Grounded Compositional Outputs for Adaptive Language Modeling

Nikolaos Pappas, Phoebe Mulcaire, Noah A. Smith

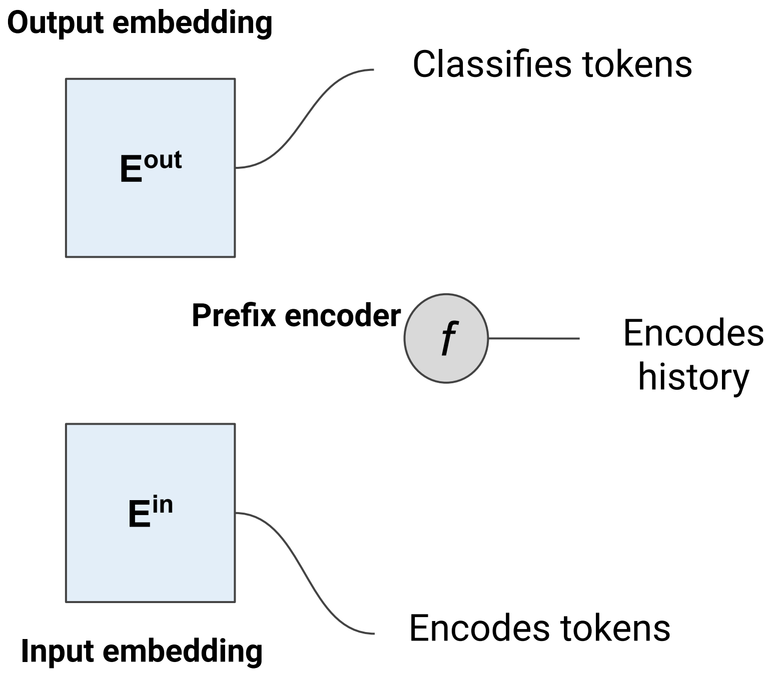
EMNLP 2020

UWNLP

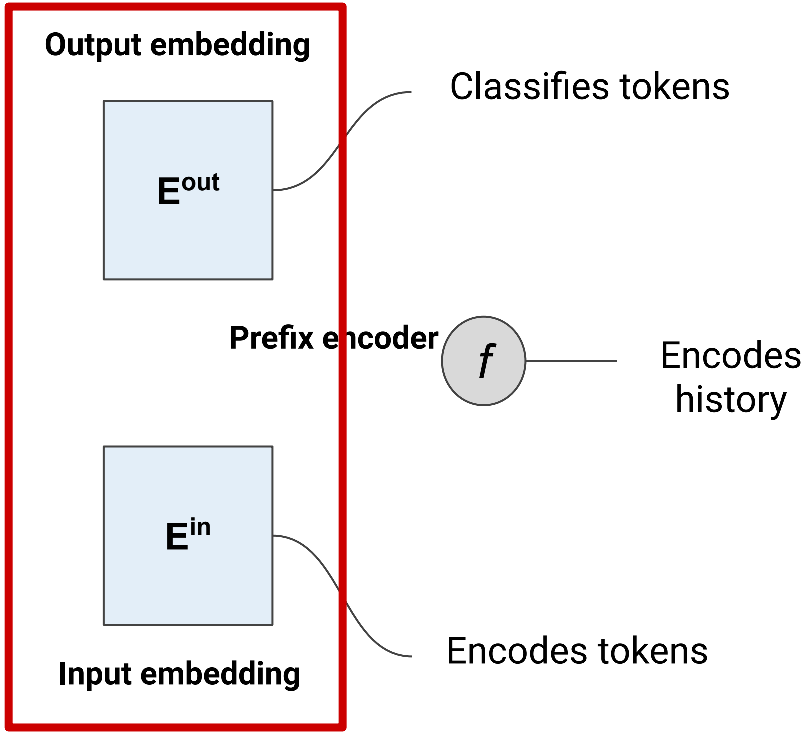
UNIVERSITY of WASHINGTON

AI2

Neural language models

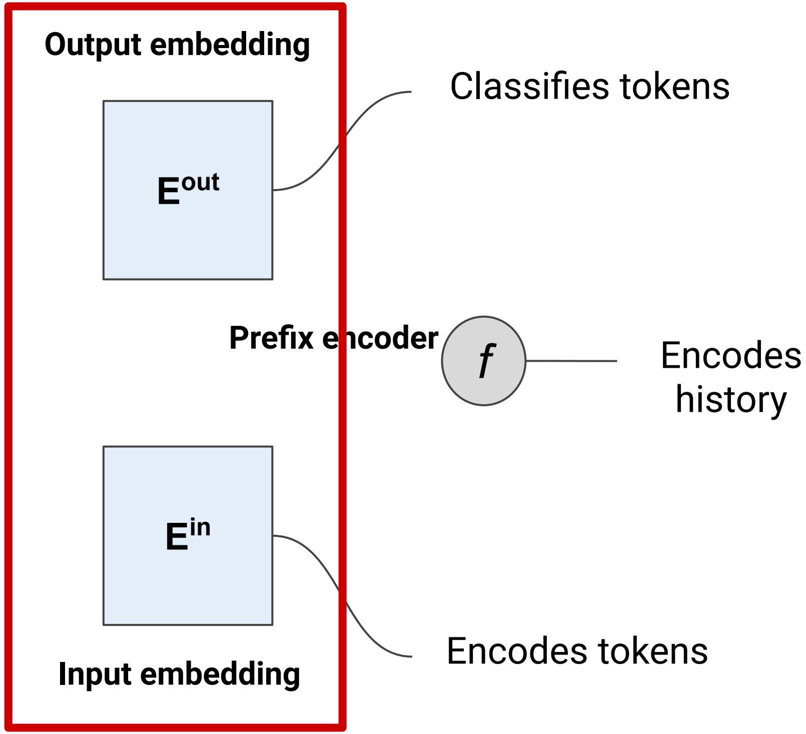


Neural language models: Limitations



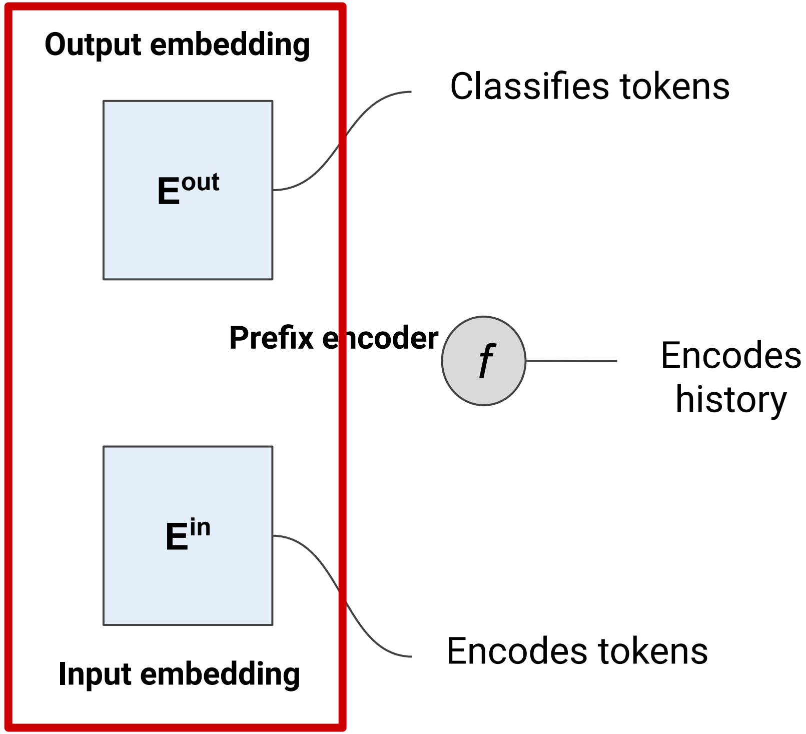
- Parameterization depends on the vocabulary

Neural language models: Limitations



- Parameterization depends on the vocabulary
- Handle rare or new words poorly

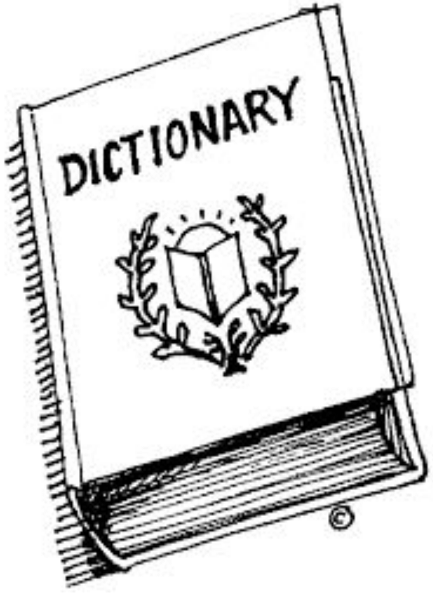
Neural language models: Limitations



- Parameterization depends on the vocabulary
- Handle rare or new words poorly
- Cannot be gracefully modified once trained

Motivation

- Can we use lexicons to better generalize to rare words?
- How to decouple training and testing vocabularies?
- What output form is most suitable for domain adaptation?



Previous work on rare words

Subword tokenization

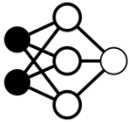
- Character-level models (Cherry et al., 2018; Al-Rfou et al., 2019)
  - ❎ Costly prefix encoders and training
- Data-driven vocabulary selection (Sennrich et al., 2016; Radford et al., 2018)
  - ❎ Linguistically simplistic
  - ❎ Rely on lookup tables

Interpolation with a neural cache

- Local or unbounded neural cache (Graves et al., 2017a,b)
  - ☑ Low-cost adaptation to rare/new words

$$p(x_t|h_{1:t}, x_{1:t}) = (1 - \lambda)p_{vocab}(x_t|h_t) + \lambda p_{cache}(x_t|h_{1:t}, x_{1:t})$$

Neural model



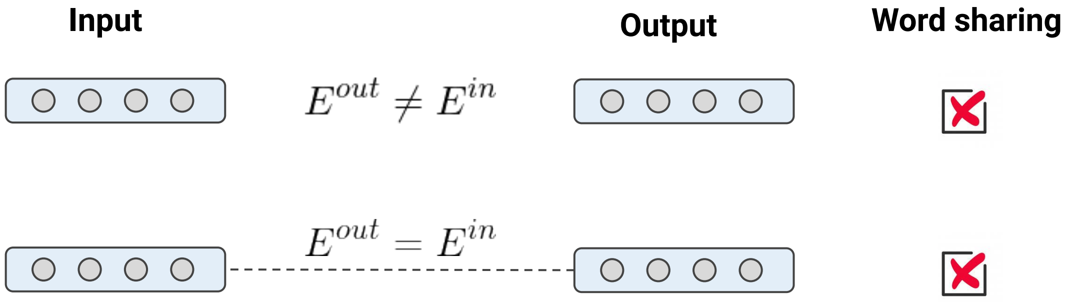
Local neural cache

$$\propto \sum_{i=1}^{t-1} \mathbb{1}_{\{w=x_{i+1}\}} \exp(\theta h_t^\top h_i)$$

Sharing across words

Sharing across words

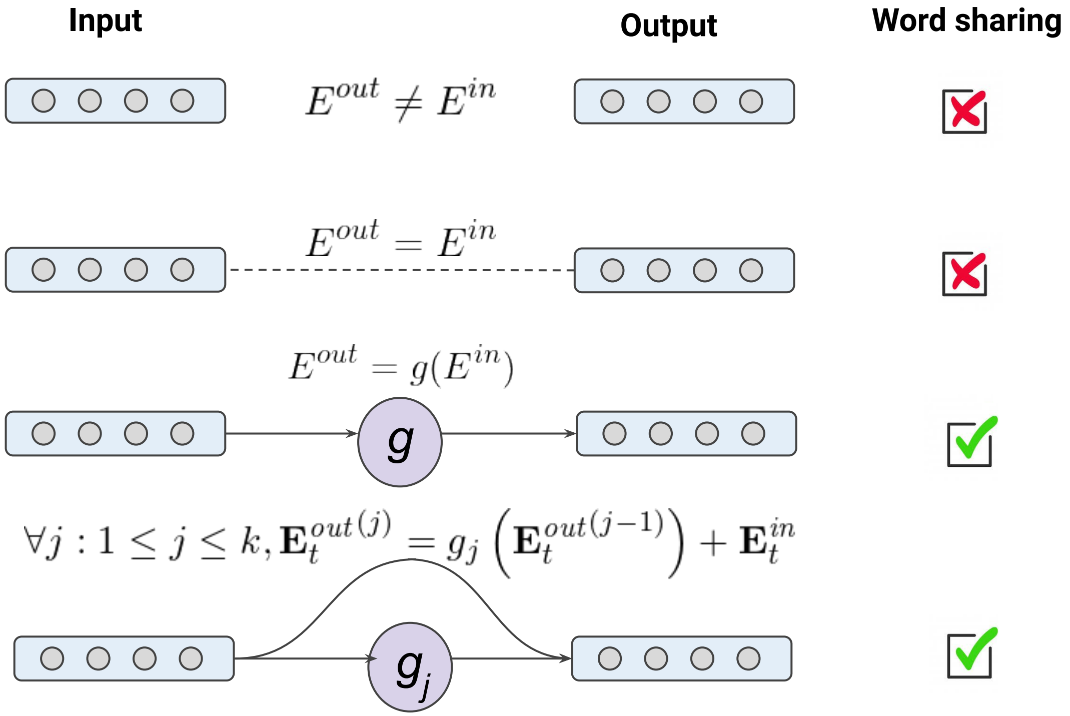
Lookup tables
(Zaremba et al., 2014;
Press & Wolf,2017)



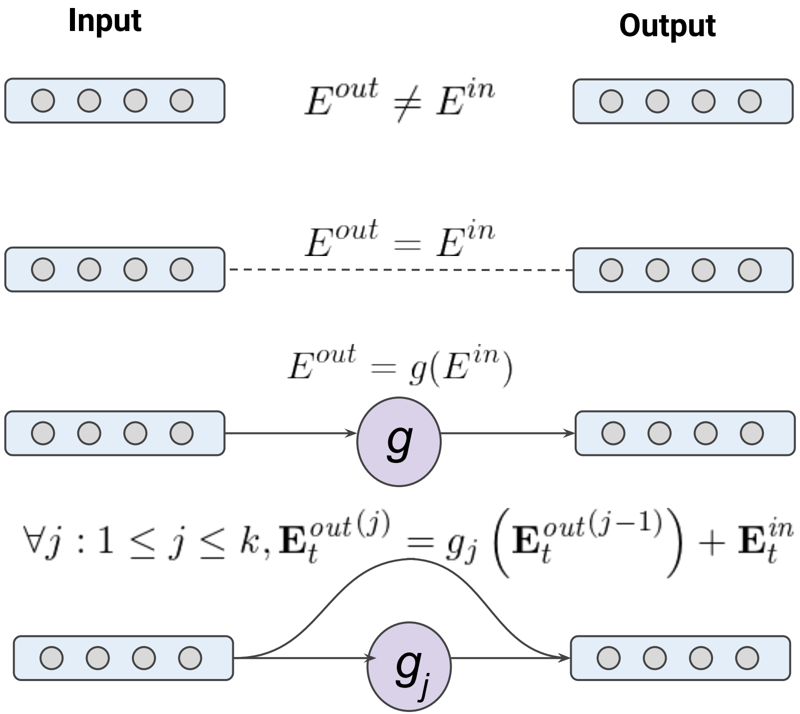
Sharing across words

Lookup tables
(Zaremba et al., 2014;
Press & Wolf,2017)

Compositional /
Functional forms
(Jozefowicz et al., 2016;
Baevski & Auli, 2019;
Pappas & Henderson, 2019)



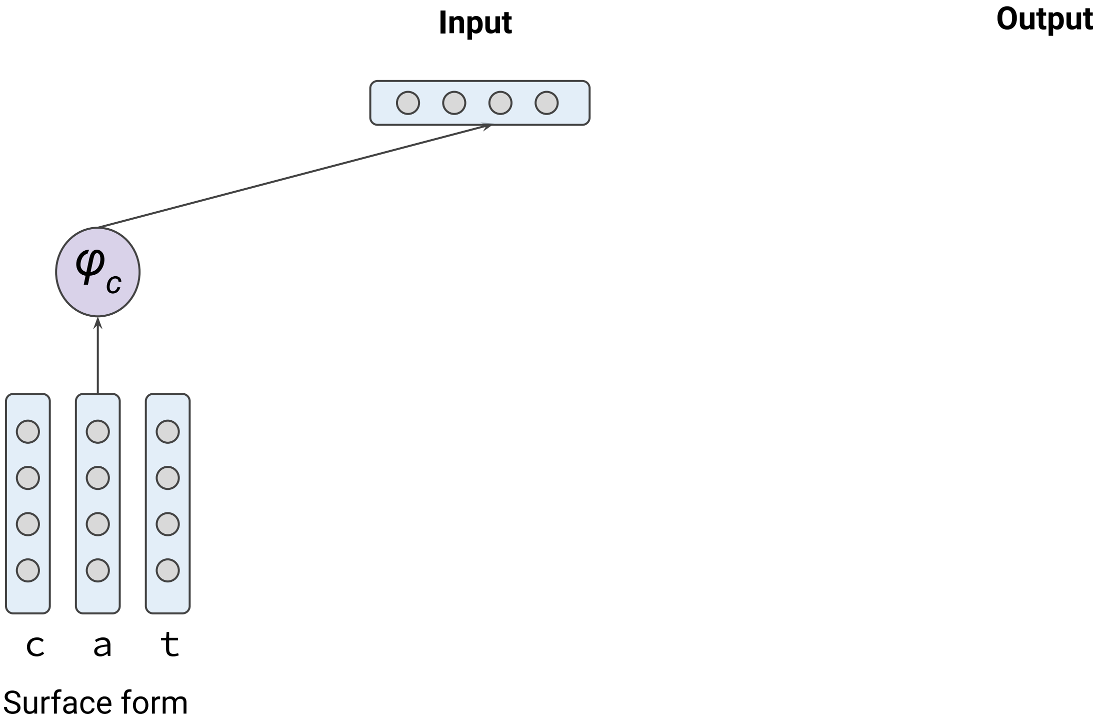
Sharing across words



	Input → Output	Word sharing	Vocabulary independent
Lookup tables (Zaremba et al., 2014; Press & Wolf,2017)	$E^{out} \neq E^{in}$	✗	✗
	$E^{out} = E^{in}$	✗	✗
Compositional / Functional forms (Jozefowicz et al., 2016; Baevski & Auli, 2019; Pappas & Henderson, 2019)	$E^{out} = g(E^{in})$	✓	✗
	$\forall j : 1 \leq j \leq k, \mathbf{E}_t^{out(j)} = g_j\left(\mathbf{E}_t^{out(j-1)}\right) + \mathbf{E}_t^{in}$	✓	✗

Our method — GroC

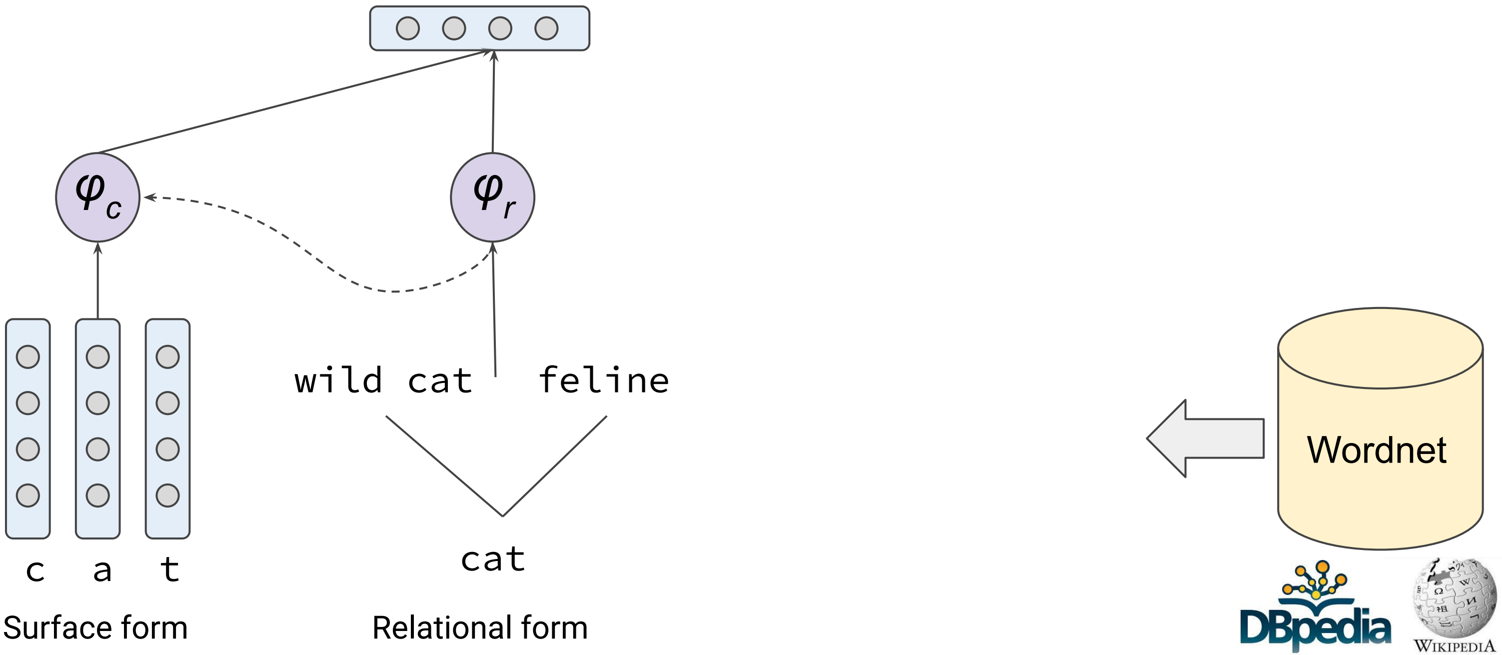
Grounded compositional forms



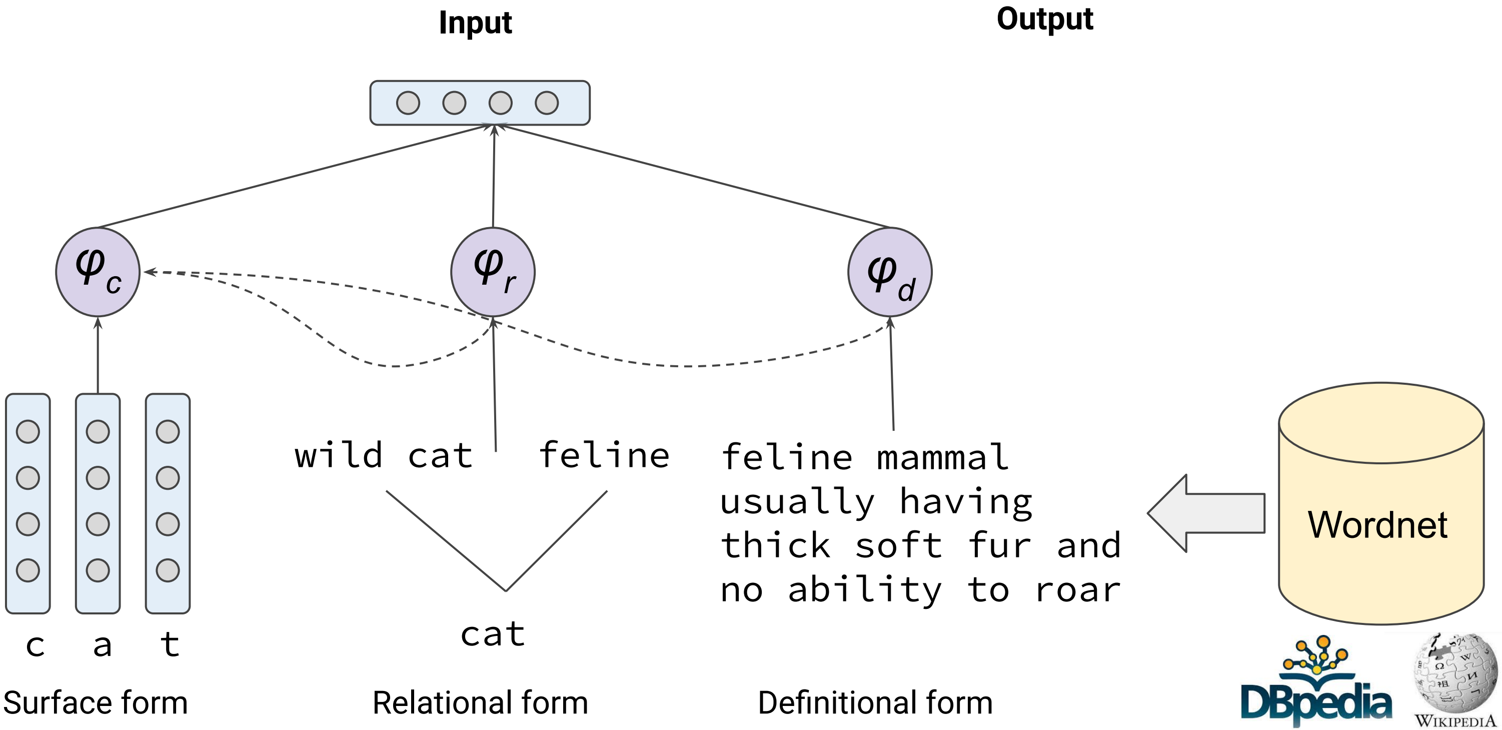
Grounded compositional forms

Input

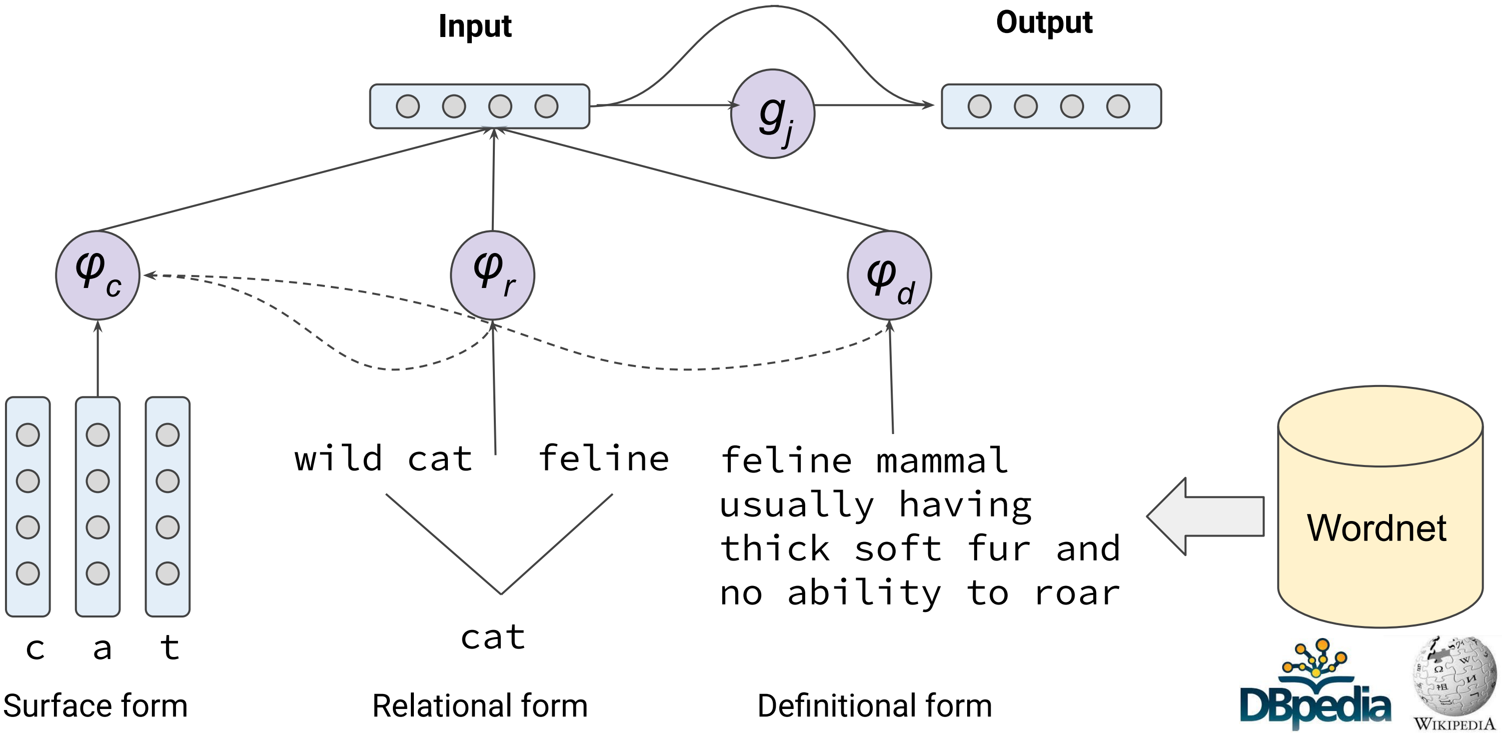
Output



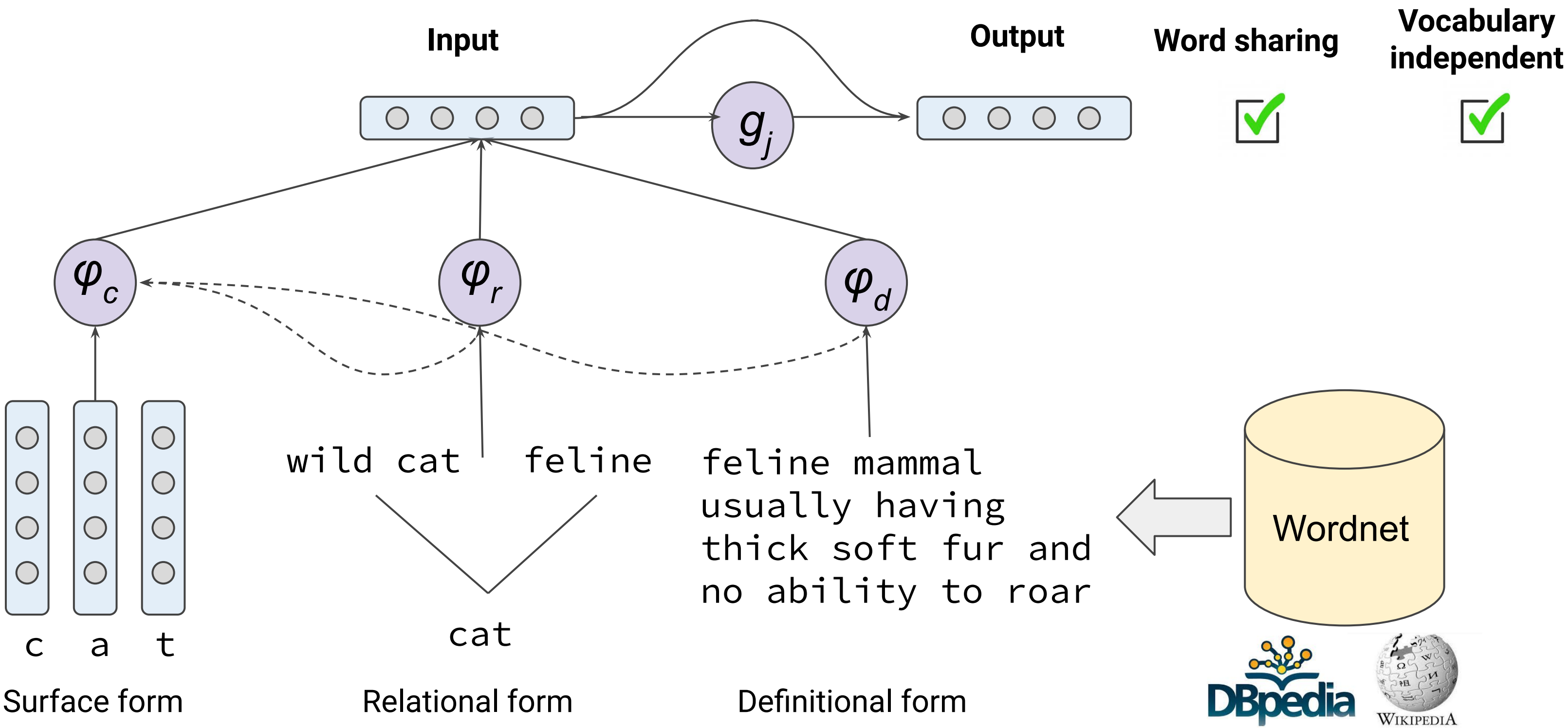
Grounded compositional forms



Grounded compositional forms



Grounded compositional forms



Adapting to any vocabulary

$$p(X_t = x_t \mid \mathbf{h}_{t-1}) \propto \exp\left(\mathbf{E}^{out}\mathbf{h}_{t-1} + \mathbf{b}\right)$$

Adapting to any vocabulary

- We first represent the vocabulary with GroC

$$\mathbf{E}^{out} = \mathtt{GroC}(\mathcal{V}^*)$$

$$p(X_t = x_t \mid \mathbf{h}_{t-1}) \propto \exp\left(\mathbf{E}^{out}\mathbf{h}_{t-1} + \mathbf{b}\right)$$

Adapting to any vocabulary

- We first represent the vocabulary with GroC

$$\mathbf{E}^{out} = \mathtt{GroC}(\mathcal{V}^*)$$

- Then we estimate the bias for each word *u*

$$b_v = \sigma\left(\mathbf{w} \cdot \mathbf{e}_v^{out} + a\right)$$

$$p(X_t = x_t \mid \mathbf{h}_{t-1}) \propto \exp\left(\mathbf{E}^{out}\mathbf{h}_{t-1} + \mathbf{b}\right)$$

GroC summary

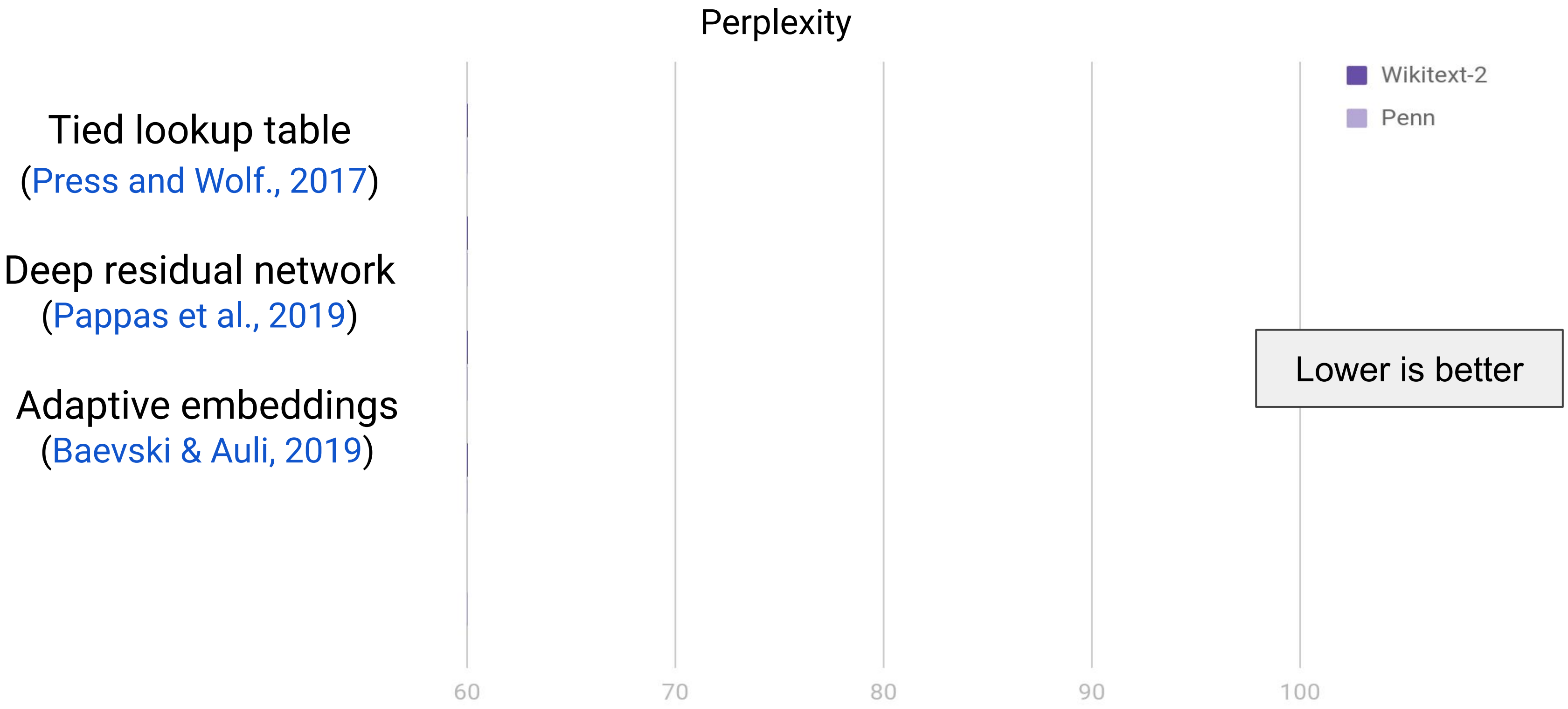
- Creates a compact representation of any vocabulary
- Grounds the language model predictions to prior knowledge
- Enables the decoupling of training and test vocabularies

Experiments

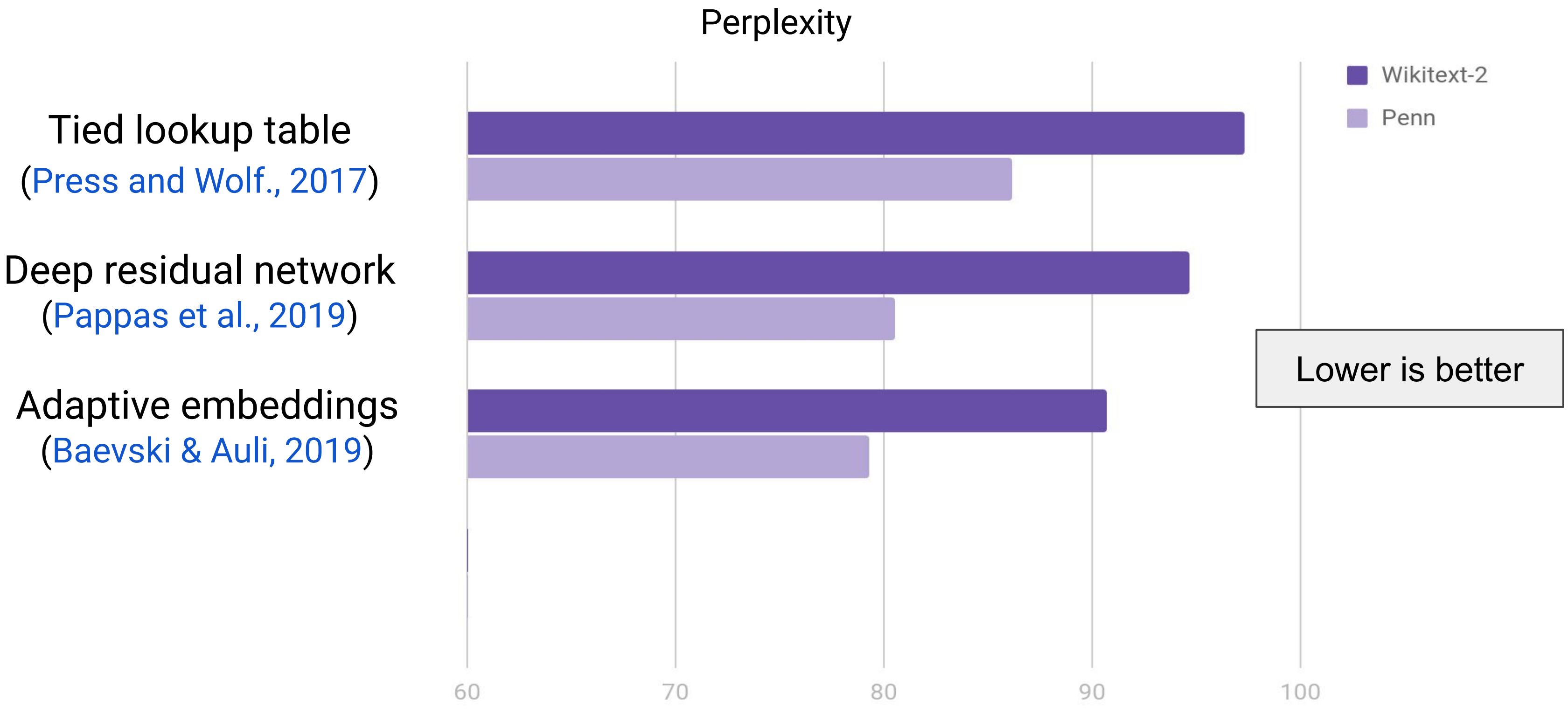
Conventional language modeling

How GroC compares to previous output embedding methods?

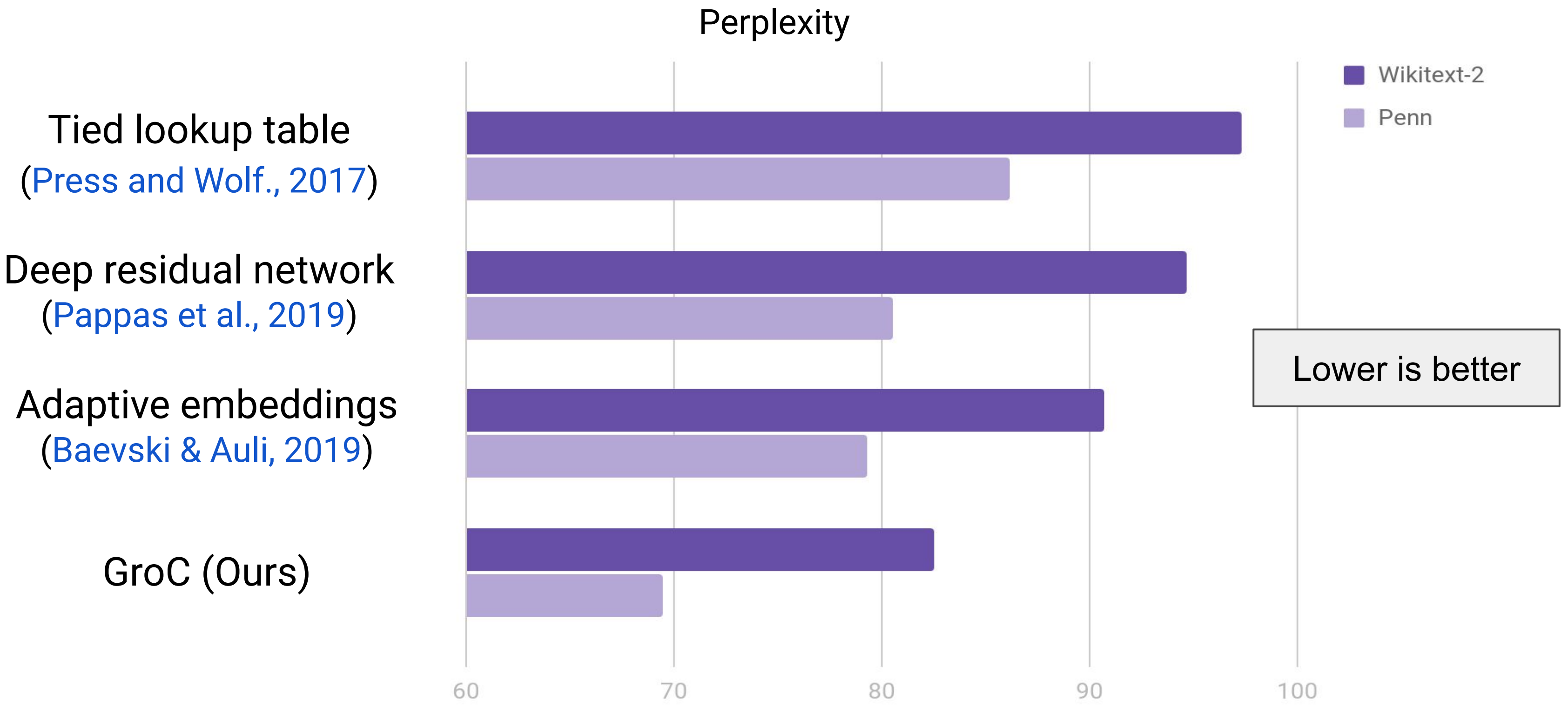
Conventional language modeling



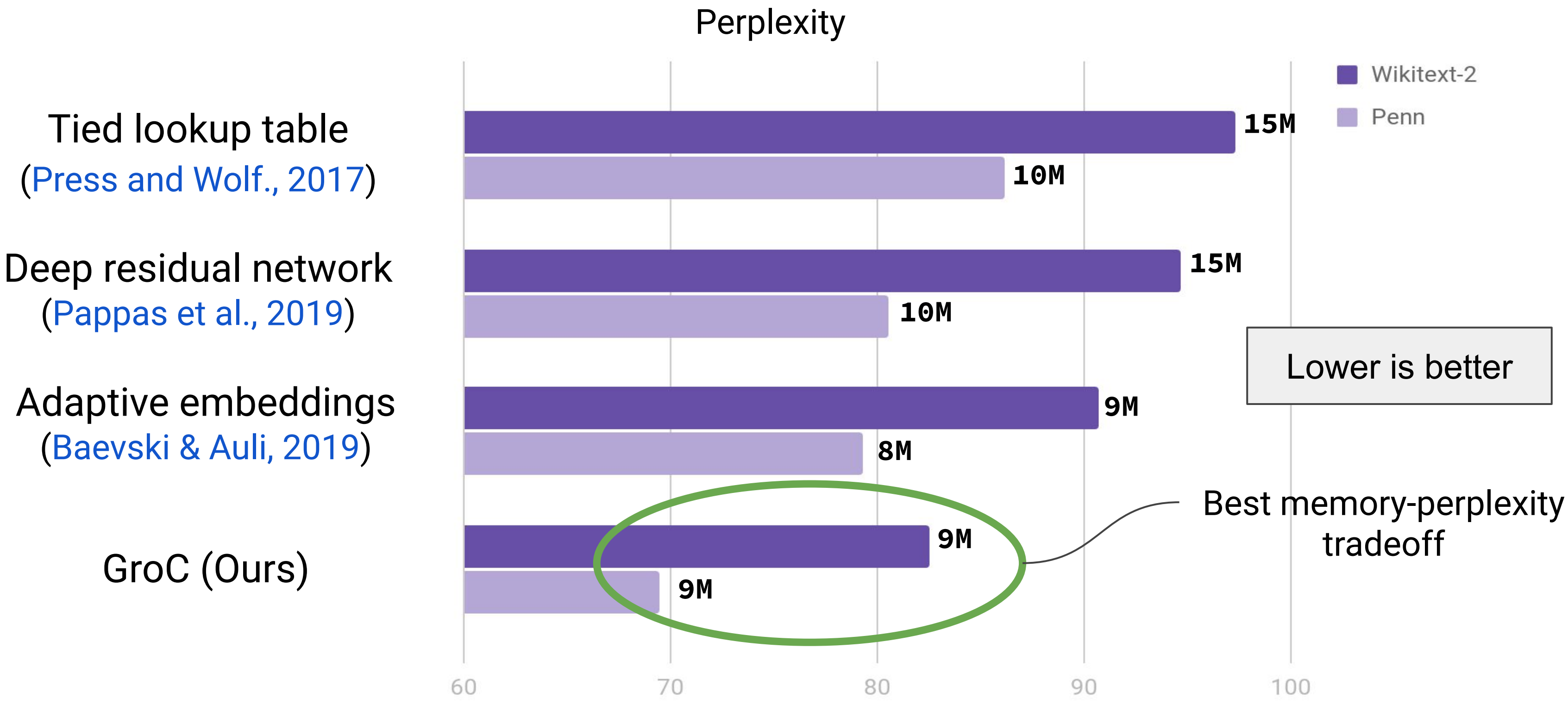
Conventional language modeling



Conventional language modeling



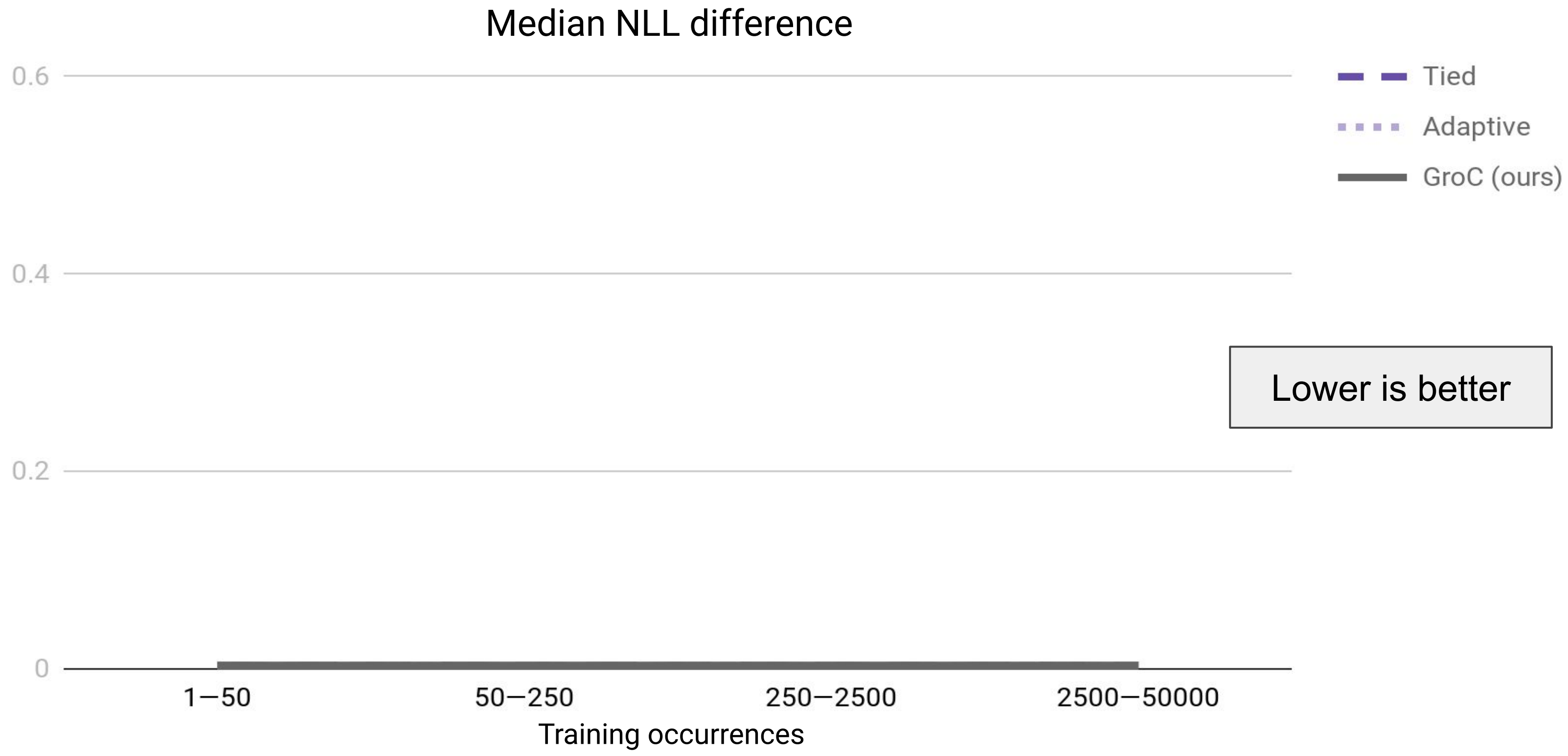
Conventional language modeling



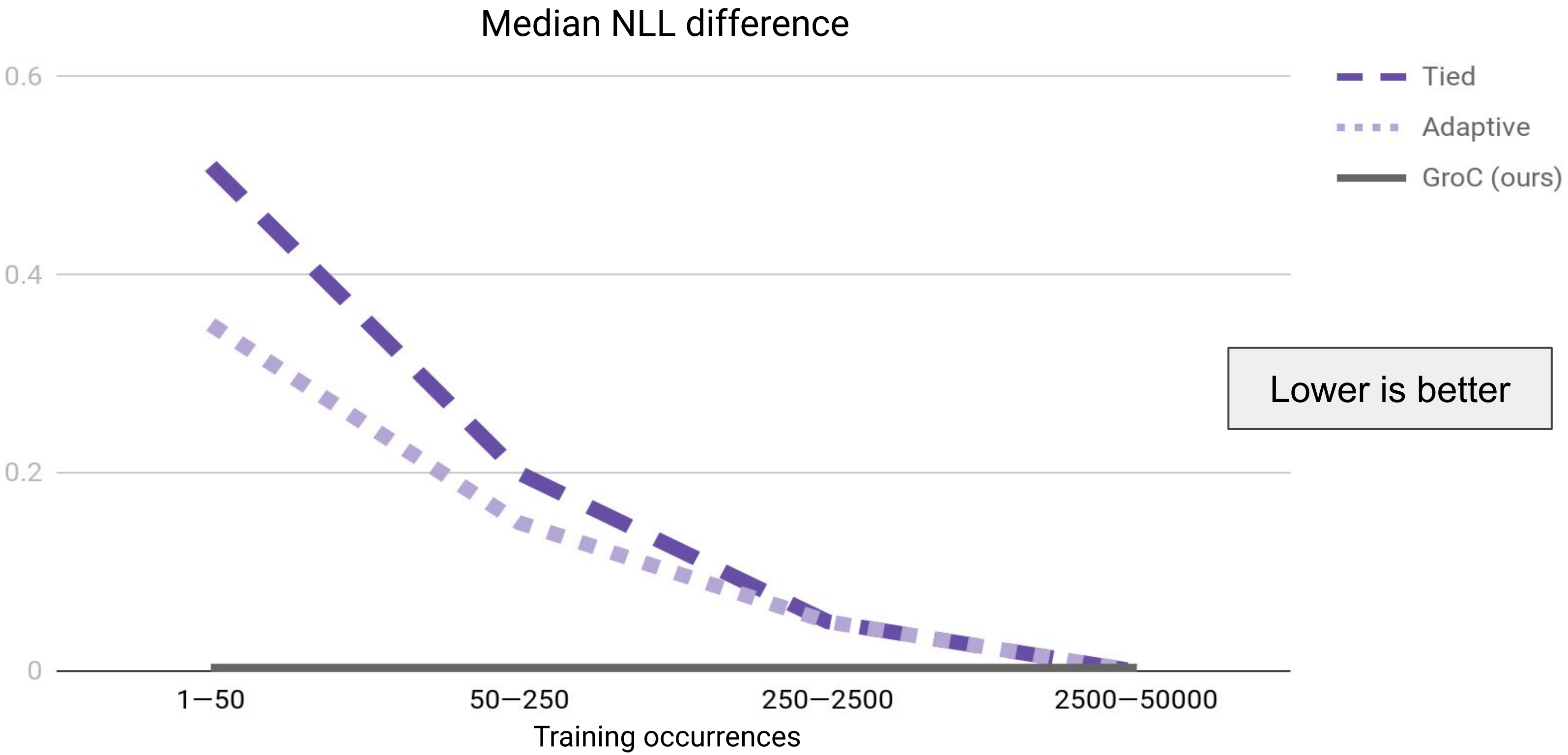
Conventional language modeling

Where does the improvement come from?

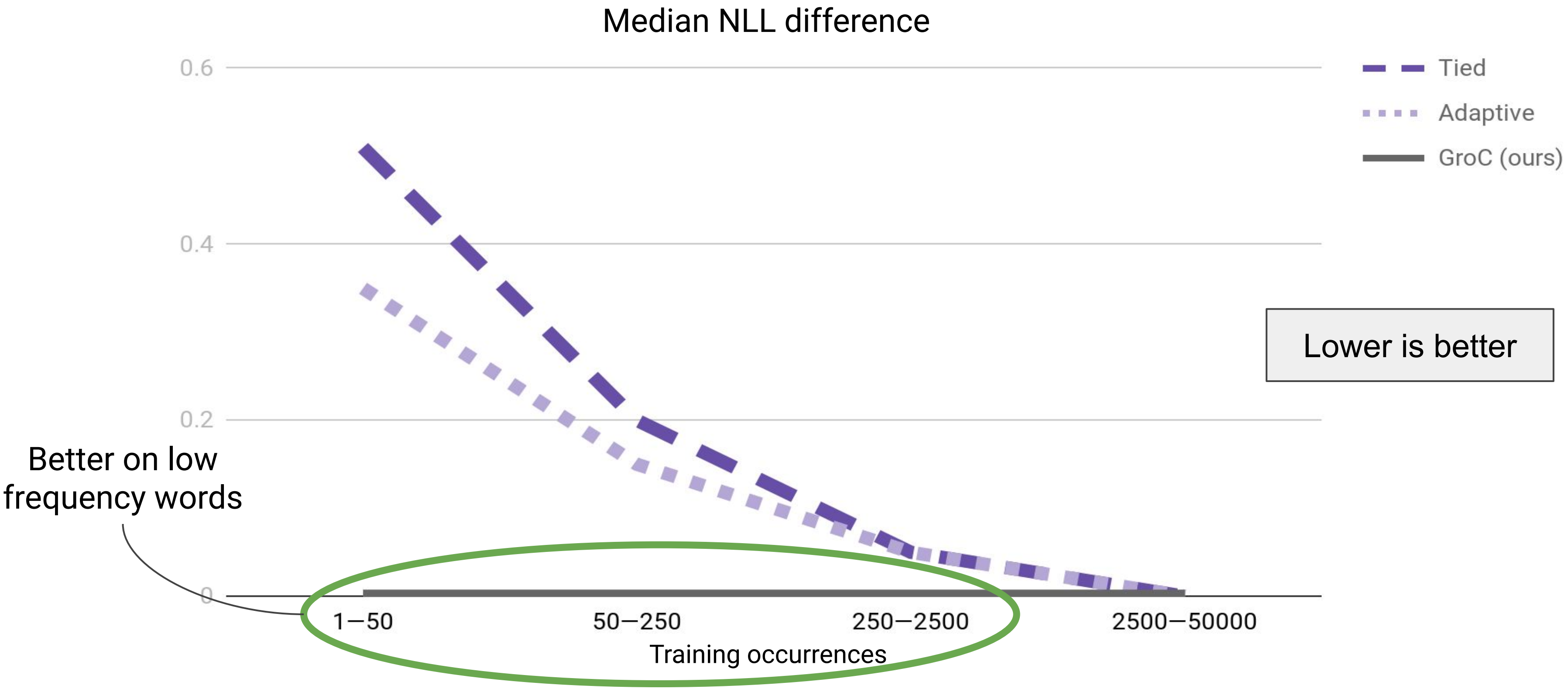
Break down by frequency



Break down by frequency



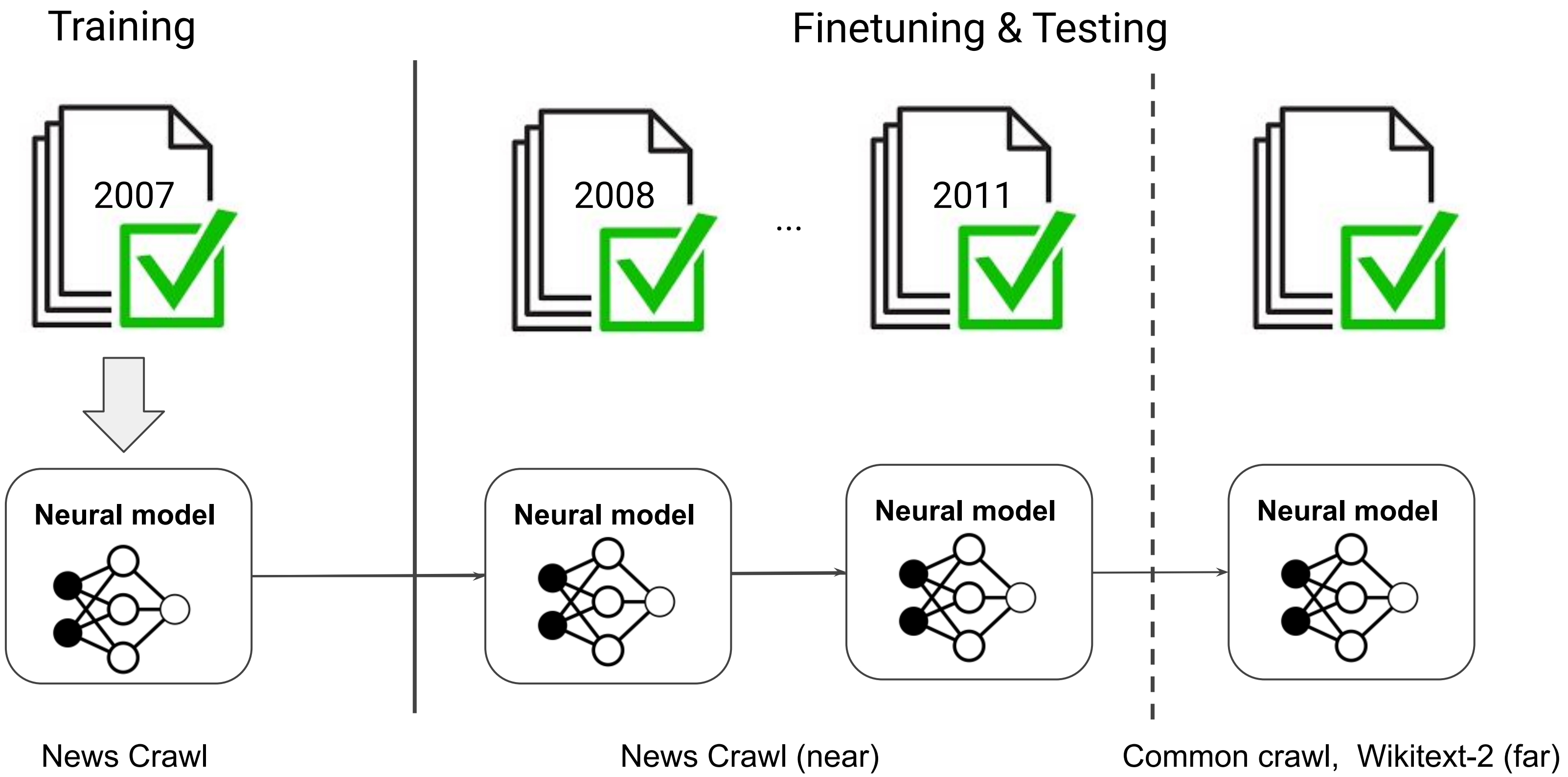
Break down by frequency



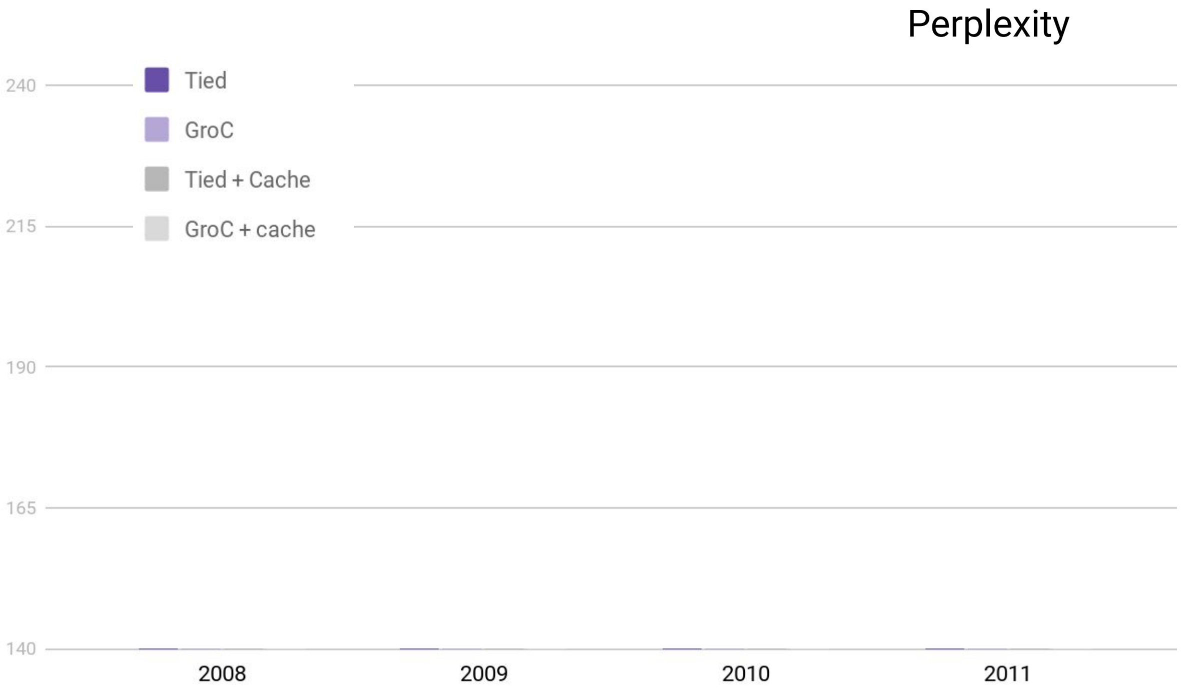
Cross-domain modeling: Zero resource

Does GroC generalize on zero resource adaptation settings?

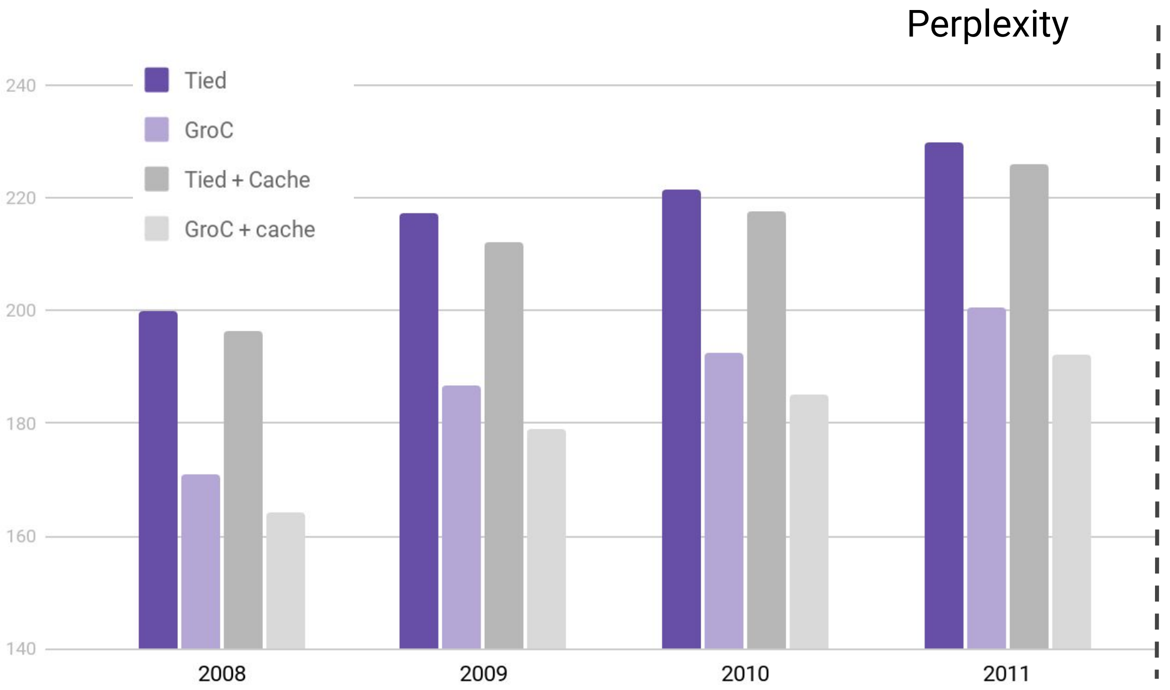
Cross-domain modeling: Zero resource



Cross-domain modeling: Zero resource

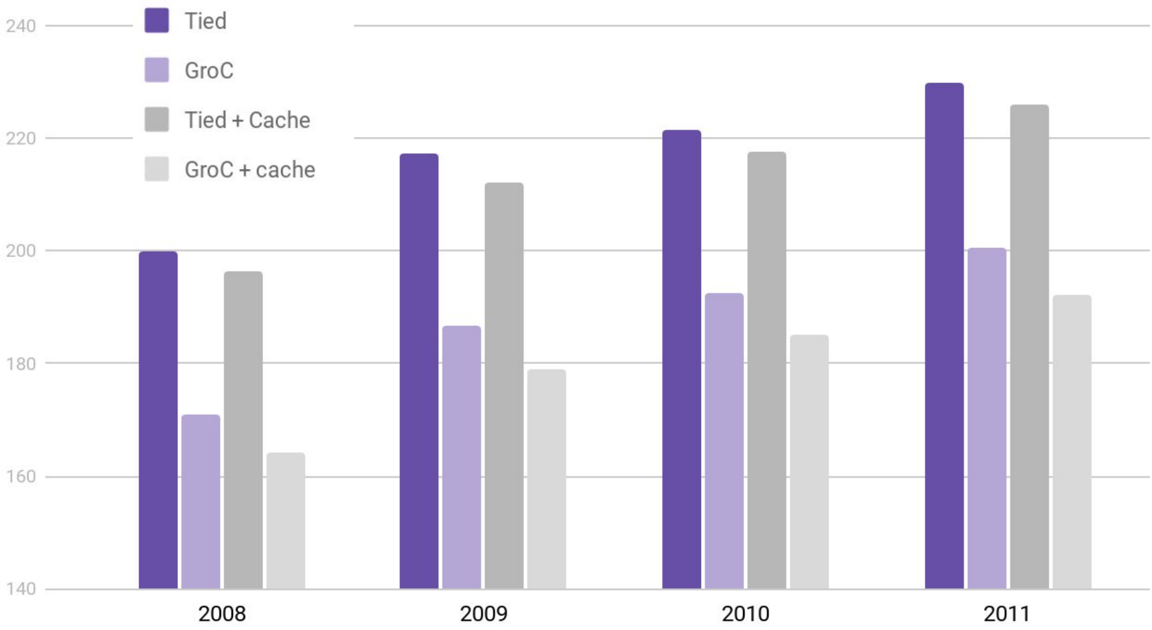


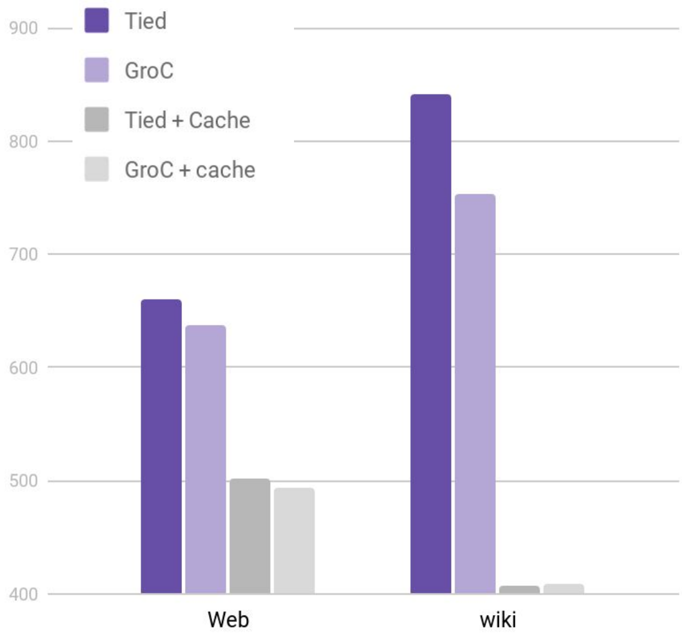
Cross-domain modeling: Zero resource



Cross-domain modeling: Zero resource

Perplexity

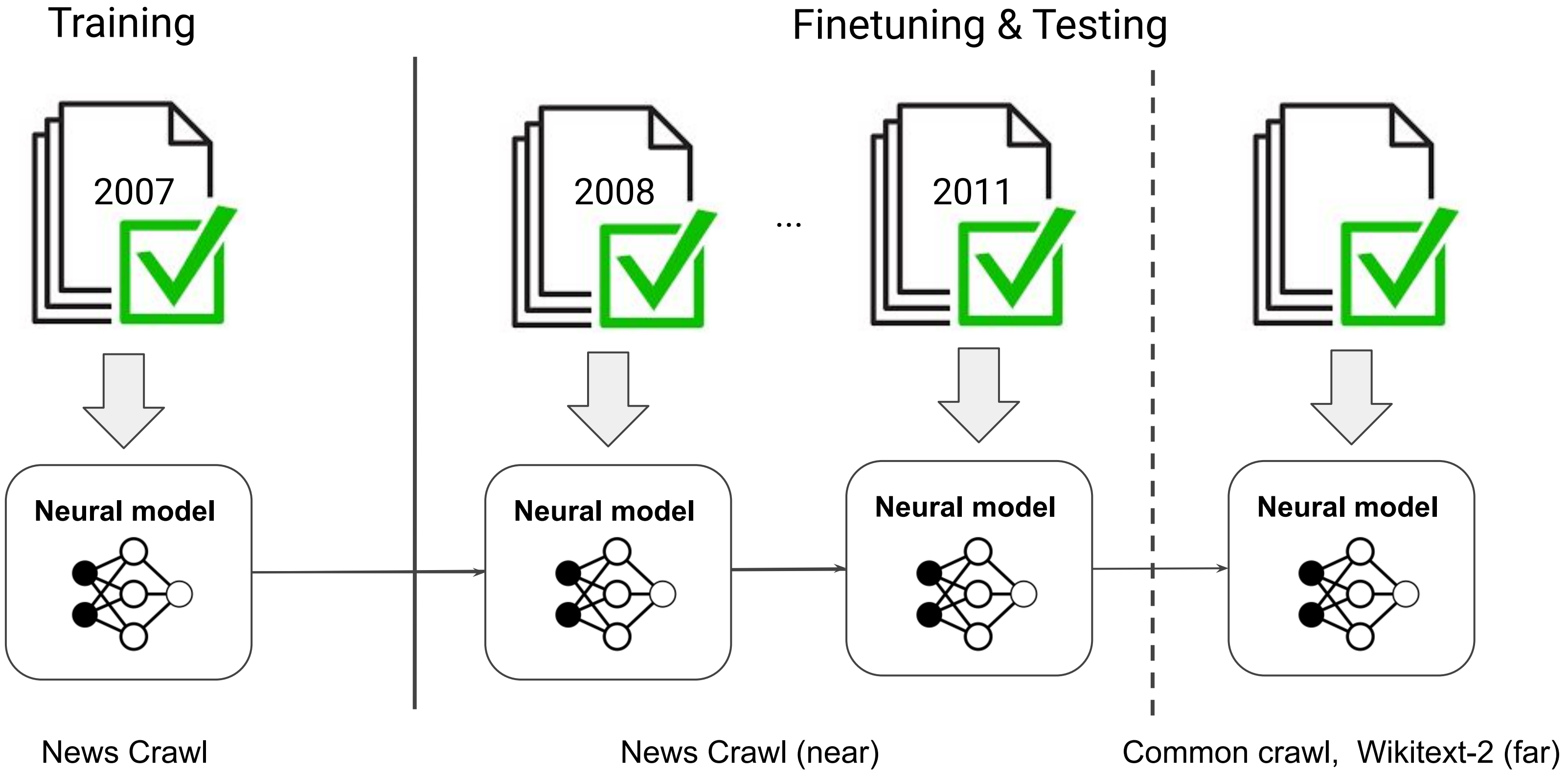




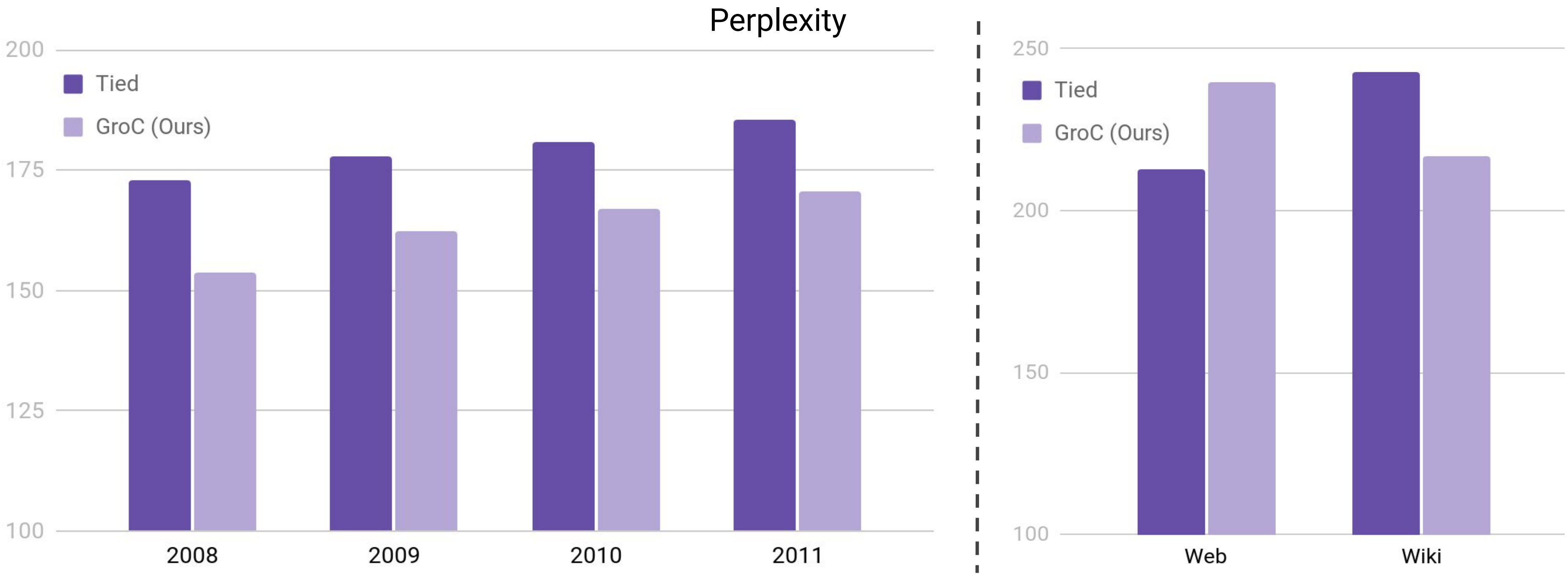
Cross-domain modeling: Low resource

Does GroC help on low resource adaptation settings?

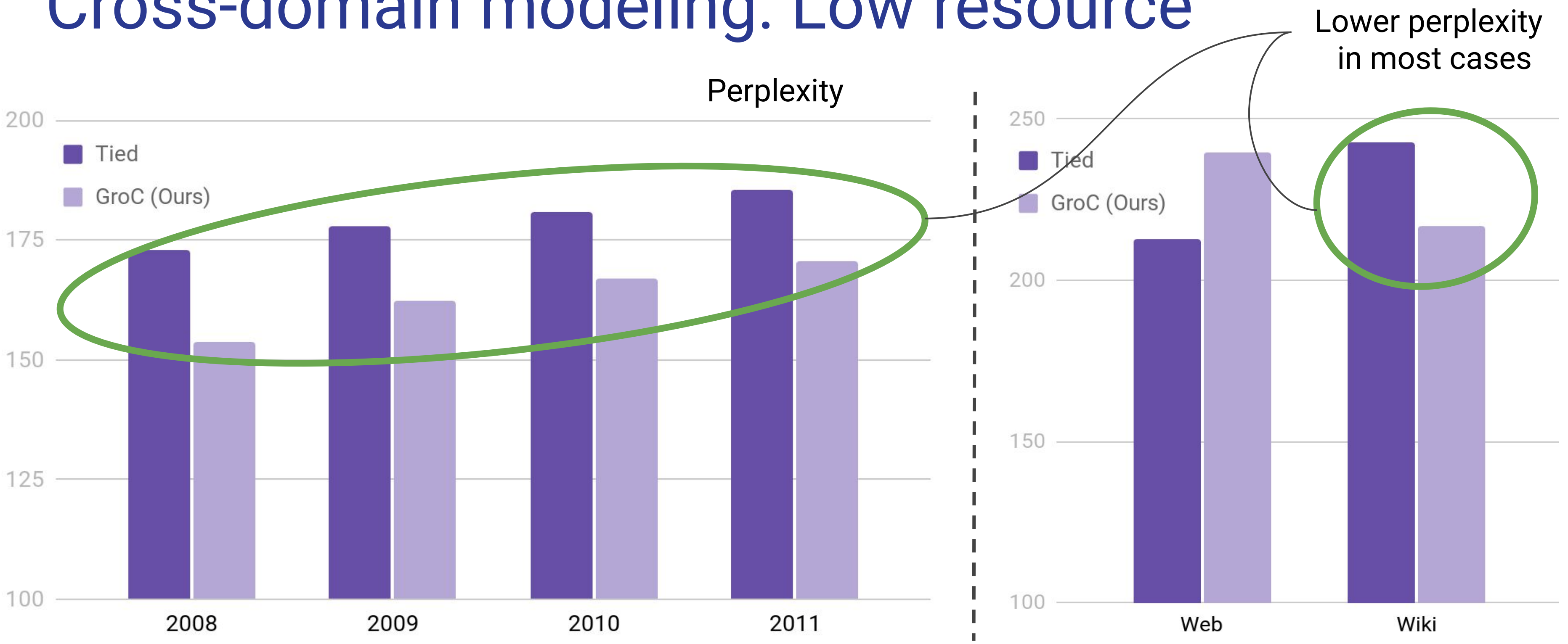
Cross-domain modeling: Low resource



Cross-domain modeling: Low resource



Cross-domain modeling: Low resource



Conclusion

Grounded compositional outputs for language models

- Outperform previous methods on conventional settings
- Achieve low perplexity on rare words
- Generalize well to previously unseen domains

Thank you

https://github.com/Noahs-ARK/groc