Language Transfer? I Hardly Know 'er: Synthetic Data Augments Transfer Learning in Low-Resource Settings

Alessio Tosolini [*†] tosolini@uw.edu
Ben Kosa [*] bkosa2@uw.edu
William Galvin [*‡] wgalvin@uw.edu

Abstract

Recent advancements in language models have been primarily realized for high-resource languages (i.e., languages with large corpora of training data) as the size of both models and training datasets increase. To apply language models to low-resource languages, transfer learning from a high-resource language (usually English) has been the de facto technique. In this work, we show that the success of transfer learning varies between target languages, and propose an alternative pretraining regime: using synthetically generated data instead of English data. Finally, we show that supplementing fine-tuning data with synthetic data can improve fine-tuning results when performing transfer learning from an English model.

1. Introduction

One of the major limitations of transformer architectures is that they're exceedingly data hungry, requiring lots of data to model a language well. This poses a problem for researchers trying to model low-resource languages, as traditional approaches using transformer architectures fall short. A couple of strategies have been used to mediate this problem with various degrees of success. Language transfer refers to the process by which a model trained on one language performs tasks in another language that shares features. Language transfer has been shown to improve accuracy in translation tasks [6] and other NLP tasks such as Named Entity Recognition and Semantic Text Similarity [10]. Another common strategy is augmenting the training data, which has also been proven to increase machine translation accuracy [7] and other NLP tasks such as hate detection [5]. The last strategy documented by [9] combines these approaches, generating synthetic training data from an artificial language with grammatical similarities to the target language to improve language transfer in a low-resource setting.

The degree to which transfer learning depends on linguistic similarity between the (high-resource) source language(s) and the (low-resource) target language remains unknown. Furthermore, it remains to be seen if a model pretrained on synthetic—yet linguistically principled—data and fine-tuned on gold standard data can overcome the problem of data scarcity when modeling low-resource languages.

2. Methods

We expand upon [9] work in the following ways: (i) we create a grammar for a probabilistic context free grammar (PCFG) that resembles the target language in the main linguistic parameters (e.g. word order, headedness, etc.), (ii) our lexicon is taken from the target language, and (iii) we compare a randomly initialized transformer with an English-adapted one to analyze whether the ability to generalize depends on the relatedness of the target language with the transformer's original training language. In this way, we are training a model on a grammatically, lexically, and semantically simplified version of the target language, attempting to model only syntactic structure in our synthetic training data. By fine-tuning on gold standard data, we hope to demonstrate that using linguistically informed methods, such as hand-coding PCFGs to generate grammatical simple sentences, can increase our ability to model extremely low resource languages.

2.1. PCFGs

To generate the synthetic data, we used Probilabilistic Context Free Grammars (PCFGs). PCFGs are useful in generating stochastic sentences that are syntactially sound, resembling human languages. Our PCFG production rules fall under one of two classes: (i) branching rules and (ii) feature rules.

Branching rules take one state and turn it into one or more states stochastically. For example, on lines 3, 5, and 7 you see the branching rules for sentences, noun phrases,

[*]Paul G. Allen Center for Computer Science & Engineering, University of Washington
[†]Department of Linguistics, University of Washington
[‡]Department of Applied Mathematics, Univeristy of Washington

and verb phrases respectively. When a state goes through a branching rule, the next states are determined stochastically. For example, a VP state representing a verb phrase may break into a verb followed by a NP for transitive constructions with probability 70% or simply a verb for intransitive constructions with probability 30%. Once a state is a part of speech, such as "noun", the following state is a terminal state representing an uninflected form of the word. Feature rules take one input state and output one state, adding a feature to the output state. These added features get passed down to all children states (e.g. NPs tagged with the feature "nom" pass the feature down to all children states). For example, on line 10, you see the feature rule adding nominative case ("nom") to a subject noun phase (sNP). The output of that rule is a noun phrase tagged with the feature nominative.

After the uninflected words are generated, any preterminal state that must agree with other words gets inflected. For example, verbs in English agree with the number and person of the nominative constituent. For brevity, pseudocode for agreement rules, inflection rules, and importing the vocabulary were not shown here.

Listing 1. Generation Rules for a Toy English Grammar

```
1  # BRANCHING RULES
2  # Sentences
3  S: [sNP, VP], 1
4  # Noun phrases
5  NP: [det, noun], 0.5, [pron], 0.5
6  # Verb phrases
7  VP: [verb, NP], 0.7, [verb], 0.3
8  # Rules for adjective, prepositions, ...
9  # FEATURE RULES
10 # Subject noun phrases
11 sNP: [NP.nom], 1
12 # Rules for number, gender, person, ...
```

We had to significantly simplify the grammars of each language, but we encoded the following grammatical features in each language, when applicable [1] :

- Nouns, pronouns, verbs, adjectives, adpositions [2] , and determiners.
- All non-periphrastic indicative tenses.
- Noun plurality, all pronoun persons,
- Subject-verb agreement, adjective-noun agreement, and determiner-noun agreement.

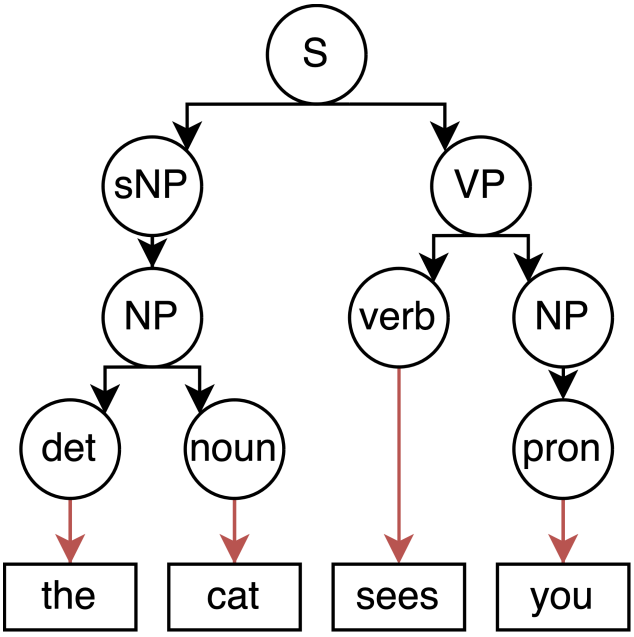

Figure 1. Image of a syntax tree generated by the PCFG pseudocode in Listing 1.

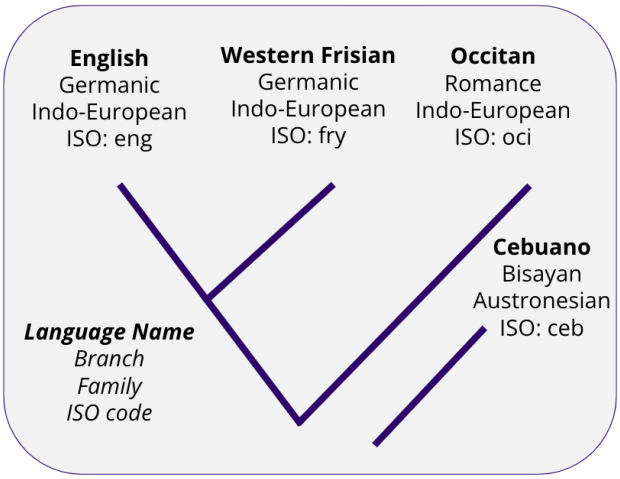

Figure 2. Language similarity by language family of English and our three low resource languages.

2.2. Data

We use three low-resource languages: Western Frisian (very closely related to English), Occitan (more distantly related), and Cebuano (very unrelated). While English and Western Frisian derive from Old English and Old Frisian

[1] Not all of our languages had all the features. For example, in Cebuano, verbs don't inflect for the number of the subject.

[2] Our Cebuano grammar does not contain adpositions, but to compensate for this syntactic simplicity we encoded the language's free word order.

respectively, Western Frisian is English's closest living relative with around 80 percent lexical similarity. Together, they form the Anglo-Frisian branch of the West Germanic language family, which is nested in the Germanic language family. Occitan, on the other hand, is under the Italic language family. Both the Germanic and Italic language families are members of the overarching Indo-European language family. Cebuano is the furthest from English, being a language that is part of the Austronesian language family, a language family separate to the Indo-European language family.

The corpora for these languages [1–3] contain between 100,000 and 200,000 unique sentences from modern language use. The corpora that we used for Western Frisian and Cebuano contained text that was taken from various community resources. The corpus for Occitan was scraped from Wikipedia. We treat sentences as the fundamental unit of language on which or models are fine-tuned—in practice, this simply means concatenating a special *end-of-sequence* token to each sentence.

2.3. Evaluation Metric

We evaluate or models on the *perplexity per word* metric [4], a standard metric in NLP derived from information theory. Perplexity is a measurement of the accuracy of a model's ability to predict a sample; a sample that is likely to occur should have a lower perplexity than an unlikely sample. More concretely, the perplexity of a discrete probability distribution, p, is given as

$$PP(p) = \prod_{i=1}^{N} p(x_i)^{-p(x_i)} = 2^{H(p)}$$

where x_i is an event in the sample space, Ω, and $H(p)$ is the standard information-theoretic Shannon entropy of the distribution, given by

$$H(p) = \sum_{i}^{N} p(x_i) \log p(x_i)$$

Entropy is considered a measurement of the information gained by actualizing a random variable.

As is standard, we use this framework to evaluate a model by computing its perplexity with respect to a testing set withheld from training—intuitively, models with low perplexity have better learned the distributions from which the testing sets are drawn, and are thus less "perplexed" by them.

In practice, the discrete probability distribution, p, that describes a natural language is unknown. Language models aim to learn $\tilde{p} \sim p$. To measure the perplexity of $\tilde{p}$ given samples $x_0, x_1, \ldots, x_N$ drawn from p, we can use the formula

$$PP(\tilde{p}) = \left(\prod_{i}^{N} \tilde{p}(x_i) \right)^{-1/N}$$

We adopt the typical NLP reformulation of this as

$$PP(\tilde{p}) = \left(\prod_{i}^{n} \tilde{p}(s_i) \right)^{-1/N}$$

where $s_0, s_1, \ldots, s_n$ are *sentences* in a corpus with N total *words*. This normlizing term N, allows us to compare perplexity accross corpora and between models.

In this work, we consider perplexity as our only evaluation criterion: models with lower perplexity are "better." While more sophisticated evaluation metrics exist—particularly for models that are able to generate text in response to prompts—we believe that perplexity offers a principled foundation on which to build.

2.4. Models and Fine-tuning

We use GPT-2 [8] as our model architecture. To evaluate how dependent transfer learning is on the language similarity between the target language and the language the transformer was pretrained on, we start with 3 different GPT-2 models in our experiment: (1) GPT-2 that has been pretrained on the full WebText dataset [8] (∼40GB of English text), (2) GPT-2 that has been pretrained on only 5 million English sentences (∼1GB of English text) that has been taken from the Book Corpus dataset [11] and (3) 3 GPT-2 models that has been pretrained only on synthetic data that was generated from our PCFG-based sentence generator. For pretraining GPT-2 on only synthetic data, we use 5,000,000 grammatically correct synthetic sentences in Western Frisian, Occitan, and Cebuano that we generated from our PCFG-based sentence generator.

To fine-tune and evaluate these models, we use the gold-standard corpora for West Frisian, Occitan, and Cebuano and withhold 1,000 sentences from each language for evaluation. We then fine-tune our models on varying quantities of the gold-standard data for one epoch, with none of the pretrained weights frozen.

3. Results

3.1. Perplexity vs Quantity of Data on Full English GPT-2

In our experiments, we first investigate whether transfer learning from a large English dataset is possible for our three low resource languages by fine-tuning our WebText pretrained GPT-2 model on varying amounts of gold standard Frisian, Occitan, and Cebuano. In 5 we see that after fine-tuning the English Web-pretrained GPT-2 model on Frisian, Occitan, and Cebuano, perplexity decreases with

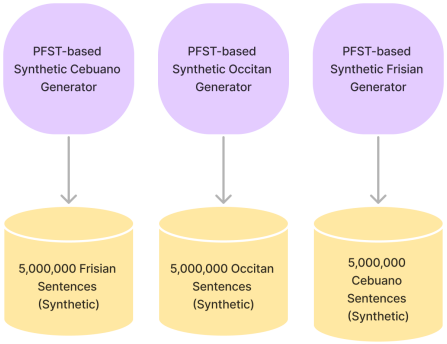

Figure 3. A visualization of how our synthetic data generation pipeline.

the amount of gold-standard training data seen[3]. This validates that transfer learning is possible in this context. See *Discussion* for more analysis.

3.2. Training GPT-2 with Synthetic Data

We then train, from scratch, GPT-2 using synthetic data to investigate the impact of language similarity on the effectiveness of transfer learning for our three low resource languages. We do so by initializing the same GPT-2 architecture with random weights and generating 5 million synthetic sentences from each language on which to train. For consistency, we also sampled 5 million English sentences from BookCorpus, to provide a baseline for the capacity of an English language model with similar training data size.

3.3. Perplexity vs Quantity of Data on Synthetic GPT-2

Next we replicate the first experiment, but using the synthetic GPT-2 models. As we can see in 6, this training routine is effective but less smooth than observed in 5 (the fully trained GPT-2 English). We can see that while the optimization curves are not particularly smooth, the synthetic data routine out performs transfer learning for each language.

3.4. Hybrid Gold-Standard and Synthetic Data

Finally, we sought to show a practical benefit to synthetic data by using it to augment gold-standard data. To do so, we create hybrid datasets with equal parts gold-standard and synthetic data, and fine-tune the fully-trained English GPT-2 model. We do not condition the model on the origin

[3]Since the corpora for each language contain varying amounts of data, the number of points on 5 is different for each language.

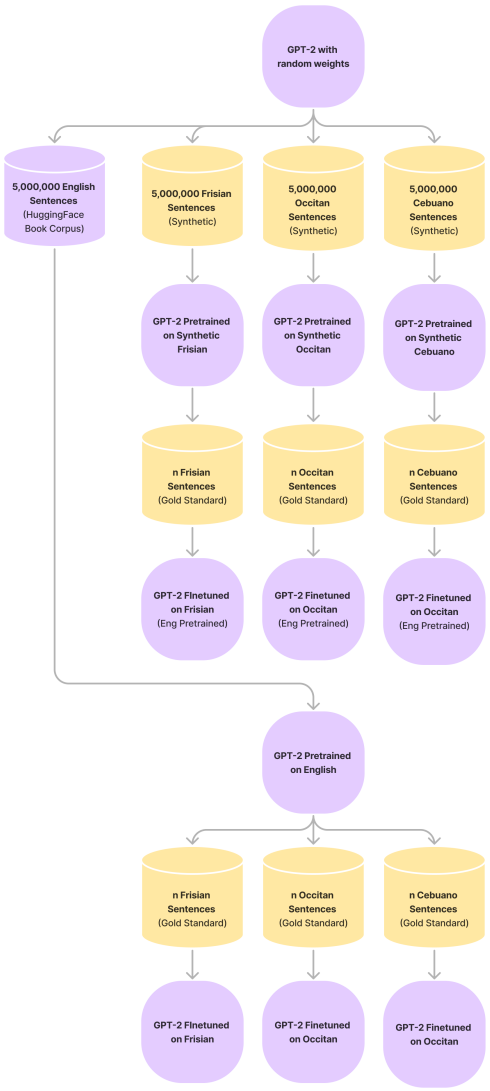

Figure 4. A visualization of our model training pipeline for all 3 of our experiments.

of each sentence. In 7, we see that this results in modest gains for Occitan and Cebuano, while severely hurting performance for Frisian.

4. Discussion

In 5 we see that for each language, perplexity decreases at approximately the same rate—i.e., the gap between the perplexity of Frisian and Occitan remains approximately constant. For Frisian and Occitan, we also observe the beginning of convergence. What remains to be seen is whether

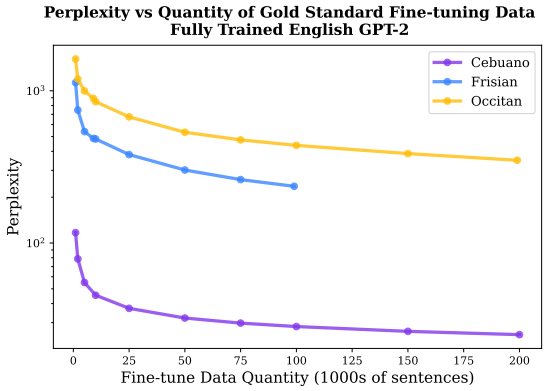

Figure 5. Perplexity vs quantity of data for fine-tuning GPT-2 pretrained on WebText (40GB of English text) on gold standard data only (no synthetic data).

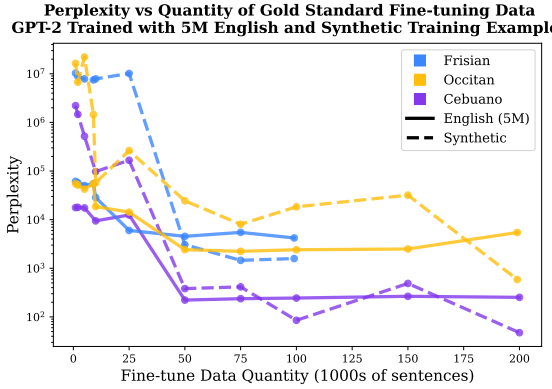

Figure 6. Perplexity vs quantity of data for fine-tuning GPT-2 on gold standard Frisian, Occitan, and Cebuano after either pretraining on 5M English sentences from BookCorpus or 5M synthetically generated sentences from Frisian, Occitan, and Cebuano respectively.

or not the gap remains as perplexity converges, or if they converge to the same value. While there is, by definition, not enough gold-standard data to thoroughly test this for low-resource languages, further study is needed.

Nonetheless, from 5 we can conclude that GPT-2 trained on English data better transfers to Frisian and Cebuano than to Occitan. We originally hypothesized that this would be the case for Frisian and Occitan, as Frisian is more closely linguistically related to English. However, the fact that Cebuano—which is very dis-similar to English—fine-tunes as-well or better than Frisian indicates that our hypothesis cannot fully describe the observed phenomena. Further study could replicate our methods but expand to dozens or hundreds of languages that need not be low-resource. In doing so, future work should take a more fine-grained approach to categorizing the similarity or lack thereof between languages, and perhaps consider foundation LLMs trained

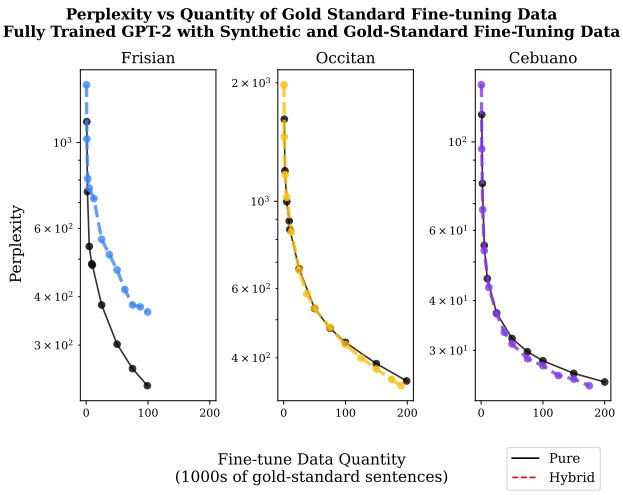

Figure 7. Perplexity vs quantity of data for fine-tuning on pure gold standard training data vs training data mixed with synthetic data for Frisian, Occitan, and Cebuano. The base model used was the fully trained GPT-2 .

in languages other than English.

When we considered GPT-2 trained on very few English sentences 6, we see that the English models tend to start with lower perplexity and plateau more quickly than their synthetic counterparts. While final perplexity values for the synthetic models are lower than the English-5M model, they are still much, much higher than the fully-trained English model. This shows that—at least at this scale—purely synthetic data may not be immediately helpful.

Our most practical contribution comes from considering the hybrid synthetic-gold-standard datasets for fine-tuning a fully trained English model, as seen in 7. We show that for Occitan and Cebuano—the two languages least similar to English—mixing in synthetic data lowers perplexity. This indicates that by beginning to learn the synthetic data distribution, the model also learns about the gold-standard distribution.

5. Future Works

Future work should consider language models that are significantly larger than GPT-2; pretraining and transfer learning using a much larger dataset of synthetically generated sentences (e.g. of equivalent size to WebText) in the likely case that the performance of our synthetic data suffered from using too little data for the architecture that was used; languages (both as sources and targets in transfer learning) that are more numerous and linguistically diverse; synthetic data generation algorithms that capture different grammatical and semantic aspects of a language; and varying quantities of synthetic data and varying sources of gold-standard corpora. If recreating our experiments, future

work should attempt to generalize between the patterns seen in figures 5 and 6 by more exhaustively testing with respect to the quantity of English pretraining data; and explore different ratios of synthetic-to-gold-standard data used in the hybrid experiments.

References

[1] Leipzig Corpora Collection. Cebuano community corpus based on material from 2017. https://corpora.uni-leipzig.de?corpusId=ceb_community_2017. Accessed: 2024-03-10. 3

[2] Leipzig Corpora Collection. Leipzig corpora collection (2020): Frisian news. https://corpora.wortschatz-leipzig.de/en?corpusId=fry_news_2020, 2020. 3

[3] Leipzig Corpora Collection. Leipzig corpora collection (2023): Occitan community. https://corpora.wortschatz-leipzig.de/en?corpusId=oci_community_2023, 2023. 3

[4] Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 1977. 3

[5] Aman Khullar, Daniel Nkemelu, V Cuong Nguyen, and Michael L Best. Hate speech detection in limited data contexts using synthetic data generation. *ACM Journal on Computing and Sustainable Societies*, 2(1):1–18, 2024. 1

[6] Toan Q Nguyen and David Chiang. Transfer learning across low-resource, related languages for neural machine translation. *arXiv preprint arXiv:1708.09803*, 2017. 1

[7] Seokjin Oh, Woohwan Jung, et al. Data augmentation for neural machine translation using generative language model. *arXiv preprint arXiv:2307.16833*, 2023. 1

[8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[9] Ryokan Ri and Yoshimasa Tsuruoka. Pretraining with artificial language: Studying transferable knowledge in language models. *arXiv preprint arXiv:2203.10326*, 2022. 1

[10] Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. Transfer to a low-resource language via close relatives: The case study on faroese. *arXiv preprint arXiv:2304.08823*, 2023. 1

[11] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 3