Tuning Zero Shot into Few shot via Self-prompting for Classification

Michael Gunn, Justin Lin, Gaozheng Liu, Lilesh Kurella, Kulbir Singh Ahluwalia

Motivation

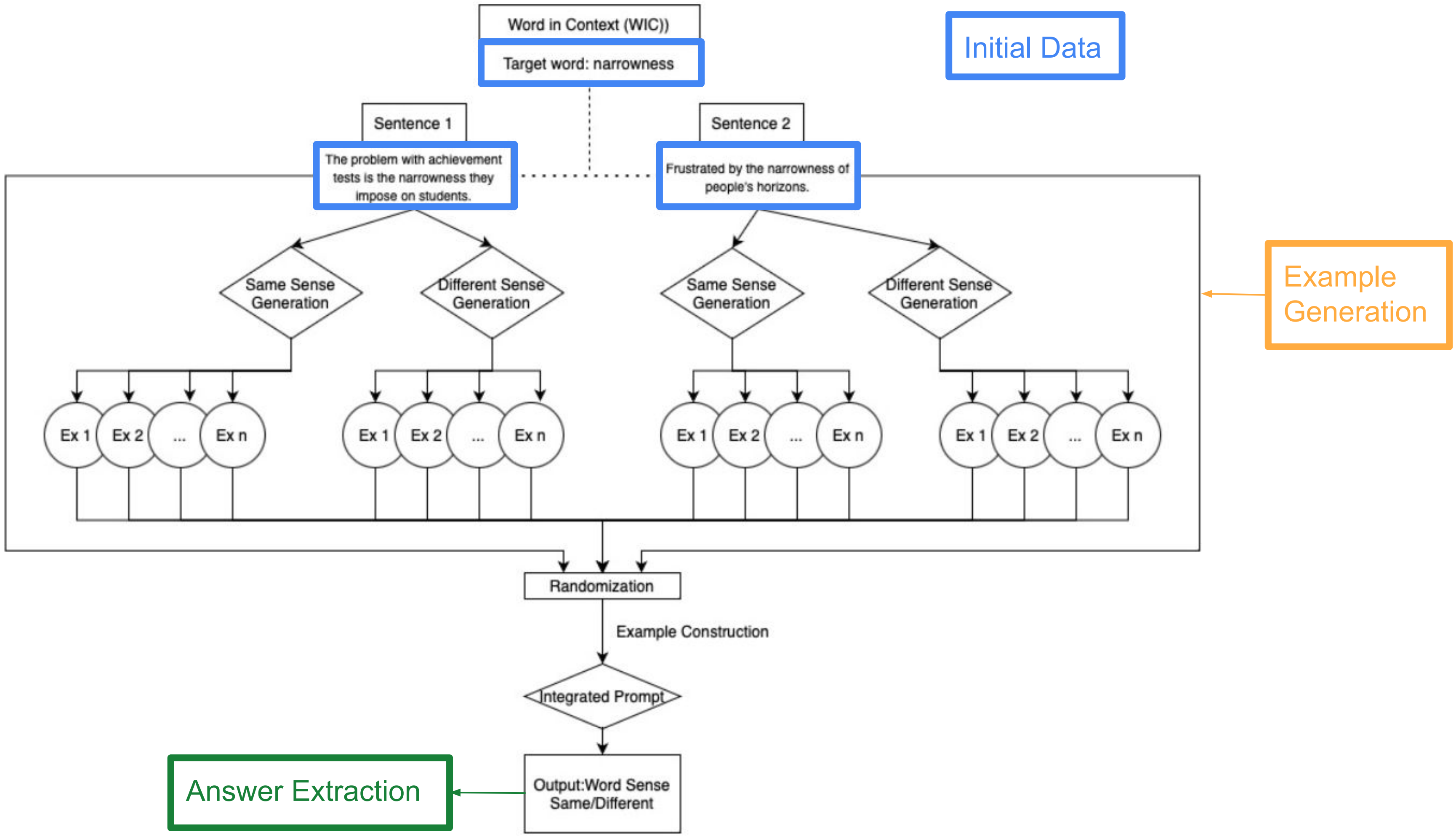
- Few shot performance while not using any outside data
- Models contain significant internal parametric knowledge
    - Can generate their own labeled data
- Previous work has fine tuned on generated pseudo-labeled data
    - Or retrieved pregenerated examples as few-shot examples
- We aim to generate highly relevant examples with labels **at inference time** to improve **classification** accuracy

Background

- **Research Objective:**
    - Addressing the gap between zero-shot and few-shot performance in language models.
    - Novel approach: Transforming zero-shot queries into few-shot by self-generating example prefixes.
- **Methodology & Key Inspirations:**
    - **Previous Approaches:** Fine-tuning models on generated examples (e.g., FewGen, SuperGen, Zero-Gen)
    - **Influential Works:** Insights from Self-Instruct and ODQA's self-prompting method
    - **Unique focus:** Employing self-prompting in a frozen language model for real-time context generation in classification tasks.
- **Benchmarking:**
    - Evaluating with SuperGLUE, particularly the Words In Context (WiC) dataset.
    - Evaluating Stanford Sentiment Treebank (SST2)

Pipeline – WiC specific

- **Key Functionalities:**
  - **Baseline Test:** Uses a basic prompt without additional context.
  - **Self-Prompting Method:** Generates and uses example prompts for additional context.
- **Self-Prompting Method Process Flow:**
  - **Load Pre-trained Model and Tokenizer:** meta-llama/Llama-2-13b-chat-hf
  - **Context Generation:** For each instance, generate example sentences for both "same sense" and "different sense" scenarios using target words.
  - **Full Prompt Generation:** Combine current prompt with the same sense and different sense examples generated by the model
  - **Answer Extraction:** Parse and Extract answer from model after passing combined prompt through model
  - **Evaluation:** compare model response with ground truth
- **Outcome Measurement:**
  - Track the number of correctly answered instances.
  - Calculate and log accuracy after processing each instance.


Word in Context (WIC))
Target word: narrowness
Initial Data
Sentence 1
The problem with achievement tests is the narrowness they impose on students.
Sentence 2
Frustrated by the narrowness of people's horizons.
Same Sense Generation
Different Sense Generation
Same Sense Generation
Different Sense Generation
Example Generation
Ex 1
Ex 2
...
Ex n
Randomization
Example Construction
Integrated Prompt
Output:Word Sense Same/Different
Answer Extraction

Experiments

- **Datasets**:
  - Words in Context (WiC), Stanford Sentiment Treebank (SST2)
- **Models**:
  - LLama2-13b-chat (4 bit quantized), Mistral
- Number of few shot generated examples

Results on WiC

Accuracy	Baseline	4-shot*	12-shot*	16-shot*	Filtered k-shot
Llama2 13b (4bit quantized)	60%	55%	59%	64%	-
Mistral v0.1	55%	-	-	-	-

*4 shot for WiC means

- 1 **same** sense example pair generated using **s1,** 1 **different** sense example pair generated using **s1**
- 1 **same** sense example pair generated using **s2,** 1 **different** sense example pair generated using **s2**

Tried using Flan-T5 large, mistral, and llama2-7b but each required making new pipeline to handle formatting differences in generation

*All tests other than baseline were performed on 250 items. Baseline used 500 items

Results on SST2

Accuracy	Baseline	2 shot*	6 shot*	Filtered k-shot
Llama2 13b	89.2%	93%	93%	-

*6 shot for SST2 means

- 3 **positive** examples generated using **s,** 3 **negative** examples generated using **s**

Tried using Flan-T5 large, mistral, and llama2-7b but each required making new pipeline to handle formatting differences in generation

*All tests other than baseline were performed on 200 items. Baseline used 500 items

Example: “Is pretty damned funny”

- Baseline: Negative - Wrong
- Few Shot: Positive - Correct

Why it doesn't work better

- Generated examples are not gold quality, some could mislead
- Previous research has indicated that examples correctness isn't very important, primarily demonstrate the label space
  - It is already understanding the label space without examples, consistently can give labels as answer
- Small models don't use few shot examples well
  - Variation of chain of thought, which has been shown to be an emergent property
  - 512 token context window means number of examples is limited
- Slows down inference, limiting how many variations we could test
  - Prompt tuning for smaller models makes large impact

WiC Bad Example

Consider the following sentence.

Sentence 1: You have a two-hour window of clear weather to finish working on the lawn

Generate 5 diverse sentences containing the word window where window has a different meaning than in the reference sentence.

Sentence 2: The window of opportunity for the new project is now open.
…
Sentence 5: The window of the store was covered in posters and signs.

Sentence 2 is a bad example

Filtering

- Our self generated examples demonstrated 3 general shortcomings:
  - Sentence structure can be very similar to provided reference sentence
    - Addressed by changing prompt to ask for “diverse sentences with different structures”
  - Sentences generated can be completely incorrect for the desired label
    - Ask for same word sense, some examples use word with different word sense
  - Sentences doesn’t contain target word or encompass the goal
    - These could be retried for generation, or passed to not use examples

Future work

- Larger models perform better with few-shot examples

- Larger models have better chain of thought reasoning capability
    - Our pipeline pulls on CoT capabilities, since asking it to generalize from examples of each label requires it to understand those are related to the problem instance

- Large models would require less parsing effort

Thank You

Filtering examples

- Our self generated examples demonstrated 3 general shortcomings:
    - Sentence structure can be very similar to provided reference sentence
        - Addressed by changing prompt to ask for “diverse sentences with different structures”
    - Sentences generated can be completely incorrect for the desired label
        - Ask for same word sense, some examples use word with different word sense
    - Sentences doesn’t contain target word or encompass the goal
        - These could be retried for generation, or passed to not use examples

- Generation of semantic embeddings for S1 and S2 using Sub-Sentence Encoder:
    - Compared the contextual similarity between the target word in S1 and S2.
    - Can Output a similarity score between the self-generated sentences [S1,S2] pair.

- Filtering of self-generated data:
    - Similar word sense: Use the top 2 similarity scores (>0.8 cosine similarity)
    - Different word sense: Use the bottom 2 similarity scores (<0.2 cosine similarity)

Category	Cosine Similarity Score	Sentence 1	Sentence 2	Ground Truth Label	Subsentence Encoder Label
Same Sense	0.947683	The flowers bloomed early this **summer**.	The sun shone brightly during the **summer**.	1	1
Same Sense	0.947683	The flowers bloomed early this **summer**.	The children enjoyed their **summer** break from school.	1	1
Different Sense	-0.033768	The **flux** prevents oxides from forming on the metal during the soldering process.	The government is in **flux**.	0	0
Different Sense	-0.026298	The **flux** prevents oxides from forming on the metal during the soldering process.	His opinions are in **flux**.	0	0

Results on WiC

Accuracy	Baseline	4-shot*	12-shot*	16-shot*
Llama2 13b (4bit quantized)	60%	55%	59%	64%
Mistral v0.1	55%	-	-	-
Llama2 13b (4bit quantized) + Filtered k-shot (Using model label)				60%
Llama2 13b (4bit quantized) + Filtered k-shot (Using subencoder label)				56%

*4 shot for WiC means

- 1 **same** sense example pair generated using **s1,** 1 **different** sense example pair generated using **s1**
- 1 **same** sense example pair generated using **s2,** 1 **different** sense example pair generated using **s2**

Tried using Flan-T5 large, mistral, and llama2-7b but each required making new pipeline to handle formatting differences in generation

*All tests other than baseline were performed on 250 items. Baseline used 500 items

Future work

- Larger models perform better with few-shot examples

- Larger models have better chain of thought reasoning capability
    - Our pipeline pulls on CoT capabilities, since asking it to generalize from examples of each label requires it to understand those are related to the problem instance

- Large models would require less parsing effort

- The similarity score threshold for filtration could be set to self-generate until semantically diverse examples are obtained.

- Subsentence encoder can be used to correct self-generated labels which can assist in increasing the accuracy.

Prompt engineering

1. Prompting for generation of examples
    a. Use the format to include self-generated examples from first inference:
        i. <s>[INST] <<SYS>>
        ii. {{ system_prompt }}
        iii. <</SYS>>
        iv. {{ user_msg_1 }} [/INST] {{ model_answer_1 }} </s><s>[INST] {{ user_msg_2 }} [/INST]
    b. **System _prompt** = The task is to label whether the word {target} is being used with the same word sense. Consider the following examples:
    c. **User_msg_1** = “Generate 2 sentences where {word} has similar meaning as Sentence 1: {sentence1}. Generate 2 sentences where {word} has similar meaning as Sentence 2: {sentence2}. Generate 2 sentences where {word} has different meaning as Sentence 1: {sentence1}. Generate 2 sentences where {word} has different meaning as Sentence 2: {sentence2}.“
    d. **Model_answer_1** = “[S1,S2,Label], [S1,S2,Label], [S1,S2,Label] ….”
    e. **User_msg_2** = Determine if the word '{target}' in each pair of sentences is used with a broadly similar meaning or if there's a significant difference in its usage. Focus on the overarching sense of the word, rather than subtle nuances.Does '{target}' have a broadly similar meaning in both sentences? Answer 'yes' for similar or 'no' for different.
    f. [TODO: Show example:]

2. Change in accuracy for the following cases:
    a. For each example pair, generate the "same" with 75% or higher similarity of the sense. generate the "difference" with 75% or higher difference in sense
    b. Generate pairs of examples based on the original pair, rather than generating sense "same" or "different" as one sentence in the pair.
    c. Generate explanation for the self-generated label (Kulbir)
    d. Different quantizations [4 bit, 8 bit, 16 bit]

Filtering and Evaluation of generated examples

1. Generation of semantic embeddings for S1 and S2:
    a. Sub-Sentence Encoder Embeddings: Can compare the contextual similarity between the target word in S1 and S2.
        i. Can Outputs a similarity score between the input [S1,S2] pair given a condition.
        ii. In our case, Condition = Target word or a description of target word from either S1 or S2.
    b. SimCSE embeddings: used to generate positive(similar semantically) or negative(different semantically) labels
        i. Embeddings are generated for the entire sentence without focus on target word.

2. Filtering of self-generated data:
    a. Similar word sense: Use the top 2 positive similarity scores
    b. Different word sense: Use the top 2 negative similarity scores

3. The similarity score threshold for filtration could be set to self-generate until semantically diverse examples are obtained.

4. Evaluation:
    a. Step1: Extraction of generated label
    b. Step2: Use ground truth label to compute accuracy
    c. [TODO: Prompt larger LLMs to:
        i. get better explanations for the generated label ⇒ see if that makes any difference in accuracy
        ii. For the self-generated [S1,S2] pairs, generate labels using a larger LLM and see if that makes a difference in accuracy
            1. (check if having a correct label matters or not)