Language Model Frame Filling for Low Resource Languages

Leon Zong
University of Rochester
lzong@u.rochester.edu

Abstract

Acceptability judgment tasks are key tools used in the understanding of the grammar of languages by linguists. A typical way to construct these tasks is to use semantic frames. We provide evidence that the fill-mask objective for language models (LMs) is a useful way to assist in generation of frames for low-resource languages. We train a model specifically for the task of mask filling on a low resource language, Zulu, and examine the performances of it and various other models on the fill-mask objective, for both high-resource and low-resource languages.

1 Introduction

Acceptability judgment data plays a crucial role in understanding the grammaticality of generated linguistic data (Schütze et al., 2013). While judgment tasks themselves can take on a variety of formats, they are all similar in that they provide a relative ranking of the generated language. This relative data can be further processed via various statistical tests into a grammaticality distribution (Bross, 2019). Parsing grammaticality information into understanding of grammar is the foundation of empirical linguistics.

The generation of language to be used in acceptability judgment tasks is also a well-studied process. Surveys often use a set of 'elicitation frames' for extracting specific types of grammaticality information from native speakers. An example elicitation frame from Berthelin (2020) is provided.

(1) anguniaq **-iaq** -tuq
hunting to.go.and IND.3SG
'He went hunting.'

(2) ? anguniaq **-hungnaq** -tuq
hunting probably IND.3SG
'I think he went hunting.'

Frames like these are often generated manually by researchers or researcher-employed native speakers once an interesting aspect of grammar has been noticed. For the example frame, the researchers wanted to check "if *hungnaq* 'probably' [could] cover epistemic modal meanings." While generating the frames themselves is a research-specific task (and often one which is foundational for the study), generating potential candidates for filling the frames to be used in an acceptability task is one that can be expanded upon.

Modern pre-trained language models like BERT (Devlin et al., 2019) and improved BERT models like RoBERTa (Liu et al., 2019) are trained on the Masked LM objective, analogous to the frame-filling described above. These analogous tasks are also referred to in linguistics as the 'Cloze task' (Taylor, 1953). Noticing the similarity between the empirical linguistic tool and the LM objective, we decided to conduct an analysis on the effectiveness of several pre-trained LMs at candidate frame filling.

Our investigation reveals that mask-filling LMs are effective in this task. We find that LMs pre-trained with a mask-filling objective can accurately fill a provided frame and that state-of-the-art chat large language models (LLMs) like ChatGPT can also fill frames accurately, even with just 0-shot prompting. We also pre-train a new LM for Zulu based on the RoBERTa architecture on the union of several Zulu datasets to measure performance on very-low-resource languages. With this model, we conclude that LMs with limited data on a language perform significantly worse on that language, suggesting that some fine-tuning or pre-training needs to be done for a model to be a useful tool.

2 Background

In this section, we provide a discussion on language models as a whole as well as brief overviews for each of the models used in the study.

2.1 Language Models

State-of-the-art performance on natural language processing (NLP) tasks has become dominated by Transformer-based models (Vaswani et al., 2023) after various implementations have shown large improvements over recurrent neural networks (RNNs). While RNNs with various memory units – long short-term memory / LSTM (Chung et al., 2014), gated recurrent units (Hochreiter and Schmidhuber, 1997) – had solid performance on NLP tasks, they are limited by their ability to train in parallel.

Transformers, on the other hand, can be trained significantly in parallel. The original Transformer model in Vaswani et al. (2023) was composed of both an encoder and decoder. Implementations of models using Transformers do not nessecarily need both an encoder and decoder stack, however.

2.2 BERT

Bidirectional Encoder Representations from Transformers, or BERT, is an encoder-only Transformer model (Devlin et al., 2019). BERT had a few key model and training features which allowed it to acheive a better language representation. Primarily, BERT was pre-trained on two unsupervised tasks – Masked LM, as discussed in the introduction, and next sentence prediction (NSP). BERT was pre-trained on two document-level corpora. Pre-training was an innovation to allow an initial language representation to be learned and re-used for various down-stream tasks.

2.3 RoBERTa

Following BERT's success on many NLP tasks, a more optimized pre-training was conducted and named RoBERTa (Liu et al., 2019). RoBERTa used the same model design as BERT. However, as improvements, the team pre-trained RoBERTa on more data, with more compute, and with dynamic masking. Most notably for our research, RoBERTa was pre-trained solely on the Masked LM objective.

2.4 GPT

Generative pre-training is a technique implemented by (Radford and Narasimhan, 2018) to pre-train a decoder-only Transformer model. The unsupervised task used in GPT models is to maximize the likelihood of the next token in a sequence of tokens. Increased scale in models by the same team led to GPT-2 (Radford et al., 2019) as well as GPT-3 (Brown et al., 2020), which show increased performance in nearly all NLP tasks. Additional training via human feedback (reinforcement learning through human feedback), proposed by Christiano et al. (2023) and implemented in InstructGPT (Ouyang et al., 2022) has increased performance even further.

OpenAI has since released GPT-3.5 and GPT-4, which each increased the model's capabilities/scale. In our research, we will be using GPT-3.5.

3 Related Work

Several studies have already been conducted regarding the performance of language models on non-English languages. Chai et al. (2022) find interesting results in the zero-shot performance of various pre-trained LMs on novel languages. They find that word order, like subject-verb order, has little effect on multilingual shared performance, but that composition, the ability of language to combine (generally), is the key to shared multilingual performance.

Doddapaneni et al. (2021) conducted a survey of many multilingual LMs, trained on between 12 and 110 languages. They also find that models have a limited ability for language crossover – in other words, models pre-trained on a set of languages can transfer some performance to tasks on an unseen language. This might suggest that LMs might be able to be used for low-resource languages in related language families, as enough performance can crossover.

Some of the key multilingual masked LMs include multilingual BERT and XLM-RoBERTa. Multilingual BERT (mBERT) was released by Devlin et al. (2019), and was trained on Wikipedia text in over 100 languages. mBERT's performance was evaluated by Pires et al. (2019). They also find an ability to perform in zero-shot non-finetuned other language tasks. XLM-RoBERTa is another multilingual masked LM which outperforms BERT in many tasks (Conneau et al., 2020). It was trained on a significantly larger amount of training data (2TB, Common Crawl), and is notable for performance on low-resource languages like Swahili.

4 Frame Filling

This section describes the novel process we used for evaluating models on their ability to fill syntactic frames. While various cloze-style tasks exist, such as those proposed by Mostafazadeh et al. (2016) and Donahue et al. (2020), we generate a simpler

cloze task for the sole purpose of analyzing models' ability to generate appropriately filled frames. We then evaluate various models on this new task, for both English and Zulu.

4.1 Frame Filling Task

The new task we propose greatly resembles both the elicitation frames discussed in the Introduction as well as the Masked LM objective discussed in the section on BERT. For each sentence in the parallel English-Zulu corpora 'No Language Left Behind' (Team et al., 2022; Schwenk et al., 2020), we randomly mask a single word or sub-word token. The selected token is chosen from a tokenization, generated by the an unbiased (not one of the selected models') tokenizer.

While this introduces potential sampling error due to random frame generation, we hope that the larger size of the initial parallel corpora ($4,000,000$ sentences) will average out the error. However, a future improvement for this type of task could be to synchronize the masked token between the two languages. This would involve having a token-level annotation for parallel tokens, or using some machine translation model.

For each model, we use HuggingFace's 'fill-mask' pipeline for generating the top set of results for a given masked frame. For RoBERTa, zulBERTa, and XLM-R, which are all BERT-based, we can fill the generated masked frames out of the box. For GPT-3.5, we used the prompt design mentioned above to fill the frame.

4.2 Evaluation

The metric chosen for evaluating the task needed to be able to describe the ability of the model to successfully fill the given frame. We take inspiration from Petroni et al. (2019), who use a similar cloze task to measure factual and common-sense knowledge and evaluated their models with mean precision at k. To quote their study, "We use the mean precision at k (P@k). For a given fact, this value is 1 if the object is ranked among the top k results, and 0 otherwise." This requires that we return k results for each frame. We chose $k = 10$.

5 Training

This section details the data and parameters used to pre-train a Zulu RoBERTa model, which we will refer to as zulBERTa. We also discuss the development of the GPT-3.5 prompt, and the other models in use.

Subcorpora Name	Sentences	Words
LEIPZIG	1,337,288	8,957,739
LORELEI	878,828	4,942,893
NCHLT	75,271	1,483,185
UKWABELENA	58,847	288,106
ZULU WIKI	57,857	535,732
Total	2,408,091	16,207,655

Table 1: Size of subcorpora in collection

5.1 Data

As much Zulu training data was collected as possible. This section lists the corpora used and the data processing steps taken. Table 1 lists the number of words and sizes for each sub-corpora.

NCHLT isiZulu Text Corpora

Eiselen and Puttkammer (2014) collected, annotated and developed tools for a corpora of 10 South African languages. They scraped data from government websites, some news and magazine articles, and some scientific articles. Minimal processing was applied.

Leipzig Corpora

Goldhahn et al. (2012) collected and publish corpora for 136 languages. Corpora were collected from crawling newspapers and websites. The data is heavily cleaned – language identification, deduplication, and non-sentence removal was used.

Ukwabalena

Spiegler et al. (2010) collected and annotated a small Zulu-specific corpora. Unfortuantely, though Ukwabalena means *to share* in Zulu, the links in the paper are dead. The unlabeled portion of the corpora was recovered from an `archive.org` backup.

Wikipedia

A dump of the Zulu Wikipedia was downloaded from the Wikimedia foundation. We wrote a Python script to parse the raw XML data into our corpora's format using the Python package `mwparserfromhell`. No cleanup of the raw text extract was performed.

DARPA LORELEI

The Low Resource Languages for Emergent Incidents (LORELEI) program collects corpora for many low-resource languages. The data in the Zulu

corpora is found from 'discussion forum, news, reference, social network, and weblogs' (Tracey et al., 2023). No information on cleaning is found.

Data processing

We chose to use a serialized data storage for our corpora collection – Concrete (Ferraro et al., 2014). Concrete is based off of Apache Thrift, and has custom definitions for NLP annotations. We used a script to automatically parse each sub-corpora based on how each's data is presented.

For example, document-level corpora like Wikipedia are maintained in documents (Concrete's *Communication*), whereas sentence-level corpora are parsed line-by-line into their own *Communication*. A unique ID and metadata from the original subcorpora are also stored. No additional data processing, like deduplication, was performed.

5.2 Model Details

As the name zulBERTa implies, we use model parameters inspired from RoBERTa (Liu et al., 2019). We trained a byte-level Byte-Pair Encoding (BPE) tokenizer (Sennrich et al., 2016) (Radford et al., 2019) with a vocabulary size of 50k. We chose a smaller hidden-layer size ($12 \rightarrow 8$) as well as a smaller batch size ($8,000 \rightarrow 128$) due to out-of-memory constraints. We also adjusted training parameters manually and arrived at those listed in table s1.

Again following Liu et al. (2019), we chose to mask dynamically (at every batch input) at 15%, where 80% of those masks were replaced with '<mask>', 10% were replaced with a random token from the vocabulary, and 10% were left unchanged.

Training was performed on Google's v2-TPU Nodes (Norrie et al., 2021) for around 18 hours. HuggingFace with PyTorch and Accelerate were the primary libraries used to implement the model along with the training. A graph of the model loss is provided in figure something.

5.3 Prompt design

We chose to use GPT-3.5-turbo as a reference generative LLM. We tested several, and while many were effective in prompting GPT-3.5 to return properly formatted and correct mask filling results, we ultimately decided on this prompt, taking advantage of 'JSON mode' in GPT-3.5-1106.

Hyperparam	**RoBERTa$_{\text{BASE}}$**	**zulBERTa**
Number of Layers	12	8
Hidden Size	768	768
Attention heads	12	12
(Attn.) Dropout	(0.1) 0.1	(0.1) 0.1
Warmup Steps	24k	0
Learning Rate	6e-4	5e-4
Batch Size	8k	64
Weight Decay	0.01	0.01
LR Decay	Linear	Linear

Table 2: Params of zulBERTa, compared to RoBERTa

System: You will behave as if you were a multilingual masked language model like BERT. Given some text with a word or token replaced with the mask token '<mask>', return the 10 most likely tokens which the mask could be. Respond in JSON.

5.4 Other models

We use 2 additional LMs for providing a reference for our novel frame filling task. These include (1) RoBERTa (Liu et al., 2019), our English-only masked LM reference model and (2) XLM-RoBERTa (Conneau et al., 2020), our multilingual masked LM refernce model. Each was used out-of-the-box from the pre-trained checkpoint on HuggingFace. Since they are both trained with the masked LM objective, they can be used to perform mask filling without any additional adjustment.

6 Results

We find that the novel frame filling task provides a good reference for masked LM performance. Results over various values of k are found in Figure 1. All models except our Zulu-only pretrained model perform very well on the English frame filling. Figure 2 shows specific performance of models at k=1. We use k=1 as the baseline result, as the model performance does not ordinally change at different values of k. We note the best-performing model on the English task is RoBERTa, and the best-performing model on the Zulu dataset is ours, zulBERTa. However, zulBERTa performs relatively worse on Zulu than even the worst-performing pre-trained model on English, GPT-3.5.

The precision @ k metric shows the expected performance increase as k increases in all models. It can be noted that GPT's performance increases relatively less after k=4. Manual inspection of the prompt results reveals that often, GPT will generate

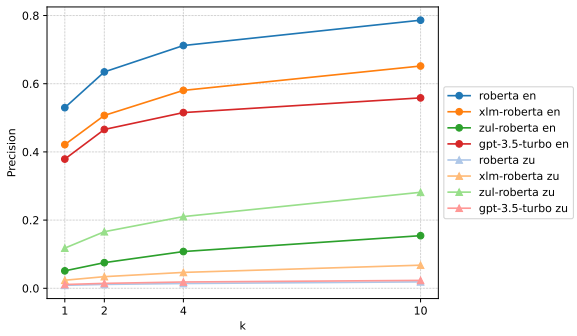

Figure 1

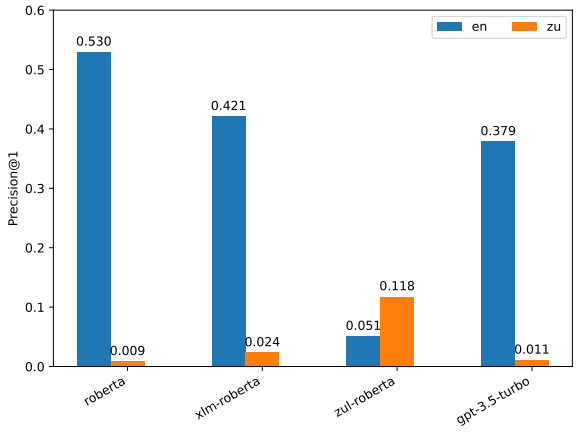

Figure 2: Enter Caption

the same fill result multiple times. This can likely be solved with better prompt design.

We additionally identify that zulBERTa has a relatively higher performance on English than the other models do on Zulu. This could indicate that additional cleaning should be done on the Zulu model's training data, as it is likely that English data was trained on.

7 Discussion

Our results shows that accurate frame filling can be achieved on high-resource languages like English with no additional tuning on state-of-the-art LMs. We find that separate pre-training provides much higher accuracy on frame filling in languages with little to no widespread data in corpora. Comparing model performance, we show that causal large language models like GPT with no masked LM objective can nearly match the performance of masked language models like RoBERTa. Additional prompt tuning, like providing 1-shot or n-shot examples, will likely improve accuracy further.

In terms of usage of language models by linguists for the purpose of filling frames, it seems likely that they can assist in early-phase development of candidate frames. It is crucial that a model becoming a tool for this situation can generate both unacceptable (easy) and acceptable (hard) results for a particular frame. As a tool, one could ask the model to generate a large number (hundreds) of candidates, and evaluate them manually with a native speaker.

One interesting development is that the most interesting acceptability judgments are those somewhere in between acceptable and unacceptable. While this is impossible to measure with our proposed task, this can be an interesting development. An interesting idea is to combine the output of a frame-filling model with real results from empirical data to combine the generation of frames with the acceptability of each frame.

We believe that these results provide groundwork for an improved connection between empirical and computational linguistics. Frame/mask filling is common to both fields, and as such can be used by both. Language models can assist empirical linguists in collecting more and/or better data and using empirical judgments can create better language models.

Limitations

There are a few limitations in our analysis. With regards to the frame filling task, there is potential sampling error in the initial generation of frames, which was done randomly. Additionally, we believe a potential improvement exists by using whole-word masking. However, it is unclear how whole-word masking would be adapted to models who unmask tokens instead of words.

We also believe significant room for improvement in the GPT-3.5 performance on this task. Additional prompt engineering, or simply using example (multi-shot) prompting could improve performance significantly. It remains difficult, as sending 40,000 frames at around 2 seconds per prompt takes around an entire day. However, fitting multiple frames into the same prompt provides degraded performance from my limited testing. There is no batch completions API for the chat models (GPT-3.5 and up).

Finally, we believe that zulBERTa also has room for improvement. Dataset cleanup, more robust hyperparameter selection (via search, etc), and increased compute can lead to better performance.

References

Signe Rix Berthelin. 2020. Semantic elicitation frames and their application.

Fabian Bross. 2019. Acceptability ratings in linguistics: A practical guide to grammaticality judgments, data collection, and statistical analysis. (1.02).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yuan Chai, Yaobo Liang, and Nan Duan. 2022. Cross-lingual ability of multilingual masked language models: A study of language structure.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. A primer on pretrained multilingual language models.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks.

Roald Eiselen and Martin Puttkammer. 2014. Developing text resources for ten South African languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3698–3703, Reykjavik, Iceland. European Language Resources Association (ELRA).

Francis Ferraro, Max Thomas, Matthew R. Gormley, Travis Wolfe, Craig Harman, and Benjamin Van Durme. 2014. Concretely annotated corpora. In *4th Workshop on Automated Knowledge Base Construction (AKBC)*.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

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.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories.

Thomas Norrie, Nishant Patil, Doe Hyun Yoon, George Kurian, Sheng Li, James Laudon, Cliff Young, Norman Jouppi, and David Patterson. 2021. The design process for google's training chips: Tpuv2 and tpuv3. *IEEE Micro*, 41(2):56–63.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert?

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

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

Carson T Schütze, Jon Sprouse, Robert J Podesva, and Devyani Sharma. 2013. Judgment data. *Research methods in linguistics*, pages 27–50.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2020. Ccmatrix: Mining billions of high-quality parallel sentences on the web.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units.

Sebastian Spiegler, Andrew van der Spuy, and Peter A. Flach. 2010. Ukwabelana - an open-source morphological Zulu corpus. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1020–1028, Beijing, China. Coling 2010 Organizing Committee.

Wilson L. Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jennifer Tracey, Stephanie Strassel, Dave Graff, Jonathan Wright, Song Chen, Neville Ryant, Seth Kulick, Kira Griffitt, Dana Delgado, and Michael Arrigo. 2023. Lorelei zulu representative language pack. website.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

A Appendix

The code and models will be made public at a later date, as I need to cleanup the training pipelines as well as write a model card, check corpora usage license, etc. For now, email me at lzong@u.rochester.edu.