CS 728: Cross-lingual In-context Learning in Multilingual LLMs

Vaibhav Singh

May 2024

1 LLMs in Study

We experiment and analyse the cross-lingual natural language inference tasks of XNLI with an encoder-decoder architecture-based `google/flan-t5`. We compare performances of `google/flan-t5-xl` and `google/flan-t5-large` which have 780M and 3B parameters respectively. For SMiLER, we use `Meta-Llama-3-8B`.

2 Experimental Setup

For generation we use `T5ForConditionalGeneraton` with the `google/flan-t5` checkpoints. Further, we use nuclues sampling with $top\_k = 4$ and $top\_p = 0.1$ instead of greedy generation to control model behaviour to output only from the desired label space. We report single model runs with a fixed seed of 42.

We keep the number of few-shot examples to 4 across all experiments as the input context length of `google/flan-t5` is fixed to 512 tokens.

3 Tasks

3.1 XNLI

For XNLI task, we project the given dataset labels to the following set:

$$\begin{gathered}\text{contradiction} \rightarrow \text{False}\\ \text{entailment} \rightarrow \text{True}\\ \text{neutral} \rightarrow \text{Unknown}\end{gathered}$$

We use translations of 'True', 'False' and 'Unknown' to map the dataset labels to the new set for French and Russian.

The few-shot examples and the test sample are presented in two different languages for the cross-lingual ICL experiments. We randomly sample 4 instances from **dev** split of source language as few-shot examples. The few-shots are presented before the test sample and are preceded by a task instruction. The prompt format takes the form:

prompt = instruction + [SEP] + '[SEP]'.join(fewshots) + [SEP] + test_sample

We use the task instruction:**Determine whether the Hypothesis is True (entailment), False (contradiction), or Unknown (neutral) based on the given Premise.** The complete prompt is demonstrated in Section 5.

For task-alignment setting, we add a sentence at the end of the few-shot examples and before the test sample, specifying the difference in the labels of the source language and the target language. For example, the task aligner for English-French NLI is **"In french, True means Vrai, False means Faux and Unknown means Inconnu."**

For semantic-alignment, we choose $top\_k = 4$ most similar premise-hypothesis pairs to the test sample using `bert-base-multilingual-cased` pooler output as contextual embeddings.

3.1.1 Results

	`Random prompting`	`Task alignment`	`Semantic alignment`
	F1 (macro)	F1 (macro)	F1 (macro)
English → French	0.02	0.18	**0.30**
French → English	0.39	0.34	**0.49**
English → Russian	0.17	0.17	**0.27**
French → Russian	0.17	0.17	**0.28**

Table 1: XNLI with `google/flan-t5-large` on 1000 test samples

	`Random prompting`	`Task alignment`	`Semantic alignment`
	F1 (macro)	F1 (macro)	F1 (macro)
English → French	0.57	**0.61**	0.57
French → English	0.70	**0.71**	0.68
English → Russian	**0.36**	0.31	**0.36**
French → Russian	0.27	**0.30**	0.26

Table 2: XNLI with `google/flan-t5-xl` on 1000 test samples

3.1.2 Discussion

For XNLI, we see a significant gap in performance as the model size increases from 780M parameters to 3B parameters. 3B Flan-T5 has better averaged macro-F1 across all language pairs between English, French and Russian. We do not report numbers for pairs with Russian as source language due to near-zero performances in both model sizes. Large language models are generally pretrained on European languages primarily covering English and French. Moreover, English and French share the**Latin** script and hence their subword level representations share the same vector space and have better transfer learning as compared to Russian which follows the **Cyrillic** script and does not share subwords with either French or English.

We observe better performance when French is used a source language and English is used as target language as compared to the opposite. We believe that along with few-shot signals, English further benefits from the world knowledge of the LLM and the model does not rely on the few-shot examples solely for prediction.

3.2 SMiLER

SMiLER has 36 relation labels. We do not choose a verbalizer for this dataset and use the dataset labels as it is. As the two entities mentioned can have any one of the 36 relation types, all relation types cannot by covered with only 4 in-context few-shot examples. However, an encoder-decoder model like `google/flan-t5` is limited by a input context length of 512 tokens. Hence, we choose a decoder-only causal language model for this task.

We use `meta-llama/Meta-Llama-3-8B-Instruct` which supports an input context length of 8000 tokens.

For the relation extraction task, we explicitly provide all the possible relation types in the instruction as all relation types cannot be covered with 4/8/16 or even 32 few-shot examples with random-sampling or semantic-alignment. We prompt the model with the instruction: **You are given a list of 36 relations that you have to choose from: [“no_relation”, “is-where”, “birth-place”, “has-type”, “movie-has-director”, “has-occupation”, “from-country”, “has-genre”, “has-author”, “has-population”, “headquarters”, “is-member-of”, “org-has-member”, “has-parent”, “org-has-founder”, “has-spouse”, “won-award”, “has-nationality”, “org-leader”, “starring”, “has-edu”, “has-child”, “event-year”, “has-sibling”, “has-length”, “invented-when”, “has-tourist-attraction”, “has-lifespan”, “first-product”, “has-height”, “has-highest-mountain”, “invented-by”, “has-weight”, “post-code”, “loc-leader”, “eats”]. Given a sentence with entities e1 and e2, choose the most appropriate relation from the given relation list.**

For complete prompt, refer Section 5. For relation extraction, we compare random-prompting against semantic-alignment.

3.2.1 Results

Given that Llama-3 is pretrained with over $15T$ tokens, we expect the model to perform well with few-shot exemplars to steer the model towards relation extraction task backed by the world knowledge the model carries from pretraining. Higher micro F1 where Russian is target language is due to two factors: (1) Fewer test samples (131 test samples) and (2) less coverage of relation types in test data.

	`Random prompting`	`Semantic alignment`
	F1 (macro)	F1 (macro)
English → French	0.65	**0.76**
French → English	0.60	**0.65**
English → Russian	0.70	**0.73**
French → Russian	**0.76**	0.75

Table 3: SMiLER with `meta-llama/Meta-Llama-3-8B-Instruct` on 200 test samples with 8 in-context examples

4 Conclusion

When comparing the three implemented methodologies irrespective of the dataset and the LLMs. Semantic alignment outperforms Task Alignment and Random few-shot examples in a scenario of cross-lingual transfer from a high resource language to a relatively low-resource langauge. Overall, we see an average increase of 19.5% and 6.5% in macro F1 scores for cross-lingual transfer from English to French and English to Russian on the XNLI and SMiLER tasks using semantically-aligned few-shot examples.

5 Appendix

Determine whether the Hypothesis is True (entailment), False (contradiction), or Unknown (neutral) based on the given Premise.

Premise: excuse me we pay for any you know the child care but we don't pay as much as they do off base
Hypothesis: Childcare costs $2000 more off base.?
Response: Unknown

Premise: They took Joe with them, and my Granny said, she said it was such a sad time in the house because, you know, everybody was missing Joe and they didn't know what to do.
Hypothesis: Everyone in the house was moping because they missed Joe so much.?
Response: True

Premise: Flanking it, a modern octagonal church to the east and a chapel and hexagonal tower to the west represent the city's post-war rebirth.
Hypothesis: The marketplace represents the city's post-war rebirth.?
Response: False

Premise: Behind the South American area you'll find the Perfume Factory, where you can create your own personal fragrance.
Hypothesis: The Perfume Factory is behind the South African Area.?
Response: True

Prémisse: Eh bien, je ne pensais même pas à cela, mais j'étais si frustré, et j'ai fini par lui reparler.
Hypothèse: Je ne lui ai pas parlé de nouveau?
Réponse:

Table 4: Random prompting prompt format for English-French pair XNLI task.

You are given a list of 36 relations that you have to choose from: [‘no_relation’, ‘is-where’, ‘birth-place’, ‘has-type’, ‘movie-has-director’, ‘has-occupation’, ‘from-country’, ‘has-genre’, ‘has-author’, ‘has-population’, ‘headquarters’, ‘is-member-of’, ’org-has-member’, ‘has-parent’, ‘org-has-founder’, ‘has-spouse’, ‘won-award’, ‘has-nationality’, ‘org-leader’, ‘starring’, ‘has-edu’, ‘has-child’, ‘event-year’, ‘has-sibling’, ‘has-length’, ‘invented-when’, ‘has-tourist-attraction’, ‘has-lifespan’, ‘first -product’, ‘has-height’, ‘has-highest-mountain’, ‘invented-by’, ‘has-weight’, ‘post-code’, ‘loc-leader’, ‘eats’]. Given a sentence with entities e1 and e2, choose the most appropriate relation from the given relation list. Sentence: [e1]Walter Röhrig[/e1] (13 April 1897 – 1945) was a German [e2]art director[/e2].
Relation: has-occupation

Sentence: [e1]Paul Kerb[/e1] (born 20 December 1929) is an Austrian fencer from [e2]Vienna[/e2].
Relation: birth-place

Sentence: [e1]Herbert Lawrence Stone[/e1] (January 18, 1871 – September 27, 1955) was a noted [e2]American[/e2] magazine editor and publisher, and a renowned sailor. He was the editor of Yachting from 1908 until 1952.
Relation: has-nationality

Sentence: [e1]Otto Erdmann[/e1] (16 November 1898 – 23 January 1965) was a German [e2]art director[/e2].
Relation: has-occupation

Sentence: [e1]Jean-Pierre Adam[/e1] (born November 24, 1937 in [e2]Paris[/e2]) is a French archaeologist, specialising in ancient architecture.
Relation: birth-place

Sentence: [e1]Sergey Kuzin[/e1] ([e2]Russia[/e2] born 18 January 1971) is a Russian motorcycle speedway rider who was a member of Russia team at 2001 and 2002 Speedway World Cup.
Relation: from-country

Sentence: [e1]Edwin Zbonek[/e1] (28 March 1928 – 29 May 2006) was an Austrian film director and [e2]screenwriter[/e2].
Relation: has-occupation

Sentence: [e1]John Lemprière[/e1] (c. 1765, [e2]Jersey[/e2] – 1 February 1824, London) was an English classical scholar, lexicographer, theologian, teacher and headmaster.
Relation: birth-place

Phrase: modifier [e1]Aarre Merikanto[/e1] est un compositeur et pédagogue finlandais, né à [e2]Helsinki[/e2] (Finlande) le 29 juin 1893, décédé au même lieu le 28 septembre 1958.
Relation:

Table 5: Random prompting prompt format for English-French pair SMiLER task.