日本語モデル構築・共有のためのプラットフォームの形成

課題ID
jh221004

研究課題の概要

- 大規模情報基盤mdx 上の GPU リソースを効率的に活用して、深層学習による日本語言語モデルを構築して公開する。

- また、 言語モデルを構築するためのノウハウを共有し、日本語言語処理の研究を幅広く支援するための 今後の方策を探る。

背景①

「**事前学習済み言語モデル**」

- テキストマイニング、機械翻訳、情報検索、対話システムなど計算機による言語処理のあらゆるタスクに不可欠
- 共通研究リソースとして公開・活用される
- 用途に応じて多様な言語モデルが提案・公開されているが、日本語の言語モデルは限られている

背景②

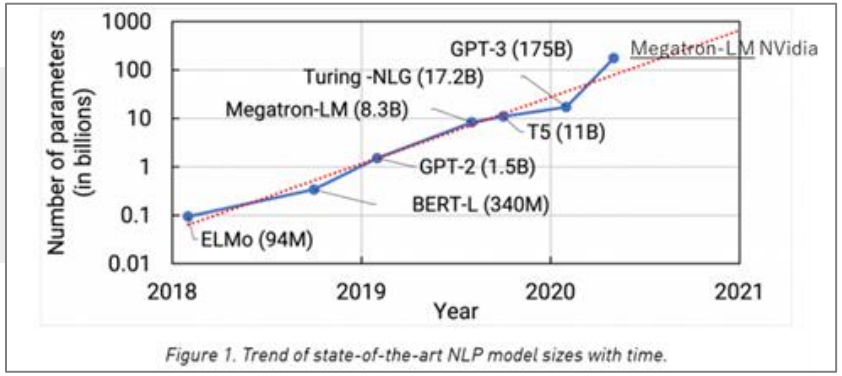

Figure 1. Trend of state-of-the-art NLP model sizes with time.

事前学習済言語モデルのパラメタ数
※　*https://developer.nvidia.com/blog/scaling-language-model-training-to-a-trillion-parameters-using-megatron/　から引用、指数的に増加している*

「**事前学習済み言語モデル**」

- 言語モデルの学習には多くのノウハウと計算資源が必要
- 言語モデルは大規模化の一途をたどっていて、従来とは桁違いの大規模なリソースが必要となる
- 単一の研究室ではリソースの確保が困難

mdxの利用

令和４年度の目標

【課題１】

分野特化型の日本語言語モデルの構築と学術分野への適用

【課題２】

汎用型の日本語言語モデルの構築と性能評価

【課題１】
分野特化型の日本語言語モデルの構築と学術分野への適用

参加研究者
国立情報学研究所　相澤、金沢、菅原
大学院生　壹岐、杉本、鈴木
奈良先端大学院大学　荒牧
東京大学　田浦

課題①：実施内容

- **医学薬学の学術ドメイン**を対象として論文テキストを収集して言語モデルを構築する。
- 言語モデルの構築における**専門用語辞書の利用やトークン化の単位の影響**を調べる
- 単語予測や文書分類などの基本的な言語タスク、医学系テキストからの情報抽出に関する共通タスク（NTCIR -MedNLP）に適用して有効性を検証する
- 構築した言語モデルを公開する

※AIPネットワークラボ日独仏「医薬品安全性監視のための言語 を越えた知識強化情報抽出」（代表者：松本裕治、分担者：荒牧、相澤）と協同

課題①：言語モデル構築方法

- 事前学習用のテキスト
  - 「医学」分野の論文（主に日本語抄録）
  - 約 1,160 万文（1文あたり平均 54.9 文字、約 1.8 GB）

- 言語モデル
  - BERT-base huggingface/transformersを使用
  - トークン化（２種類）
    1. IPA辞書 + 万病辞書（医療用語辞書）にもとづく MeCab で分かち書きし、WordPiece でサブワード化
    2. SentencePiece で直接サブワード化
  - パラメタ数 1.1億
    - （BERT-Largeだと3倍）

課題①：言語モデル評価用タスク

- 医療カルテに関するタスク
  - MedNLP2（診療データにおける固有表現抽出）
  - MedNLP3（診療データへの病名コードの付与）

- 医療論文に関するタスク
  - 医療論文の分野分類（クラス数 111 の文書分類問題）
  - 医療論文へのサブヘディングの付与（エンティティに対する「合併症」などの機能的なカテゴリの付与）

課題①：言語モデルの評価

- 医学薬学系の日本語論文から言語モデルを構築して評価、公開（2022年度、学際大規模情報基盤共同利用・共同研究拠点（JHPCN）公募型共同研究）（杉本+：2023言語処理学会年次大会）

表 2　医療ドメインタスクにおける評価結果．各モデルのスコアは，異なるランダムシードを用いた 5 回の実験結果の平均値である．太字は全モデルにおける最も高いスコア，下線は 2 種類の JMedRoBERTa のうちより高いスコアを示す．

タスク	評価指標	東北大 BERT	UTH-BERT	JMedRoBERTa（万病 WordPiece）	JMedRoBERTa（SentencePiece）
医学論文の分野分類	macro-F1	0.505	0.493	0.552	**0.566**
	micro-F1	0.622	0.627	0.671	**0.678**
医学論文の索引語への副標目の付与	macro-F1	0.451	0.456	0.487	**0.514**
	micro-F1	0.553	0.566	0.584	**0.604**
診療記録の固有表現抽出（MedNLP-2）	F1（病状）	0.855	**0.862**	0.862	0.833
	F1（時間表現）	**0.852**	0.834	0.817	0.739
診療記録への病名コード付与（MedNLPDoc）	F1（LV4 / SURE）	0.029	**0.038**	0.026	0.034
	F1（LV4 / MAJOR）	0.036	**0.067**	0.043	0.050
	F1（LV4 / POSSIBLE）	0.045	**0.078**	0.055	0.069
	F1（LV3 / SURE）	0.043	**0.059**	0.043	0.047
	F1（LV3 / MAJOR）	0.062	**0.103**	0.070	0.076
	F1（LV3 / POSSIBLE）	0.075	**0.121**	0.085	0.112
	F1（LV0 / SURE）	0.287	0.306	0.278	**0.324**
	F1（LV0 / MAJOR）	0.367	0.414	0.382	**0.432**
	F1（LV0 / POSSIBLE）	0.399	**0.448**	0.389	0.448

医学分野の言語タスクで高性能

課題①：結果まとめ

- 全般的なスコアの傾向
  - MedNLP2（診療データの固有表現抽出）では、ほとんど性能の差はない（ただし、SentencePiece は語彙の切れ目の問題でやや劣る）
  - MedNLP3（診療データへの病名コードの付与）では、UTH-BERT には劣るが東北大BERTには上回る
  - 医療論文の分野分類や医療論文へのサブヘディングの付与では、他を上回る
- 成果の公開
  - 自然言語処理で広く普及している深層学習フレームワークである Hugging Face Hub上で公開

【課題２】
汎用型の日本語言語モデルの構築と性能評価

参加研究室
京都大学 黒橋・褚・村脇研究室
東北大学 鈴木研究室
名古屋大学 武田・笹野研究室
早稲田大学 河原研究室

課題②：2022年度成果

a. Wikipediaやウェブ文書などを用いて**高性能な汎用型日本語言語モデルを構築、公開**

b. 構築した**日本語言語モデルに基づく統合的解析器を開発**

課題②-a：高性能な汎用型日本語言語モデルの構築・公開

- 学習テキスト
  - 日本語Wikipedia+ウェブテキスト(CC-100)
  - 10Bトークン
- 文字単位モデル
  - RoBERTa [Liu+ 2019] (large; 330Mパラメータ)を構築、公開
    - mdx A100 16 GPUを使用し、1か月で学習
    - https://huggingface.co/ku-nlp/roberta-large-japanese-char-wwm
- 単語単位モデル
  - GPT-2 (XL; 15億パラメータ)を構築、公開
    - mdx A100 8 GPUを使用し、2.5か月で学習
    - https://huggingface.co/nlp-waseda/gpt2-xl-japanese

課題②-b：日本語言語モデルに基づく統合的解析器の開発

- 日本語言語モデルを基盤とし、単体であらゆる言語解析を実行する日本語の統合的言語解析器KWJA (Kyoto-Waseda Japanese Analyzer)を開発

GitHub: https://github.com/ku-nlp/kwja
Demo: https://lotus.kuee.kyoto-u.ac.jp/kwja

2023年度計画

課題①：分野特化型の日本語言語モデルの構築と学術分野への適用

実施内容（１）

- パラメタ数が1B 以上の大規模言語モデルをmdx 上で動かして検証や活用が行える環境を構築する。

実施内容（２）

- 専門分野オントロジーへの対応付けにより、知識を埋め込んだ日本語言語モデルを構築・公開して、普及に努める。

課題②：汎用型の日本語言語モデルの構築と性能評価

- 外部知識を統合することによって汎用型大規模言語モデルの高性能化を検討
  - 辞書や知識グラフなどの外部知識の利用
  - 日本語言語理解ベンチマークで性能を評価
- 汎用型言語モデルを課題①に提供し、分野ごとのテキストで継続的に訓練することを検討

LLM勉強会 https://llm-jp.nii.ac.jp/

以上