PRESS RELEASE

報道関係者各位

2022年2月25日
株式会社ワークスアプリケーションズ

ワークスアプリケーションズのAI研究機関が産学共同研究成果を発表
ー大規模日本語事前学習モデル「chiTra(チータラ)」を無償公開し、自然言語処理を活用したDXを支援ー

株式会社ワークスアプリケーションズ（本社：東京都千代田区、代表取締役最高経営責任者：秦修）は、ワークスアプリケーションズ・グループのAI研究機関であるワークス徳島人工知能NLP研究所が、大学共同利用機関法人人間文化研究機構国立国語研究所（以下「国語研」）と産官の共同研究を実施し、国語研が保持する国内最大規模の日本語データベース「国語研日本語ウェブコーパス（NWJC）」と、ワークス徳島人工知能NLP研究所の形態素解析器「Sudachi（スダチ）」を用いて学習した、大規模日本語事前学習モデル「chiTra(チータラ)」を開発し、本日オープンデータとして無償公開したことをお知らせします。

１．公開の背景
ワークス徳島人工知能NLP研究所は、これまでにも大規模日本語形態素解析辞書「SudachiDict（スダチディクト）※1」、大規模日本語単語分散表現「chiVe（チャイブ）※2」等、日本語処理のための大規模な言語資源を無償公開しています。これらは曖昧性が多い日本語を正確に処理させるための言語リソースとして、ワークスアプリケーションズ・グループのみならず、幅広い企業や組織・団体においてビッグデータ活用における情報検索精度の向上やテキスト分類の精度向上に利用されています。

今後のさらなる自然言語処理研究の発展やビジネスへの利活用の促進に向けて、大規模日本語事前学習モデル「chiTra」を開発し、商用利用可能なライセンスで無償公開いたしました。

公開先URL：https://github.com/WorksApplications/SudachiTra

２．大規模日本語事前学習モデル「chiTra」とは
日本語事前学習モデルは、大規模なデータを用いて言語の学習を行い、単語の次にくる単語の予測を可能にすることで、AIによる文章理解の精度を向上させることができます。汎用性が高く、様々な自然言語処理に応用が可能なため、近年注目されている技術ですが、大規模なモデルをゼロから独自に学習させるのは非常

にコストがかかります。

今回無償公開される「chiTra」は、国語研が有する国内最大規模の「国語研日本語ウェブコーパス（NWJC）※3」を、国内最大規模の語彙をもつ「Sudachi※4」を用いて解析させ、BERT（Bidirectional Encoder Representations from Transformers）※5により学習した、大規模かつ実用的な日本語事前学習モデルであり、より手軽に高度な自然言語処理の実現ができるようになります。

大規模日本語事前学習モデル「chiTra」の特長

■ 多様な文書へ対応
国内最大規模の国語研日本語ウェブコーパス（NWJC）を訓練データに用いることで、多様な表現、様々なドメインの文書に対応しています。

■ 多様な語彙への対応
日本語では「引っ越し」「引越し」のように同一の語が多様な表記で表現されますが、こういった表記ゆれが事前学習モデルにも悪影響を及ぼします。chiTraでは形態素解析器Sudachiの豊富な語彙と表記正規化機能を利用して表記ゆれによる弊害を抑えています。

■ 使いやすいパッケージング
自然言語処理のためのディープラーニングフレームワーク「Hugging Face」※6に対応しており、様々なNLPタスクでスムーズに利用することができます。またchiTraのモデルはOpen Data on AWS※7で公開しているので簡単に入手することが可能です。

今後は日本語の性質をより捉えたモデルとするために、以下のような課題にも取り組み、継続的に更新していきます。

■ 日本語の書記法に適したトークナイズ
単純な文字ベースの分割ではなく、Sudachiの複数粒度分割の機能を利用して日本語の語構成や字種を考慮したサブワード化により、より日本語の書記法に適したモデルの構築を目指します。

■ 多様な表現への対応
Sudachi同義語辞書の情報を利用することで訓練データに現れない多様な表現を補い、機械学習と人手による知識の融合を試みます。

３．有償サポートサービスについて

ワークス徳島人工知能NLP研究所が提供するオープンソースソフトウェアおよび言語資源の有償保守サービスと、それらを利用した自然言語処理活用のためのコンサルティングサービスを提供しています。検索精度向上実現のためのSudachiDictやchiVe活用のコンサルティング等、「chiTra」を含めた自然言語処理のビジネスへの利活用をサポートしています。

有償サポートサービスに関するお問い合わせ先：
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　SaaS事業本部
E-Mail ： bizapp@worksap.co.jp

※1 大規模日本語形態素解析辞書 SudachiDict
専門家の手によりメンテナンスされた約300万語の語彙を収録する高品質な日本語自然言語処理のための辞書。継続的な拡充をすることにより、「送り仮名」「略語」「旧漢字」「同義語」「誤表記」にも対応
https://github.com/WorksApplications/SudachiDict
※2 大規模日本語単語分散表現 chiVe
大規模な単語分散表現リソースで、国内最大258億語規模のコーパスにて学習を実施
https://github.com/WorksApplications/chiVe
※3 国語研日本語ウェブコーパス（NWJC）
ウェブ上の日本語テキストから100億語を超える規模のサンプルを収集することで稀言語現象の言語学

的、心理学的および情報処理的視点からの究明の可能性を開くことを目的に構築された大規模コーパス

※4 Sudachi
https://github.com/WorksApplications/Sudachi

※5 BERT（Bidirectional Encoder Representations from Transformers）
2018年にGoogleから発表された自然言語処理のための機械学習手法。様々な自然言語処理タスクにおいて当時の最高精度を更新し話題となった

※6 Hugging Face
Hugging Face社が提供している自然言語処理に特化したディープラーニングフレームワーク
https://huggingface.co/

※7 Open Data on AWS
公益価値のある一般公開用データをAWSがホストするスポンサーシップ・プログラム
https://aws.amazon.com/jp/opendata/

【ワークス徳島人工知能NLP研究所について】

2017年2月に徳島県に開設した自然言語処理（NLP）に特化した研究機関です。人工知能、特に自然言語処理を活用した業務効率化・生産性向上を実現し、新しい働き方を提案するための研究開発を進めています。

研究成果はワークスアプリケーションズ・グループが開発するERPパッケージソフト「HUE」、SaaS製品「HUE Works Suite」等で活用されています。また、一部の成果はオープンソースソフトウェアとして商用利用可能なライセンスで公開しており、多くの企業、研究機関で活用されています。

【ワークスアプリケーションズ・グループについて】

ワークスアプリケーションズ・グループは、1996年の創業以来、日本発の業務アプリケーションのパッケージソフトウェア会社として、主に国内の大手企業向けに製品・サービスを提供してまいりました。「働く」の概念を変え、仕事をより創造的なものへ、企業の生産性を高め、企業価値を拡大する、この企業理念のもと、ERPを軸としたソリューションプロバイダーとして、大手企業に加えて中堅・中小・スタートアップ企業のDX推進のパートナーとなれるよう、さらなる発展を目指していきます。

株式会社ワークスアプリケーションズ サイト https://www.worksap.co.jp/

■本件に関するお問い合わせ先
TEL：03-3512-1400　FAX：03-3512-1401　Email：pr@worksap.co.jp
株式会社ワークスアプリケーションズ　広報担当