

Empowering Citizens. Smarter Societies.

Insight

Centre for Data Analytics



# Neural Transfer Learning for Natural Language Processing

Sebastian Ruder  
Supervisor: Dr. John Breslin

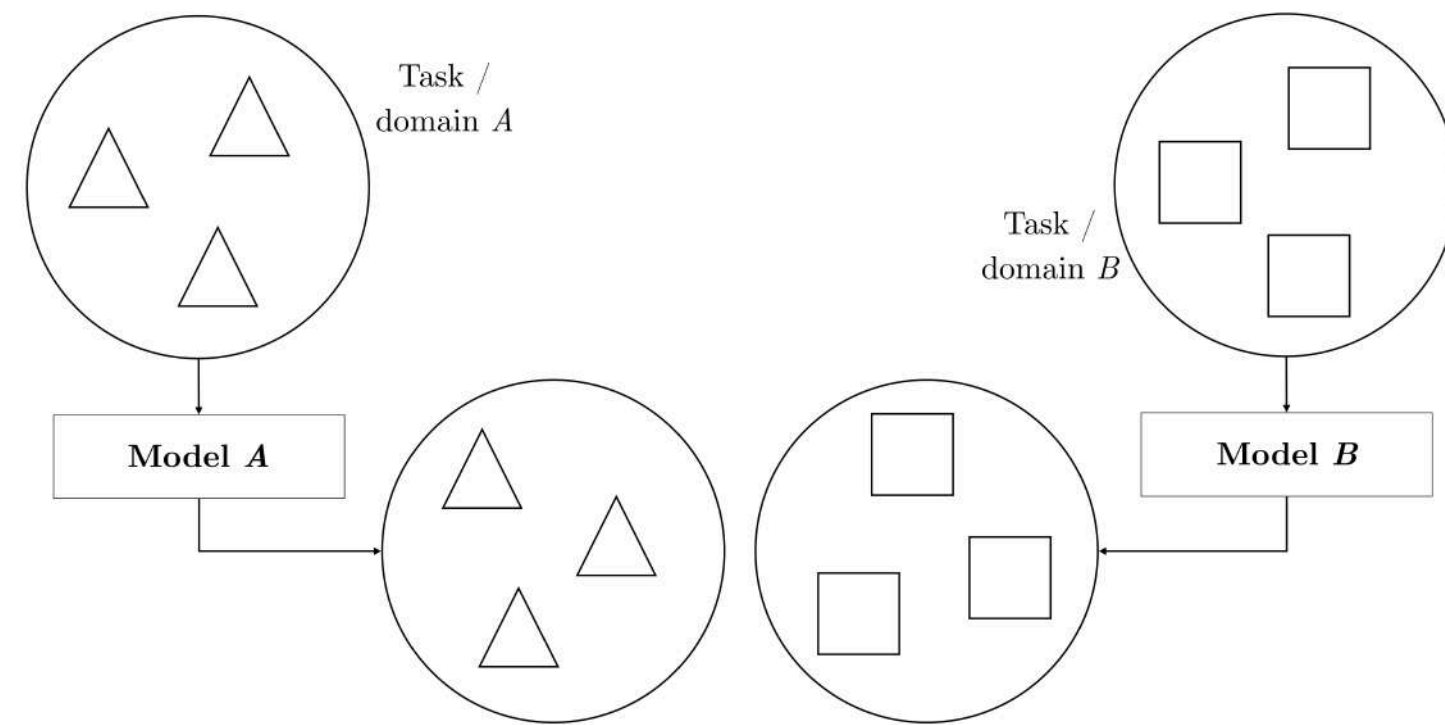
February 26, 2019  
PhD Viva Presentation

A World Leading SFI Research Centre

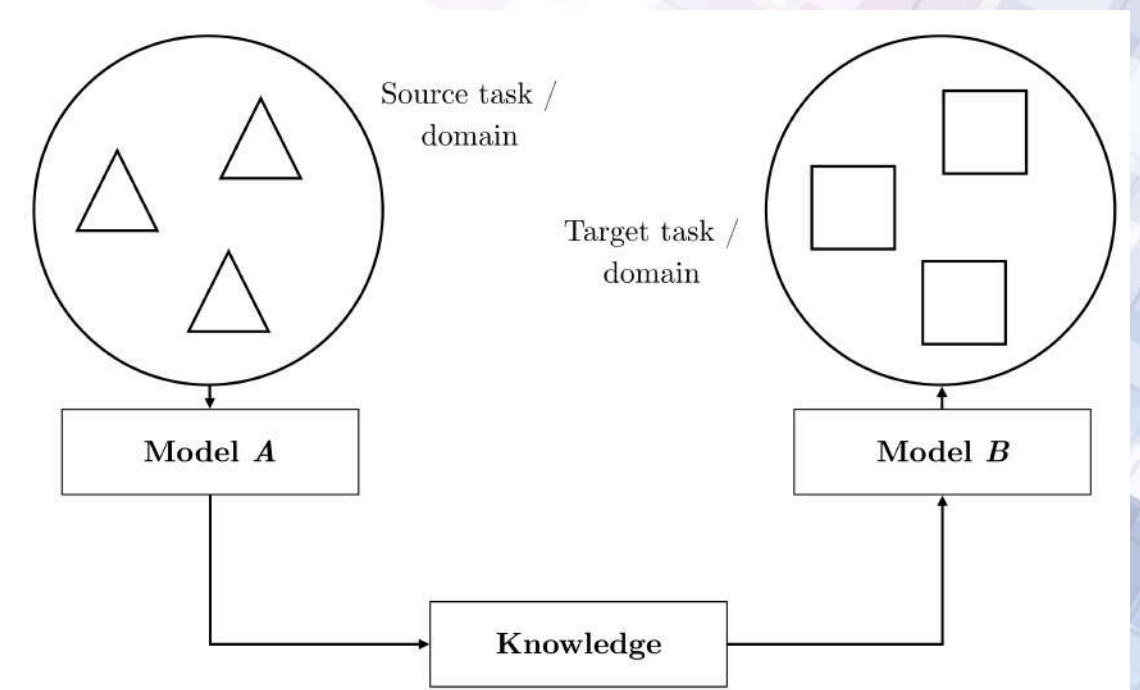




# Supervised learning vs. Transfer learning



Training and evaluation is performed on the same task and domain



Knowledge from the source setting is transferred to the target setting

# Why Transfer Learning?

- State-of-the-art supervised models are **brittle**
  - They are sensitive to **adversarial examples**

**Paragraph:** "In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses."

**Question:** "What city did Tesla move to in 1880?"

**Answer:** Prague

80% accuracy



**Paragraph:** "In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses. Tadakatsu moved to the city of Chicago in 1881."

**Question:** "What city did Tesla move to in 1880?"

**Answer:** Chicago

34.2% accuracy



Jia and Liang (EMNLP 2017)



# Why Transfer Learning?

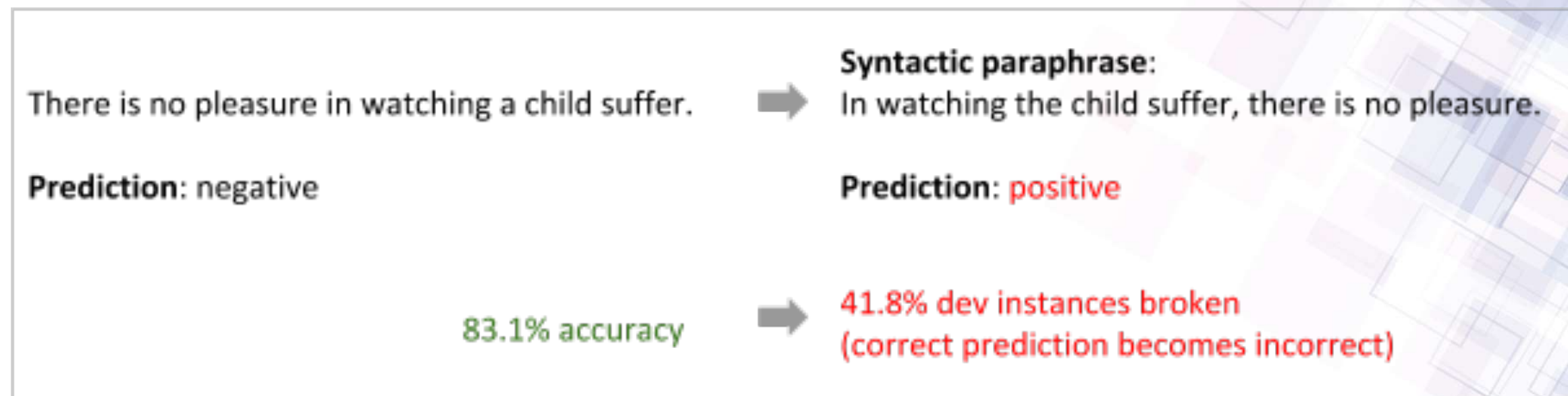
- State-of-the-art supervised models are **brittle**
  - They are sensitive to **noise**

<p><b>phonetic:</b> Tut sieht Trotzdem gekriegt Natürlich</p> <p><b>omission:</b> erfahren, Babysitter, selbst, Hausschuhe</p> <p><b>morphological:</b> wohnt, fortsetzt, wünsche</p> <p><b>key swap:</b> Eltern, Deine, nichts, Bahn</p> <p><b>other:</b> Agglomerationen Hausaufgabe Thema Detailhandelsfachfrau</p>	→	<p><b>phonetic:</b> <b>Tud</b> (devoicing of final stops) <b>zieht</b> (s = /z/ before vowel) <b>Trozdem</b> (tz = /z/) <b>gekriegt</b> (vowel length) <b>Natürlich/Nätürlich</b> (diacritics);</p> <p><b>omission:</b> <b>erfaren, Babysiter, sebst, Hausschue</b></p> <p><b>morphological:</b> <b>wonnen, forzusetzen, wünchen</b></p> <p><b>key swap:</b> <b>Eltren, Diene, nicht, Bhan</b></p> <p><b>other:</b> <b>Agromelationen</b> (omission + letter swap) <b>Hausausgabe</b> <b>Temer</b> <b>Deitellhandfachfrau</b></p>
34.79 BLEU	→	14.02 BLEU

Belinkov and Bisk (ICLR 2018)

# Why Transfer Learning?

- State-of-the-art supervised models are **brittle**
  - They are sensitive to **paraphrases**






Iyyer et al. (NAACL 2018)



# Why Transfer Learning?

- In the real world, NLP models need to be applied to a plethora of:

- domains 
- tasks 
- languages 

- Manually annotating data for every new setting is infeasible
- Transfer learning enables transferring knowledge from a related setting to the target setting

# Why Transfer Learning?

- Many of the most fundamental advances in NLP can be seen as forms of transfer learning
  - Latent semantic analysis (LSA; Deerwester et al., 1990)
  - Brown clusters (Brown et al., 1993)
  - Pretrained word embeddings (Mikolov et al., 2013)



# Limitations of previous work

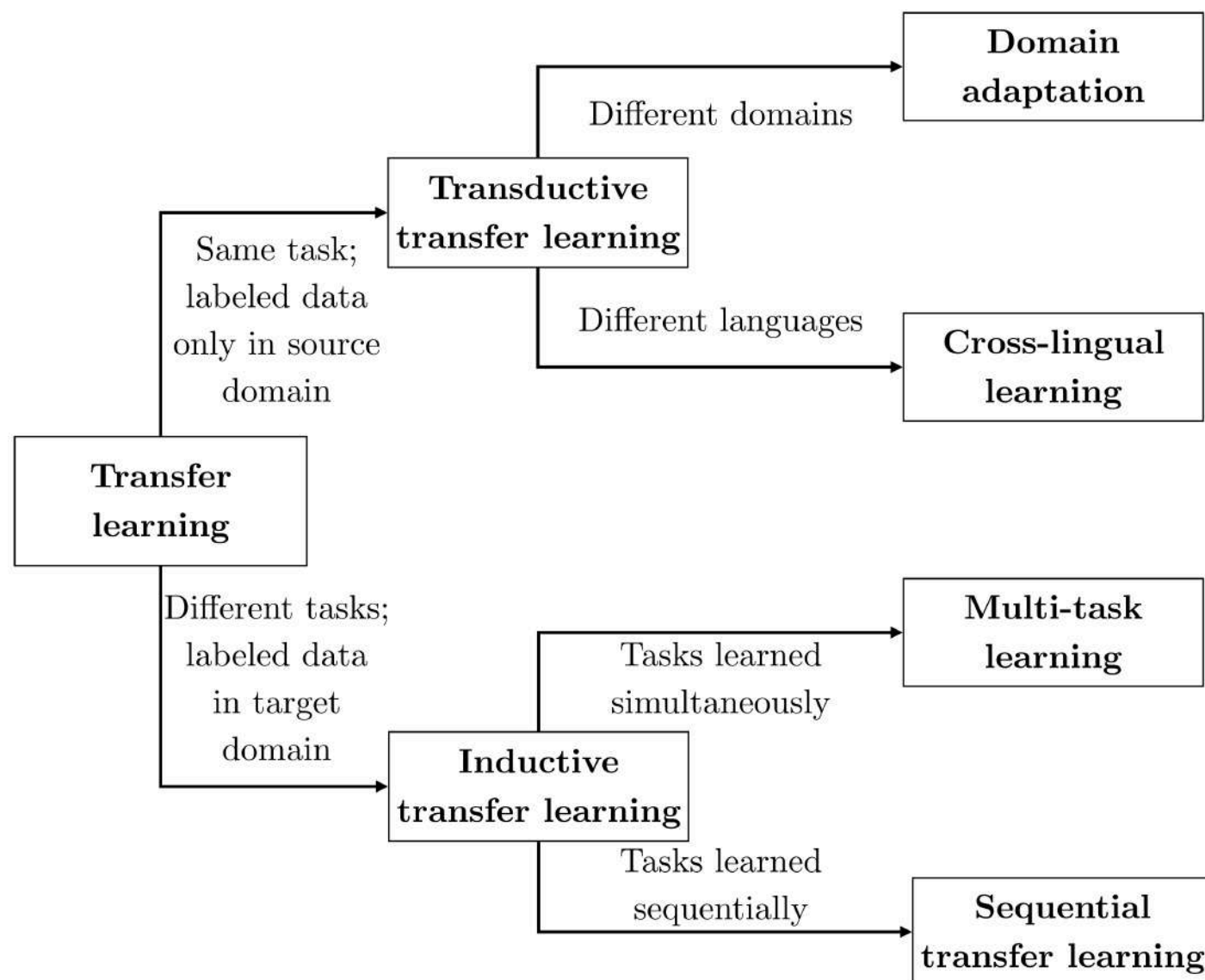
- **Overly restrictive:**  
Pre-specified similarity metrics, hard parameter sharing
- **Setting-specific:**  
Evaluation on one task, task-specific sharing schemes
- **Weak baselines:**  
Lack of comparison against traditional approaches
- **Brittle:**  
Underperform on out-of-domain data, dependent on similarity of languages/tasks
- **Inefficient:**  
Require more parameters, more time, and more samples



# Research objectives

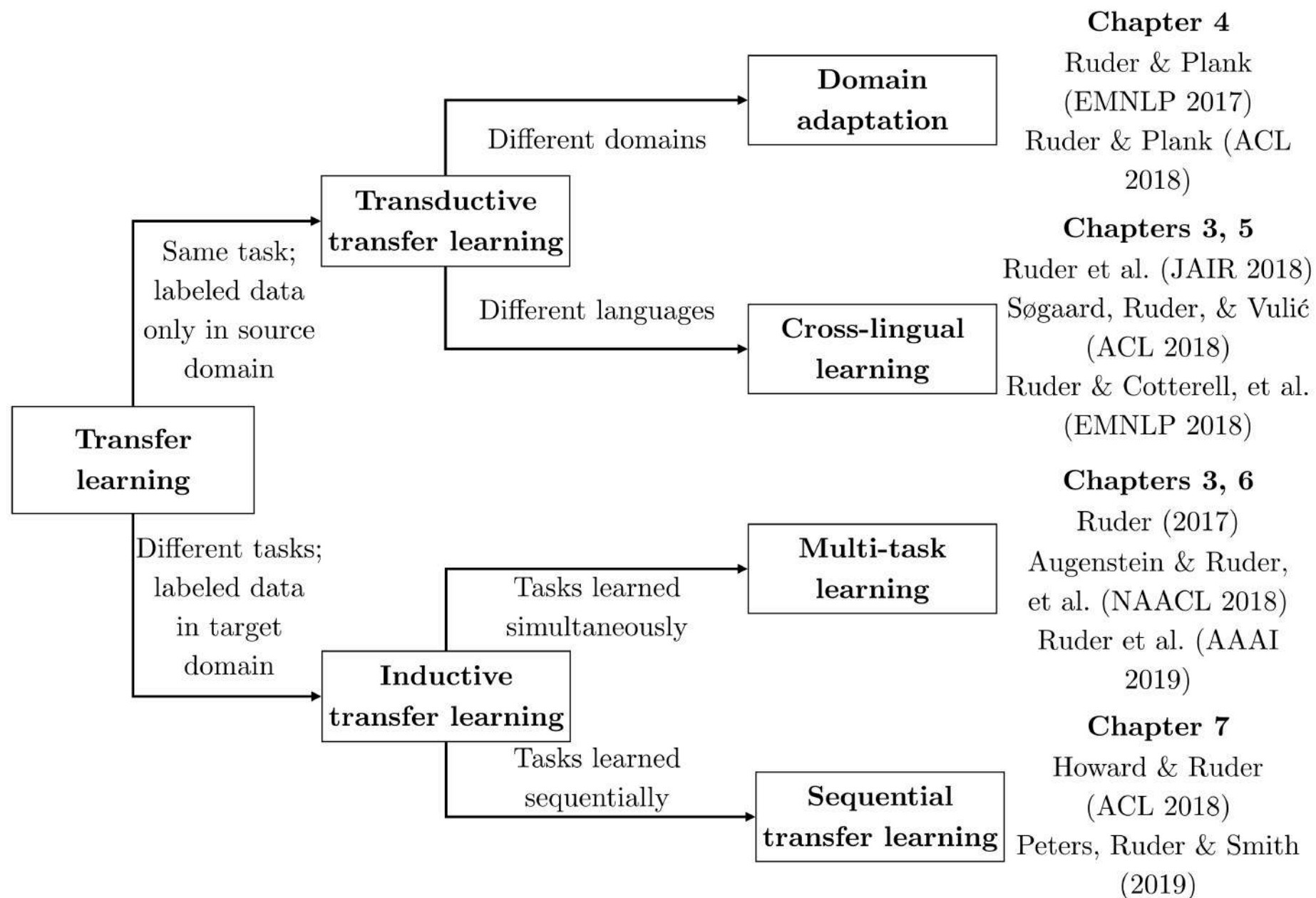
- Develop methods that:
  1. Overcome a discrepancy between the source and target setting.
  2. Induce an inductive bias.
  3. Combine traditional and current approaches.
  4. Transfer across the hierarchy of NLP tasks.
  5. Generalise across many settings.

# Taxonomy

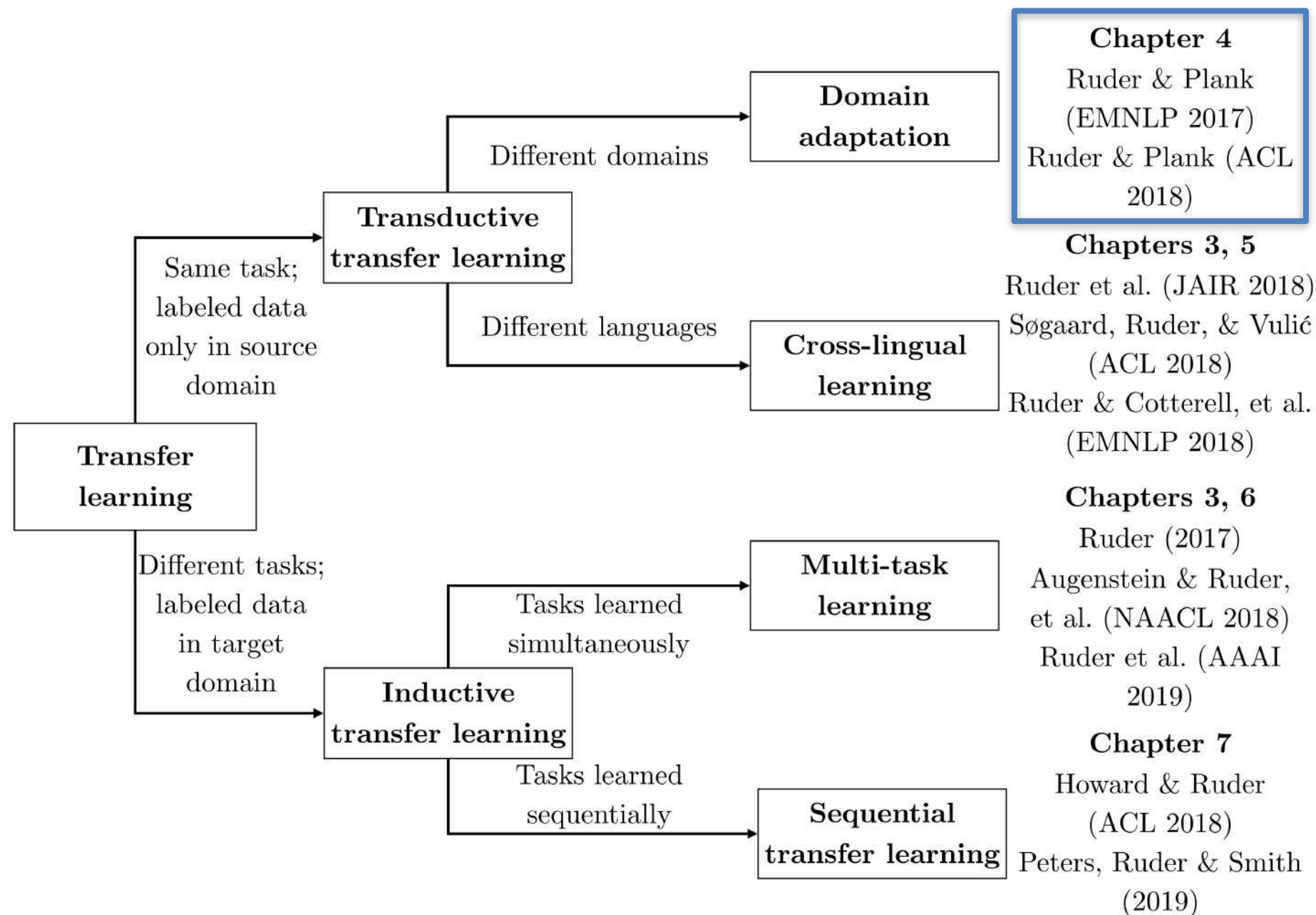




# Publications



# Contributions

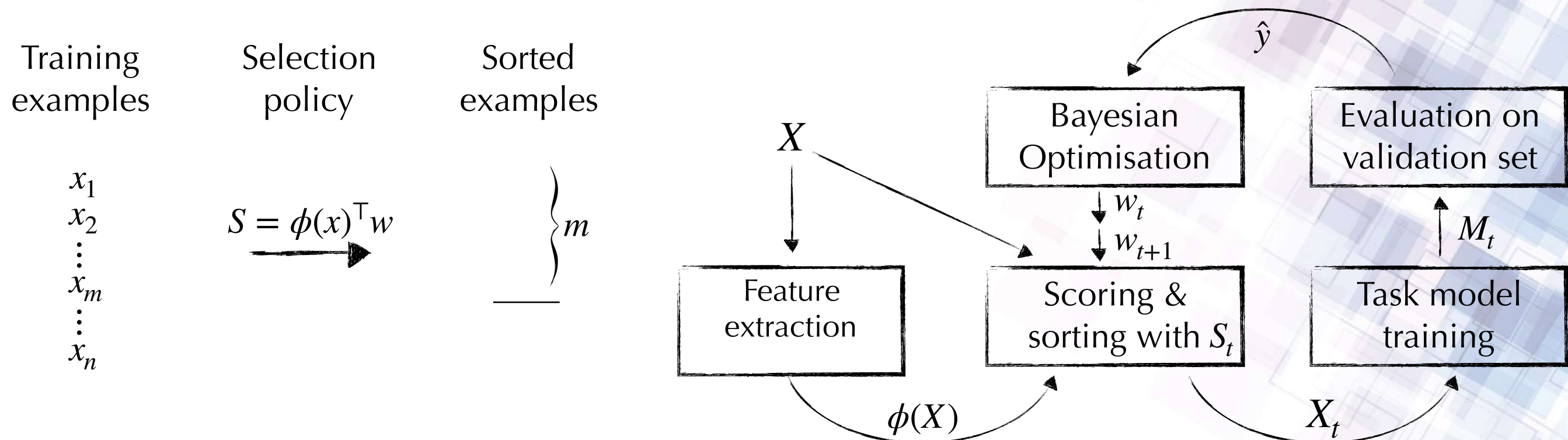




# Selecting Data for Domain Adaptation

- Propose two novel methods that bridge the domain discrepancy by selecting relevant and informative data for unsupervised domain adaptation:
  1. Based on Bayesian Optimisation ([Ruder & Plank, EMNLP 2017](#));
  2. Using semi-supervised learning and multi-task learning ([Ruder & Plank, ACL 2018](#)).

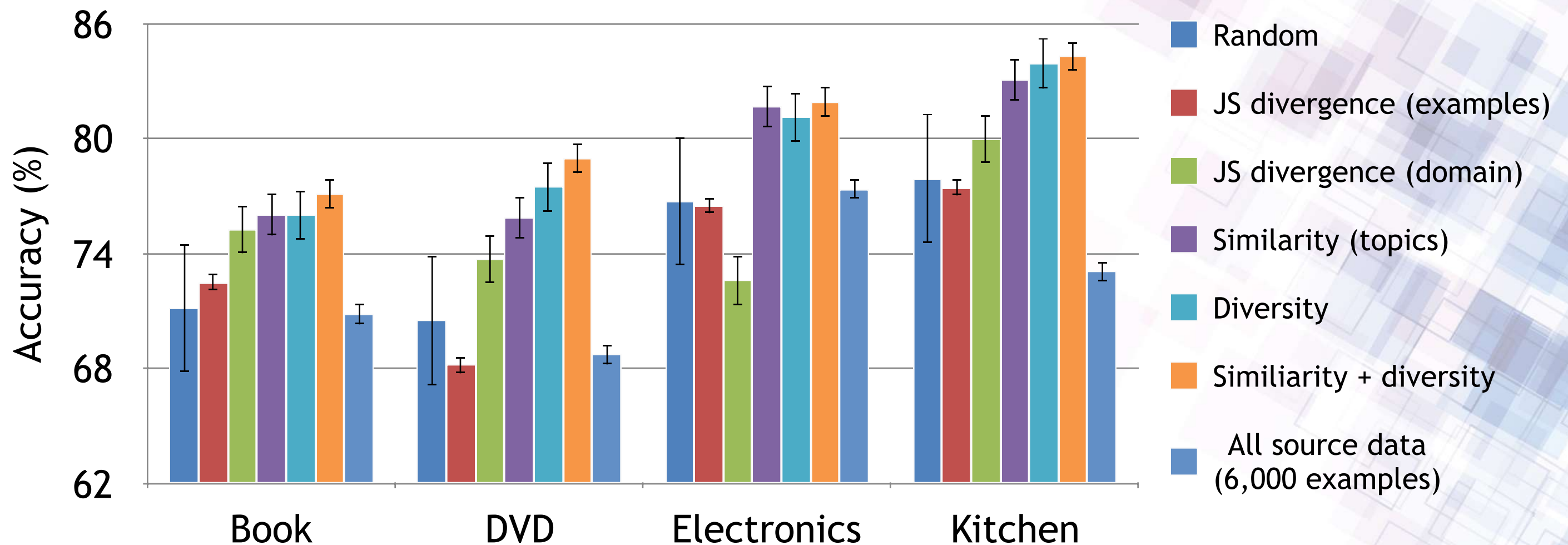
# Selecting Data with Bayesian Optimisation



(Ruder & Plank, EMNLP 2017)



## Sentiment analysis: selecting 2,000 from 6,000 source domain examples

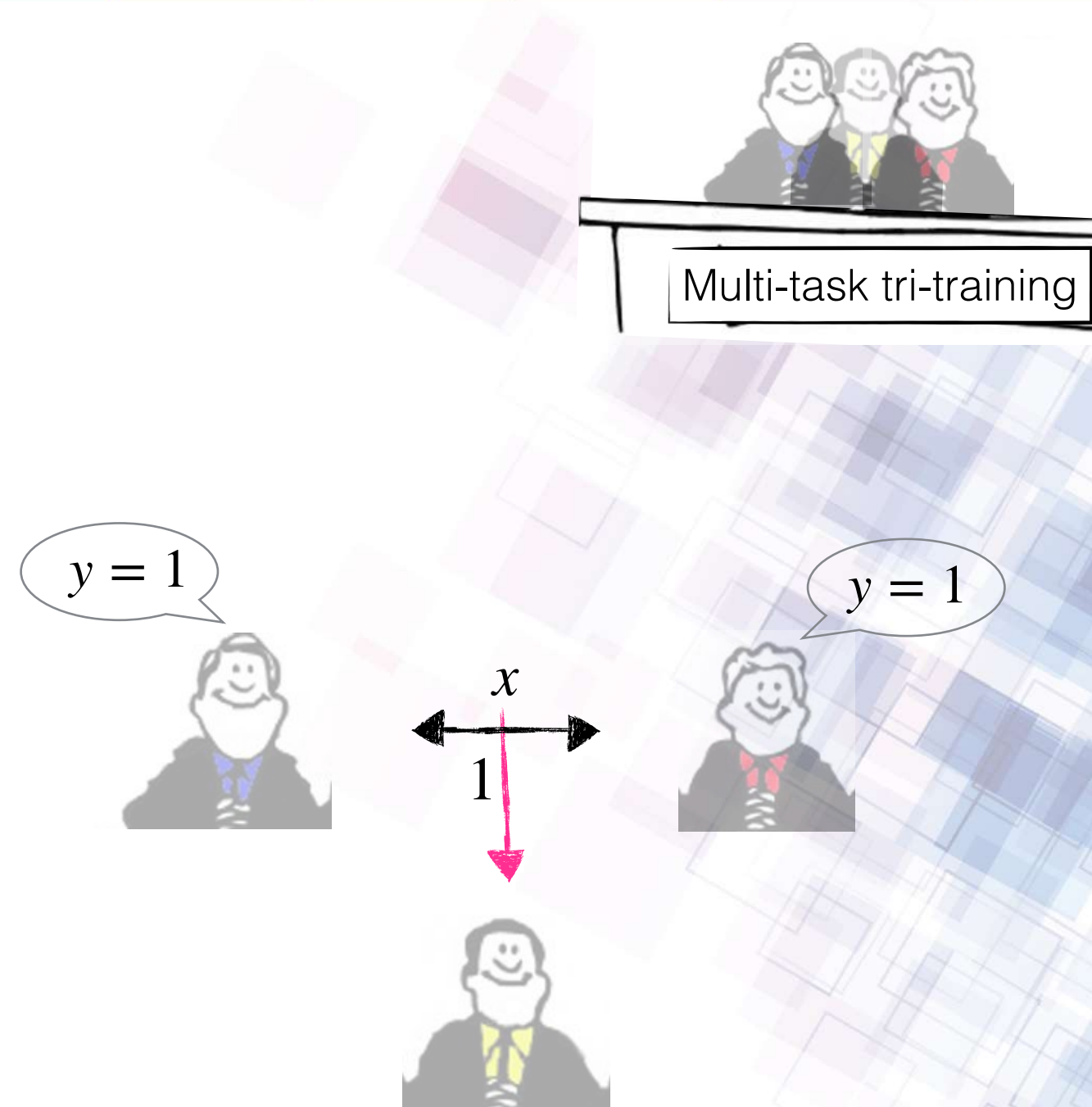


- Selecting relevant data is useful when domains are very different.

(Ruder & Plank, EMNLP 2017)

# Multi-task Tri-training I

1. Train one model with 3 objective functions.
2. Use predictions unlabeled data for third if two agree.
3. Restrict final layers to use different representations.
4. Train third objective function only on pseudo labeled data to bridge domain shift.

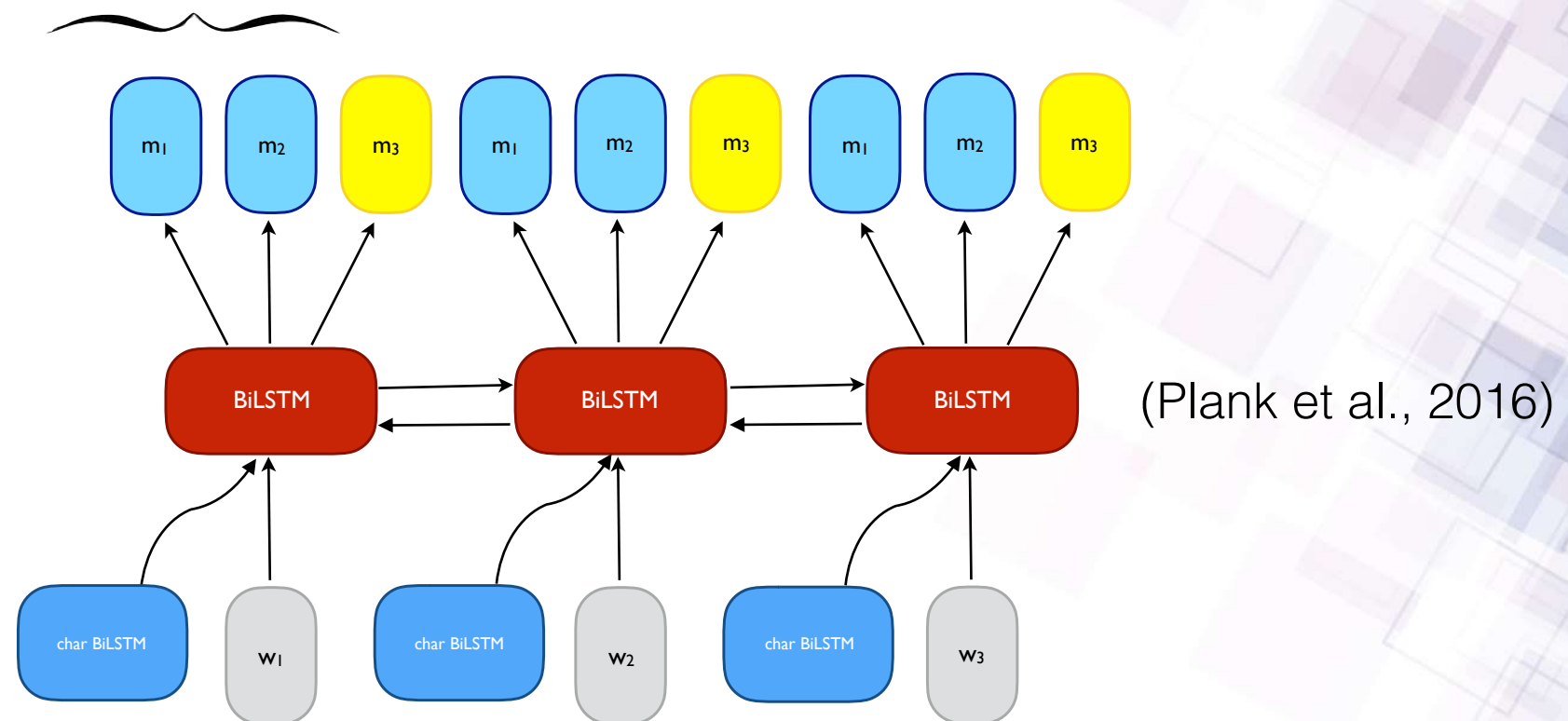


(Ruder & Plank, ACL 2018)



# Multi-task Tri-training II

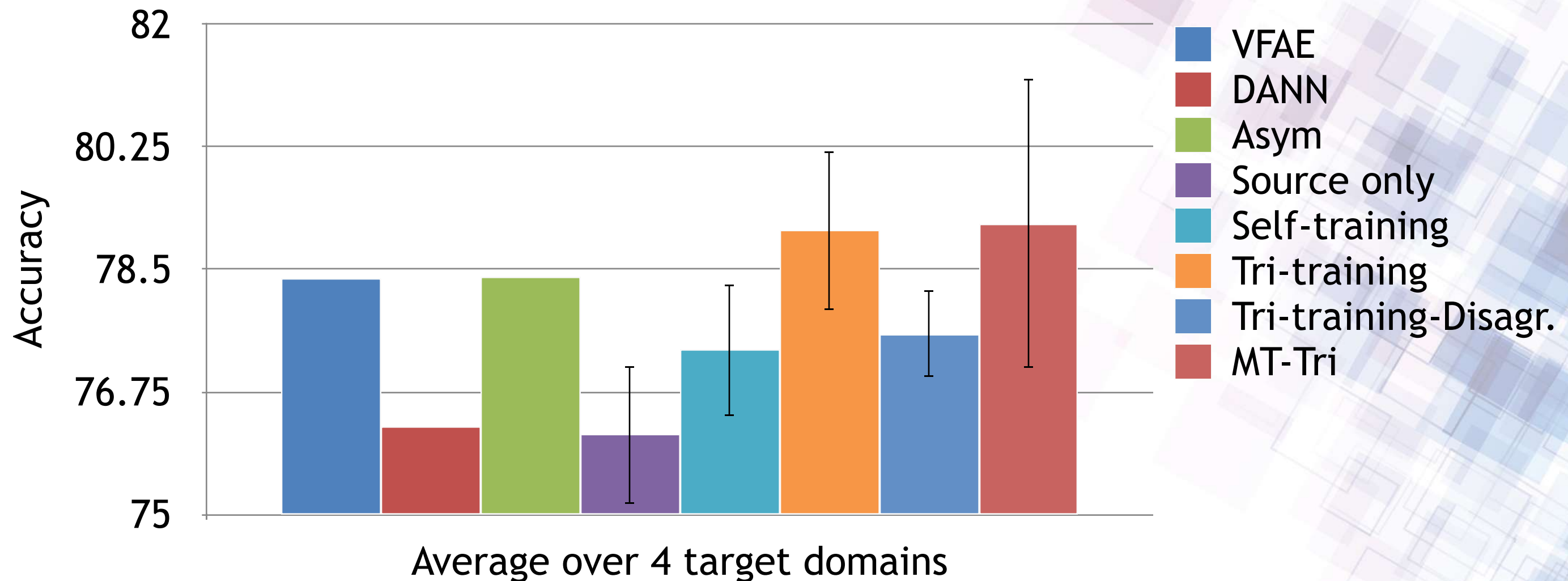
$$L_{orth} = \underbrace{\|W_{m_1}^\top W_{m_2}\|_F^2}_{\text{orthogonality constraint (Bousmalis et al., 2016)}}$$



$$\text{Loss: } L(\theta) = - \sum_i \sum_{1, \dots, n} \log P_{m_i}(y | \vec{h}) + \gamma L_{orth}$$

(Ruder & Plank, ACL 2018)

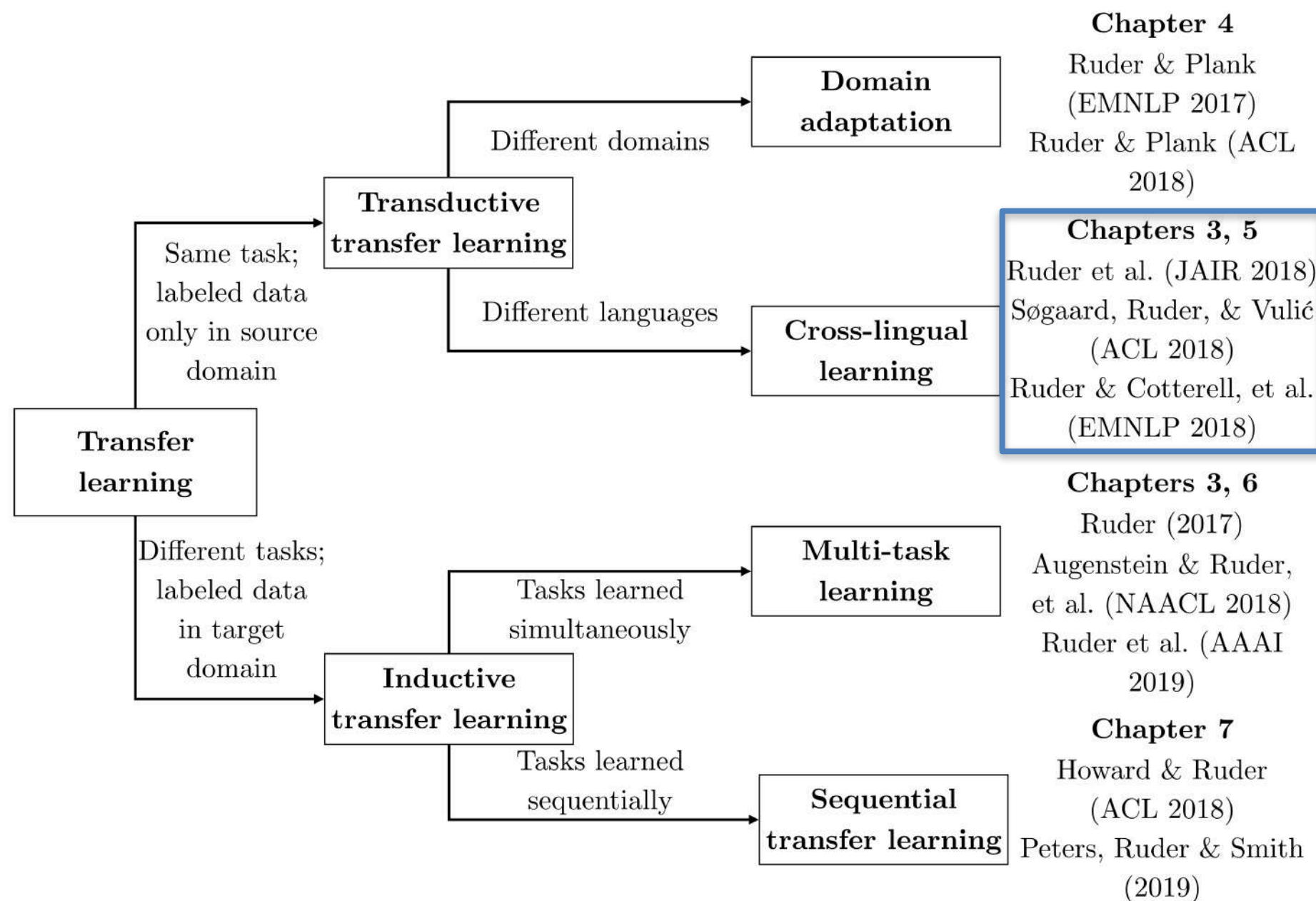
## Performance on sentiment analysis



- Multi-task tri-training slightly outperforms tri-training, while being significantly faster. However, it has higher variance.

(Ruder & Plank, ACL 2018)





# Unsupervised and weakly supervised cross-lingual learning

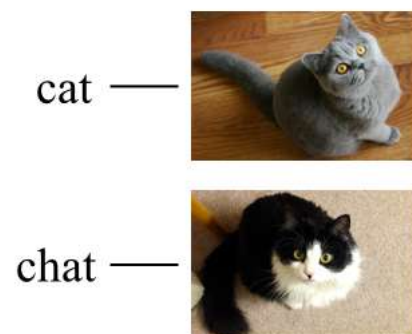
- Unsupervised cross-lingual word embedding models struggle to transfer between unrelated languages
- Analyse and propose methods to mitigate this discrepancy:
  1. Provide a taxonomy of cross-lingual word embedding models (Ruder et al., JAIR 2019).
  2. Analyse limitations of cross-lingual embedding methods (Søgaard, Ruder & Vulić, ACL 2018).
  3. Propose a novel latent-variable cross-lingual embedding model (Ruder & Cotterell et al., EMNLP 2018).



# A taxonomy for cross-lingual word embeddings

- Based on nature and type of alignment

cat — chat  
dog — chien



Word,  
parallel

Word,  
comparable

The dog chases  
the cat.  
|  
Le chien poursuit  
le chat.

Sentence,  
parallel

The dog chases the  
cat in the grass.



|  
Le chat s'enfuit  
du chien.

Sentence,  
comparable

There are a lot of  
dogs in the park. They  
like to chase cats.

|  
Les chats se relaxent.  
Ils fuient les chiens  
dès qu'ils les voient.

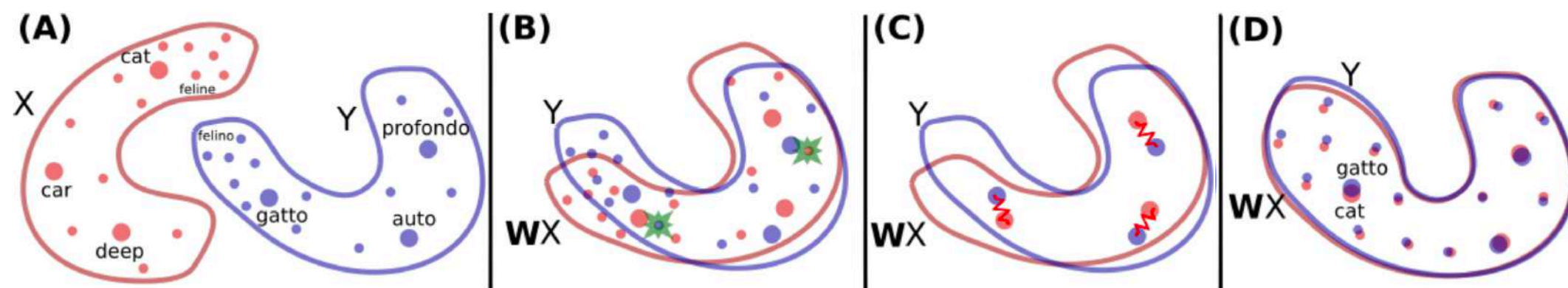
Document,  
comparable

(Ruder et al., JAIR 2019)

# Unsupervised method (Conneau et al., 2018)

4 steps:

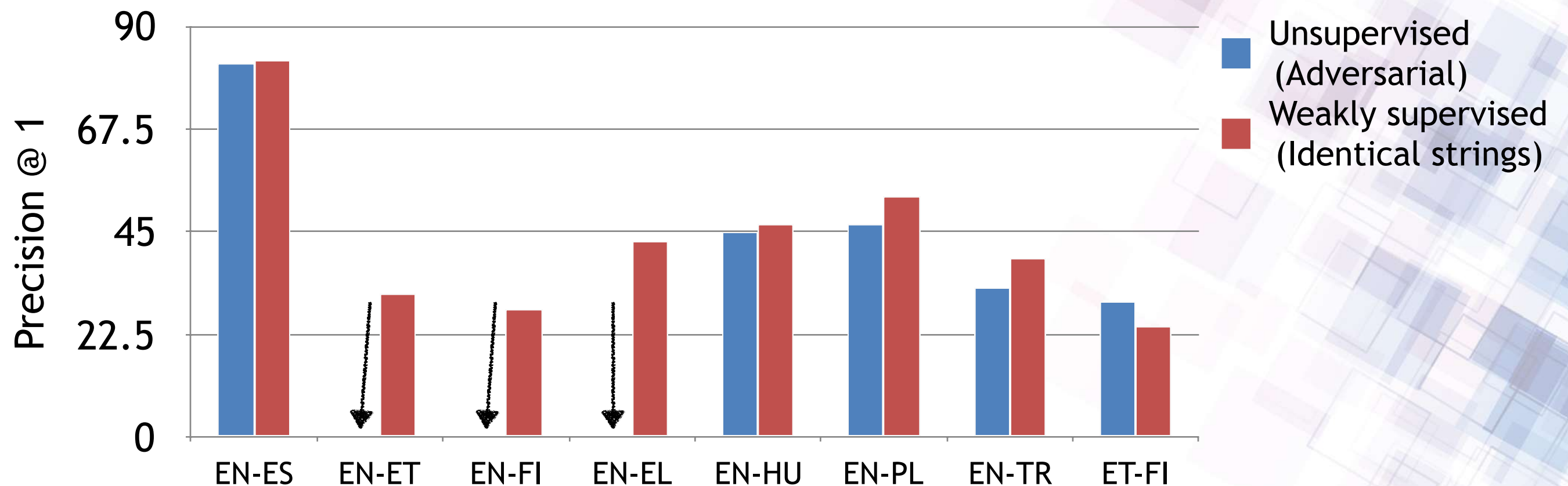
1. Monolingual word embeddings
2. Adversarial mapping
3. Refinement (Procrustes analysis)
4. Cross-domain similarity local scaling (CSLS)



(Søgaard, Ruder & Vulić, ACL 2018)



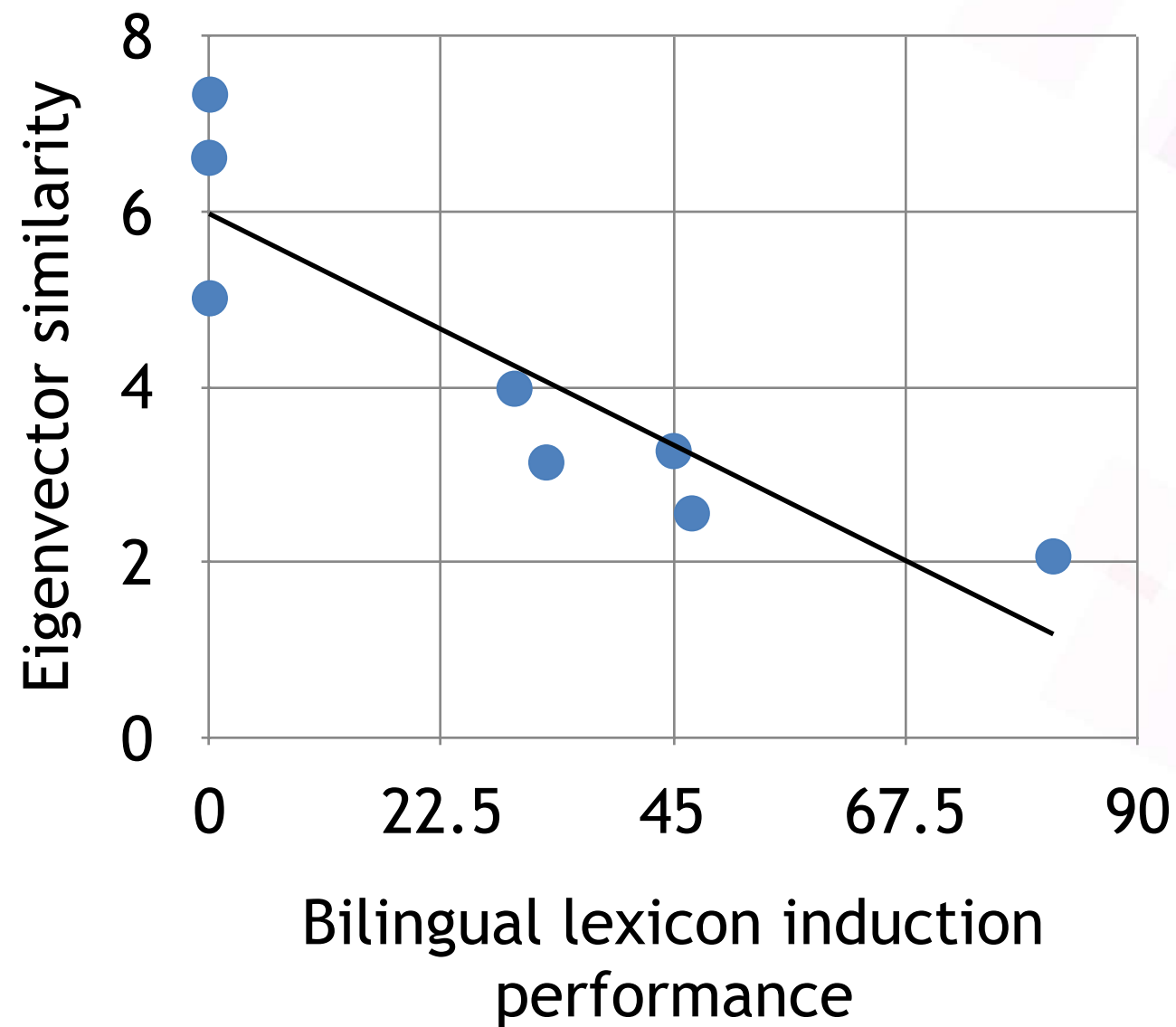
## Impact of language similarity



- Unsupervised approaches are challenged by languages that are not isolating and not dependent marking
- Weak supervision leads to competitive performance on similar language pairs and better results for dissimilar pairs

(Søgaard, Ruder & Vulić, ACL 2018)

## Impact of language similarity

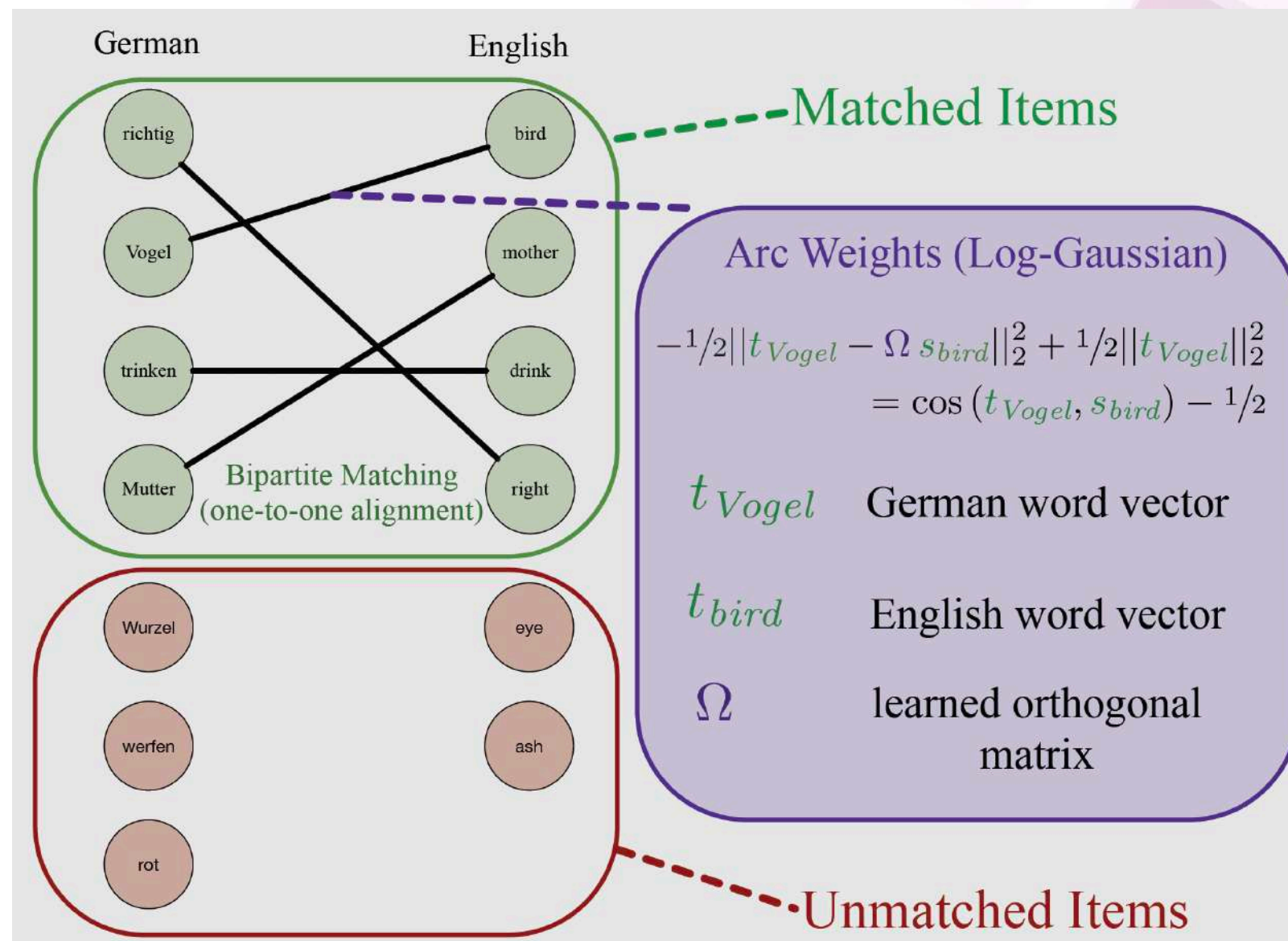


- Eigenvector similarity strongly correlates with BDI performance ( $\rho \sim 0.89$ )

(Søgaard, Ruder & Vulić, ACL 2018)

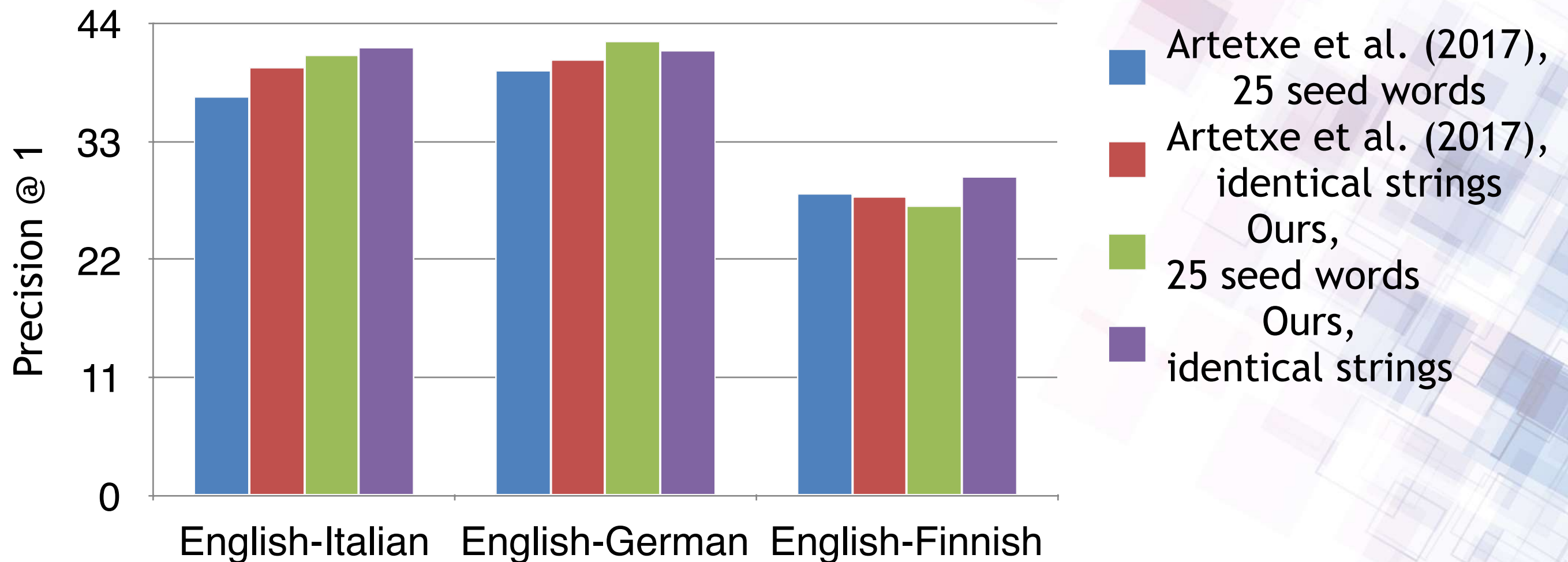


# A discriminative latent-variable model



(Ruder & Cotterell et al., EMNLP 2018)

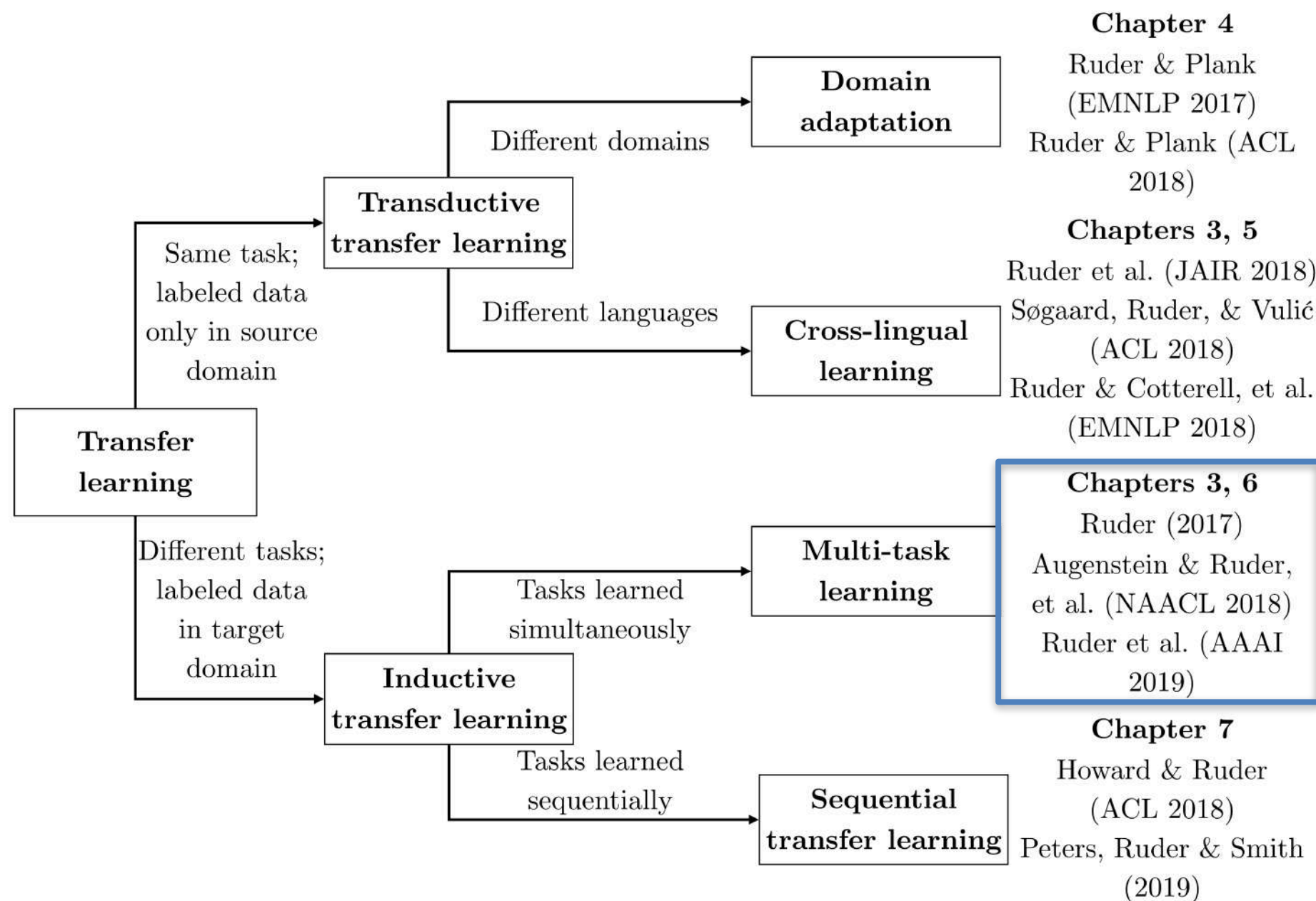
## Bilingual lexicon induction performance



- Our model outperforms the state of the art for bilingual lexicon induction

(Ruder & Cotterell et al., EMNLP 2018)

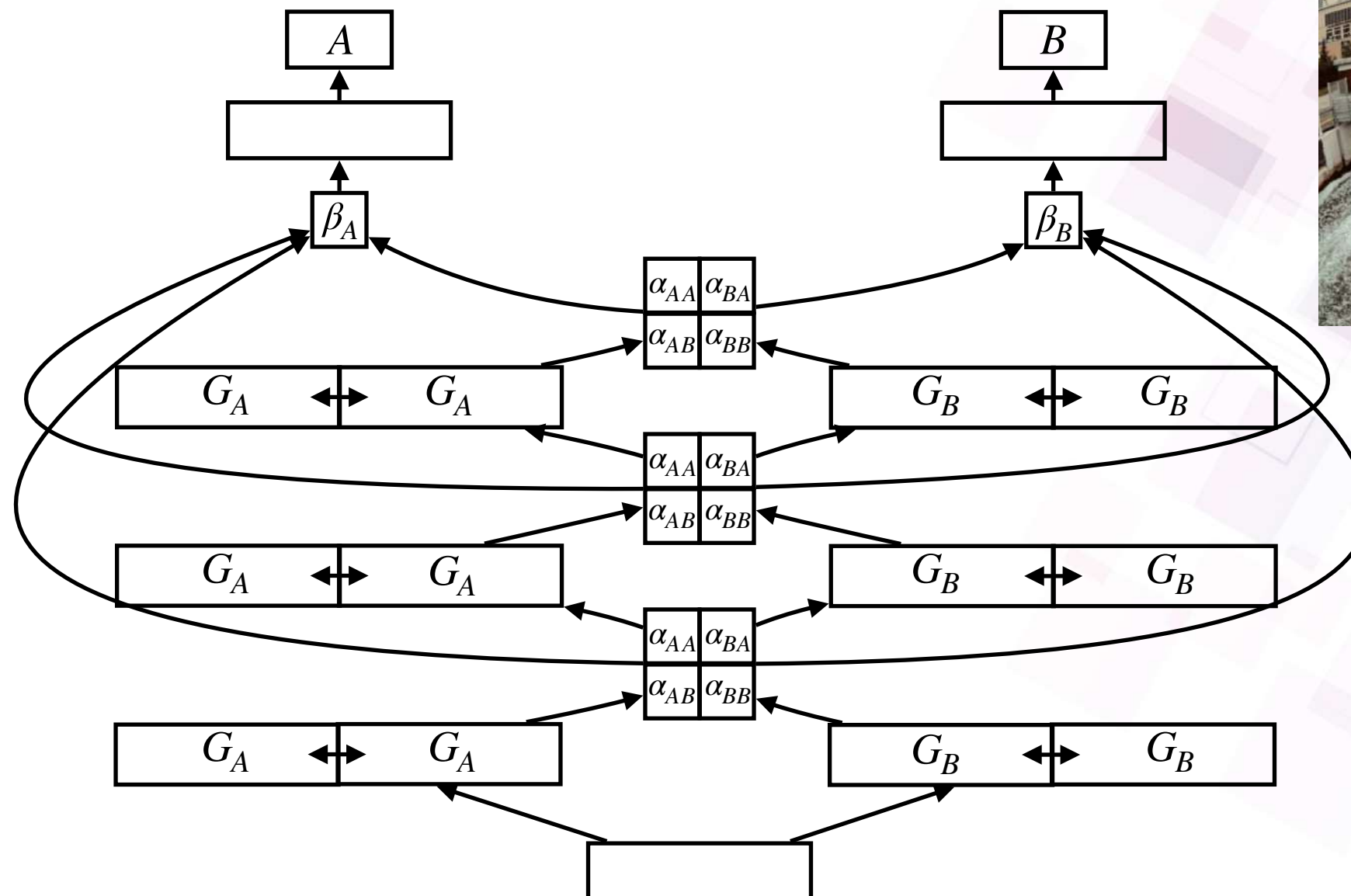




# Improved Sharing in Multi-task Learning

- Propose two novel architectures that enable more flexible sharing between tasks:
  1. Sluice networks, a meta-architecture that learns how tasks should share information ([Ruder et al., AAAI 2019](#));
  2. Label embedding layer and a label transfer network that enables using information from related label spaces more effectively ([Augenstein & Ruder et al., NAACL 2018](#)).

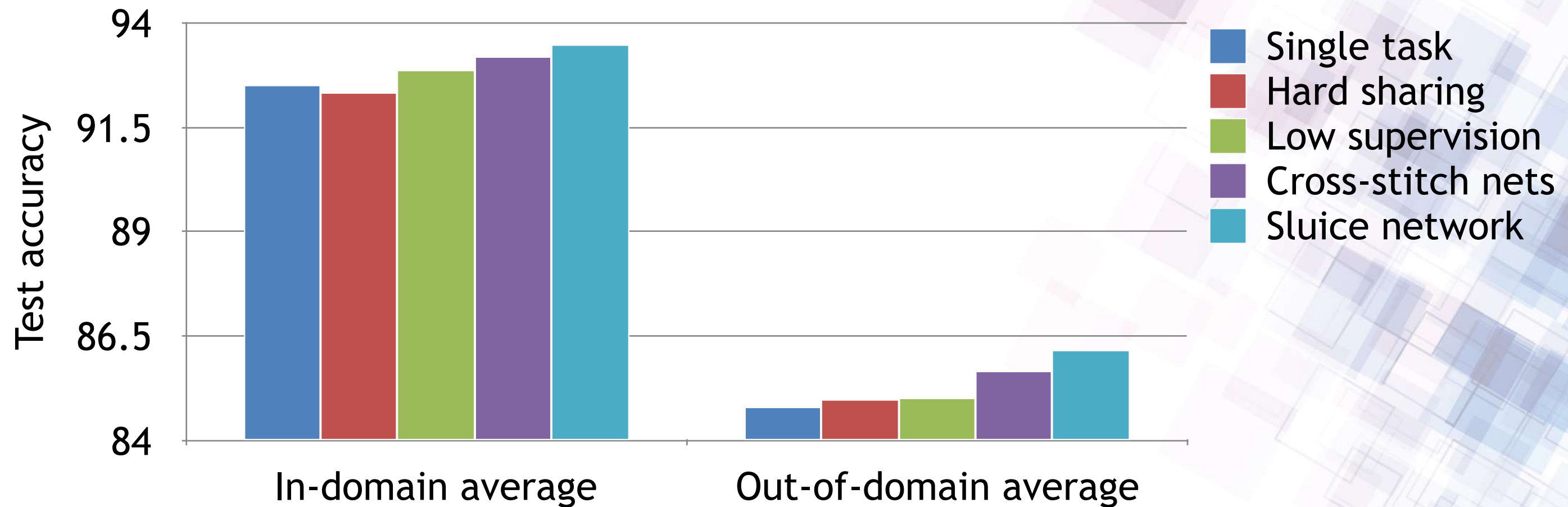




- **A sluice network:**  $\alpha$  and  $\beta$  parameters mediate traffic of information similar to how a sluice controls the flow of water

(Ruder et al., AAI 2019)

## Chunking (main task) + POS tagging (aux task)

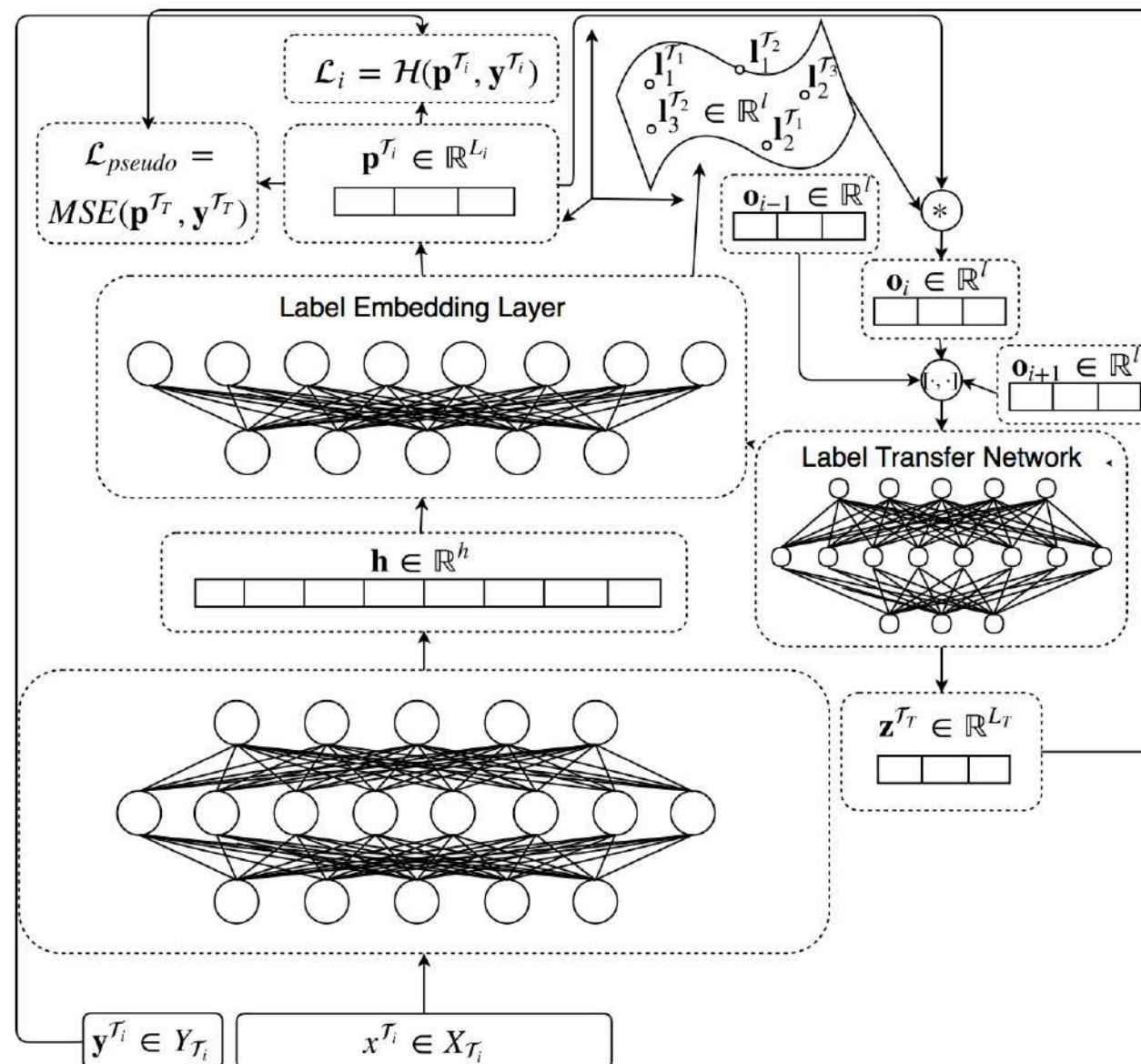


- Sluice networks significantly outperform baselines on both in-domain and out-of-domain data

(Ruder et al., AACL 2019)



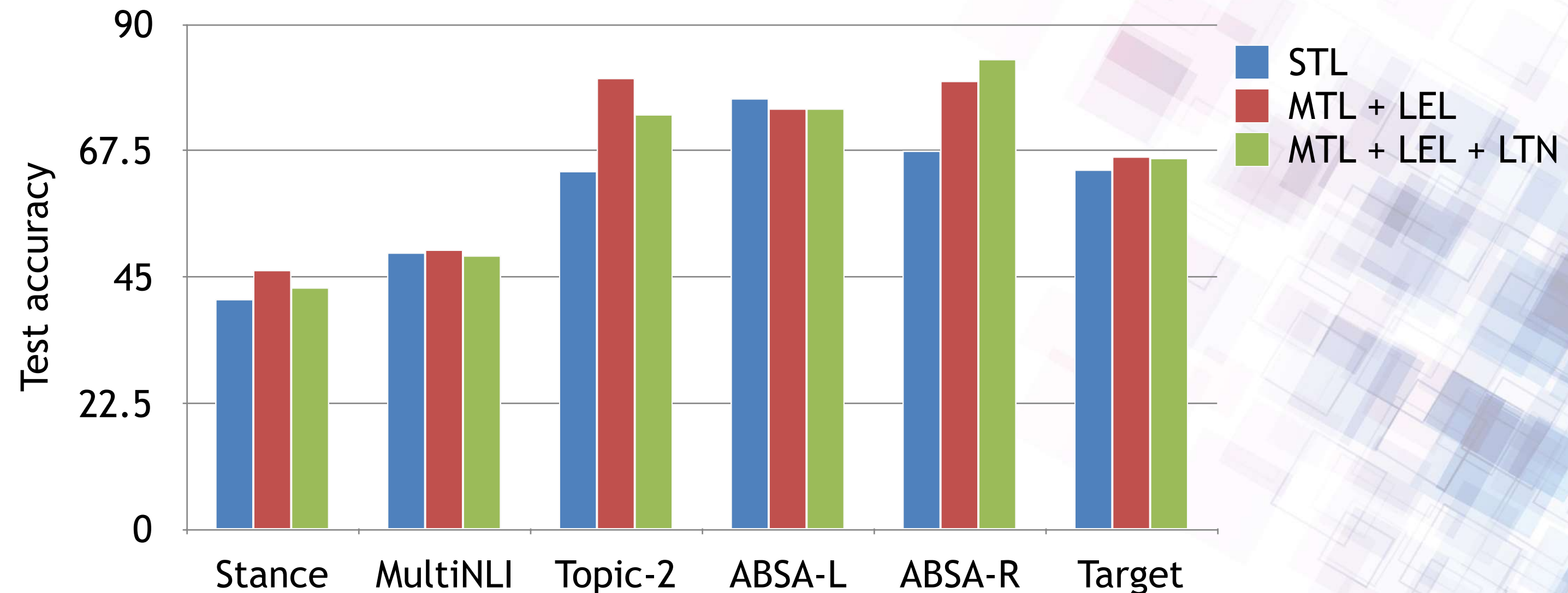
# Label Embedding Layer and Label Transfer Network



- **Label Embedding Layer (LEL):** embeds labels in a joint embedding space based on a compatibility function
- **Label Transfer Network (LTN):** learns a function to map auxiliary task labels to target task

(Augenstein & Ruder et al., NAACL 2018)

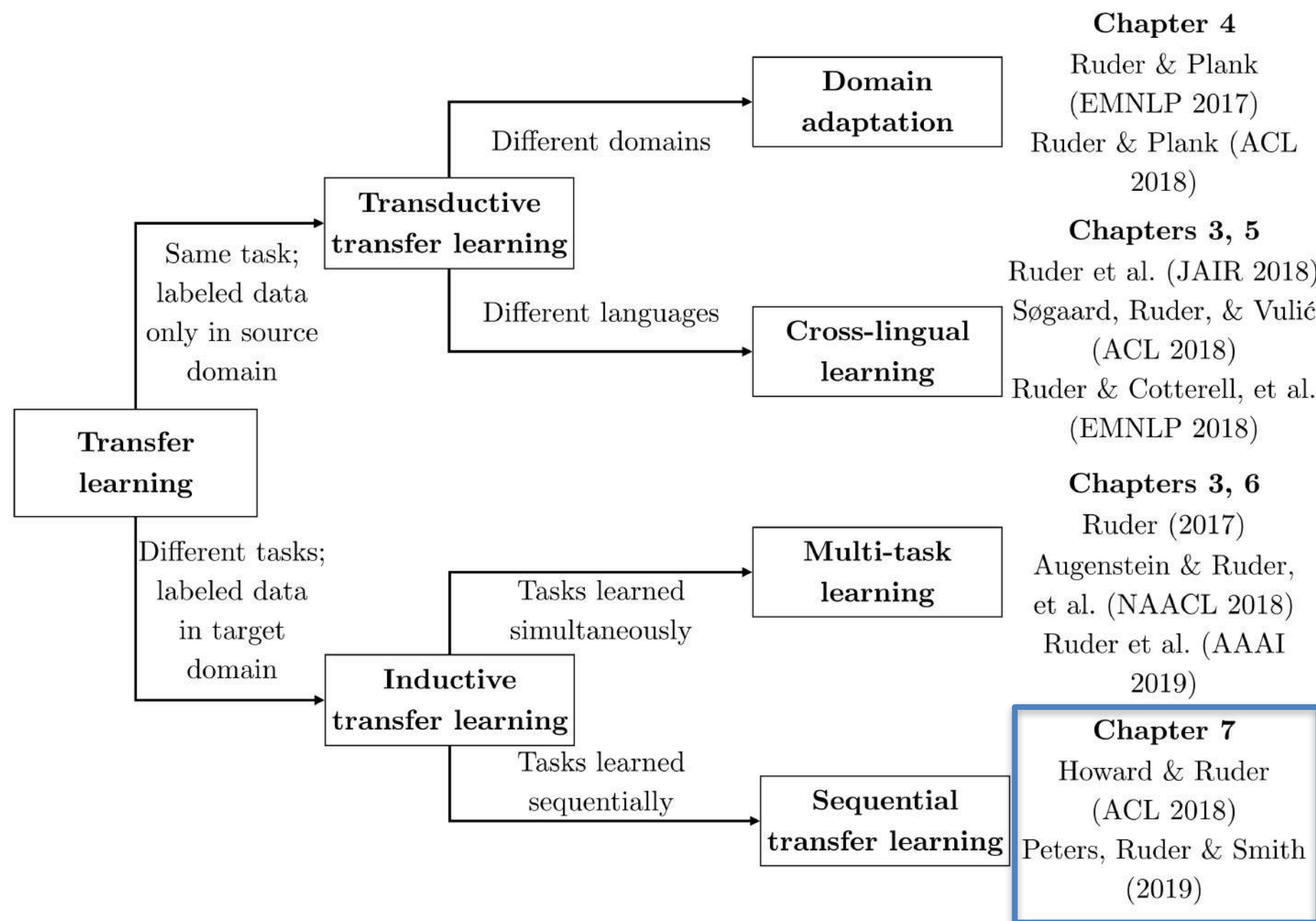
## Performance on sentence pair classification tasks



- Our multi-task learning models outperform single-task learning on most tasks
- Achieve state of the art on aspect and topic-based sentiment analysis

(Augenstein & Ruder et al., NAACL 2018)



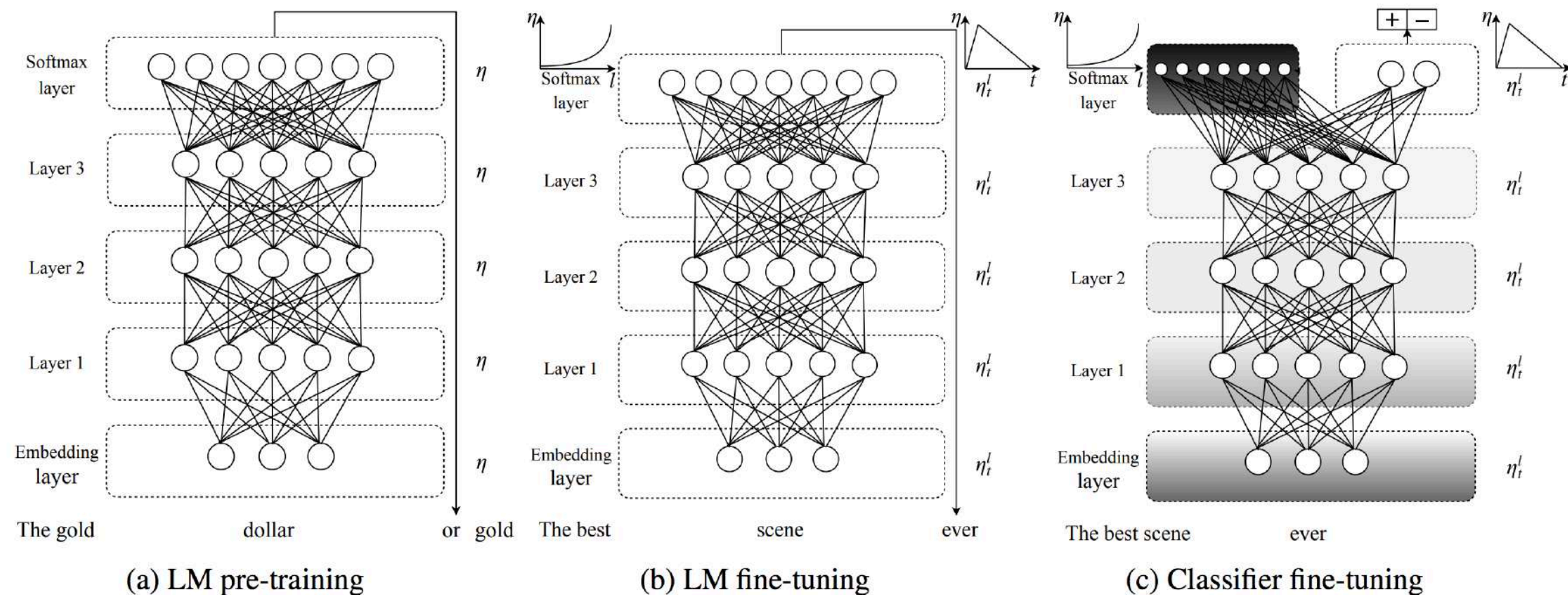


# Adapting Universal Pretrained Representations

- Analyse and propose techniques for the adaptation phase of sequential transfer learning:
  1. Universal Language Model Fine-tuning (ULMFiT), a novel framework for pretraining and adapting learned representations ([Howard & Ruder, ACL 2018](#));
  2. Compare the two main adaptation paradigms across a wide range of settings ([Peters, Ruder & Smith, 2019](#)).



# Universal Language Model Fine-tuning



- ULMFiT consists of three phases:
  - Train language model (LM) on general domain data.
  - Fine-tune LM on unlabeled target data.
  - Train classifier on top of LM on labeled target data.

(Howard & Ruder, ACL 2018)



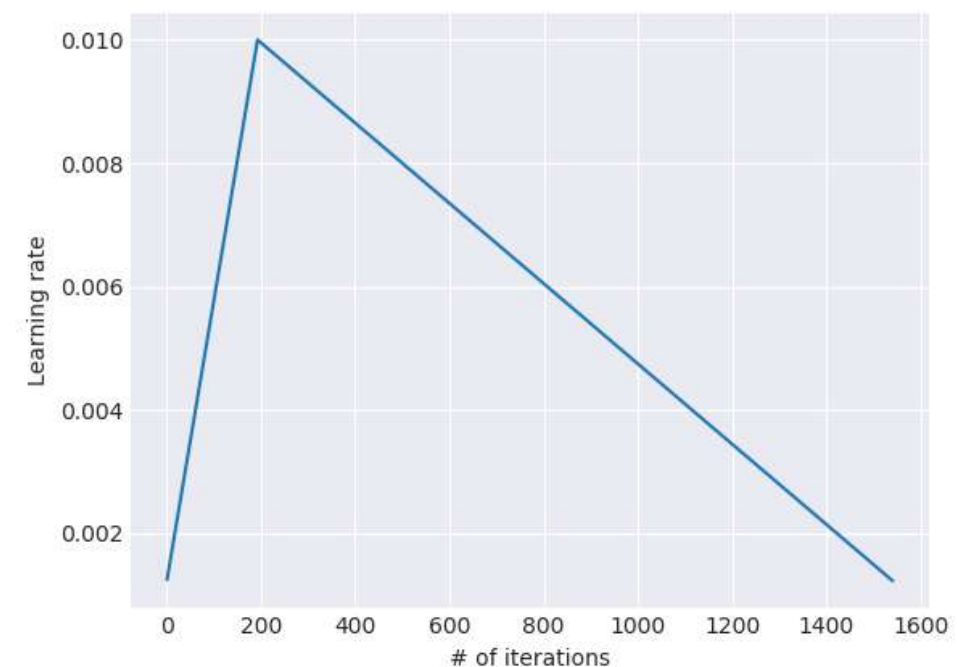
# Adaptation techniques

- **Discriminative fine-tuning**

Different layers capture *different types of information*. They should be fine-tuned to *different extents* with different learning rates:  $\theta_t^l = \theta_{t-1}^l - \eta^l \cdot \nabla_{\theta^l} J(\theta)$

- **Slanted triangular learning rates**

The model should converge quickly to a suitable region and then refine its parameters.



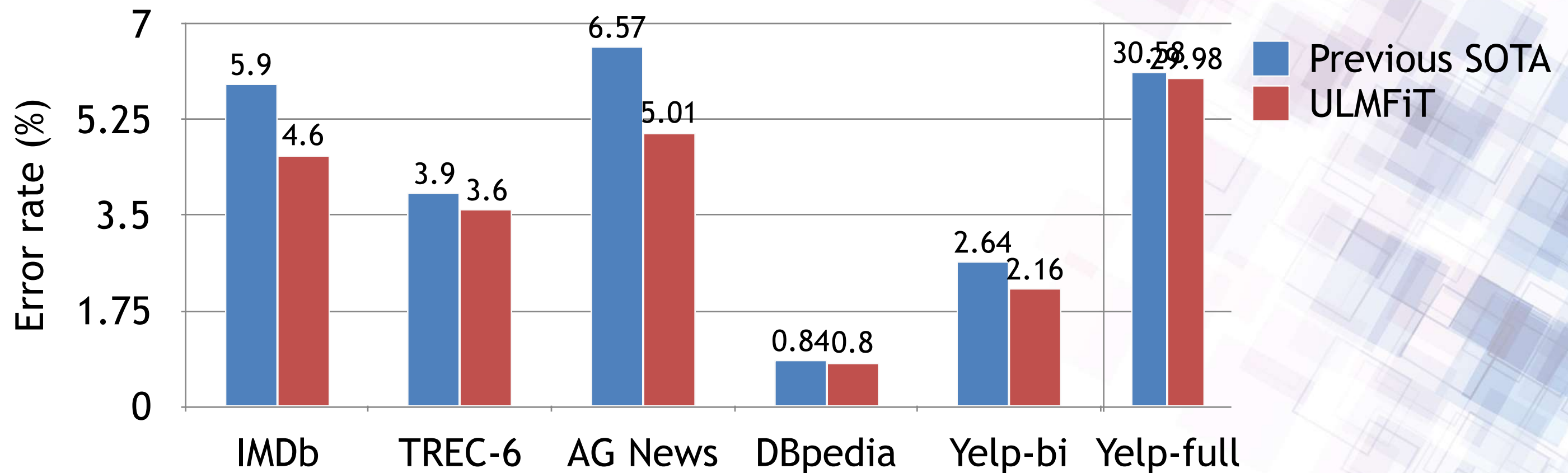
- **Gradual unfreezing**

Gradually unfreeze the layers starting from the last layer to prevent catastrophic forgetting.

(Howard & Ruder, ACL 2018)



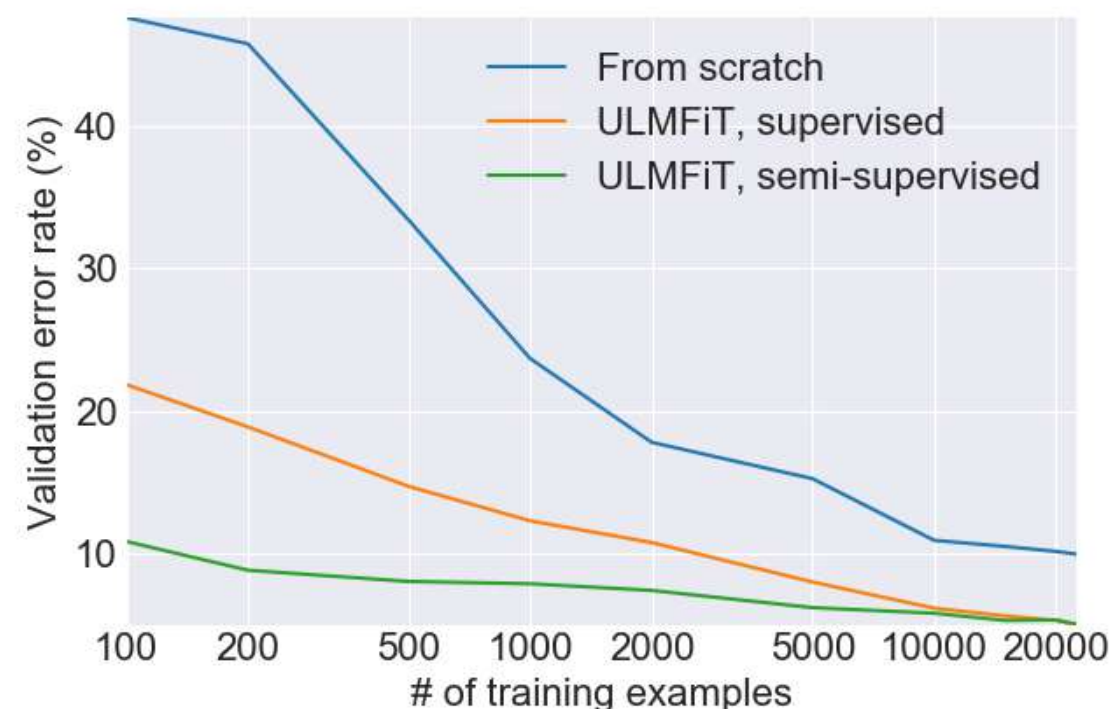
## Previous state of the art vs. ULMFiT



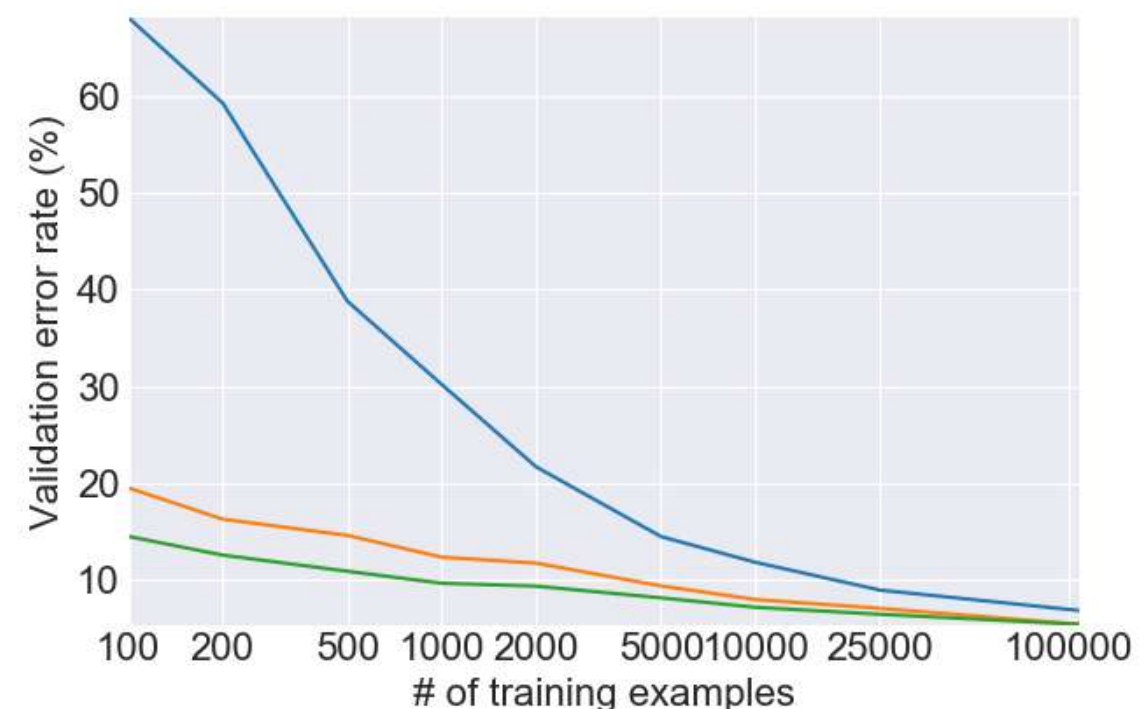
- ULMFiT outperforms the state of the art on six widely studied text classification datasets with 18-24% error reduction on the majority of datasets

(Howard & Ruder, ACL 2018)

# Few-shot learning results



IMDb



AG News

- **100 labeled examples:** ULMFiT matches performance of training from scratch with 10x and 20x more data respectively
- **100 labeled examples + 50-100k unlabeled in-domain examples:** ULMFiT matches performance of training from scratch with 50x and 100x more data respectively

(Howard & Ruder, ACL 2018)



# Guidelines for adapting pretrained representations

Pretraining	Conditions		Guidelines
	Adaptation	Task	
Any	Feature extraction	Any	Add many task parameters
Any	Fine-tuning	Any	Add minimal task parameters Hyper-parameters!
Any	Any	Seq. / clas.	Similar performance with both
ELMo	Fine-tuning	Any	Use ULMFiT techniques
ELMo	Any	Sent. pair	Use Feature extraction
BERT	Any	Sent. pair	Use Fine-tuning

(Peters, Ruder & Smith, 2019)

# Conclusion

- We have proposed neural network-based methods for NLP that transfer information from related domains, tasks, and languages.
- Our models outperformed models not using this information as well as state-of-the-art transfer learning approaches.



# Contributions

- Proposed methods that...

- 1. Overcome a discrepancy between the source and target setting.**  
*Bayesian Optimisation, weak supervision, flexible sharing, language models, ...*
- 2. Induce an inductive bias.**  
*Semi-supervised learning, multi-task learning, orthogonality constraint, ...*
- 3. Combine traditional and current approaches.**  
*Tri-training + multi-task learning, bipartite matching + word embeddings*
- 4. Transfer across the hierarchy of NLP tasks.**  
*Sharing across low-level & high-level tasks, gradual unfreezing, fine-tuning, ...*
- 5. Generalise across many settings.**  
*Evaluation on multiple domains, tasks, and languages*