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

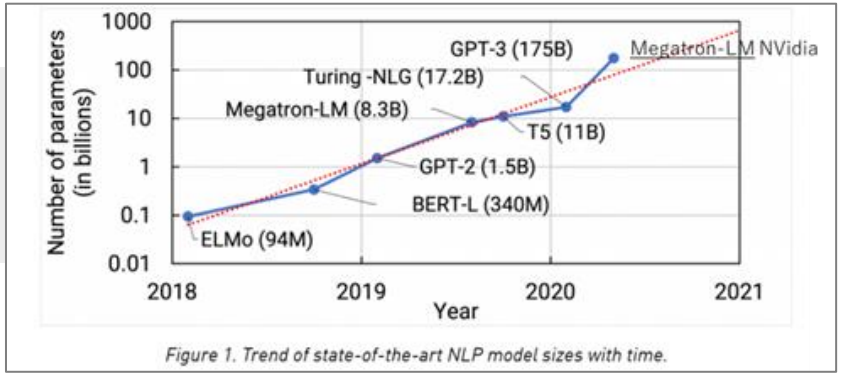

Figure 1. Trend of state-of-the-art NLP model sizes with time.

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

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：現在までの取り組み

- 予備実験
  - 少ないステップ数で試験的にモデルを作成
  - 東北大BERTやUTH-BERT（医療カルテで事前学習したBERT）と比較

- 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

【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：実施内容

- Wikipediaやウェブ文書などを用いて**高性能な汎用型大規模言語モデルを構築、公開**し、日本語言語理解ベンチマークで性能を評価する
- 言語モデルの最適なトークン単位を検討するとともに、解析器の開発も行う

サブテーマ

a. 高性能な汎用型日本語言語モデルの構築・評価・公開
b. 言語モデルの最適なトークン単位の検証
c. 文字単位言語モデルに基づく解析器の開発

課題②-a：高性能な汎用型日本語言語モデルの構築・評価・公開

- エンコーダ: RoBERTa [Liu+ 2019]
  - 日本語RoBERTa (base, large)をWikipedia+ウェブテキスト(CC-100)を用いて構築・公開済み
    - nlp-waseda/roberta-base-Japanese
    - nlp-waseda/roberta-large-Japanese
- デコーダ: GPT-3 [Brown+ 2020], HyperCLOVA [Kim+ 2021]
  - まずはGPT-2規模(15億パラメータ)の日本語モデルを構築予定
  - より大きなモデルについては、mdxでの構築、利用API提供の可能性を模索予定
- エンコーダ・デコーダ: T5 [Raffel+ 2019], BART [Lewis+ 2019]

参考: 日本語言語理解ベンチマーク(JGLUE)による評価

[Kurihara+ 2022]

	MARC-ja (acc)	JSTS (Pearson)	JNLI (acc)	JSQuAD (F1)	JComQA (acc)
人間	0.989	0.899	0.925	0.944	0.986
東北大BERTBASE	0.958	0.899	0.899	0.941	0.808
東北大BERTBASE (文字)	0.956	0.882	0.892	0.937	0.718
東北大BERTLARGE	0.955	0.908	0.900	0.946	0.816
NICT BERTBASE	0.958	0.903	0.902	0.947	0.823
早稲田大 RoBERTaBASE	0.962	0.901	0.895	0.927	0.840
早稲田大 RoBERTaLARGE	0.954	0.923	0.924	0.940	0.901
XLM-RoBERTaLARGE	0.964	0.915	0.919	-	0.840

https://github.com/yahoojapan/JGLUE

課題②-b：言語モデルの最適なトークン単位の検証

- 日本語では、テキスト処理の最初のステップとして分かち書きが行われてきたが、必要かどうかを検証
- トークン単位の候補
  - 文字

    よ い Ｂ Ｅ Ｒ Ｔ モ デ ル を 構 築 す る

  - サブワード

    よい BE RT モデル を構築 する

  - 単語+サブワード [英語の言語モデルの標準]

    よい BE ##RT モデル を 構築 する

課題②-c：文字単位言語モデルに基づく解析器の開発

- MeCab、Juman++などの既存の形態素解析器は、BERTなどの高性能な言語モデルに基づくものではない
- 文字単位言語モデルに基づく形態素解析器を開発し、既存の形態素解析器との性能比較を行う
- また、性能向上および拡張を検討する
  - 自動解析結果を用いた学習
  - タイポ修正の同時実行