

リクルートのAI研究機関、『じゃらんnet』のクチコミを活用した 日本語自然言語処理の発展を加速する学術研究用データセットを公開

株式会社リクルートホールディングスの中間持ち株会社である株式会社リクルート（本社：東京都千代田区、代表取締役社長：北村吉弘、以下リクルート）は、このたび、当社のAI研究機関であるMegagon Labsより、日本語の自然言語処理研究の発展に貢献するため、株式会社リクルートライフスタイル（本社：東京都千代田区、代表取締役社長：浅野 健、以下リクルートライフスタイル）が運営する旅行サイト『じゃらんnet』のクチコミに基づく約12万件のデータを含む学術研究用データセットを、公的研究機関および大学の研究で活用いただくために公開します。

1 日本語の自然言語処理における課題

自然言語処理とは、私たちが日常的に使っている言語（自然言語）をコンピューターに解析させる一連の処理を指し、検索エンジンや機械翻訳、対話システム、顧客の声分析など、生活・ビジネスにおけるさまざまなシーンで利用されています。

日本語の自然言語処理の場合、主要な他言語と比べて、公開されているデータセットの数が圧倒的に少ないという課題があります。日本語の自然言語処理を発展させるために、学術研究用に利用可能なデータセットを充実させることが特に望まれています。

2 公開する学術研究用データセットについて

今回、Megagon Labsが公開するデータセットは、リクルートライフスタイルが提供する旅行サイト『じゃらんnet』上で公開されている宿泊施設へのクチコミと、それを基に加工・作成した文章にラベル付け（アノテーション）を行ったものです。実際に『じゃらんnet』をご利用いただいた多くのお客様による事実の説明や感想、評価などを含む様々な文章が基となっており、幅広い日本語自然言語処理にご活用いただけます。具体的には、以下2つのデータセットをそれぞれのリポジトリにて公開します。

（1）含意関係データセット：

「部屋はオーシャンビューで景色がよかったです。」という文が成り立つとき「部屋から海が見える。」という文も成り立つかどうか（含意関係の有無）というラベルを付与したデータセットです。（約55,000件）
ポジティブ／ネガティブ／ニュートラルといった感情極性ラベルや、宿の特徴が含まれているかどうかというラベルも付与した文も含まれています。（約5,600件）

<https://github.com/megagonlabs/jrte-corpus>

（2）根拠説明データセット：

宿をお薦めする際に、ユーザーの要求に対する推薦根拠を説明するためのデータセットです。（約37,300件）
例えば、「お客様連れに優しい宿」を求めるユーザーに対して、「この宿には大きなキッズスペースがあるので、あなたにオススメです」と根拠付きで返答することを目指しています。文単位で要求に対応する根拠かどうかをラベル付けし、その後、根拠と判断された文を推薦文へと言い換えています。

<https://github.com/megagonlabs/ebe-dataset>

日本語の自然言語処理研究に取り組む研究者にこれらのデータセットを活用いただくことにより、さらなる研究の加速が期待できます。なお、本データセットは学術研究用途での利用を意図したもので、非営利目的でのみ利用可能なライセンス付与を行います。

Megagon Labsは、今後も継続的に学術研究用データセットを公開し、日本語自然言語処理技術の発展に貢献してまいります。本学術研究用データセットにつきまして何かお気づきの点がございましたら、下記お問い合わせ先へご連絡ください。

About Recruit Group

リクルートグループについて

1960年の創業以来、リクルートグループは、就職・結婚・進学・住宅・自動車・旅行・飲食・美容などの領域において、一人ひとりのライフスタイルに応じたより最適な選択肢を提供してきました。現在、HRテクノロジー、メディア&ソリューション、人材派遣の3事業を軸に、45,000人以上の従業員とともに、60を超える国・地域で事業を展開しています。2019年度の売上収益は23,994億円、海外売上比率は約45%になります。リクルートグループは、新しい価値の創造を通じ、社会からの期待に応え、一人ひとりが輝く豊かな世界の実現に向けて、より多くの『まだ、ここがない、出会い。』を提供していきます。詳しくは <https://recruit-holdings.co.jp/> をご覧ください。

本件に関する
お問い合わせ先

発表全般に関するお問い合わせ

<https://www.recruit.co.jp/support/form/>

本学術用データセットに関するお問い合わせ

Megagon Labs 学術研究用データセットページ

(<https://www.megagon.ai/jp/projects/datasets/>)