楽天グループにおける研究事例と公開データセットの紹介

中山 祐輝
楽天グループ株式会社
楽天技術研究所
2024/12/13(金)

Rakuten

自己紹介：中山 祐輝（なかやま ゆうき）石川県出身

- 自然言語処理技術を駆使して楽天サービスの改善を行う研究開発
  - 事業貢献（ユーザ体験の向上とトレンド把握）、特許出願、論文執筆
- IDRユーザフォーラムとの関わり
  - 2017（博士課程学生）：奨励賞受賞
  - 2018（入社1年目）：口頭発表セッションでその後を報告
  - 2019-2024: 楽天の研究事例とデータセットの紹介
    - 意見分析のビジネス応用
    - レシピ重複投稿の検出
    - 商品タイトル・説明文からの属性値抽出

ブランド *Apple* *iPad Pro* - *12.9インチ* サイズ

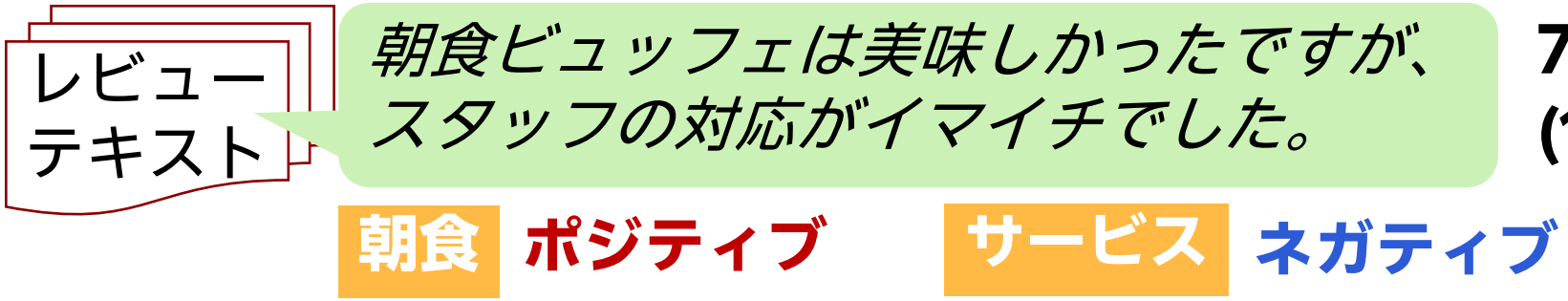
- アスペクトベース意見分析、検索クエリからの属性値抽出、有害テキストのフィルタリング、自社LLM構築

アスペクトベース意見分析（ABSA）

Rakuten Travel

■ 7種類のアスペクトカテゴリと極性を付与した日本語アスペクトベース意見分析の大規模データセットを構築 [Nakayama+ LREC2022]

- TSUKUBAコーパス（文にポジティブ、ネガティブのみ付与）の進化版
- ご使用いただいております（http://doi.org/10.32130/idr.2.14）



72,624文
(12476レビュー)

■ ABSAの現状：ChatGPTで解くのは難しい

- ChatABSA lags far behind the fully supervised baselines [Bai+ EMNLP2024]

[Nakayama+ LREC2022] **Yuki Nakayama**, Koji Murakami, Gautam Kumar, Sudha Bhingardive, and 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.

検索クエリからの属性値抽出 [Nakayama+ NAACL 2024]

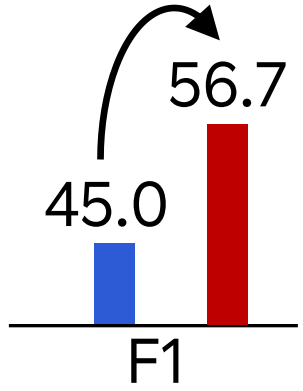
ナイキ□黒□26.5cm　検索　システム　ナイキ□黒□26.5cm
ブランド　色　サイズ

不適切な単語境界をもつ検索クエリ

単語分割

検索クエリの整形

ブランド名抽出



ブリヂ□ストン□アイアン　ブリヂストン□アイアン　ブリヂストン□アイアン

アディダスマスク　アディダス□マスク　アディダス□マスク

[Nakayama+ NAACL2024] **Yuki Nakayama**, Ryutaro Tatsushima, Erick Mendieta, Koji Murakami, and Keiji Shinzato: **Search Query Refinement for Japanese Named Entity Recognition in E-commerce Domain**, *In Proc. of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pp.447–452, Mexico City, Mexico

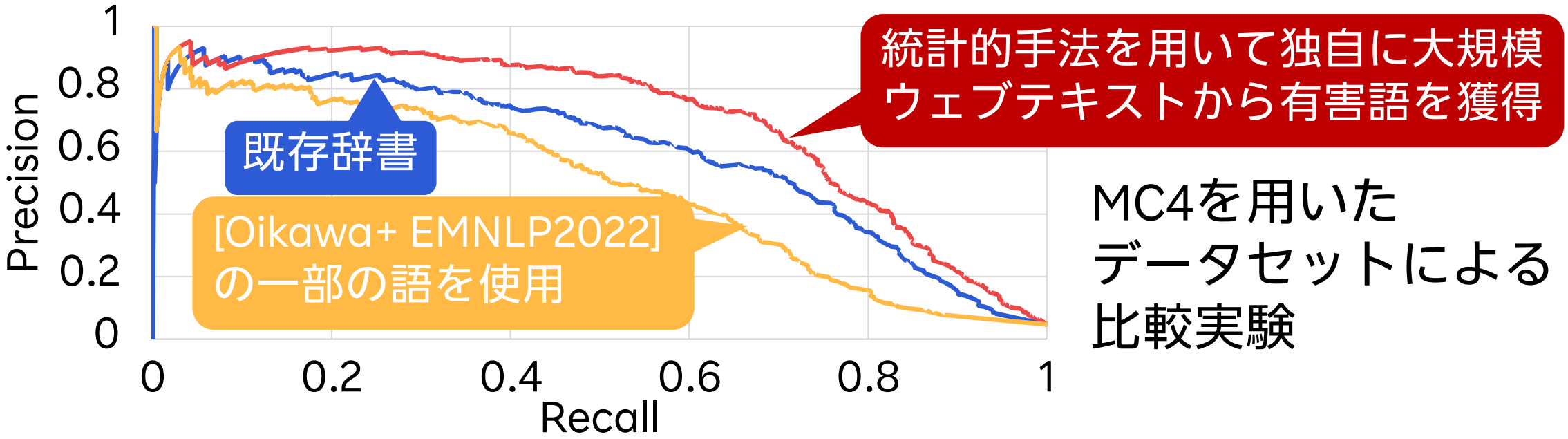
有害テキスト🔪💣💢🔞のフィルタリング

■ 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

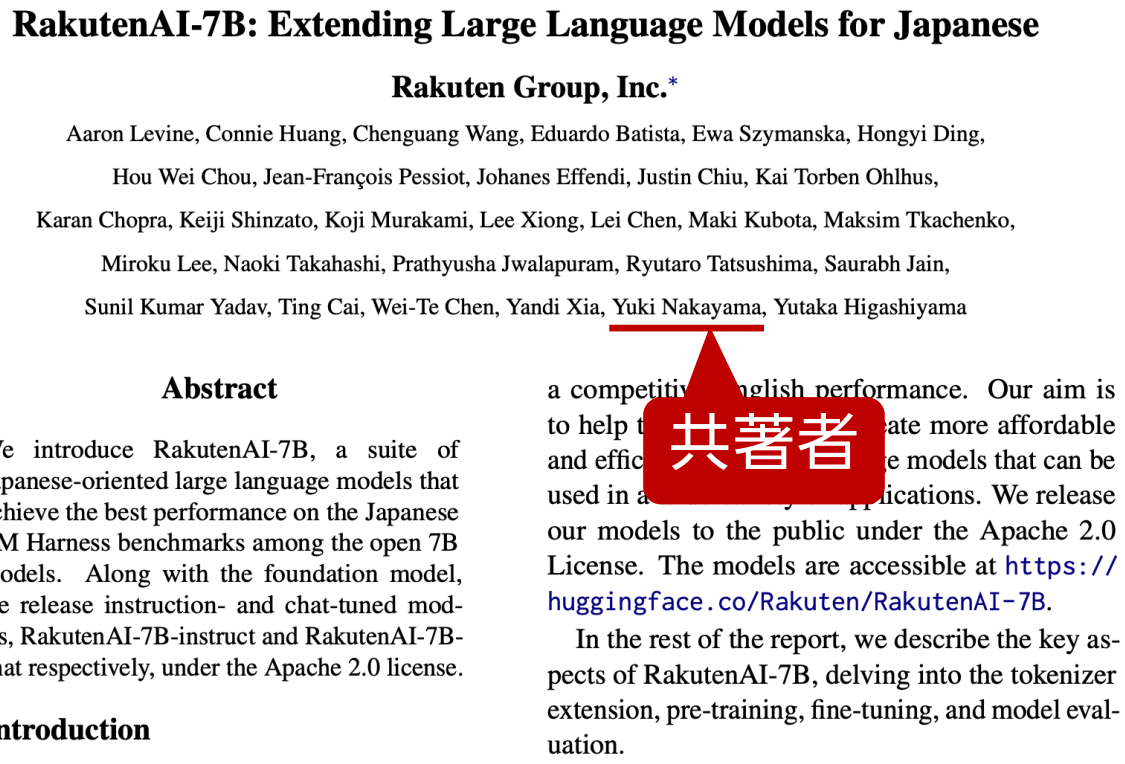
- 有害語の割合に基づいてテキストの有害スコアを計算（辞書マッチング）



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**, In *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022): Industry Track*, Online/Abu Dhabi, December 2022.

自社LLMの構築：
RakutenAI-7B (2024年3月リリース)



RakutenAI-7B: Extending Large Language Models for Japanese

Rakuten Group, Inc.*

Aaron Levine, Connie Huang, Chenguang Wang, Eduardo Batista, Ewa Szymanska, Hongyi Ding,
Hou Wei Chou, Jean-François Pessiot, Johanes Effendi, Justin Chiu, Kai Torben Ohlhus,
Karan Chopra, Keiji Shinzato, Koji Murakami, Lee Xiong, Lei Chen, Maki Kubota, Maksim Tkachenko,
Miroku Lee, Naoki Takahashi, Prathyusha Jwalapuram, Ryutaro Tatsushima, Saurabh Jain,
Sunil Kumar Yadav, Ting Cai, Wei-Te Chen, Yandi Xia, Yuki Nakayama, Yutaka Higashiyama

共著者

Abstract

We introduce RakutenAI-7B, a suite of Japanese-oriented large language models that achieve the best performance on the Japanese LM Harness benchmarks among the open 7B models. Along with the foundation model, we release instruction- and chat-tuned models, RakutenAI-7B-instruct and RakutenAI-7B-chat respectively, under the Apache 2.0 license.

1 Introduction

a competitiv nglish performance. Our aim is to help t ate more affordable and effic e models that can be used in a pplications. We release our models to the public under the Apache 2.0 License. The models are accessible at https://huggingface.co/Rakuten/RakutenAI-7B.

In the rest of the report, we describe the key aspects of RakutenAI-7B, delving into the tokenizer extension, pre-training, fine-tuning, and model evaluation.

https://arxiv.org/pdf/2403.15484
（閲覧日：2024年11月28日）

Rakuten

企業情報　Rakuten Innovation　ニュース　投資家情報　サステナビ

ホーム > ニュース > プレスリリース > 2024 > 楽天、日本語に最適化したオープンかつ高性能なLLMを公開

シェア:　Print

2024年3月21日　楽天グループ株式会社

楽天、日本語に最適化したオープンかつ高性能なLLMを公開

- 「LM Evaluation Harness」の評価基準において、基盤モデルとインストラクションチューニング済モデルがオープンな日本語LLMにてトップを獲得 -

楽天グループ株式会社（以下「楽天」）は、日本語に最適化した高性能の大規模言語モデル（以下「LLM」）

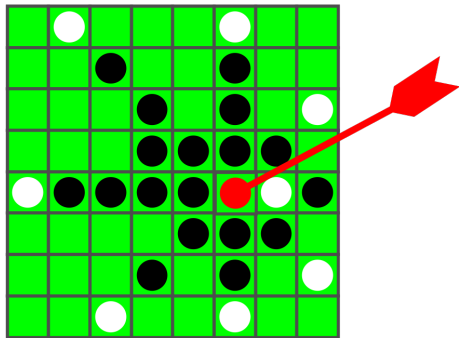
https://corp.rakuten.co.jp/news/press/2024/0321_01.html
（閲覧日：2024年11月28日）

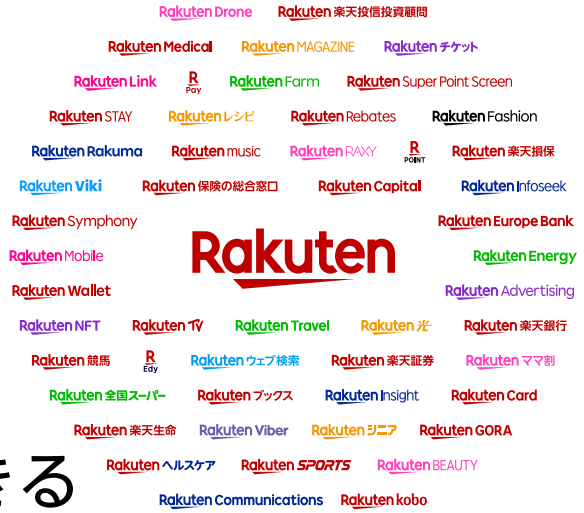
- 日本語英語両方のベンチマークで、東工大（現、東京科学大学）Swallow-7bなどの70億パラメータを持つモデルの中で最も優れた平均スコアを達成（発表当時）

おわりに：楽天で働いていて感じること

■ ☺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

■ 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も紹介します！


Rakuten