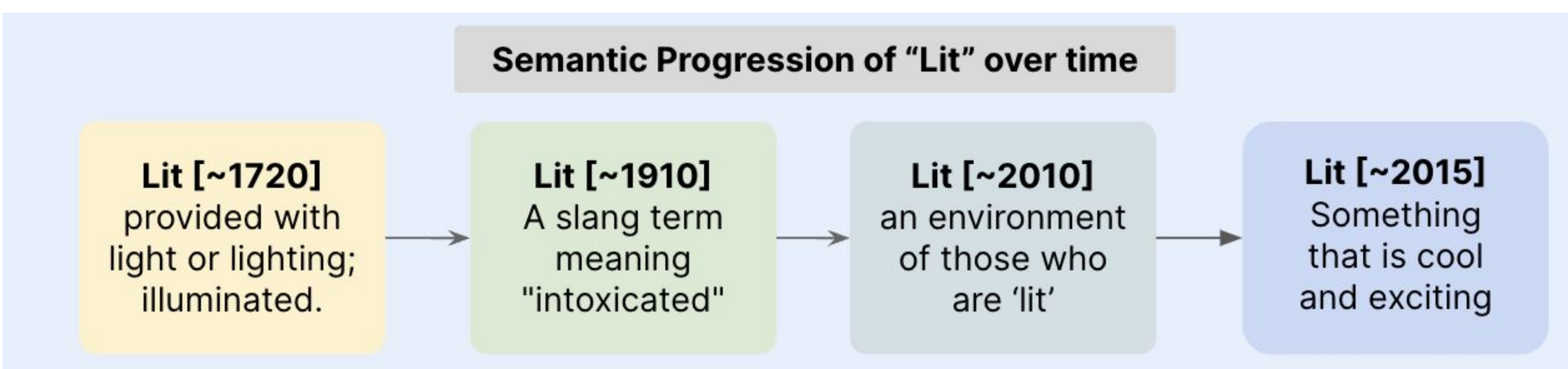
Today Years Old: Adapting Language Models to Word Shifts

Olivia Lee, Jason Chen, Zachary Xi
Stanford CS 224N Custom Project

Introduction & Motivation

- Language is constantly evolving: **existing words acquire new meanings** and new words are added to lexicons.
- Accounting for semantic shifts is crucial for language models (LMs) to **accurately model human language.**
- Current pretrained LMs face **challenges in adapting to new/modified words** due to their initialization methods.
- Our goal is to develop new approaches to **editing LMs for lexical adaptation** with Urban Dictionary data.



Approach & Methodology

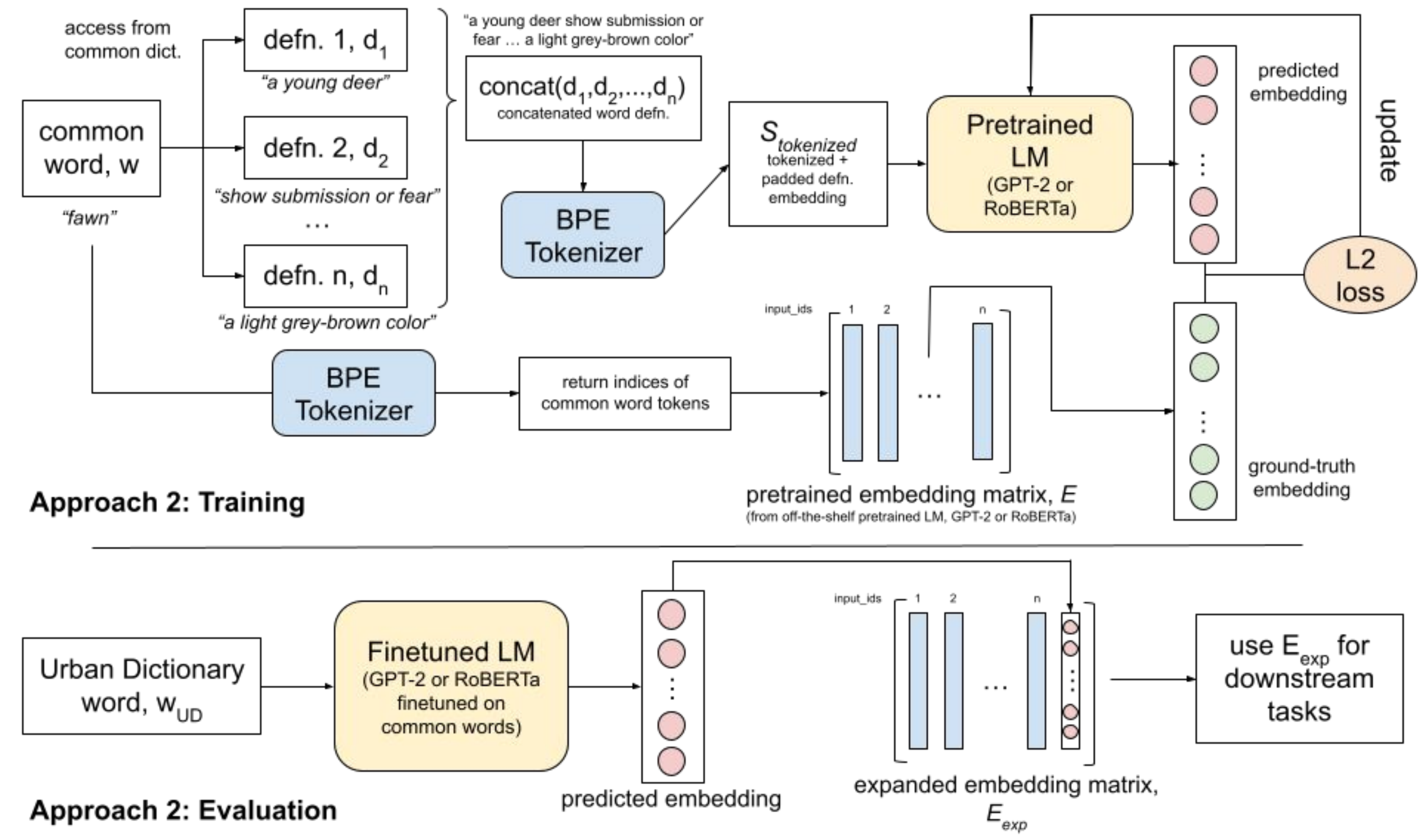
Approach 1: Adapting embedding initialization

- Initialize new embeddings by **averaging over pre and post-expansion embeddings** to construct distribution.
- Sample from distribution and **append sampled embeddings** of respective model (GPT-2, RoBERTa).
- KL divergence is then bounded; as LM vocabulary size grows, **new word probability decreases.**

Approach 2: Finetuning via Common Word Mappings

- Train separate neural network by finetuning pretrained GPT-2 or RoBERTa model via **training on dictionary of common words** to learn mapping from dictionary definitions to word embeddings.
- Perform gradient descent on L2 loss between [CLS] embedding and **ground-truth embedding of common words** (see figure for architecture)

Approach and Methodology (cont'd)



Experiments

- Evaluation method involves **masked language modeling**; masked example sentences are inputted into the model.
- Avg. ranking per word, **calculated over distribution of possible logits** (lower is better), plus number of urban dict. word appearances in the top *k* likely embeddings (GPT-2: k = 5, 10, 25; RoBERTa: k = 10, 100, 1000)

	Experiment 1	Experiment 2
Methodology	Multiple choice options tokenized and fed in as inputs into pretrained LM.	L2 Loss function used to train on common word dictionary (alongside Adam optimizer).
Learning Rate	5e-5	3e-5
Epochs	3	3

Experiment Results

Experiment 1: Novel words only

	GPT-2			RoBERTa	
	Baseline	Approach 2		Baseline	Approach 2
Top 5	123	**175**	**Top 10**	20	**46**
Top 10	287	**371**	**Top 100**	241	**410**
Top 25	**758**	734	**Top 1000**	2,451	**4,660**
Avg. Rank	25.532	**24.793**	**Avg. Rank**	**2,633.670**	2,679.187

Experiment Results (cont'd) & Analysis

- The approach **outperforms baseline on almost all metrics.** For RoBERTa, Approach 2 results in **almost 100% increase** in Urban Dict. word appearances in the top *k*.

Experiment 2: Common words only

	GPT-2			RoBERTa	
	Baseline	Approach 2		Baseline	Approach 2
Top 5	55	78	**Top 10**	31	**710**
Top 10	130	147	**Top 100**	219	**1,499**
Top 25	281	267	**Top 1000**	1,169	**3,285**
Avg. Rank	20.18329	20.32947	**Avg. Rank**	15,191.216	**12,724.62**

- The approach **outperforms baseline on all metrics.** For RoBERTa, Approach 2 results in **22x increase in word appearances** in the top 10 and **5x increase** in top 100
- Both sets of results suggest that the trained model has **successfully learned mappings** from word definitions to word embeddings.
- Model can predict word embeddings for **unseen lexical items**, used for downstream language inference tasks.

Conclusions

- Our work has demonstrated two approaches to editing LMs for lexical adaptation with Urban Dictionary data: **initialization with embedding average t**rained with gradient descent, and finetuning LMs to **learn mappings from definitions to embeddings.**

Future Work

- Investigating **overlapping words in a dictionary:** concatenating new definition to predict new embedding
- Considering **novel multi-word lexical items**: updating embeddings to incorporate new definitions

References

1. Peter Koch. Meaning change and semantic shifts. In Päivi Juvonen and Maria Koptjevskaja-Tamm, editors, The Lexical Typology of Semantic Shifts, page 21–66. De Gruyter Mouton, Berlin/Boston, 2016.
2. Merriam Webster. Accessed 11 February 2023.
3. John Hewitt. Initializing new word embeddings for pretrained language models, 2021.
4. Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
5. 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.
6. Julian Martin Eisenschlos, Jeremy R. Cole, Fangyu Liu, and William W. Cohen. Winodict: Probing language models for in-context word acquisition, 2022.
7. Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, KR'12, page 552–561. AAAI Press, 2012.
8. Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05):8732–8740, Apr. 2020.
9. Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 66–76, Hong Kong, China, November 2019. Association for Computational Linguistics.
10. Tom Bosc and Pascal Vincent. Auto-encoding dictionary definitions into consistent word embeddings. In Conference on Empirical Methods in Natural Language Processing, 2018.
11. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.