Language Transfer? Hardly Know 'er: Exploring the Utility of Synthetic Training Data in Low-Resource Settings

Alessio Tosolini, Ben Kosa, William Galvin

Problem

Currently, nothing outperforms Transformers when it comes to modeling language...

...except for low resource languages.

Transformers are data hungry!

Language transfer has shown to help improve accuracy, but two questions remain unanswered:

1. **What degree does transfer learning depend on linguistic similarity between the (high resource) source language(s) and (low resource) target language?**
2. **Can we overcome data scarcity for modeling low-resource languages if we finetune on synthetic–yet linguistically principled–data?**

How do we generate linguistically principled synthetic data?

For each low resource language, we can create a grammar for a **Probabilistic Finite State Transducer (PFST)** that resembles the target language in the main linguistic parameters (e.g. word order, headedness, etc). Inflection and agreement are handled separately.

```
# Sentences always lead to a subject noun phrase + verb
S → [sNP, VP], 1
# Subject noun phrases are nominative noun phrases
sNP → [NP.nom], 1
# A noun phrase may be a determiner + noun or a pronoun
NP → [det, noun], 0.5, [pron], 0.5
# Verb phrases may take an object with 70% probability
VP → [verb, NP], 0.7, [verb], 0.3
```

The following PFST may generate the following sentence:

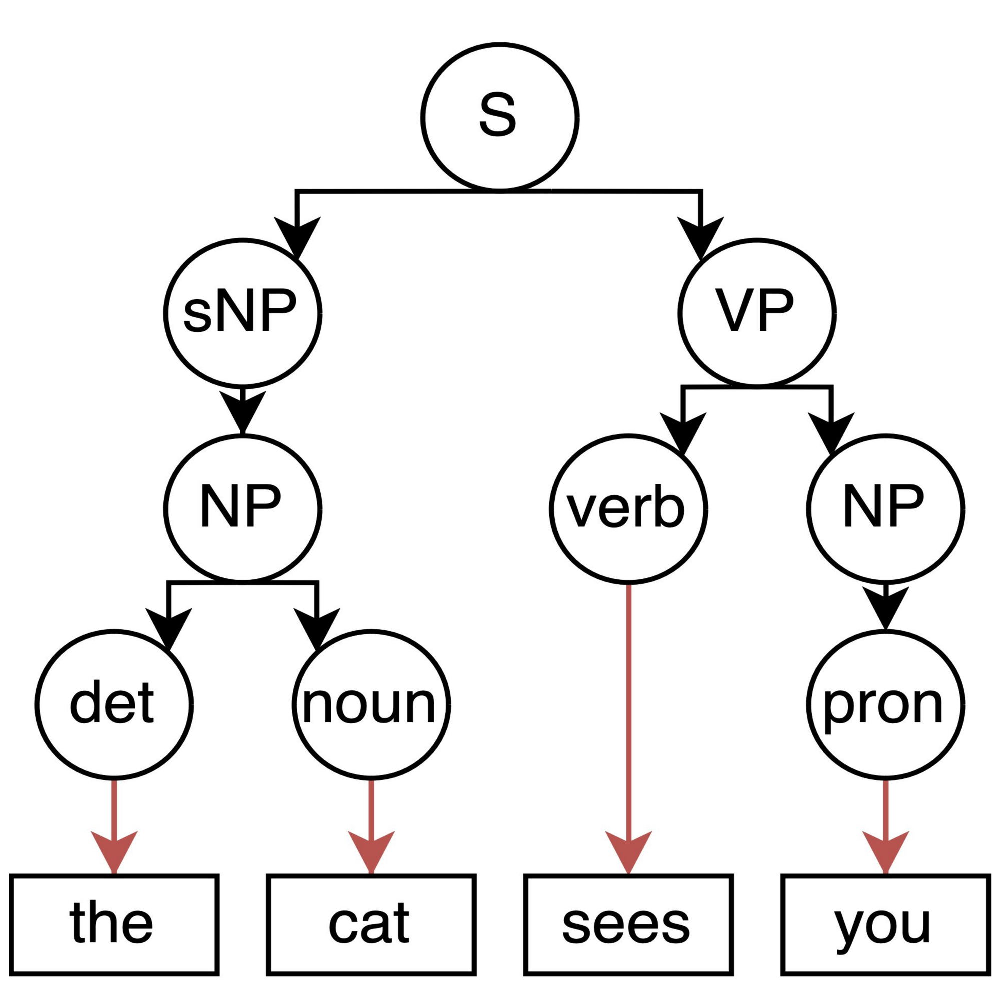

Figure 1: Example of a PFST

Data

We use 3 low resource languages with different linguistic relationships to English:

- Western Frisian (same branch)
- Occitan (same family)
- Cebuano (unrelated)

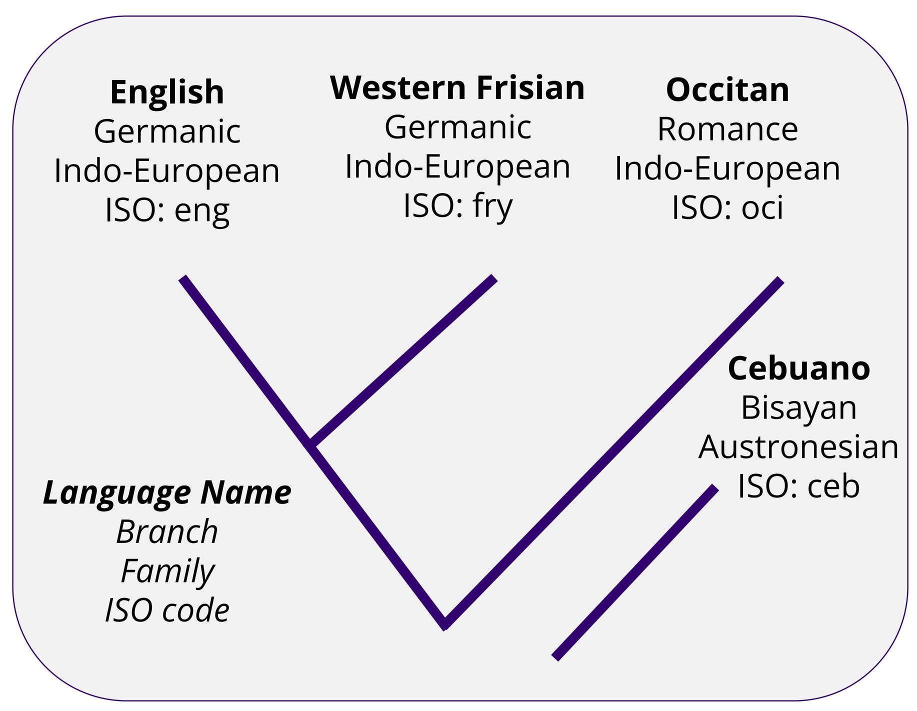

Figure 1

Evaluation Metric

We evaluate our models using **perplexity per word** on a held out testing set.

Perplexity: How well a model can predict a sample.

A sample that is more likely = Lower Perplexity

vs

A sample that is less likely = Higher Perplexity

More concretely, perplexity in NLP:

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

Where $s_0, s_1, \ldots, s_n$ are *sentences* in a corpus with *N* total *words.*

Experiment

We use the **GPT-2** Model Architecture.

We generated 5,000,000 synthetic sentences in Frisian, Occitan, and Cebuano.

We fine-tuned 4 models on each language:

1. English full pre-training→ Fine-tuning (*fig. 3*)
2. English pre-training on 5M sentences → fine-tuning (*fig. 4*)
3. Synthetic pre-training from random weights→Fine-tuning (*fig. 4*)
4. English full pretraining → Fine-tuning with 50% synthetic, hybrid sentences (*fig. 5*)

Fine-tuning always consists of training the model on gold standard language data. We evaluate model performance using the perplexity of the models on held-out gold standard data.

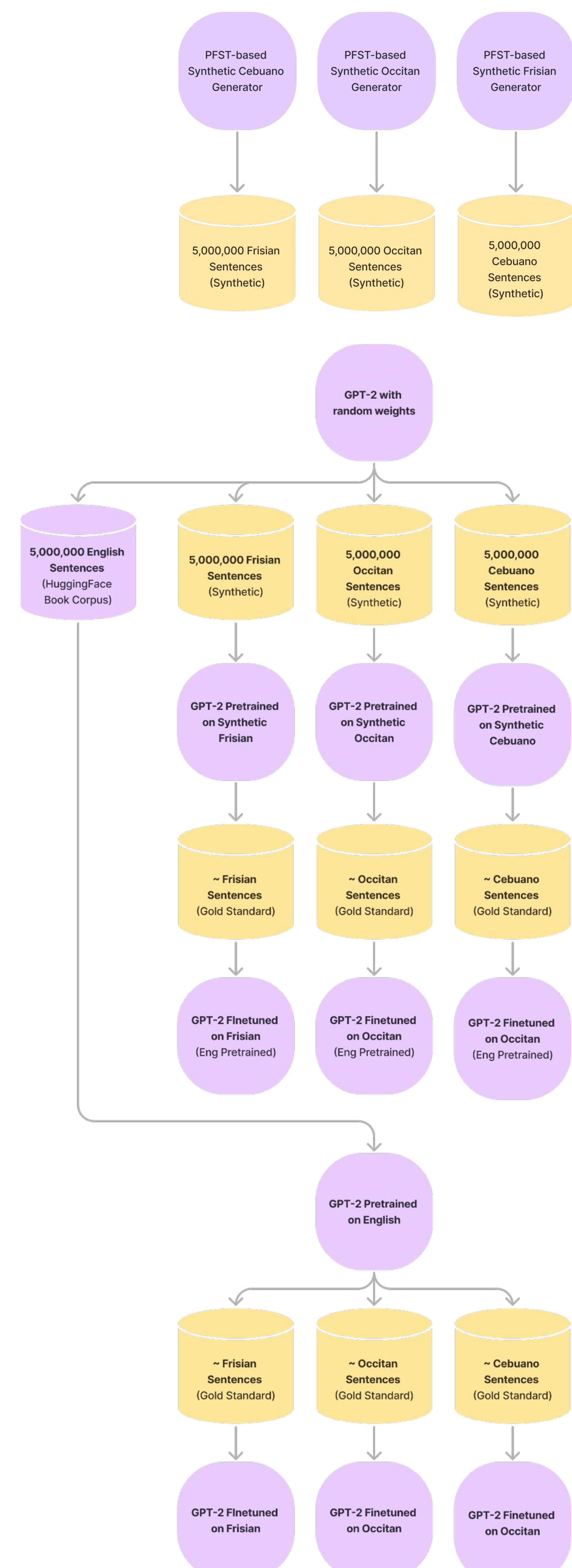

Figure 2: The training pipeline for each model we evaluated.

Results

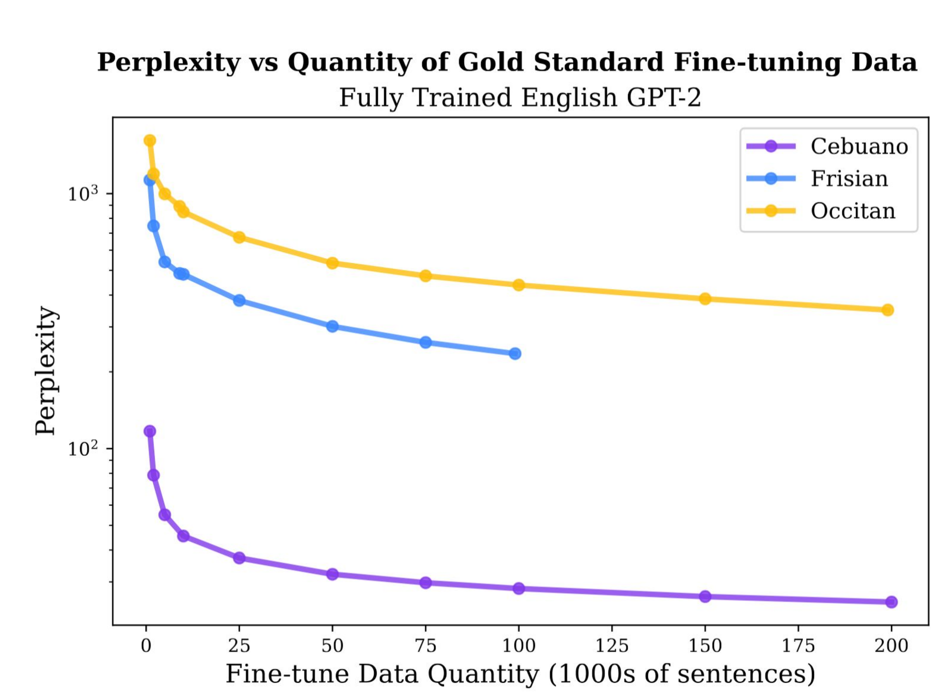

Figure 3

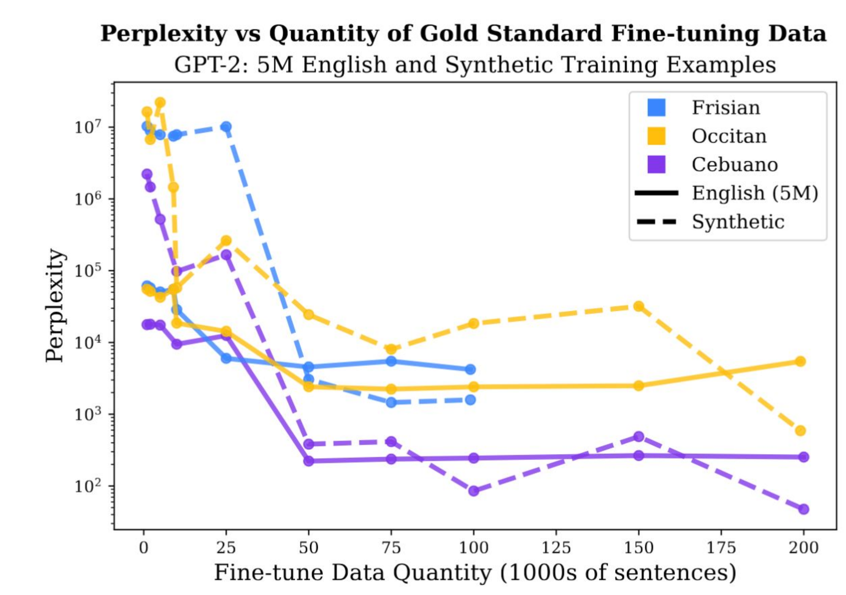

Figure 4

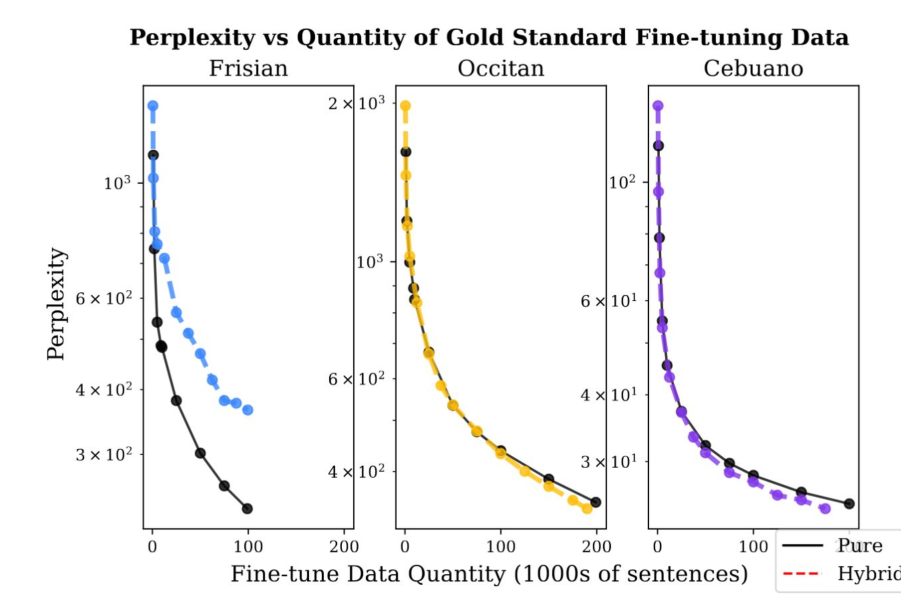

Figure 5

Discussion

Contrary to what we believed would happen, language similarity did **not** predict perplexity.

- This may be a result of not enough synthetic pretraining data
- Occitan (least closely related) benefited most from transfer learning. **Maybe language family not best metric.**

Synthetic Data Effectively Supplements Gold Standard Data:

Synthetic data can offer modest gains when mixed with gold standard data.

Future Work

Original GPT-2 was trained on ~40GB of data. We trained on ~500MB. Future work may see emergent patterns as data scales

We only considered three languages transferring from English. With more target and source languages, future work may be able to predict transfer success.

Citations

Leipzig Corpora Collection. Leipzig corpora collection (2017): Yoruba community. https : / / corpora . wortschatz - leipzig . de / en ? corpusId = yor _ community_2017, 2017.

Leipzig Corpora Collection. Leipzig corpora collection (2020): Frisian news. https : / / corpora . wortschatz - leipzig . de / en ? corpusId = fry _ news_2020, 2020.

Leipzig Corpora Collection. Leipzig corpora collection (2023): Occitan community. https : / / corpora . wortschatz - leipzig . de / en ? corpusId = oci _ community_2023, 2023.

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.

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.

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

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

lec 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

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

V´esteinn Snæbjarnarson, Annika Simonsen, Goran Glavaˇs, and Ivan Vuli´c. Transfer to a low-resource language via close relatives: The case study on faroese. arXiv preprint arXiv:2304.08823, 2023.

Language Transfer? Hardly Know 'er: Exploring the Utility of Synthetic Training Data in Low-Resource Settings

Alessio Tosolini, Ben Kosa, William Galvin