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
kailangan namin ng mas maraming dugo sa Pagasanjan.

we need more blood in Pagasanjan.
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
Direct Transfer for NER

Output: Labelled sentences in the target language

Pre-trained NER source models

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)

F1 vs Target Language

unrelated
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)

100 gold sents. In Tagalog

Source Model AR → $F_{1_{AR}}$

Source Model EN → $F_{1_{EN}}$

Source Model VI → $F_{1_{VI}}$

Source model qualities
Model 1: Few Shot Ranking and Retraining (RaRe)

20k unlabelled sents in Tagalog

Source Model AR → Dataset AR
Source Model EN → Dataset EN
Source Model VI → Dataset VI

N training sets in Tagalog
Model 1: Few Shot Ranking and Retraining (RaRe)

\[ \text{Final training set, a mixture of distilled knowledge} \]
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})
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 $P(Z|Y, \alpha, \beta)$, 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 ($BEA_{unsx2}$)

2. **Few shot scenario:**
   Given 100 gold sentences, estimate source model confusion matrices, then run BEA ($BEA_{sup}$)

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)

- Täckström et al. (2012)
- Nothman et al. (2013)
- Tsai et al. (2016)
- Ni et al. (2017)
- Mayhew et al. (2017)

Use parallel data, dictionary or wikipedia

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

- Täckström et al. (2012)
- Nothman et al. (2013)
- Tsai et al. (2016)
- Ni et al. (2017)
- Mayhew et al. (2017)
- Xie et al. (2018)

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

<table>
<thead>
<tr>
<th>Method</th>
<th>AVG F1 over de, nl and es</th>
</tr>
</thead>
<tbody>
<tr>
<td>Täckström et al. (2012)</td>
<td>50</td>
</tr>
<tr>
<td>Nothman et al. (2013)</td>
<td>60</td>
</tr>
<tr>
<td>Tsai et al. (2016)</td>
<td>65</td>
</tr>
<tr>
<td>Ni et al. (2017)</td>
<td>70</td>
</tr>
<tr>
<td>Mayhew et al. (2017)</td>
<td>75</td>
</tr>
<tr>
<td>Xie et al. (2018)</td>
<td>80</td>
</tr>
<tr>
<td>MV</td>
<td>85</td>
</tr>
<tr>
<td>RaRe uns</td>
<td>90</td>
</tr>
<tr>
<td>BEA</td>
<td>95</td>
</tr>
</tbody>
</table>

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

- Täckström et al. (2012)
- Nothman et al. (2013)
- Tsai et al. (2016)
- Ni et al. (2017)
- Mayhew et al. (2017)
- Xie et al. (2018)
- MV
- RaRe uns
- BEA
- RaRe
- hsup

AVG F1 over de, nl and 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

Transfer from 40 source languages

Tagalog
L.O.O. over 41 languages
L.O.O. over 41 languages

Transfer from 40 source languages

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

**Word representation: FastText/MUSE**

Conneau et al. (2017)
Results: WikiANN

Supervised: no transfer

Low-resource

High-resource

AVG F1 over 41 languages
Results: WikiANN

Many low quality source models

AVG F1 over 41 languages

Zero shot
Low-resource
High-resource
Results: WikiANN

Single source (en)

- Low-resource
- High-resource
- Zero shot
Results: WikiANN

Bayesian ensembling

- MV
- BWET
- BEA

Zero shot

- LSup
- HSup

Low-resource
High-resource

AVG F1 over 41 languages
Results: WikiANN

+ spammer removal

Zero shot

Low-resource

High-resource

AVG F1 over 41 languages
Results: WikiANN

MV between top 3 sources

MV (sup)
BEA (spam.)
BEA
BWET
MV

Zero shot
Few shot
Low-resource
High-resource

AVG F1 over 41 languages
Results: WikiANN

Estimate BEA confusion & prior from annotations

Zero shot

Few shot

Low-resource

High-resource

AVG F1 over 41 languages
Results: WikiANN

Ranking Retraining Method (using character info)

- MV
- BWET
- BEA
- BEA (spam.)
- MV (sup)
- BEA (sup)
- RaRe
- LSup
- HSup

AVG F1 over 41 languages

Zero shot
Few shot
Low-resource
High-resource
Effect of increasing #source languages

Methods robust to many varying quality source languages.

Even better with few-shot supervision.
Transfer from multiple source languages helps because for many languages we don’t know the best source language.
With multiple source languages, you need to estimate their qualities because uniform voting doesn’t perform well.
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
Thank you!

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.