Evaluating GPT-2 NPI ”Knowledge” in the French Language

Joseph Wargo
Former Student, University of Georgia
josephawargo@gmail.com

Abstract

There is healthy discussion and research on language model ”knowledge” of linguistic phenomena. Said research is typically limited to the English language. In this paper, we will evaluate a French language model's (Belgian GPT-2 (Louis, 2020)) ”knowledge” of Negative Polarity Items (henceforth referred to as NPIs). To do so, we generated 100 *minimal pairs*. Each pair features both a correct sentence with proper NPI licensing and an incorrect sentence with improper NPI licensing. We then judged the efficacy of the language model's NPI comprehension using the rate at which the model assigned a lower perplexity to the correct member of a *minimal pair*.

1 Introduction

Inquiries into language model grammatical ”knowledge” have ranged from studies of limited linguistic paradigms to multi-phenomena analysis. Several of these papers have aimed to quantify language model ”knowledge” by analyzing the model's preferences between *minimal pairs*. This paper draws heavily from BLiMP: The Benchmark of Linguistic Minimal Pairs for English (Warstadt et al., 2020), which is a paper that utilizes benchmarks for linguistic analysis.

In the style of BLiMP, analysis here will be done using *minimal pairs*. In this context, *minimal pairs* are two nearly identical sentences that differ in only one linguistic or grammatical attribute. For example:

I *am* walking down the street.

and

I *is* walking down the street.

are *minimal pairs* that differ on the main verb.

In BLiMP (Warstadt et al., 2020) and related studies, *minimal pair* analysis of a linguistic phenomenon is done by generating a dataset of *minimal pairs* that all differ on the same phenomenon. Researchers then assign a score to each sentence using a language model. If the language model assigns a better score to the correct *minimal pair* member, then that is deemed acceptable. This type of analysis is referred to as an *acceptability judgement*. The rate of correct *acceptability judgements* is then calculated. This percentage is somewhat indicative of the language model's ”knowledge” of the relevant linguistic phenomenon.

This paper's *minimal pair* analysis is done on a single phenomenon: NPIs. Analyses on language model ”knowledge” are rare in non-English languages. This is true for French. Thus, this paper aims to fill this void with such an analysis.

2 Background

2.1 NPIs

NPIs are words (or in some cases, two words) that are grammatically acceptable in only certain negative contexts. For example, three English NPIs are *any*, *ever*, and *all*. Take a look at the following sentence:

I did not eat any cookies.

This sentence makes grammatical sense. However, if we remove the word *not* like so:

I did eat any cookies.

the sentence no longer makes sense. Thus, *any* is an NPI. In the above sentence, it is *licensed* by negation. This means that negation is the element that enables the sentence containing *any* to make grammatical sense.

2.2 French NPIs

Like in most languages, NPIs are present in French. They can be tricky to identify. This is primarily because French negation in general is tricky. For instance, the basic form of negation in French involves two words, *ne* and *pas*. For example:

J'aime la télé. (I like TV.)

becomes

Je n'aime pas la télé. (I do not like TV.)

when negated.

As a non-native speaker, this is tricky to navigate. The dataset used for this analysis has been created and evaluated with this complication in mind.

An example of a French NPI would be the word *personne*. In the following sentence:

Je *n*'aime *personne*.

contains *personne*, which is licensed by *ne* (negation). Its *minimal pair*:

J'aime *personne*.

still contains the NPI, but there is no licensing. Thus, the sentence is not acceptable in French; it does not make grammatical sense.

3 Data

Using the frTenTen corpus and our own French grammatical knowledge, we generated 100 French *minimal pairs*. Each acceptable sentence in the pair sentences contained an NPI and differed from the unacceptable sentence only in the licensing of said NPI.

There are several ways NPIs can be licensed. Simple negation is often the most common in everyday language. Thus, each *minimal pair* in the dataset is through simple negation. Additionally, there are only two NPIs utilized, *personne* and *aucun*. A limited dataset restricts the conclusions that can be reached from this evaluation. This will be further discussed in the ”Conclusion” section.

To avoid mistakes due to limited French ”knowledge”, we had a native French speaker affirm that the acceptable sentence made sense and the unacceptable did not. The dataset was then adjusted per their feedback before being tested.

4 Model

Transformers are the cutting edge of language models. GPT-2 is a Transformer made by OpenAI (Radford et al., 2019). We chose to evaluate Belgian GPT-2, a GPT-2 language model pre-trained on a 60 gigabyte French corpus (Louis, 2020). Although the name indicates a language model exclusive to Belgian French, this is not the case. It was trained on a variety of French dialects. We ultimately chose to test Belgian GPT-2 because it had a sufficiently large training corpus.

Future editions of this paper will aim to evaluate other models, such as the delightfully named CamemBERT. CamemBERT (Martin et al., 2019) is a French language model based on Facebook's RoBERTa (Liu et al., 2019), which itself is a replication study of Google's BERT (Devlin et al., 2018). However, CamemBERT, just like BERT, is a Masked Language Model (MLM). MLMs are not designed to output probability distributions. [1] This makes *acceptability judgements* difficult to produce. There are methods with which researchers have calculated sentence probability using BERT (Wang and Cho, 2019). This will ideally be replicated for CamemBERT in future editions of this paper.

5 Methodology

This paper uses *minimal pair* analysis, the methodology discussed in this paper's introduction, to determine language model ”knowledge”. This section will add supplementary information.

5.1 Evaluation Metric

To evaluate Belgian GPT-2, we used perplexity as our metric. Perplexity measures the degree to which a probability distribution predicts a sample. Language models are essentially large probability distributions. Thus, perplexity is often used to evaluate overall language model performance by calculating the metric for a large corpus. It can also be calculated for a single sentence. We chose perplexity primarily because it is normalized by length. The licensing used for every *minimal pair* in the dataset is negation. Negation involves adding an additional word. Controlling for sentence length allows for a better comparison between *minimal*

[1] One of the creator's of BERT says here that doing so produces meaningless results (https://github.com/google-research/bert/issues/139)

pairs that differ in length. For perplexity, a lower score indicates better performance.

5.2 Code

The code and dataset used in this paper can be found here:

`https://github.com/josephwargo/NLPFinalProject/`

6 Results and Analysis

In evaluation, the Belgian GPT-2 language model assigned a lower perplexity score to the correct sentence of the *minimal pair* in 96 out of 100 cases, for an accuracy of 96%. This is unexpectedly high when compared to studies like BLiMP, in which GPT-2 assigned a lower perplexity to the correct sentence in only 78.9% of cases for the NPI dataset.

We have identified two potential reasons for this high accuracy. Firstly, the dataset is very simplistic. The sentences are relatively short and contain common words. In comparison, other benchmarking studies use longer and more grammatically complex sentences. Additionally, Belgian GPT-2 is trained on an enormous corpus. It could simply be a very effective language model, specifically in this context.

Of the 4 incorrectly chosen sentences, all contained the NPI *personne*. This could indicate that Belgian GPT-2 has an harder time with *personne* than *aucun*. Additionally, 3 of the incorrectly chosen sentences contained the first person singular pronoun *je*. This could indicate that Belgian GPT-2 may have a tougher time understanding sentences with this context. However, further analysis would be needed to draw any conclusions from this.

7 Conclusion

Ultimately, the conclusions that can be reached from this paper are limited. The dataset was small and uniform. Only one language model was evaluated. Future editions of this paper will ideally contain an expanded dataset and different models. Until then, the efficacy of this paper will be limited.

Still, if these results can be trusted, Belgian GPT-2 is quite adept at assigning a better score to sentences containing licensed NPIs than their unlicensed counterparts (in the specific parameters used when developing this dataset). That is a novel result which speaks to the power of the language model.

Acknowledgments

I would like to thank Dr. Hale for giving me the idea for this project and leading me down this rabbit hole. It has been a difficult yet rewarding journey, and I have learned so much. I would also like to thank Al Baker, who was kind enough to read over my dataset to ensure it made sense. A final thanks to Mitchell Ostrow for coding assistance.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Antoine Louis. 2020. BelGPT-2: a GPT-2 model pre-trained on French corpora. `https://github.com/antoiloui/belgpt2`.

Louis Martin, Benjamin Müller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *CoRR*, abs/1911.03894.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english.