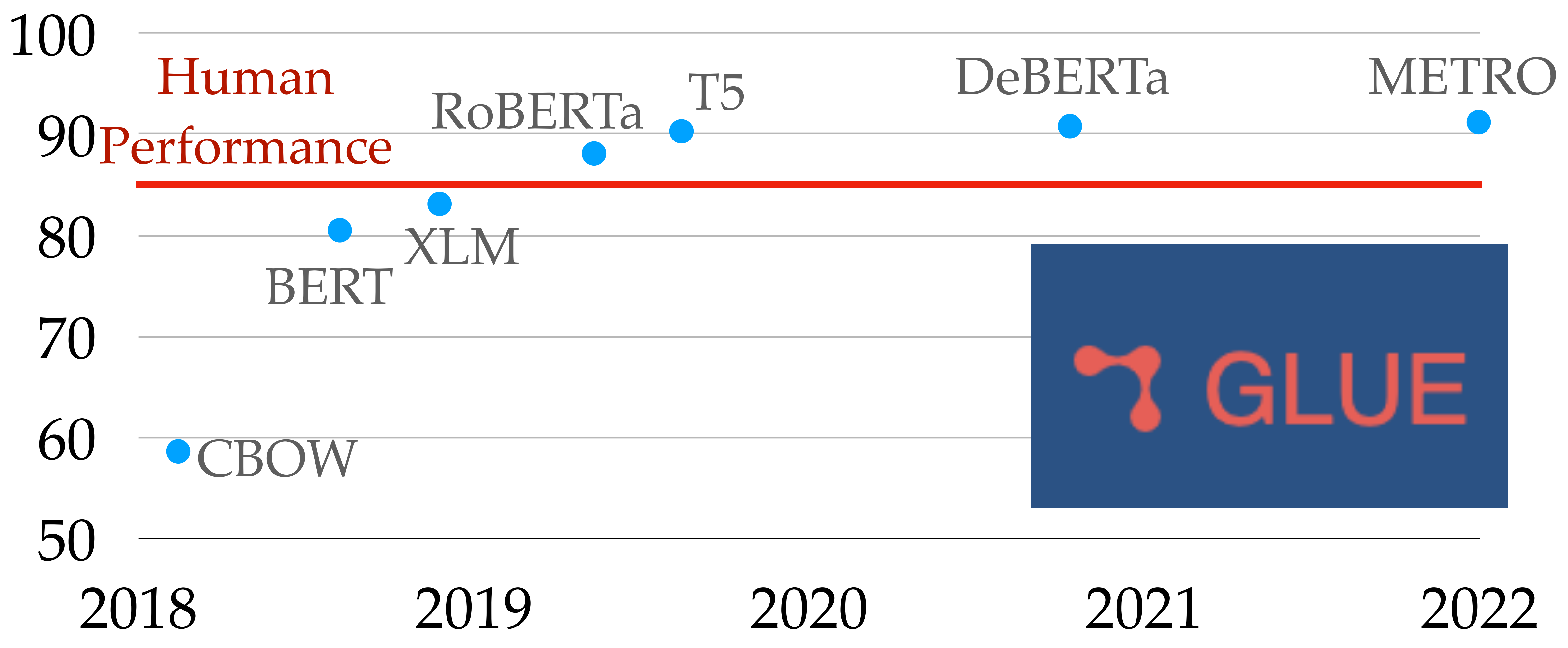
Crosslingual Sharing for Low-Resource Natural Language Processing

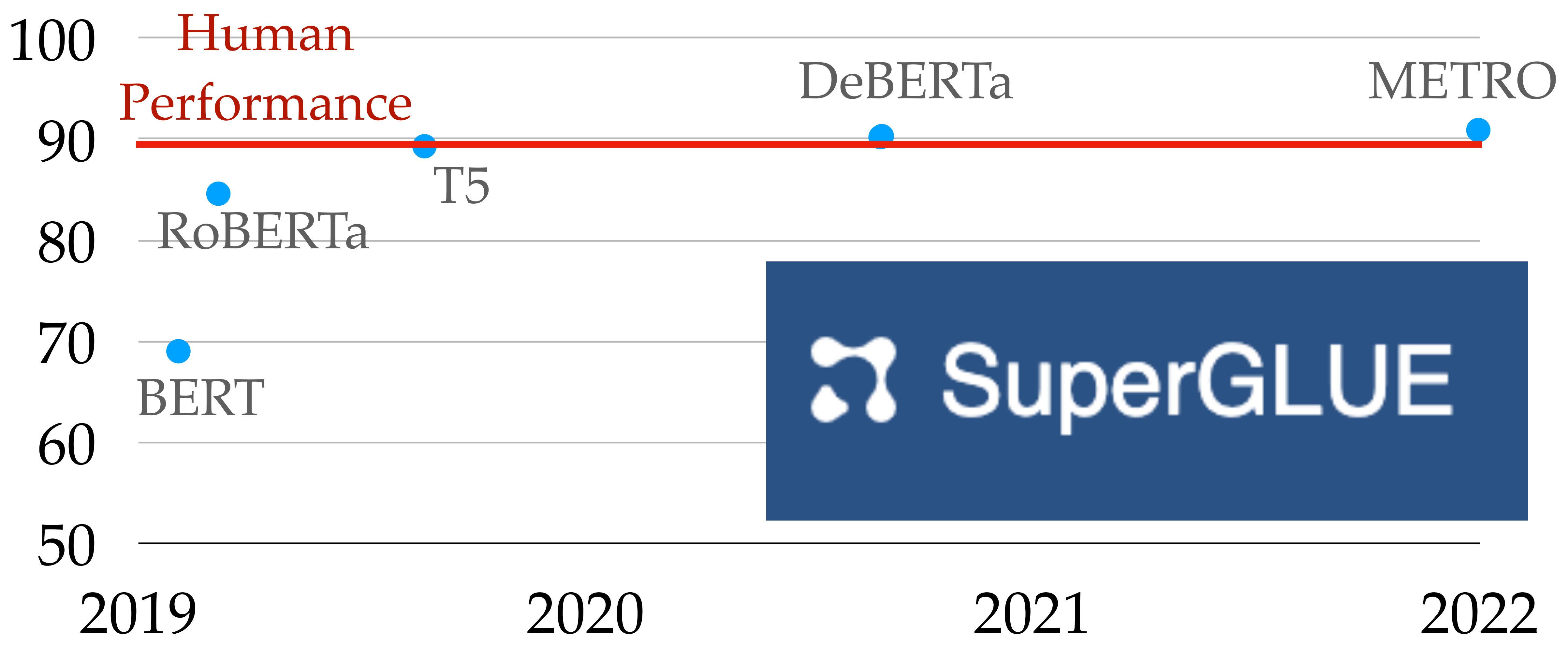
Phoebe Mulcaire

(in collaboration with Swabha Swayamdipta, Jungo Kasai, Nikolaos Pappas and Noah A. Smith)

large-scale NLP is wildly successful

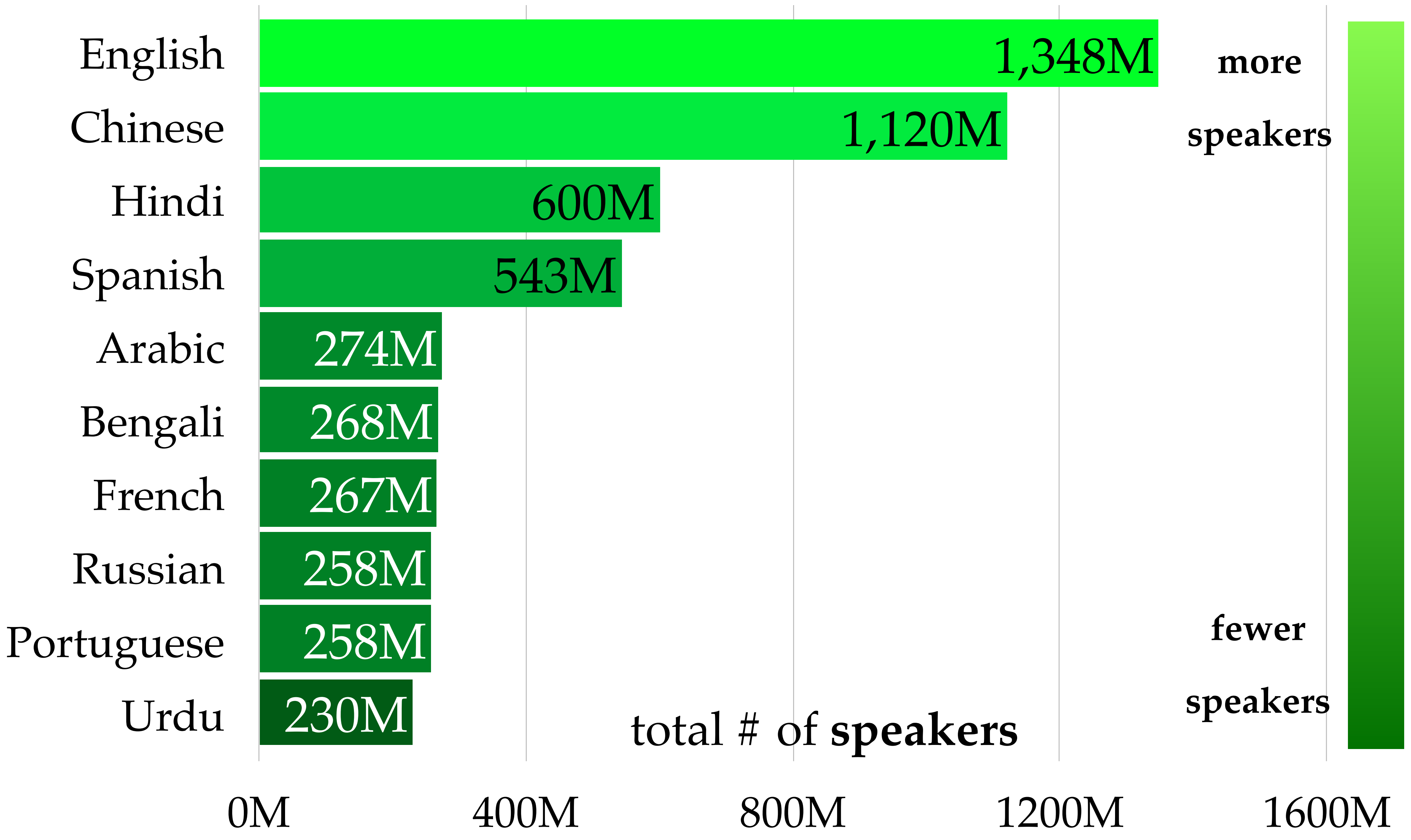
- BERT / GPT-3 / OPT / etc.: billions of params, trained on billions (even hundreds of billions!) of tokens
- result: English NLP is "solved"

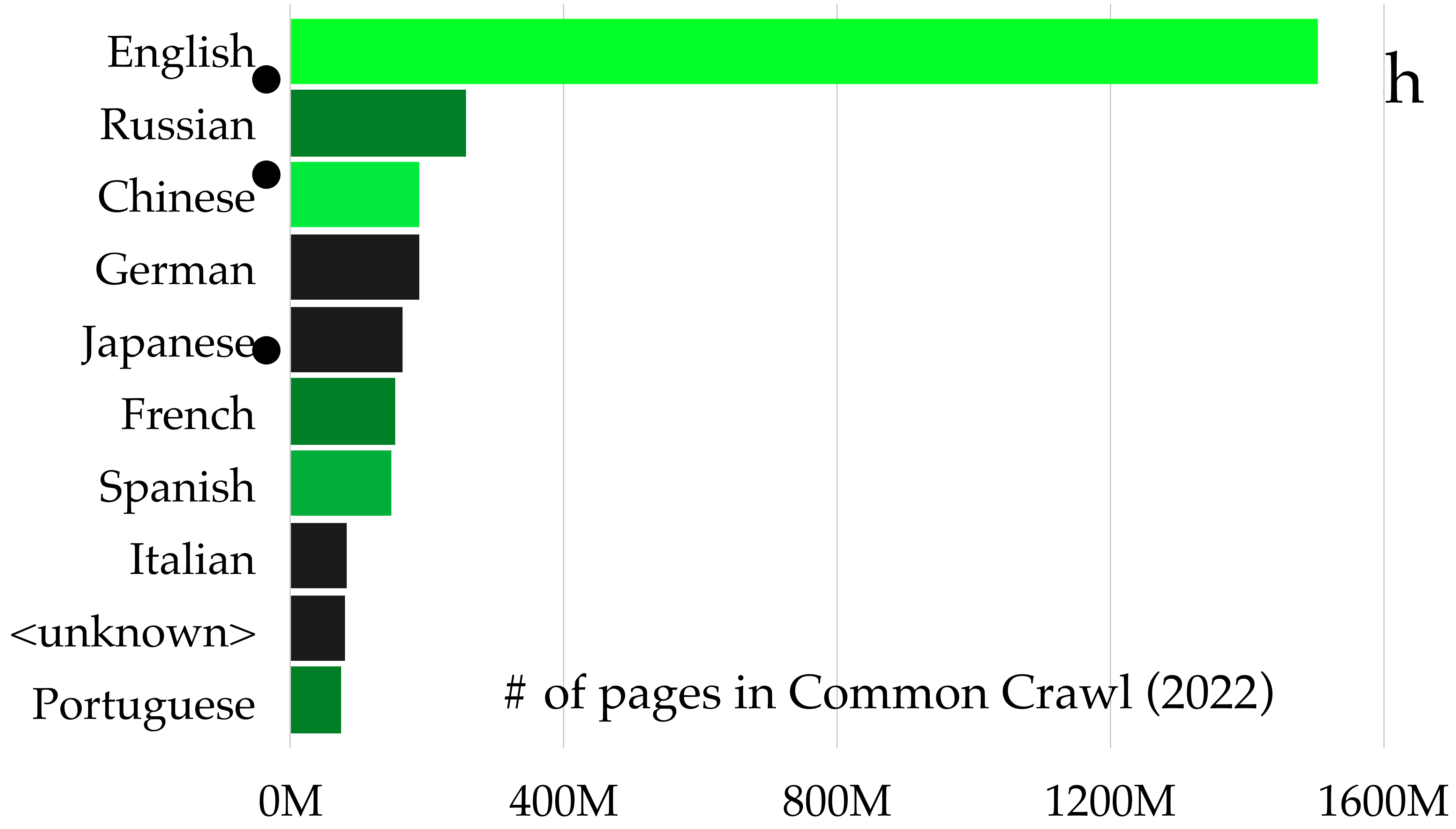




but… there's a resource gap

- Ethnologue records >7000 living languages[1]
- English is the most widely spoken language… but there's a fat tail
- the most widely *used* languages ≠ the ones with the most *resources*[2] (or research)

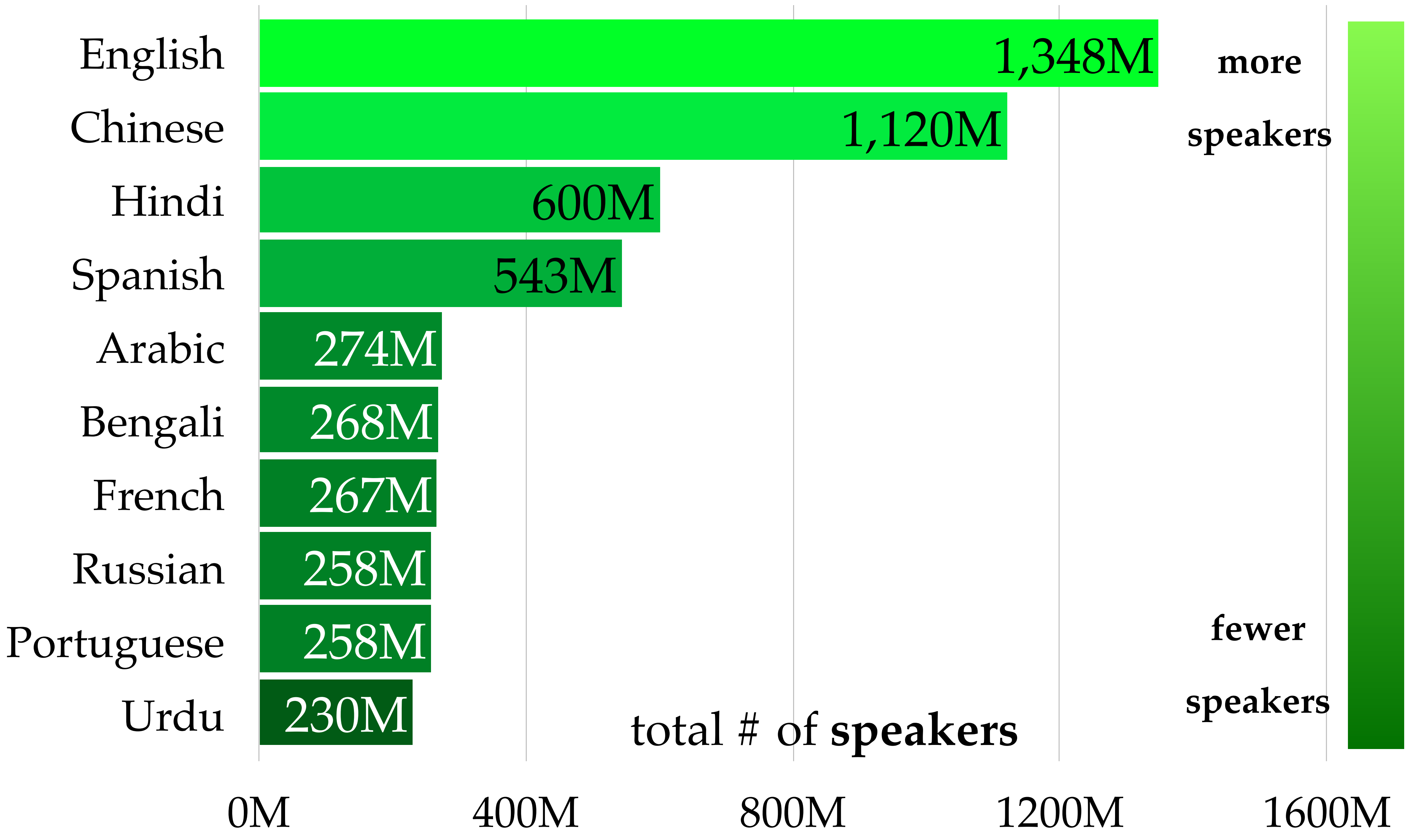


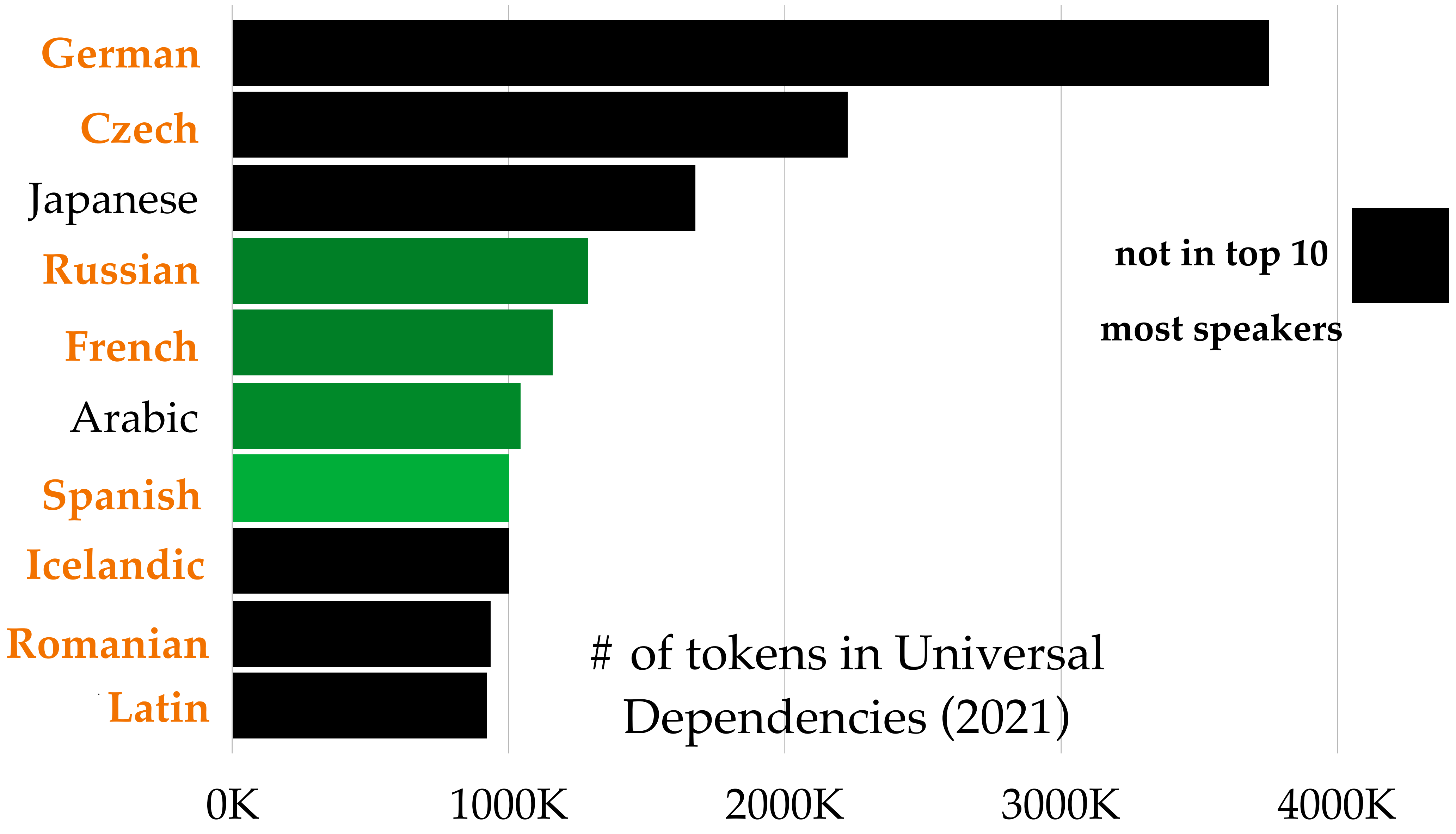


[1] Ethnologue: www.ethnologue.com
[2] Common Crawl

but… there’s a resource gap

- Ethnologue records >7000 living languages[1]
- English is the most widely spoken language… but there’s a fat tail
- the most widely *used* languages ≠ the ones with the most *resources*[2] (or research)
- most-resourced languages don’t reflect world’s linguistic diversity
- we can’t replicate English NLP for every language





[1] Ethnologue: www.ethnologue.com
[2] Common Crawl

language-universal NLP

we want systems that:

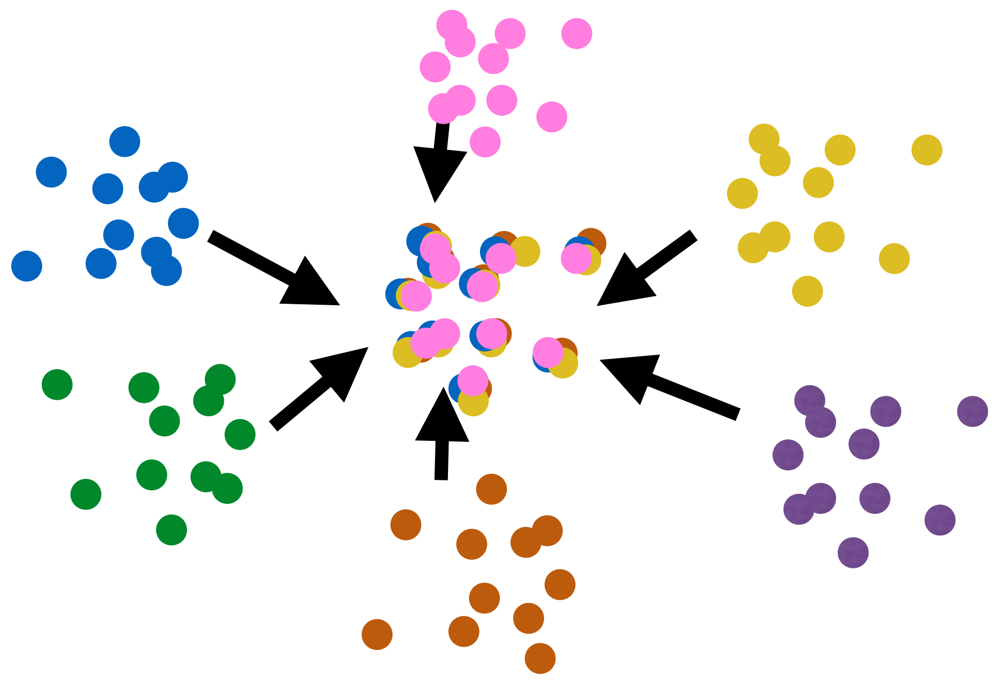
- don’t rely on large amounts of language-specific resources
- don’t rely on large amounts of language-specific researcher effort (e.g. custom architecture choices)

our focus:

- *crosslingual sharing*
- …via *polyglot* models
- …for *low-resource* settings

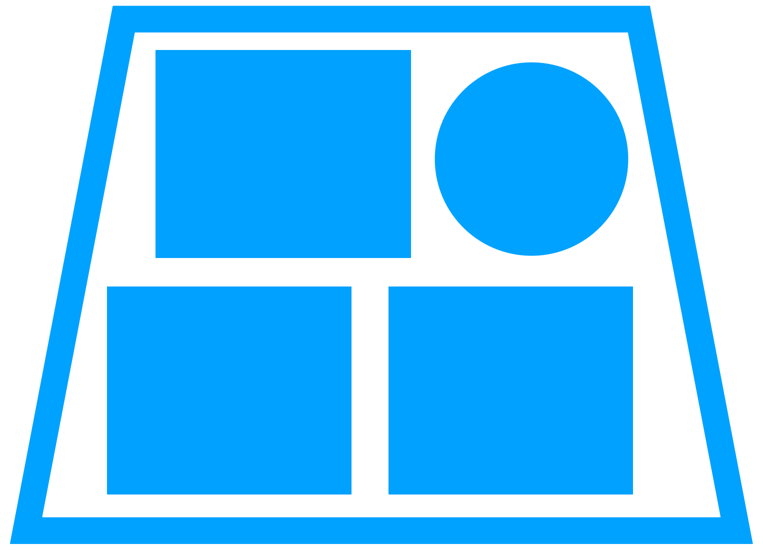
outline

- *Polyglot Semantic Role Labeling (Ch. 2)*
  - supervised, linguistic structure prediction



- *Polyglot Language Modeling (Ch. 3)*
  - language models for word representations

- *Grounded Compositional Output Embeddings (Ch. 4-5)*
  - low-resource language models



semantic role labeling

I think Peter even made some deals with the gorillas .

O O A0 AM-ADV O O A1 AM-ADV O O

Pero el suizo difícilmente atacará a Rominger en la montaña .

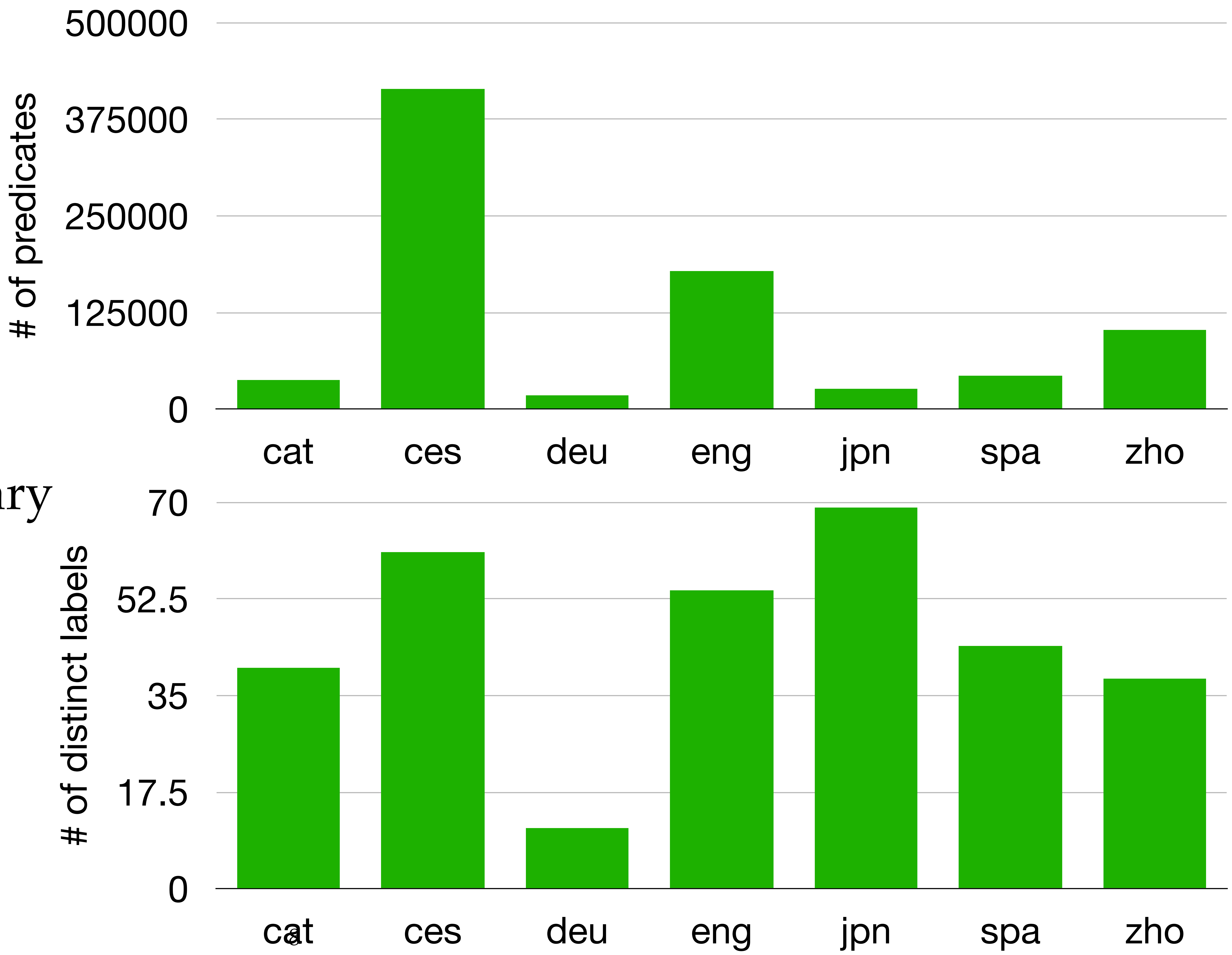
O O arg0-agt argM-adv O O arg1-pat argM-loc O O

Četrans oslovil sedm velkých evropských výrobců nákladních automobilů.

O O RSTR RSTR RSTR O O PAT

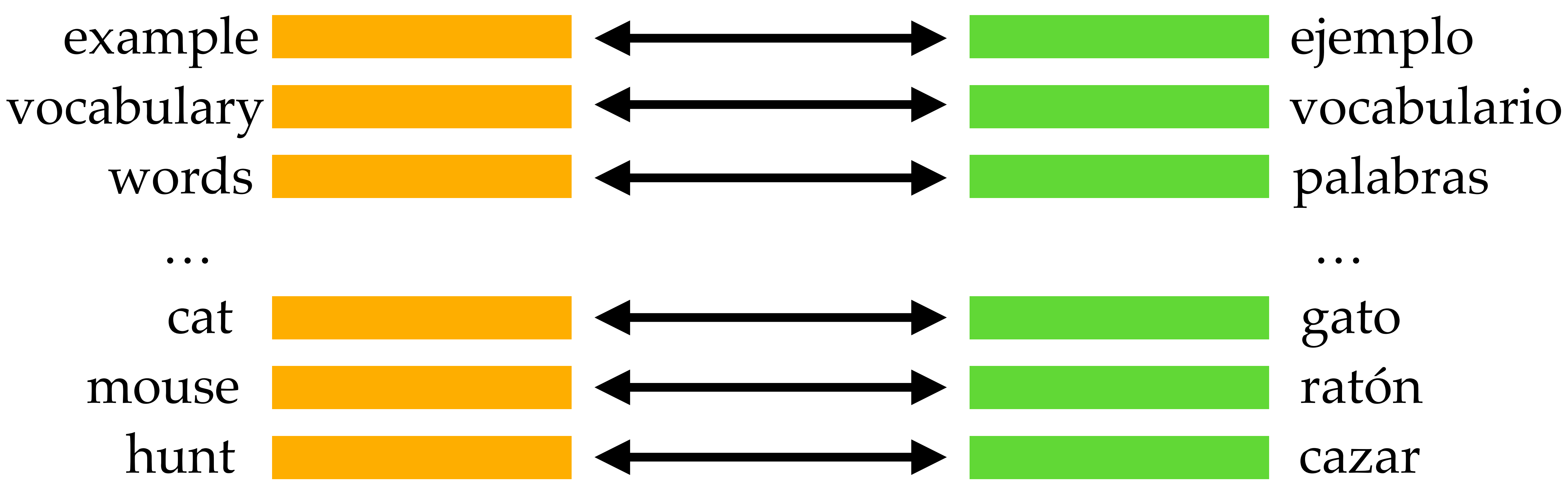
CoNLL 2009

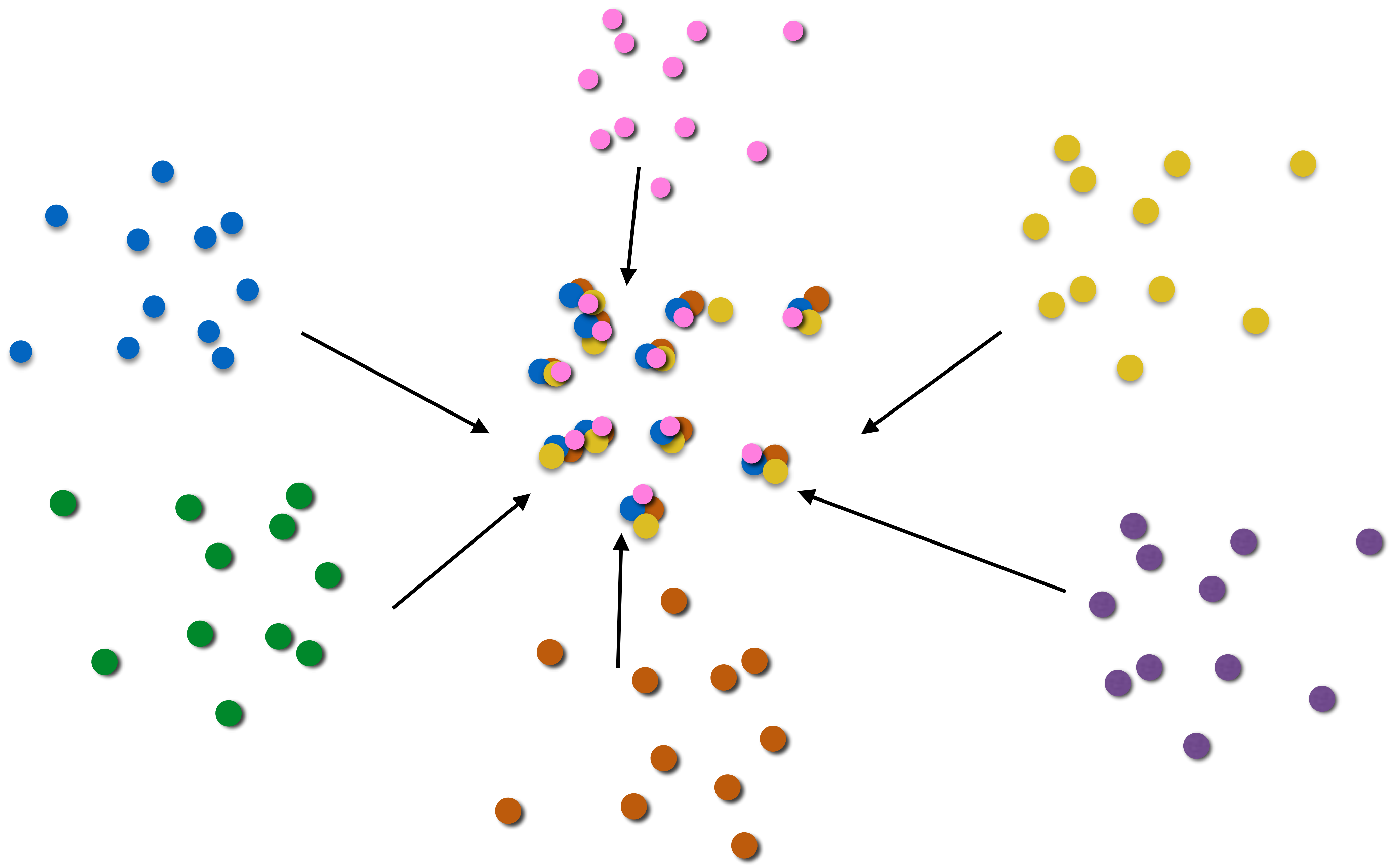
- format is the same, but:
- data is wildly imbalanced between languages—independent models (e.g. Zhao et al., 2009) would vary
- output labels vary—annotation projection (e.g. Padó and Lapata, 2005) is ruled out



our approach: polyglot semantics

- multilingual word vectors
- produce word vectors for each language based on co-occurrence statistics
- align to match English using a bilingual dictionary[1]

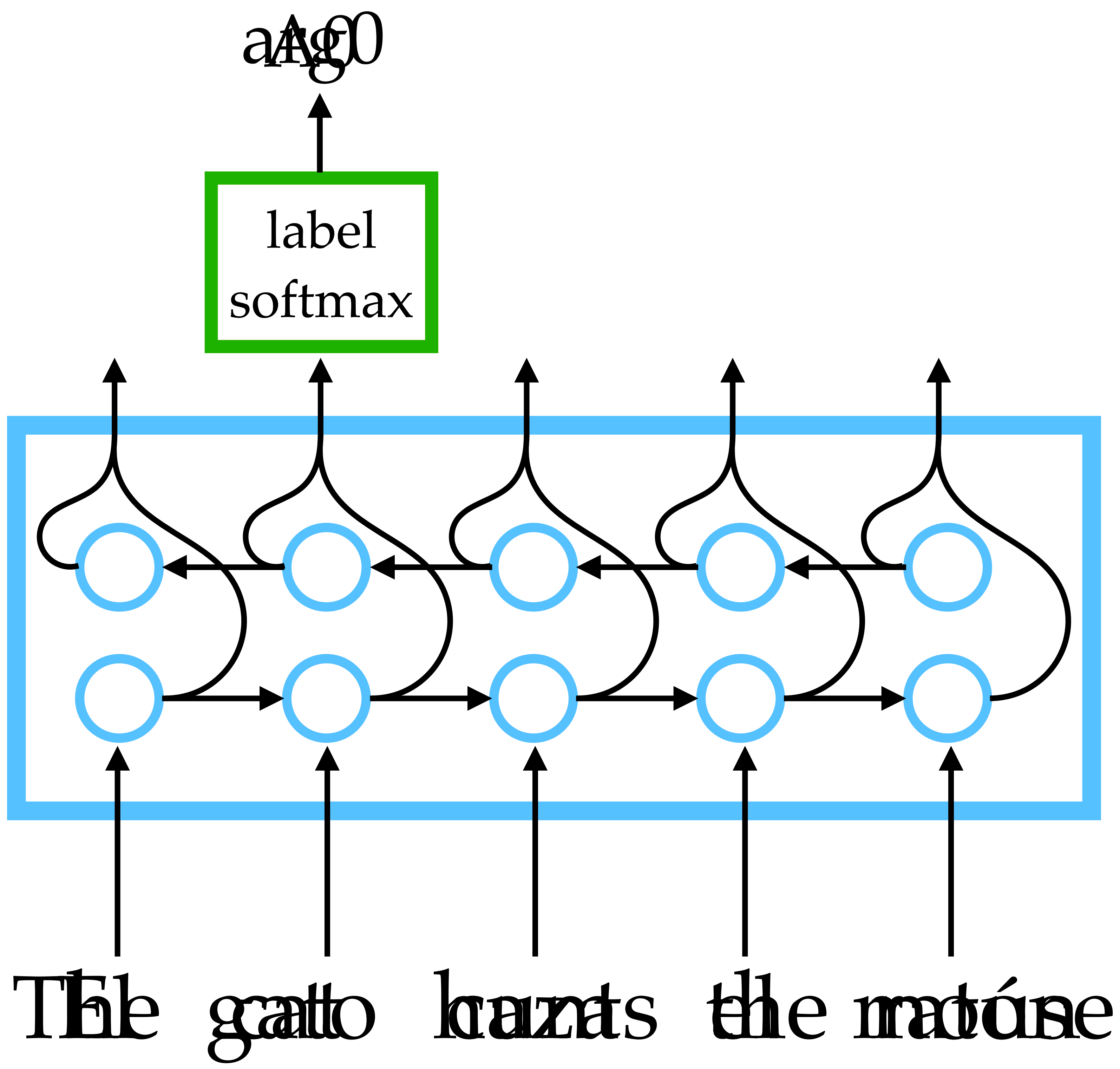




[1] Faruqui et al. (2014); Ammar et al. (2016)

our approach: polyglot semantics

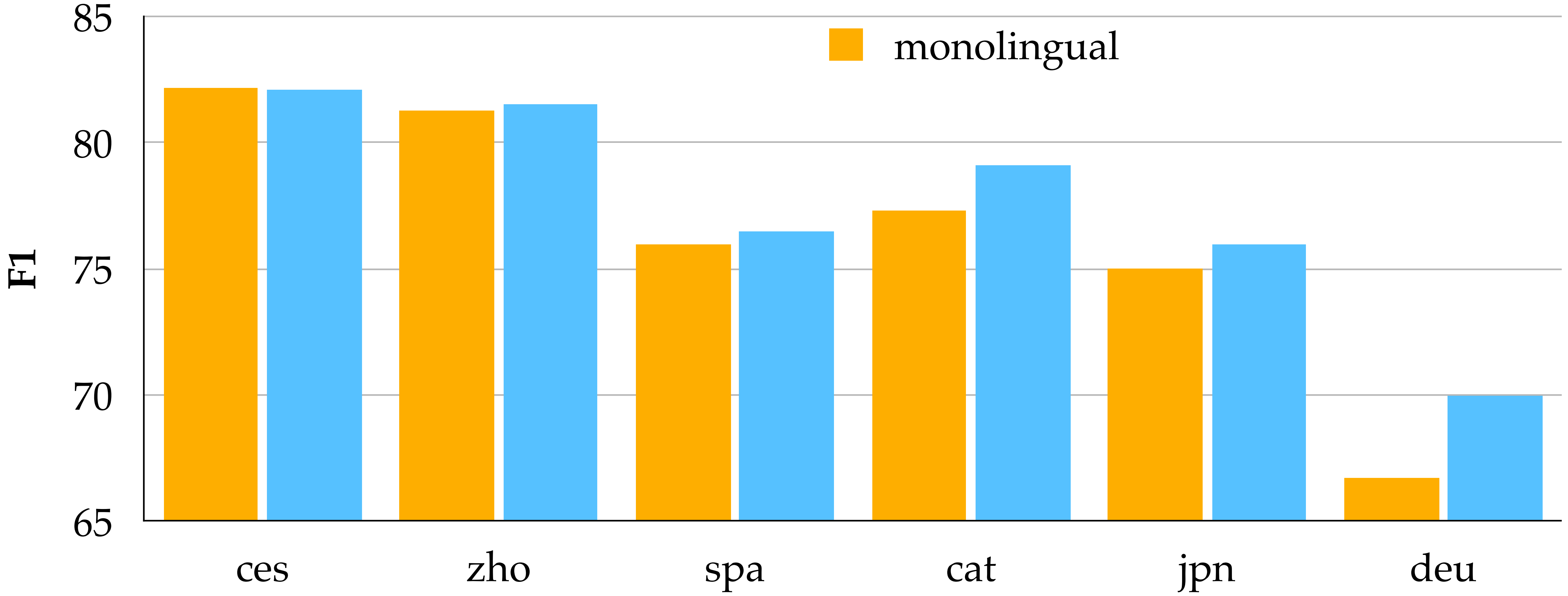
- task model based on a (then) SOTA monolingual model[1]
- multilingual word vector inputs
- sharing in parameters: deep bi-LSTM
- independent label embeddings

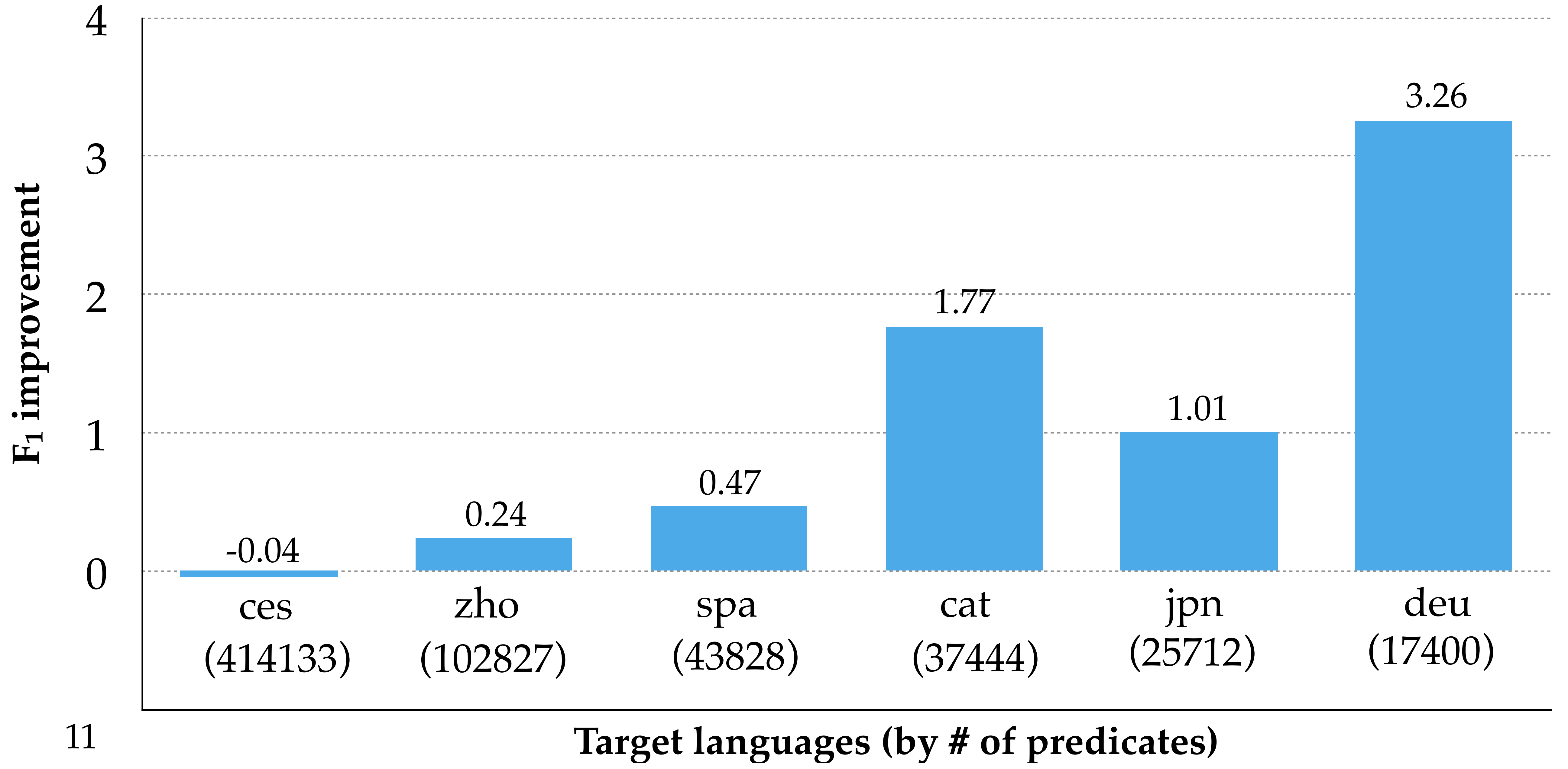


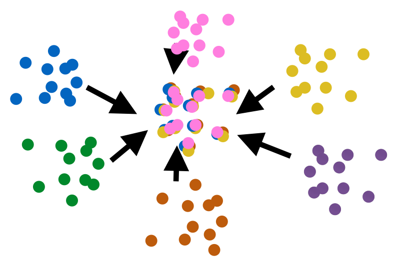
[1] He et al. (2017)

polyglot SRL: experiment and results

- for each non-English language, train
  - a monolingual model
  - a polyglot model with English
- most languages improve from polyglot training
- lower-resource languages benefit more





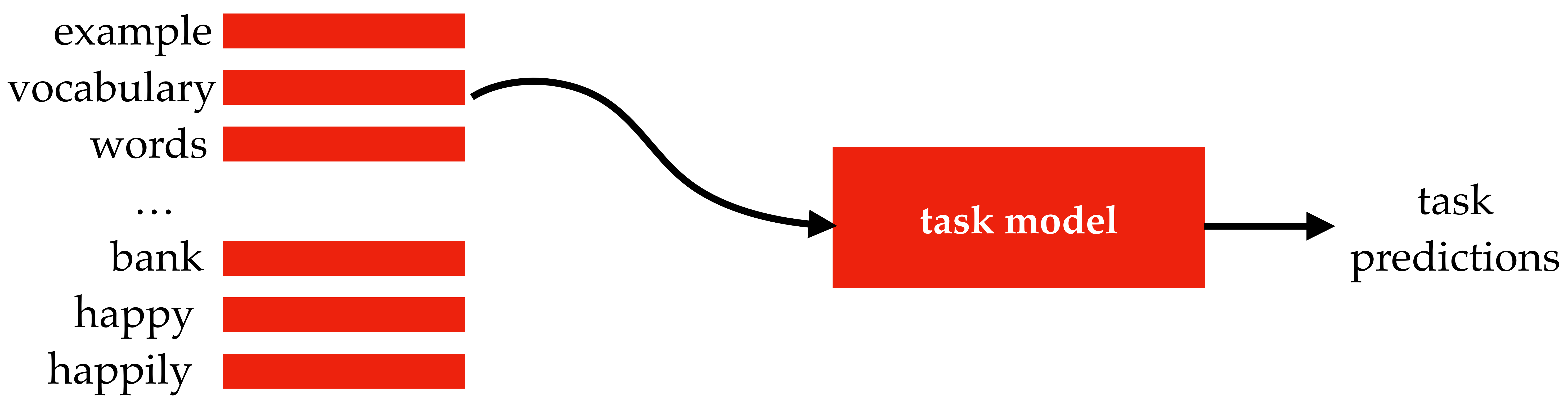


polyglot SRL: takeaways

- can represent data from multiple languages in a shared representation space
- by sharing data across languages, you can improve performance
- lower-resource languages benefit more
- different annotation schemes are not a strict barrier

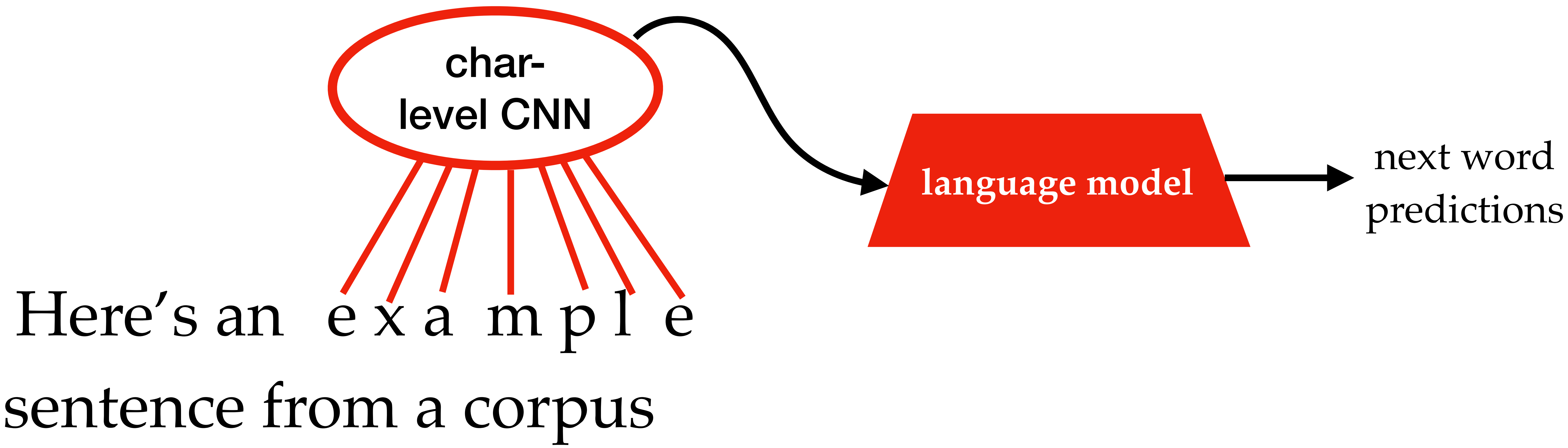
problems with word representations

- word vectors are great, but limited:
  - poorly handle polysemy
  - similar words trained independently



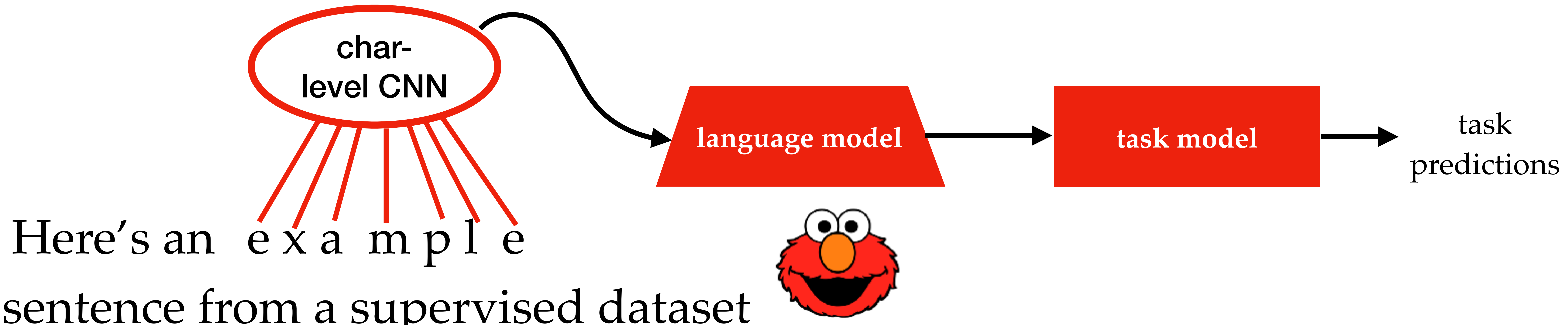
solution: contextualized word representations

- train a language model first



solution: contextualized word representations

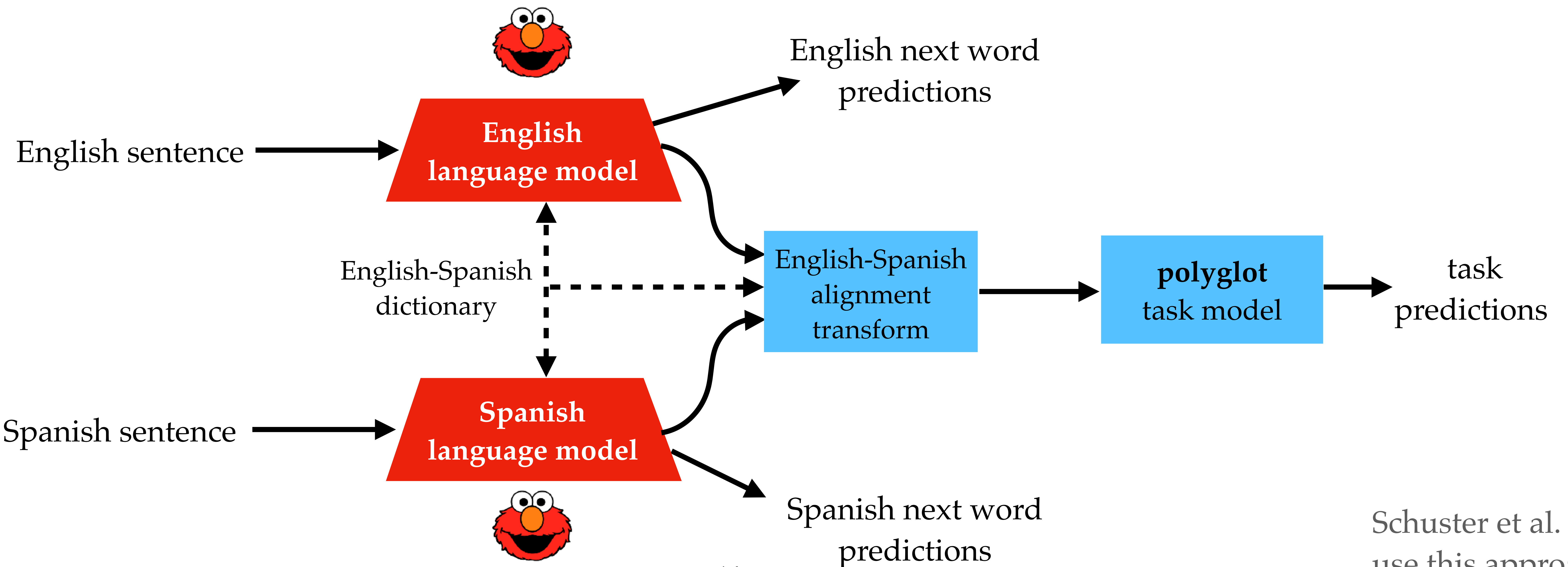
- train a language model first
- feed hidden states to the task model as input



ELMo (Peters et al. 2018); c.f. BERT (Devlin et al. 2018)

an intuitive approach: alignment of averages

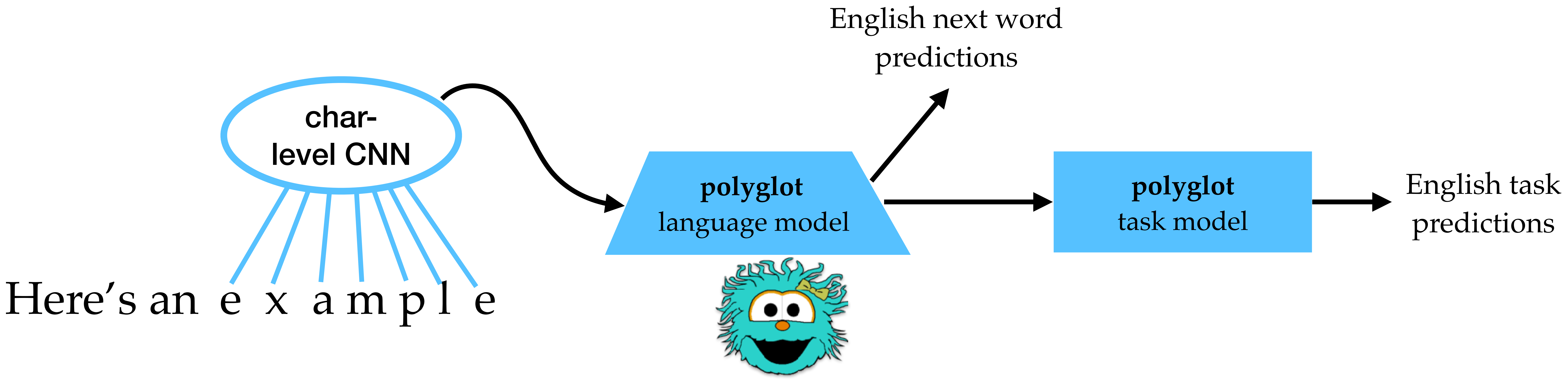
- train separate language models for each language
- align “average” embedding across contexts with a bilingual dictionary



Schuster et al. (2019) use this approach

polyglot contextualization: Rosita

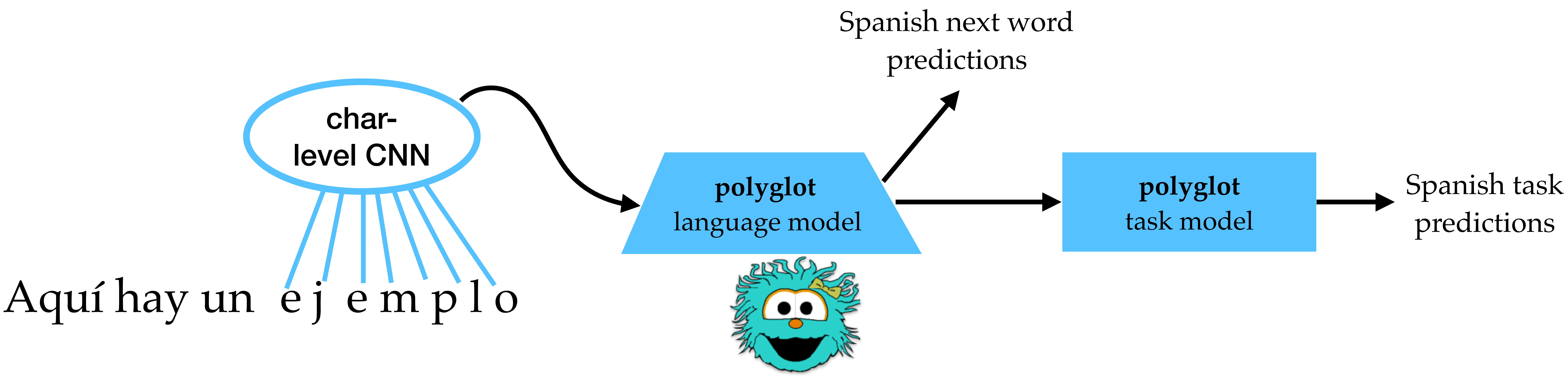
- train a language model first
- feed hidden states to the task model as input
- for a multilingual model, we need a multilingual language model!



multilingual BERT (Devlin et al. 2018) and XLM (Lampele and Conneau, 2019) also use this approach

polyglot contextualization: Rosita

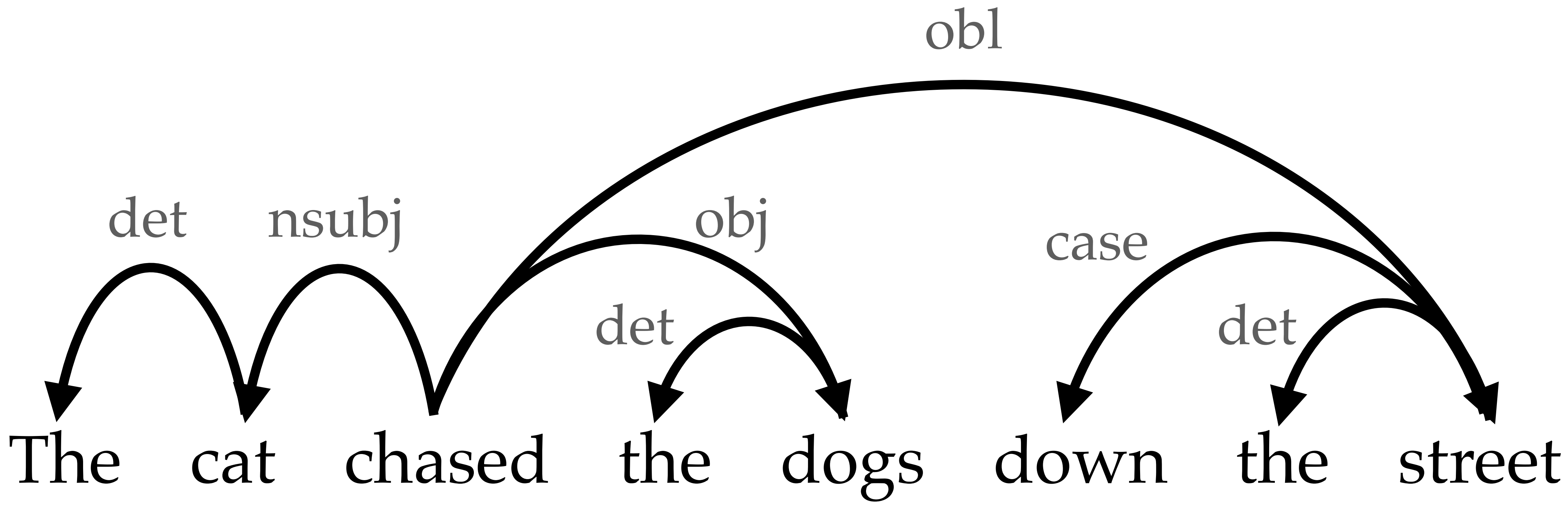
- train a language model first
- feed hidden states to the task model as input
- for a multilingual model, we need a multilingual language model!



multilingual BERT (Devlin et al. 2018) and XLM (Lample and Conneau, 2019) also use this approach

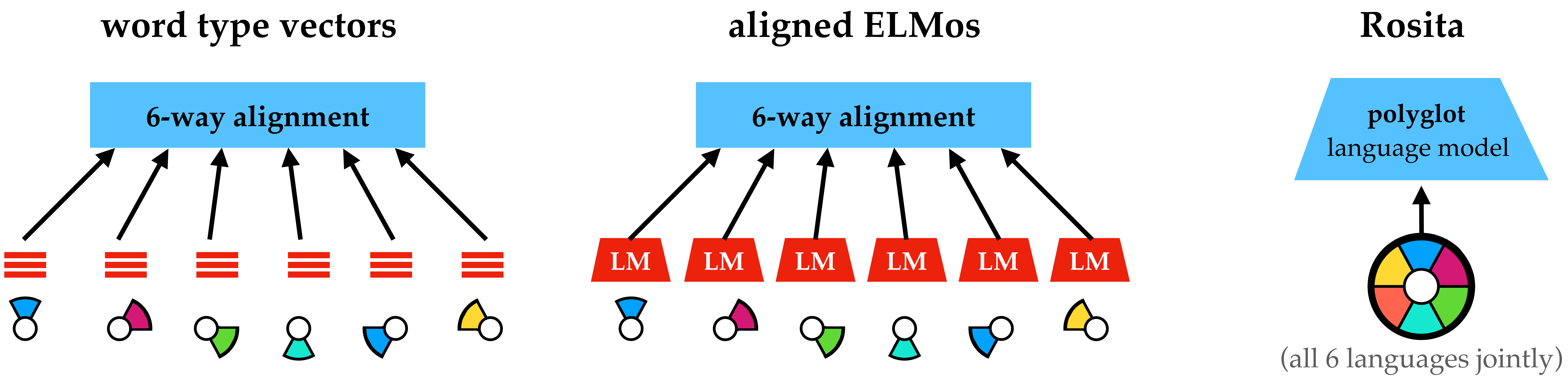
polyglot LMs: experiments

- Universal Dependencies syntax parsing (which *does* match across languages)



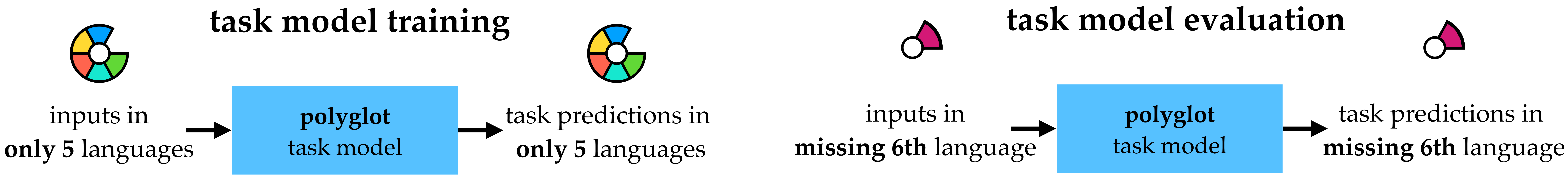
polyglot LMs: experiments

- Universal Dependencies syntax parsing (which *does* match across languages)
- “zero-target” evaluation:
  - language models (or word vectors) combining six languages

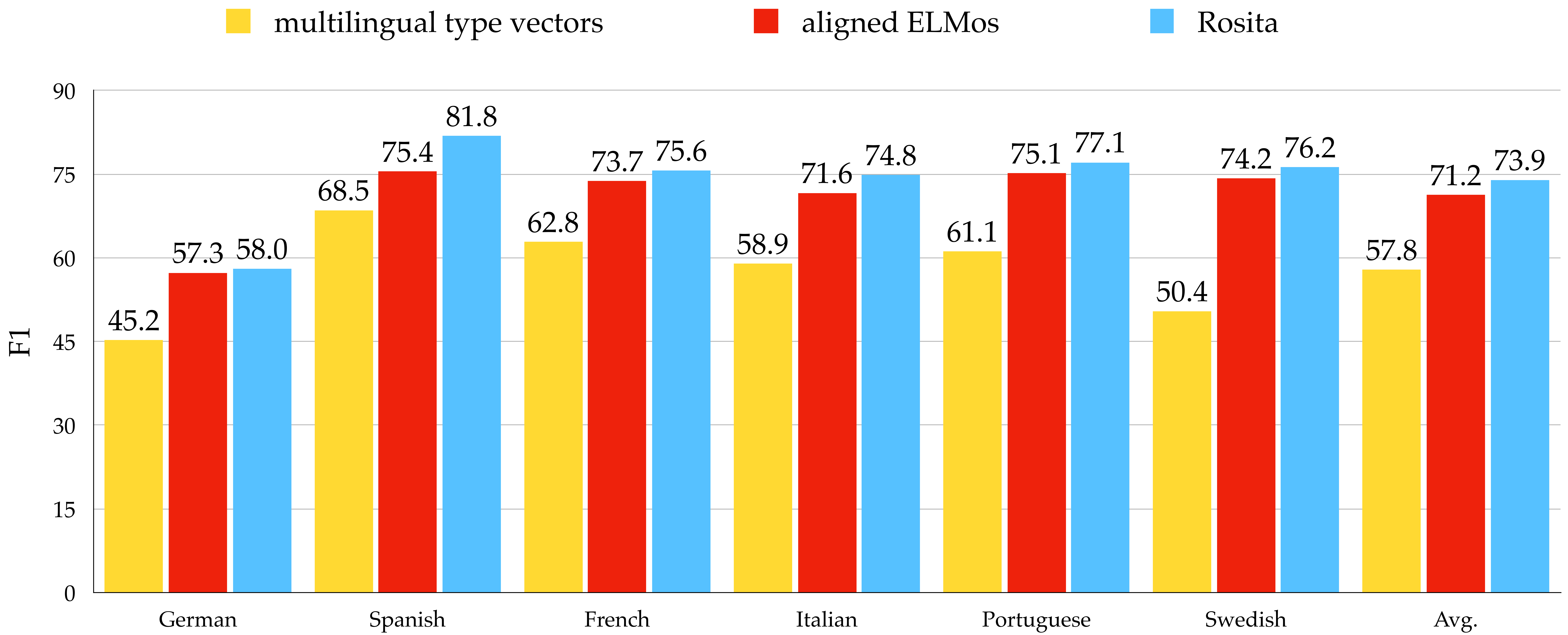


polyglot LMs: experiments

- Universal Dependencies syntax parsing (which *does* match across languages)
- “zero-target” evaluation:
  - language models (or word vectors) combining six languages
  - six parsers, each trained on only five—evaluate on the missing language

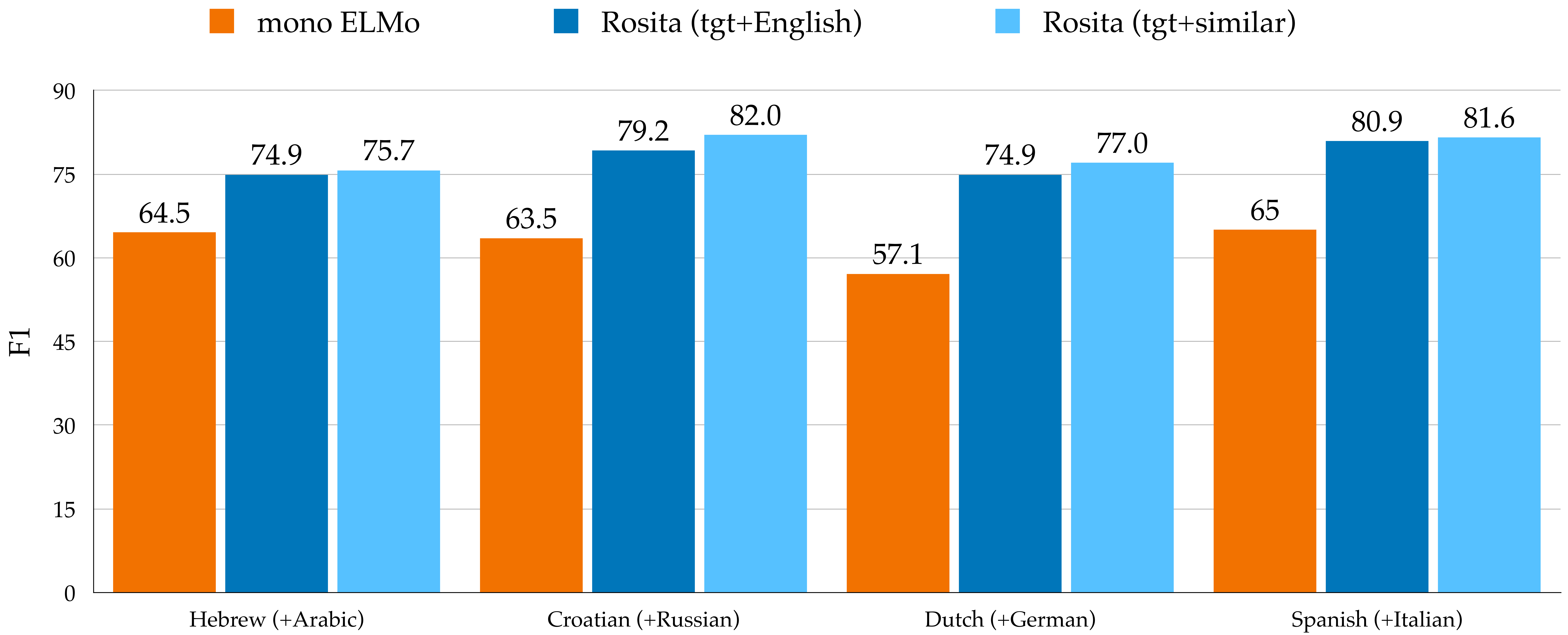


polyglot LMs: zero-target results



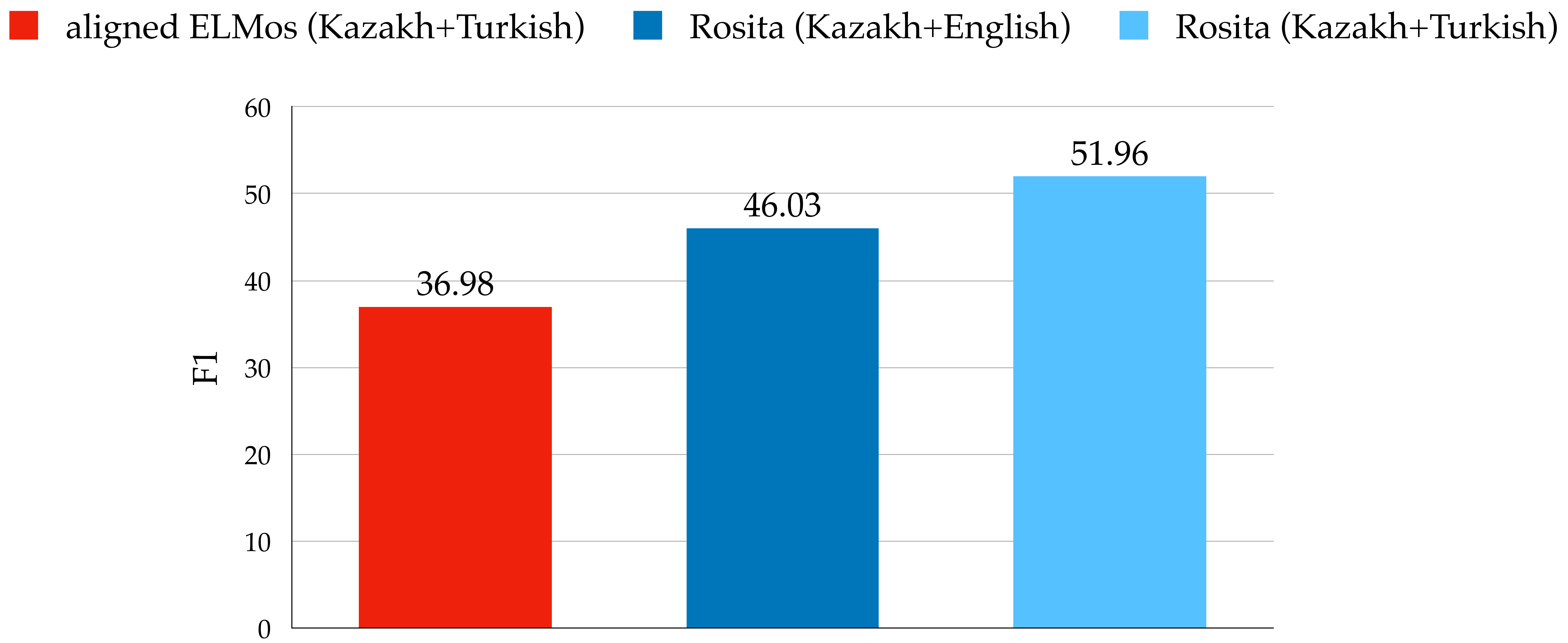
type vectors vs aligned LMs vs polyglot LMs: Universal Dependencies parsing F1

polyglot LMs: diverse languages



sharing with English vs a similar language: Universal Dependencies parsing F1

polyglot LMs: true low-resource



Kazakh, a real low-resource language: Universal Dependencies parsing F1

polyglot LMs: takeaways

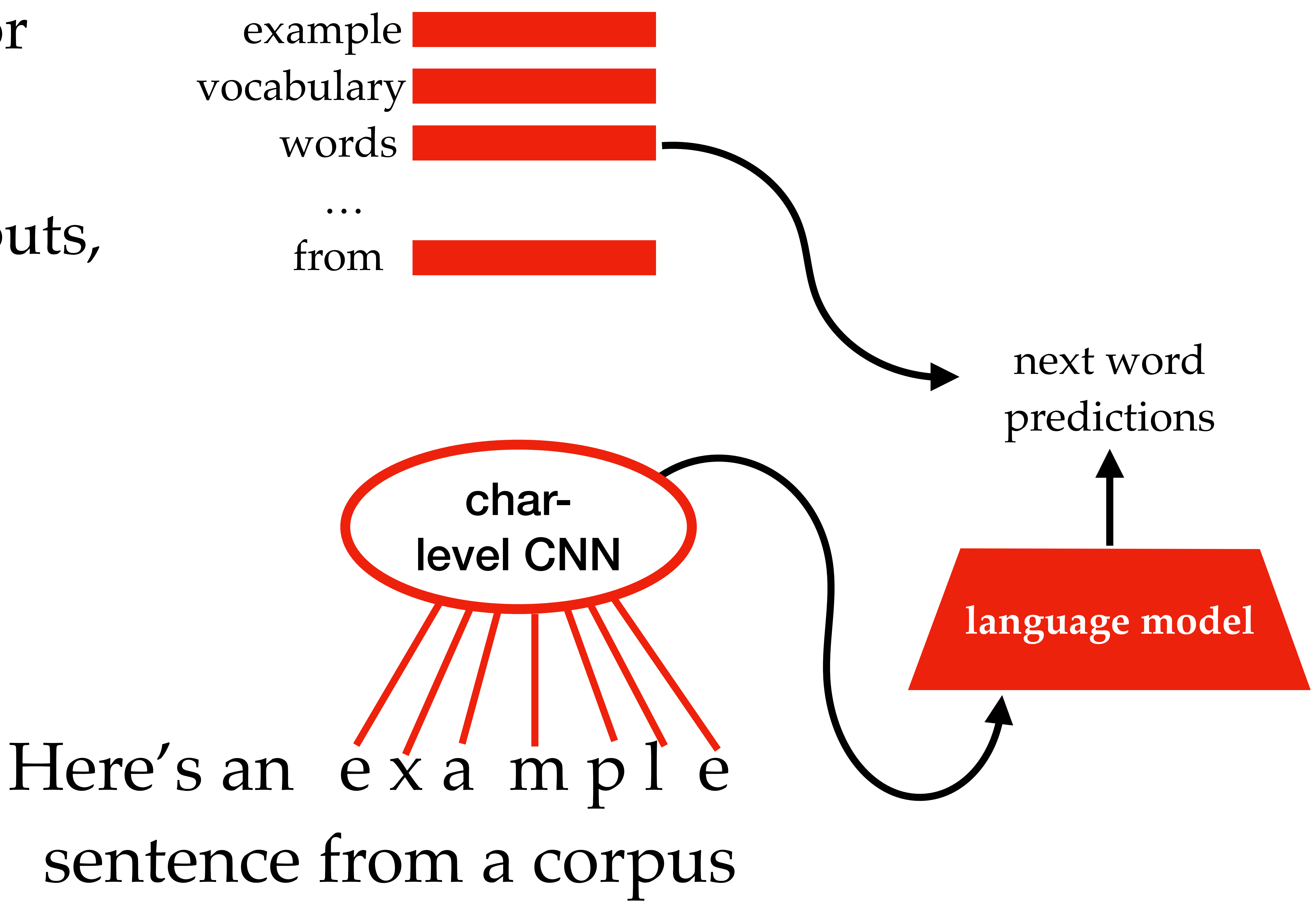
- polyglot modeling with contextualized representations works!
- don’t need *any* explicit crosslingual supervision for multilinguality!
- polyglot training captures something alignment doesn’t
- lots more experiments in the papers/Chapter 3

Mulcaire et al. 2019a: Polyglot Contextualized Representations Improve Crosslingual Transfer

Mulcaire et al. 2019b: Low-Resource Parsing With Crosslingual Contextualized Representations

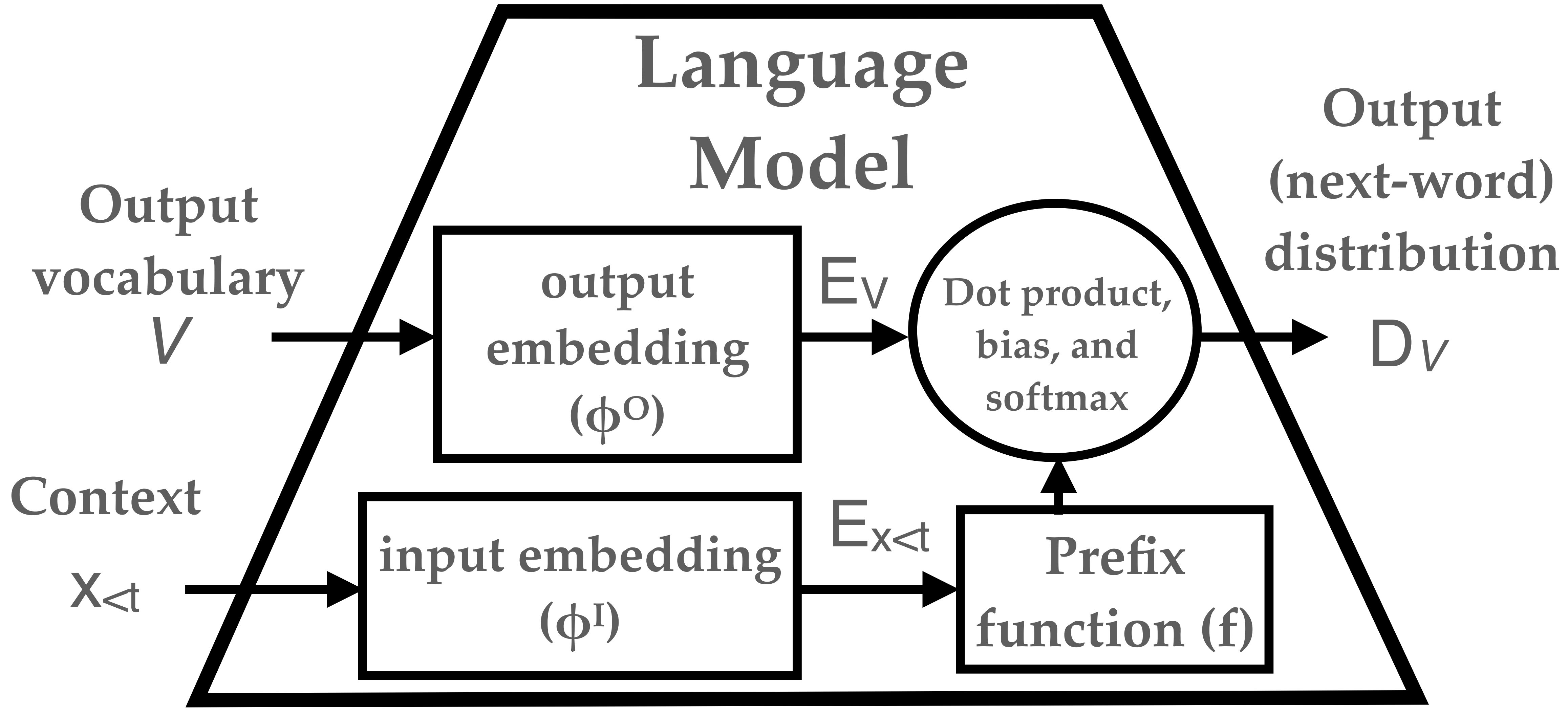
what if our language model training data is small?

- rare/out-of-domain words might get poor representations
- ELMo and Rosita have compositional inputs, but outputs are just type embeddings
- improve language models:
  - handle unknown words in test
  - improve rare word representations
  - sample-efficient learning

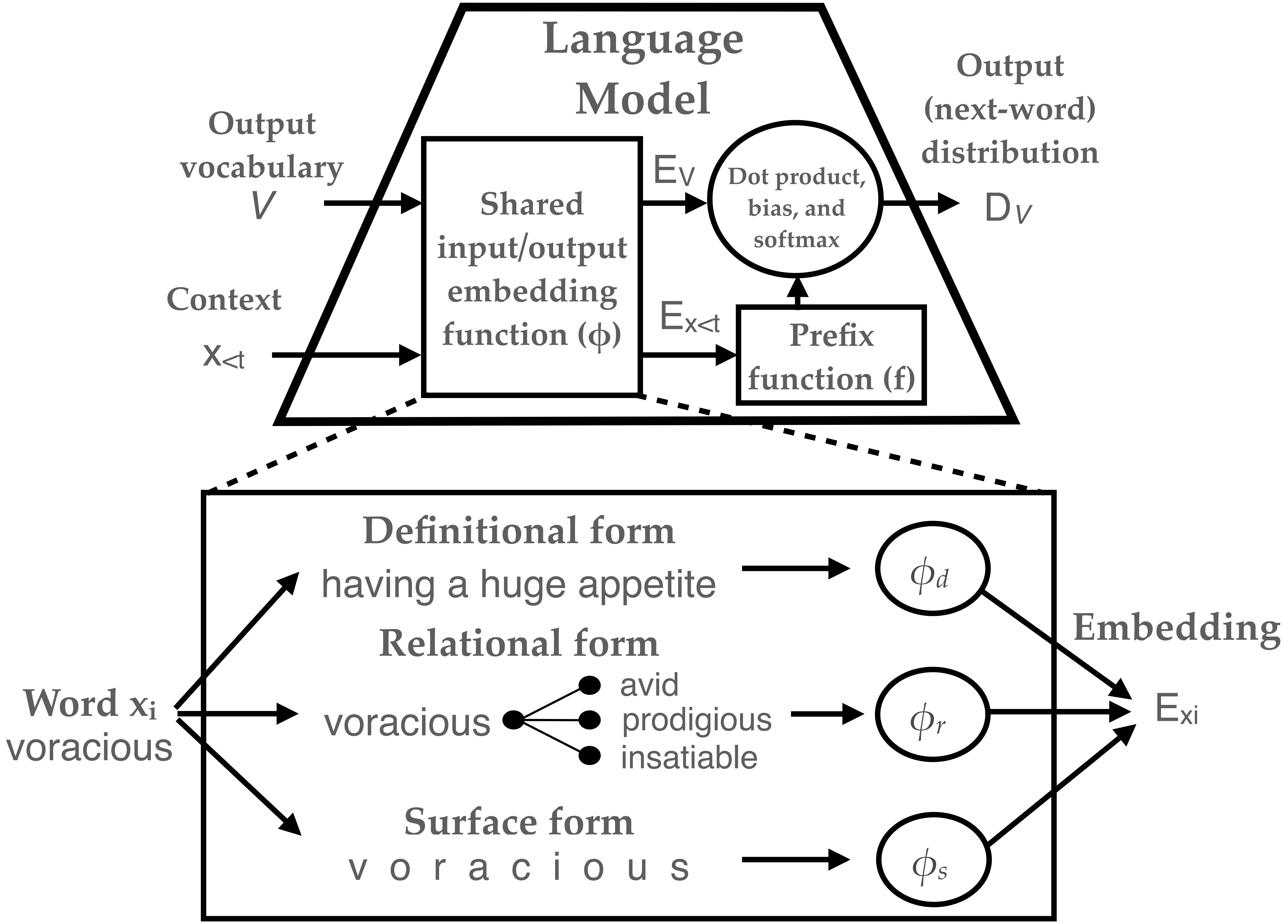


pieces of a language model

- input embedding, output embedding, prefix function
- traditional/lookup: input and output are lookup tables
- ELMo: input is a CNN, output is lookup
- many other possibilities: tied, bilinear, adaptive



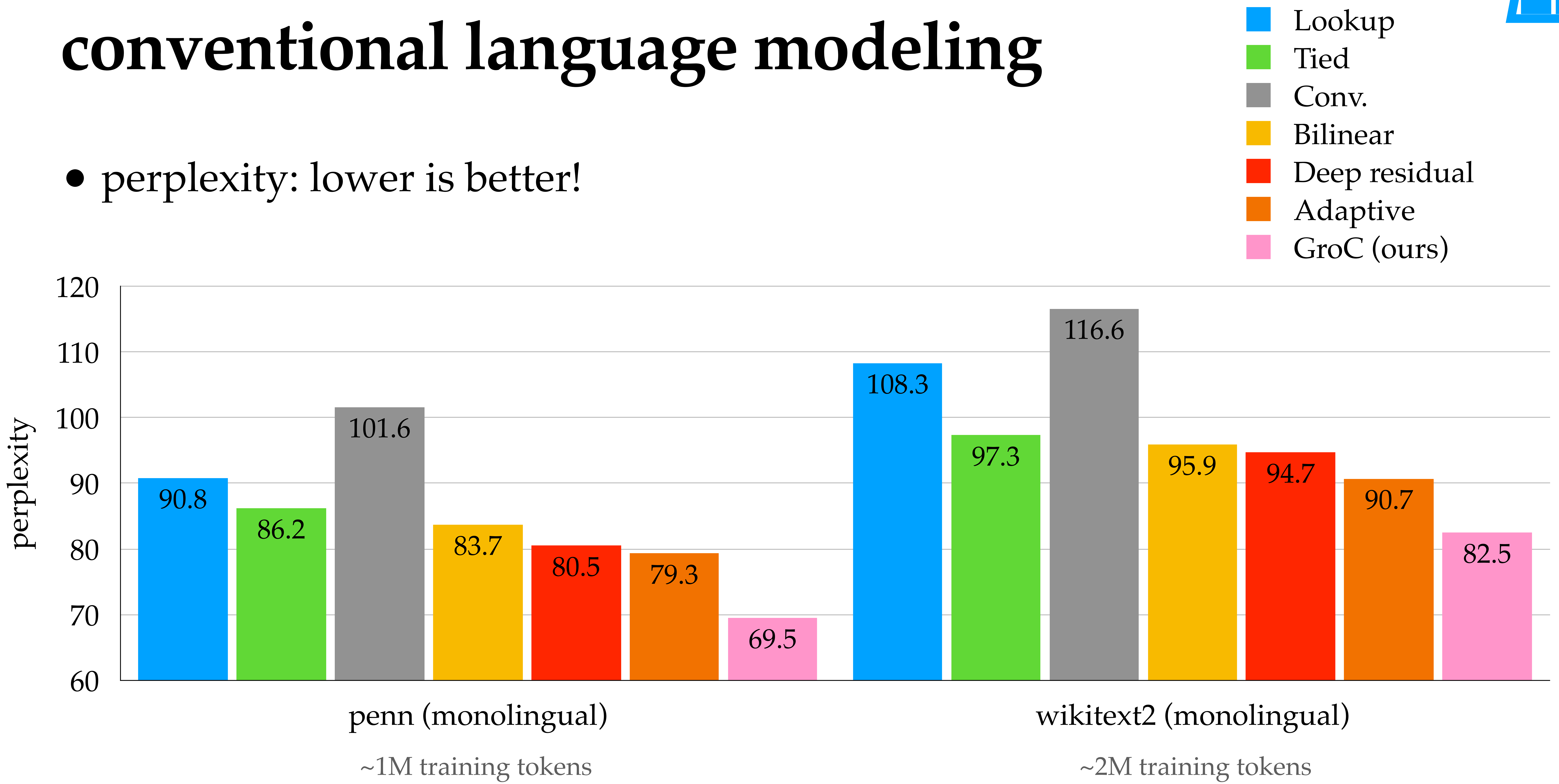
grounded compositional outputs (GroC)



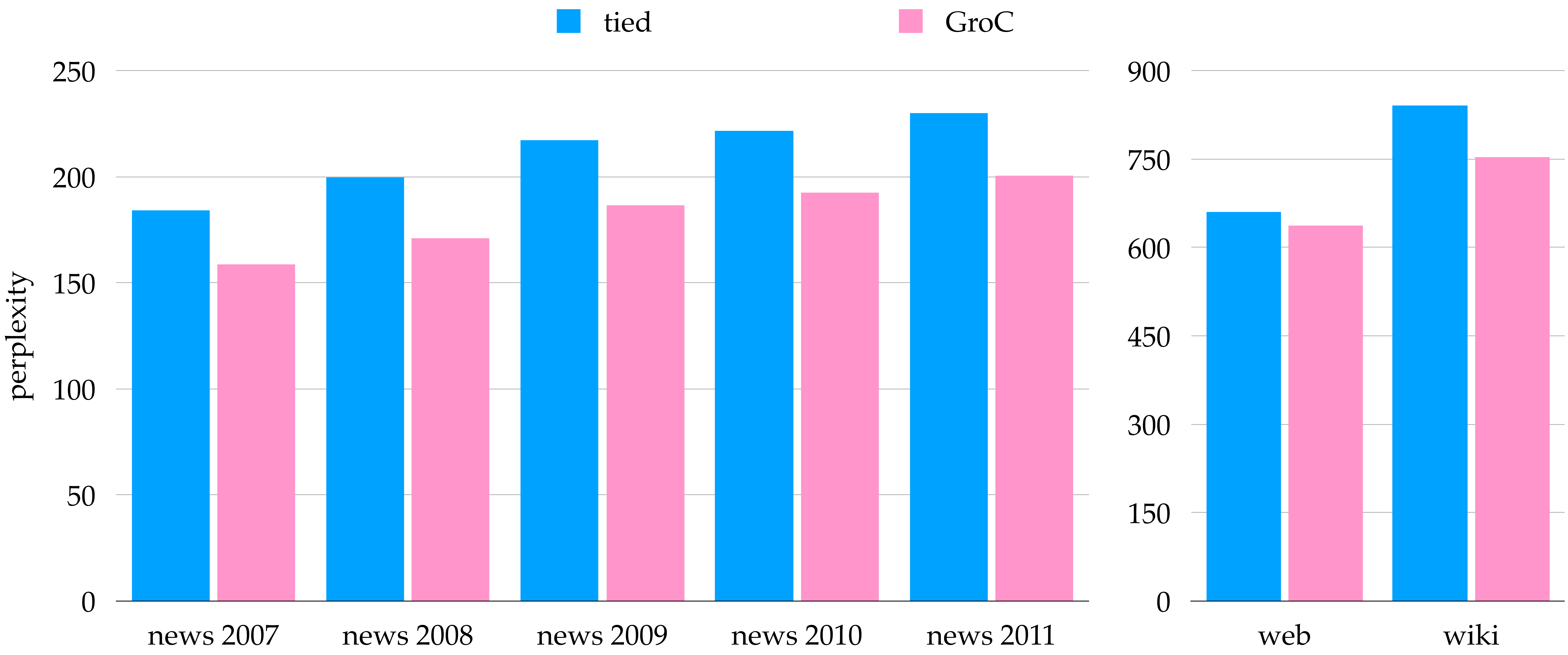
- use the same composition function for input and output
- combine surface form with relational and definitional features (from WordNet)
- (also have a residual network applied to output in some cases)

conventional language modeling

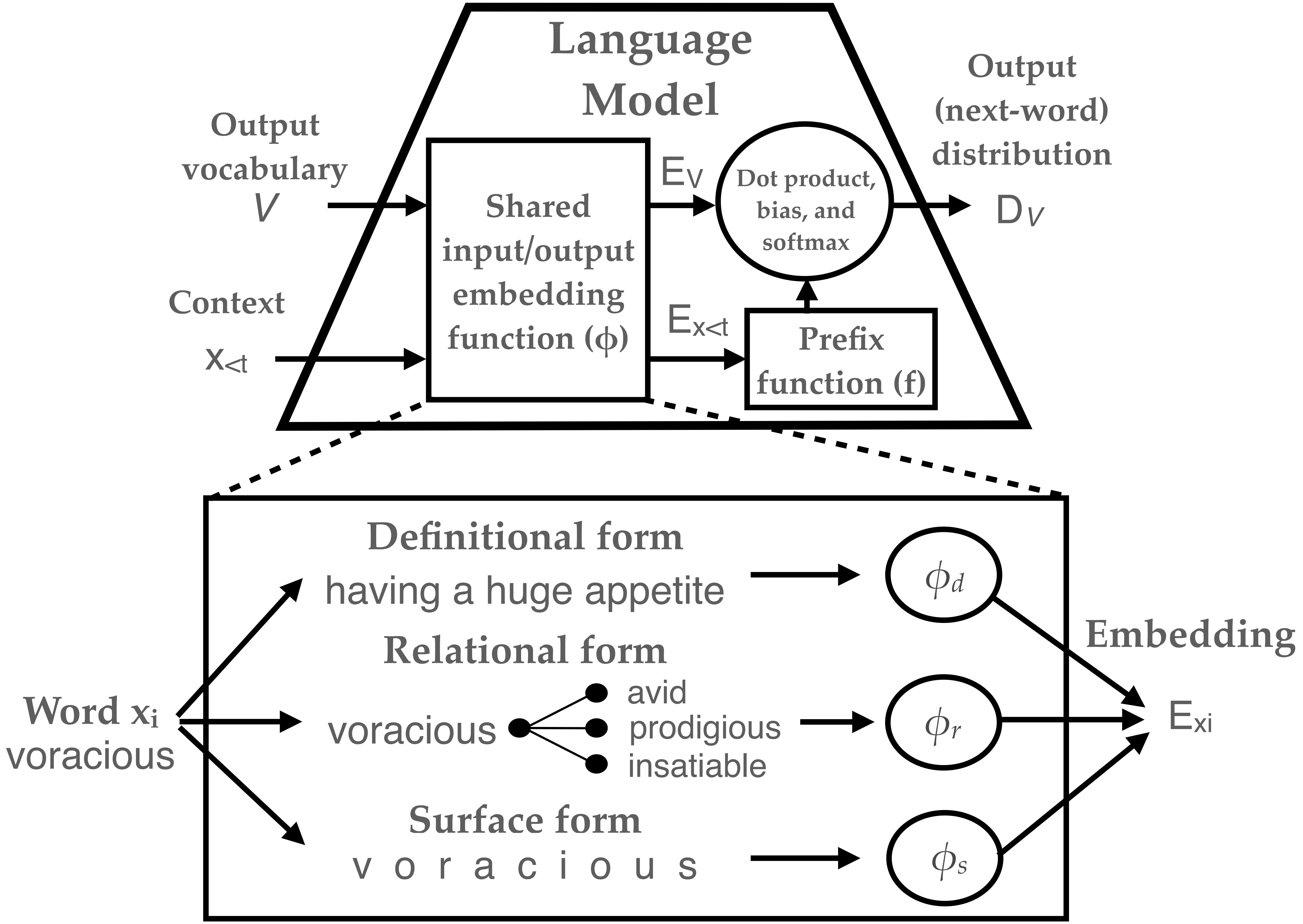
- perplexity: lower is better!



zero-resource cross-domain adaptation



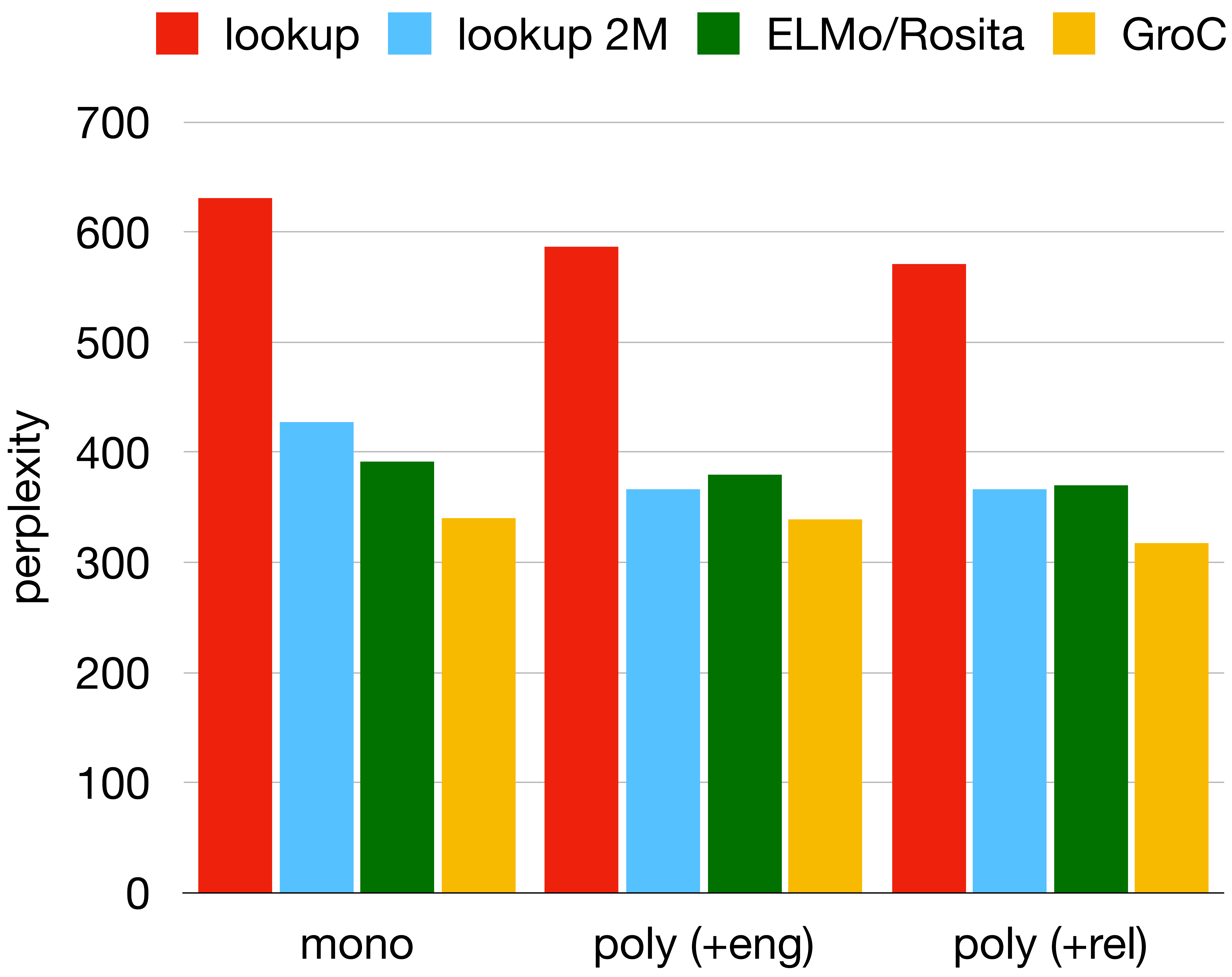
polyglot vocab-independence: multilingual GroC



- share all parameters
- use a multilingual lexicon for relational and definitional features (Open Multilingual WordNet)—this only covered some languages

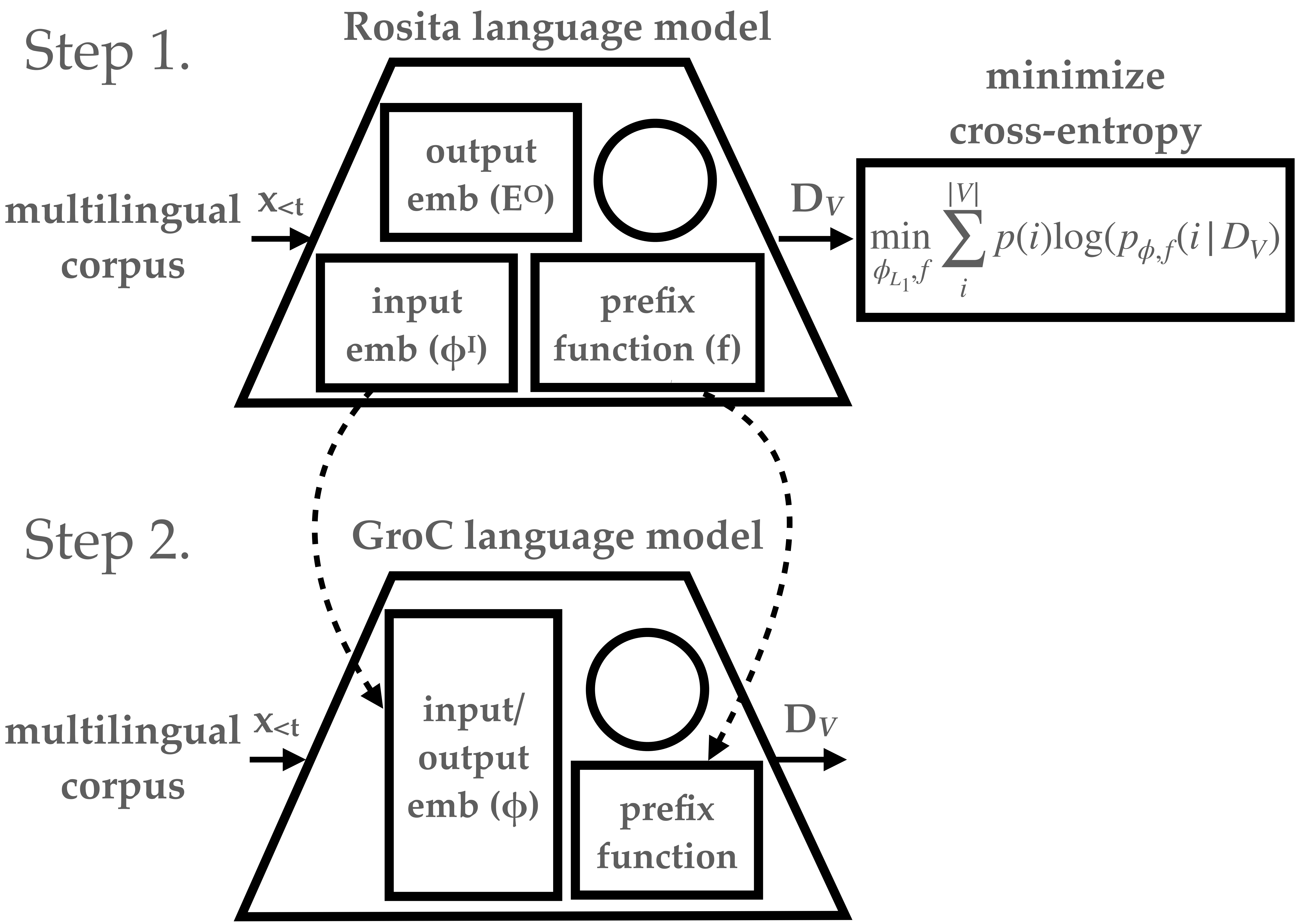
multilingual GroC: results

- lookup vs ELMo/Rosita vs GroC
- monolingual / +English / +related
- multilingual GroC is reliably the best method across 9 target languages
- related languages help more
- still outperforms lookup with 0.5x the data!



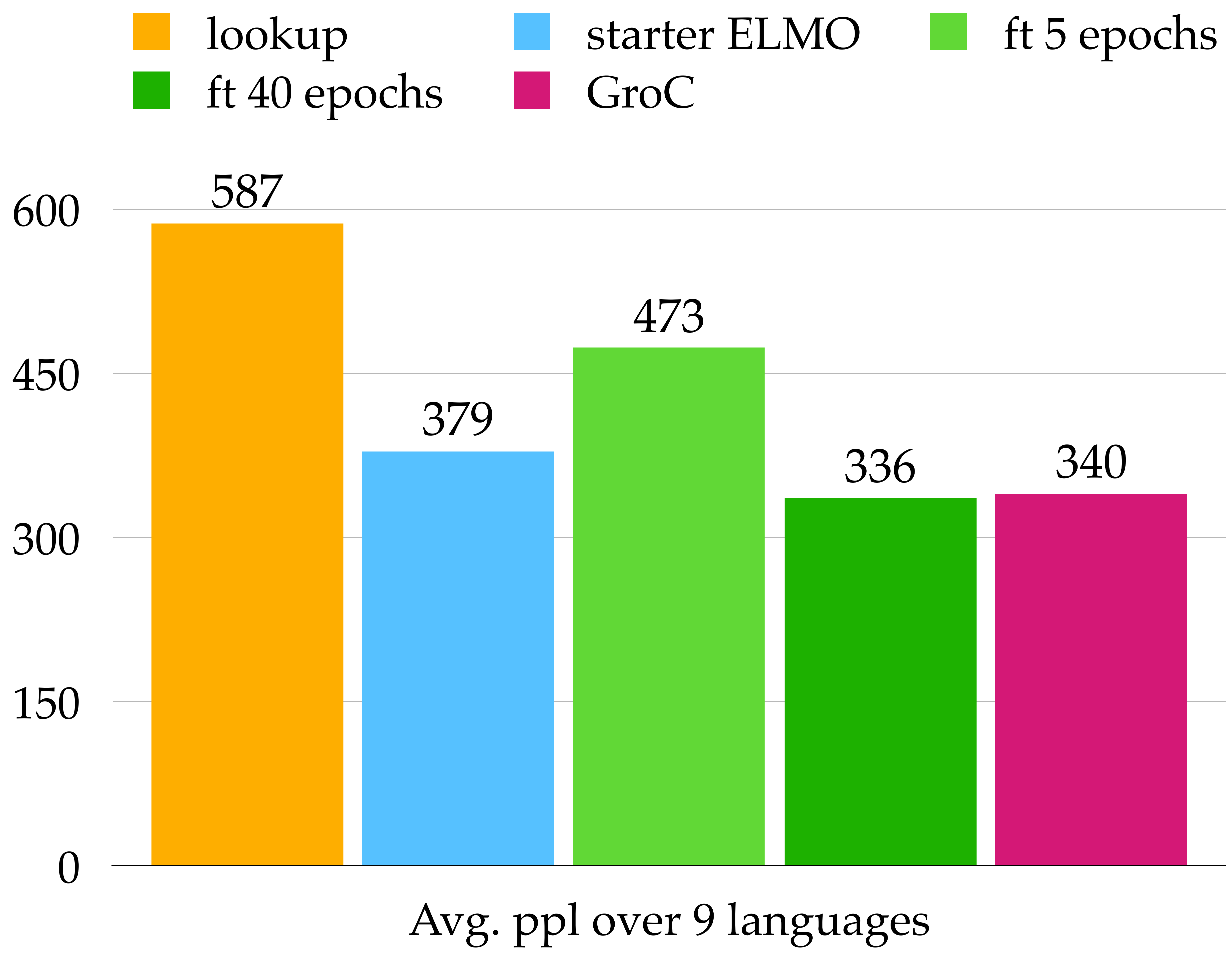
initializing compositional outputs

- ELMo/Rosita trains faster than GroC
- train an Rosita-like LM
- turn the compositional *input* embedding into a *shared* input-output embedding
- produce a GroC model cheaply—but need to finetune



initializing compositional outputs: results

- needs finetuning, but not much
- can beat GroC-from-scratch with less total training time!
- holds promise for application of GroC-like representations to large-scale language models!



conclusion

- crosslingual sharing works
- low-resource NLP is hard, but tractable—*if* we use sharing
- related languages and vocab-independence are useful

thank you!

Collaborators: