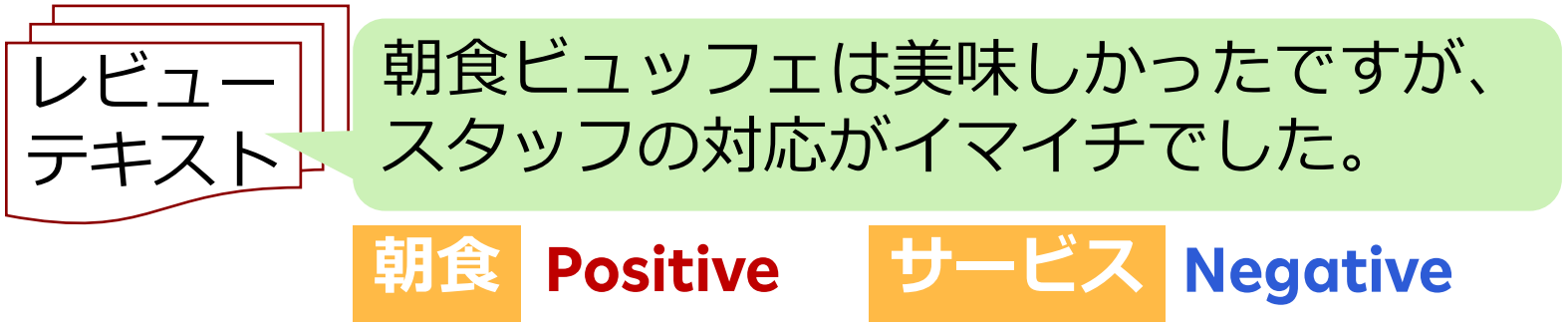
アスペクトベースの意見分析
(ABSA: Aspect-based Sentiment Analysis)

Rakuten Travel

■ これまで：7種類のアスペクトカテゴリと極性を付与した日本語のABSA大規模データセットを構築 [Nakayama+ LREC2022]

- IDRのレポジトリで公開されており、すでにご使用いただいております

72,624文 (12476レビュー)



■ これから：ABSAの高精度化に向けたデータセットの構築を目指す

[Nakayama+ LREC2022] **Yuki Nakayama**, Koji Murakami, Gautam Kumar, Sudha Bhingardive, Ikuko Hardaway: A Large-Scale Japanese Dataset for Aspect-based Sentiment Analysis, In Proceeding of Language Resource and Evaluation (LREC 2022), Online/Marseille, Jun 2022.

ABSAの高精度化に向けたデータセットの構築

英語のベンチマーク

タスク＼データセット	SemEval ’14	SemEval ’15, 16	楽天データ
アスペクトカテゴリ	☑	☑（拡張）	☑
アスペクトカテゴリ極性	☑	☑（拡張）	☑
アスペクト表現（対象、属性）	☑	☑（拡張）	
アスペクト表現極性	☑	N/A	

これから

レビューテキスト

朝食ビュッフェは美味しかったですが、スタッフの対応がイマイチでした。

<朝食ビュッフェ, NULL, **Positive**>
<スタッフ，対応，**Negative**>

<対象, 属性, 極性> の三つ組みをタグ付け
[村上+ NLP2023]

■ タグ付けの課題: ゼロ照応、名詞句、節をなすアスペクト表現の扱い

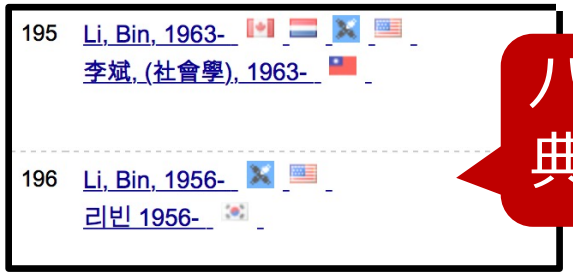
[村上+ NLP2023] 村上 浩司, **中山 祐輝**, Ikuko Hardaway: 日本語意見分析コーパスの構築に向けたアスペクト情報の特徴分析. 言語処理学会 第29回年次大会, 沖縄/オンライン, 2023年3月

書籍の著者名曖昧性解消: 同姓同名の著者名による書籍を個人に区別するクラスタリング

Rakuten ブックス

37人の田中実

国立国会図書館典拠データ検索



バーチャル国際典拠ファイル

- 網羅的に区別されておらず、人手の付与は多大な時間を要する
  - 著者を区別する自動手法が必要 ⇨しかし、評価コーパスがない☹
  - ⇨評価コーパスの構築 [中山+ NLP2020]
- コーパスの改良（リリース予定）
  - データセットを用いた研究求む！
- どのようにチャレンジングなのか？
  - Few shot/Zero shotのタスク, 扱えるデータはメタデータのみ

- 収録著者数を2から最大6に拡大
- 著者名・書籍の数を増やす

カテゴリをまたいで様々な著者とコラボする著者もいる

[中山+ NLP2020] **中山 祐輝**, Sudha Bhingardive, 村上 浩司: 書籍の著者名曖昧性解消における評価コーパスの自動構築. 言語処理学会第26回年次大会, オンライン, 2020年3月

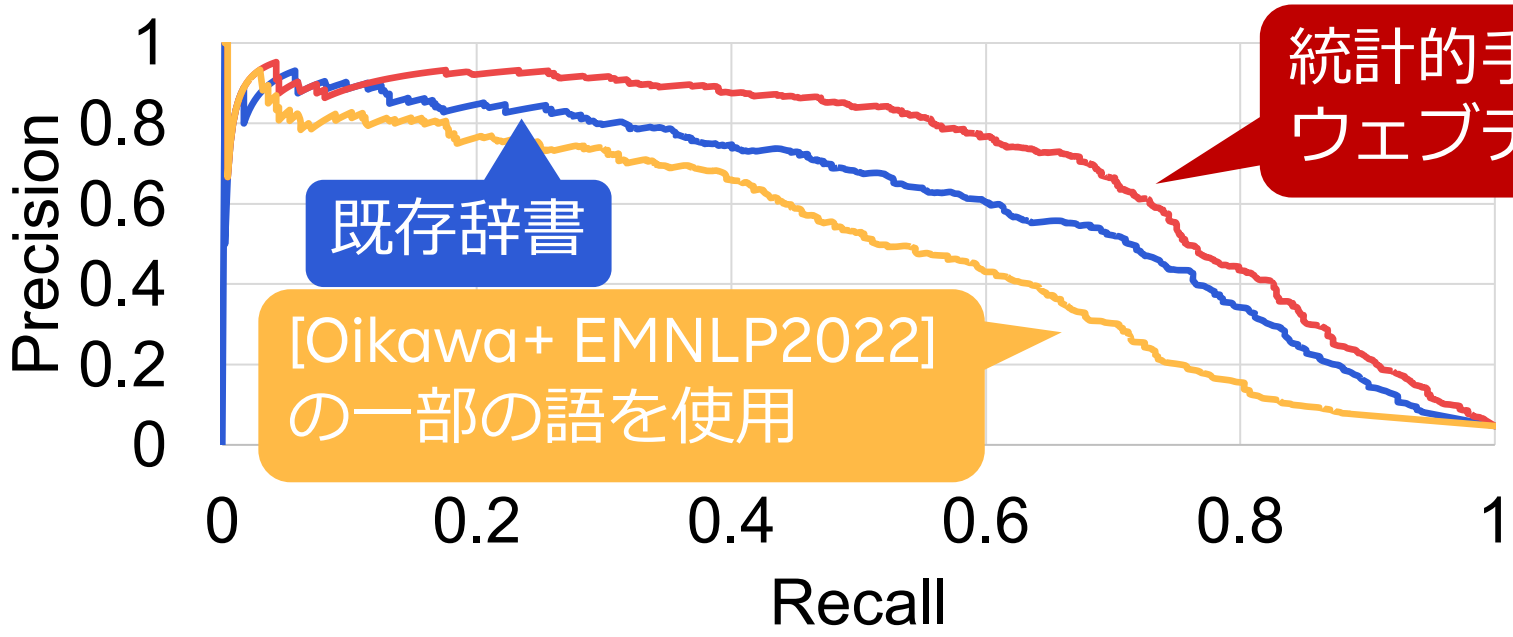
有害テキスト🔪💣💢🔞のフィルタリング

■ ライブ配信チャット内の有害語の検出 Rakuten DRAGON

- オンライン処理と低コストの計算資源を意識した手法 [Oikawa+ EMNLP2022]

■ 大規模言語モデルにおける事前学習データの前処理

- 有害語の割合に基づいてテキストの有害スコアを計算（辞書マッチング）



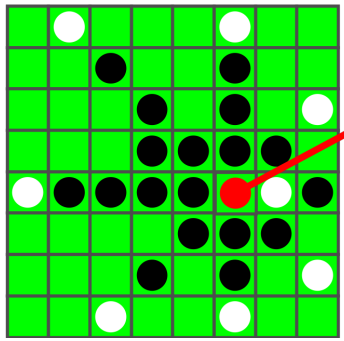
統計的手法を用いて独自に大規模ウェブテキストから有害語を獲得

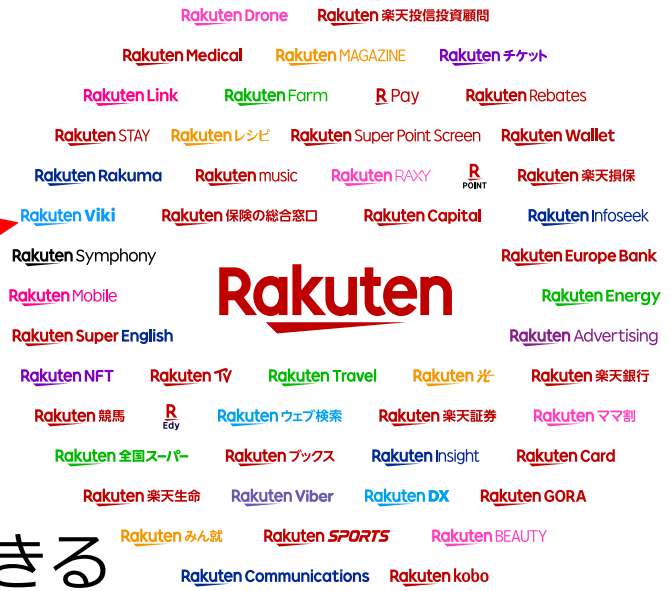
MC4を用いたデータセットによる比較実験

[Oikawa+ EMNLP2022] Yuto Oikawa, **Yuki Nakayama**, Koji Murakami: A Stacking-based Efficient Method for Toxic Language Detection on Live Streaming Chat, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022): Industry Track, Online/Abu Dhabi, December 2022.

おわりに：楽天で働いていての感想・伝えたいこと

- ☺様々なサービスからのデータを扱える
  - 会心の一撃で種々の問題を解決できる可能性を秘めている
  - 様々な部署、従業員とコラボできる





- ☺実践的なビジネス英語を学べる
  - 外国籍従業員（100を超える国・地域から）と仕事ができる
- ☺実サービスに応用できる研究を遂行できる
  - 今回発表した内容は全てサービス改善に利用されています/される予定です
- 公開データセットをぜひご活用ください！
- ポスターセッションでお話しましょう！
  - 自然言語処理以外のその他分野の研究事例も紹介します！


Rakuten