IT UNIVERSITY OF COPENHAGEN

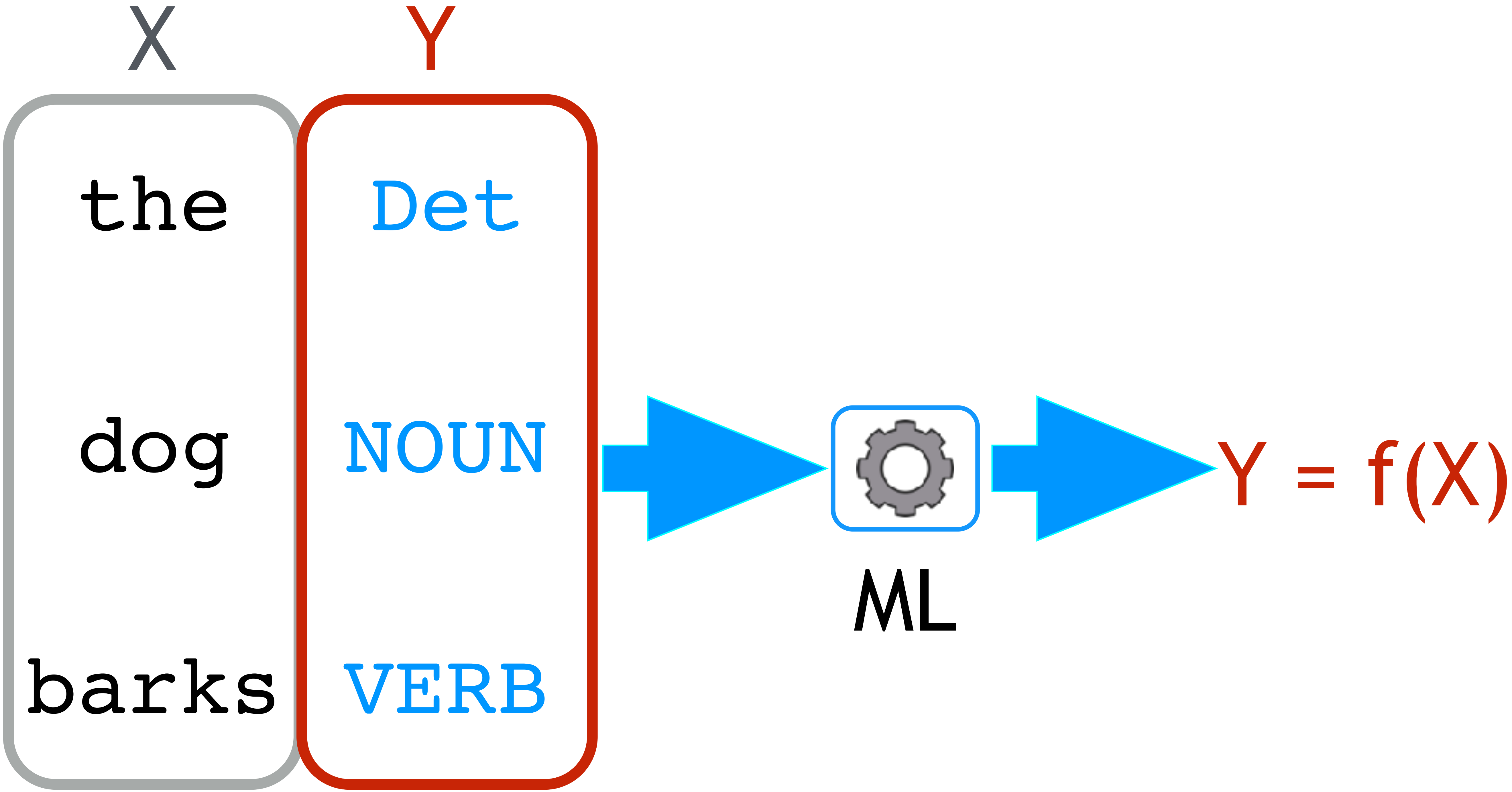
Transferring NLP models across languages and domains

Barbara Plank
ITU, Copenhagen, Denmark

SYNTAX FEST
Paris, August 26-30 2019

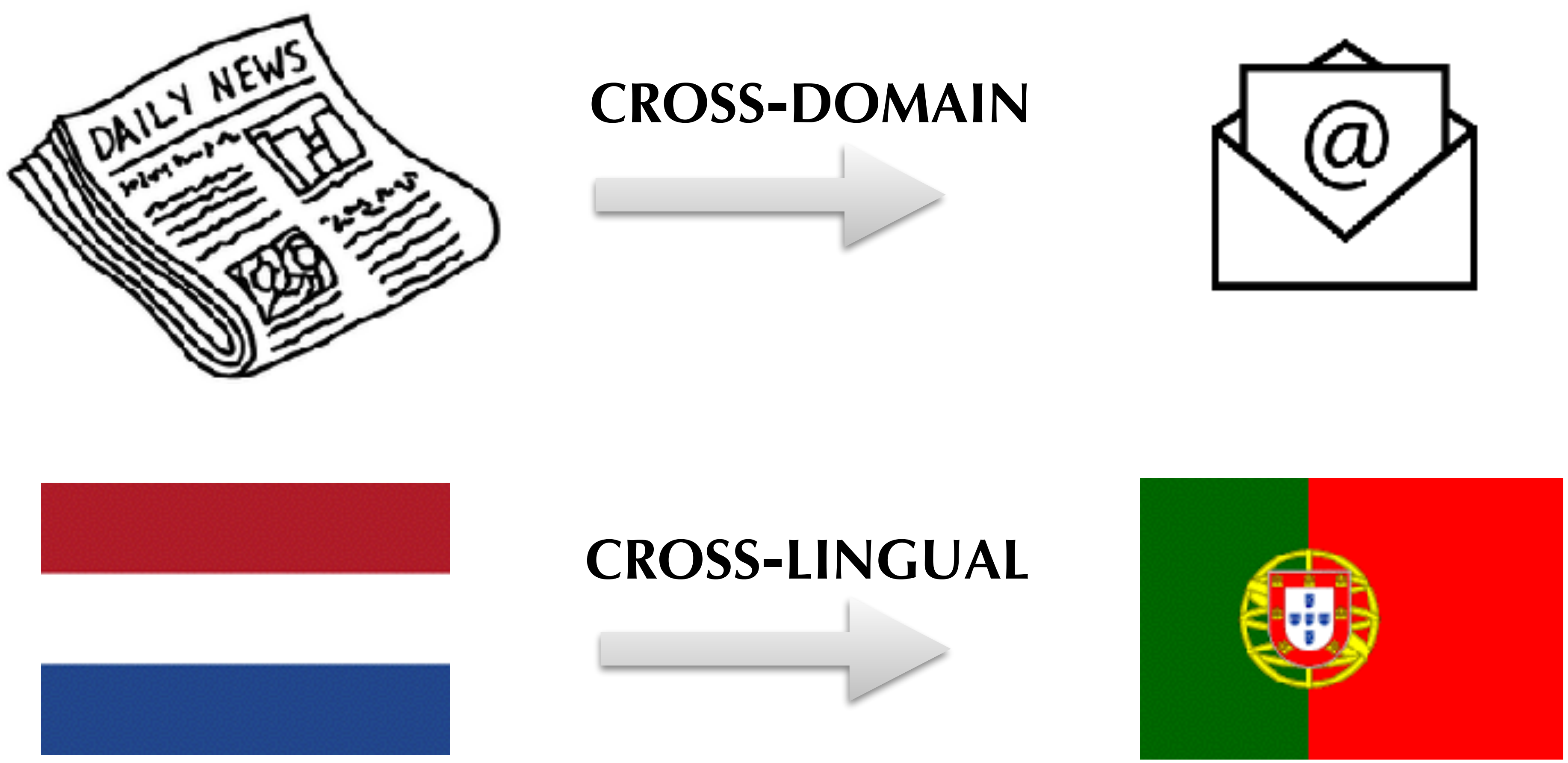
August 28, 2019,
#SyntaxFest2019 Paris

Statistical NLP: The Need for Data



Adverse Conditions

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



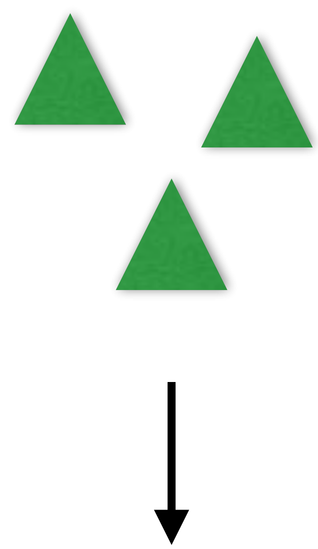
Data variability

- 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?

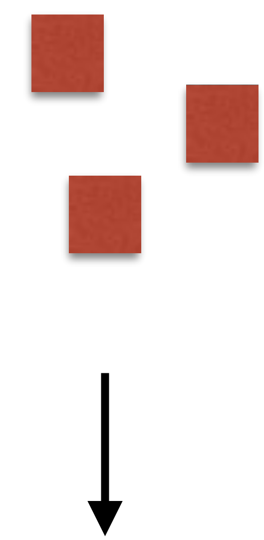
Typical setup



Model A

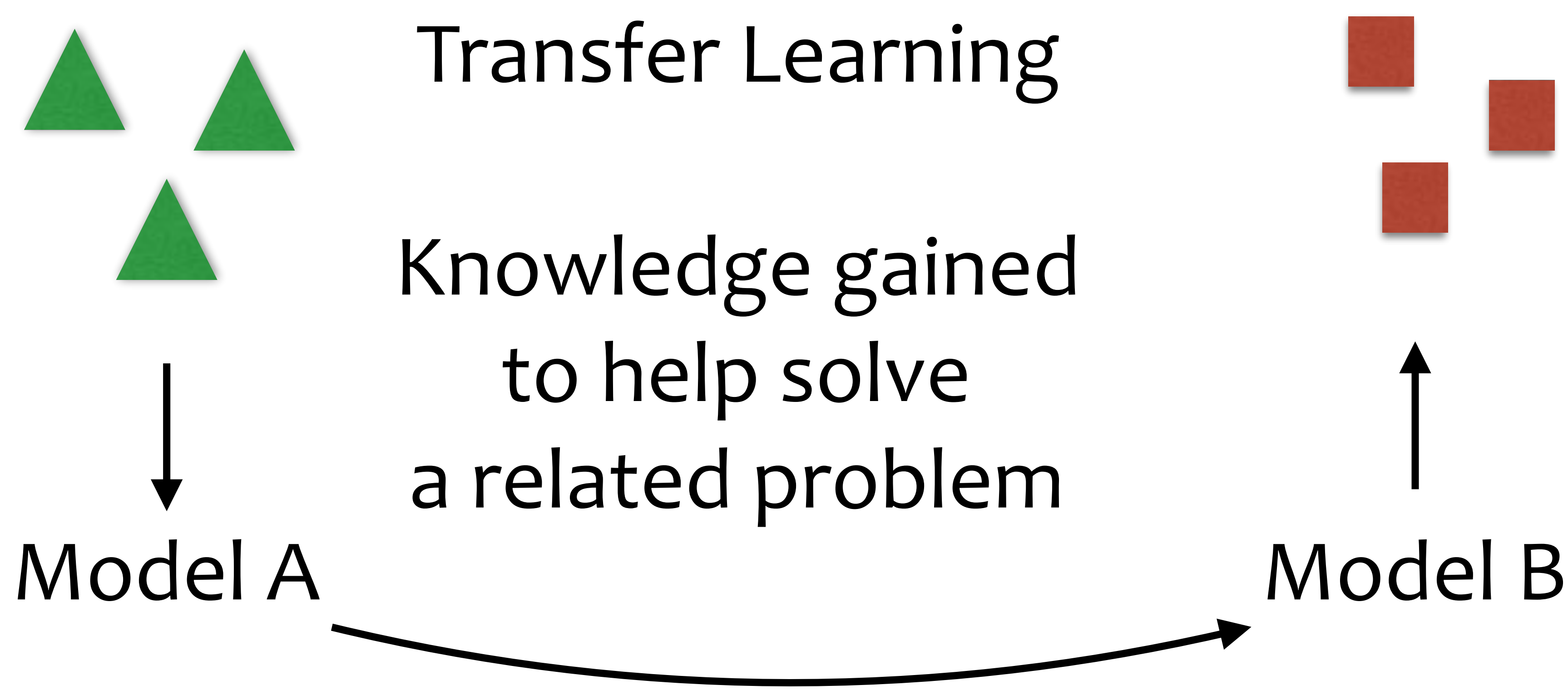
Traditional ML:

Train & evaluate
on same
domain/task/language



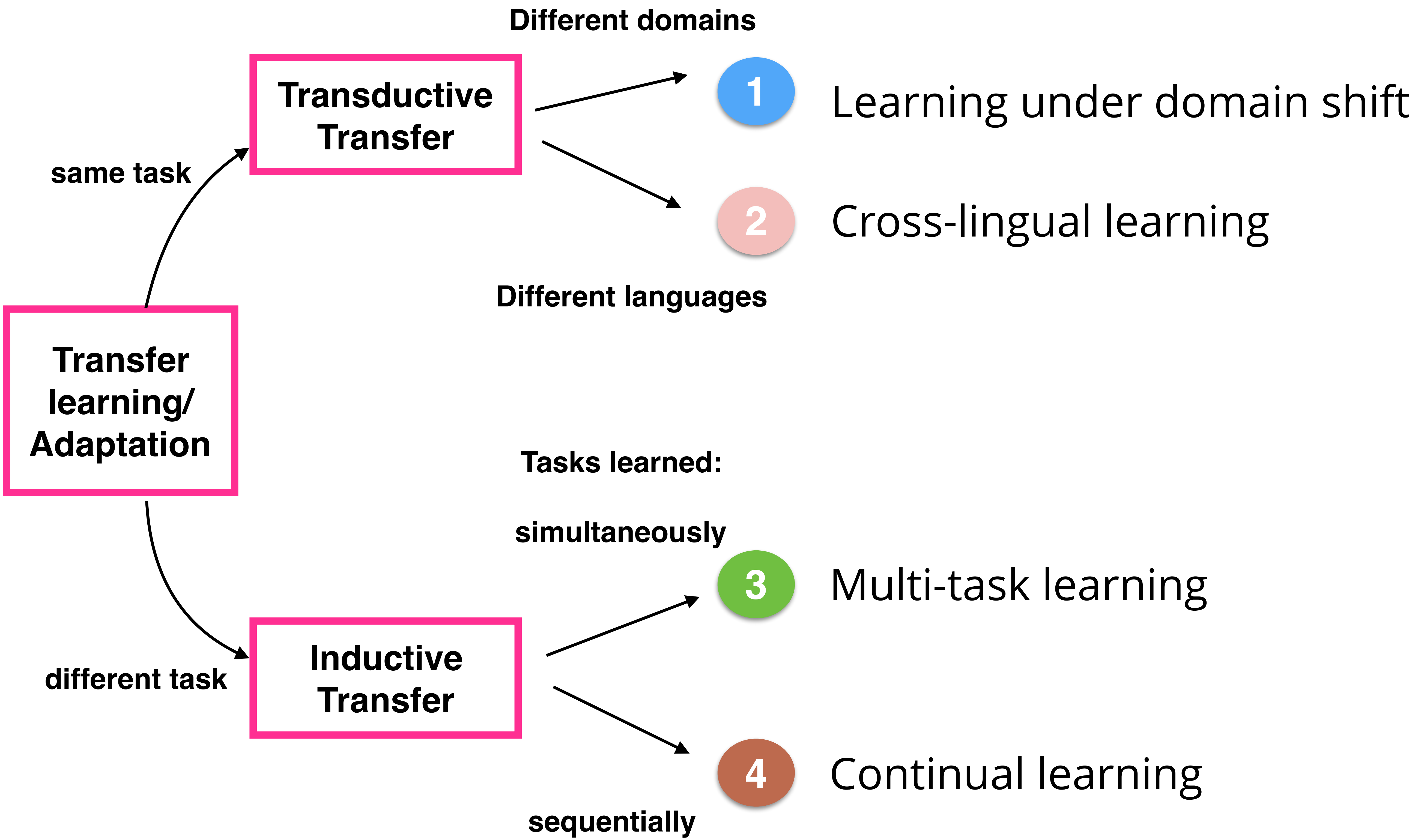
Model B

Adaptation / Transfer Learning



Adapted from Ruder (2019)

Transfer Learning - Details (1/2)



Transfer Learning - Details (2/2)

- $P(\mathcal{X}_{src}) \neq P(\mathcal{X}_{trg})$ different text types

Domain Adaptation (DA)

- $\mathcal{X}_{src} \neq \mathcal{X}_{trg}$ different languages

Cross-lingual Learning (CL)

- $\mathcal{Y}_{src} \neq \mathcal{Y}_{trg}$ different tasks

Multi-task Learning (MTL)

- Timing/Availability of tasks

Notation:

- **Domain** $\mathcal{D} = \{\mathcal{X}, P(\mathcal{X})\}$
  where $\mathcal{X}$ is the feature space, $P(\mathcal{X})$ prob. over e.g., BOW
- **Task** $\mathcal{T} = \{\mathcal{Y}, P(\mathcal{Y}|\mathcal{X})\}$
  where $\mathcal{Y}$ is the label space (e.g., +/-)

Roadmap

1. Domains: Learning to select data
2. Languages: Cross-lingual learning
3. Multi-task learning

Learning to select data for transfer learning with Bayesian optimization

Sebastian Ruder and Barbara Plank

EMNLP 2017

Data Setup:
Multiple Source Domains

Target domain

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

Training examples | Selection policy | Sorted examples

$$x_1$$
$$x_2$$
$$\vdots$$
$$x_m$$
$$\vdots$$
$$x_n$$

$$\xrightarrow{S = \phi(x)^\top w}$$

$$\left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} m$$

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

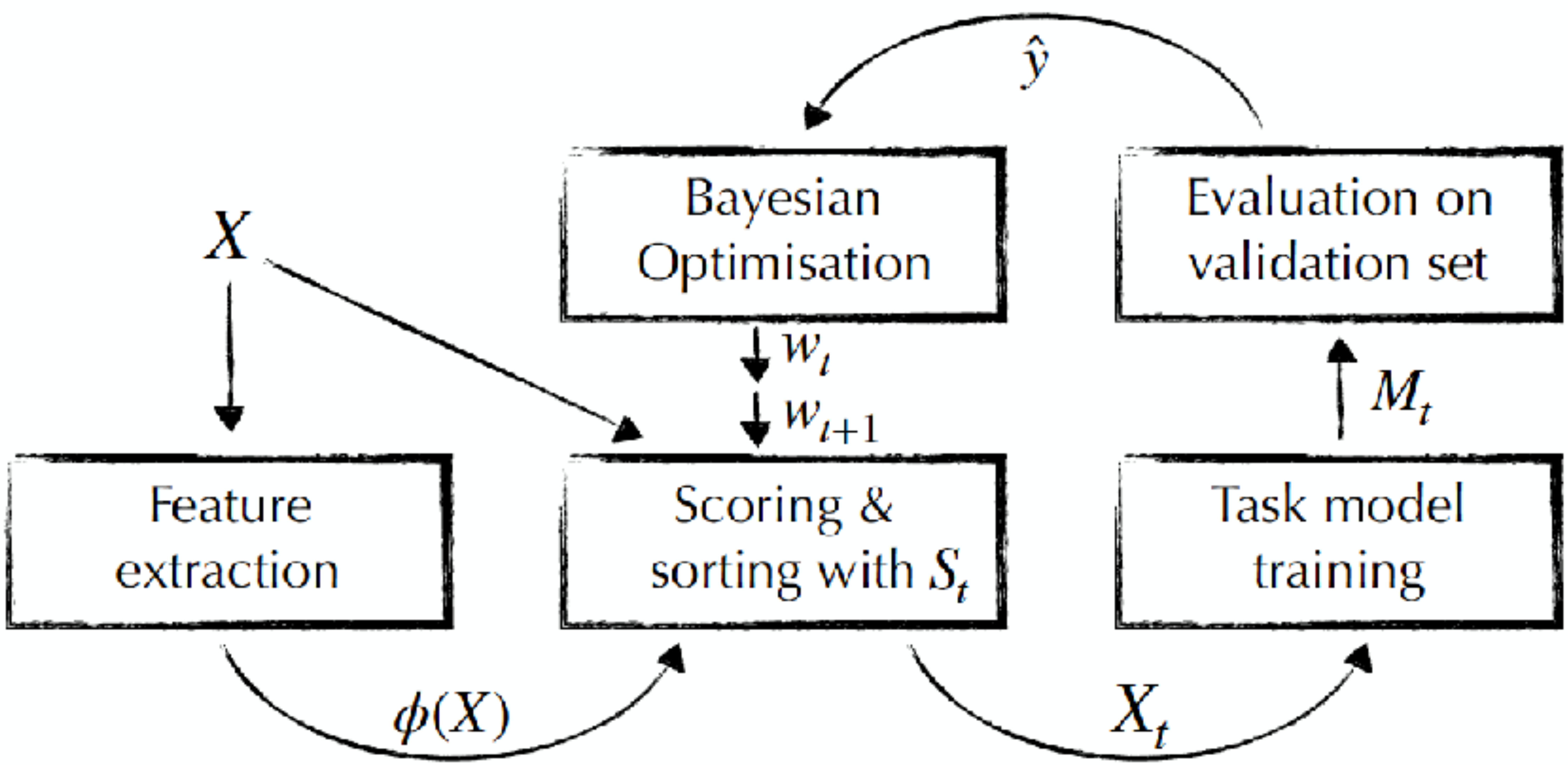
Tsvetkov, Y., Faruqui, M., Ling, W., & Dyer, C. (2016). Learning the Curriculum with Bayesian Optimization for Task-Specific Word Representation Learning. In *Proceedings of ACL 2016*.

Bayesian Data Selection Policy

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

learned feature weights

different similarity/ diversity features



Features $\phi(X)$

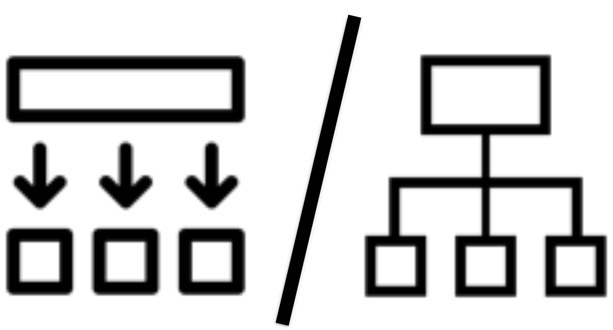
- **Similarity:**
  Jensen-Shannon, Rényi div, Bhattacharyya dist, Cosine sim, Euclidean distance, Variational dist
  - **Representations**:
    Term distributions, Topic distributions, Word embeddings (Plank, 2011)

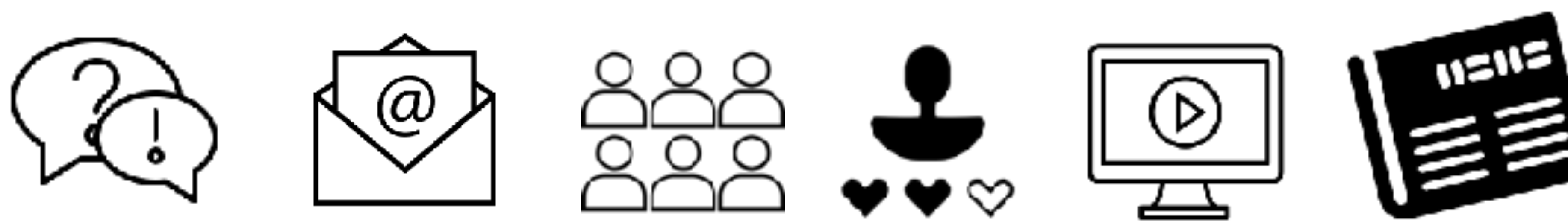
- **Diversity**: #types, TTR, Entropy, Simpson's index, Rényi entropy, Quadratic entropy

Data & Tasks

Three tasks: **Domains:**

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

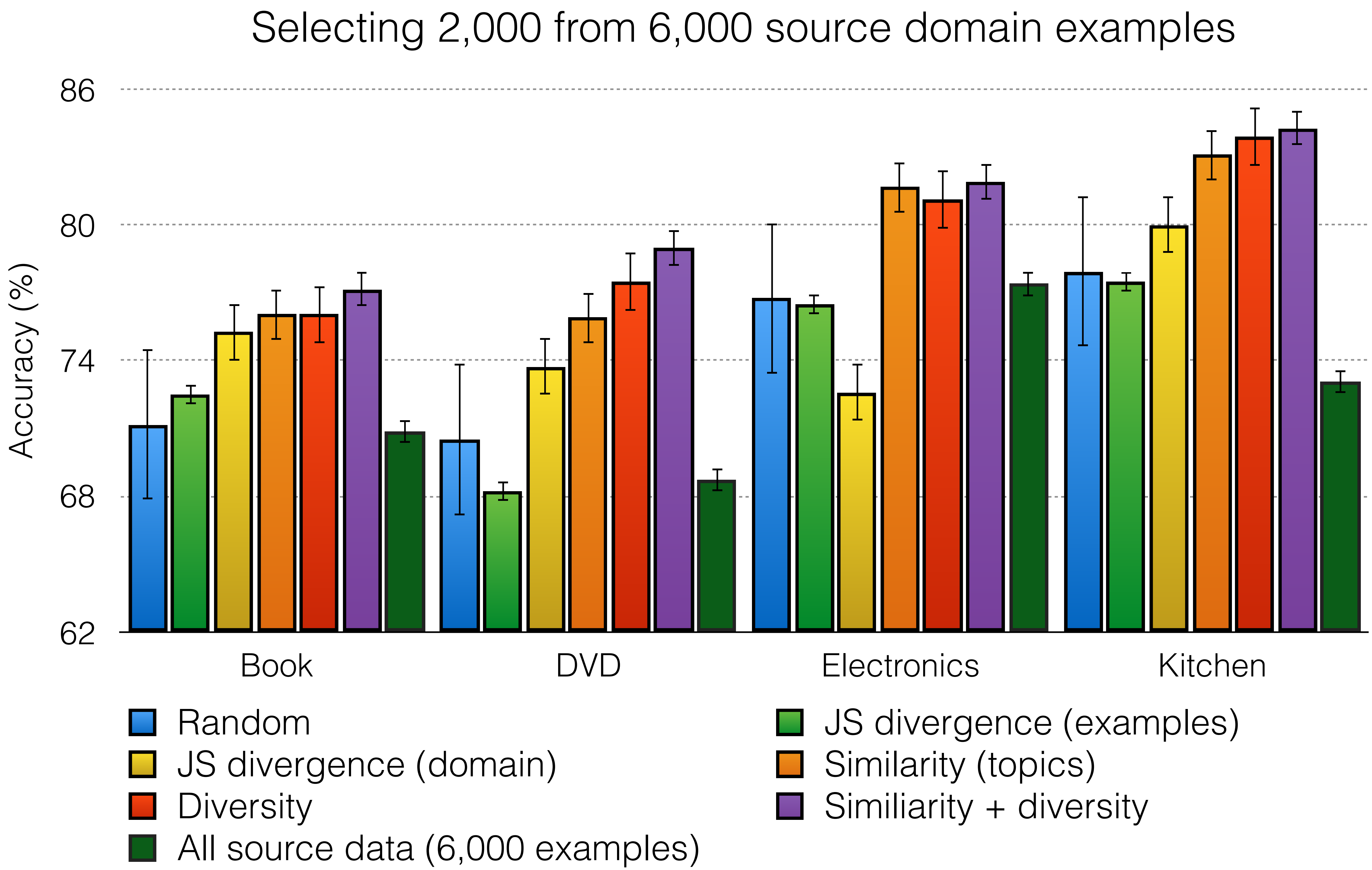




POS tagging and dependency parsing on SANCL 2012 (Petrov and McDonald, 2012)

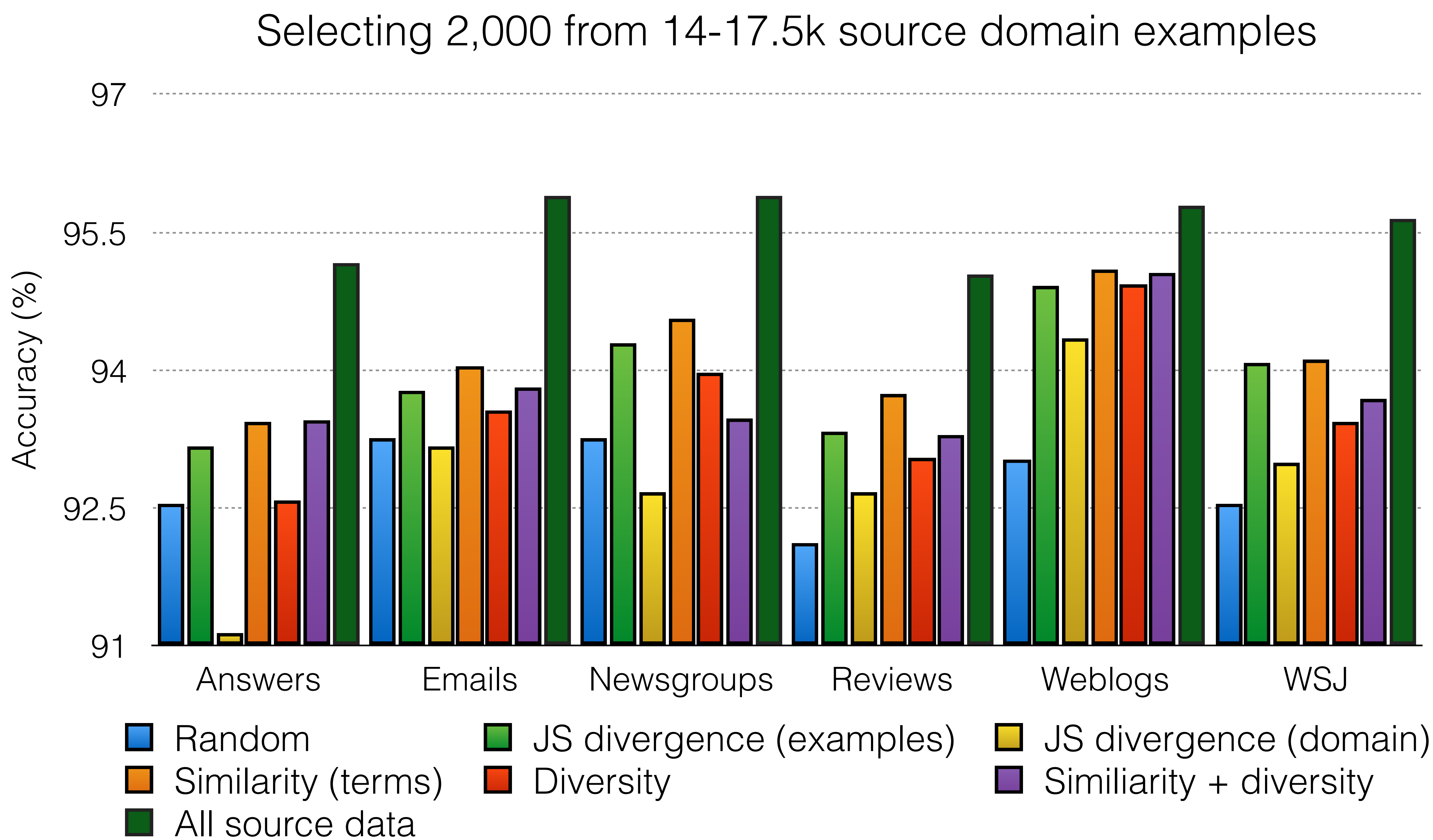
Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL 2007*.
Petrov, S., & McDonald, R. (2012). Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic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

Sentiment Analysis Results



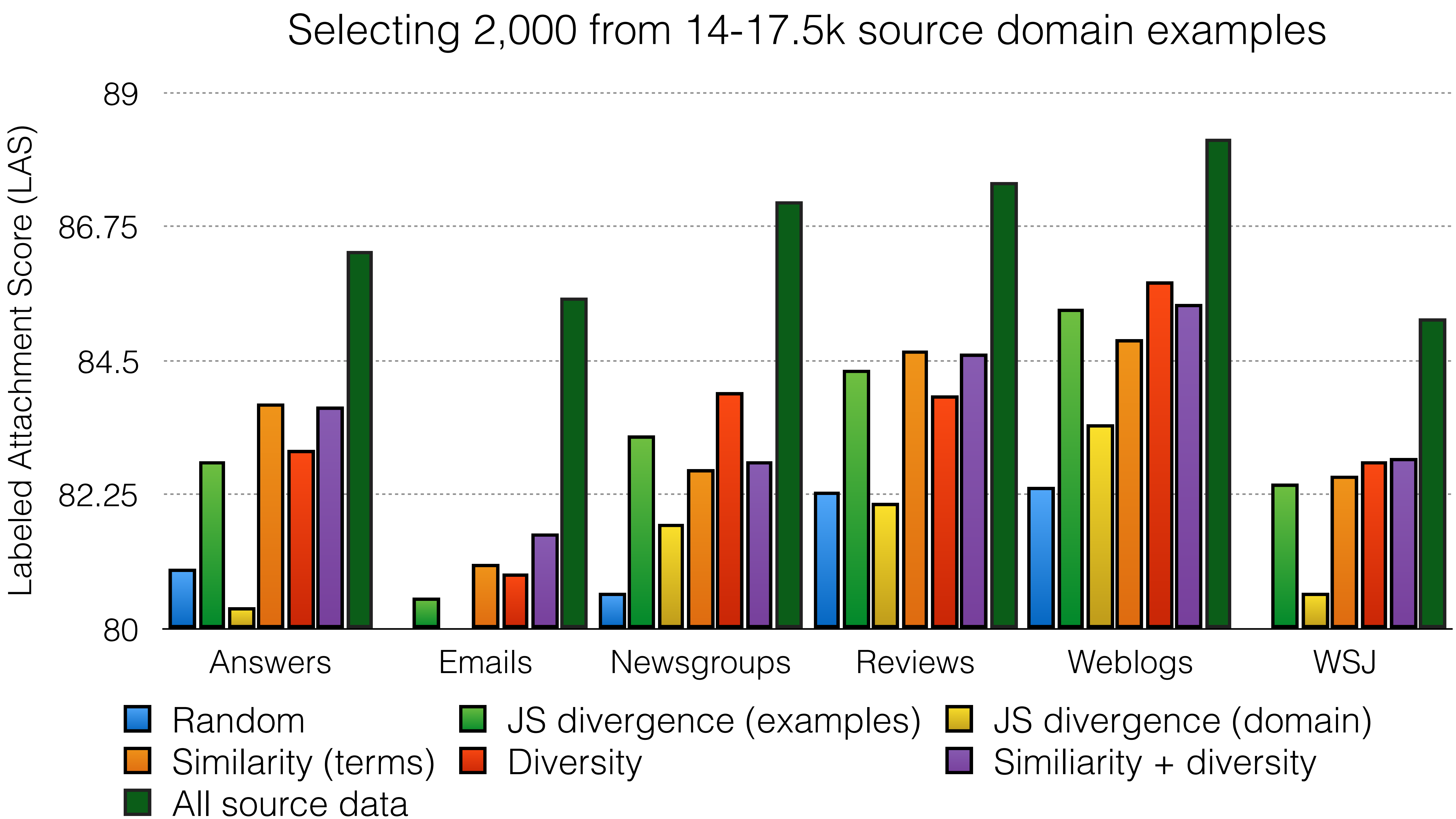
- Selecting relevant data is useful when domains are very different.

POS Tagging Results



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

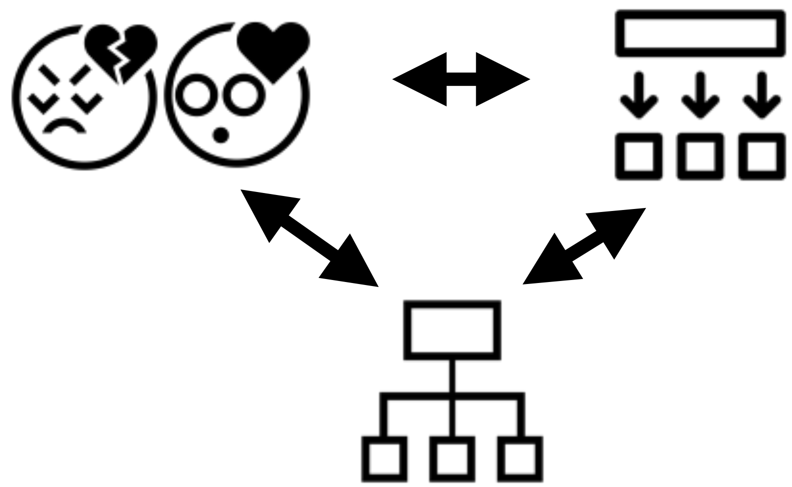
Dependency Parsing Results



(BIST parser, Kiperwasser & Goldberg, 2016)

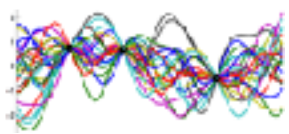
Do the weights transfer?

Cross-task transfer



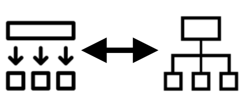
		Target tasks		
Feature set	$\mathcal{T}_S$	**POS**	**Pars**	**SA**
Sim	POS	93.51	83.11	74.19
Sim	Pars	92.78	83.27	72.79
Sim	SA	86.13	67.33	79.23
Div	POS	93.51	83.11	69.78
Div	Pars	93.02	83.41	68.45
Div	SA	90.52	74.68	79.65
Sim+div	POS	93.54	83.24	69.79
Sim+div	Pars	93.11	83.51	72.27
Sim+div	SA	89.80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



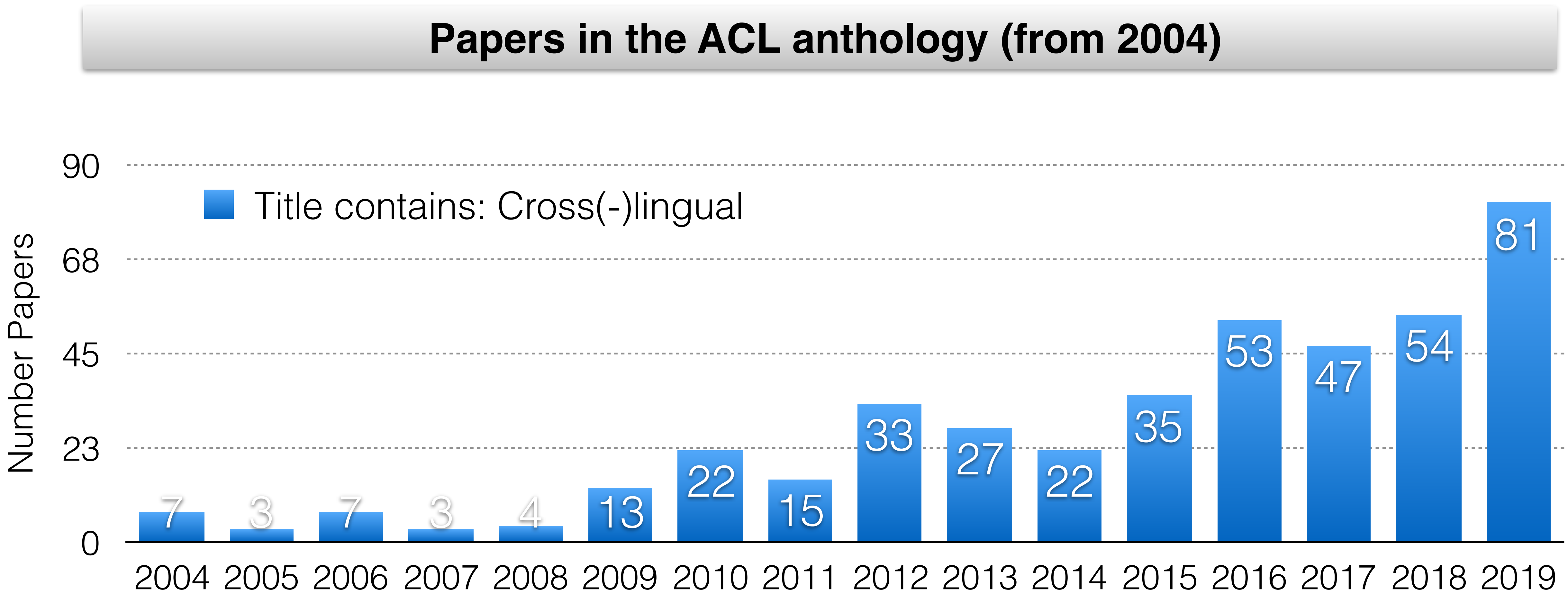
- The learned policy **transfers** (to some extent) across models, tasks, and domains

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

Roadmap

1. Domains: Learning to select data
2. Languages: Cross-lingual learning
3. Multi-task 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

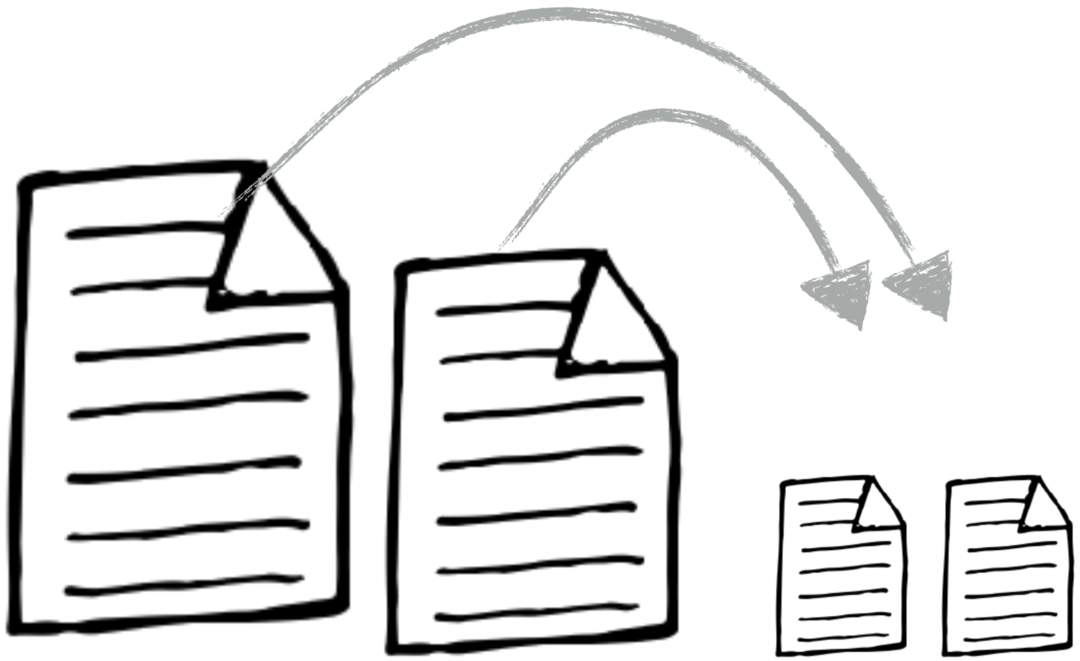
We want to process **all** languages.
Most of them are severely under-resourced.

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

Approaches

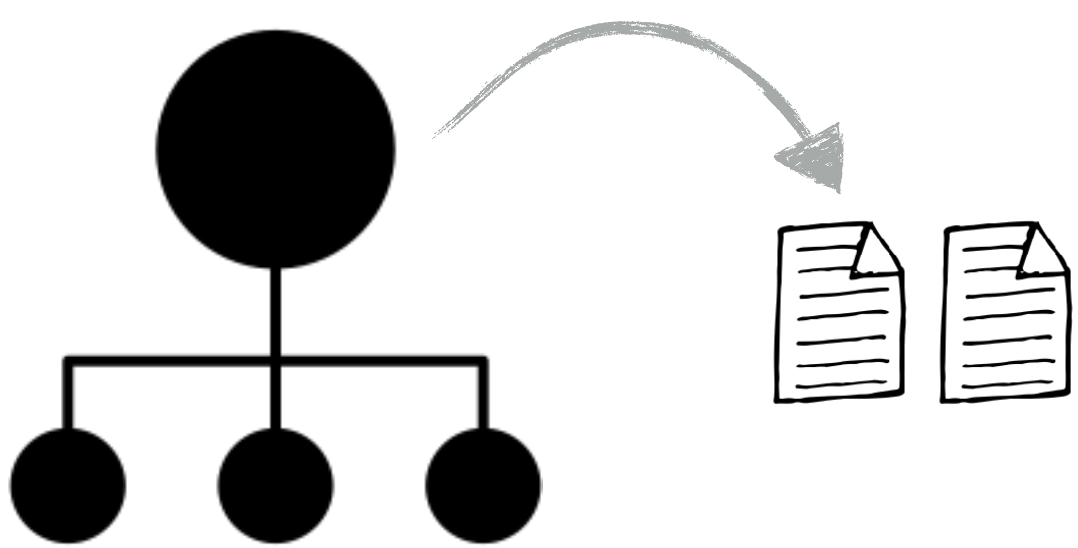
1

2



annotation transfer
(annotation projection)

3



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

Multi-Source Annotation Projection for Dependency Parsing

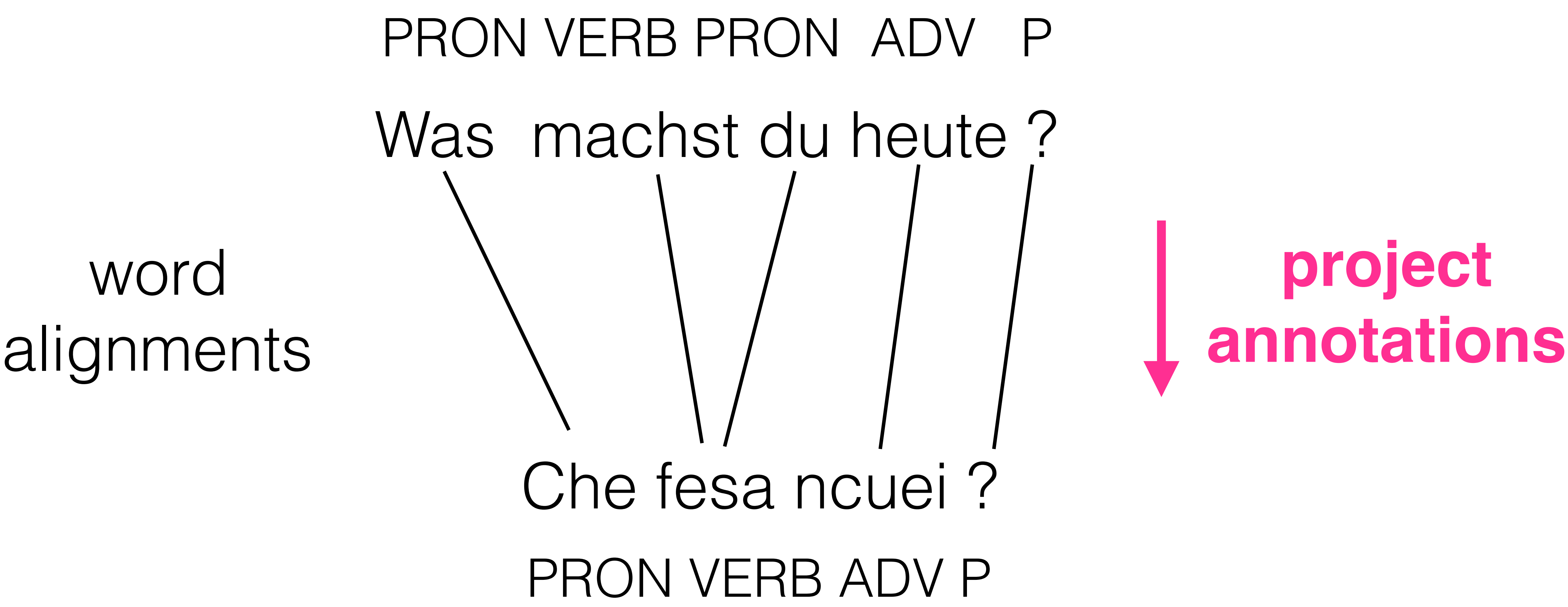
Multilingual Projection for Parsing Truly Low-Resource Languages

Željko Agić[♡] **Anders Johannsen**[♡] **Barbara Plank**[♡♣]
Héctor Martínez Alonso[♡♠] **Natalie Schluter**[♡◇] **Anders Søgaard**[♡]
♡ Center for Language Technology, University of Copenhagen, Denmark
♣ Center for Language and Cognition, University of Groningen, The Netherlands
♠ Univ. Paris Diderot, Sorbonne Paris Cité – Alpage, INRIA, France
◇ MobilePay, Copenhagen, Denmark
{zeljko.agic,soegaard}@hum.ku.dk

TACL, 2016

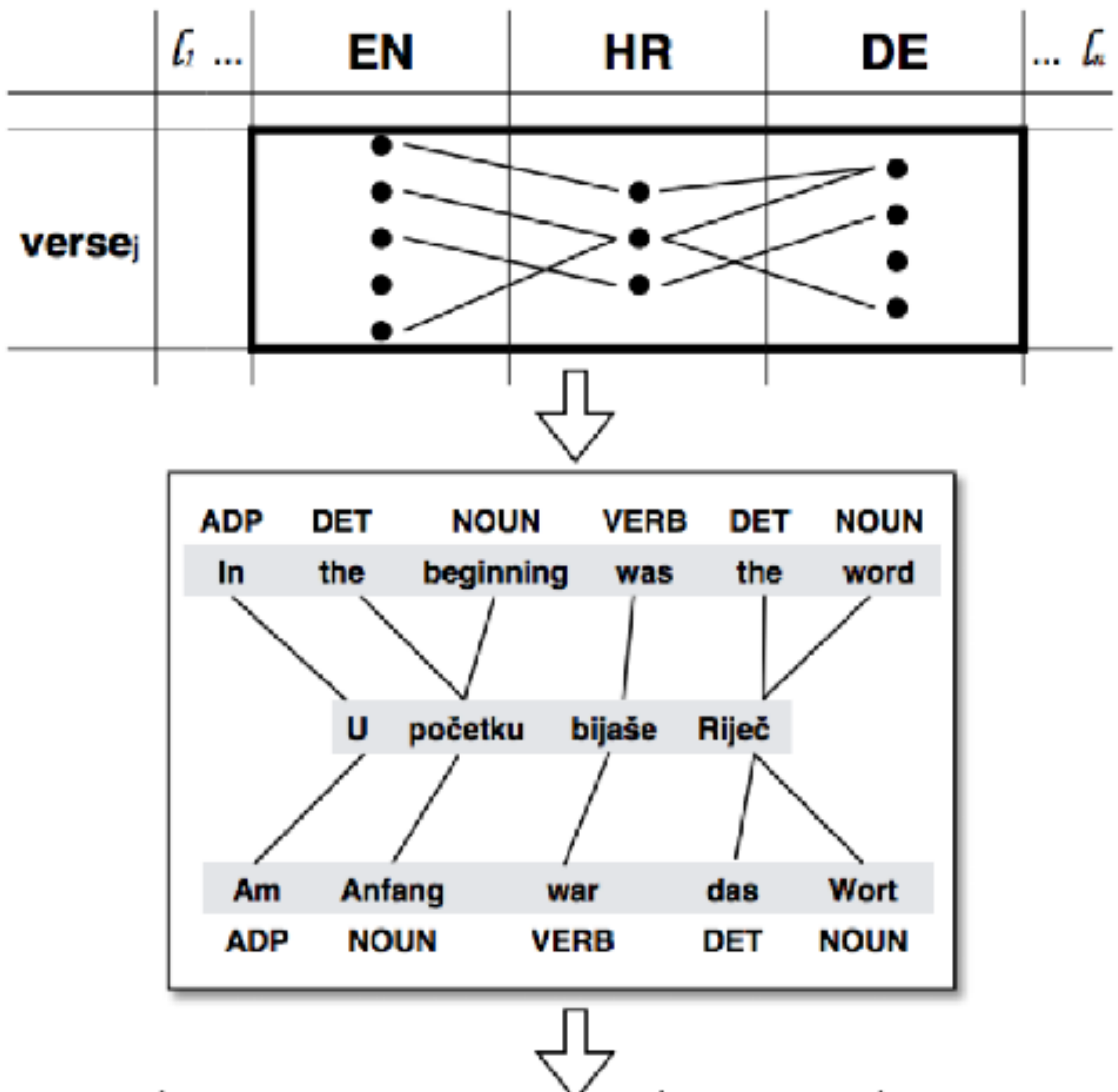
1

Annotation projection



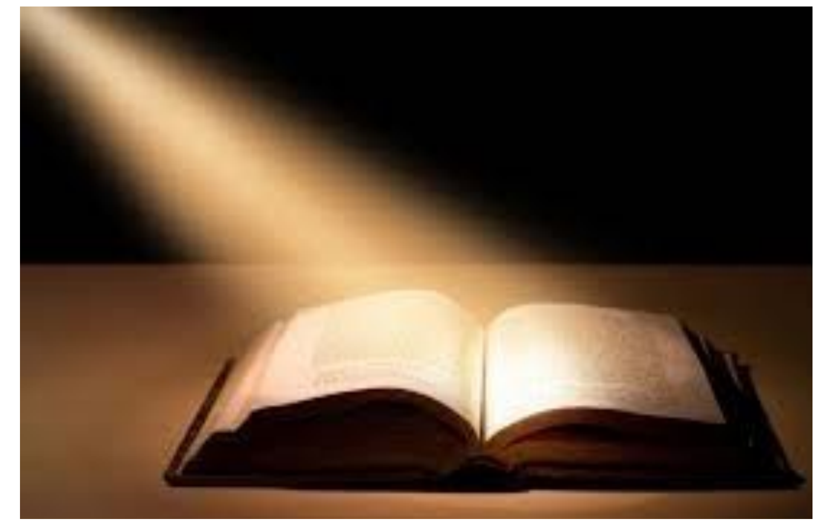
e.g., Hwa et al. (2005)

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

Bible:



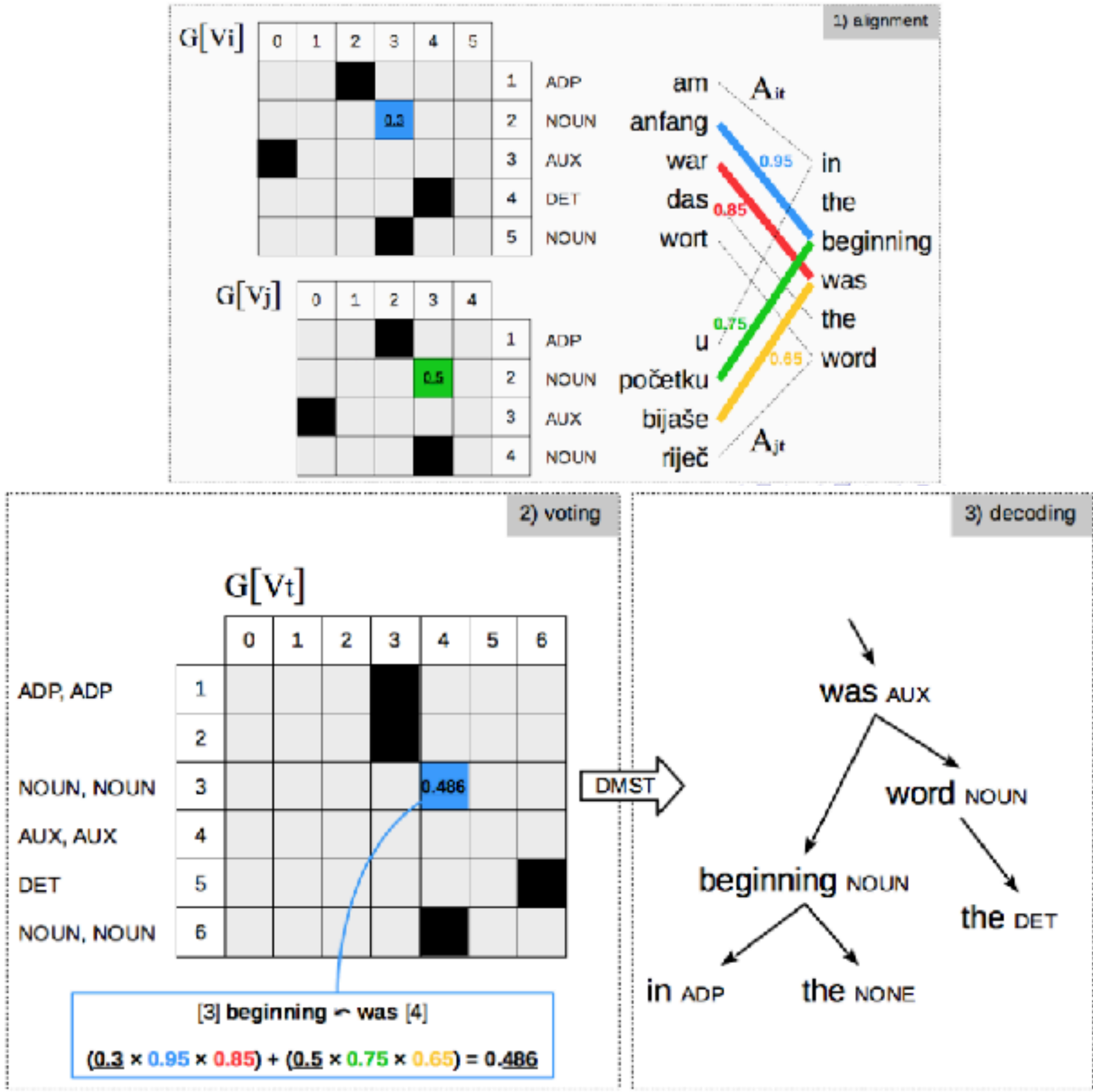
← 100 →

(data x languages)

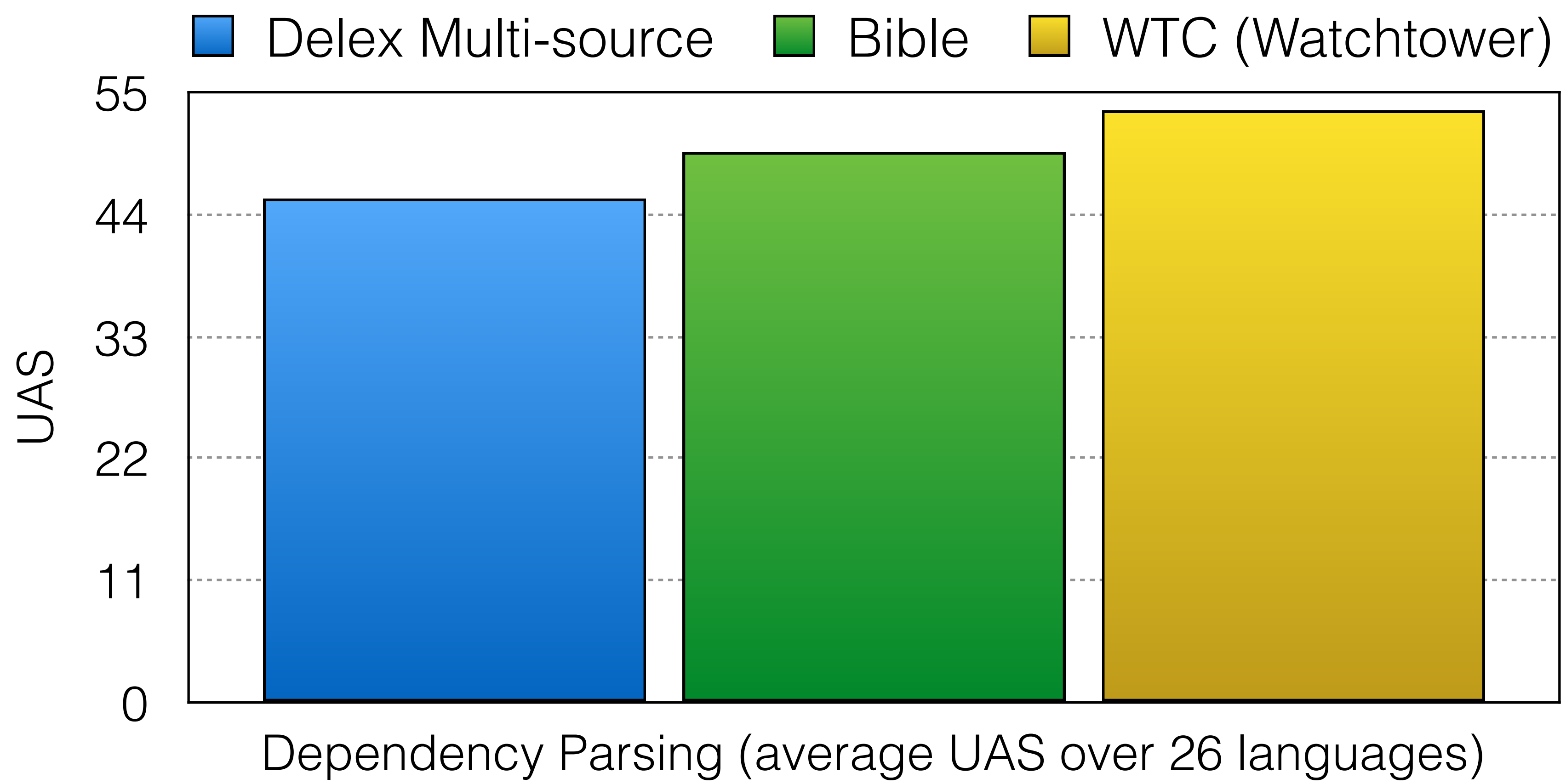
- **Project** from 21 source languages

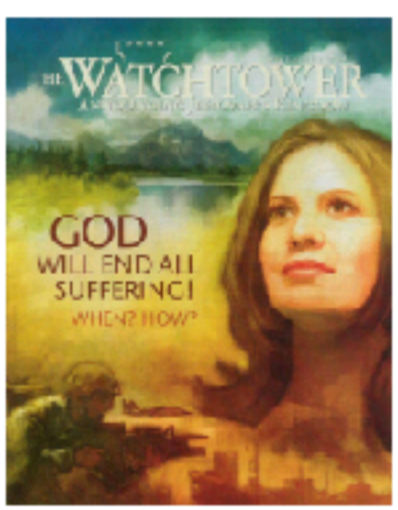
(Agić et al., 2015; 2016)

Approach: Projecting dependencies



Results

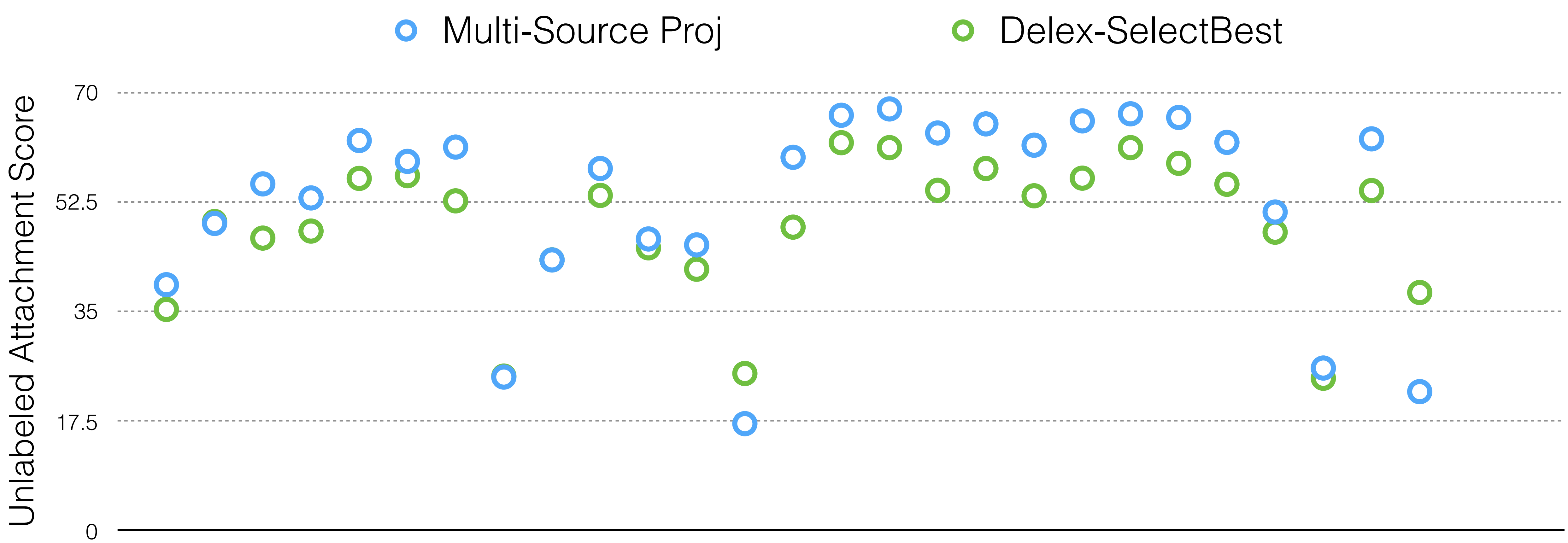




EBC: *hath, saith, hast, spake, yea, cometh, iniquity, wilt, smote, shew, begat, doth, lo, hearken, thence, verily, neighbour, goeth, shewed, giveth, smite, didst, wherewith, knoweth, night*	
WTC: *bible, does, however, says, today, during, show, human, later, important, really, humans, meetings, personal, states, future, fact, relationship, result, attention, someone, century, attitude, article, different*	

Table 1: The 25 most frequent words exclusive to the English Bible or Watchtower.

Best single source



- Single best can be better than multi-source
- Typologically closest language is not always the best (Lynn et al., 2014) (Indonesian is best for Irish in delexicalized transfer)
  - Similar recent findings on NER

Interim discussion (1/2)

Massively Multilingual Transfer for NER

Afshin Rahimi* **Yuan Li*** **Trevor Cohn**
School of Computing and Information Systems
The University of Melbourne
yuanl4@student.unimelb.edu.au
{rahimia,t.cohn}@unimelb.edu.au

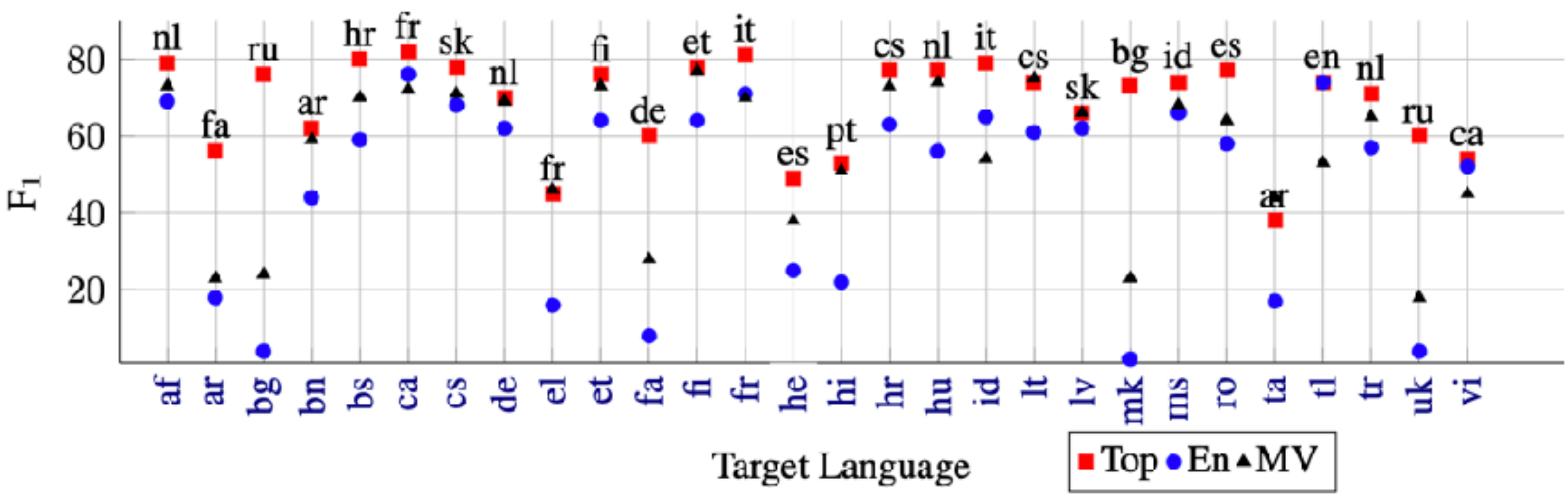
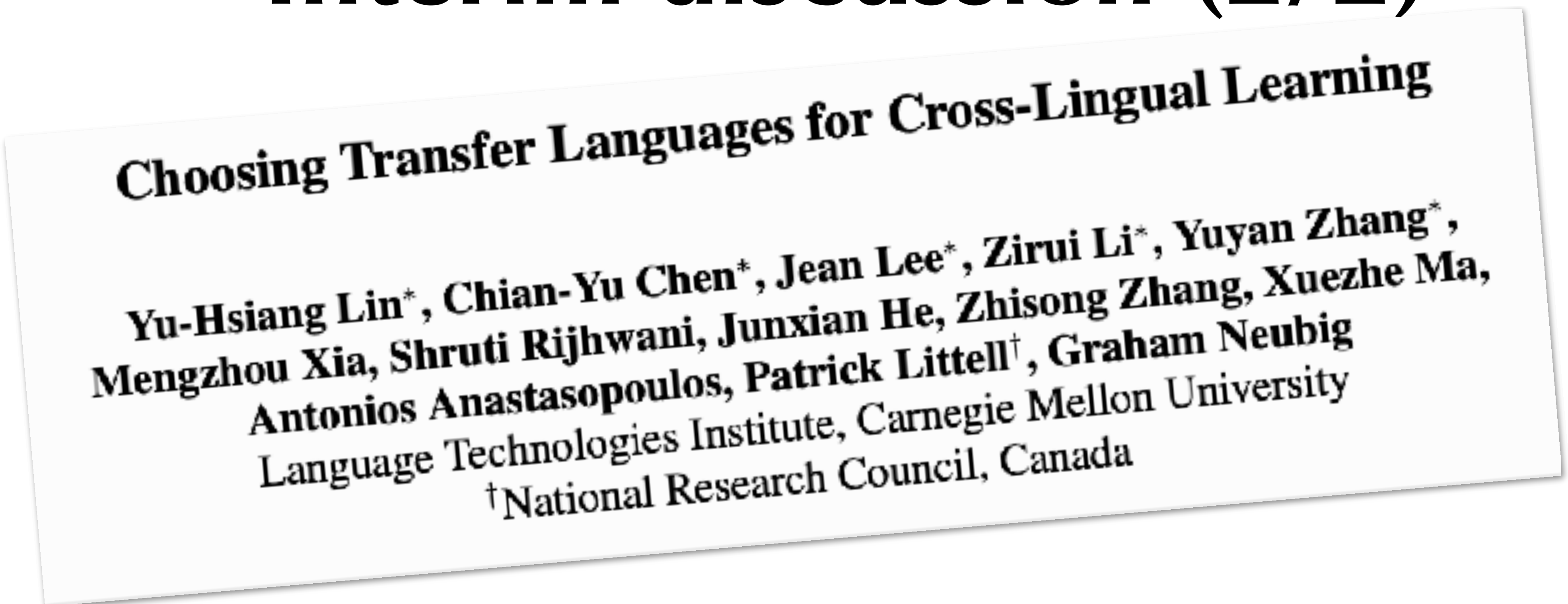


Figure 2: Best source language (■) compared with en (●), and majority voting (▲) over all source languages in terms of F_1 performance in direct transfer shown for a subset of the 41 target languages (x axis). Worst transfer score, not shown here, is about 0. See §3 for details of models and datasets.

How to automatically select the best source parser?

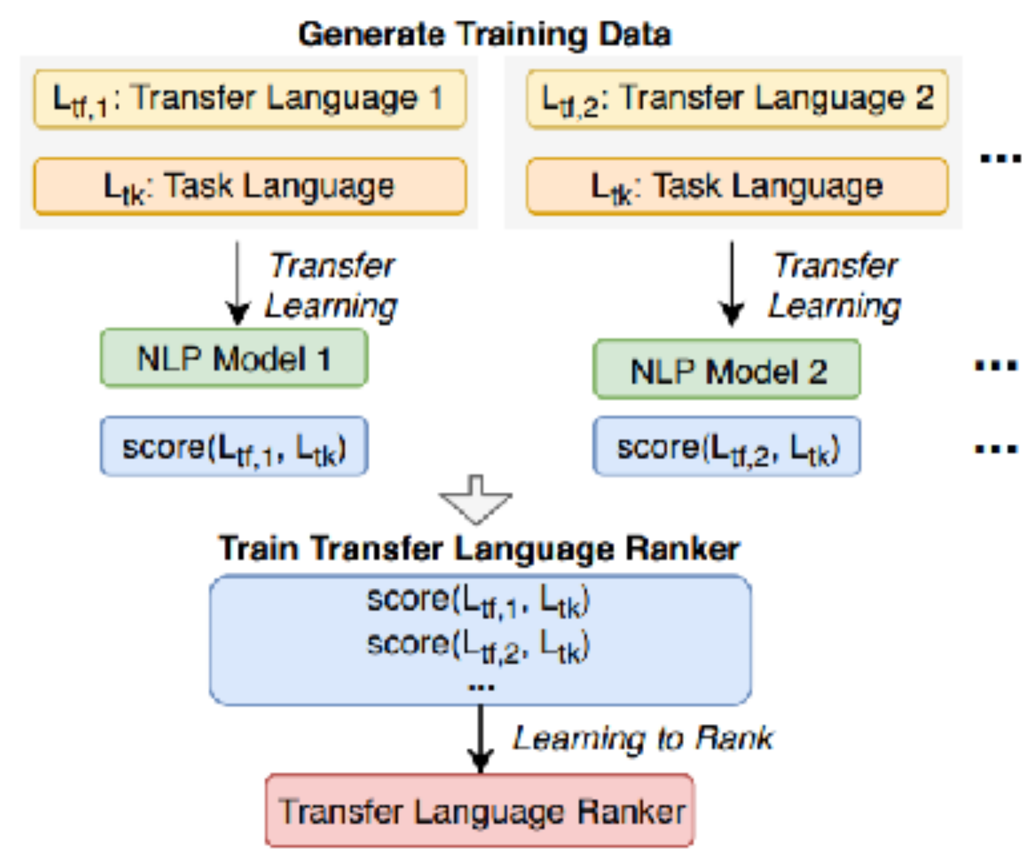
Interim discussion (2/2)


Choosing Transfer Languages for Cross-Lingual Learning

Yu-Hsiang Lin*, Chian-Yu Chen*, Jean Lee*, Zirui Li*, Yuyan Zhang*,
Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma,
Antonios Anastasopoulos, Patrick Littell†, Graham Neubig
Language Technologies Institute, Carnegie Mellon University
†National Research Council, Canada

- Data-dependent features (some similar to Ruder & Plank, 2017) including word/subword overlap, data size

- Data-independent features (Geographic/Genetic distance etc)



Interim discussion: Results

- Evaluation on 4 NLP tasks, including parsing (DEP)
- For Dependency Parsing:
  - geographic > WALS syntactic features
  - Geographic and word overlap most indicate features

	Method	MT	EL	POS	DEP
dataset	word overlap o_w	28.6	30.7	13.4	52.3
	subword overlap o_{sw}	29.2	–	–	–
	size ratio s_{tf}/s_{tk}	3.7	0.3	9.5	24.8
	type-token ratio d_{ttr}	2.5	–	7.4	6.4
ling. distance	genetic d_{gen}	24.2	50.9	14.8	32.0
	syntactic d_{syn}	14.8	46.4	4.1	22.9
	featural d_{fea}	10.1	47.5	5.7	13.9
	phonological d_{pho}	3.0	4.0	9.8	43.4
	inventory d_{inv}	8.5	41.3	2.4	23.5
	geographic d_{geo}	15.1	49.5	15.7	46.4
	LANGRANK (all)	51.1	**63.0**	**28.9**	**65.0**
	LANGRANK (dataset)	**53.7**	17.0	26.5	**65.0**
	LANGRANK (URIEL)	32.6	58.1	16.6	59.6

Table 1: Our LANGRANK model leads to higher average NDCG@3 over the baselines on all four tasks: machine translation (MT), entity linking (EL), part-of-speech tagging (POS) and dependency parsing (DEP).

Labeled data

Overview

Amount of supervision

3 (some) gold annotated data?

4 (Just a couple of rules?)

lexicons?

2

embeddings?

Have parallel data?

1

multi-parallel?

Unlabeled only

Lexical Resources for Low-Resource POS tagging in Neural Times

NoDaLiDa 2019 & EMNLP 2018

Plank & Klerke, 2019; Plank & Agic, 2018

2

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?

Lexicons

Wiktionary

Wiktionary
The free dictionary

Unimorph

Secure | https://unimorph.github.io

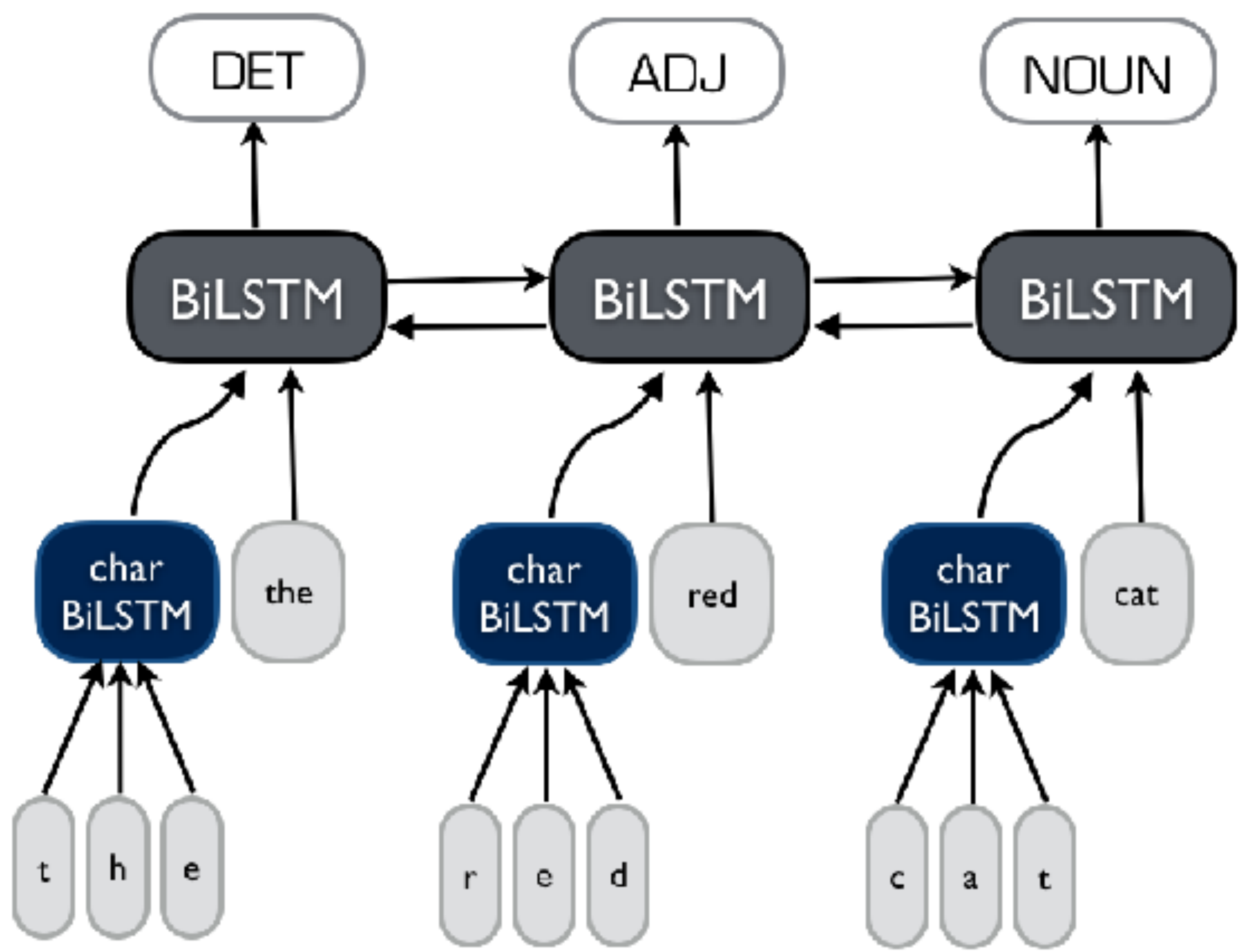
Annotated Languages

The following 51 languages have been annotated according to the UniMorph schema. Missing parts of speech will be filled in soon.

Language	ISO 639-3	Forms	Paradigms	Nouns	Verbs	Adjectives	Source	License
Albanian	sqi	33483	589	✔	✔		W	
Arabic	ara	140003	4134	✔	✔	✔	W	
Armenian	hye	338461	7033	✔	✔	✔	W	
Basque	eus	11889	26		✔			
Bengali	ben	4443	136	✔	✔		W	
Bulgarian	bul	55730	2468	✔	✔	✔	W	
Catalan	cat	81576	1547		✔		W	
Central Kurdish	ckb	22990	274	✔	✔	✔		
Czech	ces	134527	5125	✔	✔	✔	W	
Danish	dan	25503	3193	✔	✔		W	
Dutch	nld	55467	4993		✔	✔	W	

Base bi-LSTM model

- Hierarchical bi-LSTM with word & character embeddings (Plank et al., 2016)

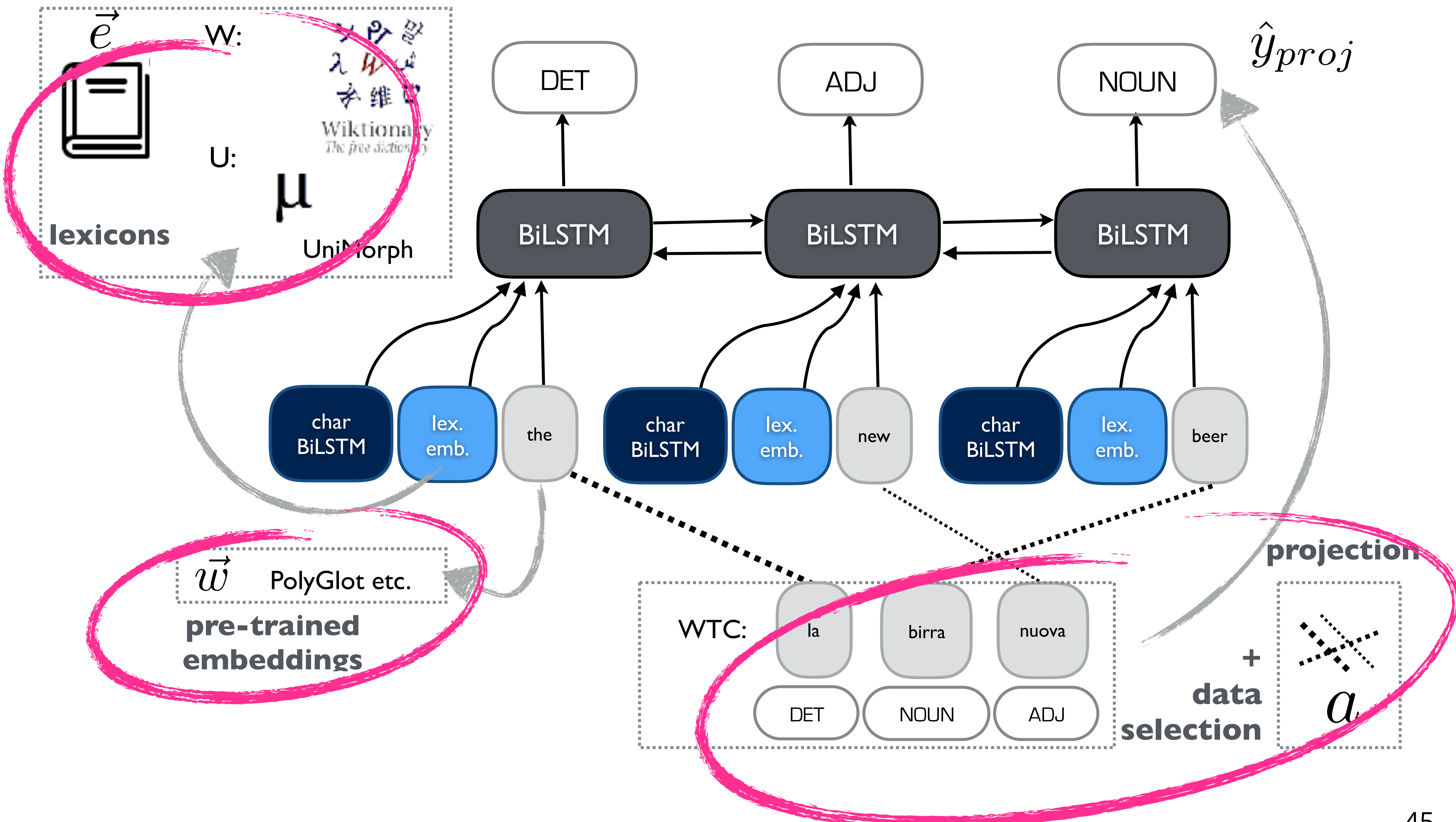


**able* (98% adj in WSJ)

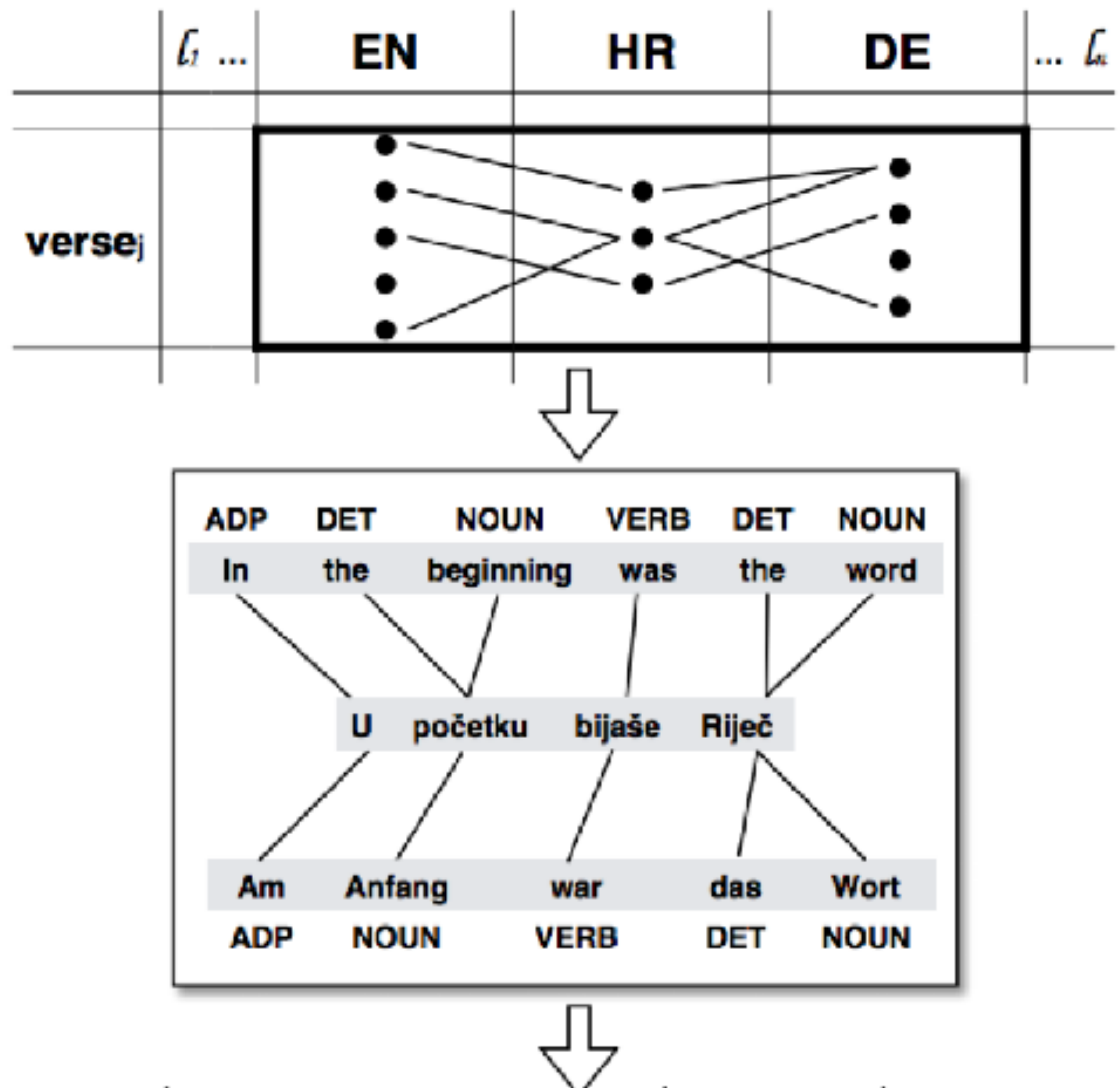
*bi** (85% noun in Danish)

How far do we get with an "all-you-can-get" approach to low-resource POS tagging?

Distant Supervision from Disparate Sources (DsDs)



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)

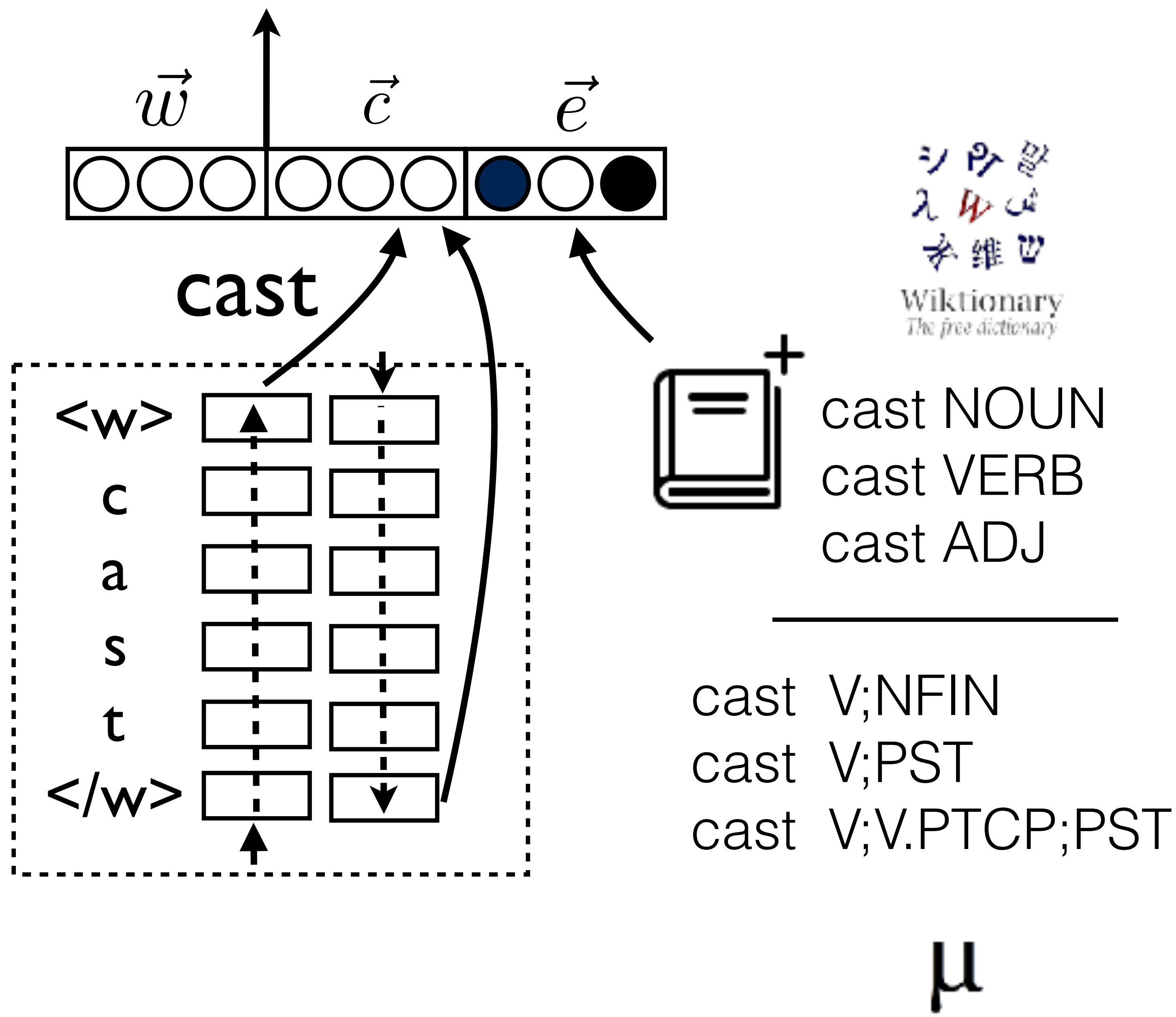
- *Watchtower corpus* (WTC), 300+ languages
- **Project** from 21 source languages
- **Select** instances by word-alignment ***coverage***

Integrating lexical information

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

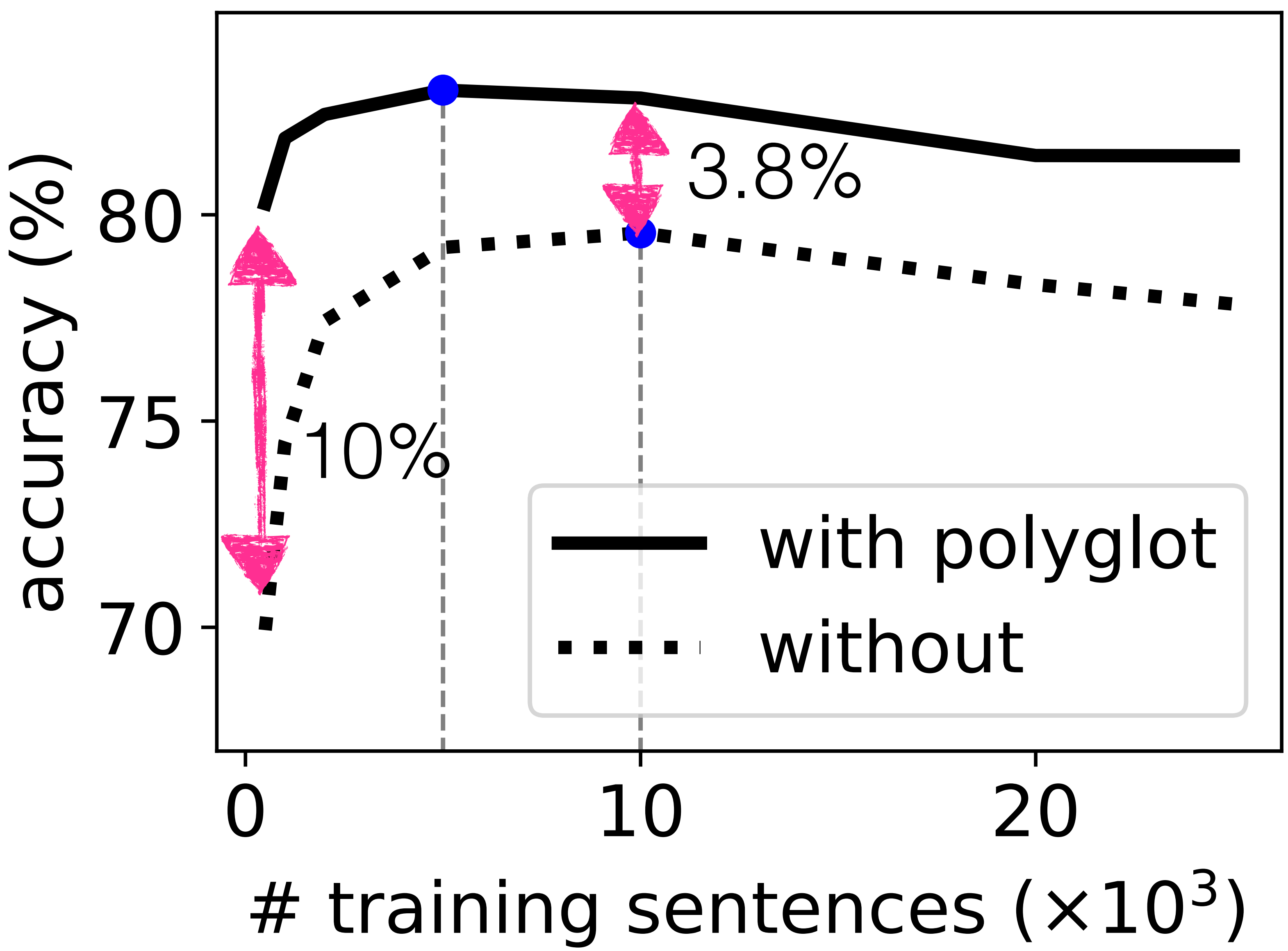
- Our approach:
  embed the lexicon

- **Sources**:
  Wiktionary
  and Unimorph



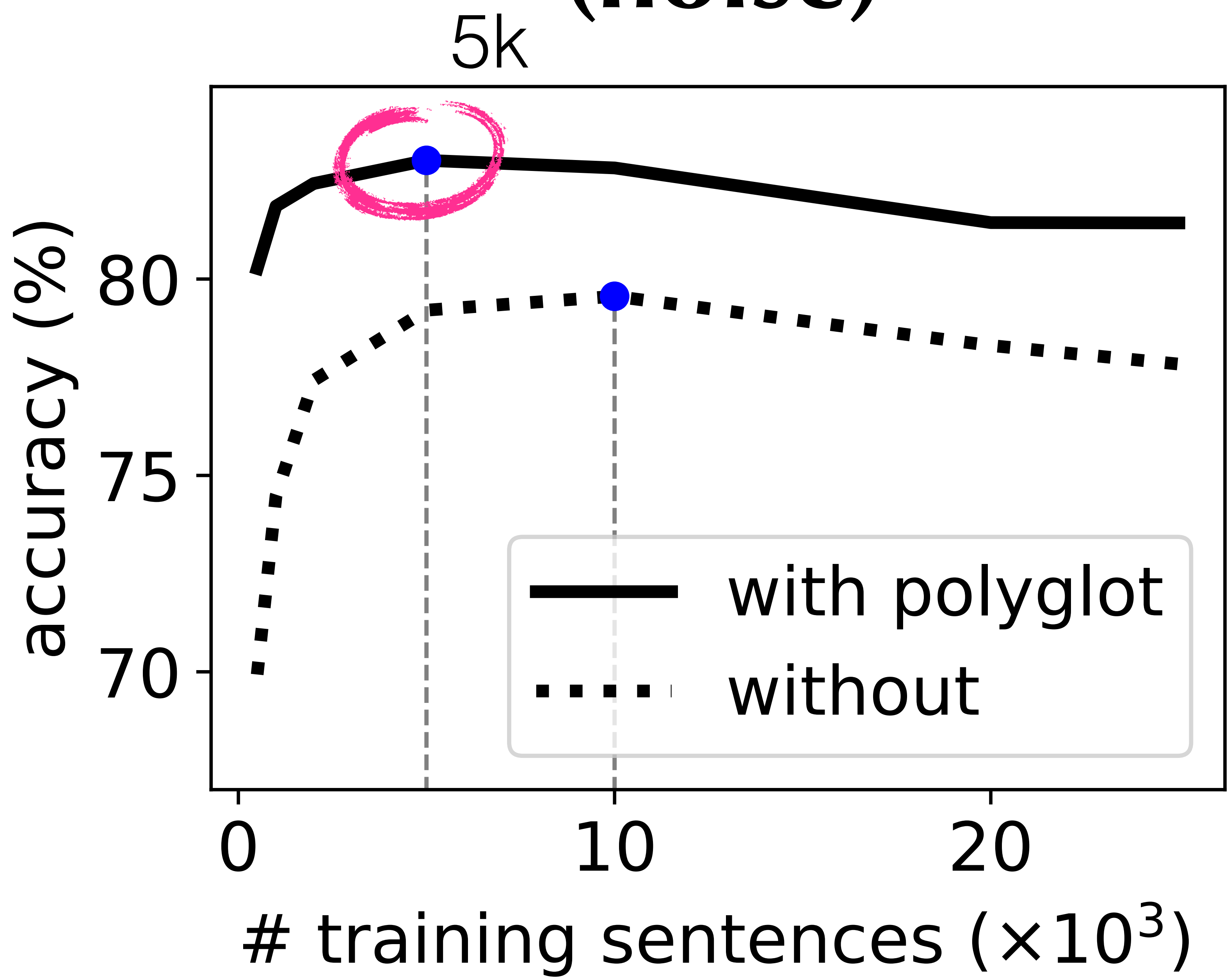
Results

Embedding initialization



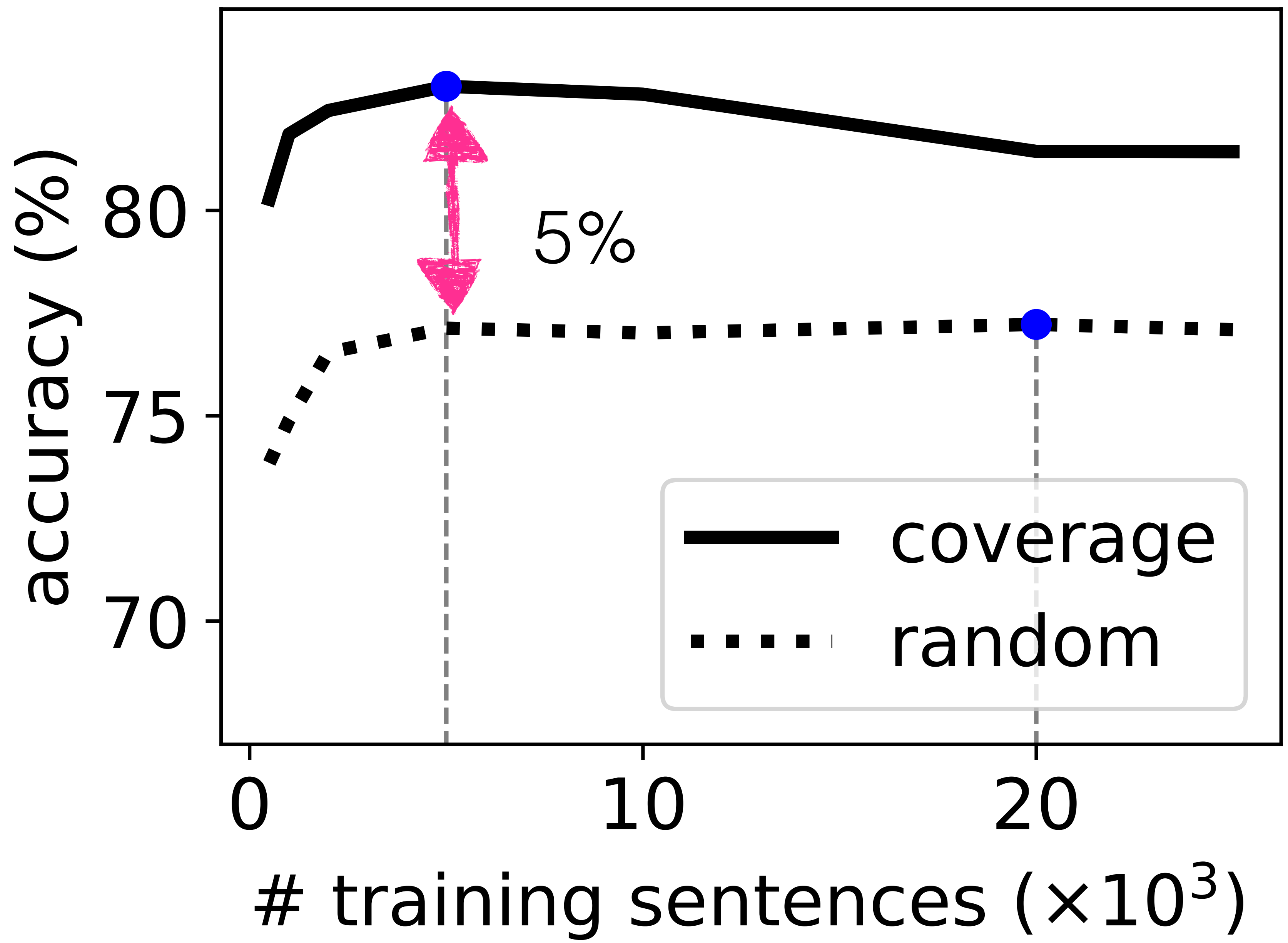
Means over 21 languages
(each point is an average over 3 runs, for random: with 5 random samples)

Less data is better than adding more (noise)



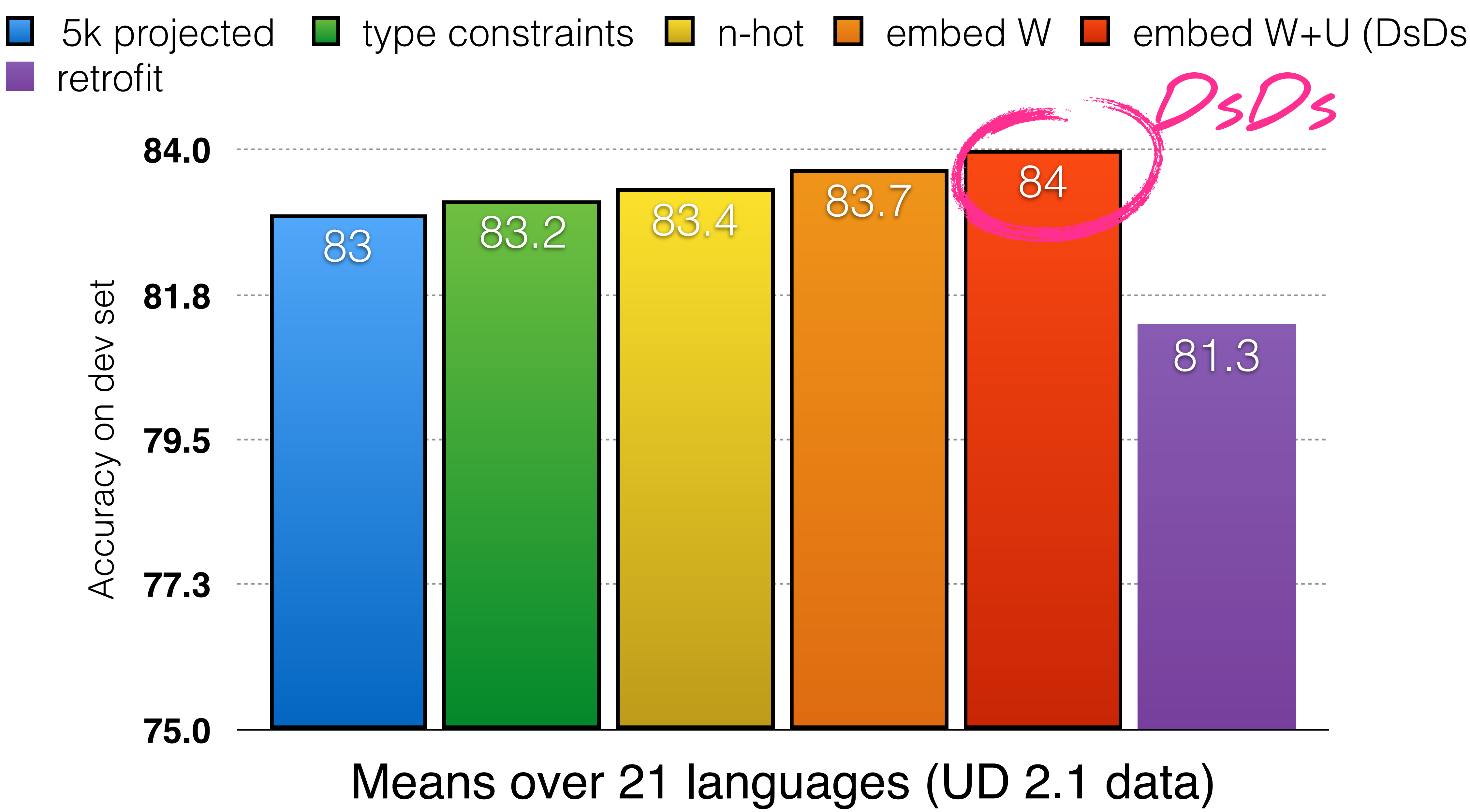
Means over 21 languages
(each point is an average over 3 runs, for random: with 5 random samples)

Coverage-based Data Selection



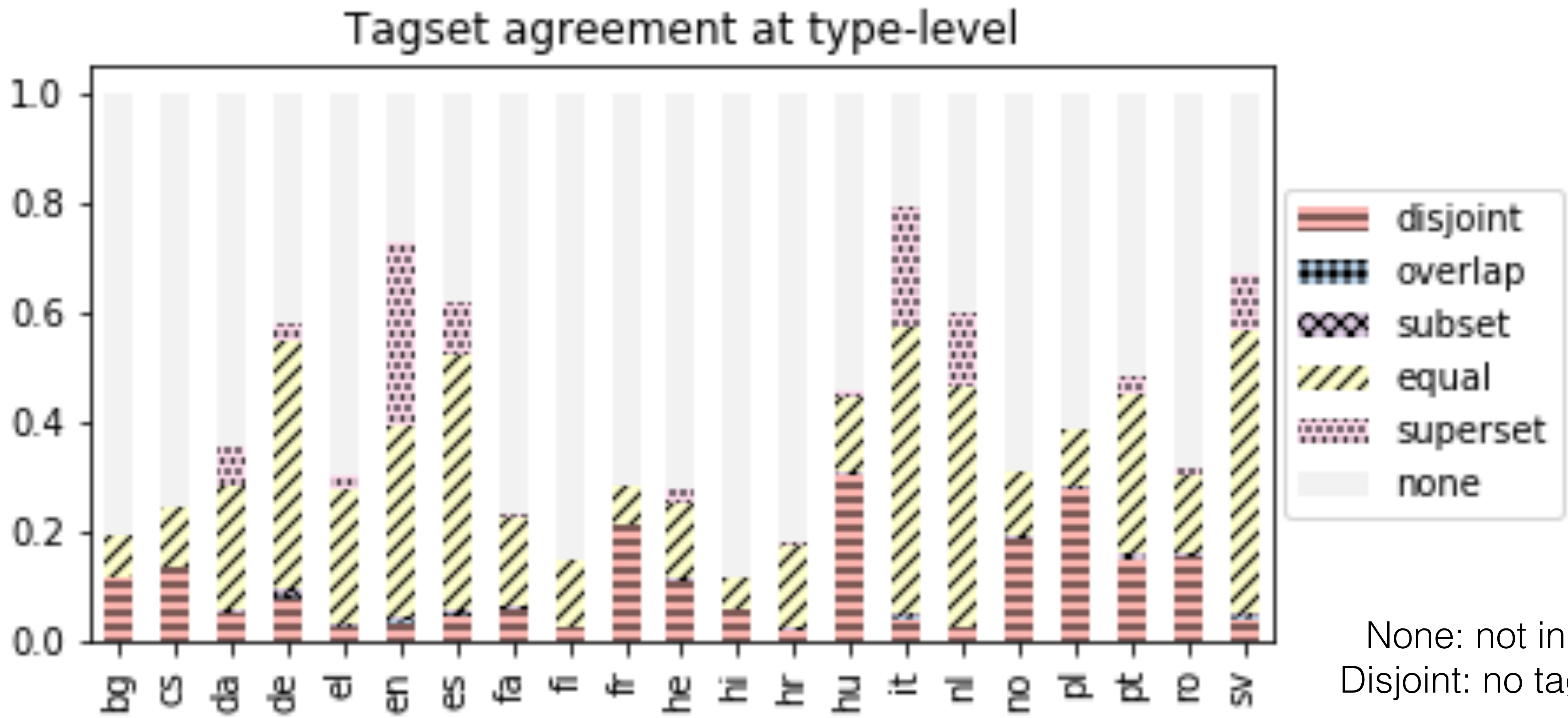
Means over 21 languages
(each point is an average over 3 runs, for random: with 5 random samples)

Inclusion of Lexical information



Analysis: Treebank tag set vs lexicon

(inspired by Li et al., 2012)

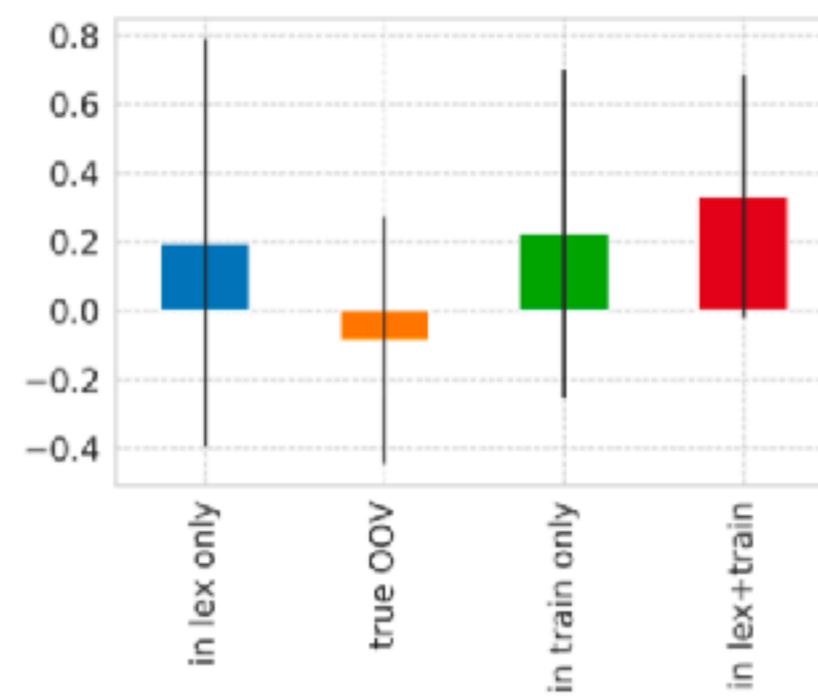


None: not in lexicon
Disjoint: no tag overlap

- 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: Coverage?

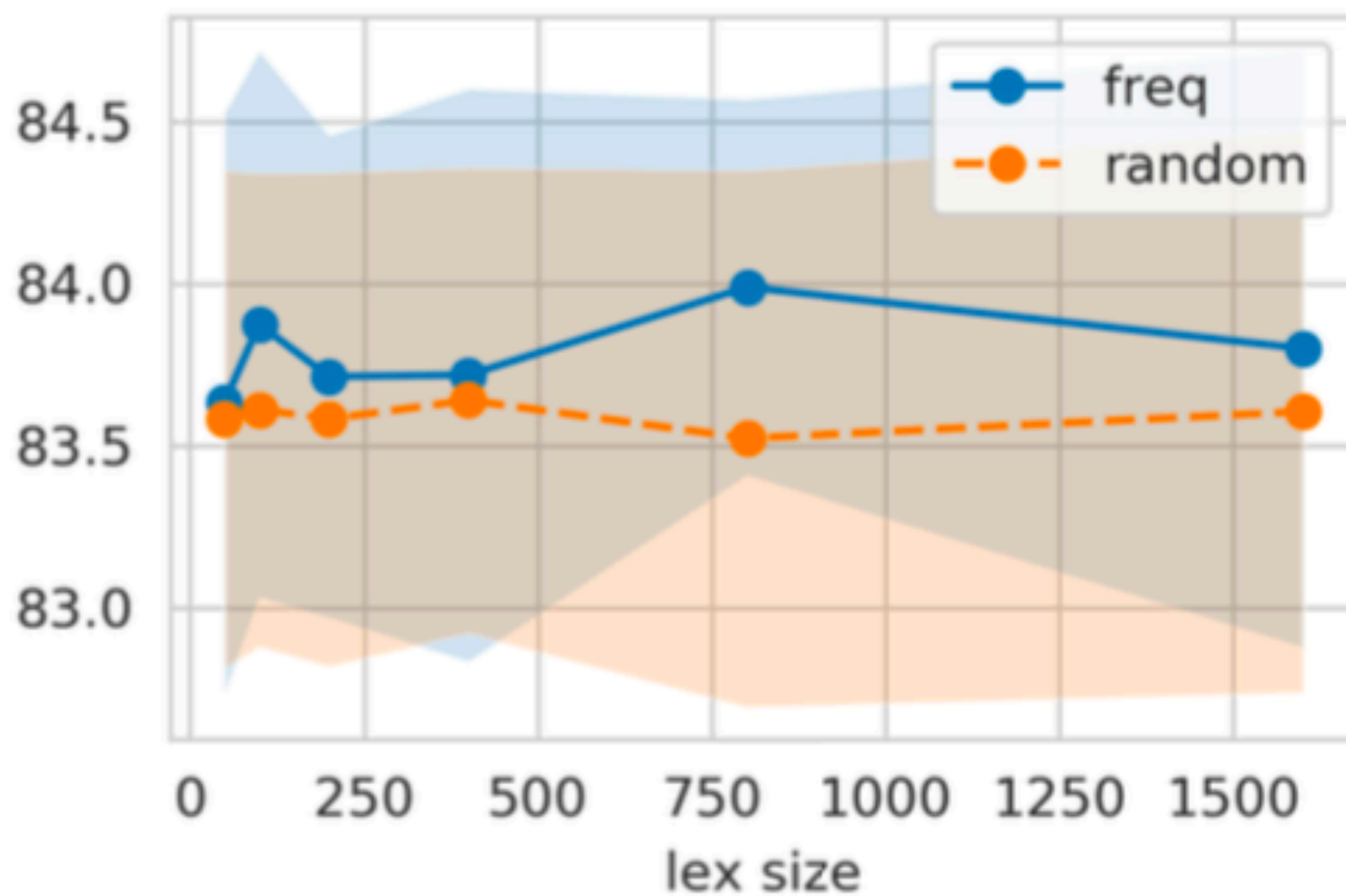
(a) Absolute improvement (delta) vs number of dictionary properties (ρ=0.08).



(b) Absolute improvement per OOV category (21 languages).

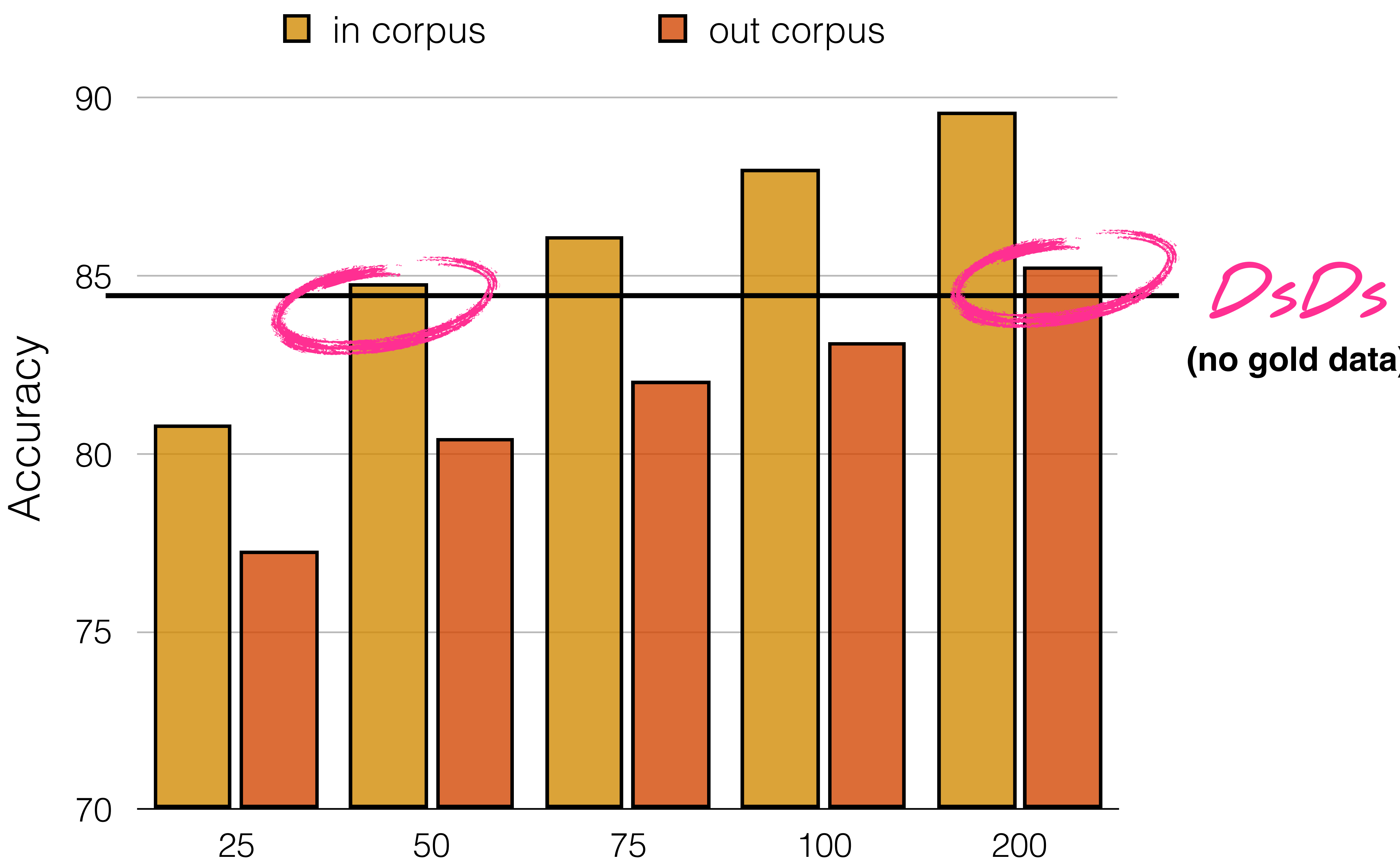
- **Coverage** is only part of the explanation

Analysis: Learning curves over dictionary size



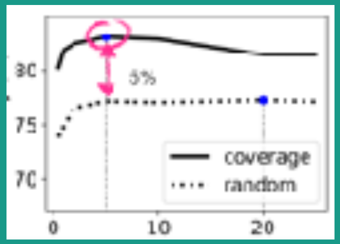
(a) Average effect over 21 languages of high-freq and random dictionaries

How much gold data?

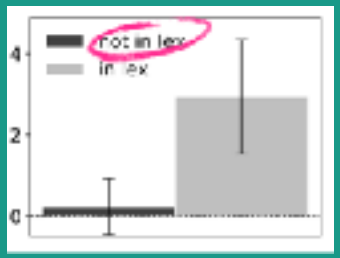


(Means over 18 languages for which we had both in- and out-corpus gold data)

Take-aways



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



2. **Lexical information** improves neural POS tagging beyond the lexicon’s coverage

Our approach so far

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

3 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

to appear in NoDaLiDa 2019

* slide title inspired by Alisa Meechan-Maddon & Joakim Nivre's SyntaxFest presentation :-)

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?
- **RQ3**: How accurate are existing NER systems on Danish?

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)

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)

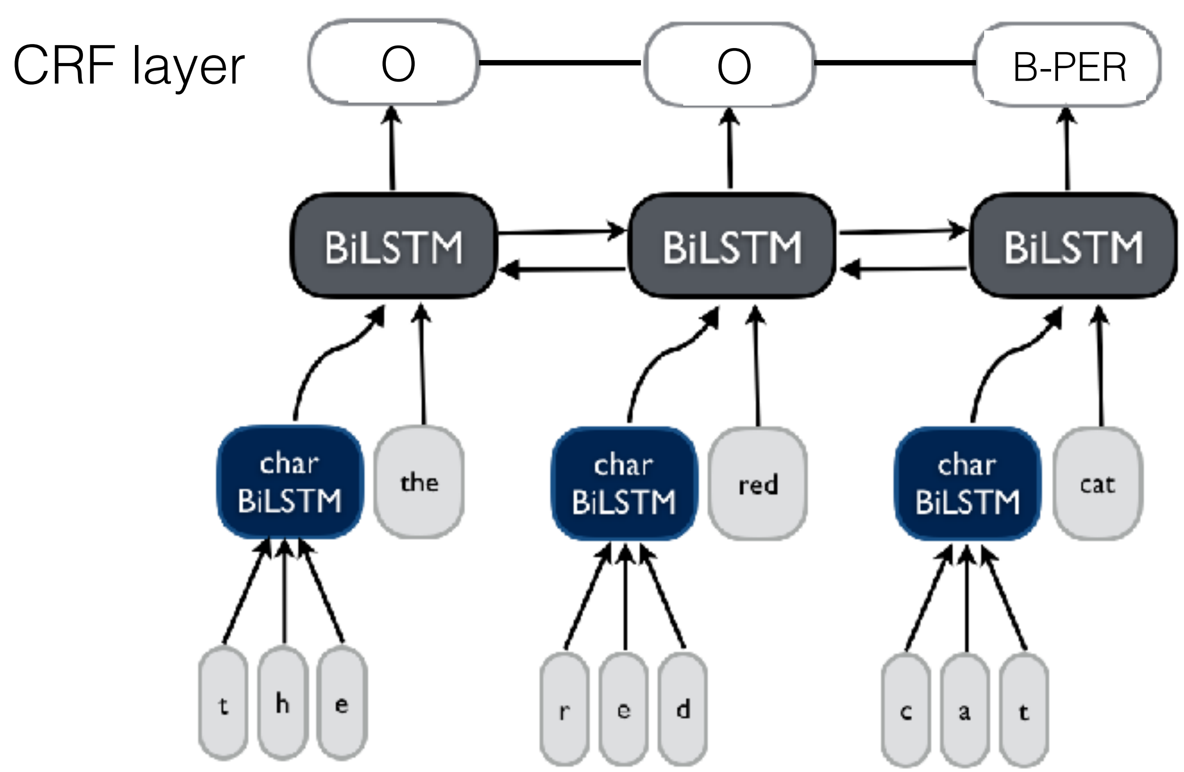
Data Setups: Data & DataAugment

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

Model and Approach

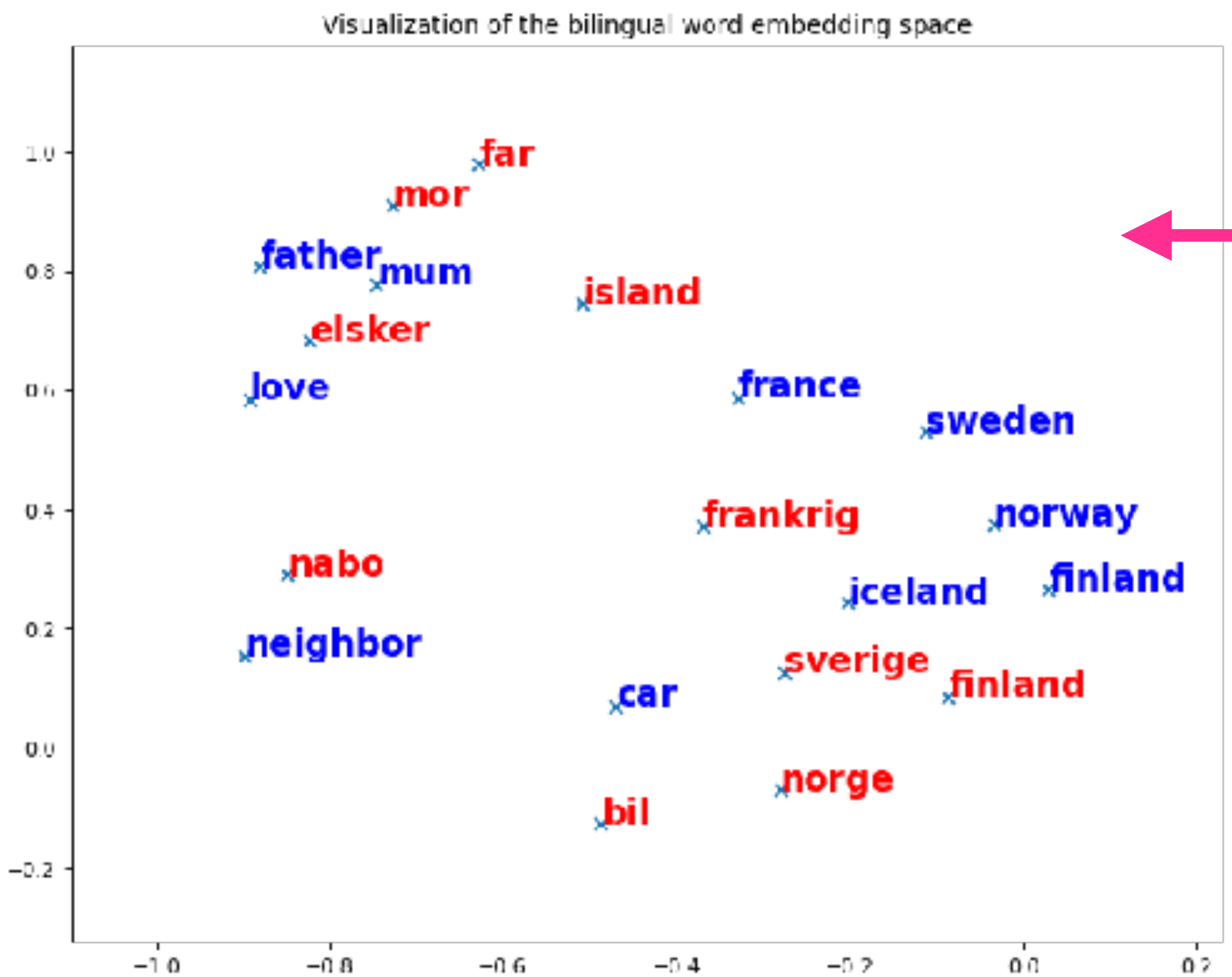
- Similar to Ma and Hovy (2016) but with a character-level bilstm

bilstm-CRF



Bilingual embeddings

- Monolingual English and Danish Polyglot embeddings
- Align with Procrustes rotation method introduced in MUSE (Conneau et al., 2017; Artetxe et al., 2017)

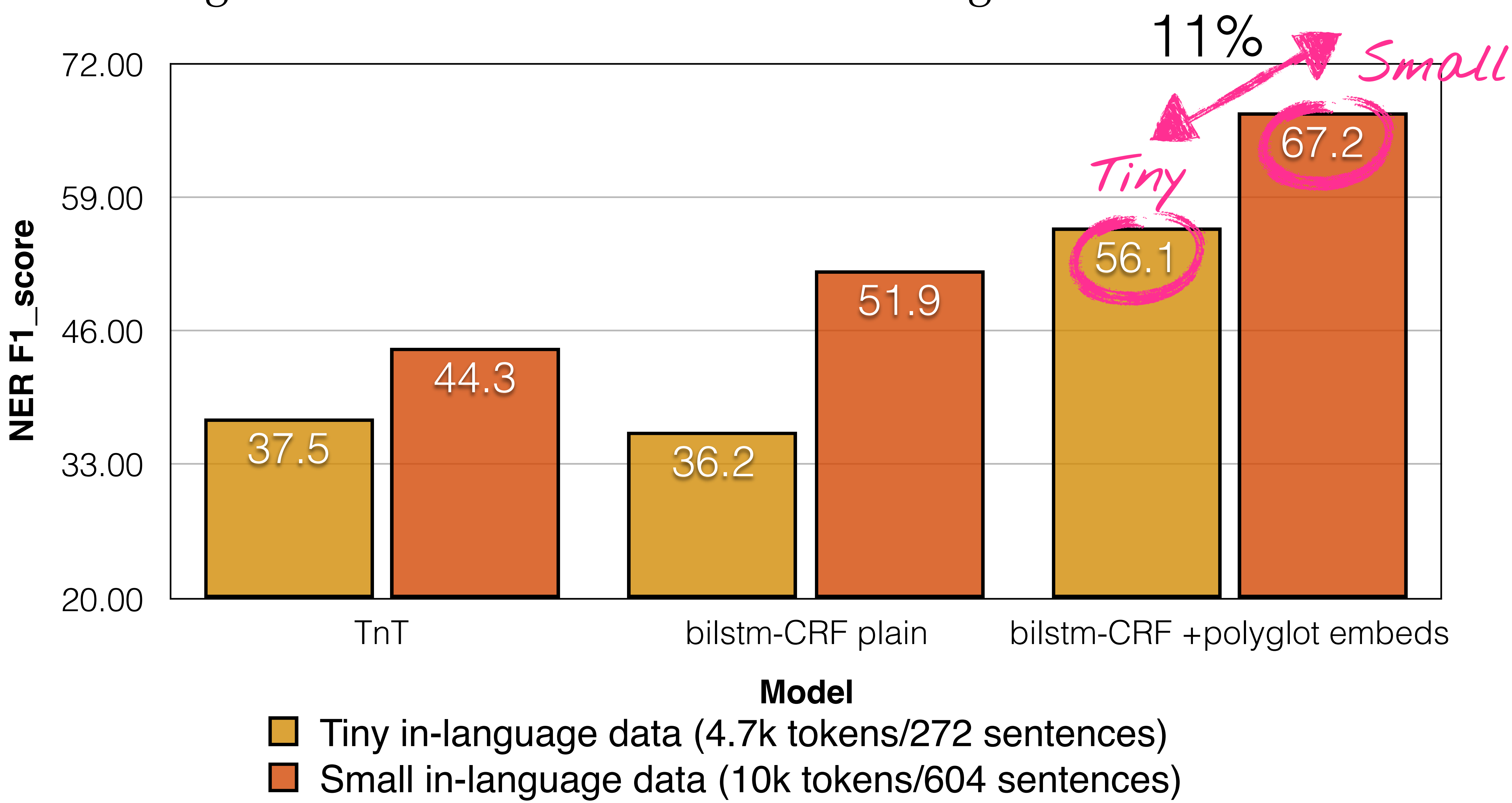


project embeddings

(many other possibilities, like joint data generation)

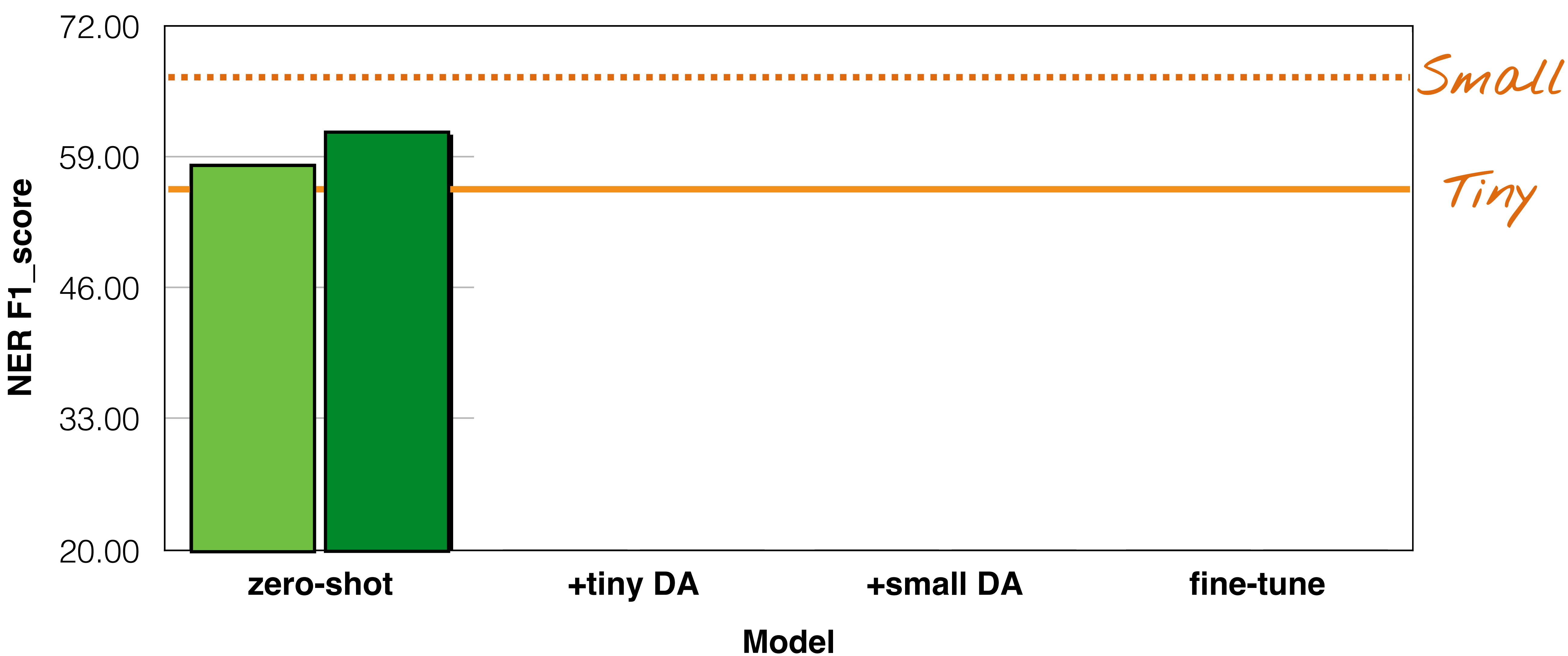
Results: Baselines

- Training on small amounts of annotated target Danish data



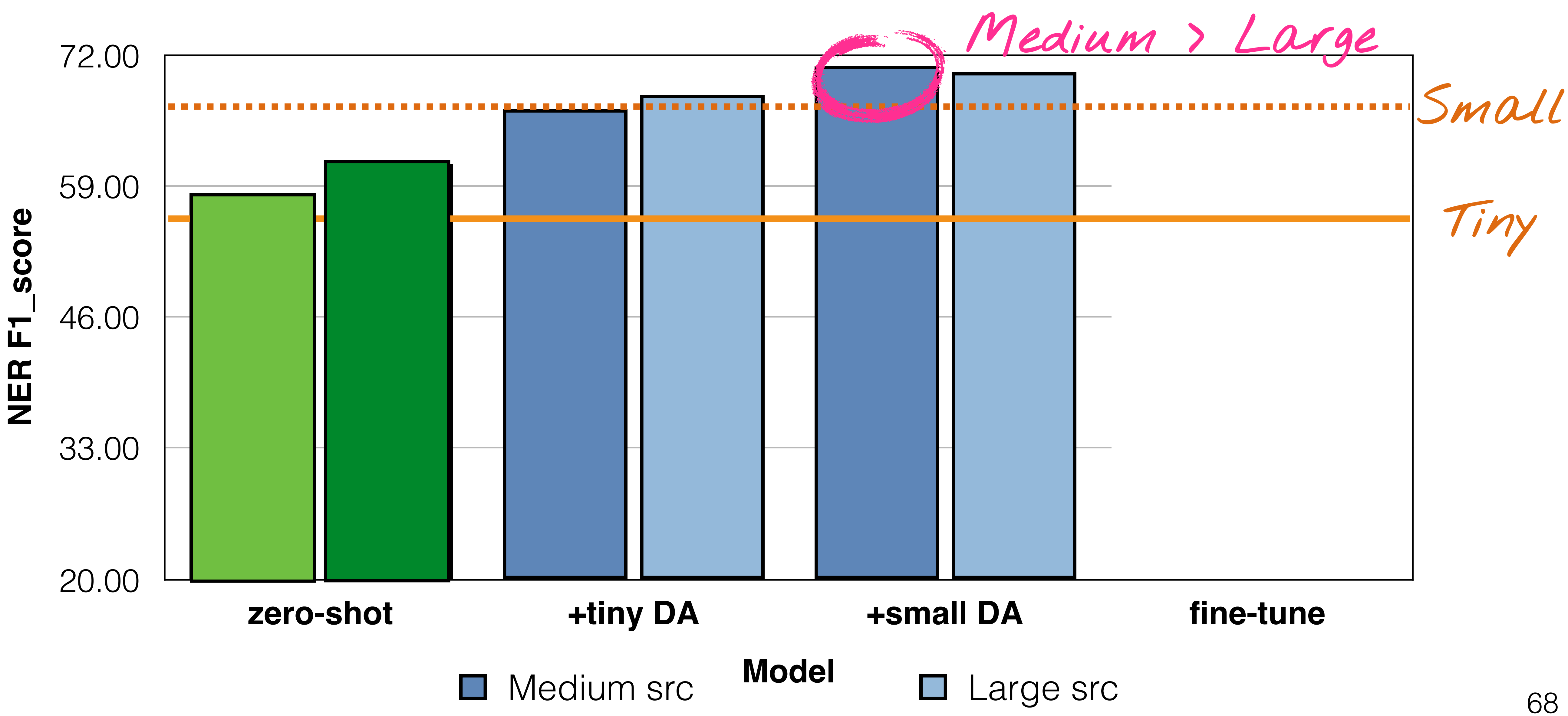
Results: Cross-lingual transfer

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



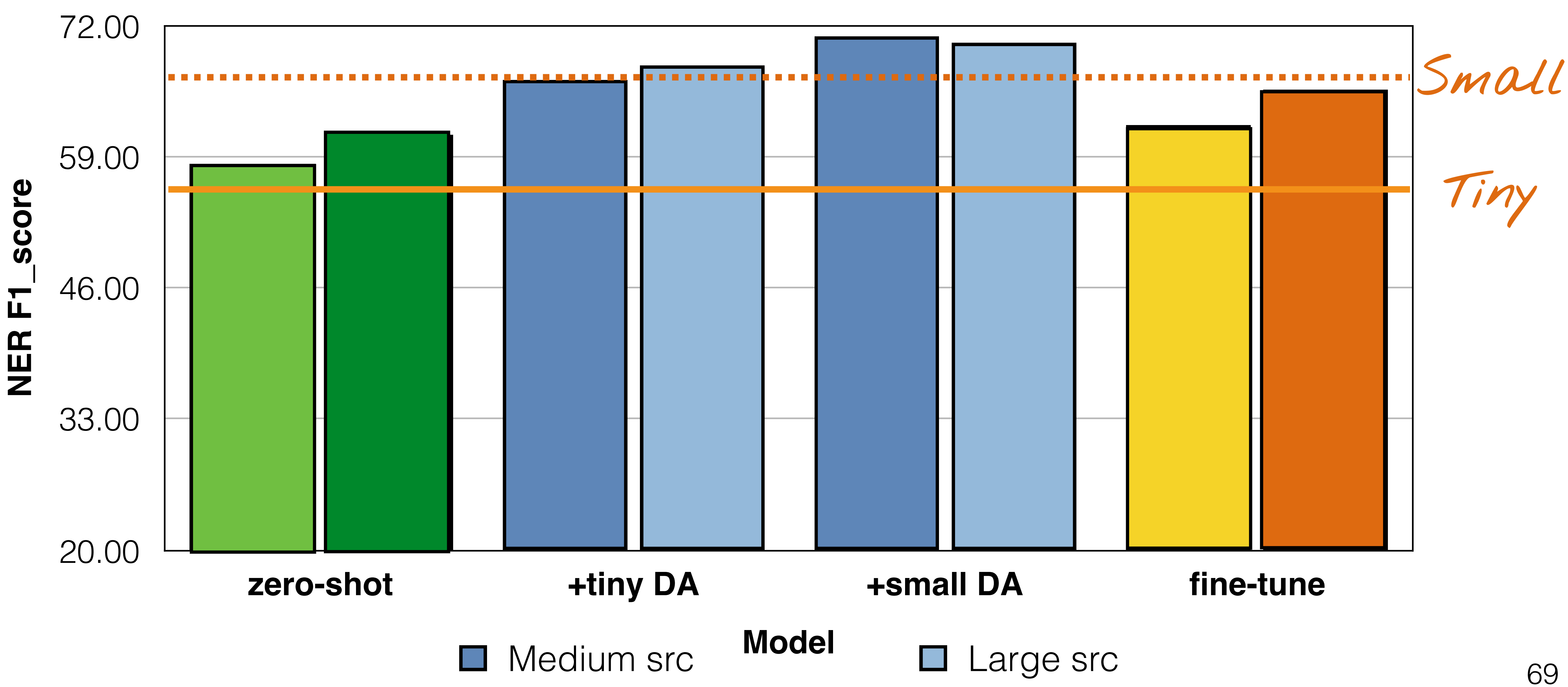
Results: Cross-lingual transfer

- **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 the 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.

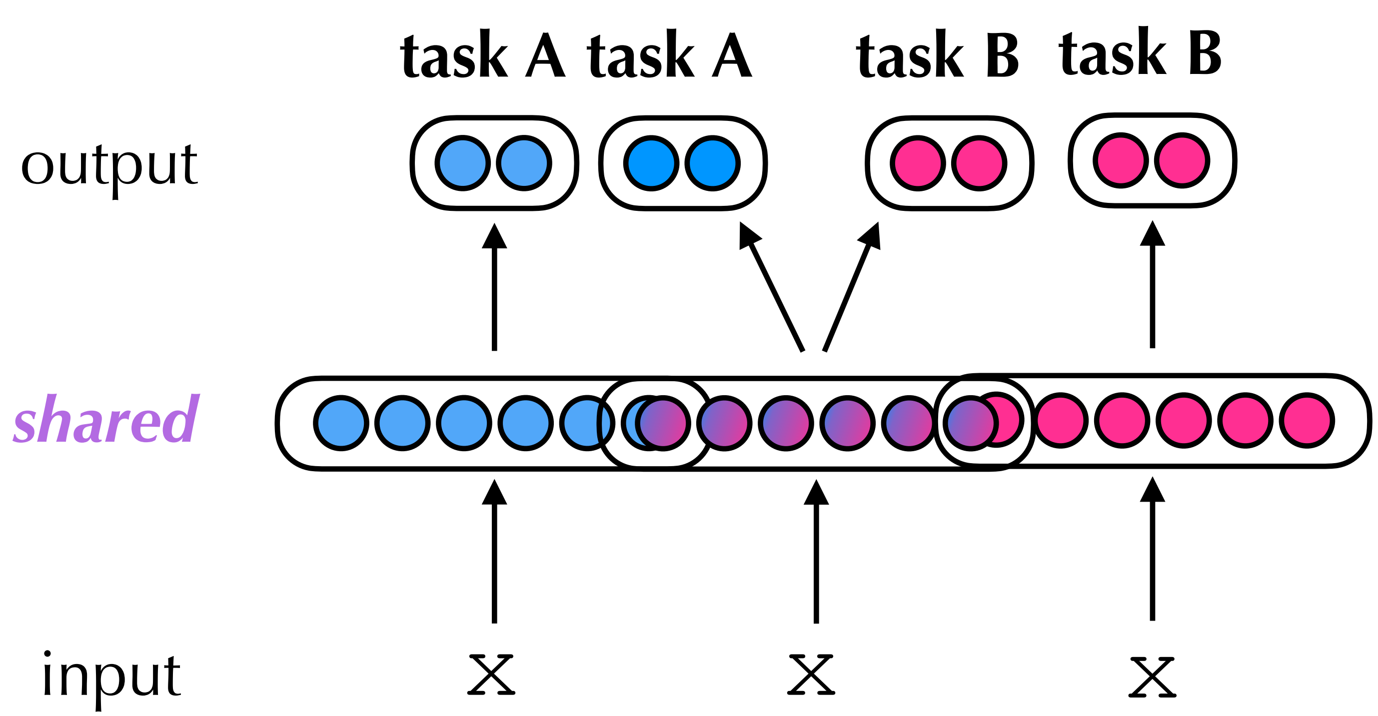
Roadmap

1. Domains: Learning to select data
2. Languages: Cross-lingual learning
3. Multi-task learning

Cross-Lingual word representations: MTL sharing at the lowermost level

Multi-task Learning (MTL): Key Idea

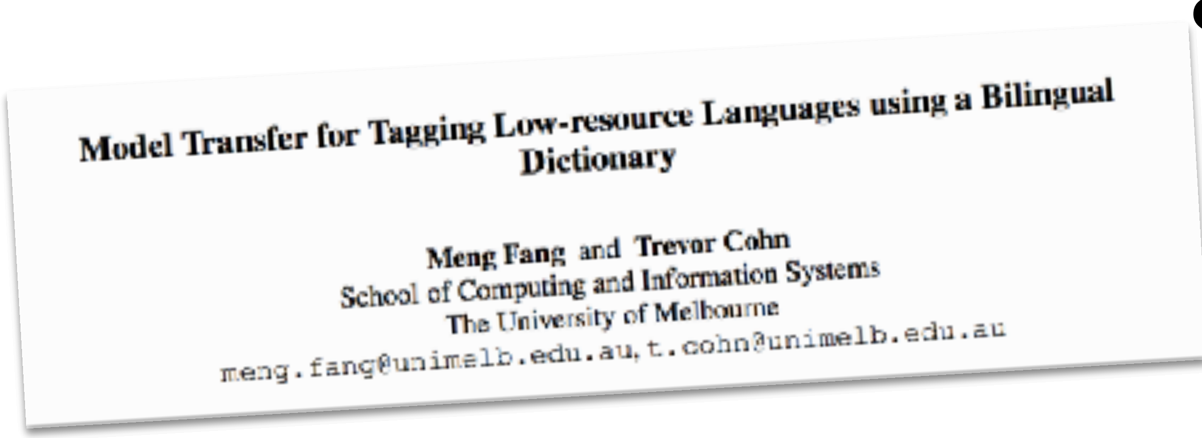
"**learning tasks in parallel while using a shared representation**; what is learned for each task **can help other tasks be learned better**" (Caruana, 1997)



single-task learning (STL)

multi-task learning (MTL)

MTL as distant supervision for low-resource tagging (Feng & Cohn, 2017, EACL)


Model Transfer for Tagging Low-resource Languages using a Bilingual Dictionary

Meng Fang and Trevor Cohn
School of Computing and Information Systems
The University of Melbourne
meng.fang@unimelb.edu.au, t.cohn@unimelb.edu.au

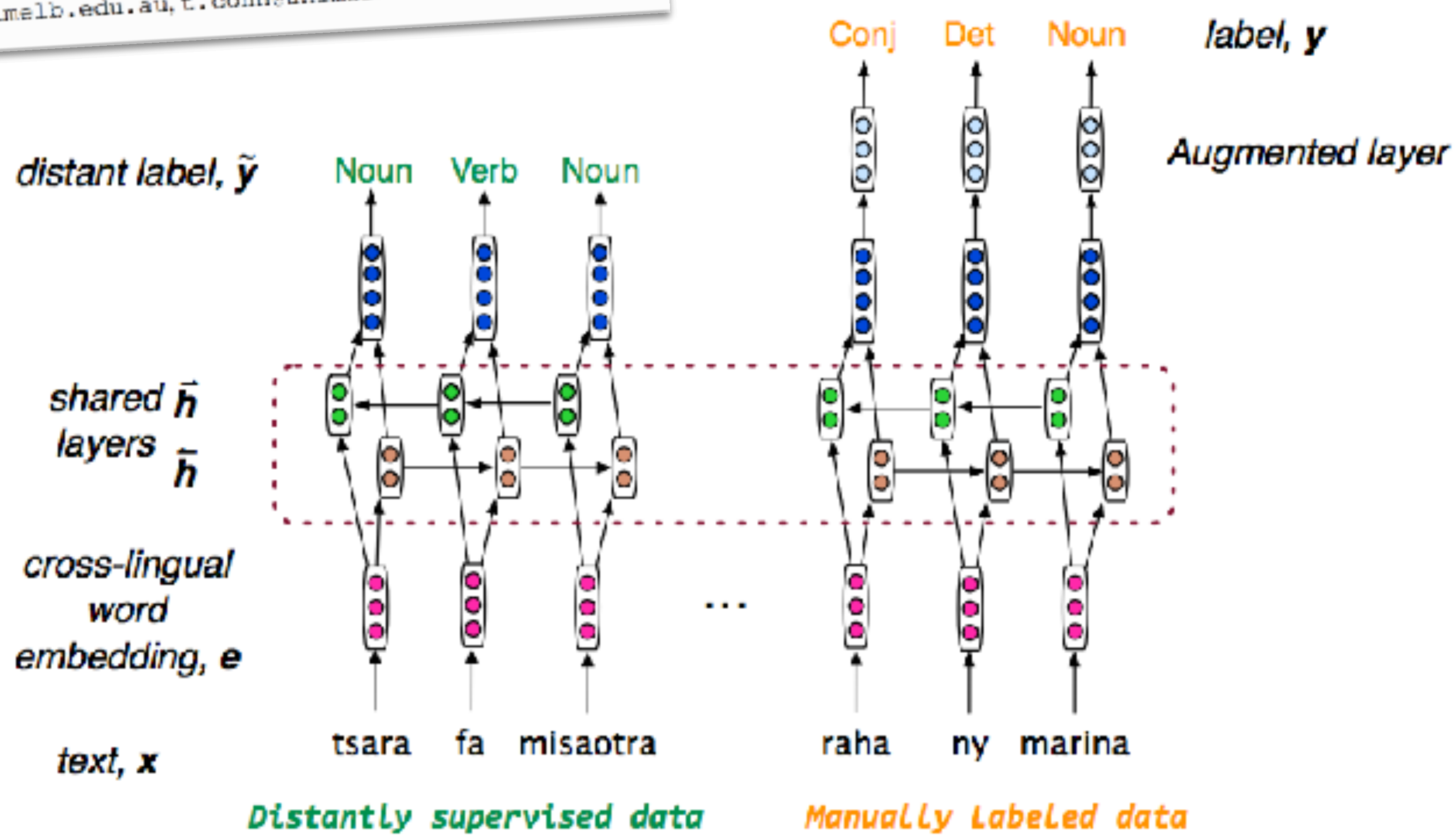
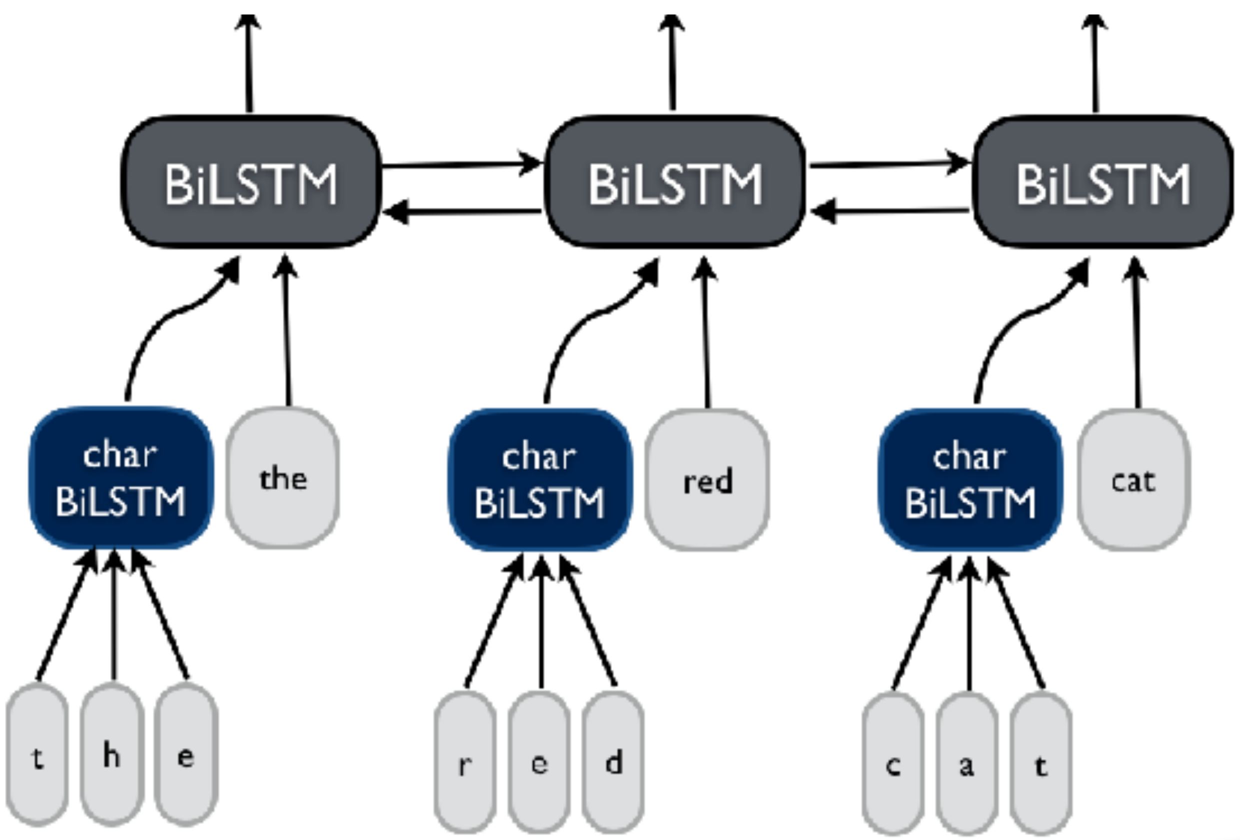


Figure 1: Illustration of the architecture of the joint model, which performs joint inference over both distant supervision (left) and manually labelled data (right).

What to share in dependency parsing?

(de Lhoneux et al., 2018, EMNLP)

(assume this is a transition-based parser)



BEST	✗	✓	ID
CHAR	✓	✗	✗
WORD	✗	✓	✗
STATE	✗	✗	✓
	⋮		


Parameter sharing between dependency parsers for related languages

Miryam de Lhoneux[1*] Johannes Bjerva[2] Isabelle Augenstein[2] Anders Søgaard[2]
[1]Department of Linguistics and Philology
Uppsala University
Uppsala, Sweden
[2] Department of Computer Science
University of Copenhagen
Copenhagen, Denmark

.. the power of contextualized word embeddings & MTL

To appear at EMNLP, 2019

75 language, one parser: UDify

75 Languages, 1 Model: Parsing Universal Dependencies Universally

Daniel Kondratyuk
Charles University
Institute of Formal and Applied Linguistics
Saarland University
Department of Computational Linguistics
dankondratyuk@gmail.com

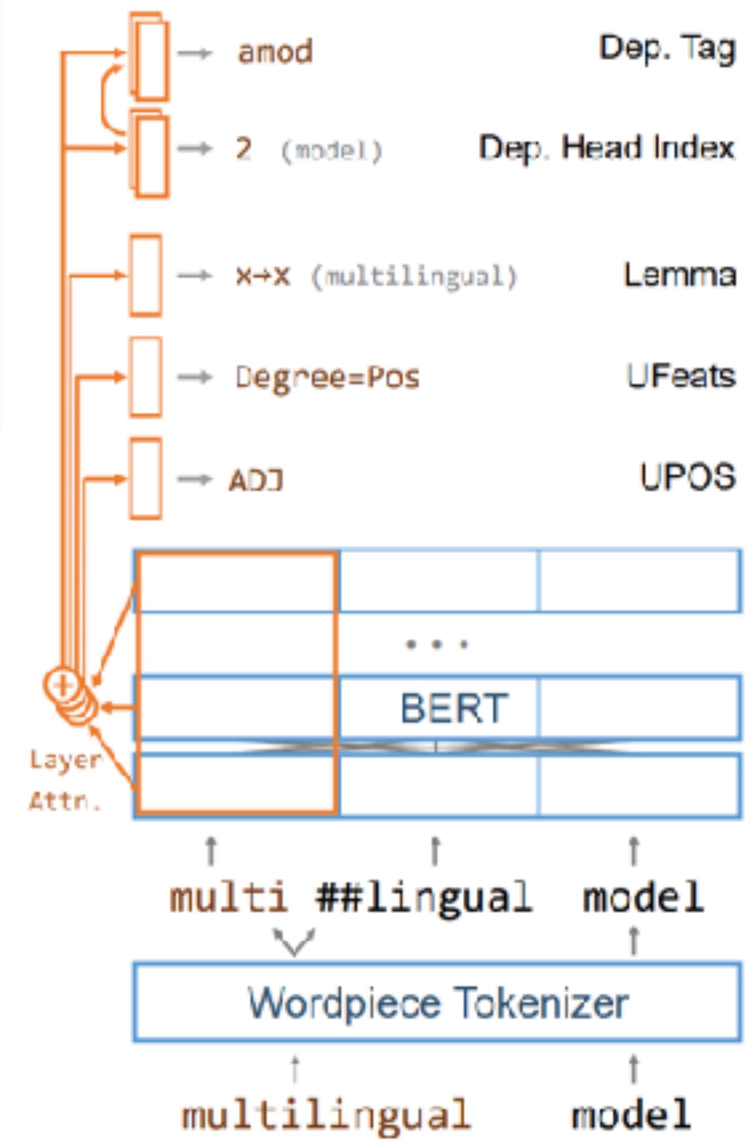


Figure 2: An illustration of the UDify network architecture with task-specific layer attention, inputting word tokens and outputting UD annotations for each token.

https://arxiv.org/pdf/1904.02099v2.pdf

To appear at EMNLP, 2019

UDify: Let's look at their results


TREEBANK
Afrikaans AfriBooms af_af
Akkadian PISANDUB akk_p
Amharic ATT
Ancient Greek PROIEL g
Ancient Greek Perseus grc
Arabic PADT
Arabic PUD
Armenian ArmTDP hy
Bambara CRB
Basque BDT
Belarusian HSE
Breton KEB
Bulgarian BTB
Buryat BDT
Cantonese HK
Catalan AnCora ca
Chinese CFL
Chinese GSD
Chinese HK
Chinese PUD
Coptic Scriptorium cop_scri
Croatian SET
Czech CAC
Czech CLTT
Czech FicTree c
Czech PDT
Czech PUD
Danish DDT
Dutch Alpino
Dutch LassySmall nl_las
Tal

Treebank		Model	UPOS	UFeats	Lemmas	UAS	LAS	CLAS	MLAS	BLEX
Afrikaans AfriBooms	af_afribooms	UDPipe	**98.25**	**97.66**	**97.46**	**89.38**	**86.58**	**81.44**	**77.66**	**77.82**
		UDify	97.48	96.63	95.23	86.97	83.48	77.42	70.57	70.93
Akkadian PISANDUB	akk_pisandub	UDify	**19.92**	**99.51**	**2.32**	**27.65**	**4.54**	**3.27**	**1.04**	**0.30**
Amharic ATT	am_att	UDify	**15.25**	**43.95**	**58.04**	**17.38**	**3.49**	**4.88**	**0.23**	**2.53**
Ancient Greek PROIEL	grc_proiel	UDPipe	**97.86**	**92.44**	**93.51**	**85.93**	**82.11**	**77.70**	**67.16**	**71.22**
		UDify	91.20	82.29	76.16	78.91	72.66	66.07	50.79	47.27
Ancient Greek Perseus	grc_perseus	UDPipe	**93.27**	**91.39**	**85.02**	**78.85**	**73.54**	**67.60**	**53.87**	**53.19**
		UDify	85.67	81.67	70.51	70.51	62.64	55.60	39.15	35.05
Arabic PADT	ar_padt	UDPipe	**96.83**	**94.11**	**95.28**	87.54	**82.94**	**79.77**	**73.92**	**75.87**
		UDify	96.58	91.77	73.55	**87.72**	82.88	79.47	70.52	50.26
Arabic PUD	ar_pud	UDify	**79.98**	**40.32**	**0.00**	**76.17**	**67.07**	**65.10**	**10.67**	**0.00**
Armenian ArmTDP	hy_armtdp	UDPipe	93.49	**82.85**	**92.86**	78.62	71.27	65.77	**48.11**	**60.11**
		UDify	**94.42**	76.90	85.63	**85.63**	**78.61**	**73.72**	46.80	59.14
Bambara CRB	bm_crb	UDify	**30.86**	**57.96**	**20.42**	**30.28**	**8.60**	**6.56**	**1.04**	**0.76**
Basque BDT	eu_bdt	UDPipe	**96.11**	**92.48**	**96.29**	**86.11**	**82.86**	**81.79**	**72.33**	**78.54**
		UDify	95.45	86.80	90.53	84.94	80.97	79.52	63.60	71.56
Belarusian HSE	be_hse	UDPipe	93.63	73.30	**87.34**	78.58	72.72	69.14	46.20	58.28
		UDify	**97.54**	**89.36**	85.46	**91.82**	**87.19**	**85.05**	**71.54**	**68.66**
Breton KEB	br_keb	UDify	**62.78**	**47.12**	**51.31**	**63.52**	**39.84**	**35.14**	**4.64**	**16.34**
Bulgarian BTB	bg_btb	UDPipe	**98.98**	**97.82**	**97.94**	93.38	90.35	87.01	**83.63**	**84.42**
		UDify	98.89	96.18	93.49	**95.54**	**92.40**	**89.59**	83.43	80.44
Buryat BDT	bxr_bdt	UDPipe	40.34	32.40	58.17	32.60	18.83	12.36	1.26	**6.49**
		UDify	**61.73**	**47.45**	**61.03**	**48.43**	**26.28**	**20.61**	**5.51**	11.68
Cantonese HK	yue_hk	UDify	**67.11**	**91.01**	**96.01**	**46.82**	**32.01**	**33.35**	**14.29**	**31.26**

UDify zero-shot results

Treebank		UPOS	Feats	Lem	UAS	LAS
Breton KEB	**br_keb**	63.67	46.75	53.15	63.97	40.19
Tagalog TRG	**tl_trg**	61.64	35.27	75.00	64.73	39.38
Faroese OFT	fo_oft	77.86	35.71	53.82	69.28	61.03
Naija NSC	pcm_nsc	56.59	52.75	97.52	47.13	33.43
Sanskrit UFAL	sa_ufal	40.21	18.45	37.60	41.73	19.80

Table 5: Test set results for zero-shot learning, i.e., no UD training annotations available. Languages that are pretrained with BERT are bolded.

https://arxiv.org/pdf/1904.02099v2.pdf

Huh!

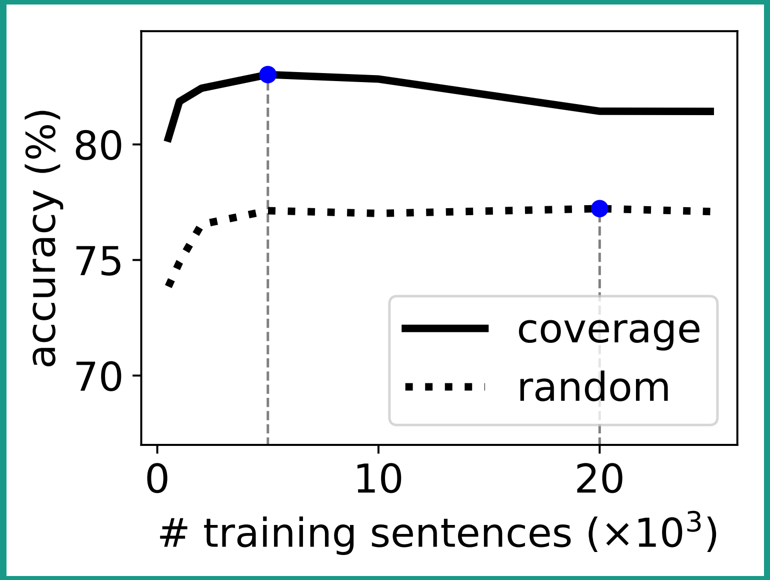
- ... Massively multi-lingual learning with contextualized embeddings and careful fine-tuning: big leaps forward
- ... Is MTL & Sequence Labeling with Attention all we need?
- More work needed (sharing what, data selection, pacing of learning)

To wrap up...

Take-away 1: Less is more

Data selection is beneficial in cross-lingual and cross-domain learning

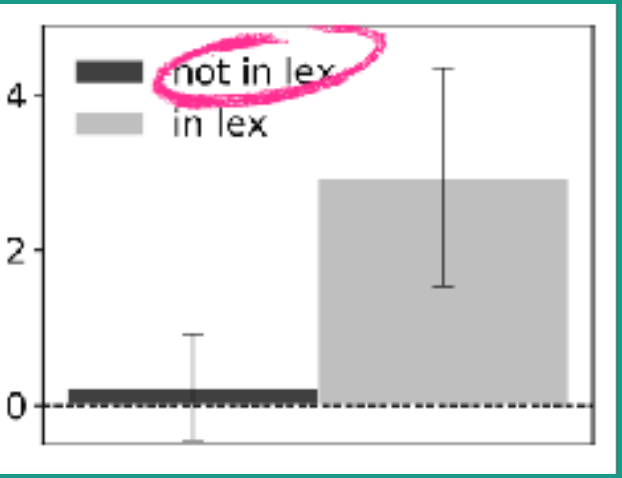
Cross-domain



Cross-lingual

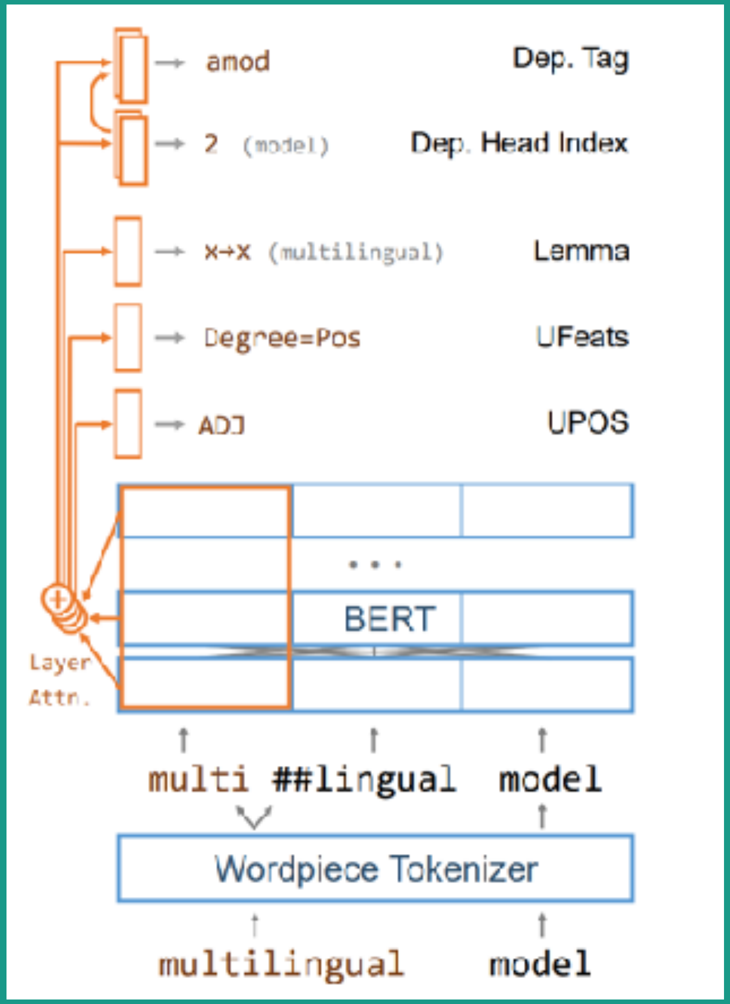
Take-away 2: Symbolic inductive bias

Neural models can benefit from inductive bias from **symbolic information.**



Take-away 3: MTL flexibility

Multi-task learning provides many opportunities (and challenges) and there is more to be discovered (especially in relation to multilingual modeling)



IT UNIVERSITY OF COPENHAGEN

Transferring NLP models across languages and domains

Barbara Plank, ITU, Denmark

SYNTAX FEST
Paris, August 26-30 2019

Questions?
Thanks!

@barbara_plank
bplank.github.io
https://nlp.itu.dk/

Supported by: NVIDIA amazon