2019年4月2日
株式会社リクルート

リクルートのAI研究機関、国立国語研究所との共同研究成果を用いた日本語の自然言語処理ライブラリ「GiNZA」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、国立国語研究所との共同研究成果の学習モデルを用いたPython（※1）向け日本語自然言語処理オープンソースライブラリ「GiNZA」（ギンザ）を公開しました。

1．背景

自然言語処理技術は、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。自然言語処理を行うには、言語ごとに異なる語彙や文法体系を保持する言語リソースが必要です。日本語テキストを解析するには、形態素解析（※2）や文節係り受け解析（※3）など複数の機能を統合する必要がありますが、日本語の自然言語処理技術の多くは、機能別に独立したライブラリとして提供されており、エンジニアが個々のライブラリを組み込み、統合する必要があります。また、高度な自然言語処理には品詞体系や文法理論に関する専門知識が要求されるため、文節係り受け解析等を応用に組み込むことができるのは、技術に精通する一部のエンジニアに限定されています。さらに、国際化を前提とするシステムの開発では、リソースファイルを差し替えることで他言語に対応させることが一般的ですが、日本語の自然言語処理技術の国際化対応は形態素解析によるアプローチが中心であり、単語依存構造解析（※4）レベルでの国際化対応への要望が高まっています。

こうした背景のもと、Megagon Labsはエンジニアやデータサイエンティストによる自然言語処理の応用を容易にすることを目指し、Python向け日本語自然言語処理オープンソースライブラリ「GiNZA」を、オープンソースソフトウェアの開発管理などを行うためのプラットフォーム「GitHub」上で公開しました。同時に、国立国語研究所との共同研究成果として、日本語のテキストを高い精度で解析できる「GiNZA日本語Universal Dependencies（UD）モデル」を「GitHub」上で公開しました。

GiNZAのGitHubページはこちらをご参照ください。
https://megagonlabs.github.io/ginza/

2．「GiNZA」の概要

「GiNZA」は、ワンステップでの導入、高速・高精度な解析処理、単語依存構造解析レベルの国際化対応などの特長を備えた日本語自然言語処理オープンソースライブラリです。「GiNZA」は、最先端の機械学習技術を取り入れた自然言語処理ライブラリ「spaCy」（※5）をフレームワークとして利用しており、また、オープンソース形態素解析器「SudachiPy」（※6）を内部に組み込み、トークン化処理に利用しています。「GiNZA日本語UDモデル」にはMegagon Labsと国立国語研究所の共同研究成果が組み込まれています。

「GiNZA」の主な特長は以下のとおりです。

▼「GiNZA」の主な特長

①高度な自然言語処理をワンステップで導入完了

これまで、高度な自然言語処理を行うためには複雑な導入作業が必要でしたが、「GiNZA」はワンステップでモジュールとモデルファイルの導入を完了できます。これにより、エンジニアは即座に解析が可能です。

②高速・高精度な解析処理と依存構造解析レベルの国際化に対応

産業用途で自然言語処理技術を活用するには、一定の処理速度を保ちながら解析精度を高めるためにチューニングを行うことが一般的です。「GiNZA」は、「spaCy」が提供する高速・高精度な依存構造解析器を使用して、産業用途に耐える性能を備えた高度な自然言語処理機能をライブラリとして提供します。同時に、「spaCy」の国際化機能により、複数の欧米言語と日本語の言語リソースを切り替えて使用することが可能となり、エンジニアは複数言語の解析を単一のライブラリで行うことができます。

③国立国語研究所との共同研究成果の学習モデルを提供

自然言語処理系の学会を中心に、人類が用いる多様な言語を、一貫した構文構造・品詞体系で解析可能にする「Universal Dependencies」の取組みが、2014年から全世界で始まっています。日本においても当初からUDの日本語への適用に関する研究と日本語版UDコーパス（データ）構築が同時に進められてきました。Megagon Labsは、国立国語研究所と共同で、日本語版UDに基づいた高精度な依存構造解析技術の研究を行い、その成果である学習済みモデルを「GiNZA日本語UDモデル」に組み込みました。

「GiNZA日本語UDモデル」は、国立国語研究所が長年の研究を通じて蓄積してきた大規模かつ高品質なテキストコーパスに加えて、日本語Wikipediaテキストも同時に用いて機械学習に適用することで、幅広い分野に適応可能なモデルを構築しています。

3．今後の展望

Megagon Labsは今後、オープンソースソフトウェアとして公開した「GiNZA」を、最先端の機械学習技術を取り入れた自然言語処理のフレームワークである「spaCy」の標準言語リソースの一つとして提供する計画です。こうした取組みを通じ、全世界の産業における日本語の自然言語処理応用の促進に貢献していきます。

4．Megagon Labsについて

リクルートのAI研究機関は、2018年4月に「Recruit Institute of Technology」から「Megagon Labs」へ名称変更しました。リクルートグループ各社と連携し、イノベーティブなサービス創出につながる自然言語処理・データマネジメント・機械学習などの新技術の研究開発に取組んでいます。

※1　Python: プログラミング言語の一つで、シンプルで記述力の高い言語として人気があります。
データサイエンス領域だけでなく、ウェブアプリケーション開発等でも広く利用されています。

※2　形態素解析: 自然言語のテキストを、言語として意味を持つ最小単位（形態素）に分割し、それぞれの品詞を推定する工程のことです。日本語ではMeCab等のオープンソースライブラリが広く利用されています。

※3　文節係り受け解析: 構文解析の一種で、文を構成する文節間の修飾・被修飾の関係を解析する工程です。
顧客の声分析で良い／悪いなどの評価の対象を特定するために利用されています。

※4　単語依存構造解析: 構文解析の一種。日本語の文節係り受け解析と比較して、形態素に相当するトークンを単位として前後両方向への依存関係を扱い、依存関係にラベルを付与することで、主語や目的語といった文法的関係を出力する点などが異なります。

※5　spaCy: ExplosionAI GmbHが開発する最先端の機械学習技術を取り入れた高機能な自然言語処理フレームワーク。

※6　SudachiPy: 株式会社ワークスアプリケーションズの自然言語処理研究に特化したAI研究機関「ワークス徳島人工知能NLP研究所」が開発するオープンソースソフトウェア。

リクルートグループではこれからも、働く、学ぶ、住む、結婚、旅、車、飲食、美容など、
さまざまな場面でユーザーが新しい発見・機会創出できるサービスを提供し、
一人ひとりにあった「まだ、ここにない、出会い。」を届けることを目指していきます。

【本件に関するお問い合わせ先】
https://www.recruit.co.jp/support/form/