Press Release

2021 年 8 月 26 日

リクルートの AI 研究機関、Transformers 事前学習モデルを構築し解析精度を向上した日本語自然言語処理ライブラリ「GiNZA version 5.0」を公開

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、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として、Transformers 事前学習モデルを用いることで、解析精度を大幅に向上させた日本語自然言語処理オープンソースライブラリ（以下、OSS ライブラリ）「GiNZA version 5.0」を無料公開しました。

1. 日本語自然言語処理 OSS ライブラリ「GiNZA」について

自然言語処理技術とは、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処理させる一連の技術を指し、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。

リクルートの AI 研究機関・Megagon Labs が開発・提供する「GiNZA」は、機械学習を利用した日本語の自然言語処理に関心があり解析を行いたいと考えている研究者やエンジニア、データサイエンティストに向けて開発された、無料で利用可能なライセンスの下で公開された OSS ライブラリです。ビジネスの現場で広く利用されることを想定し、ワンステップでの導入や高速・高精度な解析処理、単語依存構造レベルの国際化対応などの特長を備えています。「GiNZA」を使えば、構文構造の解析から、人名・組織名・地名・商品名・数値といった固有表現抽出まで統合的に解析でき、また、日本語文法に根ざした、日本語の文節を単位とする解析結果が容易に得られます。

2020 年 1 月の version 3.0 公開以降、19 ヵ月で 10 万ダウンロードを超え、Universal Dependencies(UD、※1)の日本語解析系として、学術機関だけでなく、頑健かつ柔軟な応用が可能な実用的ライブラリを望む産業界の多くの方々にご利用いただいています。Megagon Labs は今後も「GiNZA」をアップデートしていくことで、より速く、高精度な日本語の自然言語処理を可能にし、あらゆる産業において自然言語処理の活用が促進される世界を目指し、研究活動を進めてまいります。

★「GiNZA」公開ページ ：https://megagonlabs.github.io/ginza/

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、一人一人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、4 万 6,000 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。2020 年度の売上収益は 2 兆 2,693 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、一人一人が輝く豊かな世界の実現に向けて、より多くの『まだ、ここにない、出会い。』を提供していきます。

詳しくはこちらをご覧ください。
リクルートグループ：https://recruit-holdings.co.jp/　　リクルート：https://www.recruit.co.jp/

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

2. 「GiNZA version 5.0」アップデートの主な特長

（1） 20 億文以上の Web テキストで事前学習を行った Transformers モデルを用いて解析精度を飛躍的に向上

・大規模テキストで事前学習した Transformers モデルを独自に構築

近年、多くの自然言語処理タスクで最高精度記録を更新し続けている Transformers モデルの多くは、大量のテキストデータによる単語の穴埋め問題(Masked Language Model)を事前学習タスクに用いています。Megagon Labs は、インターネット上の大量のテキストを収集した mC4 データセット(※2)から抽出した日本語テキスト 20 億文以上を利用して、広範な分野をカバーする Transformers 事前学習モデルを独自に構築しました。Transformers モデルには事前学習効率が高い ELECTRA(※3)を、そのトークン化処理には日本語 Universal Dependencies と同じ国立国語研究所 UniDic 短単位をベースとする SudachiTra(※4)を、それぞれ採用しました。構築した Transformers 事前学習モデルは「transformers-ud-japanese」(※5)として別途公開します。

・処理パイプラインへの Transformers モデルの組み込み

「GiNZA」が使用する Python(※6)向け自然言語処理フレームワーク spaCy(※7)では、2021 年 1 月にリリースされた version 3 での機能拡張により、処理パイプラインへ Transformers モデルを容易に組み込むことができるようになりました。「GiNZA version 5.0」では spaCy の処理パイプラインの最前段に組み込んだ「transformers-ud-japanese」から得られる単語(サブワード)の意味ベクトル表現を用いることで、後段の依存構造解析・固有表現抽出・品詞推定の精度を大幅に向上することができました。

・国立国語研究所との共同研究成果の依存構造解析モデルを提供

2014 年から全世界で取り組みが始まった「Universal Dependencies」は、人類が用いる多様な言語を、一貫した構文構造・品詞体系で分析可能にすることを目的とする取り組みです。日本においても、当初から、Universal Dependencies 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開発、および、日本語版 UD コーパス中の固有表現への正解ラベル付与などの取り組みを行い、これらの成果を組み込んだ「GiNZA 日本語 UD モデル」を公開しています。

「GiNZA version 5.0」で使用する「GiNZA 日本語 UD モデル」は、国立国語研究所の大規模かつ高品質な「現代日本語書き言葉均衡コーパス」を Universal Dependencies 体系に変換した UD_Japanese-BCCWJ r2.8 と、広範囲なインターネット上のテキストで事前学習された「transformers-ud-japanese」を組み合わせて依存構造解析モデルの学習を行うことで、幅広い分野に適応可能な解析モデルを構築しています。

（2） 用途に応じて複数の解析モデルを提供

Transformers モデルは解析精度を大幅に向上できる反面、計算量の増大により処理速度が低下するデメリットがあります。「GiNZA version 5.0」では解析精度重視、または、処理速度重視のように用途に応じてモデルを切り替えて使用することができます。提供するモデルは次の 2 種類です。（Python 3.6 以上と対応する pip 環境が必要です。GiNZA の過去のバージョンをインストール済みの場合は事前にアンインストールしてください。）

解析精度重視モデル（ja-ginza-electra）

インストールコマンド： pip install -U ginza ja-ginza-electra

処理速度重視モデル（ja-ginza）

インストールコマンド： pip install -U ginza ja-ginza

※1 全世界の多様な言語を一貫した文法・品詞体系で解析可能にすることを目指した国際的学術プロジェクト

※2 インターネット上のテキストを収集した Common Crawl テキストデータセットに対して、Google が開発したフィルタを適用して構築した多言語テキストデータセット(事前学習には mC4 の日本語テキスト全体をさらに文らしさで絞り込んだ約 20 億文を使用)

※3 ICLR2020 で Stanford 大学と Google Research が発表した敵対的学習を模した機構で事前学習効率を大幅に向上した Transformers モデル(学習用ライブラリには NVIDIA の DeepLearningExample の TensorFlow2 による実装を、解析フレームワークには Hugging Face の transformers をそれぞれ使用)

※4 株式会社ワークスアプリケーションズ・エンタープライ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 Hugging Face Transformers 向けトークナイザライブラリ（形態素解析器として GiNZA と同じ SudachiPy を使用）

※5 Hugging Face Inc.が公開する機械学習モデル共有リポジトリ Hugging Face Hub から公開中(https://huggingface.co/megagonlabs/)

※6 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などでも広く利用されています

※7 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