Language model with presuppositions

Ali Taylan Akyürek
aakyurek17@ku.edu.tr

1 Problem statement

My project's central goal is to enhance the reasoning inference capabilities of language models through the identification and integration of presuppositions from textual data.

Presuppositions are assumptions or inferences that are implicitly suggested in a statement or question. Identifying and understanding these can be crucial for a language model to fully grasp the context and implications of a sentence, leading to more accurate and nuanced responses.

My research is motivated by the observation that most language models, while performing well on a surface level, often miss out on these nuanced presuppositions, which can lead to mistakes in interpretation and subsequent reasoning tasks. Also, a presupposition is actually a valid data instance that includes true world knowledge and a coherent structure like all other sentences, so generating presupposition may help to generate more data if there is not enough data to train.

To address this, I'm proposing a two-model system:

a) A model responsible for identifying both static and contextual presuppositions from the sentences in a dataset. This model is actually a prompt engineering module that uses gpt-3.5-turbo. I instructed it in order to:

1) Presupposition generated must be significantly different from original sentence.

2) Chosen presuppositions should give best information about the external world or situation that the sentence occured

b) A language model that will be fed by presuppositions.

My research aims to push the boundaries of language models' inference capabilities by exploring and leveraging the often-overlooked aspect of presuppositions in textual data.

2 What you proposed vs. what you accomplished

1) Building a model to generate presuppositions (Done)

2) Building a language model to make inferences given inputs (Done)

3) Creating datasets based on external datasets + presuppositions of them (Done)

4) Preprocessing data (Done)

5) Evaluating inference (text generation) capacities of language model that has fed by data + presuppositions (Done)

6) Feeding model with different type of datasets for finetuning (I will discuss detaily but some examples are: using 1 or 2 or 3 presuppositions, putting presuppositions just after relevant sentence vs putting presuppositions to a separate paragraph), then evaluating the inference capacities of different models and comparing(Done)

7) Creating different models and comparing them: I could not do this because I had to use opeanai to generate presuppositions since free models do not perform well (I will explain more), so with a limited data (only 3600 data examples that has around 10-30 sentences). This will be my next step for this project.

8) Training a model from scratch: I could not do this because of lack of data again.

9) Evaluating reasoning capacity of model: I could not do this, I tried but there is limited source and my knowledge and time was limited, I found some projects about evaluating commonsense reasoning but could not make it work. This is also my next step for this project.

3 Related work

I perceive this project's task as more of a data augmentation technique since the models' main difference is the datasets that models finetuned.

There is of course the aspect of how these presuppositions will affect the reasoning but since I could not evaluate my model in terms of reasoning capacity, I will not explain related work about reasoning with NLP.

In data augmentation end, Natural Language Processing frequently grapples with the challenge of data augmentation. Mainly because defining text transformations that maintain the original meaning is no easy task, as cited in multiple studies (Kobayashi et al., 2018). A variety of methods have been explored in the realm of research to tackle this issue. These include techniques like word replacement with synonyms (Zhang X et al. 2015) closely associated embeddings Alzantot et al., EMNLP 2018) and words forecasted by a language model (Fadaee et al., ACL 2017). Other techniques involve deletion (Huong TH et al. 2020) swapping (Wei and Zou, EMNLP-IJCNLP 2019), inducing spelling errors (Coulombe C et al., 2018), and paraphrasing (Kumar et al., NAACL 2019) at the word level. Some other methods are backtranslation (Sennrich et al., 2016) and task specific heuristics like in our case (Kafle et al., INLG 2017).

4 Datasets

One of the most important part of my project is my dataset. I have created 3600 data examples that has around 10-30 sentences. That is a little less than usual for finetunning but due to using openai api, I could not create big datasets unfortunately. I essentially tried many dataset including openwebtext, wiki-text, falcon-refinedweb etc. Most successfull dataset I tried was eli5. Reasons of that are: sources like wiki-text has so much factual sentences that explains a topic rather than discussing the topic. Presuppositions are not so informative for sentences or statements that explains something. Data examples at Eli5 is more natural and daily, and since the presuppositions are general assumptions in daily discussions, it was logical that presuppositions gave more information for the case of Eli5. I have no evaluation results across the datasets because I was using openai to generate presuppositions and that was costing time and money, so I could not prepare presupposition datasets for other datasets I tried. But I have tried to gain some insight manually and investigated presuppositions of examples in each dataset carefully. Based on my findings, I concluded Eli5 will perform best. If I will have some resources at future, I want to train my model with presuppositions of other datasets as well. Also, I did some significant amount of data cleaning due to problems related to openai responses and data corruptions caused by not getting response from openai due to overloaded requests etc. the codes that I wrote for cleaning data are mainly inside main project folder and MyPRESDataset folders. Also with concatenating my datasets in different combinations, I created some train datasets to feed the model. These different combinations are:

Dataset1 onlySentences: I feed the model with basically just sentences. That is the baseline model.

Some examples:

• Rings that intercepted would collide and damp down into a single ring. Similar any kind of dust cloud - although if the dust is diffuse enough (i.e., there's not much of it), the timescale could be longer than the age of the planet. For instance, all the giant planets have a diffuse cloud of irregular satellites, allowable since they're so few of them.Rings viscously spread, so any ring would eventually encounter any other ring. Again, the question becomes "how long will this take?", which depends on the properties of the ring. Immediately after the formation of a ring (however that happens), they needn't be aligned. But eventually, they'll be aligned.Rings can only effectively extend out to the Roche Radius - beyond that, their self-gravity will result in them accumulating into moons. This has probably happened with Saturn, at least (which has rings extending to the Roche Radius, then small, icy moons of the same composition beyond.

• You are probably talking about herbicides with 2,4-Dichlorophenoxyacetic acid/2,4-D as the active ingredient. 2,4-D works by dramatically increasing auxin production. Auxins are a class of plant growth regulators that in correct amounts promote cell growth. However, the overproduction of auxin caused by 2,4-D causes uncontrollable growth that the plant can't keep up with, leading to deleterious distribution of photosynthates, and eventually death. Dicots (e.g. broadleaf weeds) are particularly sensitive to auxins, while monocots (turf, cereal crops) are much less sensitive. Consequently, correctly applied 2,4-D will cause uncontrollable growth and death in dicots while having little or minimal effect on

monocots. It's possible to kill grass with 2,4-D, but would require a very strong application.

Dataset2 mixedPres (with 1 or 2 or 3 presuppositions): I feed the model with a dataset that contains presuppositions just near the corresponding sentences.

Some examples:

One presupposition:

• Rings that intercepted would collide and damp down into a single ring. There are multiple rings present in the situation. Similar any kind of dust cloud - although if the dust is diffuse enough (i.e., there's not much of it), the timescale could be longer than the age of the planet. There is a possibility of a dust cloud existing in the universe. For instance, all the giant planets have a diffuse cloud of irregular satellites, allowable since they're so few of them. There are giant planets in the universe. Rings viscously spread, so any ring would eventually encounter any other ring. There are multiple rings present in the situation. Again, the question becomes "how long will this take? There is a task or project that needs to be completed. ", which depends on the properties of the ring. The ring has unique properties that set it apart from other rings. Immediately after the formation of a ring (however that happens), they needn't be aligned. A ring has been formed. But eventually, they'll be aligned. There was a previous misalignment. Rings can only effectively extend out to the Roche Radius - beyond that, their self-gravity will result in them accumulating into moons. The Roche Radius is a well-known astronomical term. This has probably happened with Saturn, at least (which has rings extending to the Roche Radius, then small, icy moons of the same composition beyond. Saturn's rings are made up of the same composition as its small, icy moons beyond the Roche Radius.

Three presupposition:

• You are probably talking about herbicides with 2,4-Dichlorophenoxyacetic acid/2,4-D as the active ingredient. There is a high likelihood that herbicides with 2,4-Dichlorophenoxyacetic acid/2,4-D as the active ingredient are being discussed. The use of herbicides with this ingredient is common in agriculture. The potential benefits and drawbacks of using herbicides with this ingredient are being considered. 2,4-D works by dramatically increasing auxin production. Auxin production is low without the use of 2,4-D. The increase in auxin production is significant enough to cause visible changes in plant growth. The use of 2,4-D is necessary for optimal plant growth. Auxins are a class of plant growth regulators that in correct amounts promote cell growth. Plants cannot grow without auxins. The amount of auxins needed for cell growth varies. Other plant growth regulators do not promote cell growth as effectively as auxins. However, the overproduction of auxin caused by 2,4-D causes uncontrollable growth that the plant can't keep up with, leading to deleterious distribution of photosynthates, and eventually death. The plant was healthy before the overproduction of auxin caused by 2,4-D. The plant was unable to control its growth due to the overproduction of auxin. The plant's death was a direct result of the uncontrollable growth caused by the overproduction of auxin. Dicots (e.g. There are other types of plants besides dicots. Dicots are a common type of plant. The speaker has some knowledge or interest in botany. broadleaf weeds) are particularly sensitive to auxins, while monocots (turf, cereal crops) are much less sensitive. Auxins are commonly used in agriculture to control weed growth. Turf and cereal crops are the most commonly grown monocots. Broadleaf weeds are a major problem in agriculture due to their sensitivity to auxins. Consequently, correctly applied 2,4-D will cause uncontrollable growth and death in dicots while having little or minimal effect on monocots. 2,4-D is commonly used in agriculture. Dicots and monocots are two types of plants. The use of 2,4-D can have harmful effects on dicots. It's possible to kill grass with 2,4-D, but would require a very strong application. 2,4-D is commonly used to kill weeds. The grass in question is particularly resilient. The use of 2,4-D on the grass is necessary for the situation at hand.

Dataset3 presBelowSentence (with 1 or 2 or 3 presuppositions): This dataset is essentially a dataset that includes one example of original data, one example of presuppositions of each sentence as a differt data point at index next to it. I will just provide an example with 2 presuppositions since I think now my different number of presuppositions approach is clear and all datasets contain a subdataset that has 1, 2 and 3 presupposition version of each of my train data creation style.

An example of side-by-side data examples:

• This is a difficult question to answer. I assume you're referring to the fact that gravity is the

weakest of the four fundamental forces. Is your question why it is weakest? If that's your question the answer is basically "because it is." One of the forces had to be the weakest and it happens to be gravity.If your question is "why is it referred to as weak, when it seems so strong." The answer to that is more fun. You're probably sitting in a chair right now. You're held to the chair by gravity, but the electromagnetic forces of the bonds in the chair (and your butt) keep you from falling through the chair. So yes, gravity is the weakest, and that weakness allows objects to be stacked, and for your butt to stay above the chair seat. Here's some more info on the fundamental forces:URL0 Perhaps you can clarify your question? Or the motives for your question?

• There is a question that needs to be answered. Gravity is one of the four fundamental forces. The speaker believes that the listener has a question. The speaker is aware of the question being asked. Gravity is the only force that is weak. The term "weak" is commonly used to describe things that appear strong but have hidden vulnerabilities. There was a previous question or statement made. You may have been sitting in a chair for a while. The chair is made of materials with strong electromagnetic bonds. Gravity is a fundamental force in the universe. There is a need for more information on fundamental forces. The person being asked the question has a specific motive for asking.

And also there is a dataset that contains only-Sentences in one index and corresponding mixed-Pres in next index. I created it only for three presupposition.

5 Baselines

Baseline of my project is basically a pretrained distilgpt2 model that finetuned with 3600 regular(presuppositions excluded) data examples that contains 10-30 sentence each. I chosed this baseline model because I want to train a text generation model for casual language modelling and distilgpt2 was a popular pretrained model fort his task. The pretrained distilgpt2 model is a prominent choice for text generation tasks due to its small size, which allows for faster training and inference, as well as its impressive performance, which has been demonstrated on numerous NLP benchmarks. I will also try different models that is specialized in different tasks. Model generates completion text based on given input text. I did not tuned any hyperparameter since I was using pretrained models. My train/test split is essenntially I have 3600 examples for training and 800 example for testing.

6 Your approach

To tackle the task of integrating presuppositions into the training of language models, the approach I used involved a two-model system: a model to generate presuppositions (gpt-3.5-turbo), and a model to make inferences given these inputs (distilgpt2).

The first step of this process was to take the sentences from the dataset and use gpt-3.5-turbo to generate presuppositions for each sentence. The outputs of this model were then fed into the second model (distilgpt2), which was tasked with making inferences based on these presuppositions.

In order to ensure that the presuppositions were significantly different from the original sentence, I provided specific instructions to the gpt-3.5-turbo model to generate presuppositions that offered new information about the external world or situation that the sentence occurred in. Then I cleand the data.

Afterwards, I trained and fine-tuned the distilgpt2 model with different types of datasets to evaluate the inference capacities of different models and compare them. These datasets include original sentences, sentences with one, two or three presuppositions, and presuppositions arranged in different ways (either directly after the relevant sentence or in a separate paragraph).

I mainly used huggingface libraries. I did not implement a model myself, I finetuned distilgpt2. Corresponding files are the ipynb files and LMFineTune.py. I did run my experiments on Google colab.

Now, let's delve into the metrics I used and how presuppositions might have influenced them.

Perplexity: Perplexity is a commonly used metric for evaluating the performance of language models. It measures how well a language model predicts a sample. A lower perplexity score indicates that the language model is better at predicting the sample. Essentially, a model with a lower perplexity has less uncertainty about the next word in a sequence, given the previous words.

When it comes to presuppositions, the intuition is that if a model has been trained to handle them correctly, it should have less uncertainty when

dealing with sentences that include them. Consequently, the model should achieve a lower perplexity score, indicating a higher level of performance.

BERTScore: BERTScore is a metric that leverages the BERT language model to evaluate the quality of generated text. BERTScore computes similarity between the generated text and the reference text at the token level, where similarity is quantified as the cosine similarity between contextualized embeddings from the BERT model.

The main advantage of BERTScore over traditional metrics like BLEU is that it takes into account the contextual information of words. This means it has the ability to capture semantic similarity rather than just n-gram overlap.

Presuppositions are highly contextual in nature – they rely on understanding the context of the sentence to make sense. Therefore, a model trained on presuppositions should ideally show better performance when evaluated with BERTScore. This is because it should be better at handling the added context of the presupposition, leading to better alignment between the generated text and the reference text at a semantic level.

Why Use These Metrics:

Perplexity and BERTScore, combined, offer a holistic evaluation of my language models. Perplexity gives a measure of how well your model can predict the next word, i.e., the fluency of generated text. On the other hand, BERTScore gives a sense of how well your model understands and captures the semantic similarity in the presence of presuppositions. By using both, you get a fuller picture of how well the model handles the task, both in terms of language fluency and semantic comprehension.

These metrics were ideal for my project because they allowed me to effectively measure the impact of integrating presuppositions into the training of language models. Given that presuppositions add an additional layer of context to the data, using a metric like BERTScore that takes this context into account was crucial for accurately evaluating the models' performance.

6.1 Results:

Results were not quite like I expected. Below you can find evaluation results for each variant of model.

Perplexity:

Sentences only (baseline): 44.66

	Sentence-presuppositions mixed	Presu
One presupposition	54.08	
Two presupposition	57.23	
Three presupposition	57.15	

Table 1: Perplexity

On overleaf, I cannot make the table smaller, the column on the right is not visible so I also will provide description of the table:

Sentence-presuppositions mixed and One presupposition = 54.08

Sentence-presuppositions mixed and Two presupposition = 57.23

Sentence-presuppositions mixed and Three presupposition = 57.15

Presuppositions below sentences and One presupposition = 46.27

Presuppositions below sentences and Two presupposition = 46.75

Presuppositions below sentences and Three presupposition = 47.39

Normal sentence below sentence-presupposition mix = 47.49

BERTScore:

Sentences only (baseline): 0.339

I also finetuned a model with more training data (1440 examples) to see if how sentence number will affect the model since 3 presupposition datasets has more or less 3 times more data but the results are pretty similar to original sentence only baseline so I am not adding it.

	Sentence-presuppositions mixed	Presu
One presupposition	0.331	
Two presupposition	0.335	
Three presupposition	0.335	

Table 2: Table2

Sentence-presuppositions mixed and One presupposition = 0.331

Sentence-presuppositions mixed and Two presupposition = 0.335

Sentence-presuppositions mixed and Three presupposition = 0.335

Presuppositions below sentences and One presupposition = 0.331

Presuppositions below sentences and Two presupposition = 0.335

Presuppositions below sentences and Three presupposition = 0.333

Normal sentence below sentence-presupposition mix: 0.335

Results were suprising actually. I can understand why sentence-presuppositions mixed experiments gave low perplexity since presuppositions between actual sentences may influence coherency negatively. But I would expect models that fed by presuppositions would achieve higher results in BERTScore but baseline model was actually performed best with a 0.339 BERTScore. **Although one observation is that even though mixed models performed significantly lower in terms of perplexity, they have not much differance with baseline model and other models in terms of BERTScore it achieved. That situation may suggest that its contextual understanding is indeed developed because even though it has significantly (around 20 percent) less coherence (low perplexity) if it still performs nearly same with baseline model in terms of BERTScore.**

7 Error analysis

It is really difficult to compare the results of different experiments with eye test. However, all the models has same issues such as repeated sentences. Examples of inputs outputs are exists in ipynb files. On ipynb files, there is output of model (on top) and real completion of sentence. Prompts can be seen via executing this code:

for line in testData[:200]: prompt, expectedCompletion = splitString(line) print(prompt)

Also, models that has trained with presuppositions sometimes may generate presuppositions instead normal sentences such as:

İnput: (someone explaining a topic) Output: The speaker is knowledgeable about topic

8 Contributions of group members

All the work has done by me and by feedbacks of my instructor Gözde Gül Şahin.

9 Conclusion

The purpose of this project was to explore the impact of integrating presuppositions into the training of language models. This approach was novel and promising, as presuppositions provide additional context that could potentially enhance the model's ability to understand and generate text.

However, the results were not as expected. Although the models trained with presuppositions exhibited comparable BERTScores to the baseline model, the perplexity was generally higher, indicating a lower predictive performance. It was surprising to see that the baseline model achieved the highest BERTScore, as it was expected that the models trained with presuppositions would have a better semantic understanding and thus a higher score.

An interesting observation was that, even though the mixed models had significantly lower coherence, they still performed nearly as well as the baseline model in terms of BERTScore. This could suggest that while the sentence coherence may have been negatively affected by the insertion of presuppositions, the model's contextual understanding might have indeed been enhanced.

While this experiment did not yield the desired results, it does provide valuable insights for future work. For instance, it might be beneficial to further investigate the reason why presuppositions did not lead to an improvement in BERTScore, and how they negatively impacted coherence. Additionally, the idea of training models with presuppositions is still relatively unexplored, and it's possible that different implementation strategies or more complex models might yield better results.

The project also highlighted some recurring issues such as repetition in the generated sentences. Identifying and addressing these issues in future work could further improve the performance of language models.

Overall, this project sheds light on the potential and challenges of integrating presuppositions into the training of language models. It provides a foundation for future work in this area and underlines the importance of continued research to explore new ways to improve the semantic understanding and coherence of language models.

10 AI Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
  - I used little amount of chatgpt-4.

If you answered yes to the above question, please complete the following as well:

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.

- Prompts:

  * I gave my problem statement and what you proposed vs what you achieved part to chatgpt and asked to write a skeleton for mmy approach (which I edited dramatically based on my actual approach and I did not write any prompt for results subsection), followup to to that prompt, I wrote: can you add that I used perplexity and bertscore to evaluate models? also explain these methods briefly.

  * I gave My Approach section as prompt and asked a conclusion section, I was looking for a skeleton template to edit but actually AI performed well on this, I edited this less

- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?

  - It surely needs navigation and output should clearly be edited. Because it doesn't and can't know my project better than me.

Limitations

Limitations are caused by mainly openai and prompt engineering. it takes money and time to generate this presuppositions. Also I don't know if it counts as limitations but my model that has fed by presuppositions actually did performed worse than baseline model.

Ethics Statement

Scientific work published at ACL 2023 must comply with the ACL Ethics Policy.[1] We encourage all authors to include an explicit ethics statement on the broader impact of the work, or other ethical considerations after the conclusion but before the references. The ethics statement will not count toward the page limit (8 pages for long, 4 pages for short papers).

[1] https://www.aclweb.org/portal/content/acl-code-ethics

References:

Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations (Kobayashi, NAACL 2018)

Zhang X, Zhao J, Lecun Y (2015) Character-level convolutional networks for text classification

Generating Natural Language Adversarial Examples (Alzantot et al., EMNLP 2018)

Data Augmentation for Low-Resource Neural Machine Translation (Fadaee et al., ACL 2017)

Huong TH, Hoang VT (2020) A data augmentation technique based on text for Vietnamese sentiment analysis

EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks (Wei and Zou, EMNLP-IJCNLP 2019)

Coulombe C (2018) Text data augmentation made simple by leveraging NLP cloud APIs.

Submodular Optimization-based Diverse Paraphrasing and its Effectiveness in Data Augmentation (Kumar et al., NAACL 2019)

Improving Neural Machine Translation Models with Monolingual Data (Sennrich et al., ACL 2016)

Data Augmentation for Visual Question Answering (Kafle et al., INLG 2017)

References