報道関係各位

Press Release

効く、AI を。

Laboro

オーダーメイド AI ソリューション、
『カスタム AI』開発

株式会社 Laboro.AI

2020 年 4 月 17 日

オリジナル日本語版 BERT モデルを公開

～　260 万超の Web ページからテキスト情報を事前学習　～

株式会社 Laboro.AI は、近年 AI 自然言語処理の分野で注目を集めるアルゴリズム BERT を独自に事前学習させた日本語版モデルを開発し、オープンソースとして公開いたしました。

＜今回のポイント＞

✔ 約 4,300 の Web サイト、**計 260 万以上**の Web ページのテキスト情報を学習
✔ 既存に公開されている日本語版モデルと並んで**高い精度結果**を確認
✔ AI による文章分類や質問回答など、**自然言語処理分野での活用可能性**

株式会社 Laboro.AI
代表取締役 CEO　椎橋徹夫・代表取締役 CTO　藤原弘将

オーダーメイドによる AI・人工知能ソリューション『カスタム AI』の開発・提供およびコンサルティング事業を展開する株式会社 Laboro.AI（ラボロエーアイ、東京都中央区、代表取締役 CEO 椎橋徹夫・代表取締役 CTO 藤原弘将。以下、当社）は、研究開発の一環として、近年 AI の自然言語処理領域で注目を集めるアルゴリズム BERT（Bidirectional Encoder Representations from Transformers）を、独自に収集した Web テキスト情報をもとに事前学習させたオリジナル日本語版モデルを開発し、オープンソースとして公開いたしました。

このモデルは、約 4,300 の Web サイト、計 260 万以上の Web ページに掲載されていたテキスト情報を独自に収集したコーパス（言語データベース）を用いて事前学習させたもので、当社で行った文章分類などの検証結果では、一般的なデータに基づくモデルの精度と並んで高い性能を持つことが確認でき、この度、広く公開させていただくことといたしました。

当社では今後も、AI に関わる各種技術領域での研究開発に取り組んでいくほか、機械学習技術を用いたオーダーメイド AI ソリューション『カスタム AI』をより多くの産業の企業様に導入いいただくことを目指すとともに、イノベーション創出のパートナーとして、引き続き精進してまいります。

＜ - 背景 -　AI 自然言語処理と BERT ＞

AI（機械学習）の技術領域のひとつである自然言語処理は、人が日常的に使用する言葉や文字など、テキスト情報を AI に処理させる分野です。手書き文字の読み取りを行う OCR やテキストでの会話を実現するチャットボットのほか、近年普及しているスマートスピーカーにもこの自然言語処理技術が活用されており、AI 活用の主要領域のひとつと言えます。

2018 年 10 月に Google が発表した自然言語処理モデル BERT（Bidirectional Encoder Representations from Transformers）は、この自然言語処理に大きなブレイクスルーをもたらしたと言われる自然言語処理アルゴリズムです。それまでのものと比較して BERT は、

・文章の文脈を理解することに長けている
・文章分類や感情分析など様々なタスクに応用できる（ファインチューニング）
・インターネット上にある大量のデータから事前学習でき、データ不足を課題としにくい

などの画期的な特徴がある上、実際に様々な検証で高い精度を示すアルゴリズムであることが証明されています。

※BERT については、学術研究論文"*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*"（https://arxiv.org/pdf/1810.04805.pdf）などで、詳細を確認いただけます。

＜ - 開発内容 -　Laboro.AI 日本語版 BERT モデルについて ＞

上記のような優れた特徴をもつ自然言語処理技術である BERT を、日本の多様なビジネスシーンでも活用いただくため、今回 Laboro.AI では、主に英文への対応が中心であった BERT を日本語の文章にも対応できるよう、またより精度の高い処理を実現できるよう研究開発を行い、この度、独自の BERT 事前学習モデル（以下、Laboro.AI BERT モデル）を開発し、オープンソースとして広く公開することといたしました。

Laboro.AI BERT モデルは、インターネット上で公開されているニュースサイトやブログなど、フォーマルなものからインフォーマルなサイトまで、計 4,307 の Web サイト、ページ数にして 2,605,280 ページに掲載されているテキスト情報を収集し、事前学習させたものです。Google が公開したオリジナルの英語版 BERT が 13GB 分の英語文献データセットで学習させているのに対して、Laboro.AI BERT モデルもほぼ同量の 12GB に相当する日本語の言語情報データで学習を行なっており、当社で実施した検証※でも高い精度でのタスク処理が可能であることを確認いたしました。

※Laboro.AI BERT モデルの性能評価やその検証内容については、別紙をご覧ください。

＜ -今後の展開-　Laboro.AI BERT モデルの活用可能性 ＞

Laboro.AI BERT モデルは、現在も AI 活用が積極的に行われている次のようなシーンでのタスク処理の精度をより高めることが期待されます。

・社内に大量に蓄積された文書の整理や分類

・専門的なキーワードやそれに類似するワードを含む文書、メールなどテキストデータの分類
・チャットボットなど、テキスト情報をベースにした Q&A システムへの応用
・スマートスピーカー等、声による入力・出力など、音声技術への応用

また、当社はオーダーメイドによる AI「カスタム AI」の開発を主力事業としており、様々な業界・企業様との AI プロジェクトで今般の研究開発の成果を活かしてまいります。

＜ Laboro.AI BERT モデルのご利用について ＞

Laboro.AI BERT モデルは、国際的な著作権ライセンスであるクリエイティブコモンズの CC BY-NC 4.0（Attribution-NonCommercial 4.0 International）の下で利用いただくことができ、**非商用目的に限り無料で公開**しております。利用方法およびダウンロードは、弊社 Web サイト（https://laboro.ai/column/laboro-bert/）にてご確認いただけます。

商用目的での利用をご希望の方は、当社ホームページのお問い合わせフォーム（https://laboro.ai/contact/other/）よりご連絡ください。

株式会社 Laboro.AI について

(株)Laboro.AI は、「効く、AI を」をコンセプトに、オーダーメイドの AI ソリューション「カスタム AI」の開発・提供を事業とし、アカデミア（学術分野）で研究される先端の AI・機械学習技術のビジネスへの実用化をミッションに掲げています。業界に隔たりなく、様々な企業のコアビジネスの改革を支援しており、その専門性から支持を得る国内有数の AI スペシャリスト集団です。

＜会社概要＞

社　名：株式会社 Laboro.AI（ラボロ エーアイ）
事　業：機械学習を活用したオーダーメイド AI 開発、
　　およびその導入のためのコンサルティング
所在地：〒104-0061 東京都中央区銀座 8 丁目 11-1
　　GINZA GS BLD.2 3F
代表者：椎橋徹夫（代表取締役 CEO）
　　藤原弘将（代表取締役 CTO）
設　立：2016 年 4 月 1 日
ＵＲＬ：https://laboro.ai/

＜本リリースに関するお問い合わせ＞
株式会社 Laboro.AI リードマーケター 和田 崇
Mail：press@laboro.ai　Tel：03-6280-6564（代表）
※昨今の新型コロナウイルス感染拡大の状況を鑑み、当社では、当プレスリリース発信時点で原則、全社的なリモートワーク体制を敷いております。当社代表電話番号へのご連絡をお受けできない可能性がございますため、その際は、メールにてご連絡いただけますようお願いいたします。関係各位にはご不便をお掛けいたしますが、何卒ご理解賜れますようお願い申し上げます。

【別　紙】

＜ Laboro.AI BERT モデルの精度評価 ＞

Laboro.AI BERT モデルの性能を評価するため、今回、以下２つのタスクで検証を行いました。

＜ タスク（A）　文章分類 ＞

NHN Japan 株式会社が収集し、クリエイティブ・コモンズライセンスのもと公開している livedoor ニュースのコーパス※を用い、特定のニュース記事を９つのカテゴリー（トピックニュース、Sports Watch、IT ライフハック、家電チャンネル、MOVIE ENTER、独女通信、エスマックス、livedoor HOMME、Peachy）に正しく分類できるかを検証・評価しました。

※livedoor ニュースコーパスについてはこちらをご覧ください。http://www.rondhuit.com/download.html#ldcc
※livedoor は NHNJapan 株式会社の登録商標です。

＜ タスク（B）　質問回答 ＞

与えられた文章の中から質問に対する答えを抽出・回答するタスクで、正しい回答ができるかの精度を評価しました。今回は「運転ドメイン QA データセット※」という、インターネット上で公開されている運転に関するブログ記事を元に構成されたデータセットのうち、文章読解のための Q&A データセットである「RC-QA データセット」というものを引用しています。例えば、

・文章　：私の車の前をバイクにまたがった警察官が走っていた。
・質問　：警察官は何に乗っていた？
・答え　：バイク

といった一群がセットになっています。

※「運転ドメイン QA データセット」は、京都大学大学院 情報学研究科 黒橋禎夫教授・河原大輔准教授・村脇有吾助教 研究室が公開するものです。詳しくはこちらをご覧ください。http://nlp.ist.i.kyoto-u.ac.jp/index.php?Driving%20domain%20QA%20datasets

＜ 精度評価 ＞

上記の２つのタスクそれぞれについて、以下の３つのモデルでその精度を比較しました。

① 公開されている日本語版 Wikipedia のコーパスを事前学習させたモデル※
② Laboro.AI BERT Base モデル　（12 層、ハイパーパラメーター数 110M）
③ Laboro.AI BERT Large モデル（24 層、ハイパーパラメーター数 340M）

複数回の検証結果を平均した比較表がこちらの次表です。

	コーパスサイズ (corpus size)	タスク（A） 文章分類の正解率 (accuracy)	タスク（B） 質問回答の一致率 (exact match)
① 日本語版 Wikipedia モデル	2.9GB	97.2%	76.3%
② Laboro.AI BERT Base モデル	12GB	97.7%	75.5%
③ Laboro.AI BERT Large モデル	12GB	98.1%	77.3%

タスク（A）文章分類・タスク（B）質問回答ともに、いずれのモデルも僅差で高い精度を示している中、③ Laboro.AI BERT Large モデルがとくに高い結果を示していることが確認できました。

※日本語版 Wikipedia のコーパスを事前学習させたモデルとしては、「BERT with SentencePiece for Japanese text」(Yohei Kikuta 氏、https://github.com/yoheikikuta/bert-japanese)で公開されているものを使用。