Endangered Languages—How much data do we need to model them well?

Emily Chang
Department of Computer Science
University of Virginia
Charlottesville, VA
ec5ug@virginia.edu

Caroline Gihlstorf
Department of Computer Science
University of Virginia
Charlottesville, VA
czm5kz@virginia.edu

Jade Gregoire
Department of Computer Science
University of Virginia
Charlottesville, VA
dze3jz@virginia.edu

UNIVERSITY of VIRGINIA

Introduction

- More than 43% of languages spoken in the world are endangered (Zhang et al., 2022)
- What if we could use NLP to preserve these languages?
  - Difficult to train a model from scratch on minimal data
  - What about using a pre-trained model in a similar language?
- Can we find the minimum amount of tokens required for a pre-trained model to perform well in another language?
- Use a pre-trained English model, fine-tune it on French data

What is considered an endangered language?

- Open Super-large Crawled Aggregated coRpus (OSCAR)
- 153 languages
- 13% are considered vulnerable or endangered

Language Endangerment Level	Average Number of Tokens	Standard Deviation of Tokens
Not endangered	8.130 billion	46.938 billion
Vulnerable	13.878 million	48.027 million
Definitely endangered	28.353 million	54.083 million
Severely endangered	949 thousand	941 thousand
Critically endangered	6,347	-

Christopher Moseley. Atlas of the world's languages in danger. UNESCO. 2010.

UNIVERSITY of VIRGINIA

Main Resources

- Source language: English
- RoBERTa
  - “roberta-base”
  - Monolingual and not fine tuned
- SQuAD
  - Stanford Question Answering Dataset

- Target language: French
- CamemBERT
  - “illuin/camembert-base-fquad”
  - Use to benchmark a good performance
- FQuAD
  - French SQuAD equivalent

UNIVERSITY of VIRGINIA

The literal meaning of Durbar Square is a "place of palaces". There are three preserved Durbar Squares in Kathmandu valley and one unpreserved in Kirtipur. The Durbar Square of Kathmandu is located in the old city and has heritage buildings representing four kingdoms (Kantipur, Lalitpur, Bhaktapur, Kirtipur); the earliest is the Licchavi dynasty. The complex has 50 temples and is distributed in two quadrangles of the Durbar Square. The outer quadrangle has the Kasthamandap, Kumari Ghar, and Shiva-Parvati Temple; the inner quadrangle has the Hanuman Dhoka palace. The squares were severely damaged in the April 2015 Nepal earthquake.

When did a notable earthquake occur that damaged Kathmandu's Durbar Square?

Les deux tableaux sont certes décrits par des documents contemporains à leur création mais ceux-ci ne le font qu'indirectement car ils concernent principalement La Vierge aux rochers. Aussi demeurent-ils objets de spéculations pour les chercheurs quant à leur statut de première ou seconde version de l'œuvre, leur création, leur attribution, leur datation, leur disposition exacte sur le retable et les raisons qui ont poussé à leurs modifications au cours du temps — notamment pour ce qui concerne la couleur du fond.

Que concerne principalement les documents?

Excerpts from SQuAD & FQuAD, respectively

Methodology

Method: TSDAE

Data: OSCAR

Unsupervised Finetuning

Evaluation

Metric: Similarity scores

Data: Massive Text Embedding Benchmark

Method: Backpropagation

Data: FQuAD Training Data

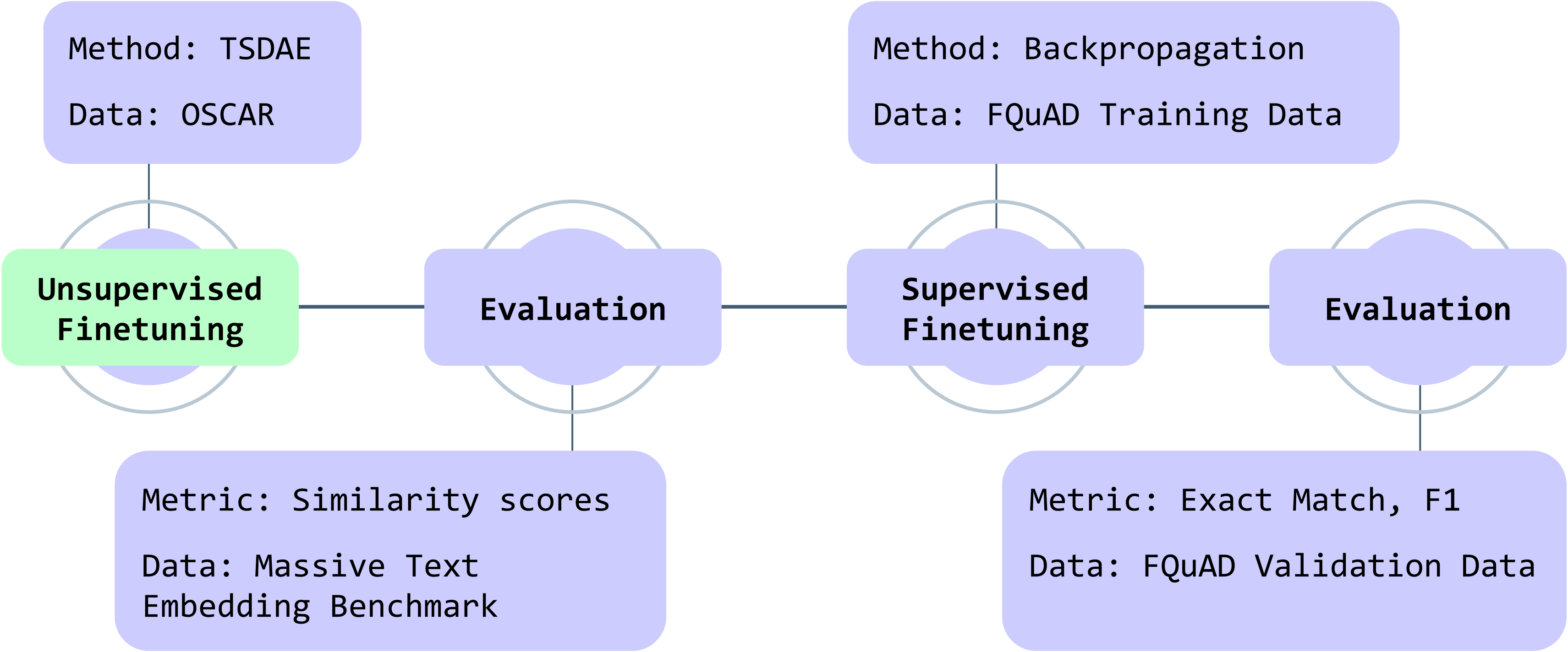
Supervised Finetuning

Evaluation

Metric: Exact Match, F1

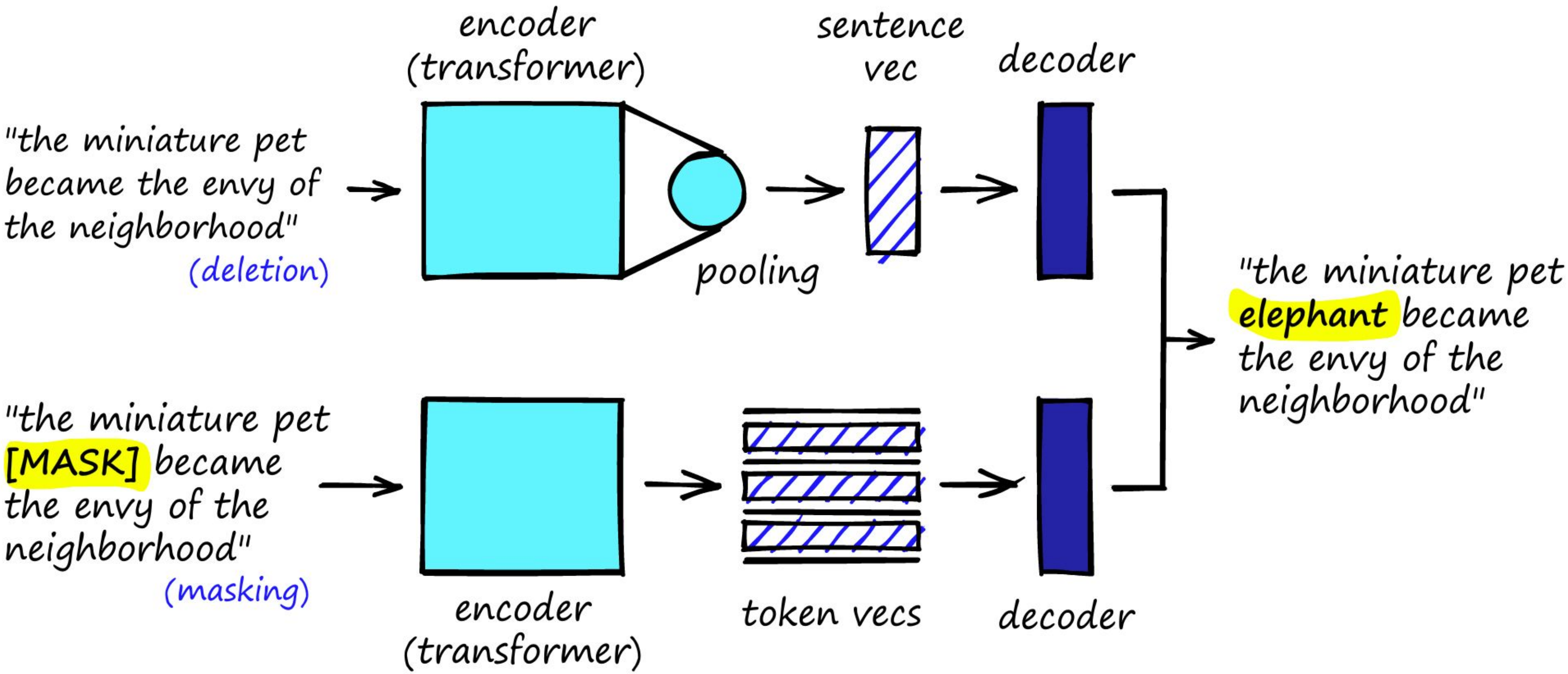
Data: FQuAD Validation Data

UNIVERSITY of VIRGINIA


Method: TSDAE
Data: OSCAR
Unsupervised Finetuning
Evaluation
Metric: Similarity scores
Data: Massive Text Embedding Benchmark
Method: Backpropagation
Data: FQuAD Training Data
Supervised Finetuning
Evaluation
Metric: Exact Match, F1
Data: FQuAD Validation Data

OPTIONS BINAIRES EN LIGNE. EBOOK. Cher trader, Bienvenue sur Options Binaires en ligne, ceci est le premier volet d’une série de 5 Ebooks pour apprendre à négocier des options binaires en ligne, nous sommes heureux de vous introduire dans le monde du négoce financier grâce aux options binaires. Les options binaires sont connues comme étant un moyen super rapide, simple et accessible pour investir et gagner de l’argent en ligne.En effet, vous pouvez vous lancer dans les options binaires, et ce, même si vous êtes nul en trading. Voyez comment il est possible d’investir facilement dans les options binaires. Nous allons y voir ce qu’est une option binaire, les pièges à éviter pour un investissement dans l’option binaire, les différentes options binaires ainsi que toutes les astuces pour maximiser vos chances de réussir votre investissement. En même temps, nous allons également vous donner de bons conseils sur le choix des plateformes d ... Avantages à investir dans les options binaires. Le premier avantage des options binaires réside dans leur simplicité : il suffit d’estimer la direction qu’une option va prendre. Sur les actions traditionnelles, on spécule sur une différence de prix réel, beaucoup plus difficile à prédire. Investir dans les options binaires : exemple d’option binaire À titre d’exemple, nous pourrons considérer une option binaire associée à l’or. La valeur monétaire de cet or est estimée à 1300 dollars et est associée à une option qui s’élève à 100 dollars.

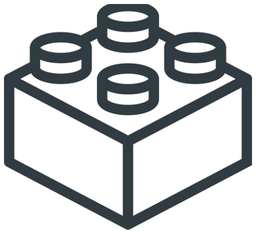
Excerpt from Open Super-large Crawled Aggregated coRpus (OSCAR)



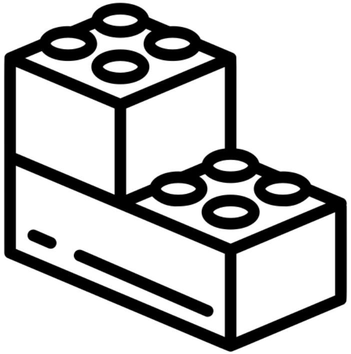
Transformer(-based) and Sequential Denoising Auto-Encoder (TSDAE)

UNIVERSITY of VIRGINIA

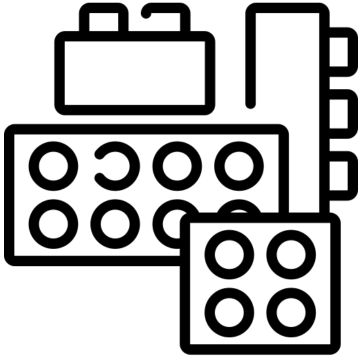
What `roberta-base` was finetuned on



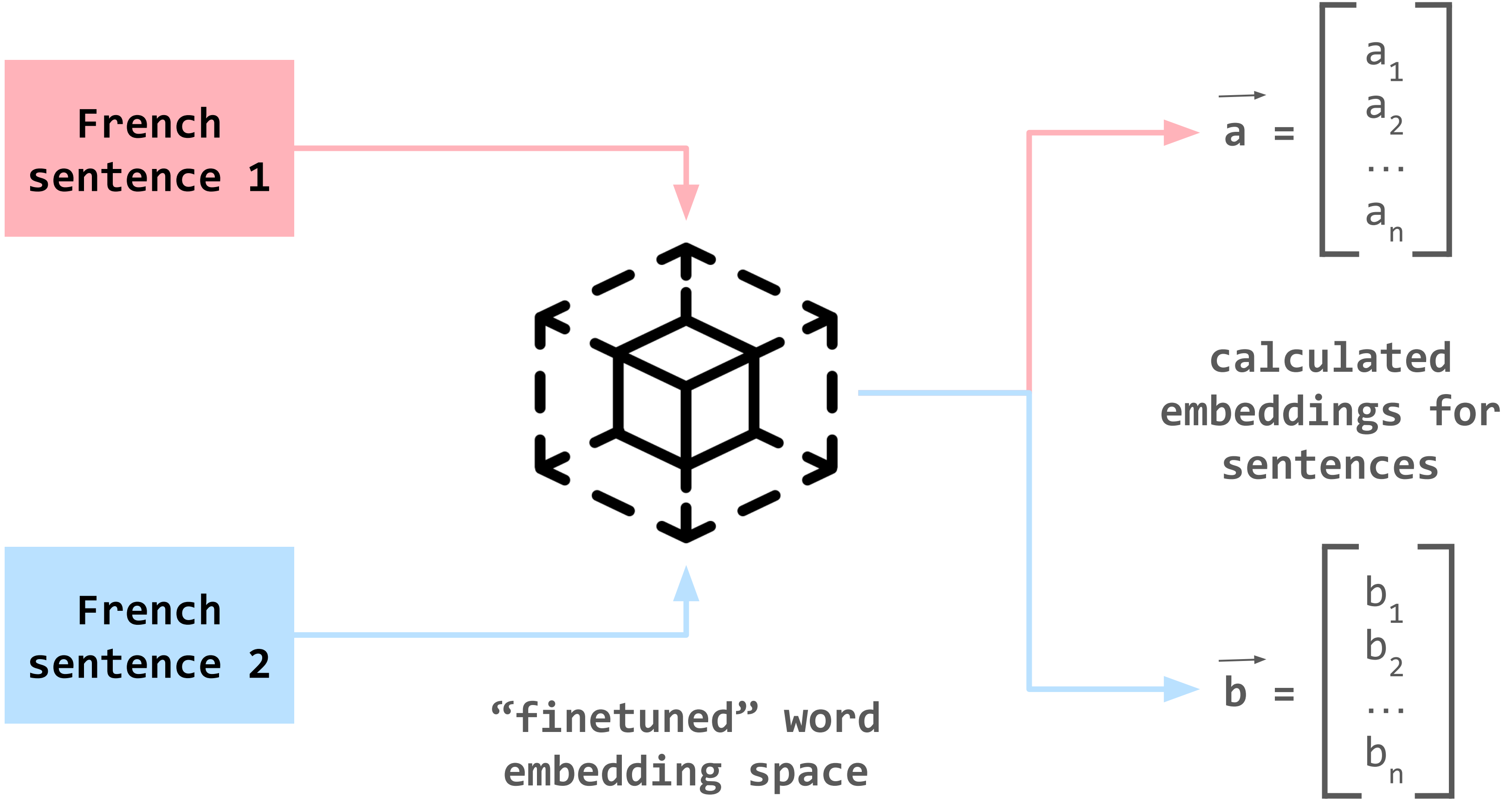
6,500 tokens



100,000 tokens



950,000 tokens


French sentence 1
French sentence 2
“finetuned” word embedding space
$\vec{a} = \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{bmatrix}$
calculated embeddings for sentences
$\vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix}$

cosine distance $= 1 - \cos(\vec{a}, \vec{b})$

how different embeddings are from one another

how similar embeddings are to another

Shkhanukova, Milana. “Cosine distance and cosine similarity.”
https://medium.com/@milana.shxanukova15/cosine-distance-and-cosine-similarity-a5da0e4d9ded

The Dangers of a High Learning Rate

Number of Tokens Finetuned on	Correlation Score
Control: roberta-base	0.382
6,500_3e-5_1epoch	0.376
6,500_3e-7_1epoch	0.398
6,500_3e-7_2epoch	0.398
100,000_3e-5_1epoch	0.349
100,000_3e-7_2epoch	0.418
100,000_3e-7_3epoch	0.406
950,000_3e-5_1epoch	0.124
950,000_3e-7_1epoch	0.405
950,000_3e-10_1epoch	0.398
950,000_3e-7_2epoch	0.194
Camembert	0.634

Table.1: Evaluation of Unsupervised Finetuning

*Model naming convention: token amount, learning rate, followed by epoch amount

UNIVERSITY of VIRGINIA

More Tokens, More Complex Hyperparameters

Lowering learning rate and increasing epochs improve performance

Number of Tokens Finetuned on	Correlation Score
Control: roberta-base	0.382
6,500_3e-5_1epoch	0.376
6,500_3e-7_1epoch	0.398
6,500_3e-7_2epoch	0.398
100,000_3e-5_1epoch	0.349
100,000_3e-7_2epoch	0.418
100,000_3e-7_3epoch	0.406
950,000_3e-5_1epoch	0.124
950,000_3e-7_1epoch	0.405
950,000_3e-10_1epoch	0.398
950,000_3e-7_2epoch	0.194
Camembert	0.634

No marked difference

Increasing epochs decreases performance

Table.1: Evaluation of Unsupervised Finetuning

*Model naming convention: token amount, learning rate, followed by epoch amount

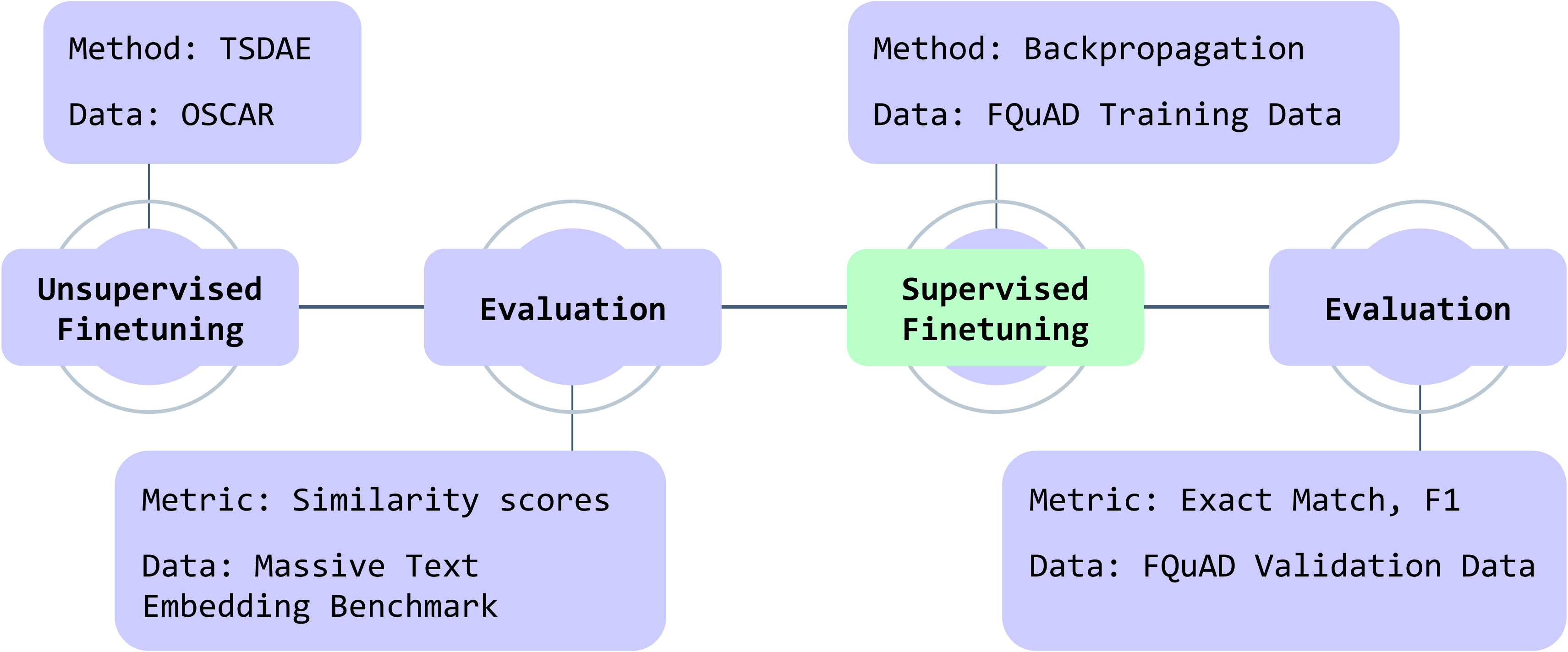
Word Embeddings did not improve significantly

Number of Tokens Finetuned on	Correlation Score
Control: roberta-base	0.382
6,500_3e-5_1epoch	0.376
6,500_3e-7_1epoch	0.398
6,500_3e-7_2epoch	0.398
100,000_3e-5_1epoch	0.349
100,000_3e-7_2epoch	0.418
100,000_3e-7_3epoch	0.406
950,000_3e-5_1epoch	0.124
950,000_3e-7_1epoch	0.405
950,000_3e-10_1epoch	0.398
950,000_3e-7_2epoch	0.194
Camembert	0.634

Table.1: Evaluation of Unsupervised Finetuning

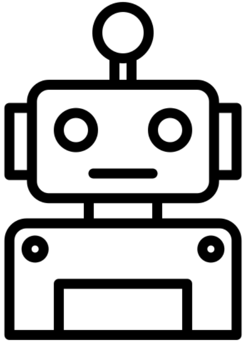
*Model naming convention: token amount, learning rate, followed by epoch amount

UNIVERSITY of VIRGINIA


Method: TSDAE
Data: OSCAR
Unsupervised Finetuning
Evaluation
Metric: Similarity scores
Data: Massive Text Embedding Benchmark
Method: Backpropagation
Data: FQuAD Training Data
Supervised Finetuning
Evaluation
Metric: Exact Match, F1
Data: FQuAD Validation Data

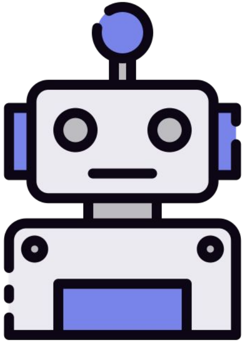
Context	Piazzi observa Cérès 24 fois, la dernière fois le 11 février. Le 24 janvier 1801, Piazzi annonça sa découverte par des lettres à plusieurs collègues italiens, parmi lesquels Barnaba Oriani à Milan. Il la décrivit comme une comète, mais remarqua que « puisque son mouvement est lent et uniforme, il m'a semblé à plusieurs reprises qu'il pourrait s'agir de quelque chose de mieux qu'une comète. » En avril, Piazzi envoya ses observations complètes à Oriani, Bode et Lalande à Paris. Elles furent publiées dans l'édition de septembre 1801 du Monatliche Correspondenz.
Question	Pourquoi Cérès n'était pas directement assimilable à une comète ?
Answer	son mouvement est lent et uniforme

UNIVERSITY of VIRGINIA



roberta-base

- Clean copy of `roberta-base`
- Never undergone unsupervised finetuning
- Never undergone supervised finetuning



roberta-base-fquad-finetuned

- **Modified** copy of `roberta-base`
- Never undergone unsupervised finetuning
- **Has undergone supervised finetuning**

Supervised finetuning improves score the most

Our method of improving the word embeddings does not have much impact

Model	FQUAD Exact Match	FQUAD F1
Control: roberta-base	0.063%	7.58%
roberta-base-fquad-finetuned	21.6%	31.9%
6,500_3e-5_1epoch	21.1%	31.9%
6,500_3e-7_1epoch	22.2%	32.4%
6,500_3e-7_2epoch	21.8%	31.8%
100,000_3e-5_1epoch	21.4%	31.8%
100,000_3e-7_2epoch	21.0%	31.5%
100,000_3e-7_3epoch	21.7%	32.9%
950,000_3e-10_1epoch	21.6%	32.2%
950,000_3e-7_1epoch	21.5%	32.1%
950,000_3e-7_2epoch	21.4%	32.1%
Camembert	45.8%	68.2%

Table.2: Evaluation of Supervised Finetuning

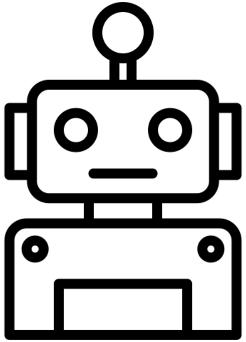
*Model naming convention: token amount, learning rate, followed by epoch amount

Varying token size does not improve performance

Model	FQUAD Exact Match	FQUAD F1
Control: roberta-base	0.063%	7.58%
roberta-base-fquad-finetuned	21.6%	31.9%
6,500_3e-5_1epoch	21.1%	31.9%
6,500_3e-7_1epoch	22.2%	32.4%
6,500_3e-7_2epoch	21.8%	31.8%
100,000_3e-5_1epoch	21.4%	31.8%
100,000_3e-7_2epoch	21.0%	31.5%
100,000_3e-7_3epoch	21.7%	32.9%
950,000_3e-10_1epoch	21.6%	32.2%
950,000_3e-7_1epoch	21.5%	32.1%
950,000_3e-7_2epoch	21.4%	32.1%
Camembert	45.8%	68.2%

Table.2: Evaluation of Supervised Finetuning

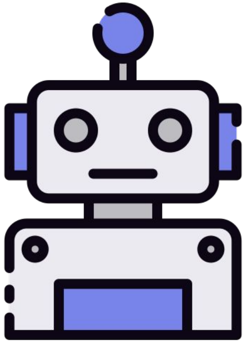
*Model naming convention: token amount, learning rate, followed by epoch amount



roberta-base responses to first 11 FQuAD QA pairs

1. mais ceux-ci ne le font qu'indirectement car ils concernent principalement La Vierge aux
2. mais ceux-ci ne le font qu'indirectement
3. mais ceux-ci ne le font qu'indirectement car ils concernent principalement La Vierge aux
4. empty
5. ans
6. empty
7. dans la version
8. dans la version londonienne du panneau
9. dans la version londonienne du panneau
10. dans la version
11. puisque ce dernier fait partie des trois artistes désignés dans le contrat de commande, chacun ayant un rôle

University of Virginia



roberta-base-fquad-finetuned responses to first 11 FQuAD QA pairs

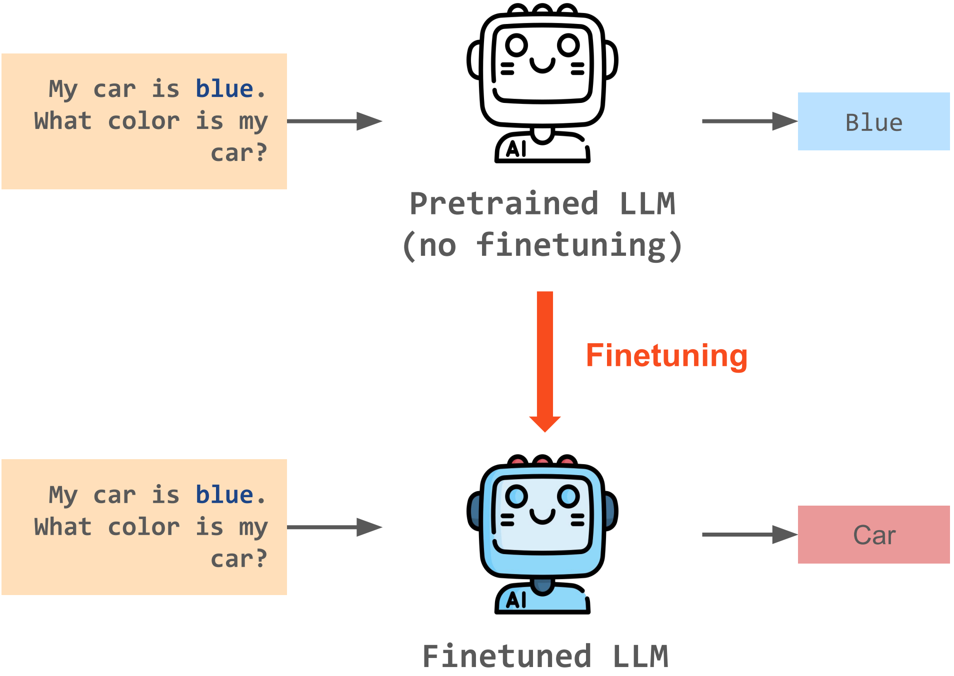
1. La Vierge aux rochers
2. documents contemporains à leur création
3. objets de spéculations
4. droite
5. gauche
6.
7. l'atelier de Léonard de Vinci
8.
9.
10. (La Vierge aux rochers)
11. trois

UNIVERSITY of VIRGINIA

Catastrophic Forgetting

UNIVERSITY of VIRGINIA

Catastrophic Forgetting



When finetuning causes models to forget what they learned during pretraining

UNIVERSITY of VIRGINIA

Results of Catastrophic Forgetting

Model	SQUAD Exact Match	SQUAD F1
Control: roberta-base	0.194%	4.33%
roberta-base-fquad-finetuned	42.0%	46.1%
6,500_3e-5_1epoch	42.6% ⇧	46.5% ⇧
6,500_3e-7_1epoch	45.0% ⇧	48.6% ⇧
6,500_3e-7_2epoch	43.8% ⇧	47.7% ⇧
100,000_3e-5_1epoch	43.0% ⇧	47.0% ⇧
100,000_3e-7_2epoch	39.4% ⬇	43.9% ⬇
100,000_3e-7_3epoch	41.7% ⬇	45.7% ⬇
950,000_3e-10_1epoch	41.8% ⬇	45.9% ⬇
950,000_3e-7_1epoch	42.5% ⇧	46.4% ⇧
950,000_3e-7_2epoch	44.0% ⇧	47.8% ⇧
roberta-base-squad2	79.5%	82.5%

Table.3: Catastrophic Forgetting

*Model naming convention: token amount, learning rate, followed by epoch amount

Minor catastrophic forgetting:

- Most models did not decrease in performance
- All performance decreases were within 3 percentage points of roberta-base-fquad-finetuned

UNIVERSITY of VIRGINIA

Ideas for Next Steps

Supervised finetuning on more epochs

Unsupervised finetuning on more tokens

Manual evaluation of outputs; qualitative analysis

What if we had a base model trained in an endangered language?

If we had ample text in a related language…	Finetune the base model on the text of the related language Would language similarities improve model performance on the original language?
Issues:	We would need an evaluation set for the original language (which is already low-resource)

UNIVERSITY of VIRGINIA

Ethical Implications

- Our current work is theoretical
- Essential to consider whether communities who use low-resource/endangered languages actually **want** technology made for their language (Wilson, 2022)
- Language is far more than just “data”

UNIVERSITY of VIRGINIA

Thank you!

Any questions?

Works Cited

Christopher Moseley. *Atlas of the world's languages in danger*. UNESCO. 2010.

d’Hoffschmidt Martin, Vidal Maxime, Belblidia Wacim, and Brendle Tom. FQuAD: French Question Answering Dataset. arXiv e-prints, art. arXiv:2002.06071, Feb 2020a.

Kexin Wang, Nils Reimers, and Iryna Gurevych. TSDAE: using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. CoRR, abs/2104.06979, 2021. URL https://arxiv.org/abs/2104.06979.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Su´arez, Yoann Dupont, Laurent Romary, ´Eric de la Clergerie, Djam´e Seddah, and Beno ˆıt Sagot. CamemBERT: a tasty French language model. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7203–7219, Online, July 2020b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.645

Pedro Javier Ortiz Súarez, Benoıt Sagot, and Laurent Romary. Asynchronous pipelines for process- ing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pp. 9 – 16, Mannheim, 2019. Leibniz-Institut f¨ur Deutsche Sprache. doi: 10.14618/ids-pub-9021. URL http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215.

Works Cited

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822, 2018. URL http://arxiv.org/abs/1806.03822.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language. *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.108

Shkhanukova, Milana. “Cosine distance and cosine similarity.” Medium, 4 Mar 2023, https://medium.com/@milana.shxanukova15/cosine-distance-and-cosine-similarity-a5da0e4d9ded

T. C. Rajapakse. Simple transformers. https://github.com/ThilinaRajapakse/simpletransformers, 2019.

“Unsupervised Training for Sentence Transformers” Pinecone. https://www.pinecone.io/learn/series/nlp/unsupervised-training-sentence-transformers/

Works Cited

Wilson, Joseph. “Why Ai Will Never Fully Capture Human Language.” SAPIENS, 12 Oct 2022, www.sapiens.org/language/ai-oral-languages/.

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

Methodology

- Show q/a pair from datasets?
- Roadmap -> make prettier flowchart
- Very brief overview, the finetuning sections will go in depth

Unsupervised Finetuning

- Give an example of what OSCAR is
- Similarity score: evaluation metric, higher is better
- List the token sizes of what we use
- TSDAE
  - Masked language modeling
  - Pretraining if you don’t have a lot of data
  - Alters word embedding

Results

- High learning rate → worse performance
- Word embedding space did not improve significantly
- More tokens you have: combination of learning rate and epoch has a bigger impact

Number of Tokens Finetuned on	Similarity Score
Control: roberta-base	0.382
6,500_3e-5_1epoch	0.376
6,500_3e-7_1epoch	0.398
6,500_3e-7_2epoch	0.398
100,000_3e-5_1epoch	0.349
100,000_3e-7_2epoch	0.418
100,000_3e-7_3epoch	0.406
950,000_3e-5_1epoch	0.124
950,000_3e-7_1epoch	0.405
950,000_3e-10_1epoch	0.398
950,000_3e-7_2epoch	0.194
Camembert	0.634

Table.1: Evaluation of Unsupervised Finetuning

*Model naming convention: token amount, learning rate, followed by epoch amount

UNIVERSITY of VIRGINIA

Supervised Finetuning

- Method & results
- FQUAD: what is it
- Controls (2)
  - Clean copy of roberta-base
  - roberta-base that was never finetuned; without the benefit of a “better” word embedding
- Finetuning for a specific task (act of finding where answer is), not learning French

Results from Supervised Finetuning

- Not printing out anything prior to finetuning
- Varying token size does not improve supervised finetuning performance; same can be said for epoch and learning rate
- Supervised finetuning is what improves scores the most

Model	FQUAD Exact Match	FQUAD F1
Control: roberta-base	0.063%	7.58%
roberta-base-fquad-finetuned	21.6%	31.9%
6,500_3e-5_1epoch	21.1%	31.9%
6,500_3e-7_1epoch	22.2%	32.4%
6,500_3e-7_2epoch	21.8%	31.8%
100,000_3e-5_1epoch	21.4%	31.8%
100,000_3e-7_2epoch	21.0%	31.5%
100,000_3e-7_3epoch	21.7%	32.9%
950,000_3e-10_1epoch	21.6%	32.2%
950,000_3e-7_1epoch	21.5%	32.1%
950,000_3e-7_2epoch	21.4%	32.1%
Camembert	45.8%	68.2%

Table.2: Evaluation of Supervised Finetuning

*Model naming convention: token amount, learning rate, followed by epoch amount

UNIVERSITY of VIRGINIA