sPhinX: Multilingual Instruction Paper Implementation

Natural Language Processing

03 / 02 / 26

Presented by

Jordan Brown

Paper Pipeline Overview

sPhinx Dataset generation

Selectively Translation an existing corpus (Orca prompt response dataset) using an LLM (gpt-4)

Translations should keep semantic meaning and logic intact

Translate into many different languages with a preference for high resource languages to prevent catastrophic forgetting

Heuristic Dataset Filtering and splitting

Normalize data by remove Redundant Whitespace, Punctuation and capital letters

Count the number english words remaining if it exceeds 90% discard the example as it likely hasn't been translated

N-shot example generating using Lang strategy

N examples from the target language are prepended to the beginning of the prompt as few-shot examples before the training example

N is selected randomly with a higher weight towards n = 0 and n = 1

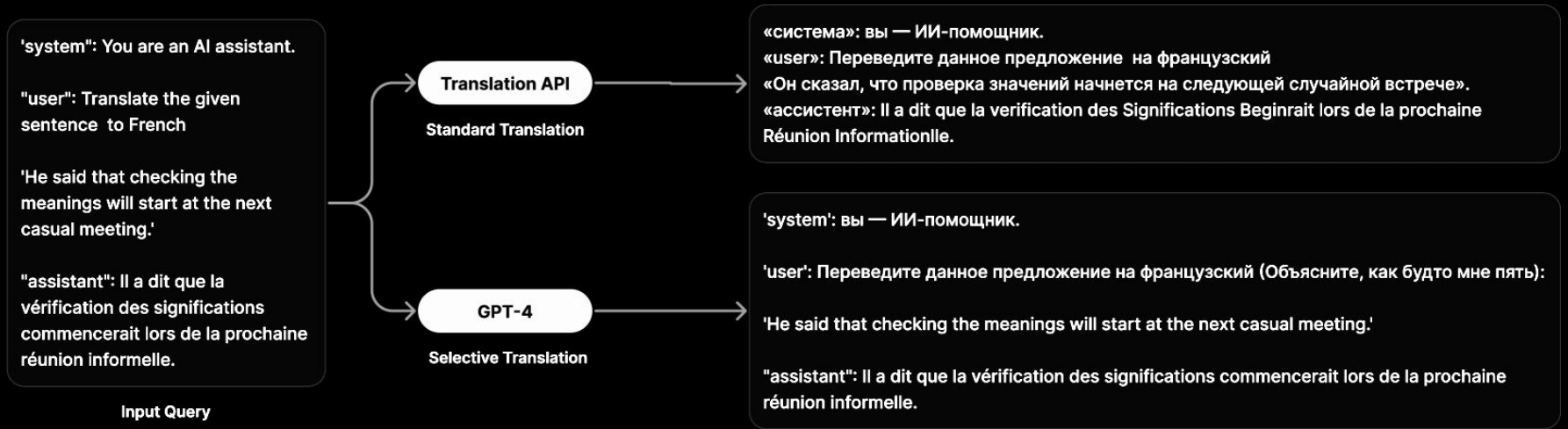
Fine-tuning using the N-shot Examples

Fine tune a smaller model such as Mistral-7B or Phi-3-small using the generated n-shot examples

Evaluate the model using various heuristics, including regression benchmarks on english

Selective Translation

- Selectively Translation an existing corpus (Orca prompt response dataset) using an LLM (gpt-4)
- Translations should keep semantic meaning and logic intact
- Translate into many different languages with a preference for high resource languages to prevent catastrophic forgetting



Step 1

Selective Translation

INPUT QUERY	MULTIALPACA DATASET	SELECTIVE TRANSLATION

{'instruction': 'Find the French equivalent of the following phrase.',
'input': 'Wishing you good luck',
'output': 'Je vous souhaite bonne chance'}
	{'instruction': 'निम्नलिखित वाक्यांश के फ्रेंच समकक्ष का पता लगाएं।',
	'input': 'आपको शुभकामनाएं',
	'output': 'Vous avez mes meilleurs vœux.'}
		{
		"system": "निम्नलिखित वाक्यांश का फ्रेंच समकक्ष खोजें।",
		"human": "Wishing you a good luck",
		"assistant": "Je vous souhaite bonne chance"
		}

{'instruction': 'You are provided with a sentence, edit it in a way that it becomes grammatically correct.',
'input': 'The wind is blowing fast and furious',
'output': 'The wind is blowing fast and furiously.'}
	{'instruction': 'आपको एक वाक्य प्रदान किया जाता है, इसे इस तरह संपादित करें कि यह व्याकरणिक रूप से सही हो जाए।',
	'input': 'हवा तेज और उग्र चल रही है',
	'id': 'alpaca-9380',
	'output': 'तेज और उग्र हवा चल रही है।'}
		{
		"system": "आपको एक वाक्य दिया गया है, इसे इस प्रकार संपादित करें कि यह व्याकरणिक रूप से सही हो जाए।",
		"human": "The wind is blowing fast and furious",
		"assistant": "The wind is blowing fast and furiously."
		}

Heuristic Dataset Filtering and splitting

- Normalize data by remove Redundant Whitespace, Punctuation and capital letters
- Count the number english words remaining if it exceeds 90% discard the example as it likely hasn't been translated
- Split the filtered data into train, test, and validation sets

```
Function dataFilter(listOfSentences):
    englishWords ← set of English words from NLTK;
    foreach sentence in listOfSentences do
        cleanedSentence ← replace all punctuations, digits, and single characters with a single space;
        cleanedSentence ← replace all sequences of whitespace with a single space;
        wordCount ← 0;
        foreach word in cleanedSentence do
            if word.lower() in englishWords then
                wordCount ← wordCount + 1;

        content ← wordCount / len(cleanedSentence);

        if content > 0.90 then

            remove sentence from listOfSentences
```

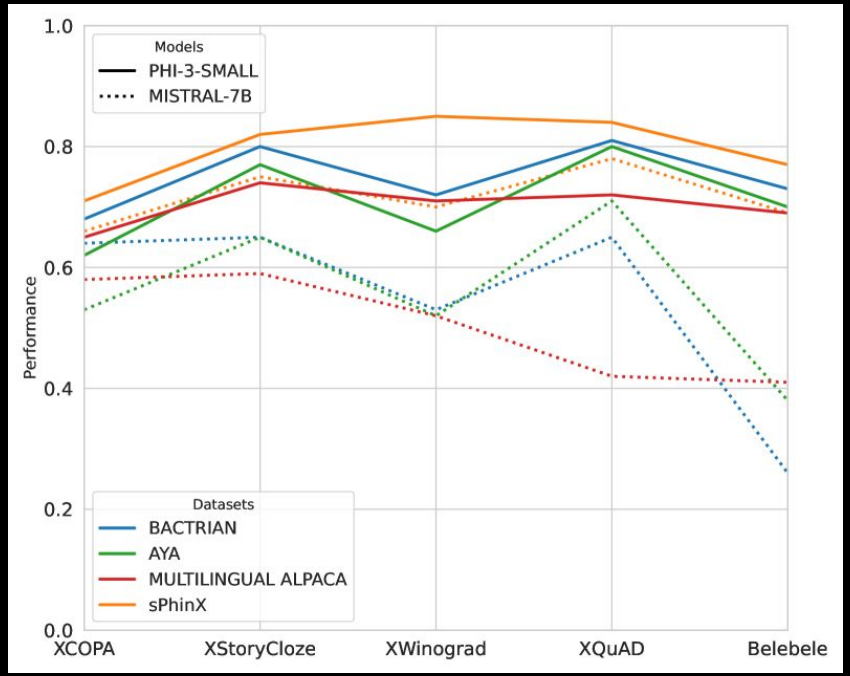
N-shot example generating using Lang strategy

- N examples from the target language are prepended to the beginning of the prompt as few-shot examples before the training example
- N is selected randomly with a higher weight towards n = 0 and n = 1

$$\left(\bigoplus_{i=1}^{N} \mathcal{T}(I^{l}_{\text{fewshot}_i}, R^{l}_{\text{fewshot}_i}) \oplus \mathcal{T}(I^{l}_{\text{train}}, R^{l}_{\text{train}})\right)$$

Fine-tuning using the N-shot Examples

- Fine tune a smaller model such as Mistral-7B or Phi-3-small using the generated n-shot examples
- Evaluate the model using various evaluation methods, including XStoryCloze, XCOPA, Belebele, XQuAD and XWinograd regression benchmarks on english



My Implementation Demo

Thank You for Listening