IT UNIVERSITY OF COPENHAGEN

Transferring NLP models across languages and domains

Barbara Plank
ITU, Copenhagen, Denmark

DeepLo-2019

November 3, 2019,
Hong Kong

Lots of exciting opportunities, but…

Adverse Conditions

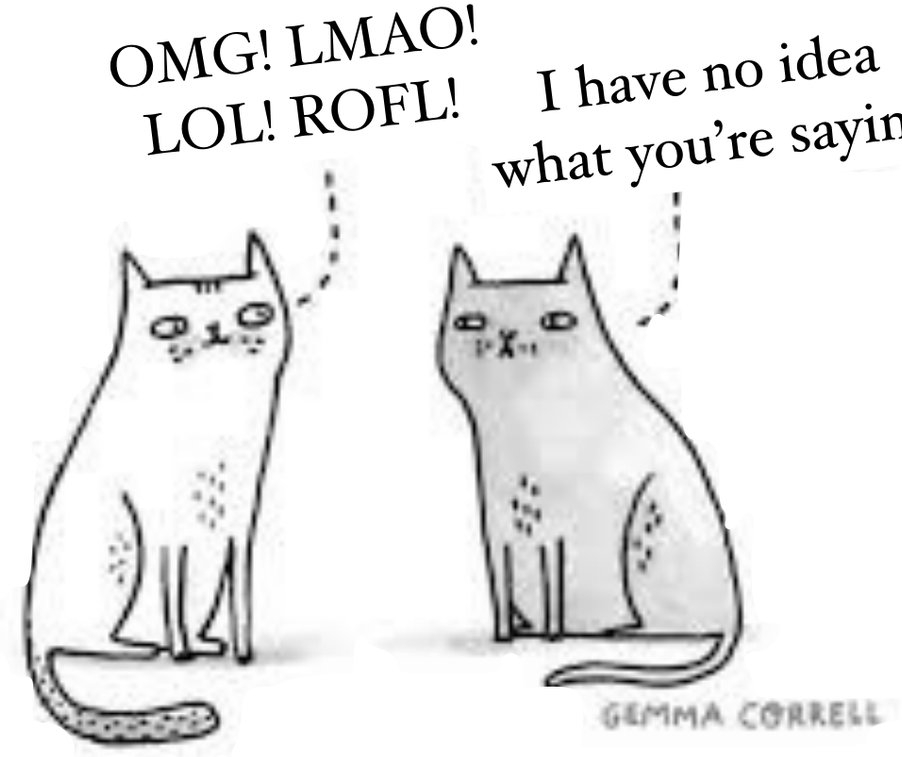
- **Data dependence:** our models dreadfully **lack** the ability to **generalize** to new conditions:

CROSS-DOMAIN

CROSS-LINGUAL

Data variability

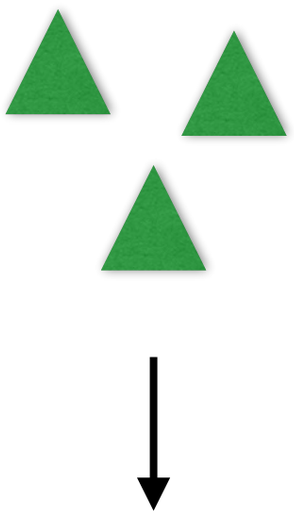
- Training and test distributions typically differ (are not i.i.d.)



- Domain changes
- Extreme case of adaptation: a new language

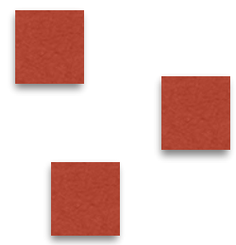
What to do about it?

Typical setup



Traditional ML:

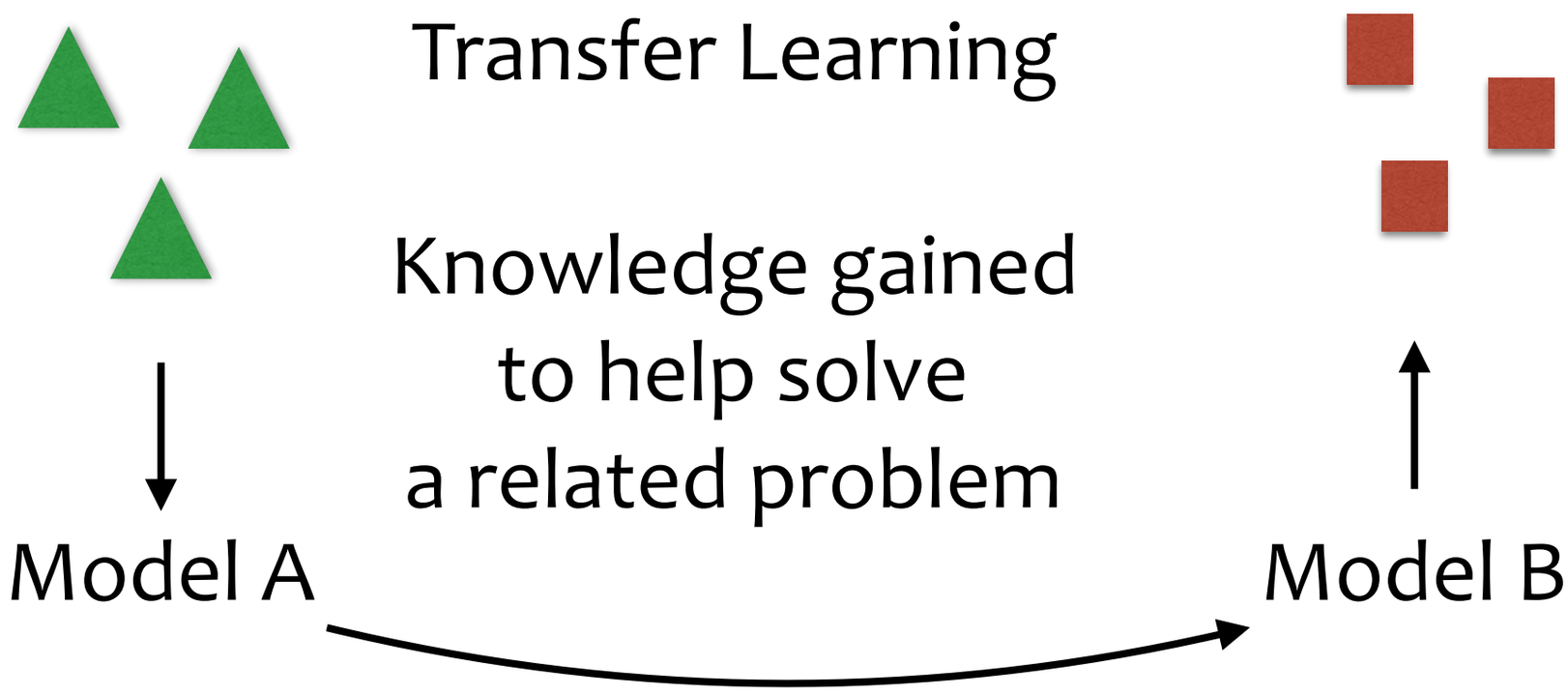
Train & evaluate
on same
domain/task/language



Model A

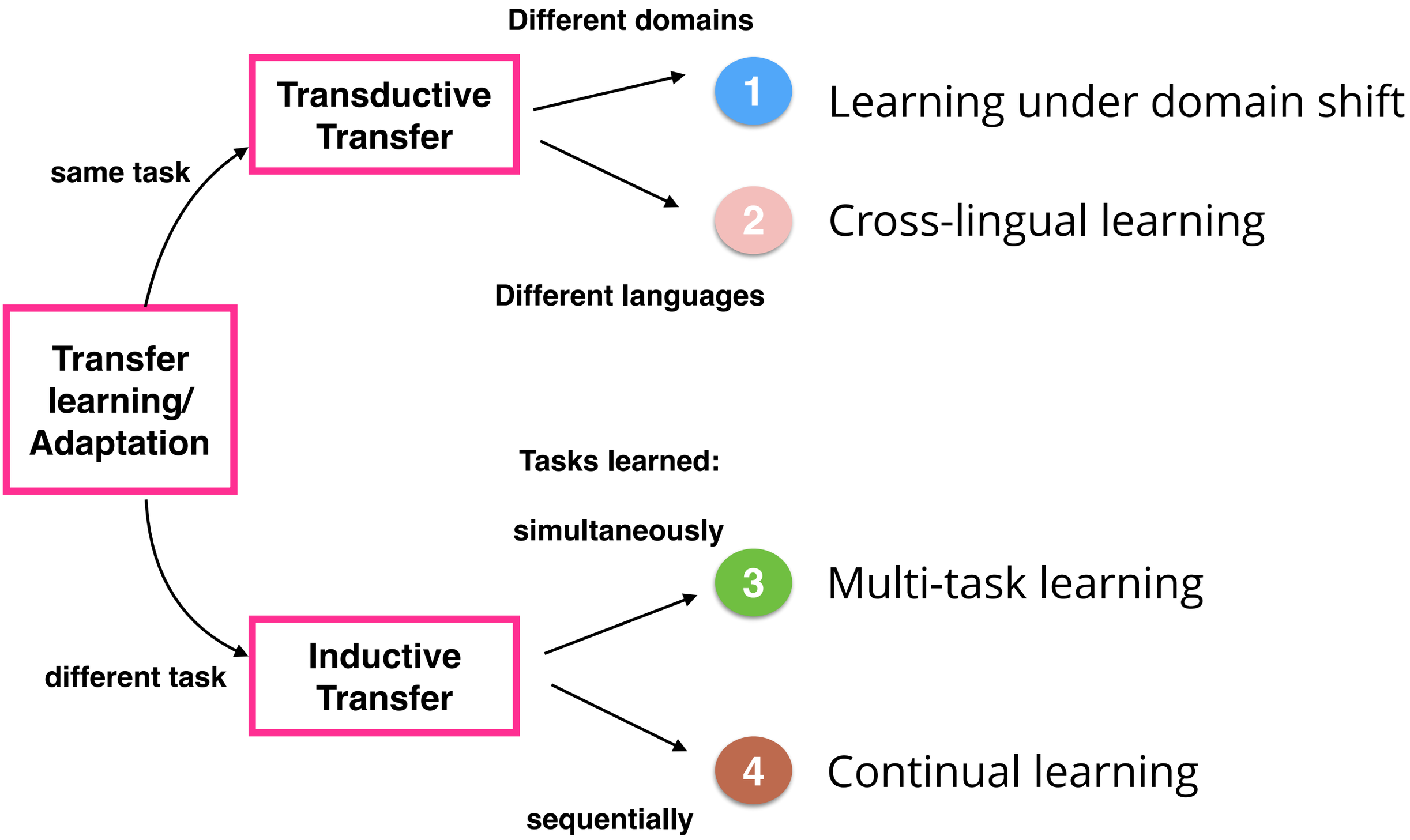
Model B

Adaptation / Transfer Learning



Adapted from Pan et al., (2009) & Ruder (2019)

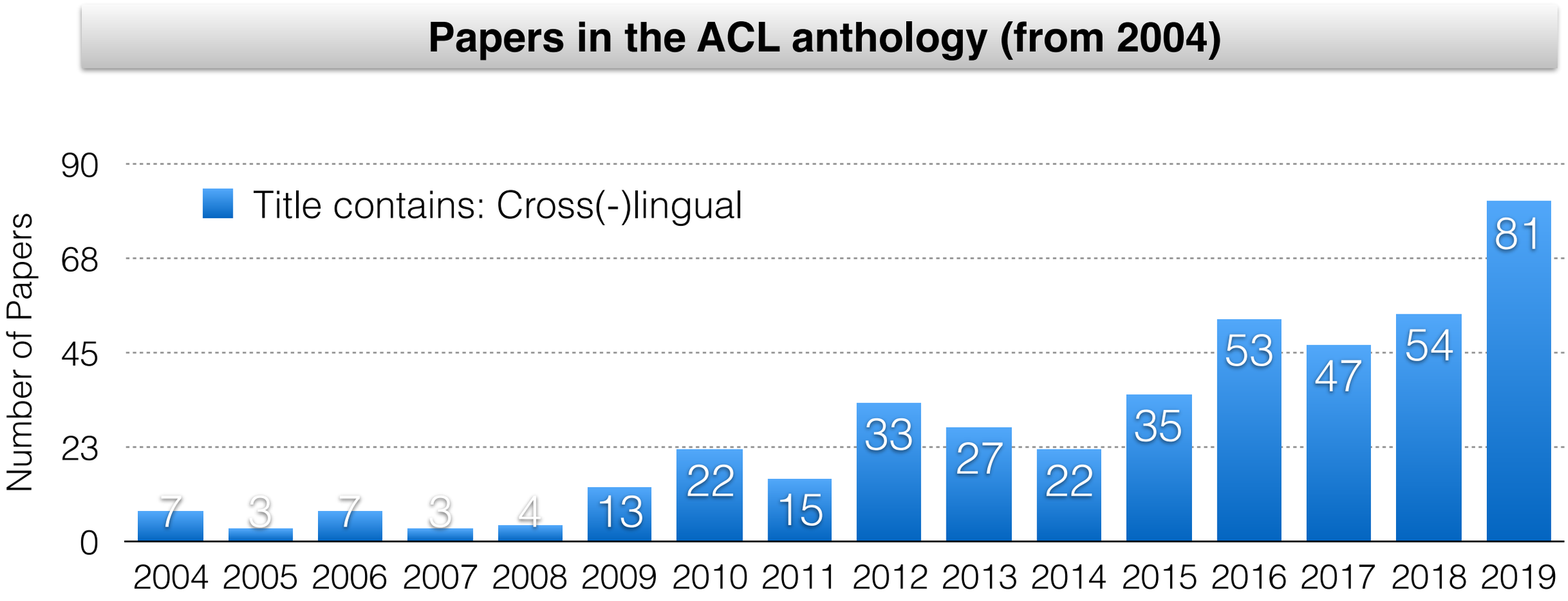
Transfer Learning



Part I:

Cross-Lingual Learning

🔥 Cross-lingual learning is on the rise 🔥



- Includes many advances on **cross-lingual representations**, e.g. see ACL 2019 tutorial (Ruder et al., 2019)

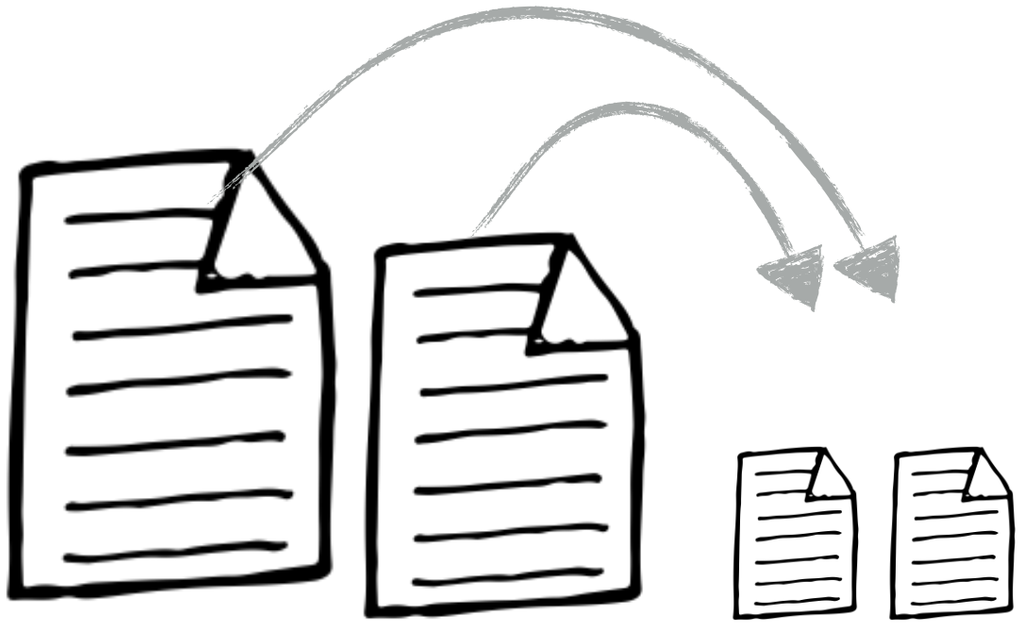
Motivation

We want to process **all** languages.

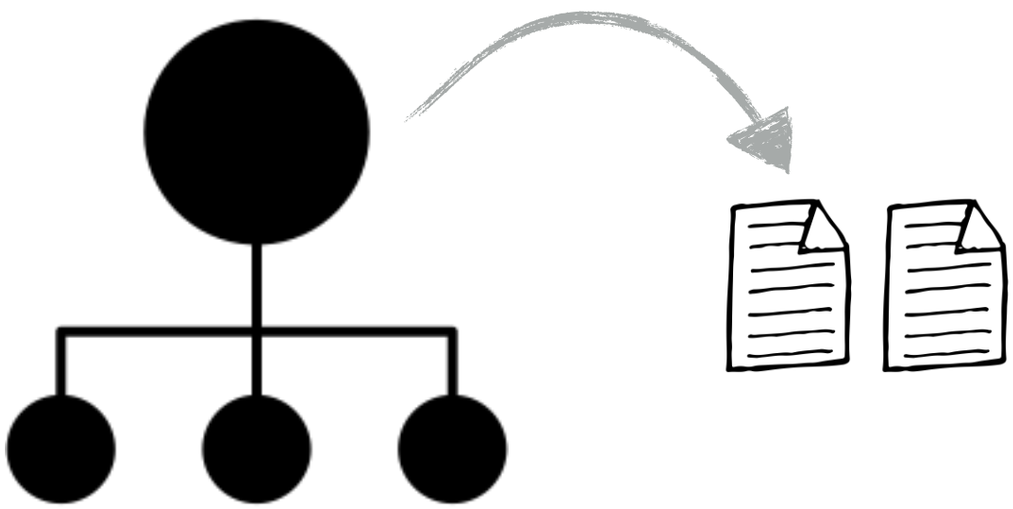
Most of them are severely under-resourced.

How to build taggers, parsers, etc. for those?

Approaches



annotation transfer
(annotation projection)
via parallel data



model transfer
(multi-lingual embeddings,
zero-shot/few-shot learning,
delexicalization,...)

Annotation projection

PRON VERB PRON ADV P

Was machst du heute ?

word alignments

project annotations

Che fesa ncuei ?

PRON VERB ADV P

e.g., Hwa et al. (2005)

1

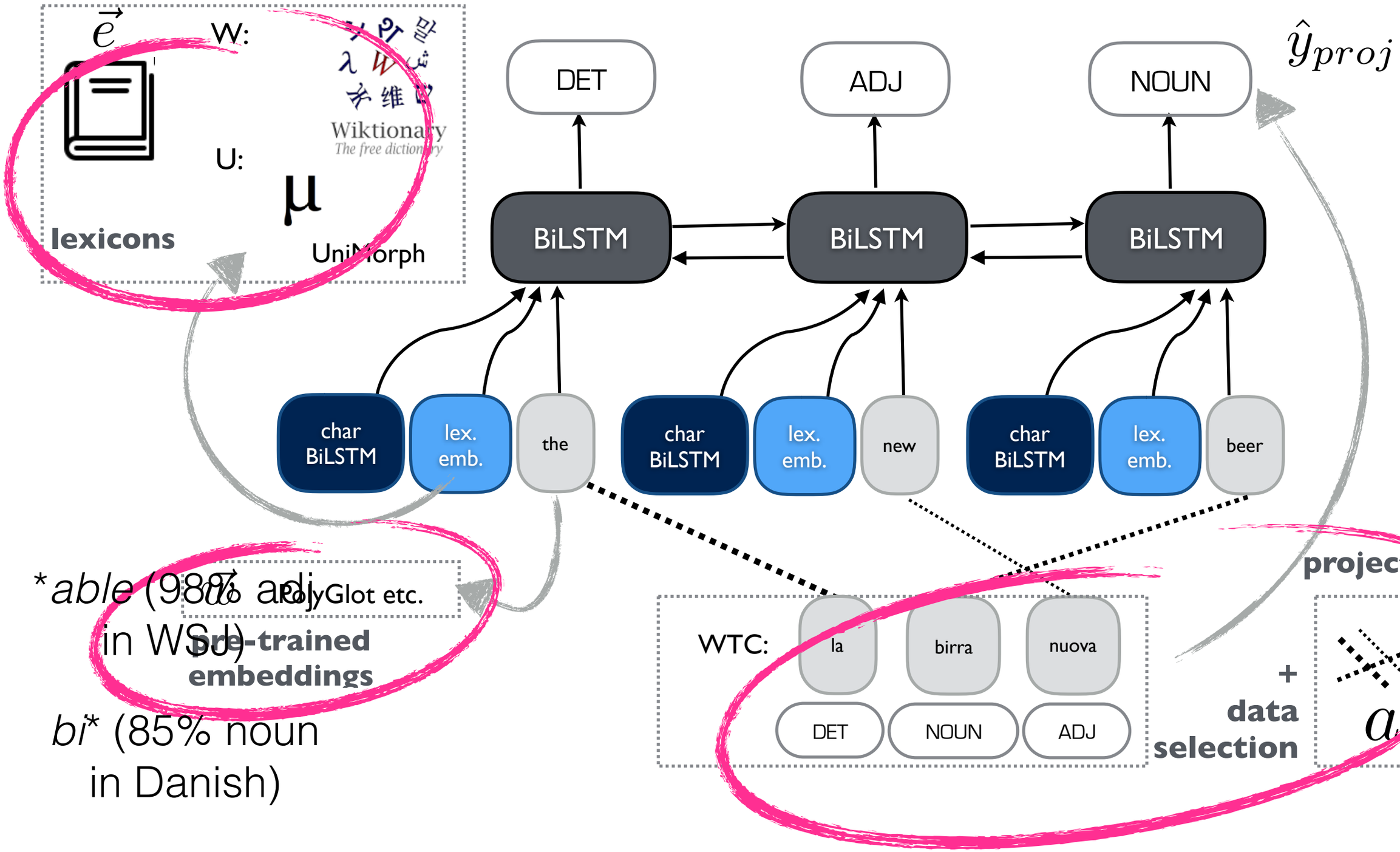
A case study: Lexical Resources for Low-Resource POS tagging in Neural Times

NoDaLiDa 2019 & EMNLP 2018
Plank & Klerke, 2019; Plank & Agic, 2018

More and more evidence is appearing that integrating **symbolic** lexical knowledge into neural models aids learning

Question: Does neural POS tagging benefit from lexical information?

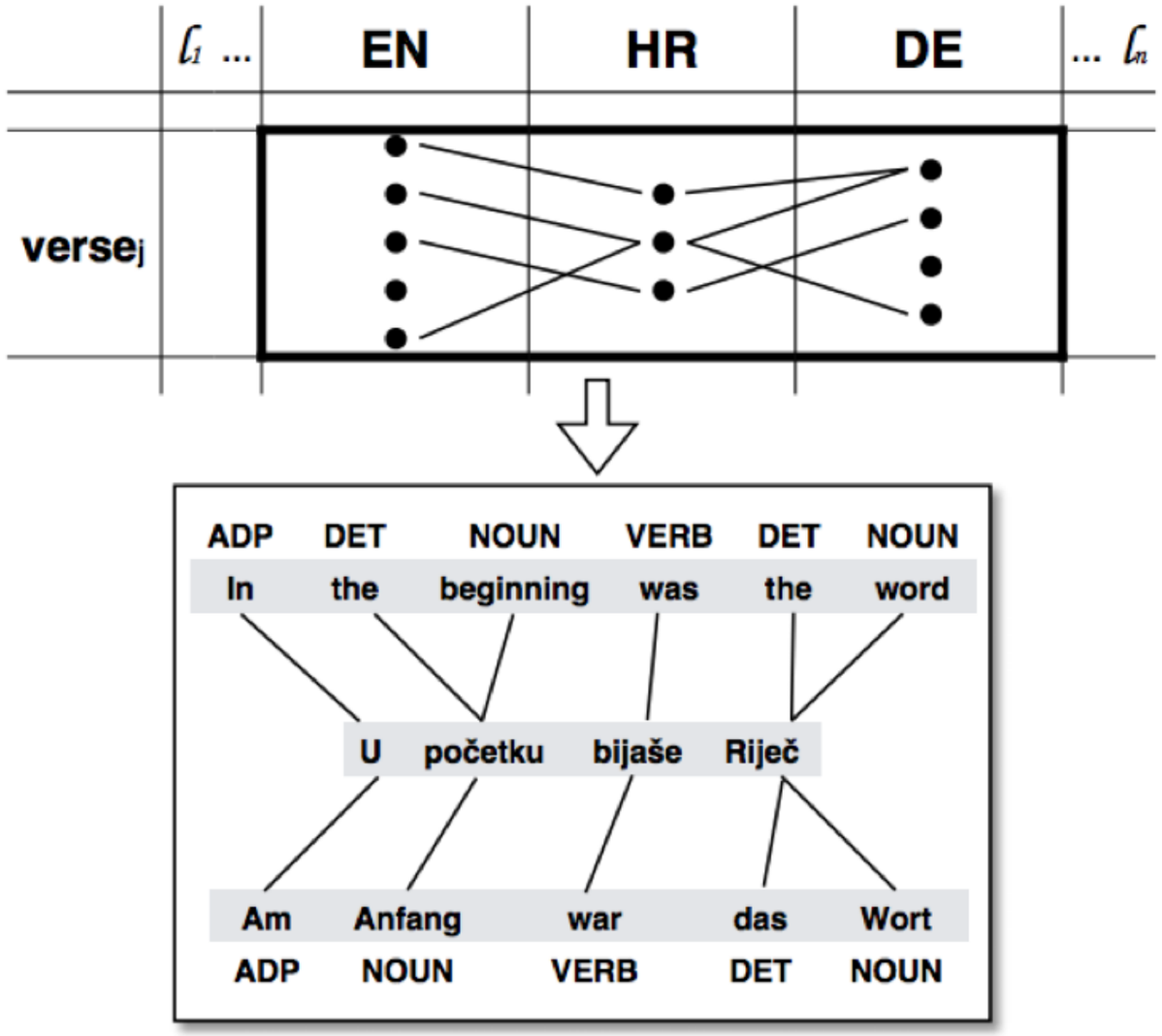
Distant Supervision from Disparate Sources (DsDs)



able (98% adj in WSJ)

*bi** (85% noun in Danish)

Multi-source Annotation Projection



HR	EN	DE	...	voted	confidence
U	ADP	ADP	...	**ADP**	0.8667
početku	NOUN, DET	NOUN	...	**NOUN**	0.7448
bijaše	VERB	VERB	...	**VERB**	0.8560
Riječ	DET, NOUN	DET, NOUN	...	**NOUN**	0.6307

(Agić et al., 2015; 2016)

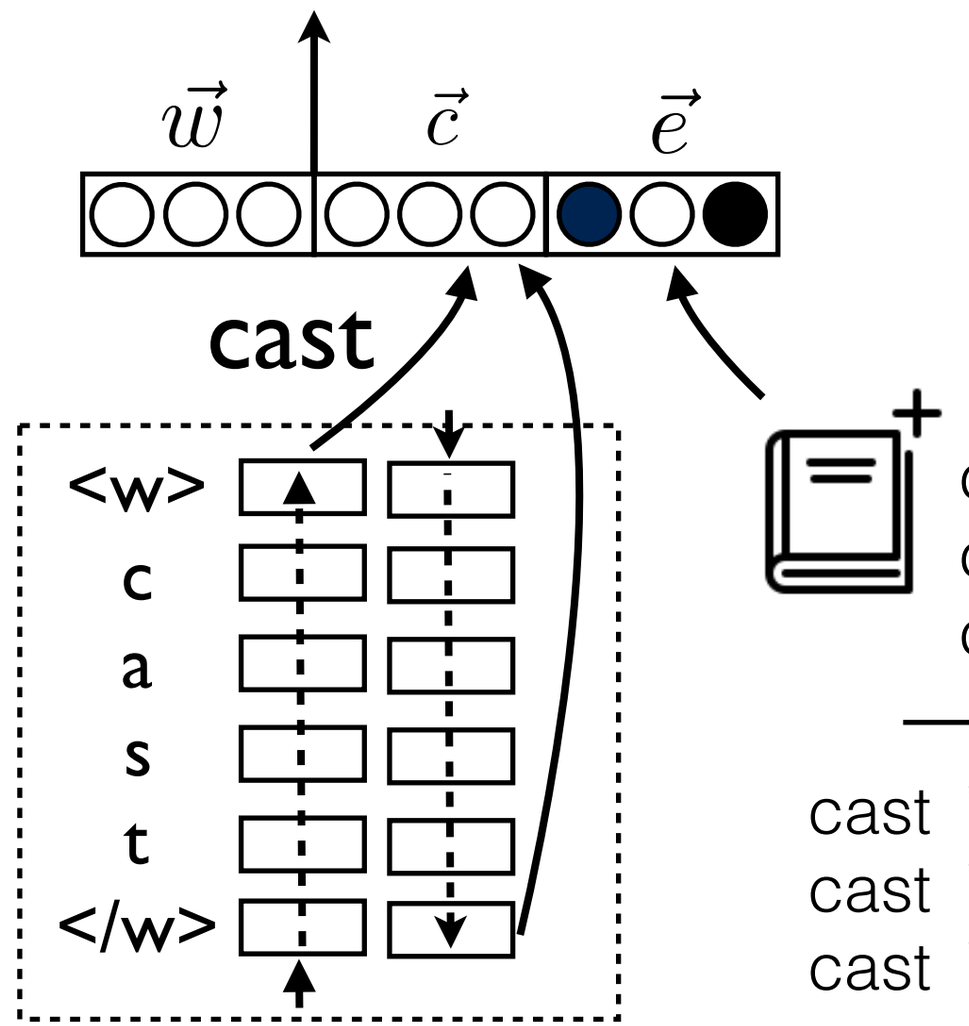
- *Watchtower corpus* (WTC), 300+ languages
- **Project** from 21 source languages (Agić et al., 2016)
- **Select** instances by word-alignment ***coverage***

Integrating lexical information

- ***n*-hot** encoding
  (Benoit & Martinez Alonso, 2017)

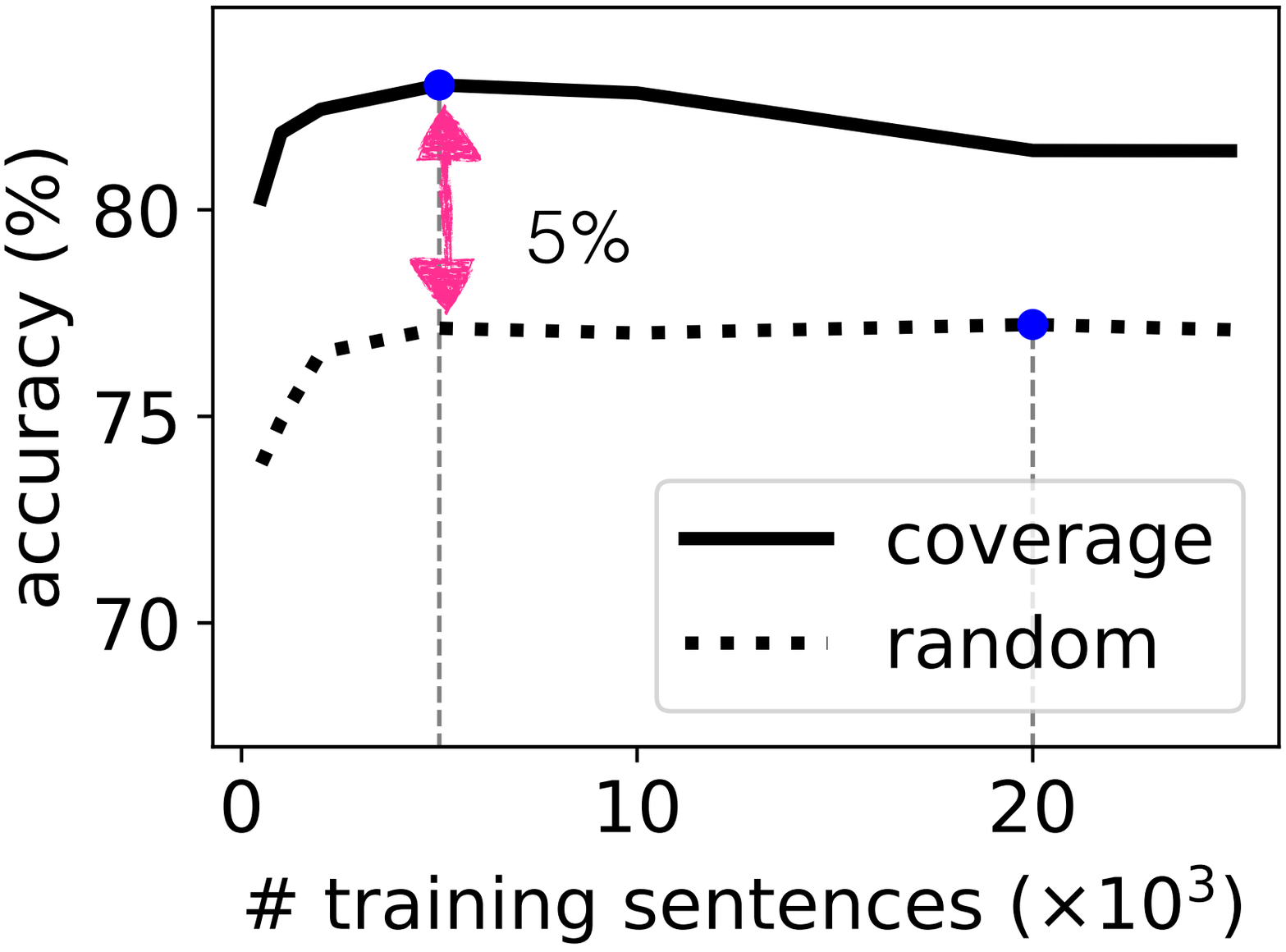
- Our approach:
  embed the lexicon

- **Sources**:
  Wiktionary
  and Unimorph



Results

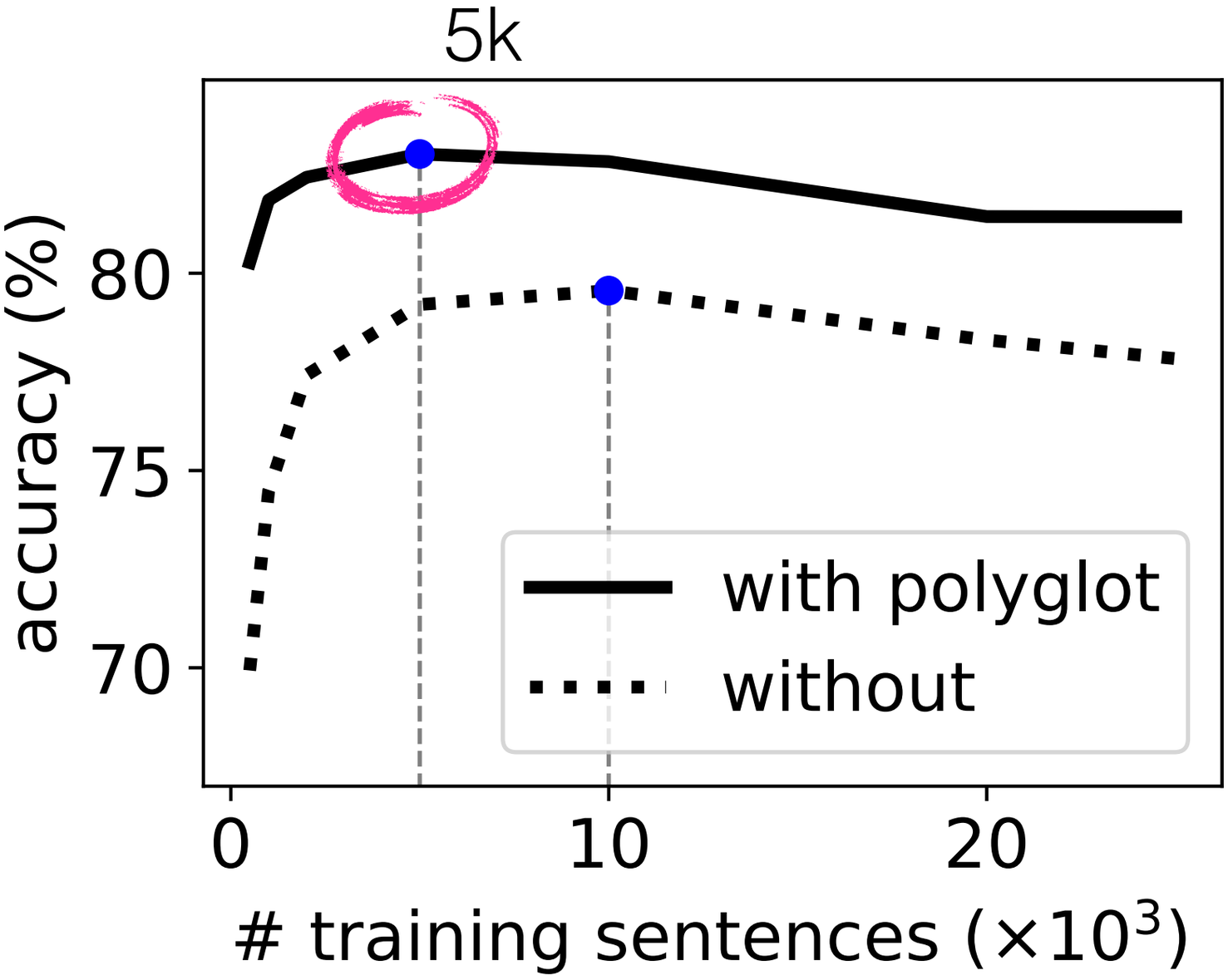
Less data is better than adding more (noise)



Means over 21 languages

(each point is an average over 3 runs, for random: with 5 random samples)

Embedding pre-initialization helps

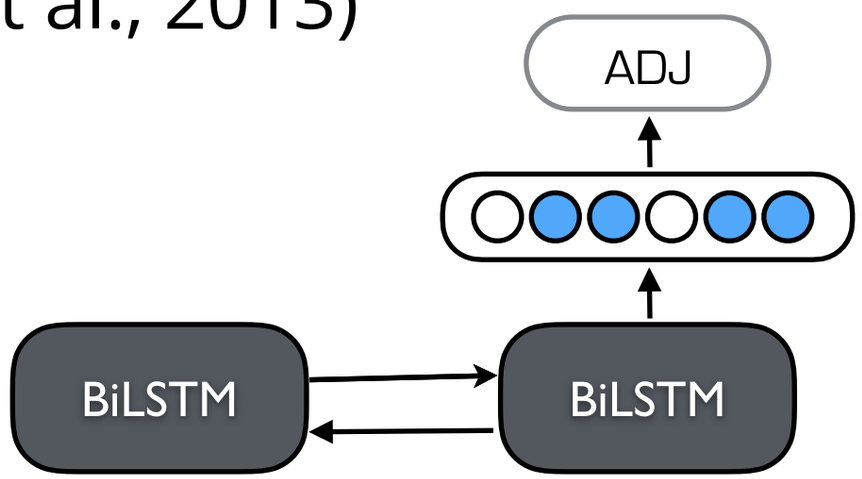


Means over 21 languages

(each point is an average over 3 runs, for random: with 5 random samples)

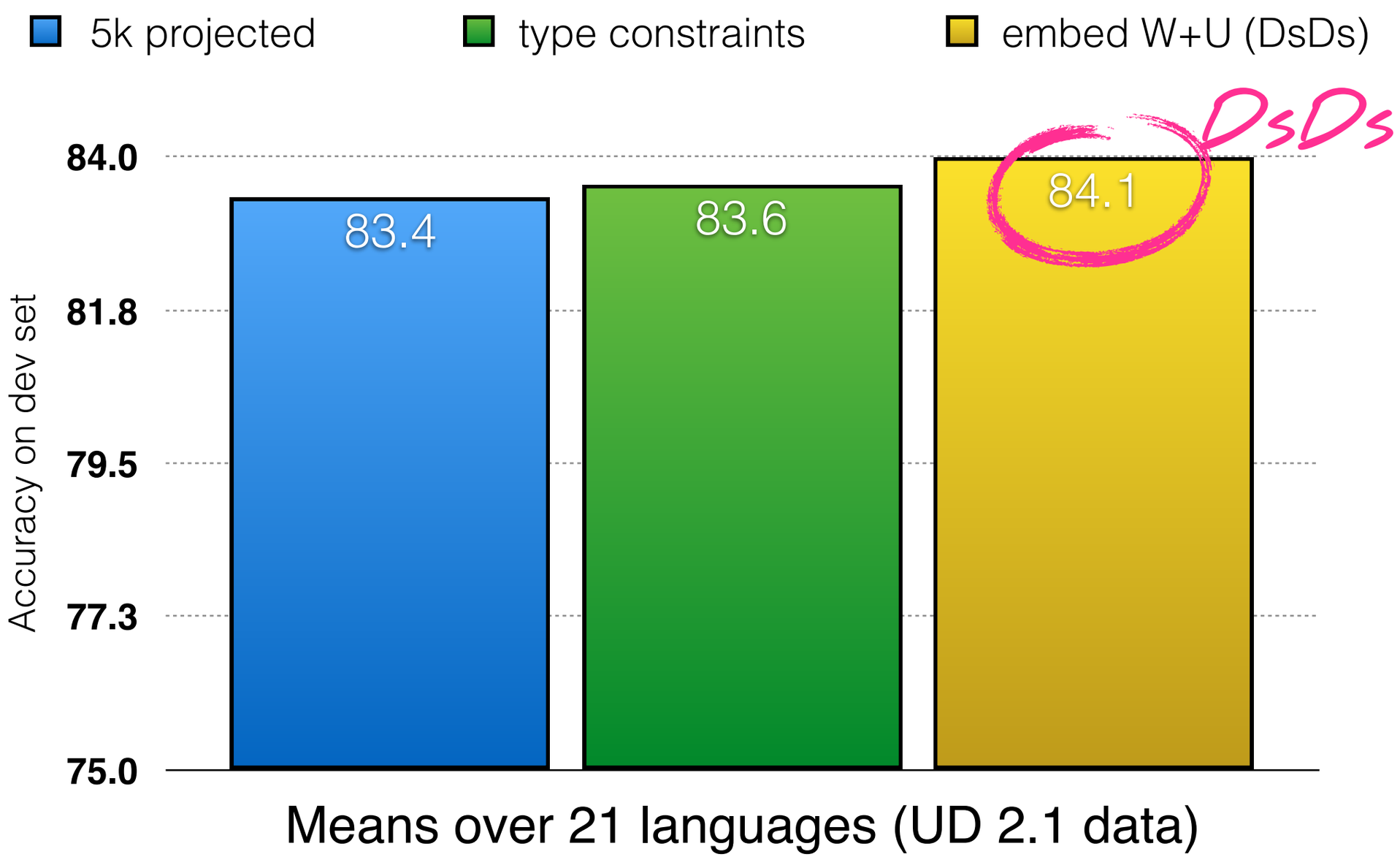
Use of lexicons: Alternatives

- Use lexicon as type constraints during decoding (e.g. Täckström et al., 2013)



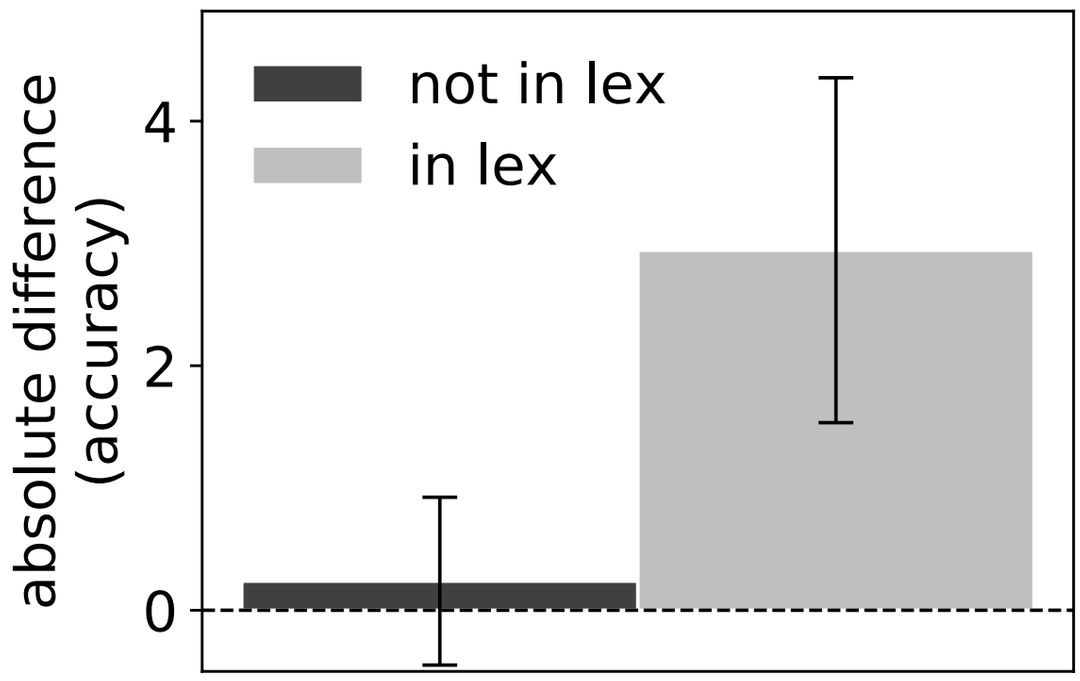
- Retrofitting (Faruqui et al., 2015) for embeddings (“a priori lexicon-based fine-tuning”) - [did not work well]

Inclusion of Lexical information



Analysis: Coverage?

(a) Absolute improvement (delta) vs number of dictionary properties (ρ=0.08).

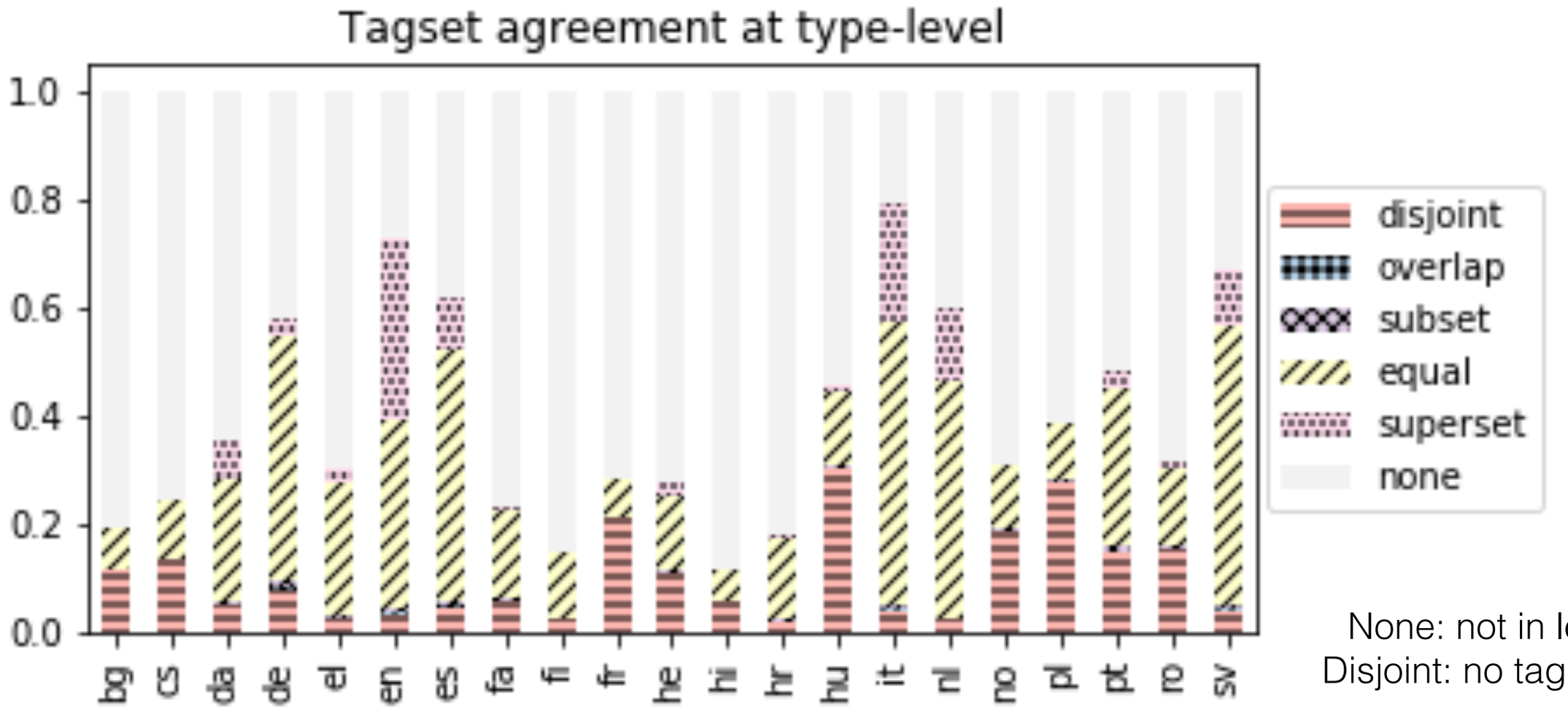


OOV accuracy

- **Coverage** is only part of the explanation

Analysis: Treebank tag set vs lexicon

(inspired by Li et al., 2012)

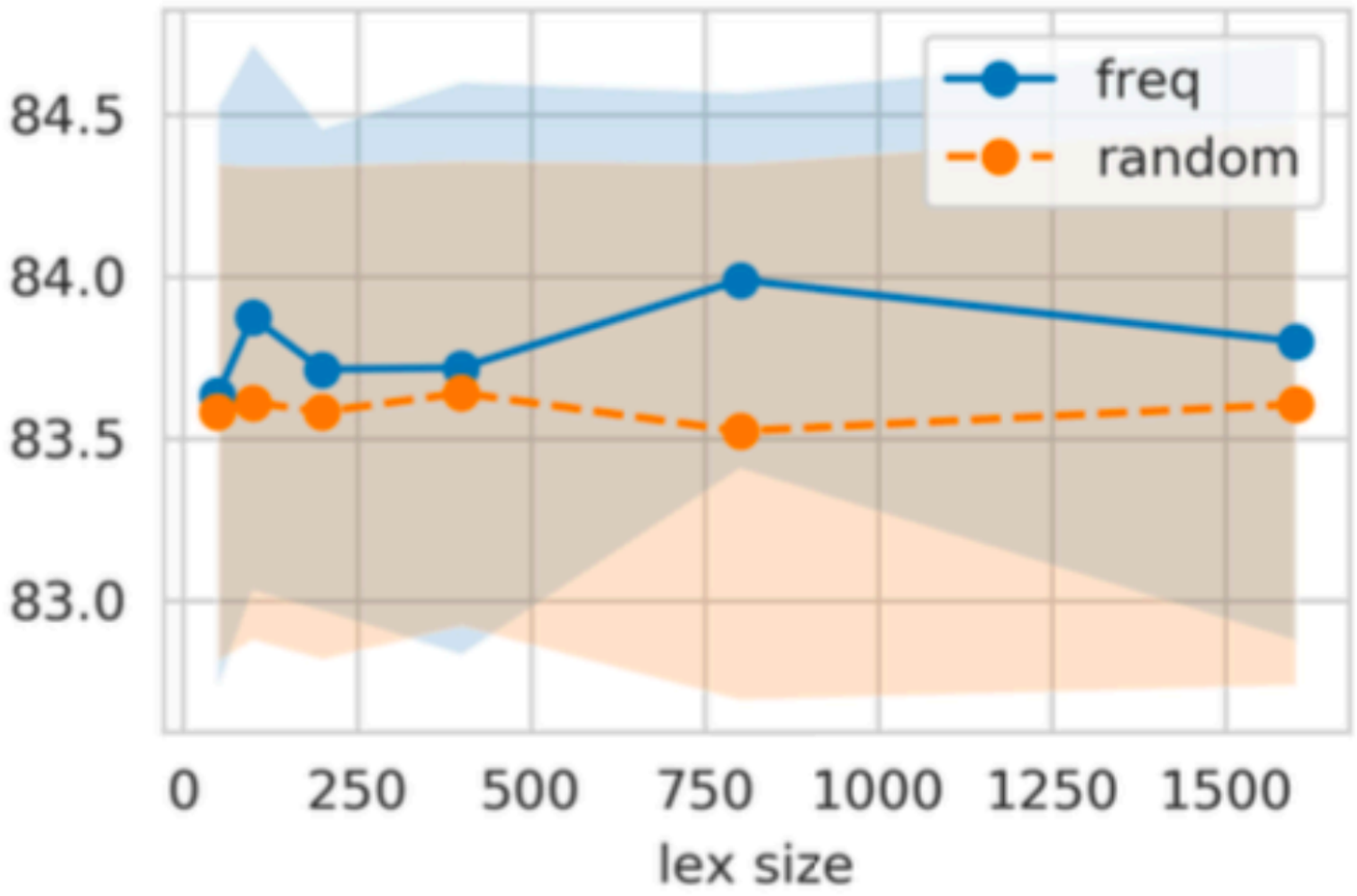


None: not in lexicon
Disjoint: no tag overlap

- For languages where disjoint is low, Type constraints help typically (Greek, English, Croatian, Dutch)

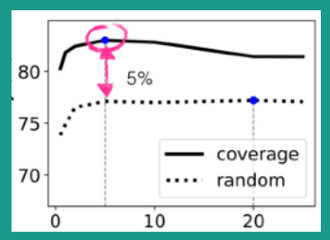
- More implicit use by DSDS helps on languages with high dict coverage and low tag set agreement (e.g., Danish, Dutch, Italian) and languages with low dictionary coverage (such as Bulgarian, Hindi, Croatian, Finnish)

Analysis: Learning curves over dictionary size

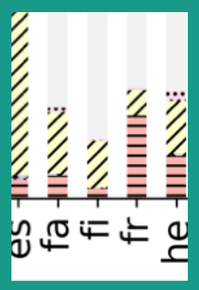


(a) Average effect over 21 languages of high-freq and random dictionaries

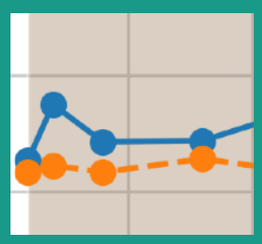
Take-aways



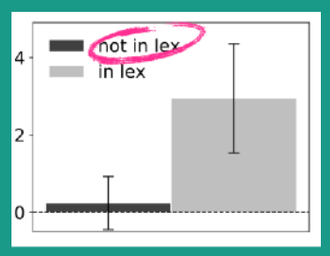
1. **Coverage-based data selection** boosts projection performance (+5% on average)



2. **DsDs (soft inclusion) better** if lexicon has **higher disjoint tag info** (not too high) or better on low-coverage cases than (hard) type constraints



3. **Smaller** dictionaries work well, too



4. Model learns properties **beyond** the lexicon's coverage

Our approach so far

- No gold data (only 5k projected data!)
- No sharing between languages during learning

Lots of alternative approaches…
For instance, what if you had **some** in-language data?

2

NER for low-resource Danish: Cross-Lingual Transfer, Target language annotation, or both?

Neural Cross-Lingual Transfer and Limited Annotated Data for Named Entity Recognition in Danish

Barbara Plank
Department of Computer Science
ITU, IT University of Copenhagen
Denmark
bplank@itu.dk

NoDaLiDa 2019

Motivation

- **RQ1**: To what extent can we transfer a NER tagger to Danish from existing English resources?
- **RQ2**: How does cross-lingual model transfer compare to annotating small amounts of gold data? And how to best combine them?

Annotation with a Limited Budget

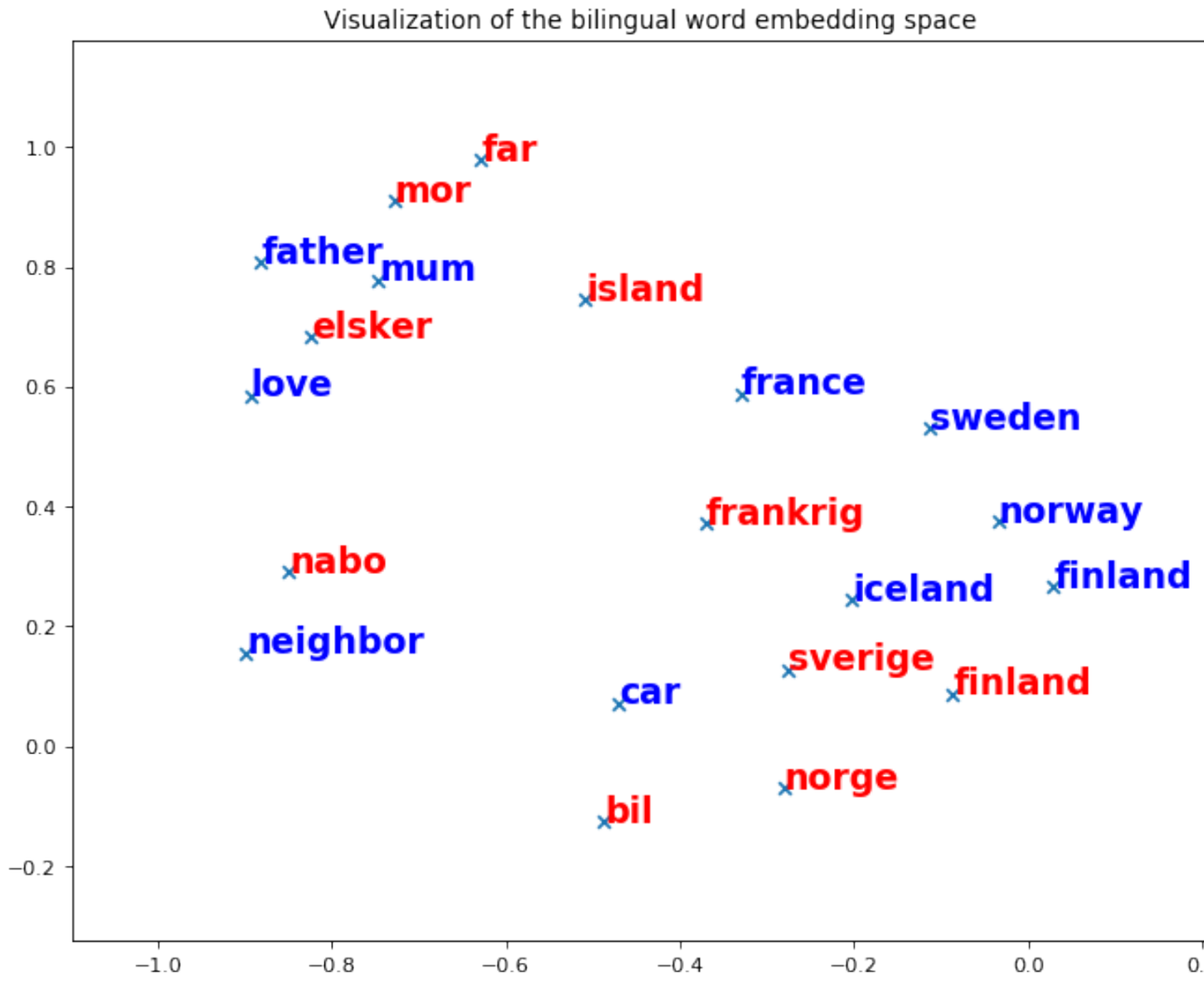
- **Data**: We annotated a subset of the Danish Universal Dependencies (UD) data for NERs
  - Dev set & Test set (both around 10k tokens, ~560 sentences)
  - Two training data set sizes: **Tiny** (272 sentences) and **Small** (604 sentences)
- Note: Lower density of NER, ~35% of the sentences contain NEs (vs 80% on the CoNLL'03 English NER data)

Data Setups: Data & DataAugment

		English Source (CoNLL 03)	
	#sentences	**Medium**	**Large (all)**
Danish (UD train subset)	**(no target)**	~3k	~14k
	Tiny	272+ ~3k	272+ ~14k
	Small	604+ ~3k	604+ ~14k

Embeddings - Alignment

- Align English and Danish off-the-shelf emeddings with Procrustes rotation method introduced in MUSE (Conneau et al., 2017; Artetxe et al., 2017)

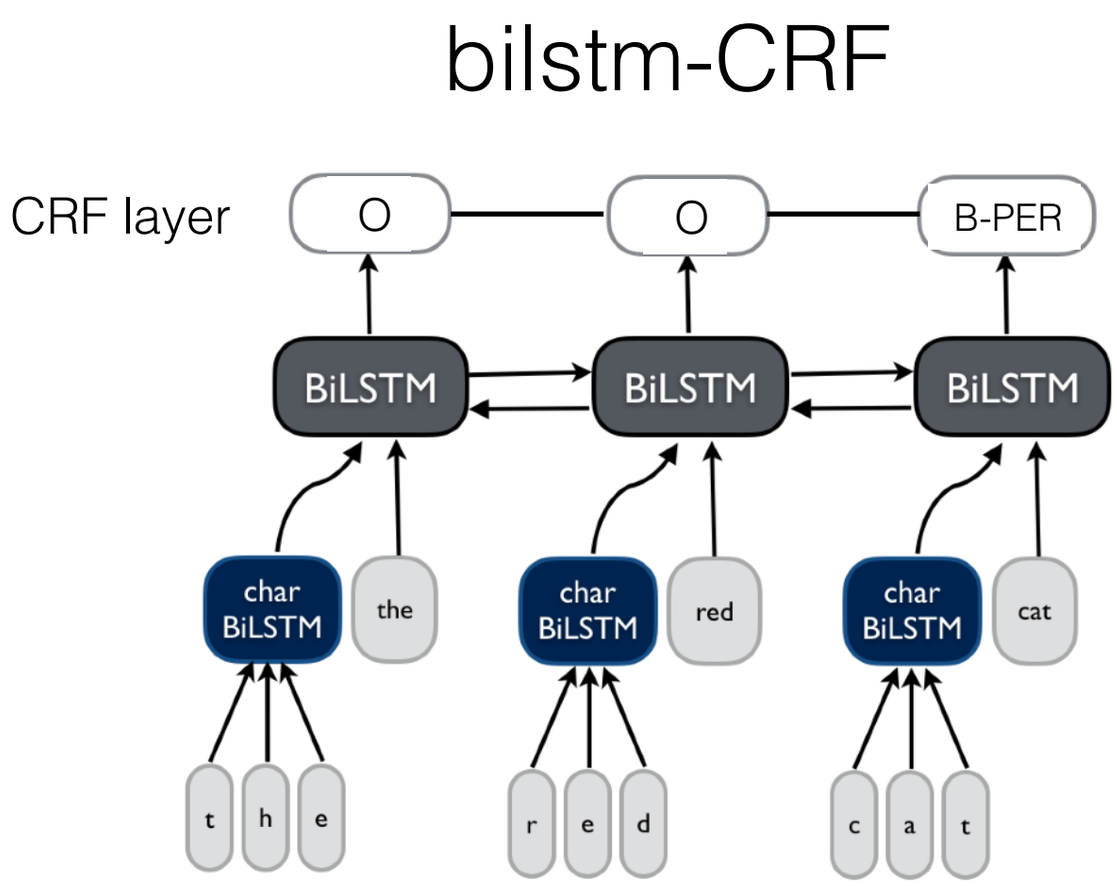


Cross-Lingual Transfer Scenarios

- **Zero-shot**: Direct model transfer CoNLL03->Danish via bilingual embeddings
- **Few-shot direct transfer (DataAug):** train on concatenation English & Danish (tiny|small)
- **Few-shot fine-tuning:** train first on English, then fine-tune on Danish
- **In-language baseline** (train on tiny|small Danish data)

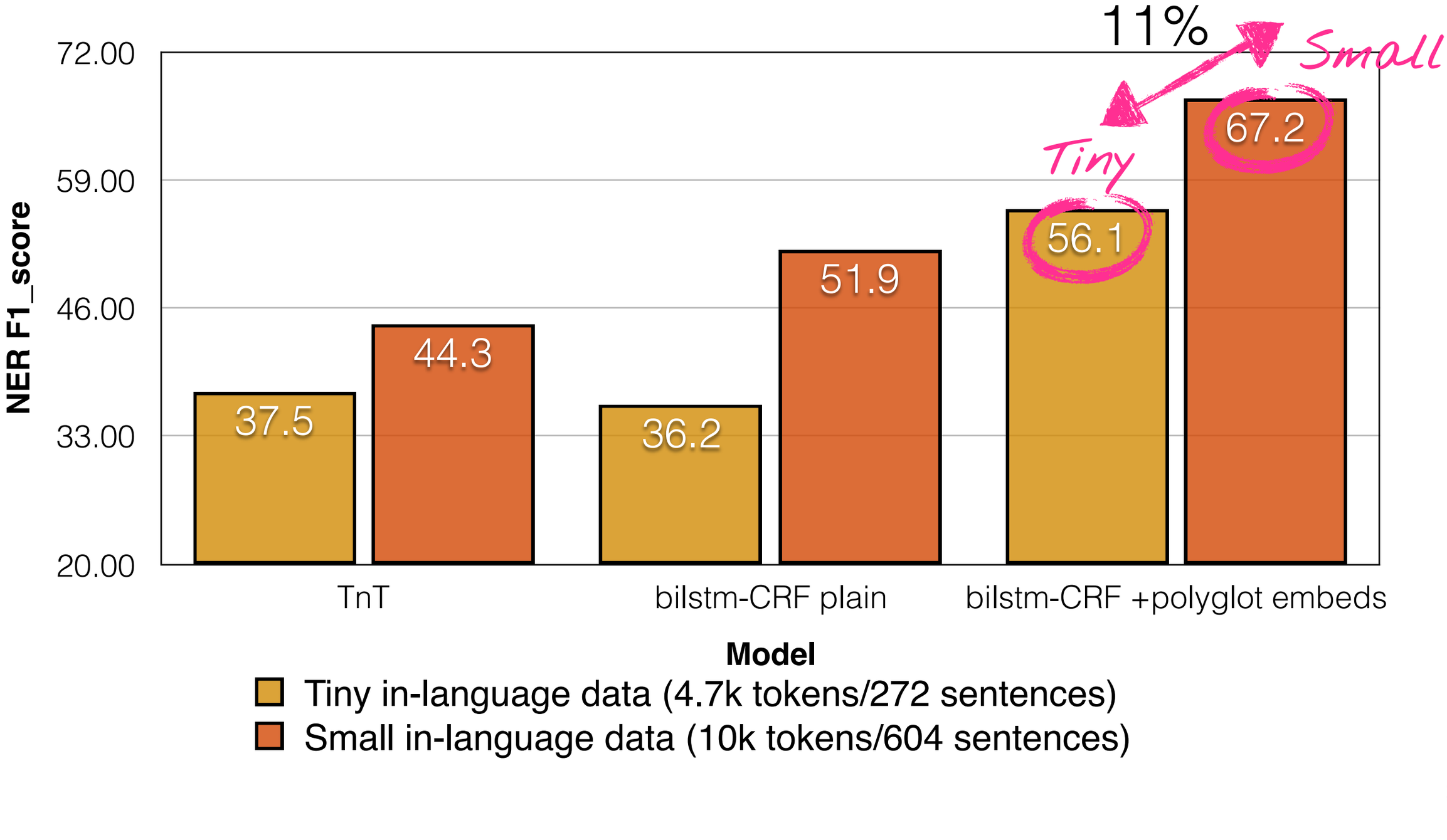
Model and Approach

- Similar to Ma and Hovy (2016) but with a character-level bilstm



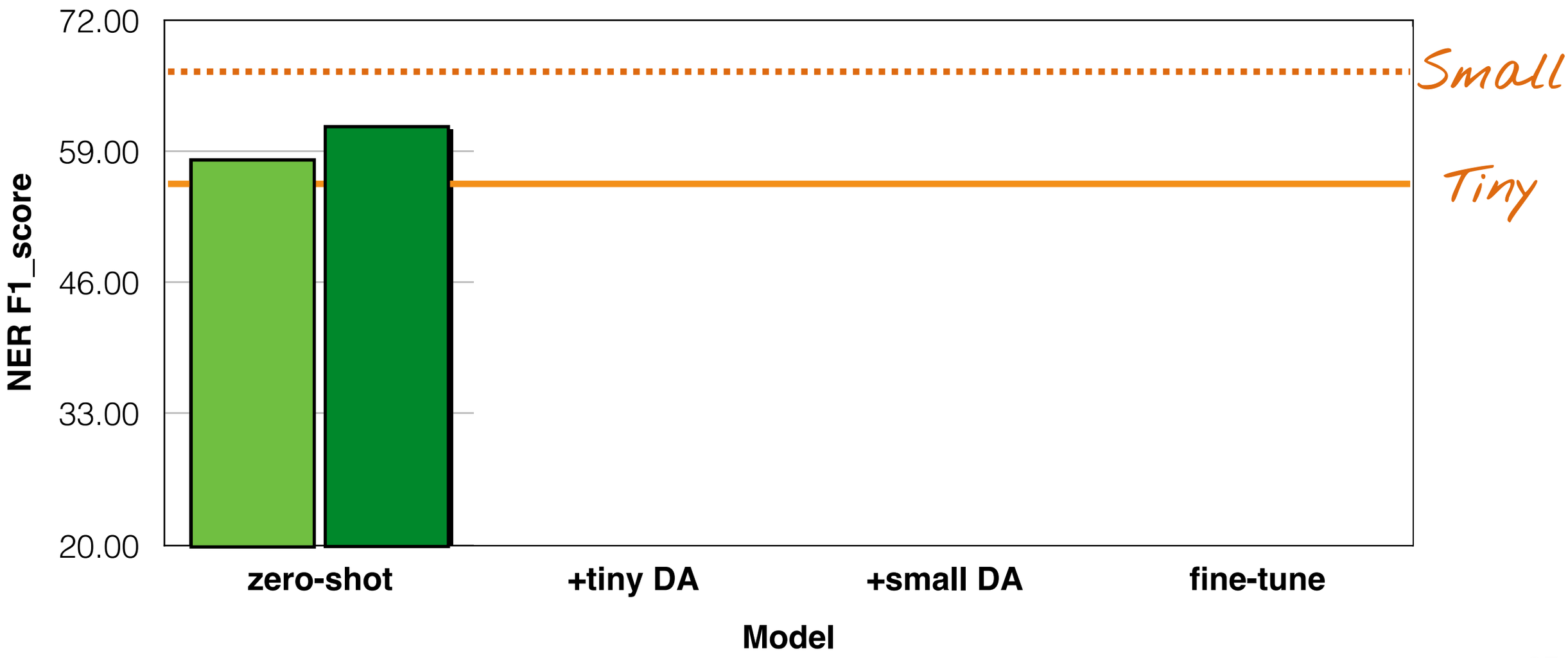
Results: Baselines

- Training on small amounts of annotated target Danish data



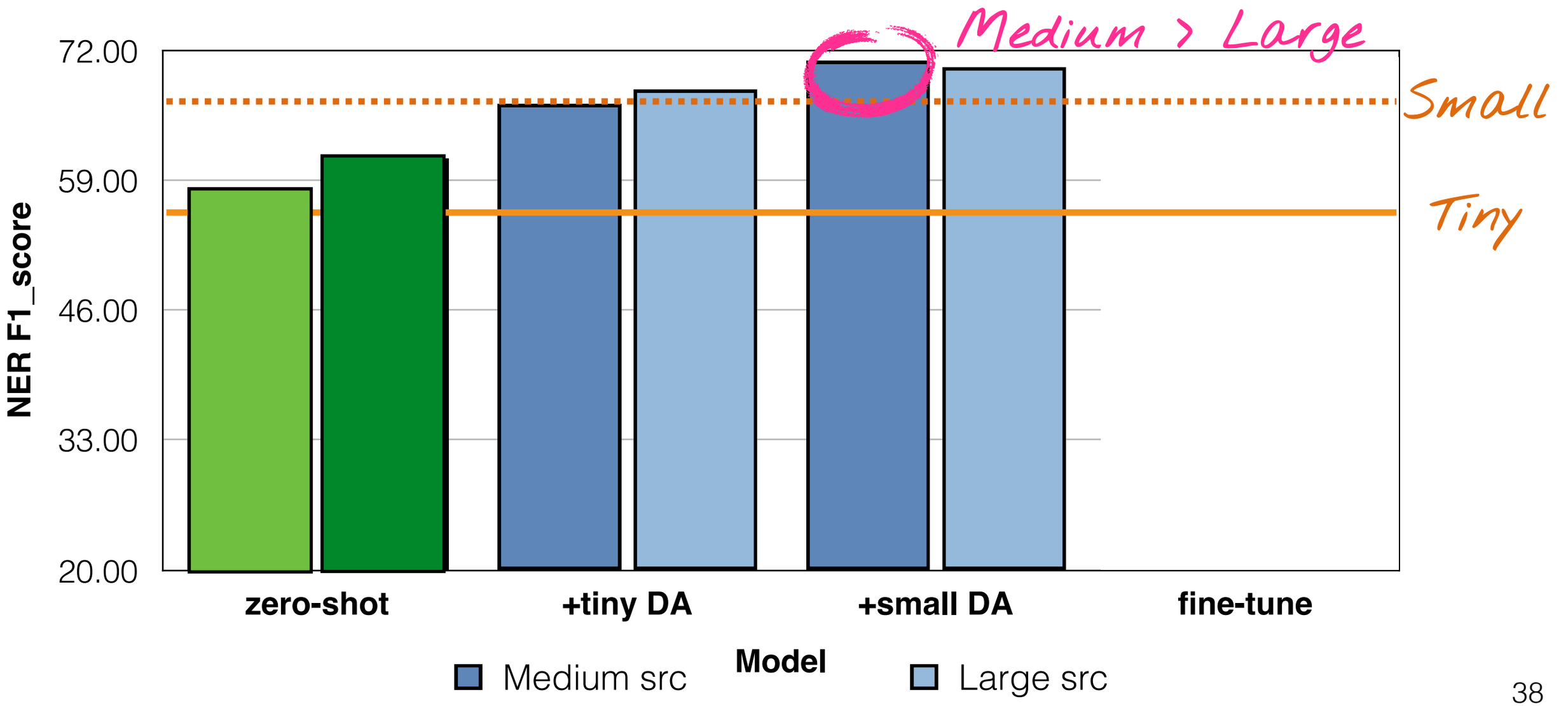
Results: Cross-lingual transfer

- **RQ1**: To what extent can we directly transfer a NER tagger from English to Danish (**zero-shot learning**)?



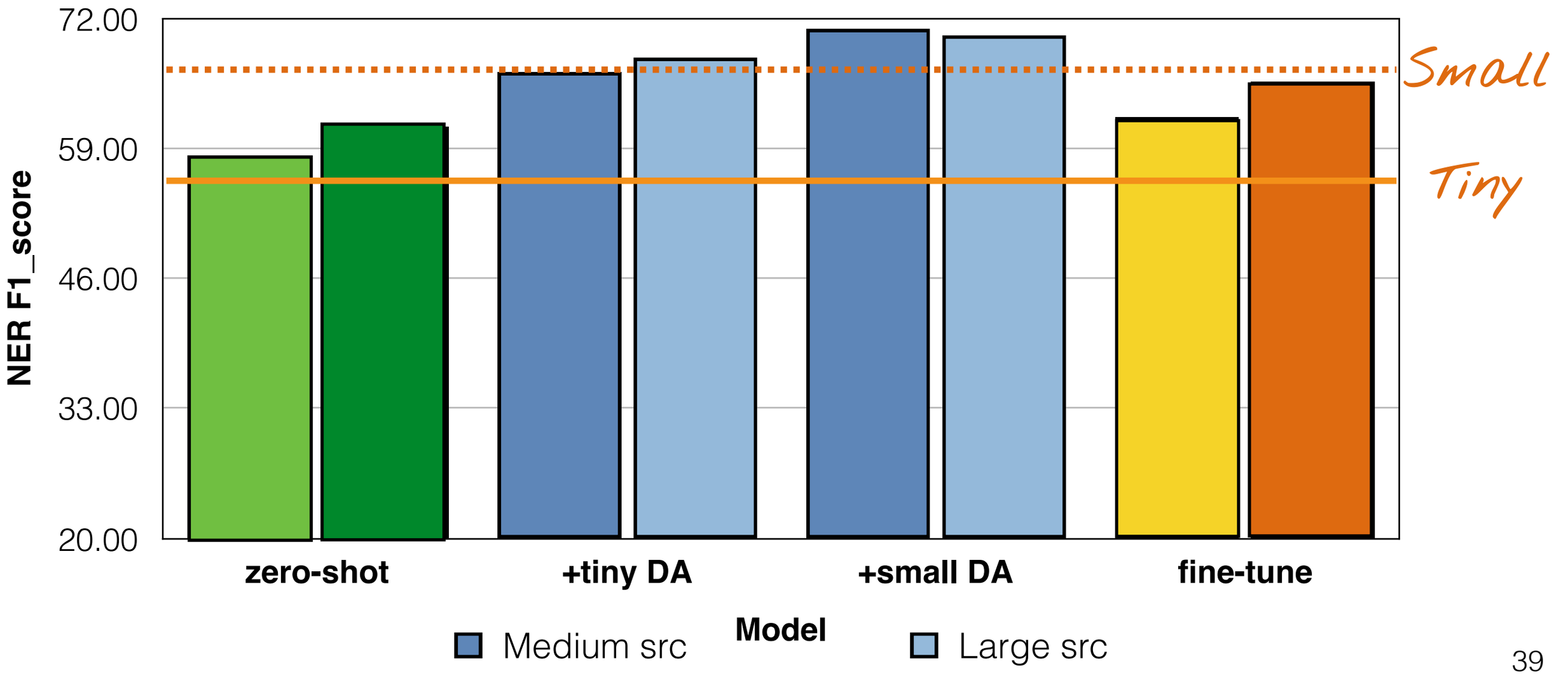
Results: Cross-lingual transfer

- **RQ2**: How does transfer compare to small amounts of annotated labeled data (**few-shot learning**)?



Results: Cross-lingual transfer

- **RQ2**: Worse results with fine-tuning.



Results: Comparison

- **RQ3**: How good are existing systems for Danish?
- Best system identified: Polyglot NER (Al-Rfou et al., 2015) build on automatically-derived data from Wikipedia & Freebase

TEST	All	PER	LOC	ORG	MISC
Polyglot	61.6	78.4	**69.7**	24.7	—
Bilstm	**66.0**	**86.6**	63.6	**42.5**	24.8

Table 4: F_1 score for Danish NER.

Take-aways

- The most beneficial way is **DataAug**: add a small amount of target data to the source; fine-tuning was inferior
- Less source (EN) data is better: best transfer from the Medium setup (rather than the entire CoNLL data)
- Very little target data paired with dense cross-lingual embeddings yields an effective NER tagger for Danish quickly.

Part II:

Cross-domain learning

Learning to select data for transfer learning with Bayesian optimization

Sebastian Ruder and Barbara Plank

EMNLP 2017

Data Setup:
Multiple Source Domains

How to select the most relevant data?

Motivation

Why? Why don’t we just train on all source data?

- **Prevent negative transfer**
- e.g. “predictable” is negative for  , but positive in 

Prior approaches:

- use a single similarity metric in isolation;
- focus on a single task.

Our approach

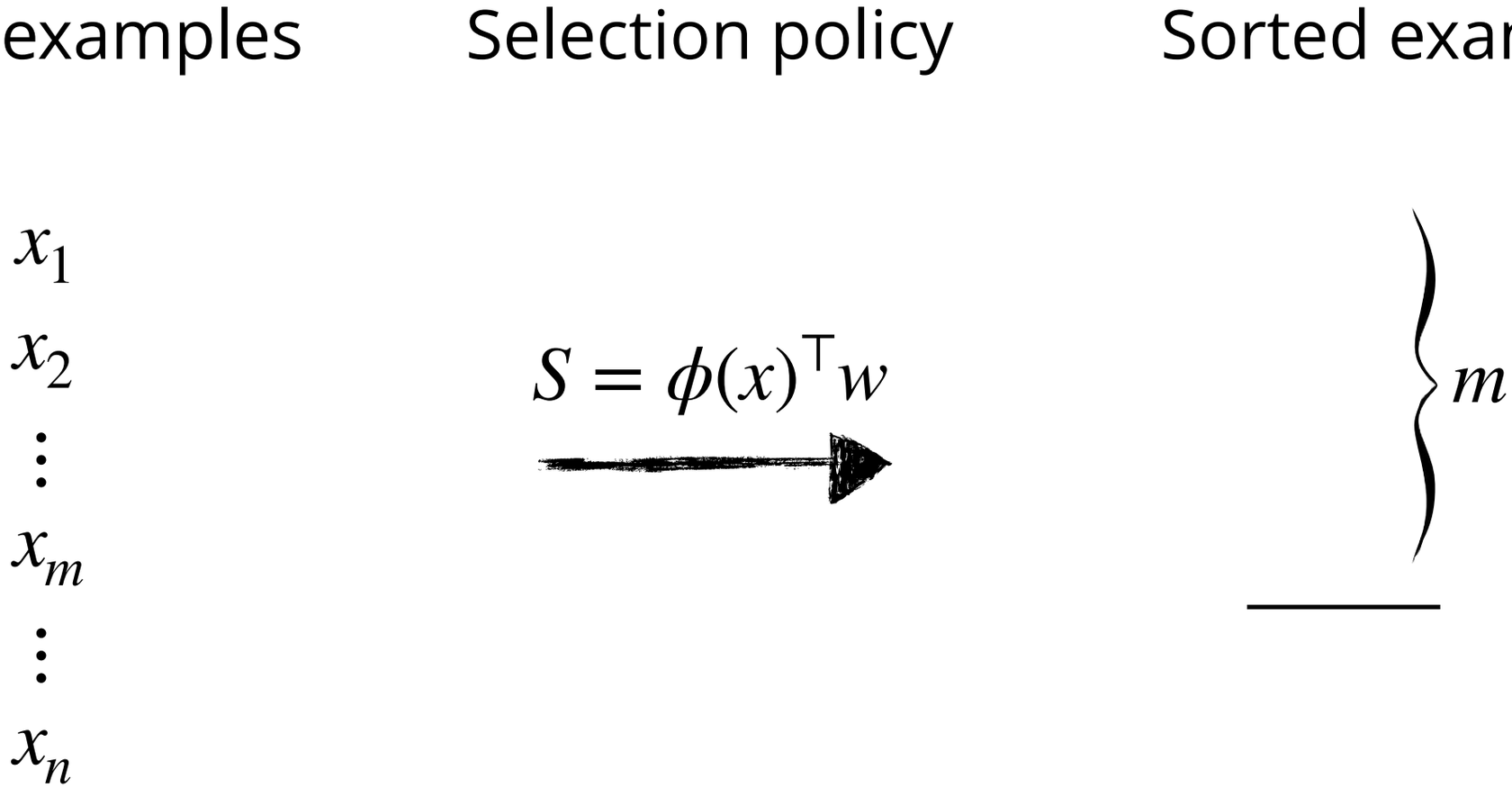
Intuition

- Different tasks and domains require different notions of similarity.

Idea

- Learn a data selection policy using Bayesian Optimization.

Our approach



- Related: curriculum learning (Tsvetkov et al., 2016)

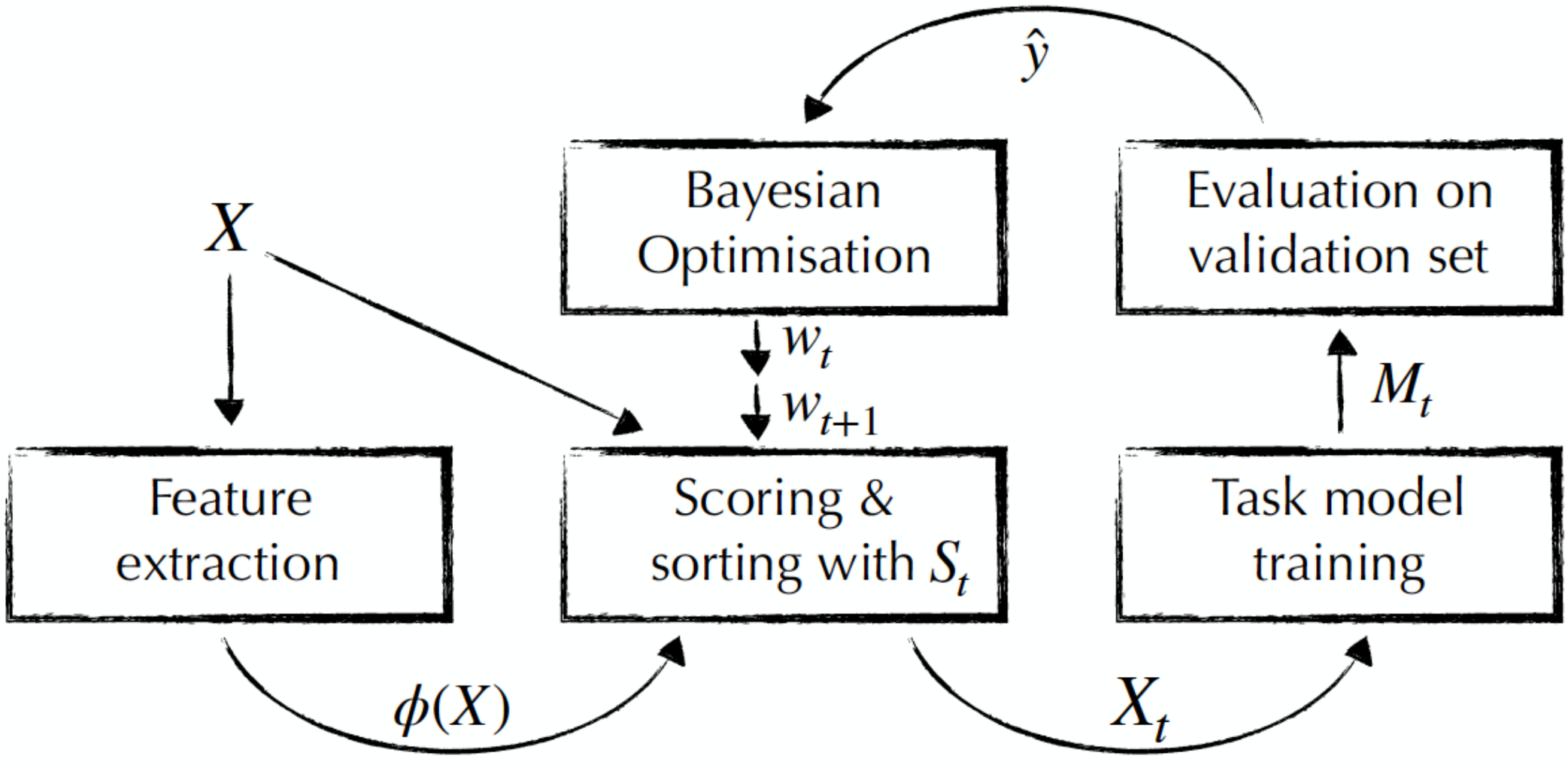
Tsvetkov, Y., Faruqui, M., Ling, W., & Dyer, C. (2016). Learning the Curriculum with Bayesian Optimization for Task-Specific Word Representation Learning. In *Proceedings of ACL 2016*.

Bayesian Data Selection Policy

$$\mathcal{S} = \phi(\mathcal{X}) \cdot w^T$$

learned feature weights

different similarity/ diversity features



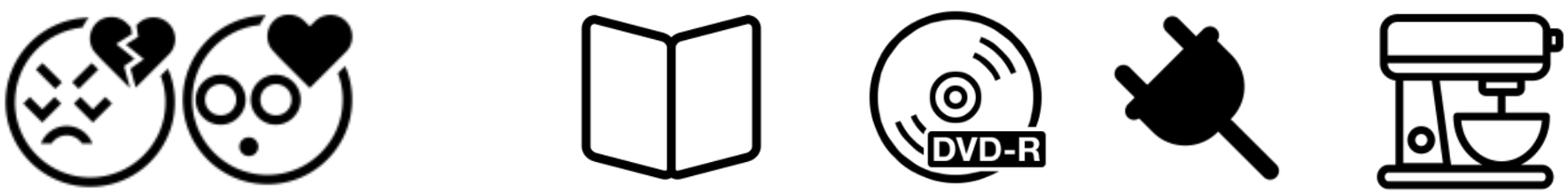
Features $\phi(X)$

- **Similarity:**
  Jensen-Shannon, Rényi div, Bhattacharyya dist,
  Cosine sim, Euclidean distance, Variational dist
  - **Representations**:
    Term distributions, Topic distributions,
    Word embeddings (Plank, 2011)

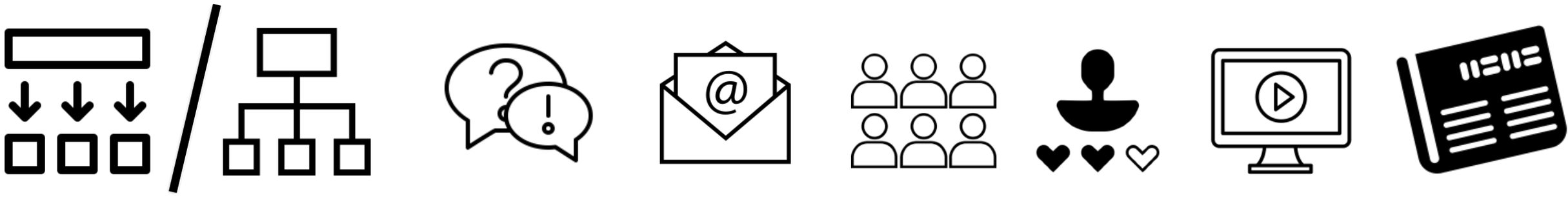
- **Diversity**: #types, TTR, Entropy, Simpson's index, Rényi entropy, Quadratic entropy

Data & Tasks

Three tasks: **Domains:**



Sentiment analysis on Amazon reviews dataset (Blitzer et al., 2007)

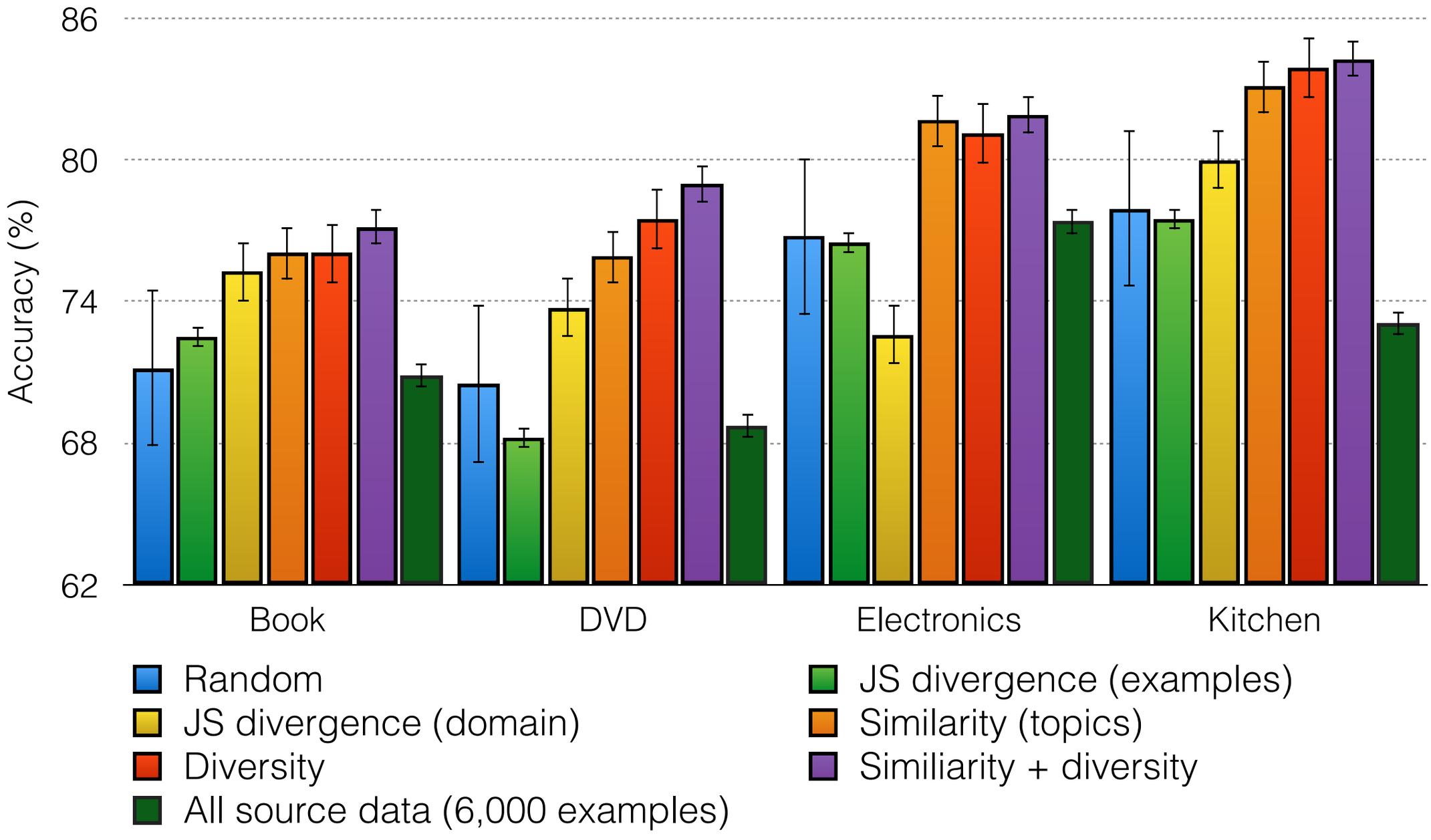


POS tagging and dependency parsing on SANCL 2012 (Petrov and McDonald, 2012)

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL 2007*.
Petrov, S., & McDonald, R. (2012). Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*.

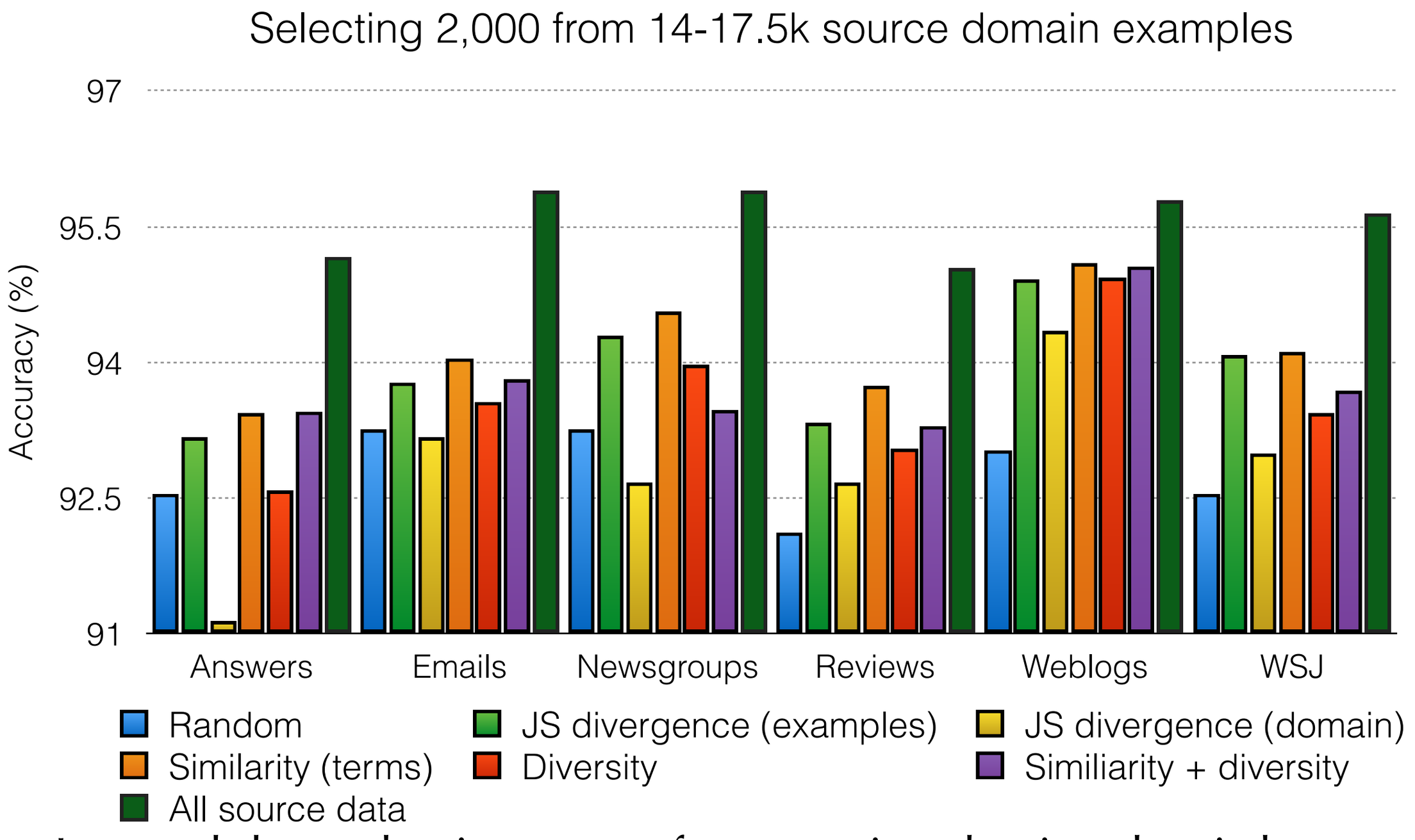
Sentiment Analysis Results

Selecting 2,000 from 6,000 source domain examples



- Selecting relevant data is useful when domains are very different.

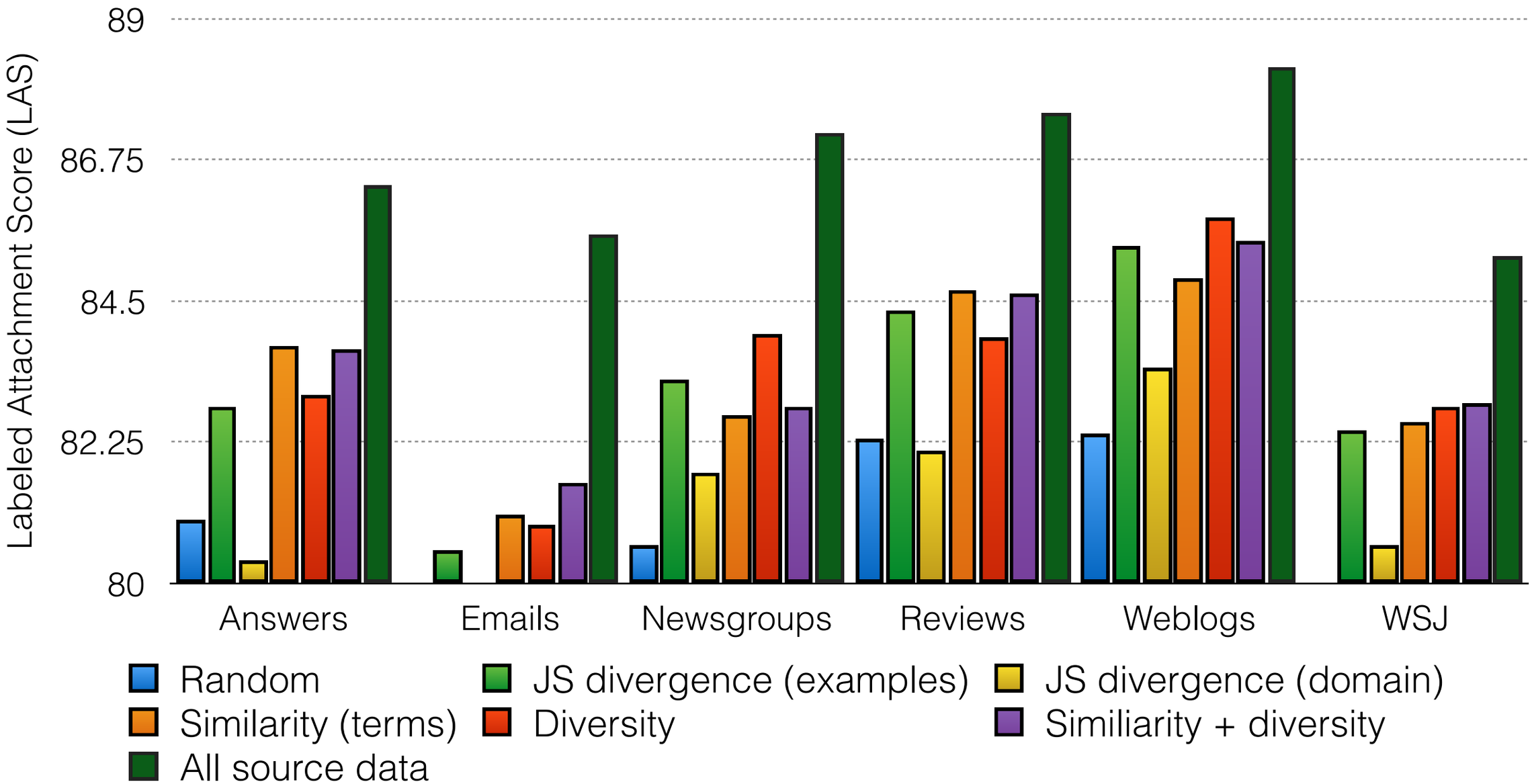
POS Tagging Results



- Learned data selection outperforms static selection, but is less useful when domains are very similar.

Dependency Parsing Results

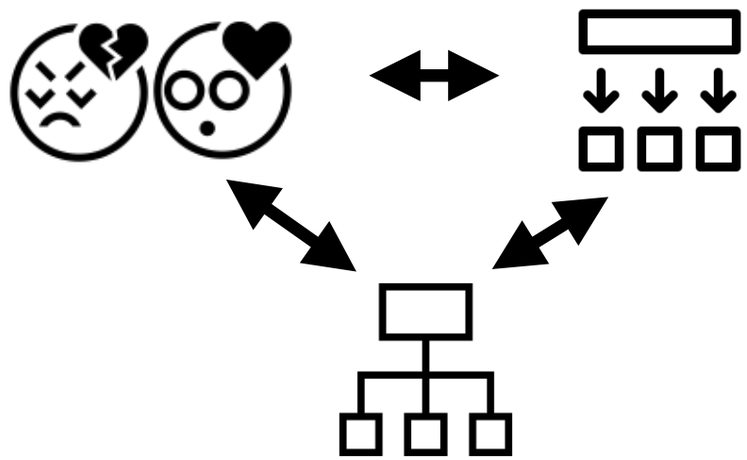
Selecting 2,000 from 14-17.5k source domain examples



(BIST parser, Kiperwasser & Goldberg, 2016)

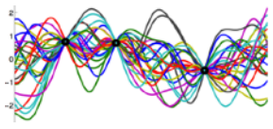
Do the weights transfer?

Cross-task transfer



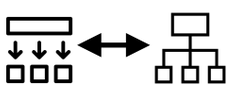
		Target tasks		
Feature set	$\mathcal{T}_S$	**POS**	**Pars**	**SA**
Sim	POS	93.51	83.11	74.19
Sim	Pars	92.78	83.27	72.79
Sim	SA	86.13	67.33	79.23
Div	POS	93.51	83.11	69.78
Div	Pars	93.02	83.41	68.45
Div	SA	90.52	74.68	79.65
Sim+div	POS	93.54	83.24	69.79
Sim+div	Pars	93.11	83.51	72.27
Sim+div	SA	89.80	75.17	80.36

Take-aways



- Domains & tasks have different notions of similarity. Learning a task-specific data selection policy helps.

- Preferring certain examples is mainly useful when **domains are dissimilar**.



- The learned policy **transfers** (to some extent) across models, tasks, and domains

Code: https://github.com/sebastianruder/learn-to-select-data

Our approach so far

- Assumed labeled data for the target domain
- What to do if no such data is available? (i.e., unlabeled data only)

Bootstrapping methods

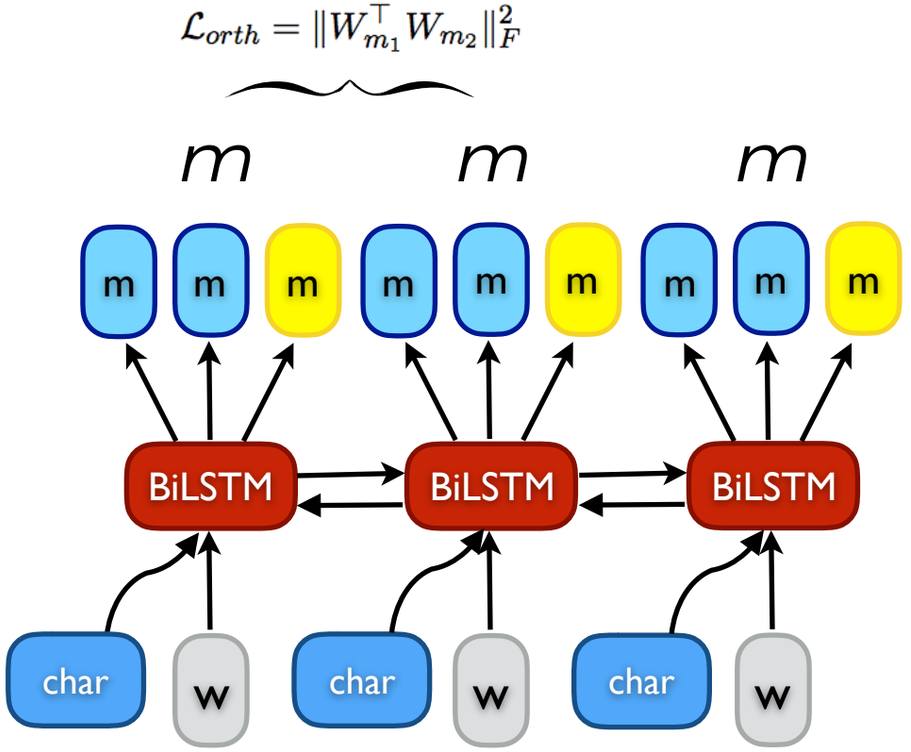
Strong Baselines for Neural Semi-supervised Learning under Domain Shift

Sebastian Ruder[♠♣] **Barbara Plank**[♡◇]

[♠]Insight Research Centre, National University of Ireland, Galway, Ireland
[♣]Aylien Ltd., Dublin, Ireland
[♡]Center for Language and Cognition, University of Groningen, The Netherlands
[◇]Department of Computer Science, IT University of Copenhagen, Denmark

sebastian@ruder.io,bplank@gmail.com

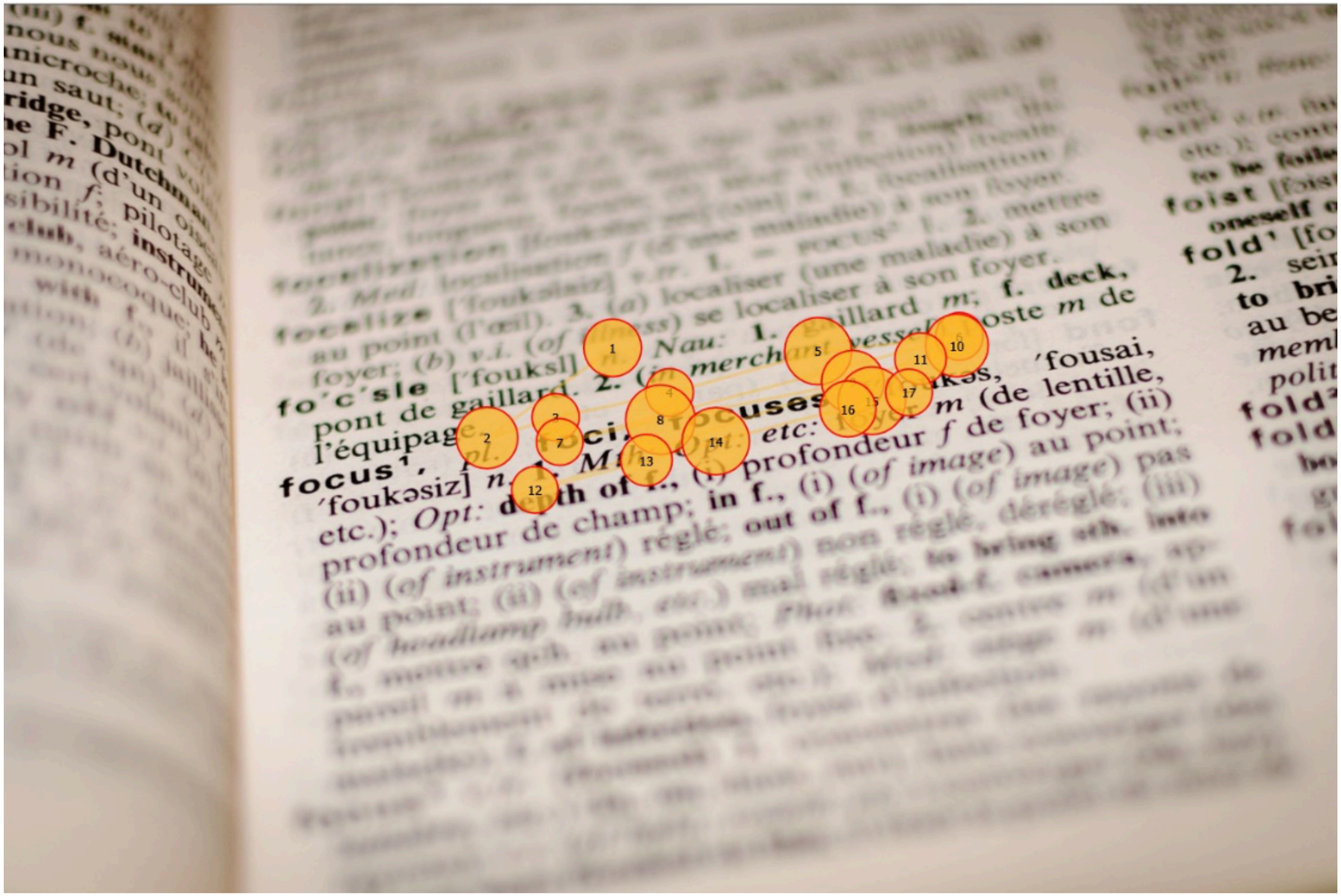
- Classic bootstrapping in SSL:
  - Self-training
  - Tri-training
  - Tri-training with disagreement
- Proposed: **Multi-task tri-training**



ACL 2018

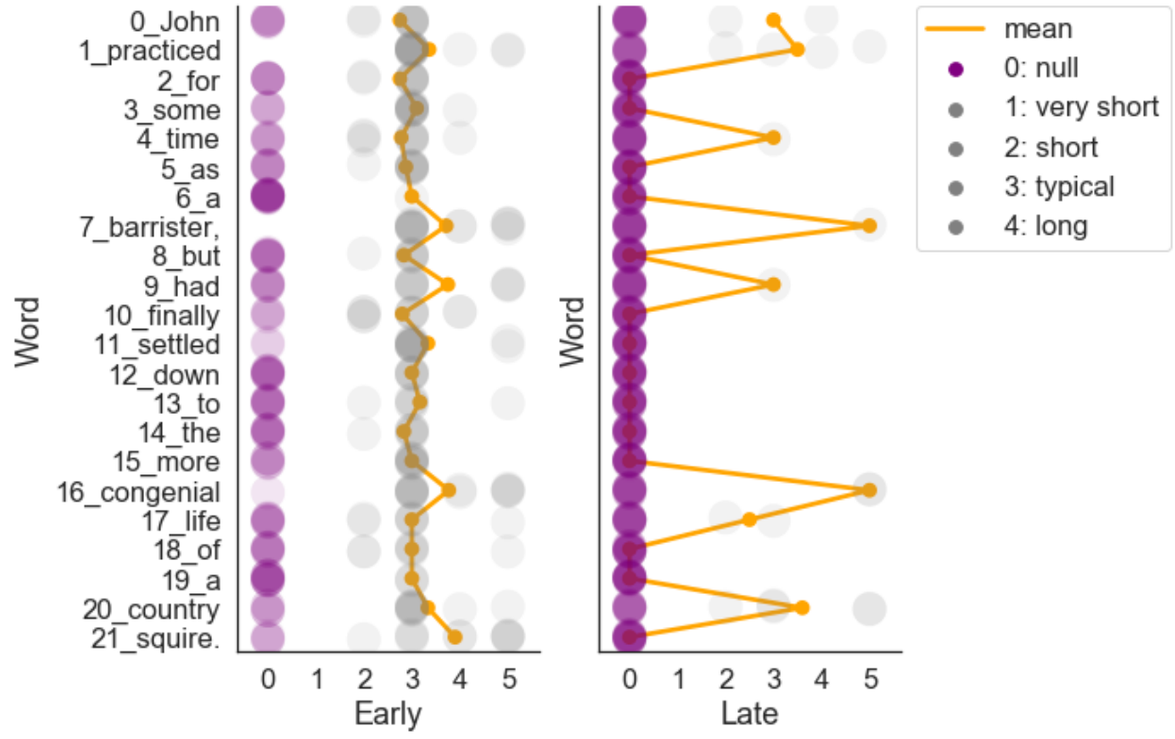
More distant, fortuitous sources?

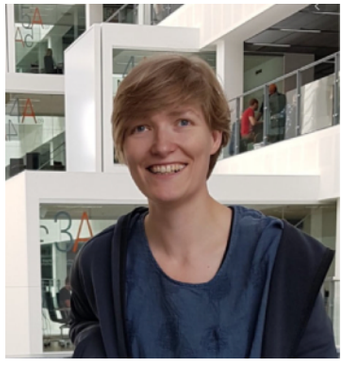
Motivation



https://imotions.com/blog/reading-eye-tracking/

At a Glance:
The Impact of Gaze Aggregation Views on Syntactic Tagging





Sigrid Klerke and Barbara Plank

LANTERN 2019 workshop

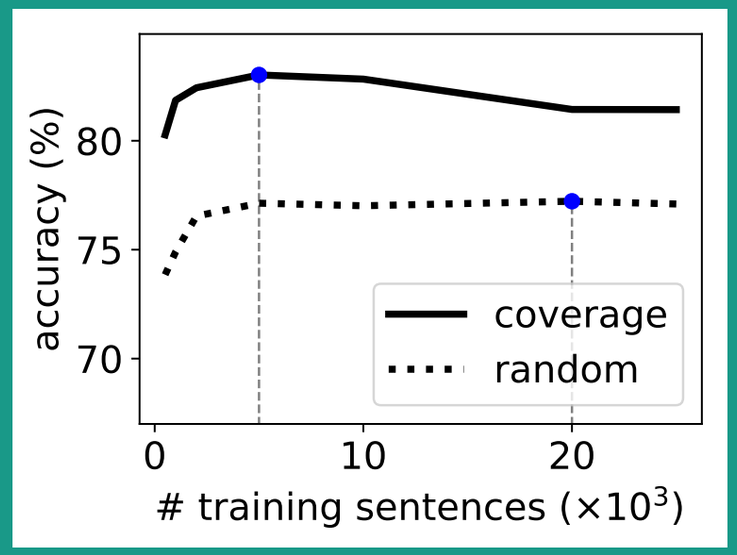
today 16:45-17:00

To wrap up...

Take-away 1

Sometimes less is more:
Data selection is useful!

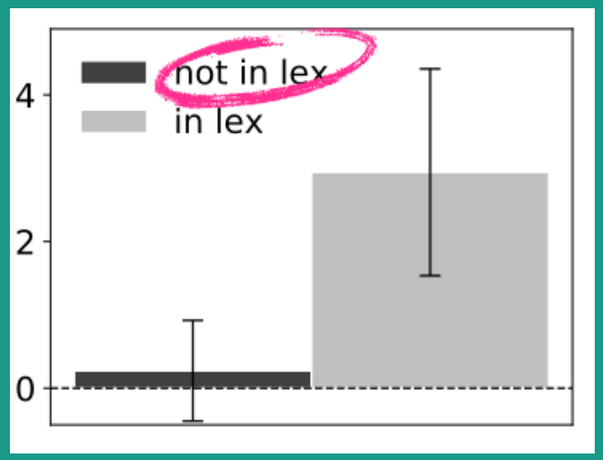
Cross-domain



Cross-lingual

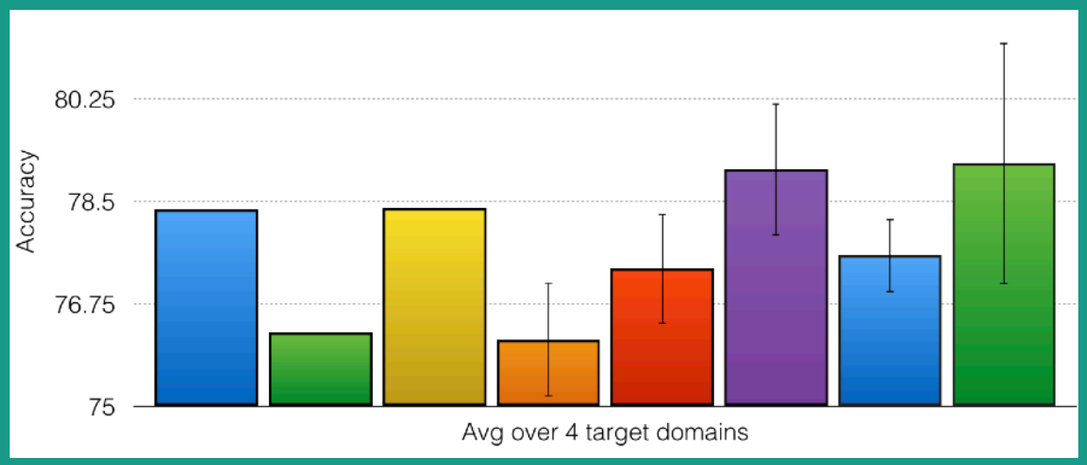
Take-away 2

Combine old & new: :)
Integration of symbolic knowledge helps neural models



Take-away 3

Use **both** old & new:
Importance of strong (traditional) baselines in neural times



It's an exciting time

WHAT?

WHY?

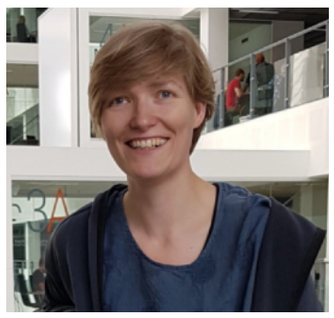
HOW?

Questions?
Thanks!

IT UNIVERSITY OF COPENHAGEN

THE END
Transferring NLP models across languages and domains

Thanks to



Barbara's research team at ITU:

You?

Thanks to sponsors & funding:

amazon

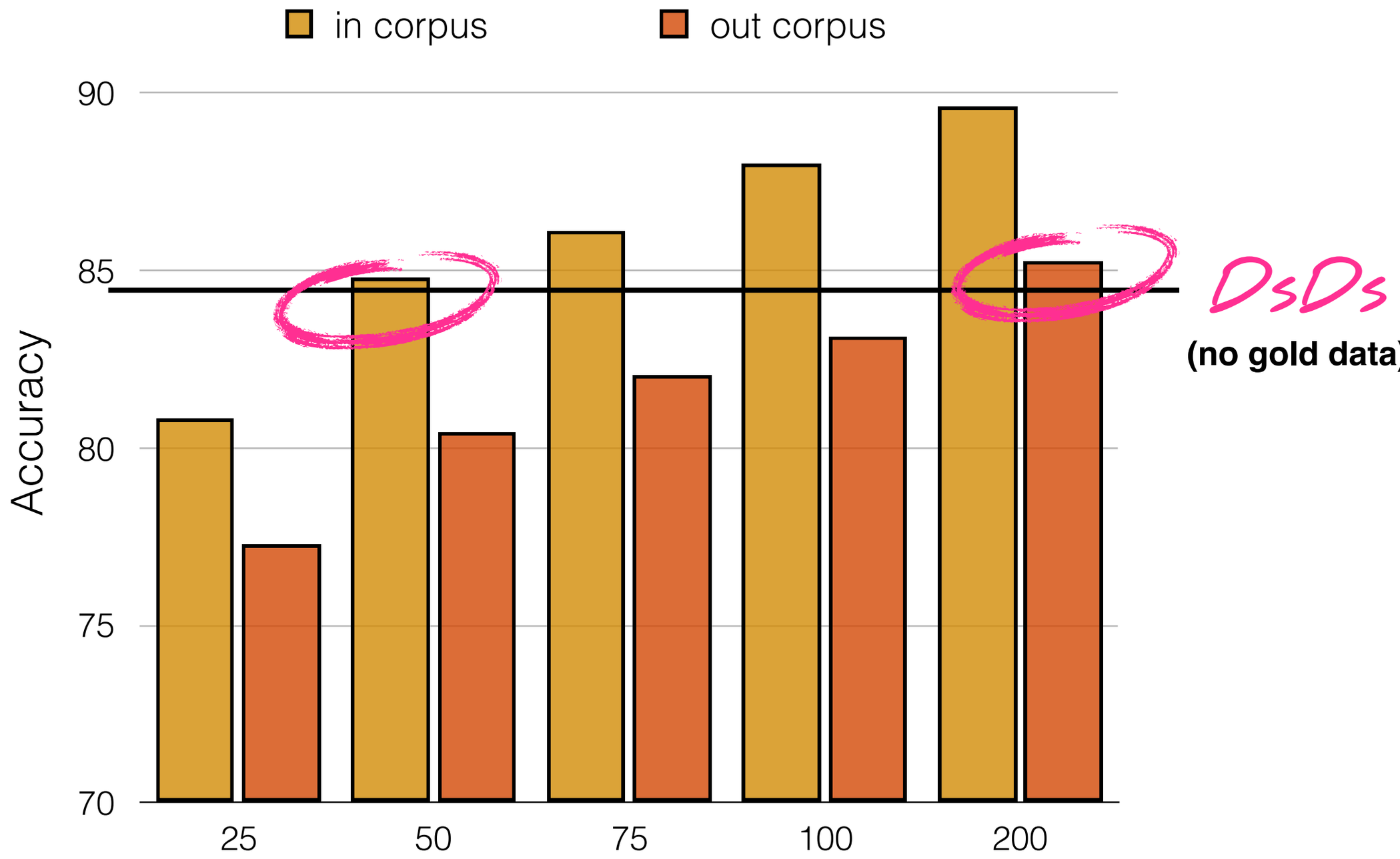
NVIDIA

DANMARKS FRIE FORSKNINGSFOND

PhD opening (spring 2020)

Appendix

How much gold data?



(Means over 18 languages for which we had both in- and out-corpus gold data)