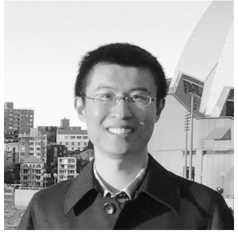
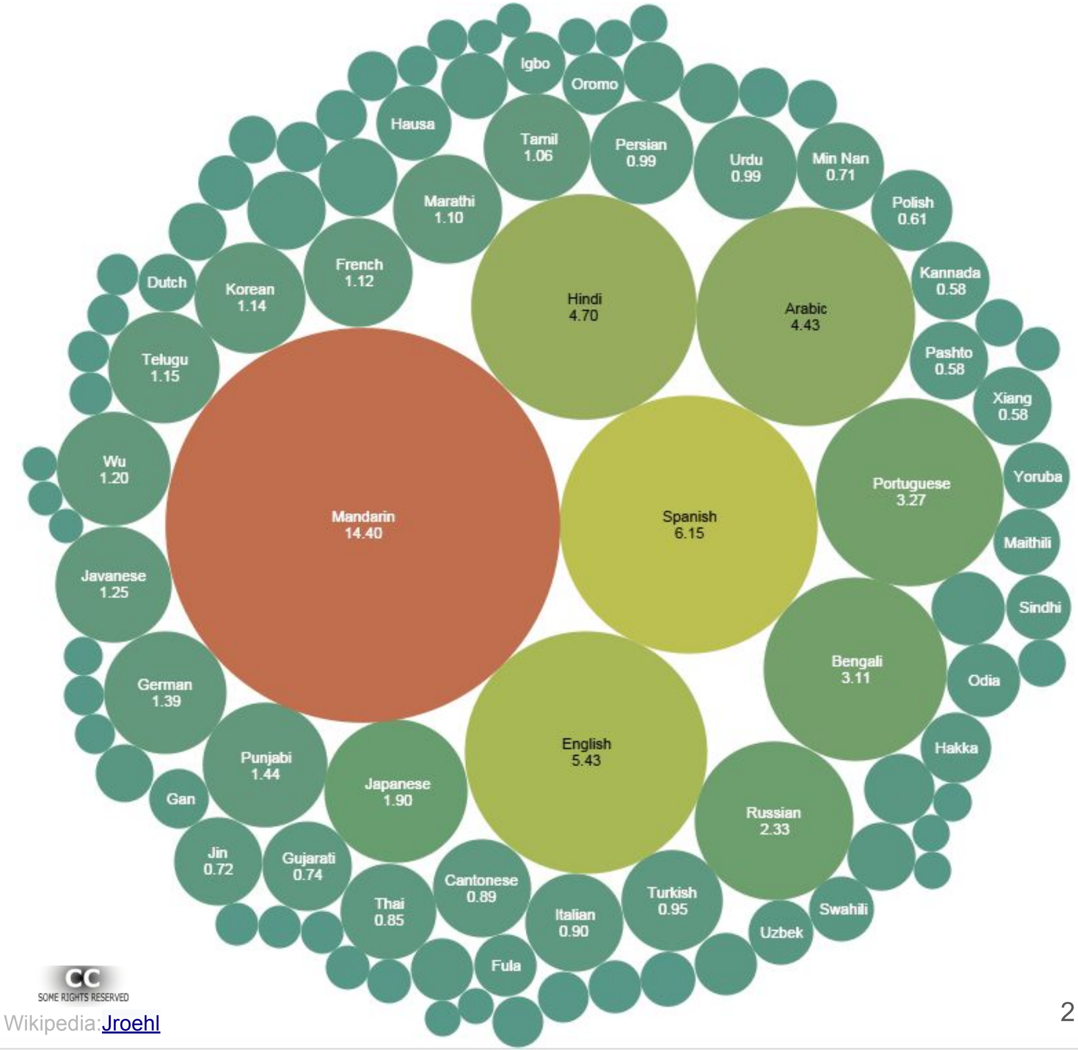
Massively Multilingual Transfer for NER

Afshin Rahimi, Yuan Li, and **Trevor Cohn**
University of Melbourne



6000+ languages

≈ **1%** with annotation



Emergency Response

Named Entity Recognition

Annotation Projection for Transfer

kailangan namin ng mas maraming dugo sa Pagasanjan . Tagalog

B-LOC

. . . .

we need more blood in Pgasanjan . English

O O O O O B-LOC O

Yarowsky et al. (2001)

Representation Projection for Transfer

kailangan namin ng mas maraming dugo sa Pagasanjan .

language independent representation

Cross-lingual word embeddings
(Lample et al., 2018)

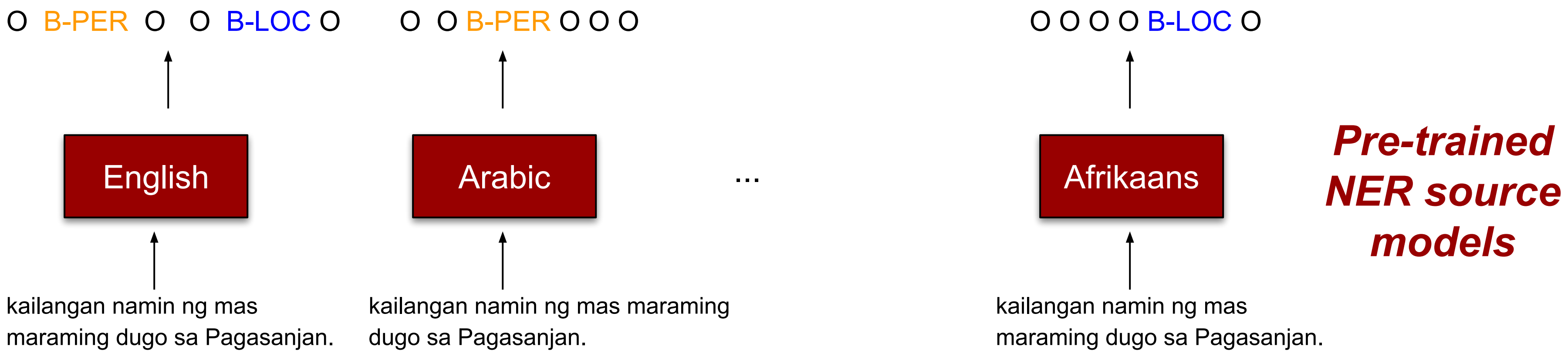
Mis-matched Model

Ideal: source-target similar in word order, script, syntax

O O O O O B-LOC O

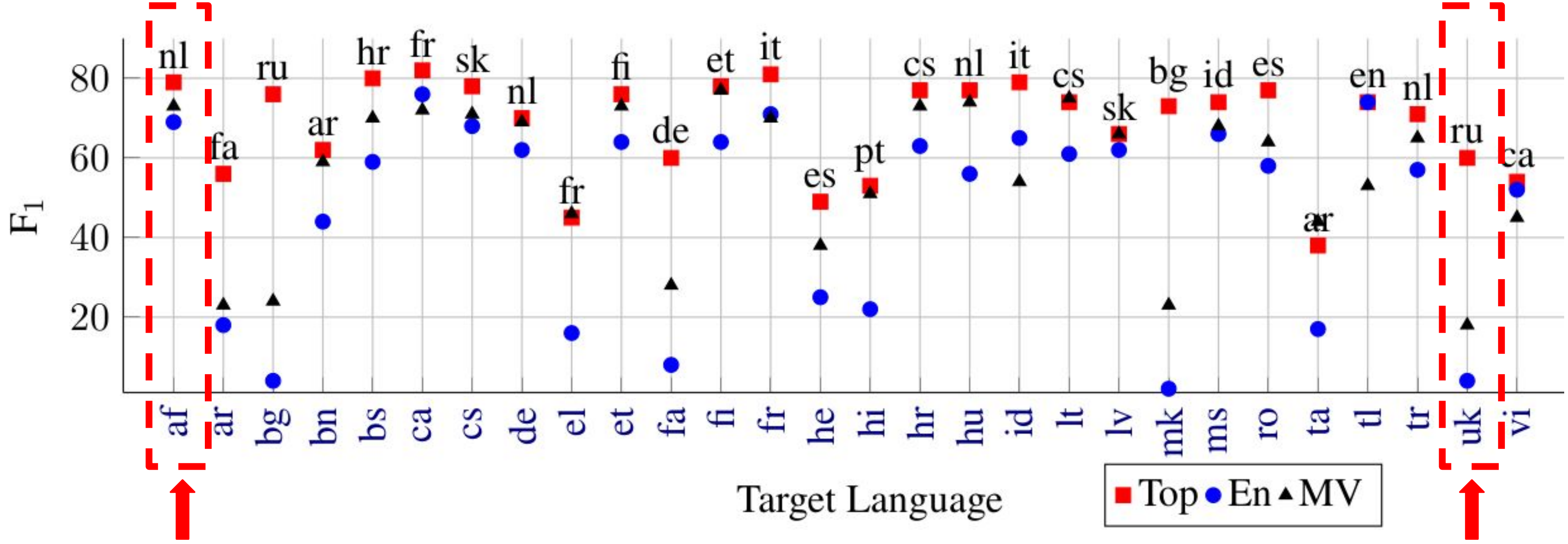
Direct Transfer for NER

Output: Labelled sentences in the target language



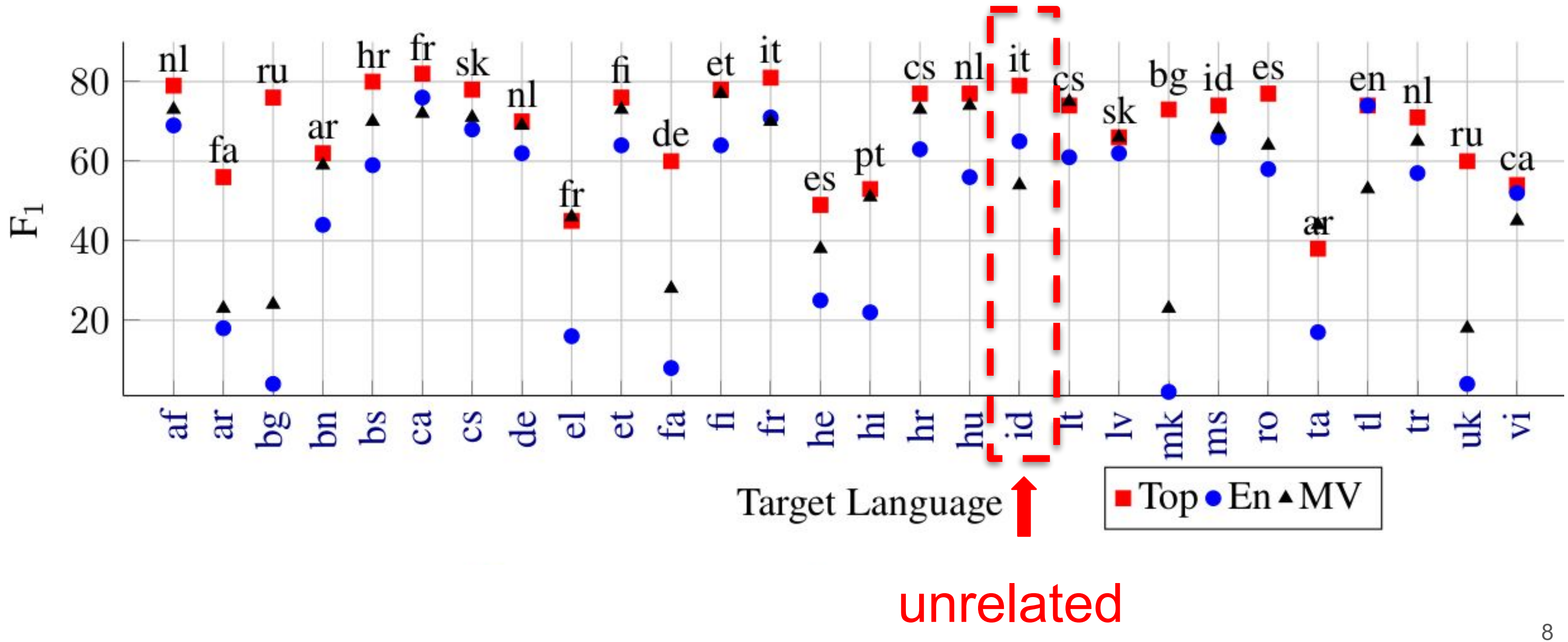
Input: Unlabelled sentences in the target language encoded with cross-lingual embeddings

Direct Transfer Results (NER F1 score, WikiANN)

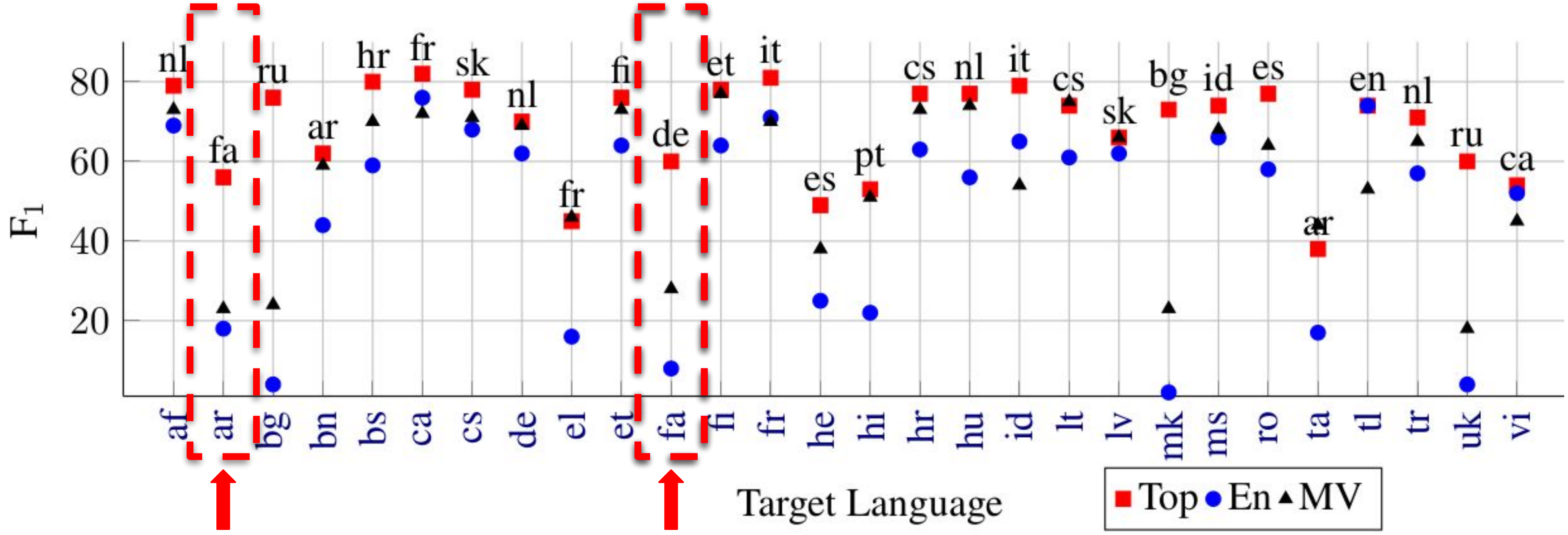


unsuprising

Direct Transfer Results (NER F1 score, WikiANN)

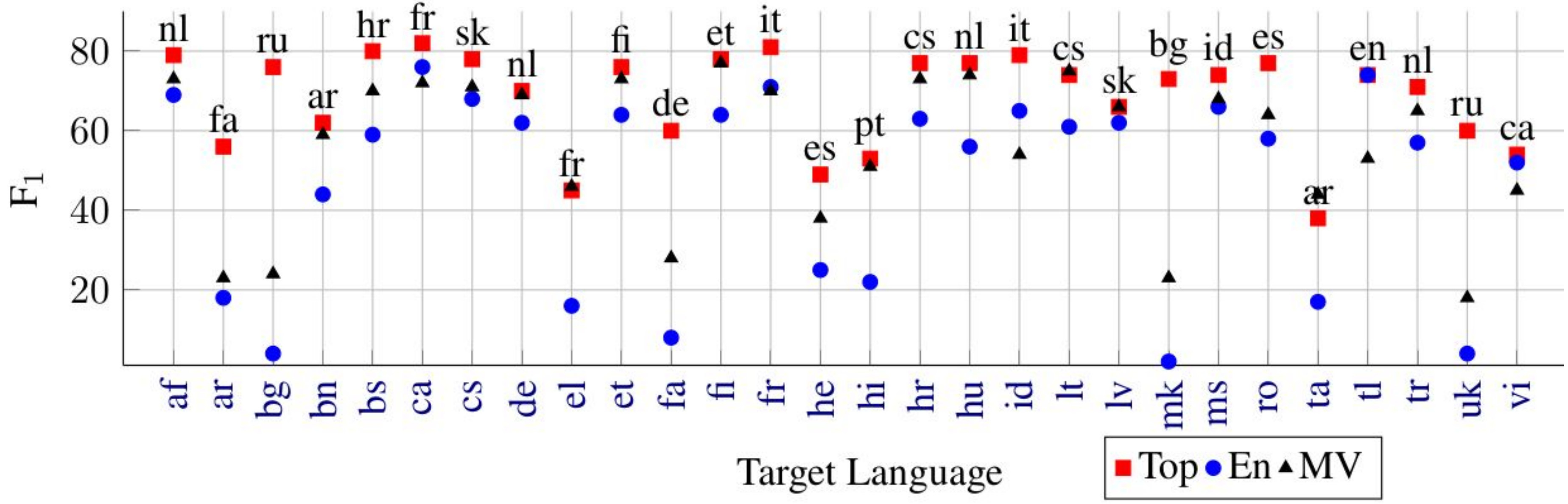


Direct Transfer Results (NER F1 score, WikiANN)



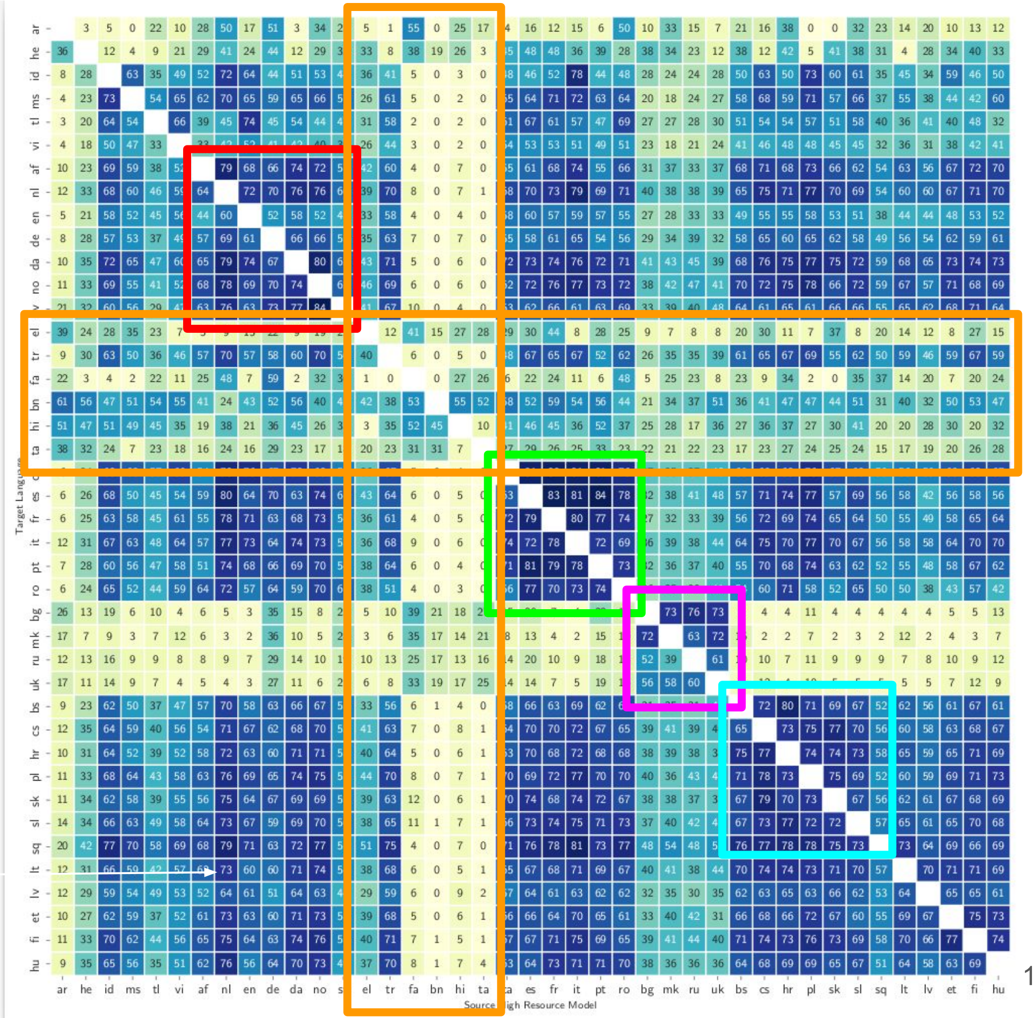
asymmetry

Voting & English are often poor!



General findings

- Transfer strongest within language family (**Germanic**, **Roman**, **Slavic-Cyr**, **Slavic-Latin**)
- Asymmetry between use as source vs target language (**Slavic-Cyr, Greek/Turkish/...**)
- But lots of odd results & overall highly noisy



Problem Statement

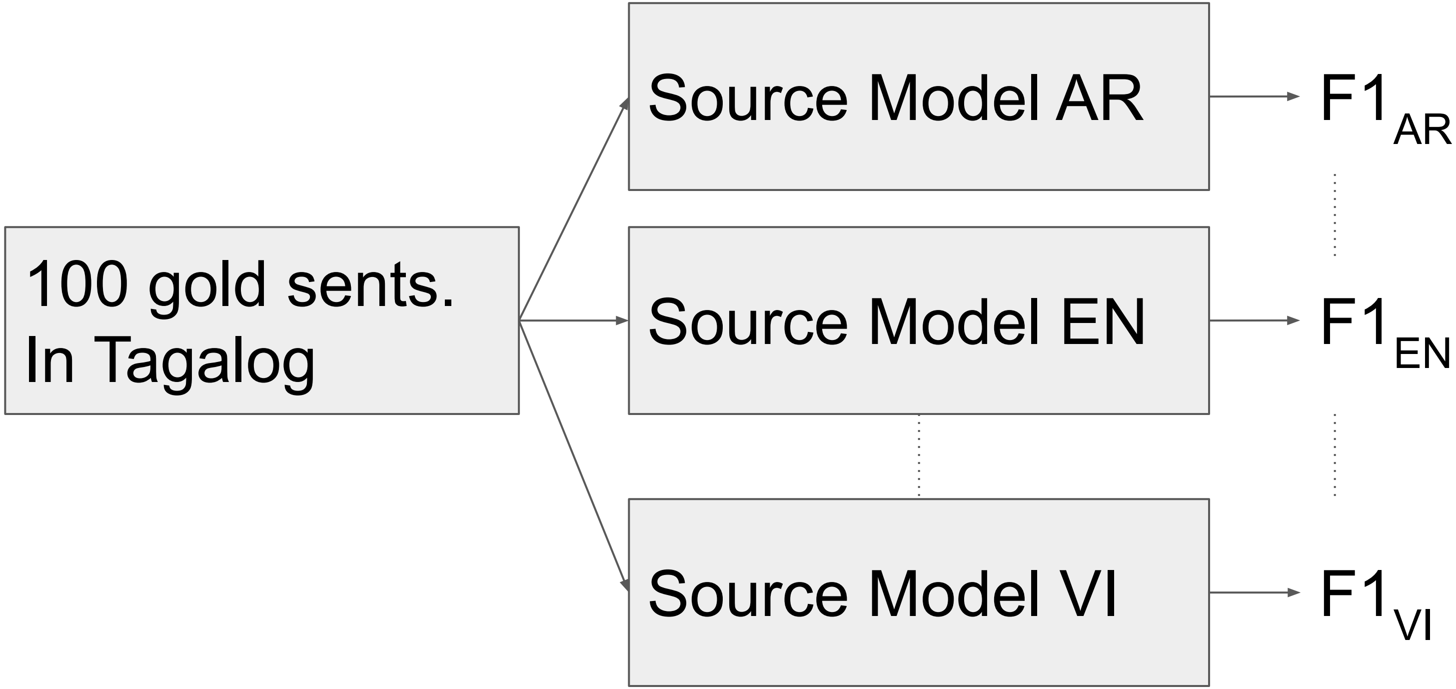
Input:

- N black-box source models
- Unlabelled data in target language
- Little or no labelled data (few shot and zero shot)

Output:

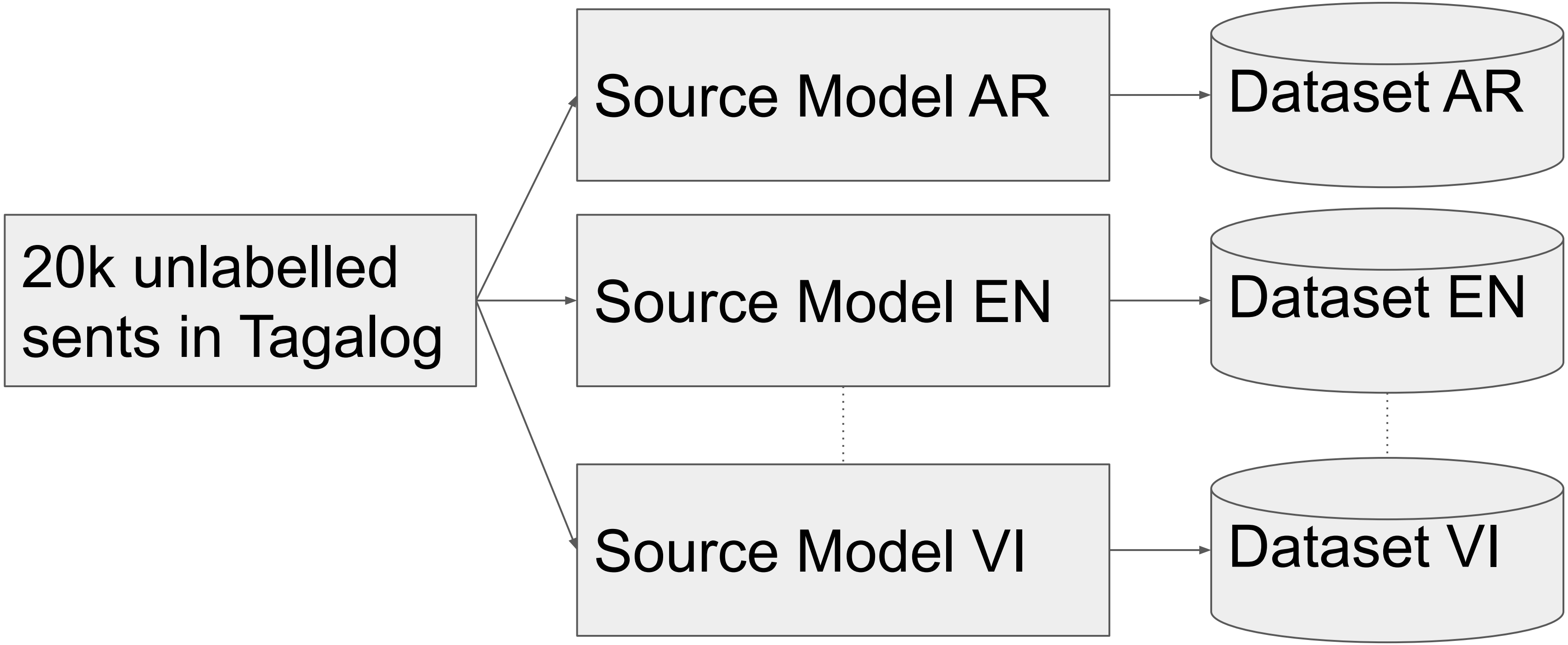
- Good predictions in the target language

Model 1: Few Shot Ranking and Retraining (RaRe)



Source model qualities

Model 1: Few Shot Ranking and Retraining (RaRe)



N training sets in Tagalog

Model 1: Few Shot Ranking and Retraining (RaRe)

$$\text{Training Set} = \sum_{l \,\in\, \text{source langs.}} g(F1_l) \times \text{Dataset } l$$

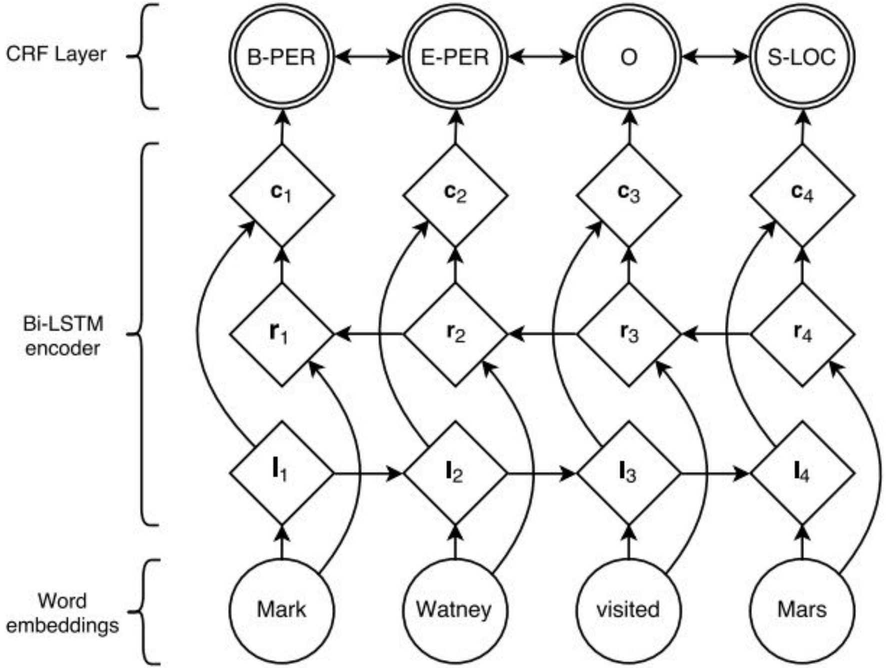
Final training set, a mixture of distilled knowledge

Model 1: Few Shot Ranking and Retraining (RaRe)

1. Train an NER model on the mixture datasets.
2. Fine-tune on 100 gold samples.

Zero-shot variant: uniform sampling without fine-tuning (**RaRe_{uns}**)

Hierarchical BiLSTM-CRF as model



Our method is **independent** of model choice.

Lample et al., (2016)

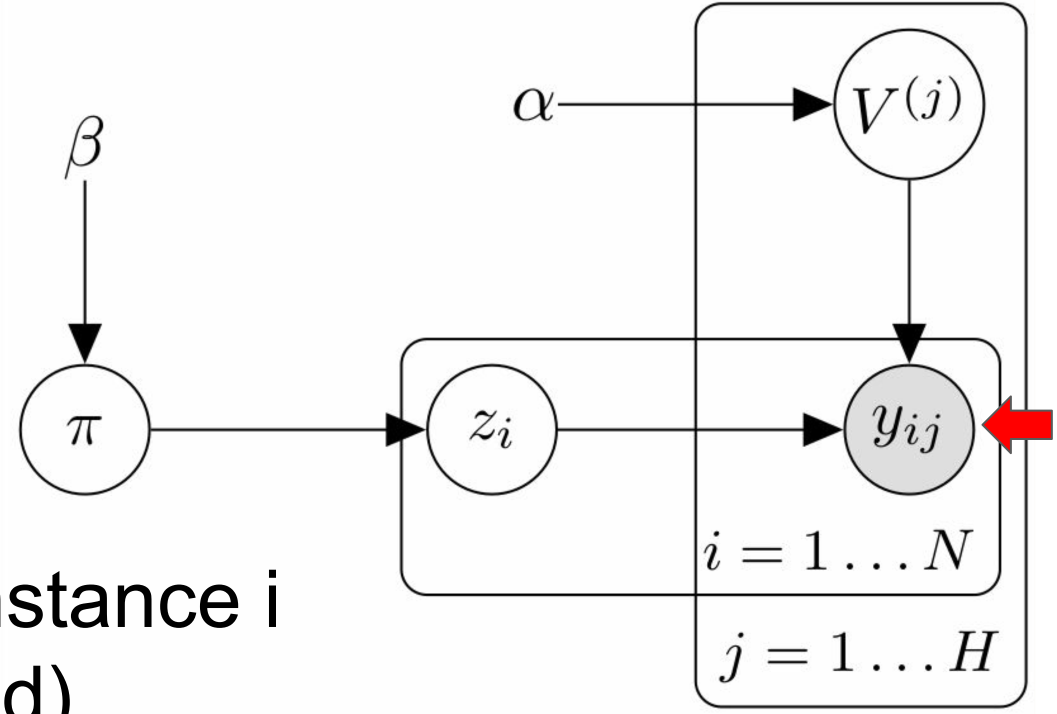
Model 2: Zero Shot Transfer (BEA)

What if no gold labels are available?

1. Treat gold labels Z as hidden variables
2. Estimate Z that best explains all the observed predictions
3. Re-estimate the quality of source models

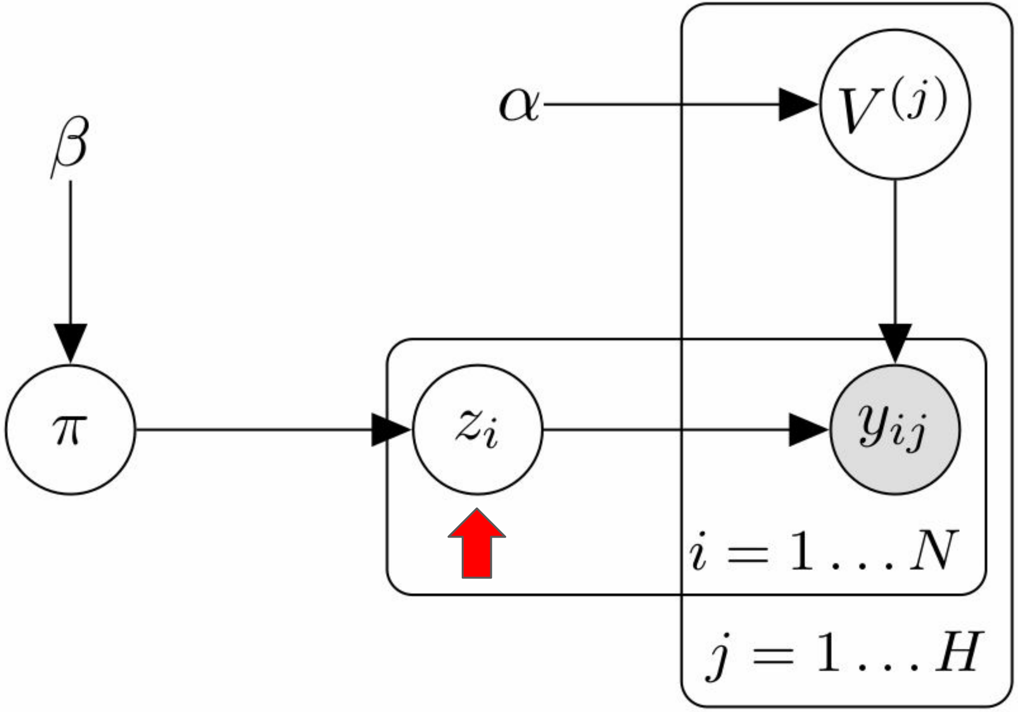
Inspired by Kim and Ghahramani (2012)

Model 2: Zero Shot Transfer (BEA)



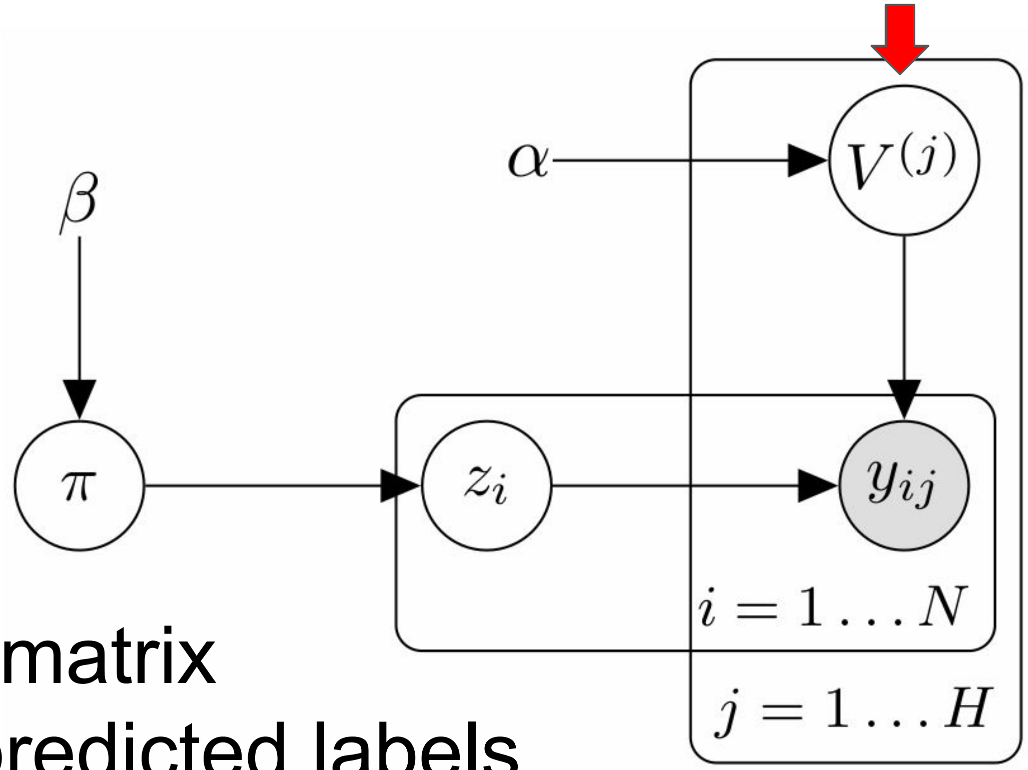
Predicted label of instance i
by model j (observed)

Model 2: Zero Shot Transfer (BEA)



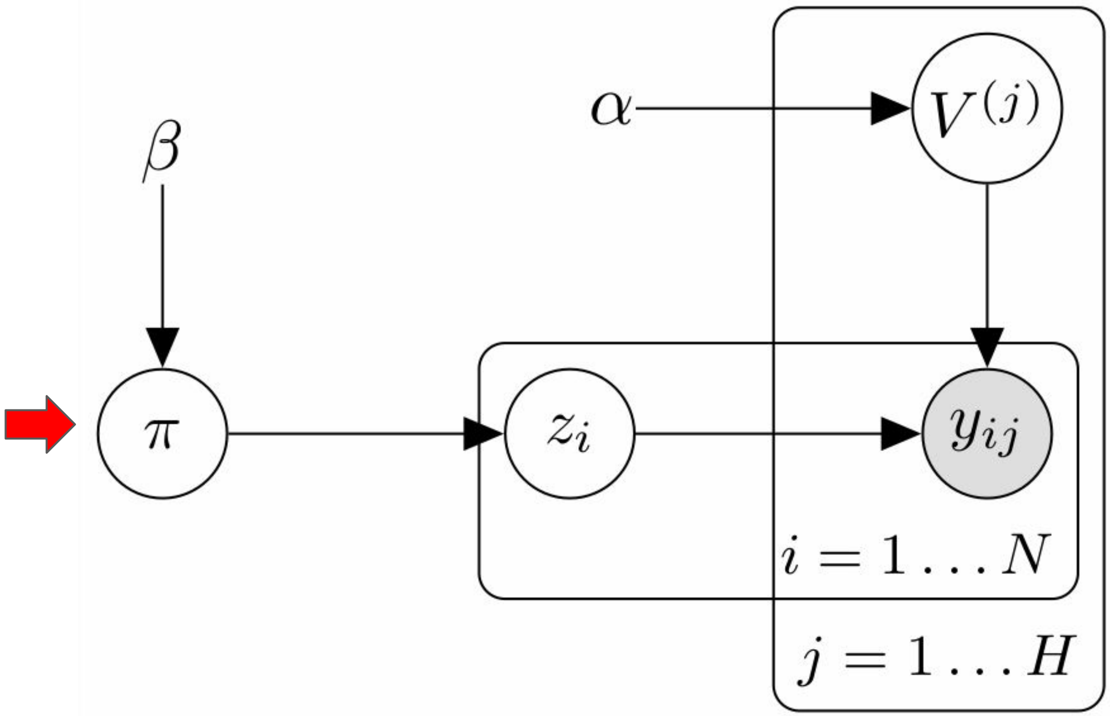
True label of instance i

Model 2: Zero Shot Transfer (BEA)



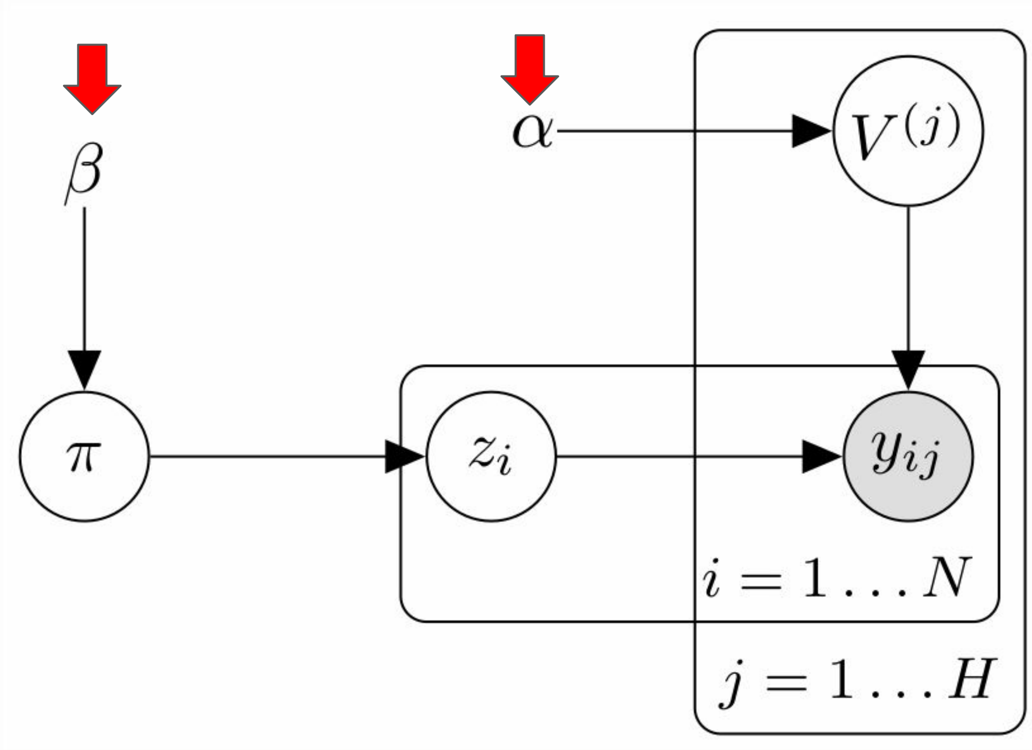
Model j's confusion matrix between True and predicted labels.

Model 2: Zero Shot Transfer (BEA)



Categorical Distribution

Model 2: Zero Shot Transfer (BEA)

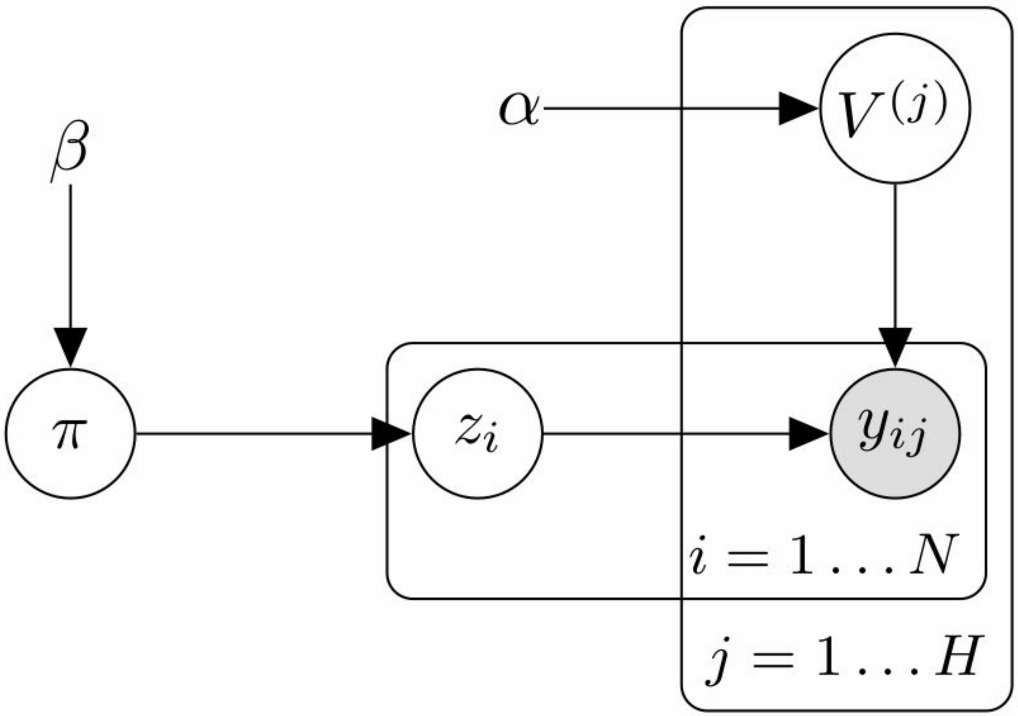


Uninformative Dirichlet Priors

Model 2: Zero Shot Transfer (BEA)

Find Z to maximises P(Z|Y,α,β), using variational mean-field approx.

Warm-start with MV.



Extensions to BEA

1. **Spammer removal:**

After running BEA, estimate source model qualities and remove bottom k, run BEA again (**BEA**$_{\textbf{unsx2}}$)

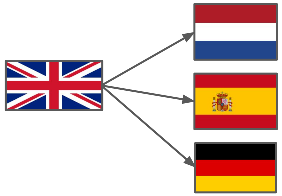
2. **Few shot scenario:**

Given 100 gold sentences, estimate source model confusion matrices, then run BEA (**BEA**$_{\textbf{sup}}$)

3. **Token vs Entity application**

Benchmark: BWET (Xie et al., 2018)

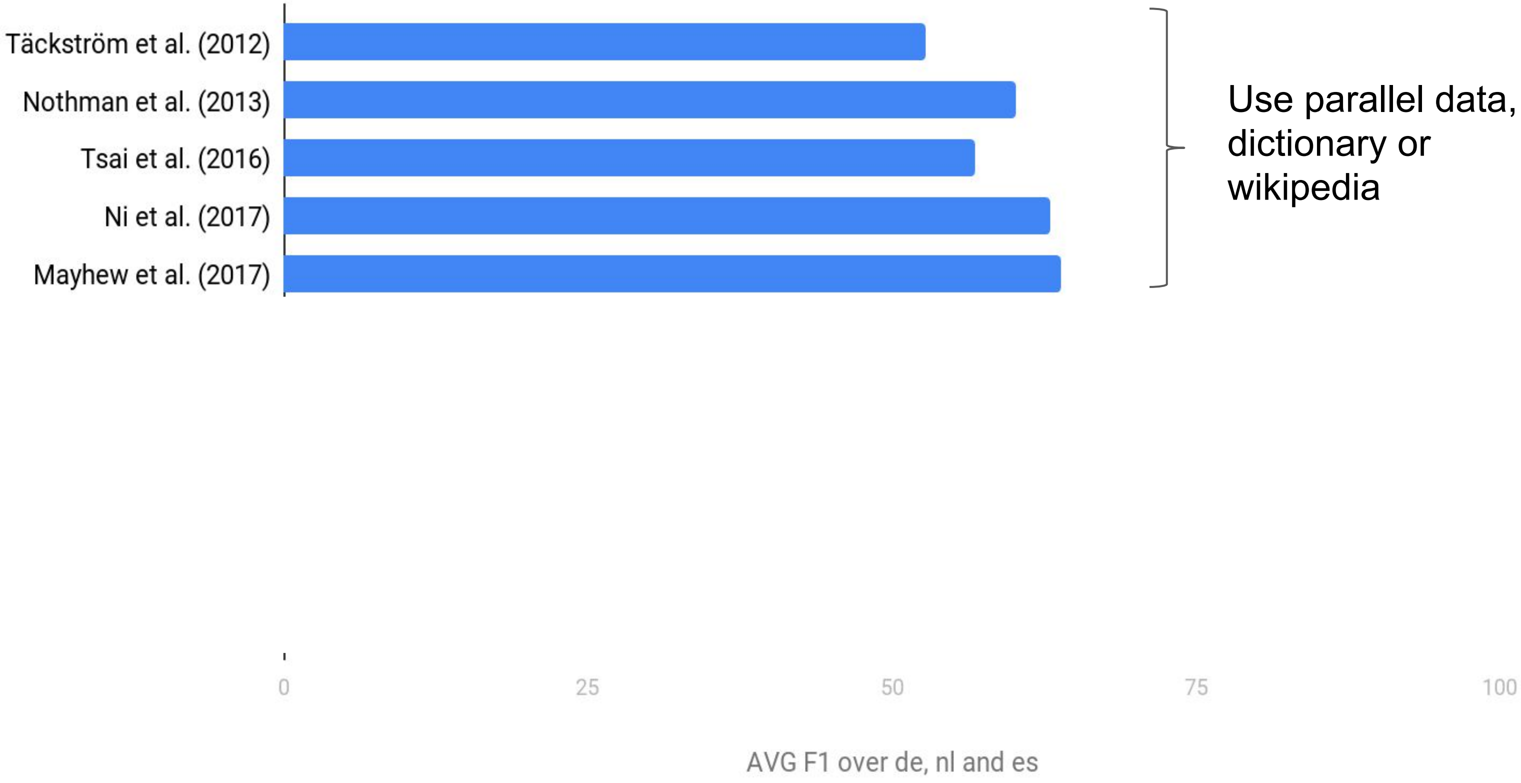
Single source annotation projection with bilingual dictionaries from cross-lingual word embeddings



- Transfer english training data to German, Dutch, and Spanish.
- Train a transformer NER on the projected training data.

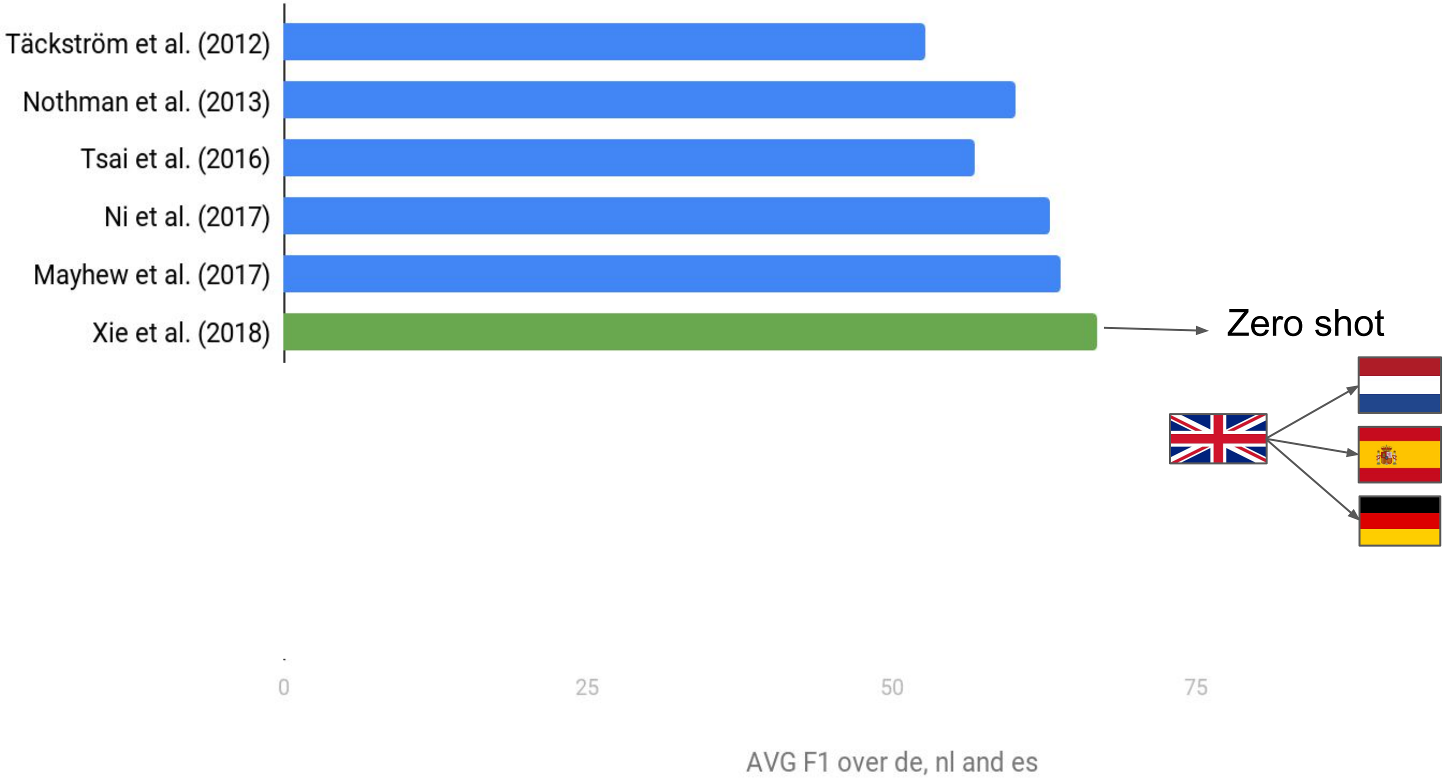
State-of-the-art on zero-shot NER transfer (orthogonal to this)

CoNLL Results (avg F1 over de, nl, es)

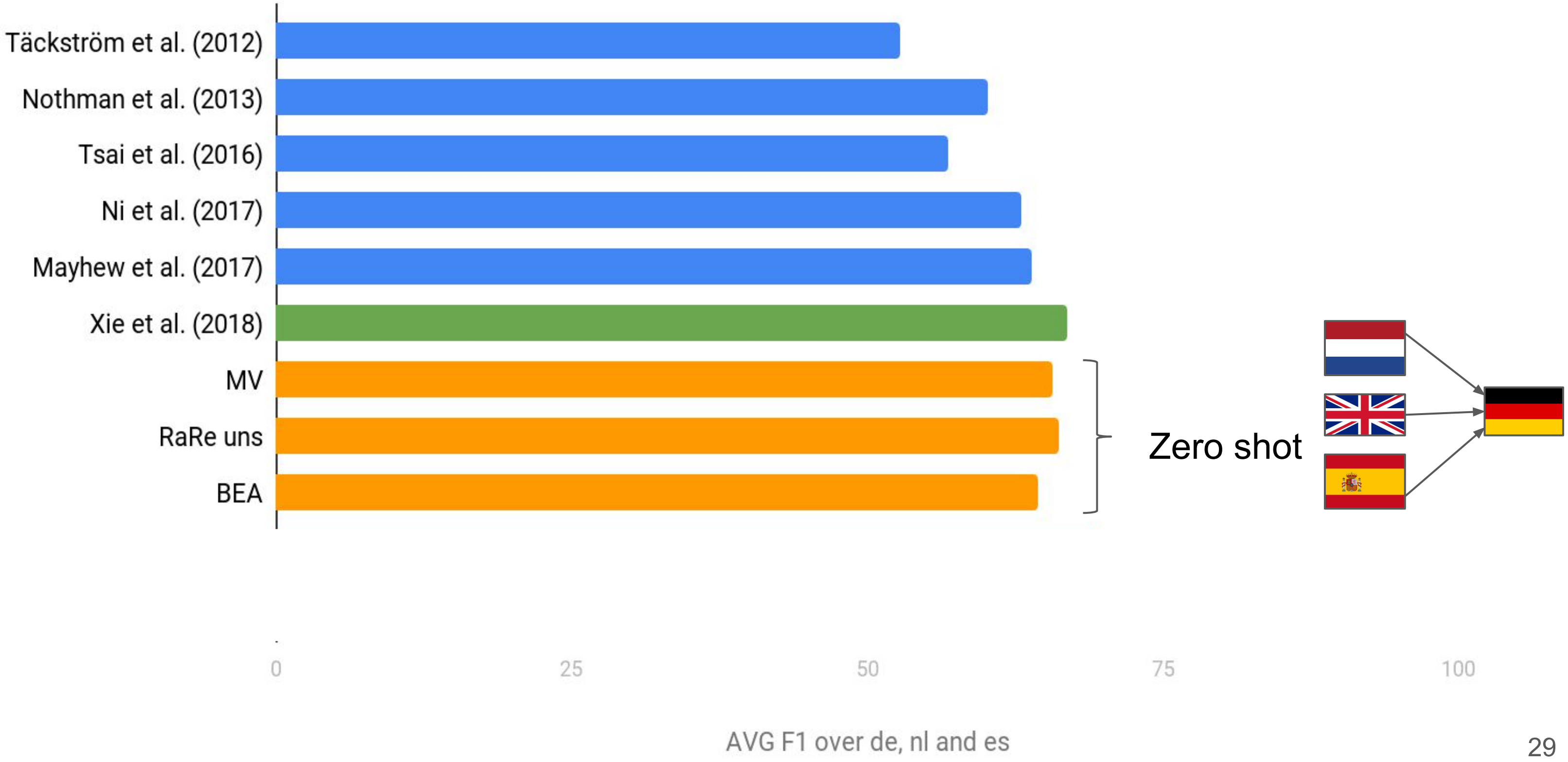


Use parallel data, dictionary or wikipedia

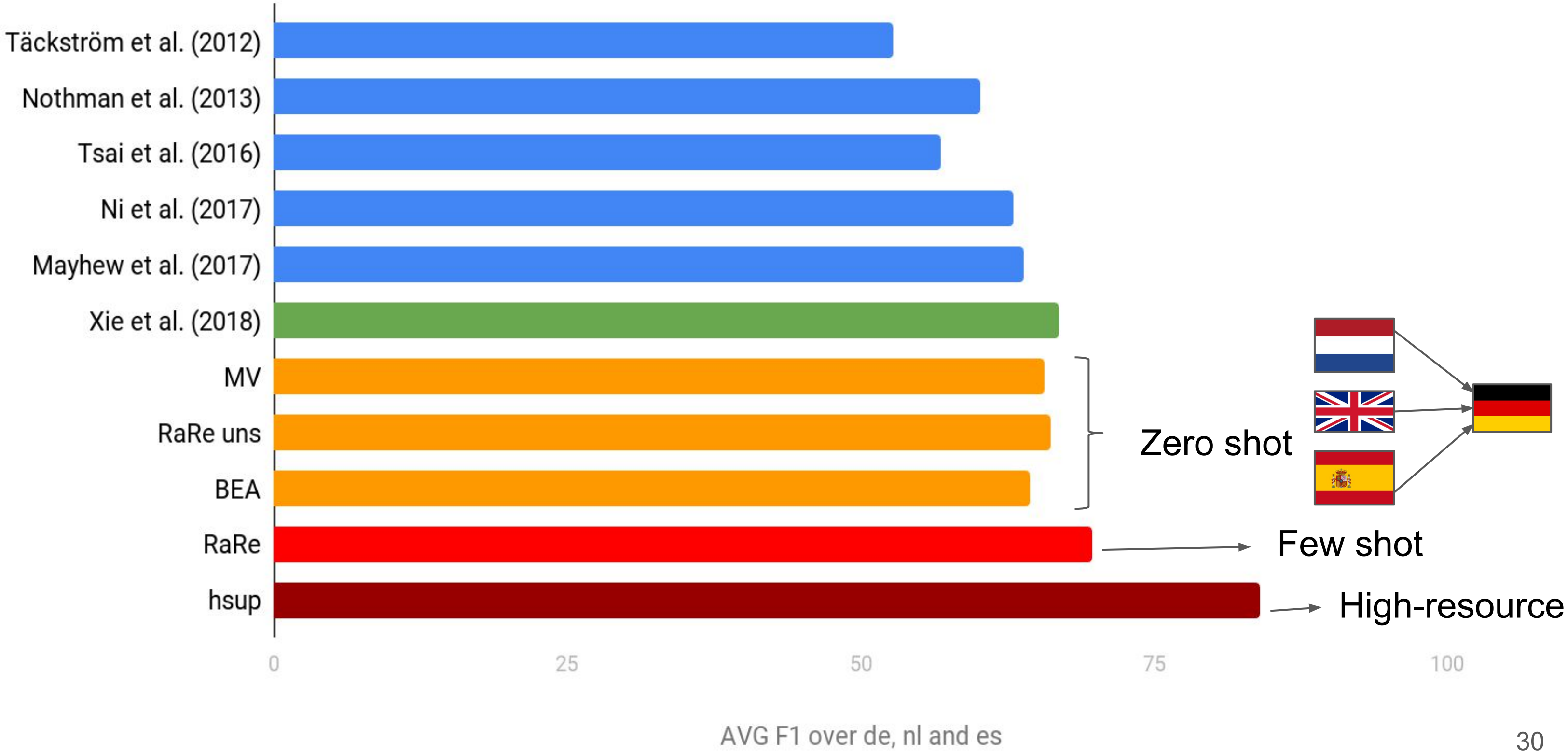
CoNLL Results (avg F1 over de, nl, es)



CoNLL Results (avg F1 over de, nl, es)



CoNLL Results (avg F1 over de, nl, es)



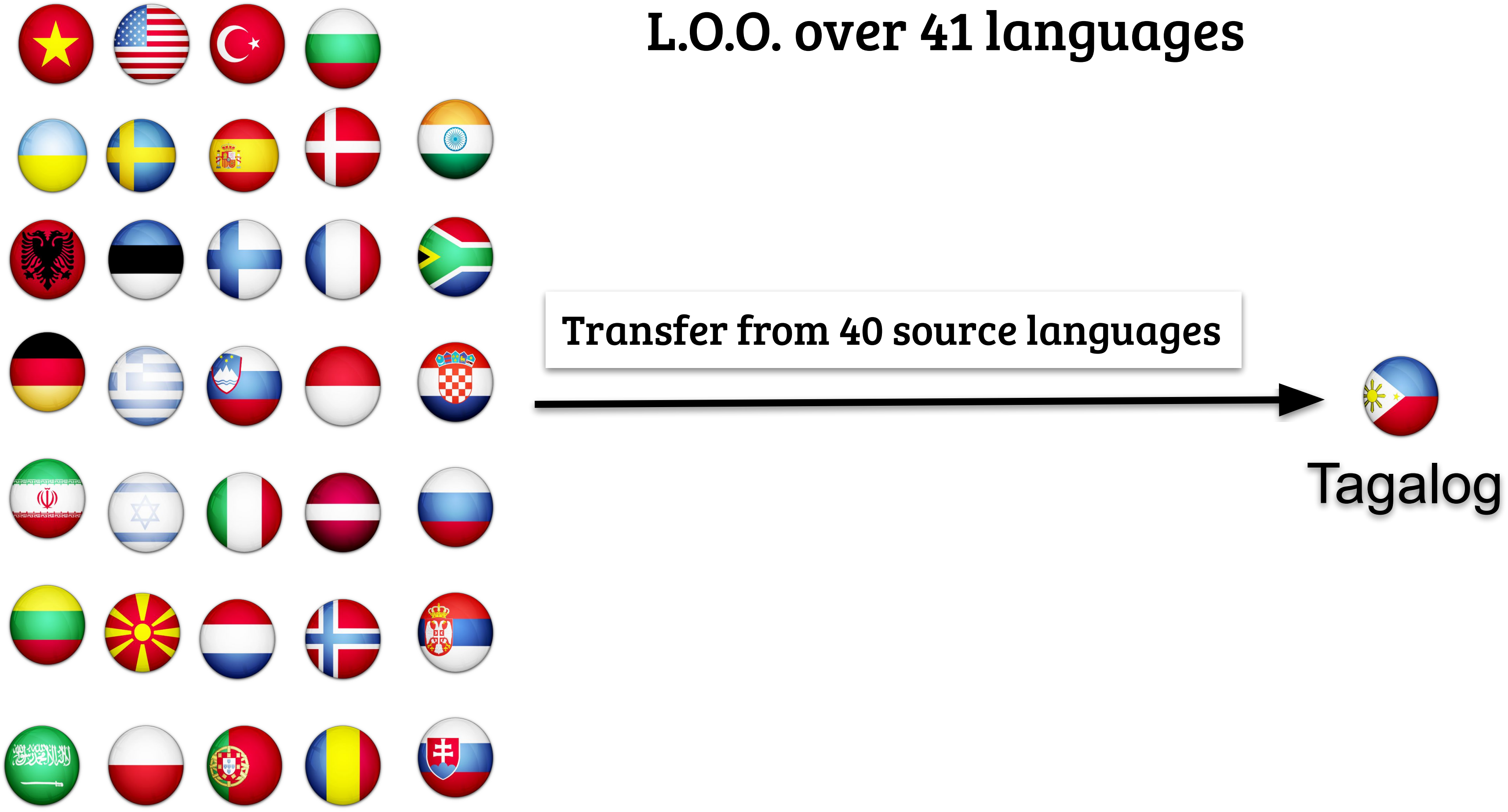
WIKIANN NER Datasets (Pan et al., 2017)

- Silver annotations from Wikipedia for **282** languages.
- We picked **41** languages based on availability of bilingual dictionaries.
- Created balanced training/dev/test partitions (varying size of training according to data availability)

github.com/afshinrahimi/mmner

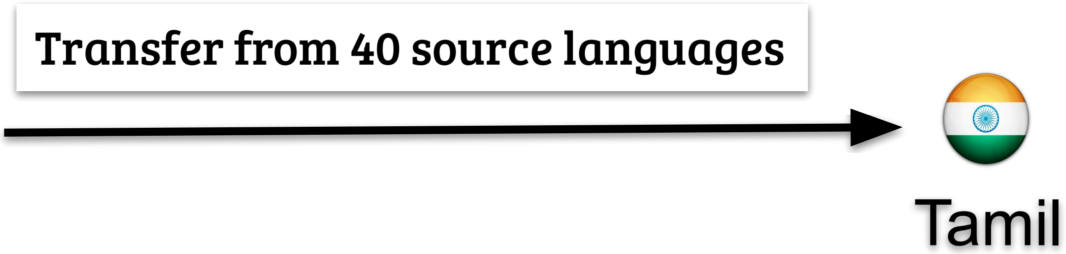
L.O.O. over 41 languages

L.O.O. over 41 languages



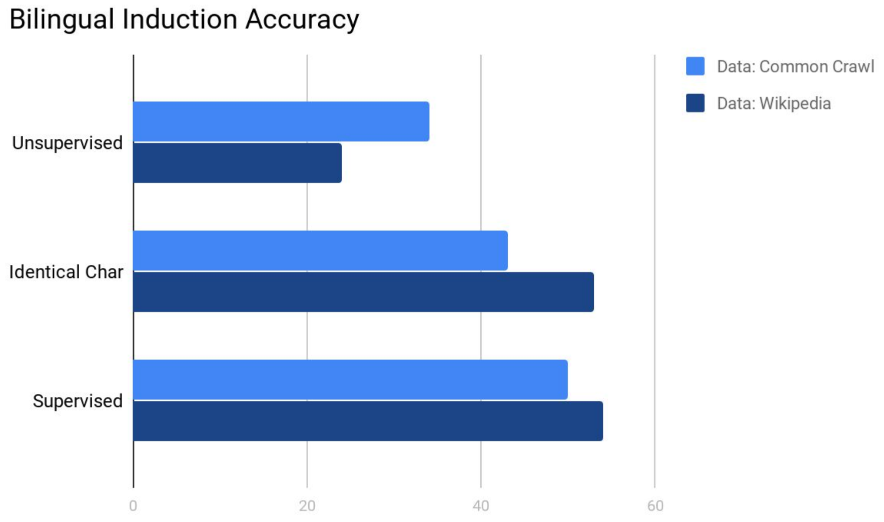
L.O.O. over 41 languages

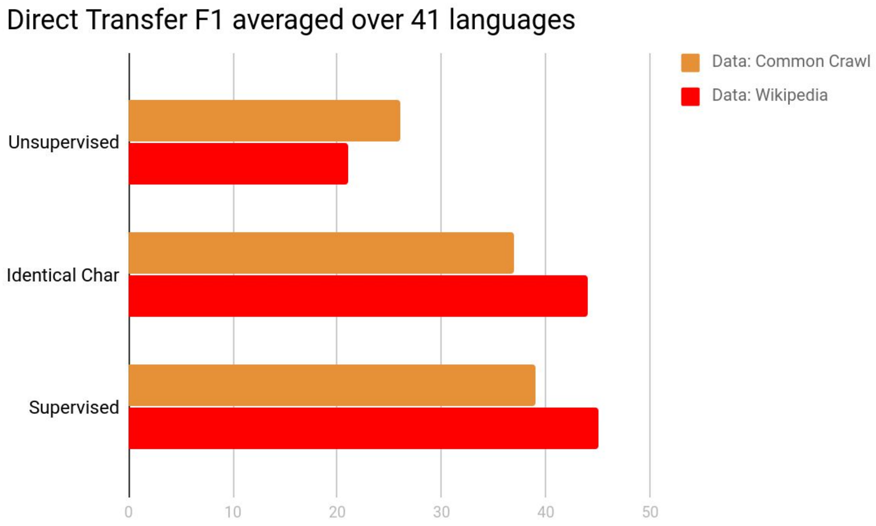
L.O.O. over 41 languages



Word representation: FastText/MUSE

Use **fasttext** monolingual **wiki** embeddings mapped to English space using **Identical Character Strings**.

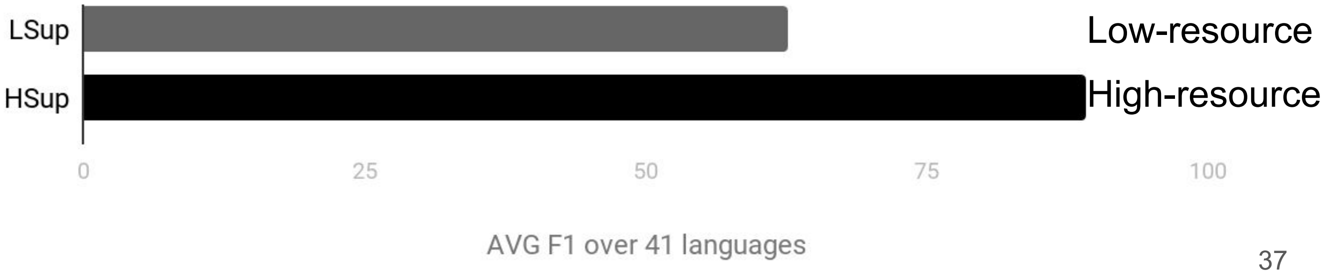




Conneau et al. (2017)

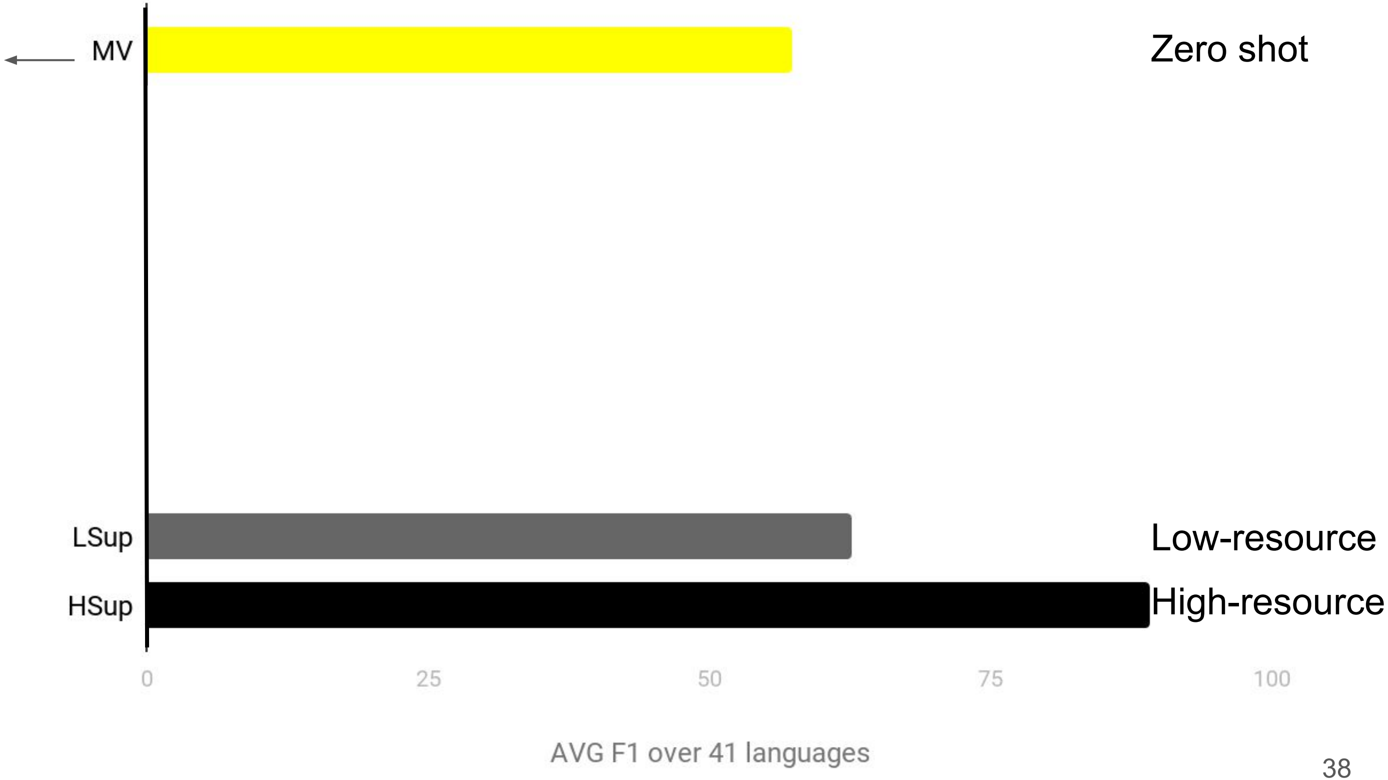
Results: WikiANN

Supervised: no transfer



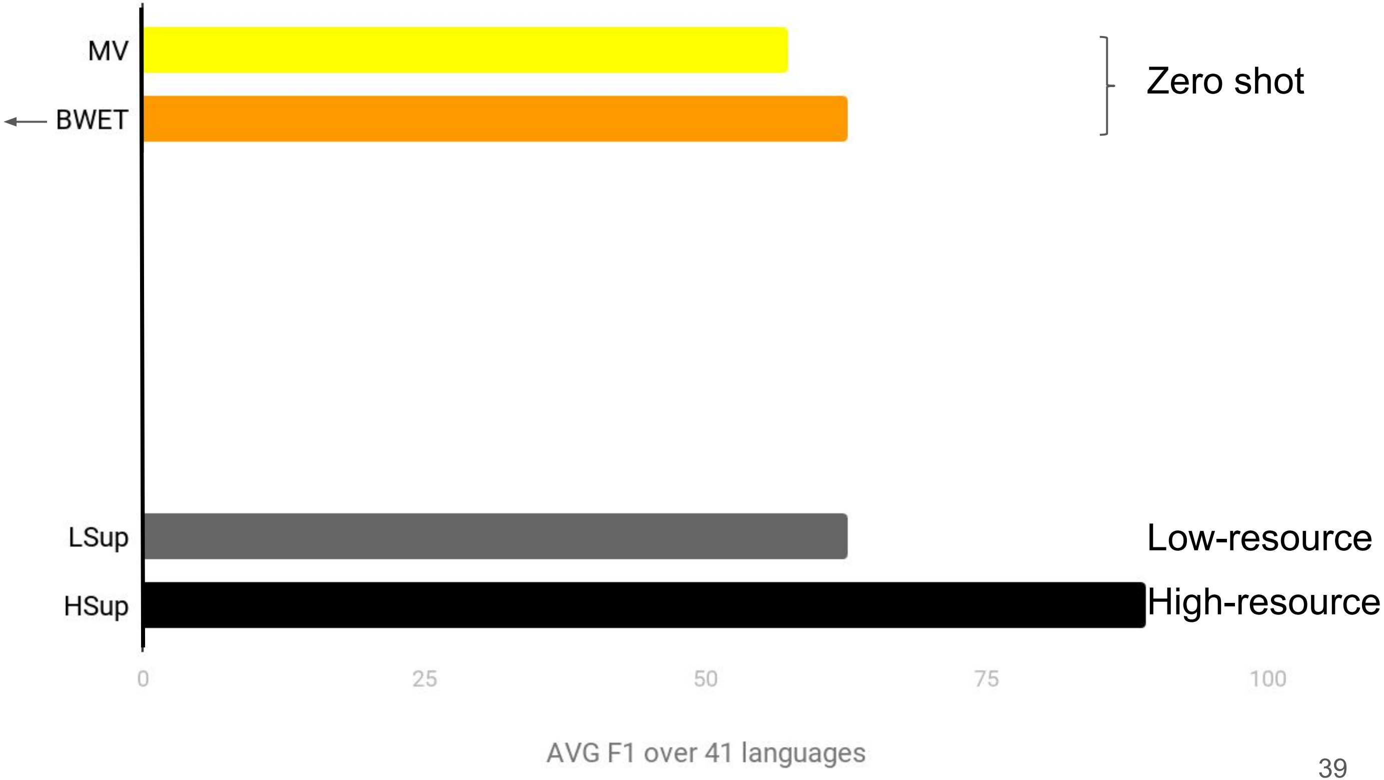
Results: WikiANN

Many low quality source models



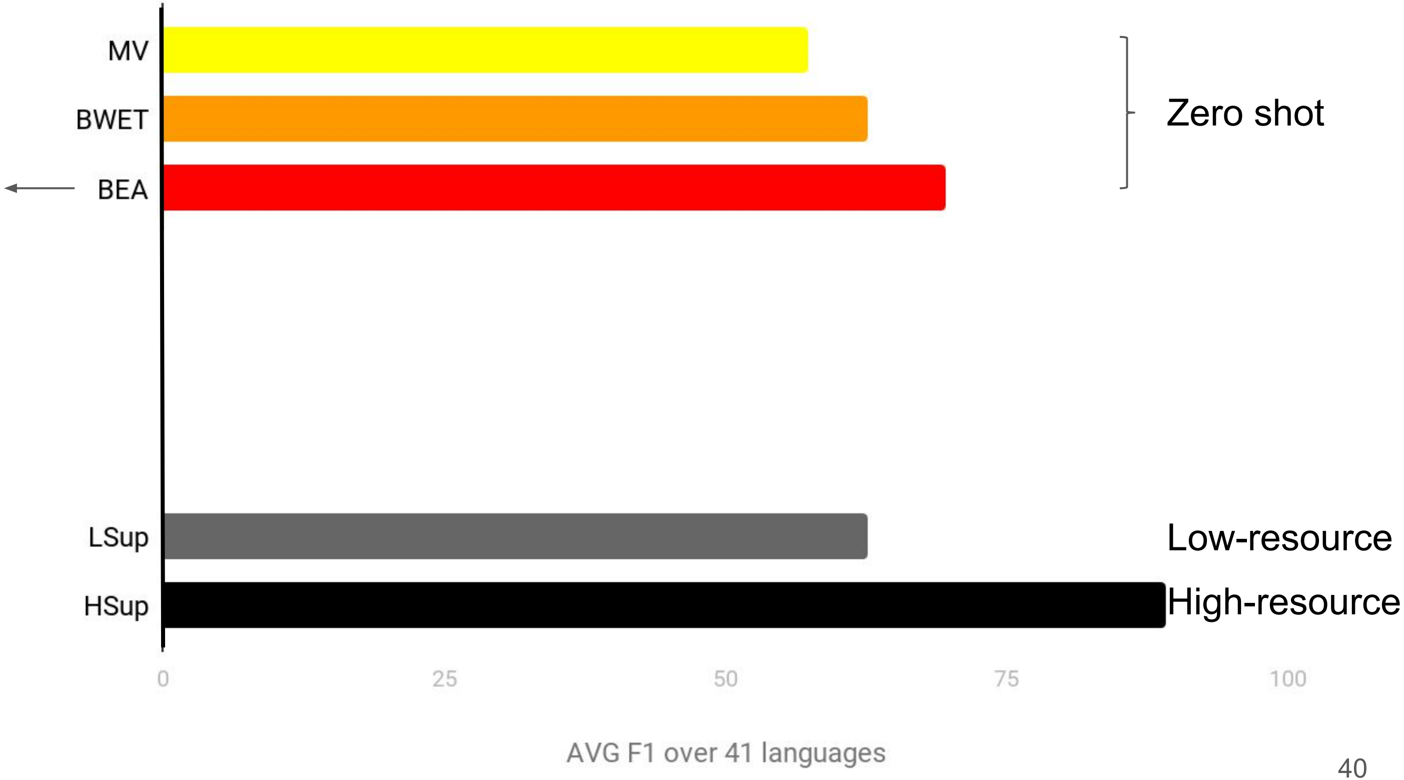
Results: WikiANN

Single source (en)



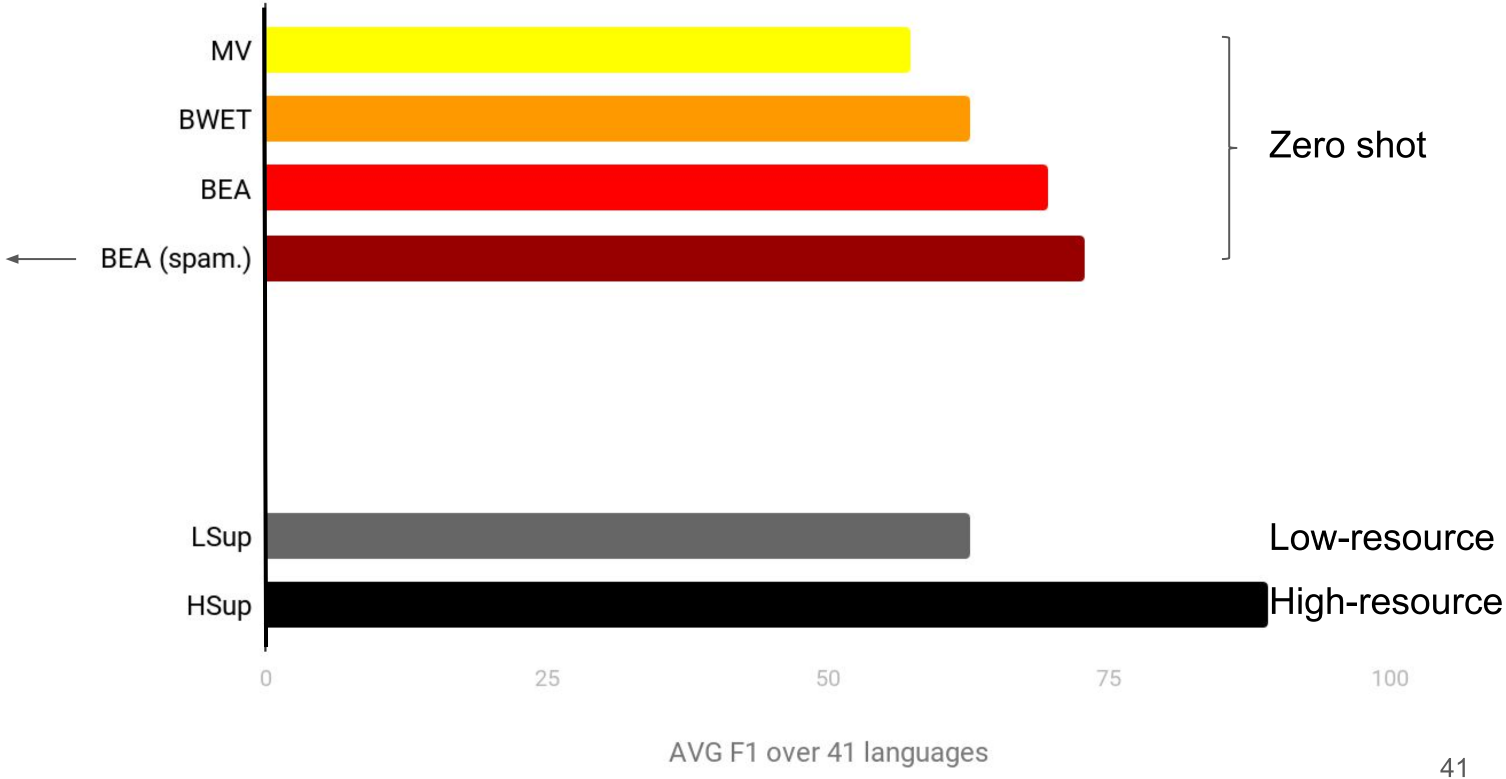
Results: WikiANN

Bayesian ensembling



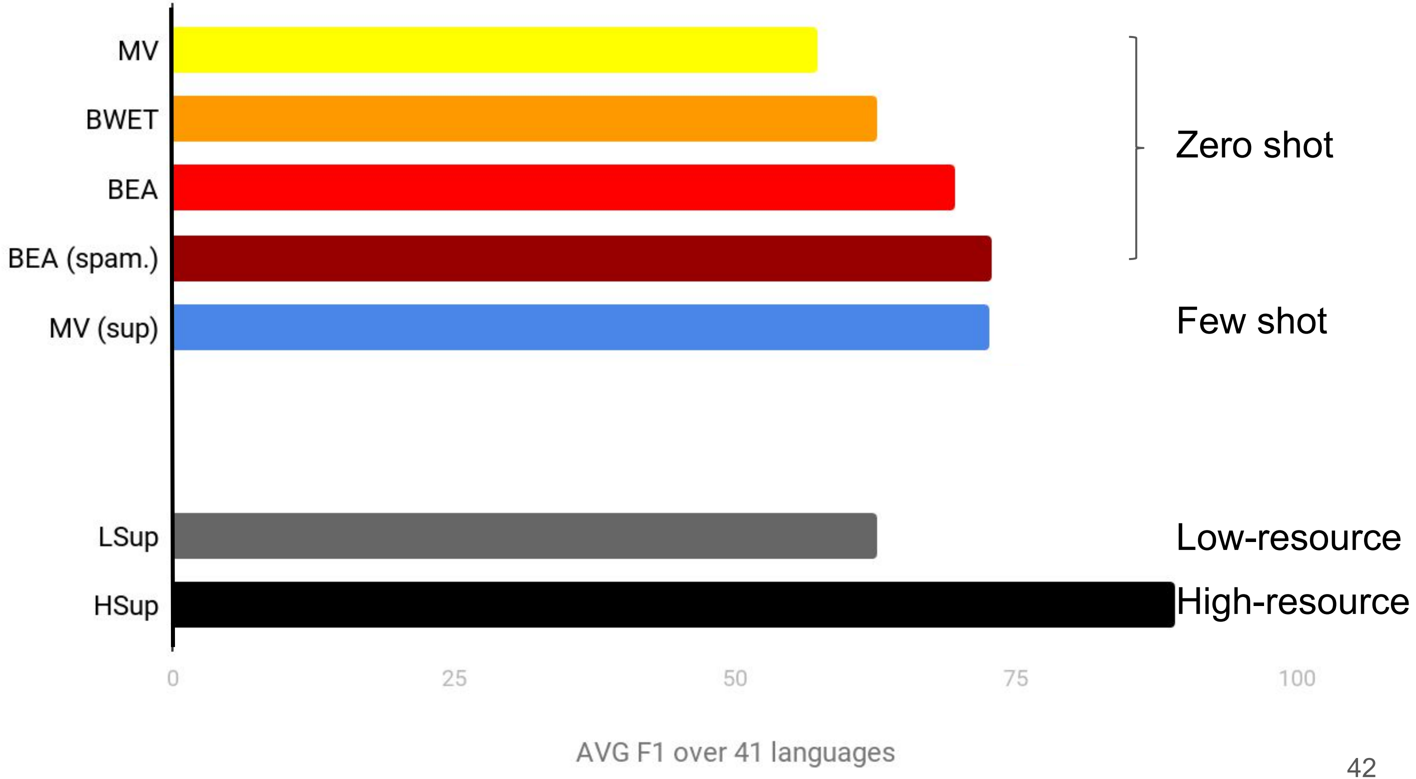
Results: WikiANN

+spammer removal



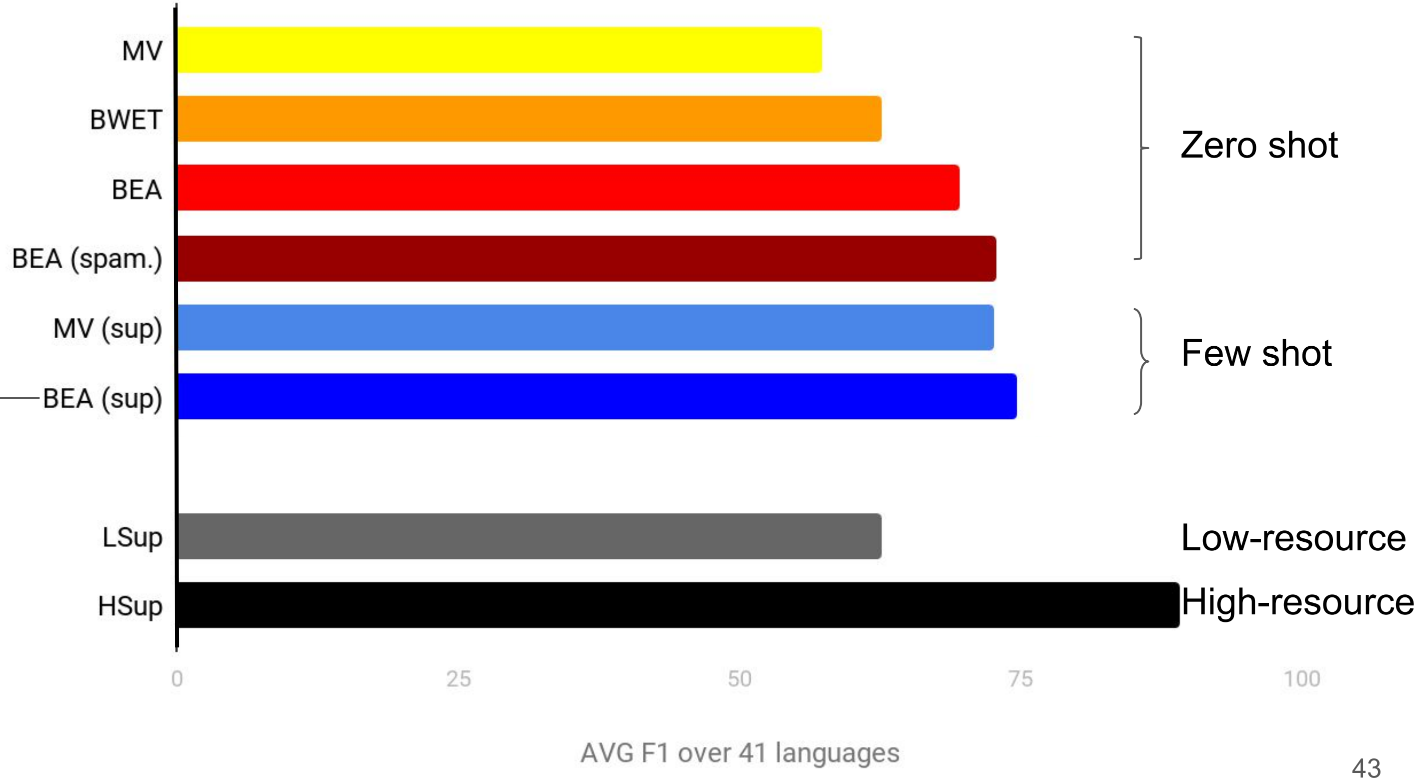
Results: WikiANN

MV between top 3 sources



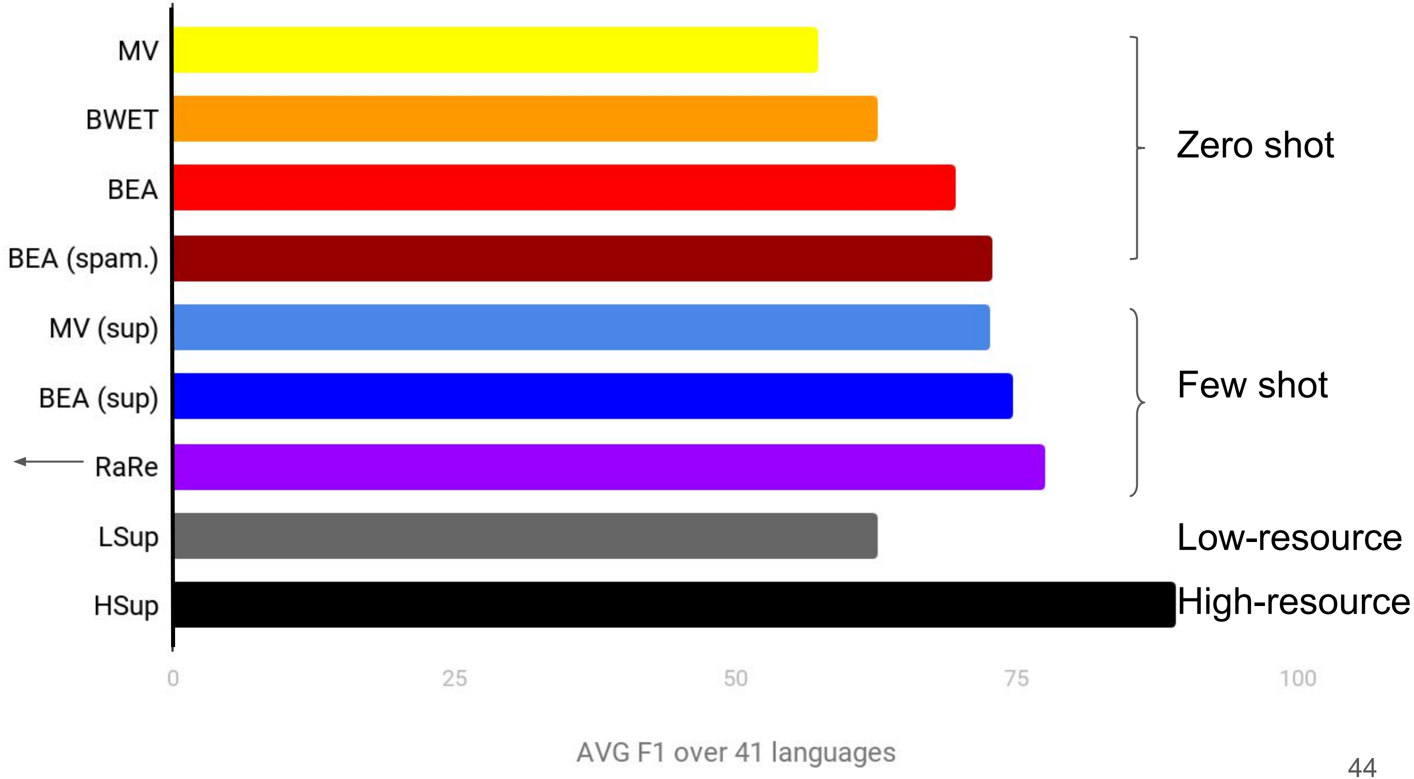
Results: WikiANN

Estimate BEA confusion & prior from annotations



Results: WikiANN

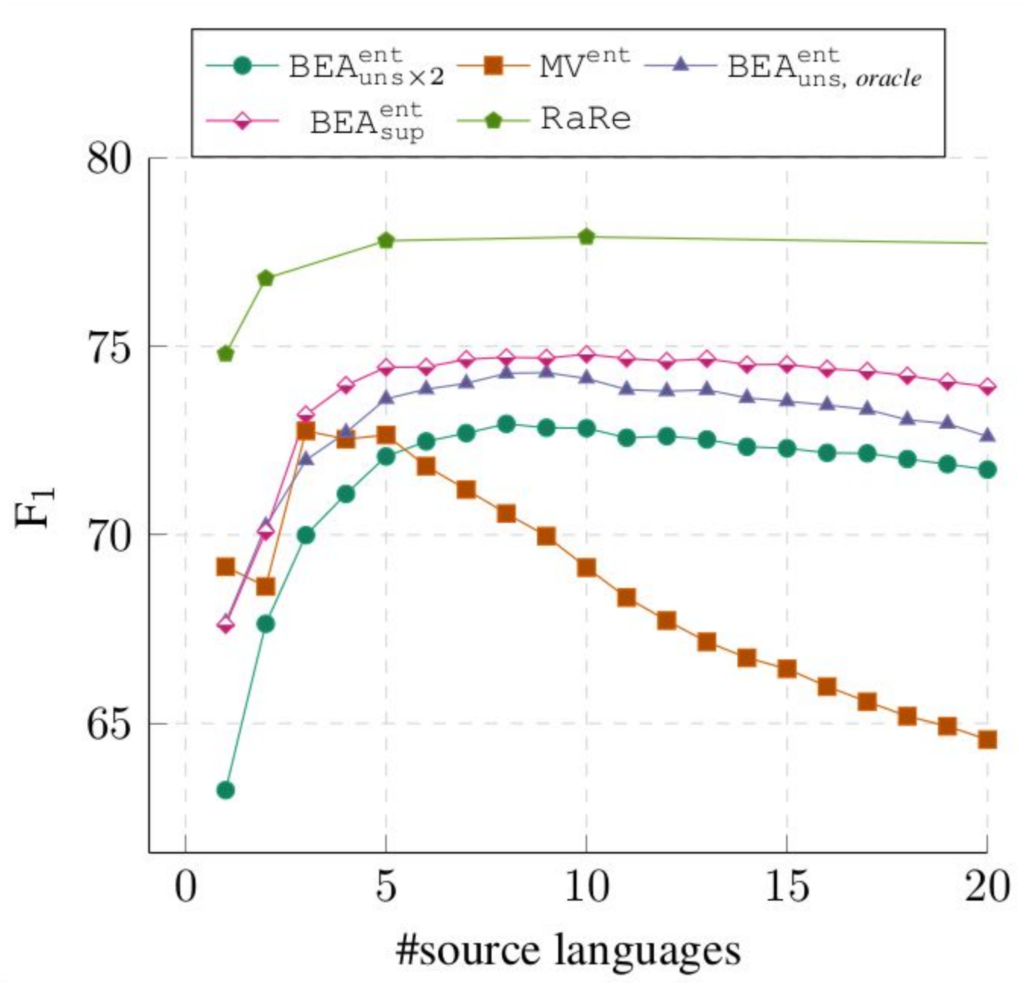
Ranking Retraining Method (using character info)



Effect of increasing #source languages

Methods robust to many varying quality source languages.

Even better with few-shot supervision.



Takeaways I

Transfer from multiple source languages helps because for many languages we don't know the best source language.

takeaway / noun [uk/aus/nz]: *a meal cooked and bought at a shop or restaurant but taken somewhere else...*
Cambridge English Dictionary

Takeaways II

With multiple source languages, you need to estimate their qualities because uniform voting doesn't perform well.

takeaway / noun [uk/aus/nz]: *a meal cooked and bought at a shop or restaurant but taken somewhere else...*
Cambridge English Dictionary

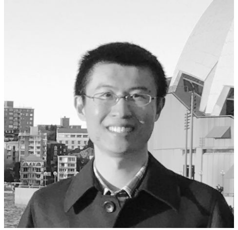
Takeaways III

A small training set in target language helps, and can be done cheaply and quickly (Garrette and Baldridge, 2013).

takeaway / noun [uk/aus/nz]: *a meal cooked and bought at a shop or restaurant but taken somewhere else...*
Cambridge English Dictionary

THE UNIVERSITY OF MELBOURNE

Thank you!



Datasets & code

github.com/afshinrahimi/mmner

Future Work

- Map all scripts to IPA or Roman alphabet
  (good for shared embeddings and character-level transfer)
    - uroman: Hermjakob et al. (2018)
    - epitran: Mortensen et al. (2018)
- Can we estimate the quality of source models/languages for a specific target language based on language characteristics (Littell et al., 2017)?
- Technique should apply beyond NER to other tasks.