DLHLP 2023 Fall
Cross-lingual

胡恩沛
enpeihu@gmail.com

Outline

- Cross-lingual transfer
- Traditional Fine-tuning v.s. In-context learning
- Homework
- Codes & instructions
- Reading list

Cross-lingual transfer

- For a downstream task, the test set may consist of more than a single language while the training data is monolingual
- Zero-shot cross-lingual transfer: Fit a Pretrained-language-model (PLM) on the monolingual training set (usually English) and inference it on multilingual data (French, German etc.)
- There has been many PLM pretrained from multilingual data such as MBert, XLM-R (Encoder-only), Mbart (Autoregressive encoder-decoder) or GPT-3, ChatGPT, GPT-4... (Autoregressive decoder) that can be applied to different cross-lingual downstream tasks

GPT-3

Davinci

Davinci is the most capable engine and can perform any task the other models can perform and often with less instruction. For applications requiring a lot of understanding of the content, like summarization for a specific audience and creative content generation, Davinci is going to produce the best results. These increased capabilities require more compute resources, so Davinci costs more per API call and is not as fast as the other engines.

Another area where Davinci shines is in understanding the intent of text. Davinci is quite good at solving many kinds of logic problems and explaining the motives of characters. Davinci has been able to solve some of the most challenging AI problems involving cause and effect.

Good at: **Complex intent, cause and effect, summarization for audience**

Curie

Curie is extremely powerful, yet very fast. While Davinci is stronger when it comes to analyzing complicated text, Curie is quite capable for many nuanced tasks like sentiment classification and summarization. Curie is also quite good at answering questions and performing Q&A and as a general service chatbot.

Good at: **Language translation, complex classification, text sentiment, summarization**

Babbage

Babbage can perform straightforward tasks like simple classification. It's also quite capable when it comes to Semantic Search ranking how well documents match up with search queries.

Good at: **Moderate classification, semantic search classification**

Ada

Ada is usually the fastest model and can perform tasks like parsing text, address correction and certain kinds of classification tasks that don't require too much nuance. Ada's performance can often be improved by providing more context.

Good at: **Parsing text, simple classification, address correction, keywords**

Note: Any task performed by a faster model like Ada can be performed by a more powerful model like Curie or Davinci.

Traditional fine-tuning

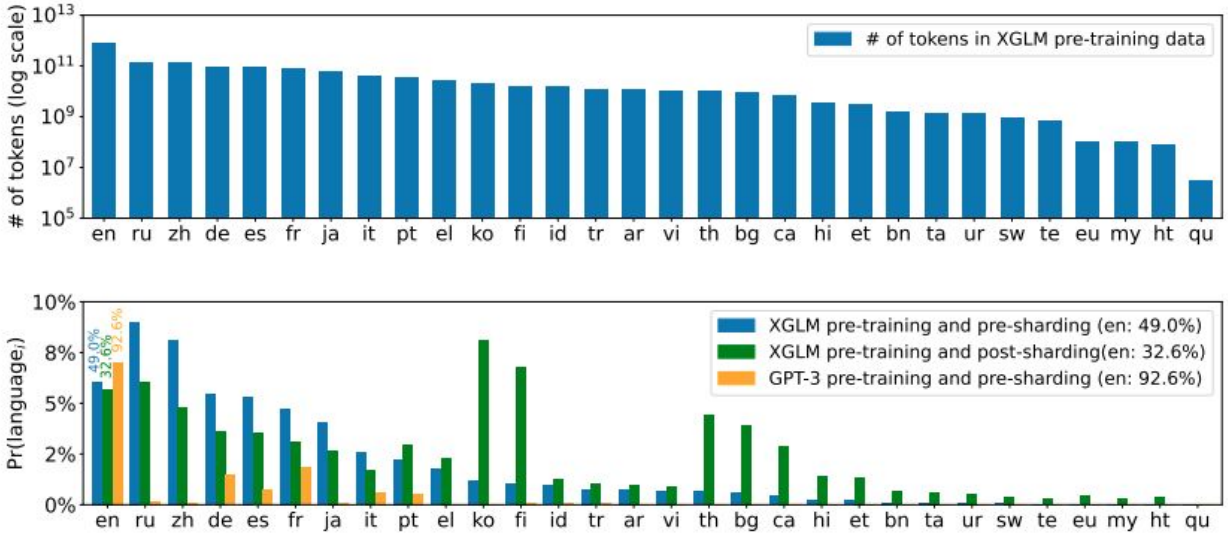
Traditional fine-tuning

"In-context" Learning

http://speech.ee.ntu.edu.tw/~tlkagk/courses/DLHLP20/GPT3%20(v6).pdf

Model - XGLM

- XGLM: A more multilingual version of GPT-3



Dataset - XNLI

- training set consists only of english data
- dev and test set can be downloaded from XNLI (choose the file XNLI 1.0 (17MB, ZIP))

Language	Premise / Hypothesis	Genre	Label
English	You don't have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d'offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Fiction	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Travel	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Telephone	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. Мы только начали этим заниматься.	Letters	Contradiction
Chinese	让我告诉你，美国人最终如何看待你作为独立顾问的表现。 美国人完全不知道您是独立律师。	Slate	Contradiction

Homework

- Part 1
- Part 2
- Other things to try

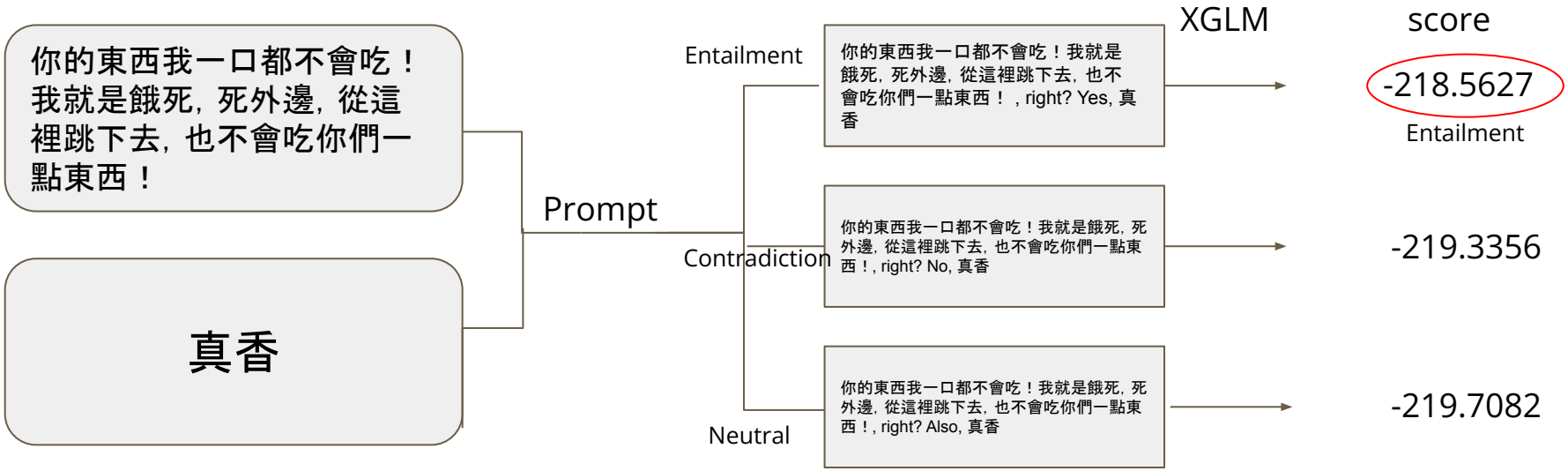
Part 1 - In-context Learning

- For XNLI, perform {0, 12}-shot in-context learning on the testing set of the specified languages and report the acc score.

List of the specified languages

- English: en
- French: fr
- Russian: ru
- Chinese: zh
- Hindi: hi
- Urdu: ur
- Bulgarian: bg
- Vietnamese: vi

Example - 0-shot on XNLI



12-shot

- There are 3 labels in XNLI: entailment, neutral and contradiction
- To balance the label distribution in the examples, you should randomly draw 4 pieces from each label on the dev set
- Trying using different examples and rearranging their **order**. Show your findings.

Part 1 - In-context Learning

1. For XNLI, perform **{0, 12}-shot** in-context learning on the testing set of the specified languages and report the acc score. (examples and inference data should be in the same language and the prompts should be in english)
2. **Translate** testing set of languages other than English with XGLM **to English** and perform **{0, 12}-shot** XNLI classification on them. Note that for k-shot translation, it is recommended to set k > 0 (Do this part first since it takes 6000s to 0-shot-translate the test data of a single language)
    - examples for k-shot translation can be drawn from the dev set since the other languages' sentences are translations of the english ones. One can also use the dev set of the flores-101 benchmark by MetaAI

- Repeat the first part but with the prompts translated to the specified language (by default the prompts are all in english)

Example of reporting accuracy scores

Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Lample and Conneau (2019)	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R$_{\text{Base}}$	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	**89.1**	**84.1**	**85.1**	**83.9**	**82.9**	**84.0**	**81.2**	**79.6**	**79.8**	**80.8**	**78.1**	**80.2**	**76.9**	**73.9**	**73.8**	**80.9**

Part 2 - Cross-lingual In-context Learning

1. For XNLI, perform 12-shot cross-lingual in-context learning on the testing set of languages other than English (7 languages). Cross-lingual means we are transfering from English to other languages. e.g. transfering from English to Chinese means the examples are in english while the “inference data” is in chinese (prompt in english)
- Try to transfer from other high-resourced languages such as Chinese to other lower-resourced languages or transfer between related language pairs (e.g. Russian to Bulgarian, French to English or Chinese to Vietnamese)

Other things to try

- Repeat part 1 - 2 on multilingual datasets such as XCOPA, PAWS-X, or even MLQA
  - Report EM & f1 score for MLQA
- Try different settings for few-shot on XNLI
  - different prompts
- Try other Multilingual Pretrained Model
  - ChatGPT, GPT-4...

Codes and instructions

- XNLI (choose the file XNLI 1.0 (17MB, ZIP))
- fairseq code to run XGLM
  - image: choose pytorch-21.06-py3:latest
  - bash experiment/req.sh
- Revision 1: dev set of the flores-101 benchmark by MetaAI for the examples of k-shot translation in part 1 (optional)
  - copy the link and download via wget
  - tar zxvf flores101_dataset.tar.gz
  - e.g. for fr-en, take flores101_dataset/dev/fra.dev and flores101_dataset/dev/eng.dev

Reading list

- Few-shot Learning with Multilingual Language Models, Lin et al.
- Unsupervised Cross-lingual Representation Learning at Scale, Conneau et al.
- Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, Liu et al.
- Language Models are Few-Shot Learners, Brown et al.

Report Date

- 10/30 : 5, 6, 7, 8
- 11/6 : 1, 2, 3, 4

Q & A

有任何問題都可以直接在 facebook 社團貼
文底下留言