Clue-Instruct: Text-Based Clue Generation for Educational Crossword Puzzles

Andrea Zugarini, Kamyar Zeinalipour, Surya Sai Kadali, Marco Maggini, Marco Gori, Leonardo Rigutini

Introduction

- ▷ Crossword puzzles are popular linguistic games often used as tools to engage students in learning.
- ▷ Educational crosswords are characterized by less cryptic and more factual clues with respect to traditional ones.
- ▷ Puzzles can enhance memory skills and make learning experience more captivating.
- ▷ Manually crafting crossword puzzles can be time consuming, which hinders their diffusion in the education system.

We propose a methodology to generate educational clues automatically.

Introduction

We make use of Large Language Models (LLMs) to generate clues.

In order to control the generation process and prevent the risk of hallucinations, we ground the generation to a given context.

In particular, we constructed `clue-instruct`, a corpus of text-keyword pairs associated with three distinct crossword clues.

We used `clue-instruct` to fine-tune different LLMs to generate educational clues from a given input content and keyword.

Method

- **Data Retrieval.**
  - Crawling Wikipedia pages to extract its content.
  - Identification of keywords.
  - Additional metadata.
- **Data-screening.** We select pages based on their number of views and importance rating. We discard too long or too short contents and keywords
- **Prompt Crafting.** Designed to guide generation according to the extracted **keyword**, **context** (wiki page) and **category**.
- **Clues Generation.** We query an LLM to generate clues automatically.

```
You are a crosswords expert.

Generate short and clever definitions for
crosswords, based on a given keyword, a
category and a keyword-related context,
following the instructions provided below.

        KEYWORD: {keyword}

        CATEGORY: {category}

        CONTEXT: {text}

Follow these steps:

1. Find parts of the given context related
to the {keyword} and {category}.

2. Select three key pieces of information
related to {keyword} and {category} that
are present in the context.

3. Create short clues from these key
facts, making sure not to include the
keyword in the clues.

4. Put these clues into a JSON file under
the key: 'clues'.
```

Clue-instruct

Clue-instruct

A dataset for English educational clues, built selecting from 258,325 Wikipedia pages the ones:

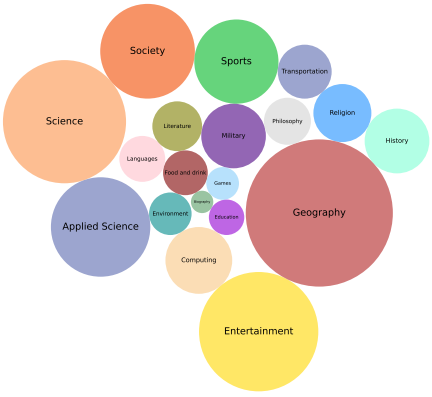
- ▷ Belonging to one of the most popular 20 distinct categories.
- ▷ With at least 10 thousands views or an importance rating equal to *top*.
- ▷ Contexts below 30 or above 1000 words were discarded.

We used GPT3.5 Turbo as clues generator. The quality of the clues was assessed with both automatic metrics and human ratings.

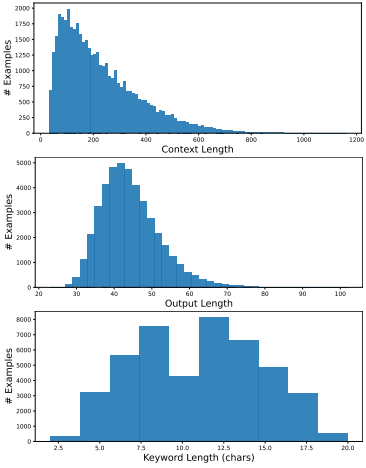
clue-instruct	
#contexts	44,075
#keywords	44,075
#categories	20
#clues	132,225

Statistics

Categories distribution:



Lengths distributions of context, clues and keywords:



Some Generated Examples

Answer	Category	Clue	Rating[1]
Robocall	Society	May be blocked by phone companies to prevent scams	A
Ministry Of Magic	Literature	Corrupt and incompetent government in J.K. Rowling's Wizarding World	A
Lovesick	Literature	Renewed for a third season, released exclusively on Netflix	C
South American tapir	Science	One of the four recognized species in the tapir family	E

[1] Five-levels rating sysyem. It goes from A (clue is valid and coherent to the context) to D (irrelevant and/or incorrect). A clue containing the answer it self is rated as E. contains the answer

Experiments

Experimental Setup

Data. We kept 600 annotated examples as test set. The remaining 43,475 examples were used for training. LLMs were instructed with the same prompt used to generate the dataset.

Baselines. Four instruction-tuned LLMs from two different families: LLAMA2-CHAT 7B and 13B sizes, and MPT-INSTRUCT in both 7B and 30B releases.

Training. All the models were fine-tuned with LORA.

Metrics. ROUGE scores and human evaluations based on the A-E ratings.

Off-the-Shelf LLMs

	model name	#params	ROUGE-1	ROUGE-2	ROUGE-L
Off-the-shelf LLMs	LLAMA2-CHAT	7B	–	–	–
	MPT-INSTRUCT	7B	23.98	11.79	19.69
	LLAMA2-CHAT	13B	**31.80**	**15.32**	**25.27**
	MPT-INSTRUCT	30B	29.92	14.47	24.30

Without instruction tuning, smaller models struggle to follow the clue-generation request.

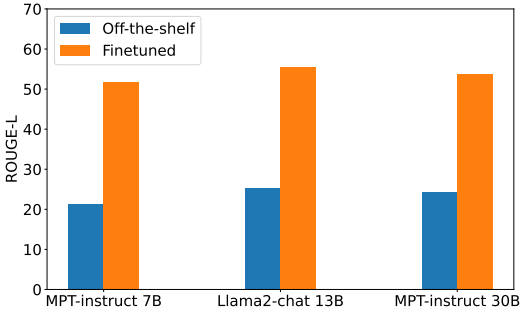
Instruction-Tuned LLMs

	model name	#params	ROUGE-1	ROUGE-2	ROUGE-L
Off-the-shelf LLMs	LLAMA2-CHAT	7B	–	–	–
	MPT-INSTRUCT	7B	23.98	11.79	19.69
	LLAMA2-CHAT	13B	31.80	15.32	25.27
	MPT-INSTRUCT	30B	29.92	14.47	24.30
Finetuned LLMs	LLAMA2-CHAT	7B	59.92	40.98	52.28
	MPT-INSTRUCT	7B	59.26	40.37	51.68
	LLAMA2-CHAT	13B	**62.97**	**44.97**	**55.40**
	MPT-INSTRUCT	30B	61.42	42.63	53.77

The generated outputs always align with the expected format.

Finetuned LLMs surpass off-the-shelf models by a large margin.

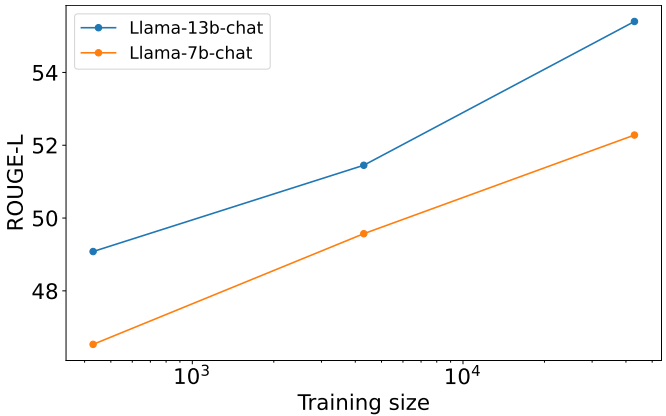
Model Size and Family



- Larger models tend to outperform smaller ones.
- LLAMA2-CHAT 13B model is particularly well-performing, surpassing MPT-INSTRUCT 30B that is more than twice its size.

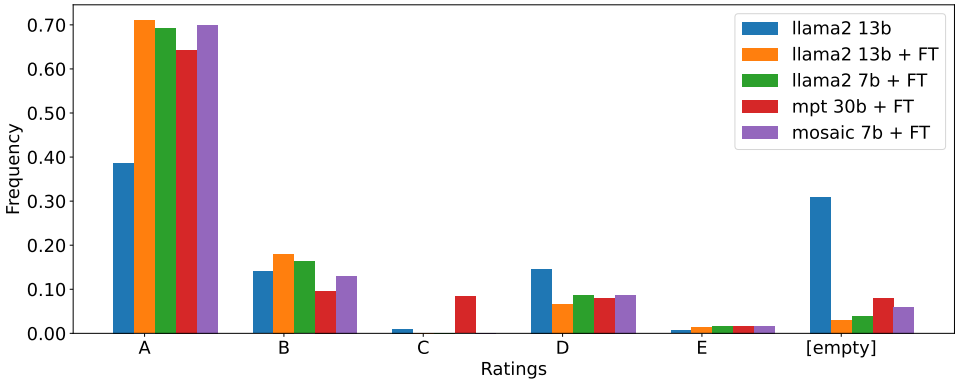
Impact of Dataset Size

`clue-instruct` is truncated at sizes corresponding to 1%,10% and 100% of the training set.



A small number of examples is enough to align the LLMs, even for LLAMA2-CHAT 7B that failed to produce valid clues when applied as zero-shot.

Human Evaluation



All the tuned models exhibit major reductions of malformed outputs and a significant increase of A-rated clues.

Conclusions

- ▷ We constructed `clue-instruct`, a dataset with keyword-clue pairs grounded on an input context, specifically designed for educational crosswords.
- ▷ Fine-tuning different Large Language Models on `clue-instruct`, shows that aligning LLMs to this kind of instructions greatly improves the quality of the clues in terms of both automatic and human evaluation.

Our methodology can be applied to non-English languages.

We plan to tune newer LLMs as well.

Thank you for listening!

Dataset

https://huggingface.co/datasets/azugarini/clue-instruct

Models

https://huggingface.co/azugarini/clue-instruct-llama-7b
https://huggingface.co/azugarini/clue-instruct-llama-13b
https://huggingface.co/azugarini/clue-instruct-mpt-7b
https://huggingface.co/azugarini/clue-instruct-mpt-30b

Andrea Zugarini

@AZugarini