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

Abstract

In cross-lingual transfer, NLP models over one or more source languages are applied to a low-resource target language. While most prior work has used a single source model or a few carefully selected models, here we consider a "massive" setting with many such models. This setting raises the problem of poor transfer, particularly from distant languages. We propose two techniques for modulating the transfer, suitable for zero-shot or few-shot learning, respectively. Evaluating on named entity recognition, we show that our techniques are much more effective than strong baselines, including standard ensembling, and our unsupervised method rivals oracle selection of the single best individual model.[1]

1 Introduction

Supervised learning remains king in natural language processing, with most tasks requiring large quantities of annotated corpora. The majority of the world's 6,000+ languages however have limited or no annotated text, and therefore much of the progress in NLP has yet to be realised widely. Cross-lingual transfer learning is a technique which can compensate for the dearth of data, by transferring knowledge from high- to low-resource languages, which has typically taken the form of annotation projection over parallel corpora or other multilingual resources (Yarowsky et al., 2001; Hwa et al., 2005), or making use of transferable representations, such as phonetic transcriptions (Bharadwaj et al., 2016), closely related languages (Cotterell and Duh, 2017) or bilingual dictionaries (Mayhew et al., 2017; Xie et al., 2018).

Most methods proposed for cross-lingual transfer rely on a single source language, which limits the transferable knowledge to only one source. The target language might be similar to many source languages, on the grounds of the script, word order, loan words etc, and transfer would benefit from these diverse sources of information. There are a few exceptions, which use transfer from several languages, ranging from multitask learning (Duong et al., 2015; Ammar et al., 2016; Fang and Cohn, 2017), and annotation projection from several languages (Täckström, 2012; Fang and Cohn, 2016; Plank and Agić, 2018). However, to the best of our knowledge, none of these approaches adequately account for the quality of transfer, but rather "weight" the contribution of each language uniformly.

In this paper, we propose a novel method for zero-shot multilingual transfer, inspired by research in truth inference in crowd-sourcing, a related problem, in which the 'ground truth' must be inferred from the outputs of several unreliable annotators (Dawid and Skene, 1979). In this problem, the best approaches estimate each model's reliability, and their patterns of mistakes (Kim and Ghahramani, 2012). Our proposed model adapts these ideas to a multilingual transfer setting, whereby we learn the quality of transfer, and language-specific transfer errors, in order to infer the best labelling in the target language, as part of a Bayesian graphical model. The key insight is that while the majority of poor models make lots of mistakes, these mistakes are diverse, while the few good models consistently provide reliable input. This allows the model to infer which are the reliable models in an unsupervised manner, i.e., without explicit supervision in the target language, and thereby make accurate inferences despite the substantial noise.

In the paper, we also consider a supervised setting, where a tiny annotated corpus is available in the target language. We present two methods to use this data: 1) estimate reliability parameters of

*Both authors have equally contributed to this work.

[1]The code and the datasets will be made available at https://github.com/afshinrahimi/mmner.

the Bayesian model, and 2) explicit model selection and fine-tuning of a low-resource supervised model, thus allowing for more accurate modelling of language specific parameters, such as character embeddings, shown to be important in previous work (Xie et al., 2018).

Experimenting on two NER corpora, with as many as 41 languages, we show that single model transfer has highly variable performance, and uniform ensembling often substantially underperforms the single best model. In contrast, our zero-shot approach does much better, exceeding the performance of the single best model, and our few-shot supervised models produce further gains.

2 Approach

We frame the problem of multilingual transfer as follows. We assume a collection of H models, all trained in a high resource setting, denoted $M^h = \{M_i^h, i \in (1, H)\}$. Each of these models are not well matched to our target data setting, for instance these may be trained on data from different domains, or on different languages, as we evaluate in our experiments, where we use cross-lingual embeddings for model transfer. This is a problem of transfer learning, namely, how best we can use the H models for best results in the target language.[2]

Simple approaches in this setting include a) choosing a single model $M \in M^h$, on the grounds of practicality, or the similarity between the model's native data condition and the target, and this model is used to label the target data; or b) allowing all models to 'vote' in an classifier ensemble, such that the most frequent outcome is selected as the ensemble output. Unfortunately neither of these approaches are very accurate in a cross-lingual transfer setting, as we show in §4, where we show a fixed source language model (`en`) dramatically underperforms compared to oracle selection of source language, and the same is true for uniform voting.

Motivated by these findings, we propose novel methods for learning. For the "zero-shot" setting where no labelled data is available in the target, we propose the $\texttt{BEA}_{\texttt{uns}}$ method inspired by work in truth inference from crowd-sourced datasets (§2.1). To handle the "few-shot" case §2.2 presents a rival supervised technique, `RaRe`, based on using very limited annotations in the target language for model selection and classifier fine-tuning.

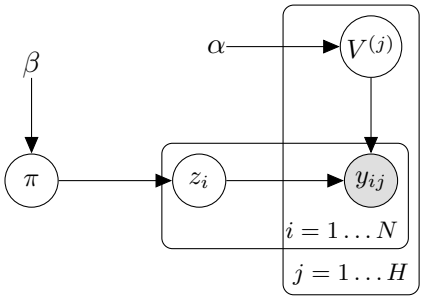

Figure 1: Plate diagram for the aggregation model.

2.1 Zero-Shot Transfer as Truth Inference

One way to improve the performance of the ensemble system is to select a subset of component models carefully, or more generally, learn a non-uniform weighting function. Some models do much better than others, on their own, so it stands to reason that identifying these handful of models will give rise to better ensemble performance. How might we proceed to learn the relative quality of models in the setting where no annotations are available in the target language? This is a classic unsupervised inference problem, for which we propose a probabilistic graphical model, inspired by Kim and Ghahramani (2012).

We develop a generative model, illustrated in Figure 1, of the transfer models' predictions, y_{ij}, where $i \in [1, N]$ is an instance (a token or an entity span), and $j \in [1, H]$ indexes a transfer model.[3] The generative process assumes a 'true' label, $z_i \in [1, K]$, which is corrupted by each worker, in producing the observation, y_{ij}. The corruption process is described by $P(y_{ij} = l | z_i = k, V^{(j)}) = V_{kl}^{(j)}$, where $V^{(j)} \in \mathcal{R}^{K \times K}$ is a worker-specific confusion matrix.

To complete the story, the confusion matrices are drawn from vague row-wise independent Dirichlet priors, with a parameter $\alpha = 1$, and the true labels are governed by a Dirichlet prior, π, which is drawn from an uninformative Dirichlet distribution with a parameter $\beta = 1$.

[2] We limit our attention to transfer in a 'black-box' setting, that is, given predictive models, but not assuming access to their data, nor their implementation. This is the most flexible scenario, as it allows for application to settings with closed APIs, and private datasets. It does, however, preclude multi-task learning, as the source models are assumed to be static.

[3] For simplicity, hereinafter we refer to y_{ij} as 'annotations' and the transfer models as 'workers'. This reflects the inspiration of our approach from models of truth inference from crowd-sourcing (Dawid and Skene, 1979).

Inference under this model involves explaining the observed Y in the most efficient way. Where several workers have identical outputs, k, on an instance, this can be explained by letting $z_i = k$,[4] and the workers' confusion matrices assigning high probability to $V_{kk}^{(j)}$. Other, less reliable, workers will have divergent labels, which are less likely to be in agreement, or else are heavily biased towards a particular class. Accordingly, the model can better explain these outputs through label confusion, using the off-diagonal elements of the confusion matrix. Aggregated over a corpus of instances, the model can learn to differentiate between those reliable workers, with high $V_{kk}^{(j)}$ and those less reliable ones, with high $V_{kl}^{(j)}$, $l \neq k$. This procedure applies per-label, and thus worker 'reliability' is with respect to a specific label, and may differ between classes. This helps in the NER setting where many poor classifiers have excellent accuracy for the outside label, but considerably worse performance for entity labels.

For inference, we use mean-field variational Bayes (Jordan, 1998), which learns a variational distribution, $q(Z, V, \pi)$ to optimise the evidence lower bound (ELBO),

$$\log P(Y|\alpha, \beta) \geq \mathbb{E}_{q(Z,V,\pi)} \log \frac{P(Y, Z, V, \pi|\alpha, \beta)}{q(Z, V, \pi)}$$

assuming a fully factorised variational distribution, $q(Z, V, \pi) = q(Z)q(V)q(\pi)$. This gives rise to an iterative learning algorithm with update rules:

$$\mathbb{E}_q \log \pi_k \tag{1a}$$
$$= \psi\left(\beta + \sum_i q(z_i = k)\right) - \psi\left(K\beta + N\right)$$
$$\mathbb{E}_q \log V_{kl}^{(j)} \tag{1b}$$
$$= \psi\left(\alpha + \sum_i q(z_i = k)\mathbf{1}[y_{ij} = l]\right) - \psi\left(K\alpha + \sum_i q(z_i = k)\right)$$

$$q(z_i = k) \propto \exp\left\{\mathbb{E}_q \log \pi_k + \sum_j \mathbb{E}_q \log V_{ky_{ij}}^{(j)}\right\} \tag{2}$$

[4] Although there is no explicit breaking of the symmetry of the model, we initialise inference using the majority vote, which results in a bias towards this solution.

	w_1	w_2	w_3	w_4	$[1,4]$	$[2,4]$	$[3,4]$
M_1^h	B-ORG	I-ORG	I-ORG	I-ORG	ORG	O	O
M_2^h	O	B-ORG	I-ORG	I-ORG	O	ORG	O
M_3^h	O	O	B-ORG	I-ORG	O	O	ORG
M_4^h	O	B-PER	I-PER	I-PER	O	PER	O
M_5^h	O	B-PER	I-PER	I-PER	O	PER	O
Agg.	O	B-PER	I-ORG	I-ORG	O	PER	O

Table 1: An example sentence with its aggregated labels in both token view and entity view. Aggregation in token view may generate results inconsistent with the BIO scheme.

where ψ is the digamma function, defined as the logarithmic derivative of the gamma function. The sets of rules (1) and (2) are applied alternately, to update the values of $\mathbb{E}_q \log \pi_k$, $\mathbb{E}_q \log V_{kl}^{(j)}$, and $q(z_{ij} = k)$ respectively. This repeats until convergence, when the difference in the ELBO between two iterations is smaller than a threshold.

The final prediction of the model is based on $q(Z)$, using the maximum a posteriori label $\hat{z}_i = \arg\max_z q(z_i = z)$. This method is referred to as $\texttt{BEA}_{\texttt{uns}}$.

In our NER transfer task, classifiers are diverse in their F1 scores ranging from almost 0 to around 80, motivating *spammer removal* (Raykar and Yu, 2012) to filter out the worst of the input models. We adopt a simple strategy that first estimates the confusion matrices for all classifiers on all labels, then ranks classifiers based on their mean recall on different entity categories (elements on the diagonals of their confusion matrices), and then runs the model again using only labels from the top k classifiers only. We call this method $\texttt{BEA}_{\texttt{uns}\times 2}$ and its results are reported in §4.

2.1.1 Token versus Entity Granularity

Our proposed aggregation method in §2.1 is based on an assumption that the true annotations are independent from each other, which simplifies the model but may generate undesired results. That is, entities predicted by different models could be mixed, resulting in labels inconsistent with the BIO scheme. Table 1 shows an example, where a sentence with 4 is annotated by 5 models with 4 different predictions, among which at most one is correct as they overlap. However, the aggregated result in the token view is a mixture of two predictions, which is supported by no classifiers.

To deal with this problem, we consider aggregating the predictions in the entity view. As shown

in Table 1, we convert the predictions for tokens to predictions for ranges, aggregate labels for every range, and then resolve remaining conflicts. A prediction is ignored if it conflicts with another one with higher probability. By using this greedy strategy, we can solve the conflicts raised in entity-level aggregation. We use superscripts `tok` and `ent` to denote token-level and entity-level aggregations, i.e. $\texttt{BEA}^{\texttt{tok}}_{\texttt{uns}}$ and $\texttt{BEA}^{\texttt{ent}}_{\texttt{uns}}$.

2.2 Few-Shot Transfer

Until now, we have assumed no access to annotations in the target language. However, when some labelled text is available, how might this best be used? In our experimental setting, we assume a modest set of 100 labelled sentences, in keeping with a low-resource setting (Garrette and Baldridge, 2013).[5] We propose two models $\texttt{BEA}_{\texttt{sup}}$ and `RaRe` in this setting.

Supervising `BEA` ($\texttt{BEA}_{\texttt{sup}}$) One possibility is to use the labelled data to find the posterior for the parameters $V^{(j)}$ and π of the Bayesian model described in §2.1. Let n_k be the number of instances in the labelled data whose true label is k, and n_{jkl} the number of instances whose true label is k and classifier j labels them as l. Then the quantities in Equation (1) can be calculated as

$$\mathbb{E}\log \pi_k = \psi(n_k) - \psi(N)$$
$$\mathbb{E}\log v_{jkl} = \psi(n_{jkl}) - \psi\left(\sum_l n_{jkl}\right).$$

These are used in Equation (2) for inference on the test set. We refer to this setting as $\texttt{BEA}_{\texttt{sup}}$.

Ranking and Retraining (`RaRe`) We also propose an alternative way of exploiting the limited annotations, `RaRe`, which first **ra**nks the systems, and then uses the top ranked models' outputs alongside the gold data to **re**train a model on the target language. The motivation is that the above technique is agnostic to the input text, and therefore is unable to exploit situations where regularities occur, such as common words or character patterns that are indicative of specific class labels, including names, titles, etc. These signals are unlikely to be consistently captured by cross-lingual transfer. Training a model on the target language with a character encoder component, can distil the signal that are captured by the transfer models, while relating this towards generalisable lexical and structural evidence in the target language. This on its own will not be enough, as many tokens will be consistently misclassified by most or all of the transfer models, and for this reason we also perform model fine-tuning using the supervised data.

The ranking step in `RaRe` proceeds by evaluating each of the H transfer models on the target gold set, to produce scores s_h (using the F_1 score). The scores are then truncated to the top $k \leq H$ values, such that $s_h = 0$ for those systems h not ranked in the top k, and normalised $\omega_h = \frac{s_h}{\sum_{j=1}^{k} s_j}$. The range of scores are quite wide, covering $0.00 - 0.81$ (see Figure 2), and accordingly this simple normalisation conveys a strong bias towards the top scoring transfer systems.

The next step is a distillation step, where a model is trained on a large unannotated dataset in the target language, such that the model predictions match those of a weighted mixture of transfer models, using $\vec{\omega} = (\omega_1, \ldots, \omega_H)$ as the mixing weights. This process is implemented as mini-batch scheduling, where the labels for each mini-batch are randomly sampled from transfer model h with probability ω_h.[6] This is repeated over the course of several epochs of training.

Finally, the model is fine-tuned using the small supervised dataset, in order to correct for phenomena that are not captured from model transfer, particularly character level information which is not likely to transfer well for all but the most closely related languages. Fine-tuning proceeds for a fixed number of epochs on the supervised dataset, to limit overtraining of richly parameterise models on a tiny dataset. Note that in all stages, the same supervised dataset is used, both in ranking and fine-tuning, and moreover, we do not use a development set. This is not ideal, and generalisation performance would likely improve were we to use additional annotated data, however our meagre use of data is designed for a low resource setting where labelled data is at a premium.

3 Experiments

3.1 Data

Our primarily evaluation is over a subset of the `Wikiann` NER corpus, using 41 out of 282 lan-

[5] Garrette and Baldridge (2013) showed that about 100 sentences can be annotated with POS tags in two hours by non-native annotators.

[6] We show that uniform sampling with few source languages achieves worse performance.

guages, where the langauges were chosen based on their overlap with multilingual word embedding resources from Lample et al. (2018).[7] The NER taggs are in IOB2 format comprising of LOC, PER, and ORG. The distribution of labels is highly skewed, so we created balanced datasets, and partitioned into training, development, and test sets, details of which are in the Appendix. For comparison with prior work, we also evaluate on the CoNLL 2002 and 2003 datasets (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), which we discuss further in §4.

For language-independent word embedding features we use fastText 300 dimensional Wikipedia embeddings (Bojanowski et al., 2017), and map them to the English embedding space using character-identical words as the seed for the Procrustes rotation method for learning bingual embedding spaces from MUSE (Lample et al., 2018).[8] Similar to Xie et al. (2018) we don't rely on a bilingual dictionary, so the method can be easily applied to other languages.

3.2 Model Variations

As the sequential tagger, we use a BiLSTM-CRF (Lample et al., 2016), which has been shown to result in state-of-the-art results in high resource settings. This model includes both word embeddings (for which we used fixed cross-lingual embeddings) and character embeddings, to form a parameterised potential function in a linear chain conditional random field. With the exception of batch size and learning rate which were tuned (details in Appendix), we kept the architecture and the hyperparameters the same as the published code.[9]

We trained models on all 41 languages in both high-resource (HSup) and naive supervised low-resource (LSup) settings, where HSup pre-trained models were used for transfer in a leave-one-out setting, i.e., taking the predictions of 40 models into a single target language. The same BiLSTM-CRF is also used for RaRe.

To avoid overfitting, we use early stopping based on a validation set for the HSup, and LSup baselines. For RaRe, given that the model is already trained on noisy data, we stop fine-tuning after only 5 iterations, chosen based on the performance for the first four languages.

We compare the supervised **HSup** and **LSup** monolingual baselines with our proposed transfer models:

MV uniform ensemble, a.k.a."majority vote";

BEA$_{\text{uns}\times 2}$, **BEA$_{\text{uns}}$** unsupervised aggregation models, applied to entities or tokens (see §2.1);

BEA$_{\text{sup}}$ supervised estimation of BEA prior (§2.2);

RaRe, **RaRe$_{\text{uns}}$** supervised ranking and retraining model (§2.2), and uniform ranking without fine-tuning, respectively; and

Oracle selecting the best performing single transfer model, based on test performance.

We also compare with BWET (Xie et al., 2018) as state-of-the-art in unsupervised NER transfer. BWET transfers the source English training and development data to the target language using bilingual dictionary induction (Lample et al., 2018), and then uses a transformer architecture to compensate for missing sequential information. We used BWET in both CoNLL, and Wikiann datasets by transferring from their corresponding source English data to the target language.[10]

4 Results

We report the results for single source direct transfer, and then show that our proposed multilingual methods outperform majority voting. Then we analyse the choice of source languages, and how it affects transfer. Finally we report results on CoNLL NER datasets.

Direct Transfer The first research question we consider is the utility of direct transfer, and the simple majority vote ensembling method. As shown in Figure 2, using a single model for direct transfer (English: en) is often a terrible choice. The oracle choice of source language model does much better, however it is not always a closely related language (e.g., Italian: it does best for In-

[7] With ISO 639-1 codes: af, ar, bg, bn, bs, ca, cs, da, de, el, en, es, et, fa, fi, fr, he, hi, hr, hu, id, it, lt, lv, mk, ms, nl, no, pl, pt, ro, ru, sk, sl, sq, sv, ta, tl, tr, uk and vi.

[8] We also experimented with other bilingual embedding methods, including: supervised learning over bilingual dictionaries, which barely affected system performance; and pure-unsupervised methods (Lample et al., 2018; Artetxe et al., 2018), which performed substantially worse. For this reason we use identical word type seeding, which is preferred as it imposes no additional supervision requirement.

[9] https://github.com/guillaumegenthial/sequence_tagging

[10] Because BWET uses identical characters for bilingual dictionary induction, we observed many English loan words in the target language mapped to the same word in the induced bilingual dictionaries. Filtering such dictionary items might improve BWET.

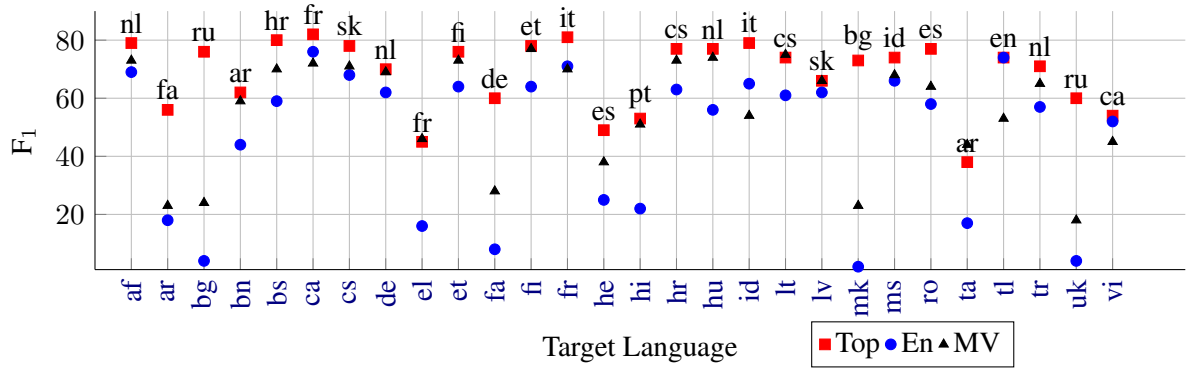

Figure 2: Best source language (■) compared with en (●), and majority voting (▲) over all source languages in terms of F_1 performance in direct transfer shown for a subset of the 41 target languages (x axis). Worst transfer score, not shown here, is about 0. See §3 for details of models and datasets.

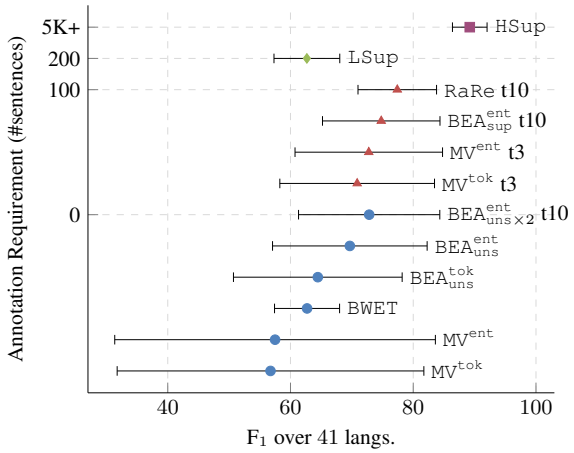

Figure 3: The mean and standard deviation for the F_1 score of the proposed unsupervised models ($\texttt{BEA}^{tok}_{uns}$ and $\texttt{BEA}^{ent}_{uns}$), supervised models (RaRe and $\texttt{BEA}^{ent}_{sup}$ t10) compared with state-of-the-art unsupervised model BWET (Xie et al., 2018), high- and low-resource supervised models HSup and LSup, and majority voting ($\texttt{MV}^{tok}$) in terms of entity level F_1 over the 41 languages (40 for BWET) summarised from Table 4. The x axis shows the annotation requirement of each model in the target language where "200" means 100 sentences each for training and development, and "5K+" means using all the available annotation for training and development sets. Points with the same colour/shape have equal data requirement.

donesian: id, despite the target being closer to Malay: ms). Note the collection of Cyrillic languages (bg, mk, uk) where the oracle is substantially better than the majority vote, which is likely due to script differences. The transfer relationship is not symmetric e.g., Persian: fa does best for Arabic: ar, but German: de does best for Persian. Figure 2 also shows that ensemble voting is well below the oracle best language, which is likely to be a result of overall high error rates coupled with error correlation between models, and little can be gained from ensembling.

Multilingual Transfer We report the results for the proposed low-resource supervised models (RaRe and $\texttt{BEA}_{sup}$), and unsupervised models ($\texttt{BEA}_{uns}$ and $\texttt{BEA}_{uns\times2}$), summarised as an average over the 41 languages in Figure 3 (see Appendix A for the full table of results). The figure compares against high- and low-resource supervised baselines (HSup and LSup, respectively), and BWET. The best performance is achieved with a high supervision (HSup, $F_1 = 89.2$), while very limited supervision (LSup) results in a considerably lower F_1 of 62.1. The results for $\texttt{MV}^{tok}$ show that uniform ensembling of multiple source models is even worse, by about 5 points.

Unsupervised truth inference dramatically improves upon $\texttt{MV}^{tok}$, and $\texttt{BEA}^{ent}_{uns}$ outperforms $\texttt{BEA}^{tok}_{uns}$, showing the effectiveness of inference over entities rather than tokens. It is clear that having access to limited annotation in the target language makes a substantial difference in $\texttt{BEA}^{ent}_{sup}$ and RaRe with F_1 of 74.8 and 77.4, respectively.

Further analysis show that majority voting works reasonably well for Roman and Germanic languages, which are well represented in the dataset, but fails miserably compared to single best for Slavic languages (e.g. ru, uk, bg) where there are only a few related languages. For most of the isolated languages (ar, fa, he, vi, ta), explicitly training a model in RaRe outperforms $\texttt{BEA}^{ent}_{sup}$, showing that relying only on aggregation of annotated data has limitations, in that it cannot exploit character and structural features.

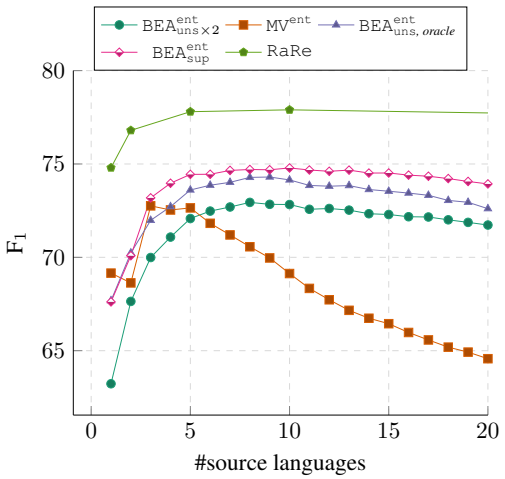

Figure 4: The mean F_1 performance of MV^{ent}, BEA^{ent}_{sup}, $BEA^{ent}_{uns\times2}$, $BEA^{ent}_{uns,\ oracle}$, and RaRe over the 41 languages by the number of source languages.

Choice of Source Languages An important question is how the other models, particularly the unsupervised variants, are affected by the number and choice of sources languages. Figure 4 charts the performance of MV, BEA, and RaRe against the number of source models, comparing the use of ideal or realistic selection methods to attempt to find the best source models. MV^{ent}, BEA^{ent}_{sup}, and RaRe use a small labeled dataset to rank the source models. $BEA^{ent}_{uns,\ oracle}$ has the access to the perfect ranking of source models based on their real F_1 on the test set. $BEA_{uns\times2}$ is completely unsupervised in that it uses its own estimates to rank all source models.

MV doesn't show any benefit with more than 3 source models.[11] In contrast, BEA and RaRe continue to improve with up to 10 languages. We show that BEA in two realistic scenarios (unsupervised: $BEA^{ent}_{uns\times2}$, and supervised: BEA^{ent}_{sup}) is highly effective at discriminating between good and bad source models, and thus filtering out the bad models gives the best results. The $BEA^{ent}_{uns\times2}$ curve shows the effect of filtering using purely unsupervised signal, which has a positive, albeit mild effect on performance. In $BEA^{ent}_{uns,\ oracle}$ although the source model ranking is perfect, it narrowly outperforms BEA. Note also that neither of the BEA curves show evidence of the sawtooth pattern, i.e., they largely benefit from more inputs, irrespective of their parity. Finally, adding supervision in the target language in RaRe further improves upon the unsupervised models.

[11]The sawtooth pattern arises from the increased numbers of ties (broken randomly) with even numbers of inputs.

lang.	de	es	nl	en
Täckström et al. (2012)^{p}	40.4	59.3	58.4	—
Nothman et al. (2013)^{w}	55.8	61.0	64.0	61.3
Tsai et al. (2016)^{w}	48.1	60.6	61.6	—
Ni et al. (2017)^{w,p,d}	58.5	65.1	65.4	—
Mayhew et al. (2017)^{w,d}	59.1	66.0	66.5	—
Xie et al. (2018)^{0}	57.8	72.4	70.4	—
our work				
$MV^{tok,\ 0}$	57.4	66.4	71.0	62.1
$MV^{ent,\ 0}$	57.7	69.0	70.3	64.6
$BEA^{tok,\ 0}_{uns}$	58.2	64.7	70.1	61.2
$BEA^{ent,\ 0}_{uns}$	57.8	63.4	70.3	64.8
$RaRe^{0}_{uns}$	59.1	71.8	67.6	67.5
$RaRe^{l}$	64.0	72.5	72.5	70.0
HSup	79.1	85.7	87.1	89.5

Table 2: The performance of RaRe and BEA in terms of phrase-based F_1 on CoNLL NER datasets compared with state-of-the-art benchmark methods. Resource requirements are indicated with superscripts, p: parallel corpus, w: Wikipedia, d: dictionary, l: 100 NER annotation, 0: no extra resources.

CoNLL Dataset Finally, we apply our model to the CoNLL-02/03 datasets, to benchmark our technique against related work. This corpus is much less rich than Wikiann used above, as it includes only four languages (en, de, nl, es), and furthermore, the languages are closely related and share the same script. Results in Table 2 show that our methods are competitive with benchmark methods, and, moreover, the use of 100 annotated sentences in the target language ($RaRe^{l}$) gives good improvements over unsupervised models.[12] Results also show that MV does very well, especially MV^{ent}, and its performance is comparable to BEA's. Note that there are only 3 source models and none of them is clearly bad, so BEA estimates that they are similarly reliable which results in little difference in terms of performance between BEA and MV.

5 Related Work

Two main approaches for cross-lingual transfer are representation and annotation projection. Representation projection learns a model in a high-

[12]For German because of its capitalisation pattern, we lowercase all the source and target data, and also remove German as a source model for other languages.

resource source language using representations that are cross-linguistically transferable, and then directly applies the model to data in the target language. This can include the use of cross-lingual word clusters (Täckström et al., 2012) and word embeddings (Ammar et al., 2016; Ni et al., 2017), multitask learning with a closely related high-resource language (e.g. Spanish for Galician) (Cotterell and Duh, 2017), or bridging the source and target languages through phonemic transcription (Bharadwaj et al., 2016) or Wikification (Tsai et al., 2016). In annotation projection, the annotations of tokens in a source sentence are projected to their aligned tokens in the target language through a parallel corpus. Annotation projection has been applied to POS tagging (Yarowsky et al., 2001; Das and Petrov, 2011; Duong et al., 2014; Fang and Cohn, 2016), NER (Zitouni and Florian, 2008; Ehrmann et al., 2011; Agerri et al., 2018), and parsing (Hwa et al., 2005; Rasooli and Collins, 2015). The Bible, Europarl, and recently the Watchtower has been used as parallel corpora, which are limited in genre, size, and language coverage, motivating the use of Wikipedia to create weak annotation for multilingual tasks such as NER (Nothman et al., 2013). Recent advances in (un)supervised bilingual dictionary induction (Gouws and Søgaard, 2015; Duong et al., 2016; Lample et al., 2018; Artetxe et al., 2018; Schuster et al., 2019) have enabled cross-lingual alignment with bilingual dictionaries (Mayhew et al., 2017; Xie et al., 2018). Most annotation projection methods with few exceptions (Täckström, 2012; Plank and Agić, 2018) use only one language (often English) as the source language. In multi-source language setting, majority voting is often used to aggregate noisy annotations (e.g. Plank and Agić (2018)). Fang and Cohn (2016) show the importance of modelling the annotation biases that the source language(s) might project to the target language.

Transfer from multiple source languages: Previous work has shown the improvements of multi-source transfer in NER (Täckström, 2012; Fang et al., 2017; Enghoff et al., 2018), POS tagging (Snyder et al., 2009; Plank and Agić, 2018), and parsing (Ammar et al., 2016) compared to single source transfer, however, multi-source transfer might be noisy as a result of divergence in script, phonology, morphology, syntax, and semantics between the source languages, and the target language. To capture such differences, various methods have been proposed: latent variable models (Snyder et al., 2009), majority voting (Plank and Agić, 2018), utilising typological features (Ammar et al., 2016), or explicitly learning annotation bias (Fang and Cohn, 2017). Our work is also related to knowledge distillation from multiple source models applied in parsing (Kuncoro et al., 2016) and machine translation(Kim and Rush, 2016; Johnson et al., 2017). In this work, we use truth inference to model the transfer annotation bias from diverse source models.

Finally, our work is related to truth inference from crowd-sourced annotations (Whitehill et al., 2009; Welinder et al., 2010), and most importantly from diverse classifiers (Kim and Ghahramani, 2012).

6 Conclusion

Cross-lingual transfer does not work out of the box, especially when using large numbers of source languages, and distantly related target languages. In an NER setting using a collection of 41 languages, we showed that simple methods such as uniform ensembling do not work well. We proposed two new multilingual transfer models (`RaRe` and `BEA`), based on unsupervised transfer, or a supervised transfer setting with a small 100 sentence labelled dataset in the target language. We also compare our results with `BWET` (Xie et al., 2018), a state-of-the-art unsupervised single source (English) transfer model, and showed that multilingual transfer outperforms it, however, our work is orthogonal to their work in that if training data from multiple source models is created, `RaRe` and `BEA` can still combine them, and outperform majority voting. Our unsupervised method, $\texttt{BEA}_{\texttt{uns}}$, provides a fast and simple way of annotating data in the target language, which is capable of reasoning under noisy annotations, and outperforms several competitive baselines, including the majority voting ensemble, a low-resource supervised baseline, and the oracle single best transfer model. We show that light supervision improve performance further, and that our second approach, `RaRe`, based on ranking transfer models and then retraining on the target language, results in further and more consistent performance improvements.

References

Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, and German Rigau. 2018. Building named entity recognition taggers via parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia.

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96. Asian Federation of Natural Language Processing.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 845–850.

Long Duong, Trevor Cohn, Karin Verspoor, Steven Bird, and Paul Cook. 2014. What can we get from 1000 tokens? A case study of multilingual POS tagging for resource-poor languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 886–897.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 118–124.

Jan Vium Enghoff, Søren Harrison, and Željko Agić. 2018. Low-resource named entity recognition via multi-source projection: Not quite there yet? In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 195–201.

Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource POS tagging using cross-lingual projection. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 178–186.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 587–593.

Meng Fang, Yuan Li, and Trevor Cohn. 2017. Learning how to active learn: A deep reinforcement learning approach. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 595–605.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara I. Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Michael Irwin Jordan. 1998. *Learning in graphical models*, volume 89. Springer Science & Business Media.

Hyun-Chul Kim and Zoubin Ghahramani. 2012. Bayesian classifier combination. In *AISTATS*, volume 22 of *JMLR Proceedings*, pages 619–627. JMLR.org.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, and Noah A. Smith. 2016. Distilling an ensemble of greedy dependency parsers into one mst parser. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1744–1753.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1470–1480.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Barbara Plank and Željko Agić. 2018. Distant supervision from disparate sources for low-resource part-of-speech tagging. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 614–620.

Mohammad Sadegh Rasooli and Michael Collins. 2015. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 328–338.

Vikas C. Raykar and Shipeng Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *J. Mach. Learn. Res.*, 13:491–518.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2009. Adding more languages improves unsupervised multilingual part-of-speech tagging: a bayesian non-parametric approach. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 83–91.

Oscar Täckström. 2012. Nudging the envelope of direct transfer methods for multilingual named entity recognition. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 55–63.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 219–228.

Peter Welinder, Steve Branson, Serge J. Belongie, and Pietro Perona. 2010. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, pages 2424–2432. Curran Associates, Inc.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier R. Movellan. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, pages 2035–2043. Curran Associates, Inc.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8.

Imed Zitouni and Radu Florian. 2008. Mention detection crossing the language barrier. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 600–609.

A Appendices

A.1 Hyperparameters

We tuned the batch size and the learning rate using development sets in four languages,[13] and then fixed these hyperparameters for all other languages in each model. The batch size was 1 sentence in low-resource scenarios (in baseline LSup and fine-tuning of RaRe), and to 100 sentences, in high-resource settings (HSup and the pretraining phase of RaRe). The learning rate was set to 0.001 and 0.01 for the high-resource and low-resource baseline models, respectively, and to 0.005, 0.0005 for the pretraining and fine-tuning phase of RaRe based on development results for the four languages. For CoNLL datasets, we had to decrease the batch size of the pre-training phase from 100 to 20 (because of GPU memory issues).

A.2 Cross-lingual Word Embeddings

We experimented with Wiki and CommonCrawl monolingual embeddings from fastText (Bojanowski et al., 2017). Each of the 41 languages is mapped to English embedding space using three methods from MUSE: 1) supervised with bilingual dictionaries; 2) supervised by identical character sequences; and 3) unsupervised using adversarial learning (Lample et al., 2018). The cross-lingual mappings are evaluated by precision at $k = 1$. The resulting cross-lingual embeddings are then used in NER direct transfer in a leave-one-out setting for the 41 languages (41×40 transfers) whose mean F_1 is shown in Table 3. CommonCrawl despite having larger text corpora, doesn't perform well in bilingual induction, and consequently in diract transfer NER. It is also evident that using the identical character strings instead of a bilingual dictionary as seed for the supervised method doesn't significantly hurt the mapping and the direct NER transfer performance, while being extensible to languages for which there is no bilingual dictionary available. Based on our experiments, RaRe with $k = 40$ achieves average F_1 of 77.9 with supervised corss-lingual embeddings versus 76.9 (as shown in Table 4) with cross-lingual embeddings seeded by identical character strings, which is not substantially higher. Experiments with unsupervised mappings performed substantially worse than supervised, and so we didn't explore further.

[13] Afrikaans, Arabic, Bulgarian and Bengali.

	Unsup		IdentChr		Sup	
	crawl	wiki	crawl	wiki	crawl	wiki
word translation accuracy	34	24	43	**53**	50	**54**
average F_1 in direct transfer	26	21	37	**44**	39	**45**

Table 3: The effect of the choice of monolingual word embeddings (Common Crawl and Wikipedia), and their cross-lingual mapping on NER direct transfer.

A.3 Direct Transfer Results

In Figure 5 the performance of an NER model trained in a high-resource setting on a source language applied on the other 40 target languages (leave-one-out) is shown. An interesting finding is that symmetry does not always hold (e.g. id vs. ms or fa vs. ar).

A.4 Detailed Low-resource Results

The result of applying baselines, proposed models and their variations, and unsupervised transfer model of Xie et al. (2018) are shown in Table 4.

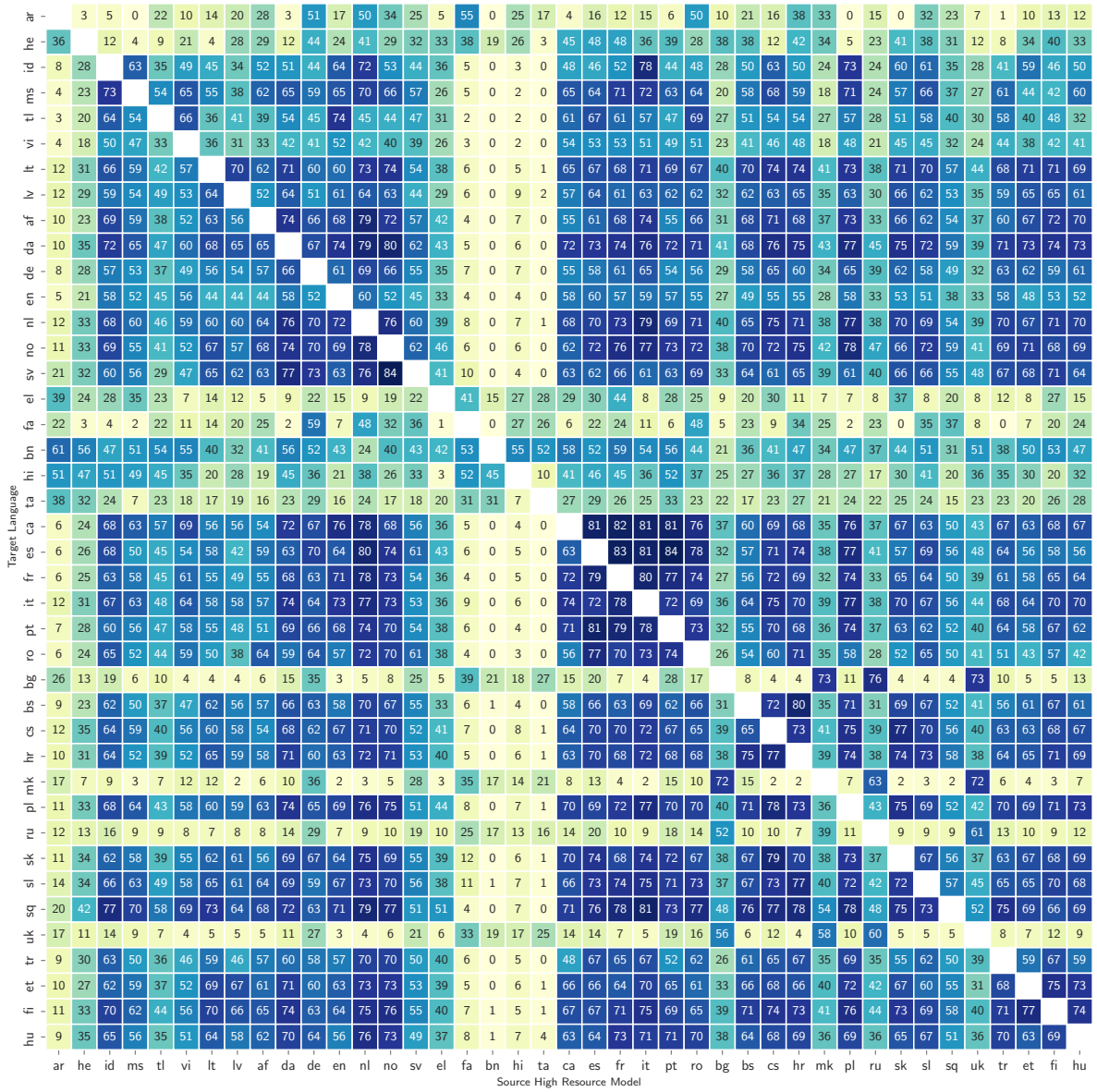

Figure 5: The direct transfer performance of a source NER model trained in a high-resource setting applied on the other 40 target languages, and evaluated in terms of phrase-level F_1.

				Supervised						Unsupervised						
	#train(k)	#test(k)	BiDic.P@1	HSup	LSup	RaRe t1	RaRe t10	RaRe all	$\text{BEA}^{\text{ent}}_{\text{sup}}$ t10	RaRe_{uns}	BWET	$\text{BEA}^{\text{ent}}_{\text{uns}\times 2}$ t10	$\text{BEA}^{\text{ent}}_{\text{uns}}$	$\text{BEA}^{\text{tok}}_{\text{uns}}$	MV^{tok}	Oracle
af	5	1	36	84	59	73	79	79	80	76	64	79	79	74	75	80
ar	20	10	46	88	64	71	74	74	65	26	19	54	45	54	12	56
bg	20	10	55	90	61	80	81	81	81	5	51	81	65	54	4	76
bn	10	1	1	95	70	68	74	74	69	65	36	67	66	60	56	63
bs	15	1	30	92	63	80	79	80	78	76	52	80	78	77	69	82
ca	20	10	70	91	62	82	86	84	86	80	62	85	80	79	72	83
cs	20	10	64	90	62	77	78	75	78	73	59	77	75	72	71	78
da	20	10	68	90	62	77	81	81	82	79	68	83	82	79	78	80
de	20	10	73	86	58	73	74	73	72	69	63	72	71	64	68	70
el	20	10	55	89	61	67	67	67	54	13	45	49	43	34	13	45
en	20	10	—	81	47	64	65	64	65	58	—	63	61	57	56	61
es	20	10	83	90	63	83	84	84	85	76	62	85	81	76	73	84
et	15	10	41	90	64	73	77	77	78	72	58	78	78	71	73	75
fa	20	10	33	93	74	78	81	79	69	30	16	65	50	52	15	60
fi	20	10	58	89	67	78	80	80	81	76	68	81	80	69	77	78
fr	20	10	82	88	57	81	81	80	84	75	59	83	79	73	71	80
he	20	10	52	85	53	61	61	60	55	40	26	54	54	46	34	50
hi	5	1	29	85	68	64	74	73	68	48	27	64	61	58	35	54
hr	20	10	48	89	61	74	79	78	80	76	49	80	79	77	73	78
hu	20	10	64	90	59	75	79	78	80	71	55	79	79	69	73	76
id	20	10	68	91	67	82	83	81	75	59	62	73	67	61	62	79
it	20	10	77	89	60	80	81	80	82	75	59	81	78	76	72	79
lt	10	10	26	86	62	72	79	80	79	76	48	80	80	75	77	74
lv	10	10	31	91	68	70	75	75	69	68	40	69	69	67	65	66
mk	10	1	50	91	67	79	82	81	80	4	38	79	66	48	3	75
ms	20	1	48	91	66	78	80	78	74	69	62	68	67	63	68	74
nl	20	10	76	89	59	78	80	80	81	77	63	82	81	78	76	79
no	20	10	67	90	65	79	82	81	83	79	59	83	83	77	79	79
pl	20	10	66	89	61	76	79	78	81	73	63	82	80	77	76	78
pt	20	10	80	90	59	79	81	80	82	77	65	82	77	74	70	82
ro	20	10	67	92	66	80	82	82	80	76	46	78	76	74	67	77
ru	20	10	59	86	53	73	71	71	56	10	38	53	40	36	11	61
sk	20	10	52	91	62	76	79	79	80	74	50	79	76	76	71	79
sl	15	10	47	92	64	76	80	80	79	76	58	79	78	76	73	78
sq	5	1	37	88	69	79	84	84	83	82	59	83	84	76	79	79
sv	20	10	61	93	69	83	83	84	82	77	60	79	80	69	76	84
ta	15	1	7	84	54	44	53	53	46	35	12	39	42	25	29	38
tl	10	1	20	93	66	75	82	80	78	65	60	62	60	57	52	76
tr	20	10	61	90	61	75	77	77	77	70	53	77	76	67	67	71
uk	20	10	45	89	60	70	78	79	70	5	35	64	58	49	6	60
vi	20	10	54	88	55	64	72	72	61	58	53	56	55	48	47	56
μ	—	—	—	89.2	62.1	74.3	77.4	76.9	74.8	60.2	50.5	72.8	69.7	64.5	56.7	71.6
σ	—	—	—	2.8	5.2	7.3	6.4	6.4	9.6	24.1	14.7	11.5	12.6	13.7	25	11.5

Table 4: The size of training and test sets (development set size equals test set size) in thousand sentences, and the precision at 1 for Bilingual dictionaries induced from mapping languages to the English embedding space (using identical characters) is shown (BiDic.P@1). F_1 scores on the test set, comparing baseline supervised models (HSup, LSup), multilingual transfer from top k source languages (RaRe, 5 runs, $k = 1, 10, 40$), an unsupervised RaRe with uniform expertise and no fine-tuning (RaRe_{uns}), and aggregation methods: majority voting (MV^{tok}), $\text{BEA}^{\text{tok}}_{\text{uns}}$ and $\text{BEA}^{\text{ent}}_{\text{uns}}$ (Bayesian aggregation in token- and entity-level), and the oracle single best annotation (Oracle). We also compare with BWET (Xie et al., 2018), an unsupervised transfer model with state-of-the-art on CoNLL NER datasets. The mean and standard deviation over all 41 languages, μ, σ, are also reported.