Entity Projection via Machine Translation for Cross-Lingual NER

Alankar Jain **Bhargavi Paranjape** **Zachary C. Lipton**
Carnegie Mellon University
Pittsburgh, USA
{alankarjain91,bhargavi22294}@gmail.com, zlipton@cmu.edu

Abstract

Although over 100 languages are supported by strong off-the-shelf machine translation systems, only a subset of them possess large annotated corpora for *named entity recognition*. Motivated by this fact, we leverage machine translation to improve annotation-projection approaches to cross-lingual named entity recognition. We propose a system that improves over prior entity-projection methods by: (a) leveraging machine translation systems twice: first for translating sentences and subsequently for translating entities; (b) matching entities based on orthographic and phonetic similarity; and (c) identifying matches based on distributional statistics derived from the dataset. Our approach improves upon current state-of-the-art methods for cross-lingual named entity recognition on 5 diverse languages by an average of 4.1 points. Further, our method achieves state-of-the-art F_1 scores for Armenian, outperforming even a monolingual model trained on Armenian source data.[1]

1 Introduction

While machine learning methods for various *Natural Language Processing (NLP)* tasks have progressed rapidly, the benefits accrue disproportionately among languages endowed with large annotated corpora. Owing to the dependence of state-of-the-art deep learning approaches on massive amounts of data, creating suitable datasets can be prohibitively expensive. This asymmetry between resource-rich and relatively under-resourced languages has inspired work on *cross-lingual* approaches that leverage annotated datasets from the former to build strong models for the latter.

This paper focuses on cross-lingual approaches to *Named Entity Recognition (NER)*, owing to NER's importance as a core component in *information retrieval* and *question answering* systems. Specifically, we focus on **medium-resource** languages. We define these to be languages for which although annotated NER corpora do not exist, off-the-shelf *Machine Translation (MT)* systems, such as Google Translate[2], do. We are motivated by the fact that although there are fewer than 50 languages for which large NER datasets (greater than 200k tokens) with gold annotations are publicly available[3], many more languages are supported by good-quality MT (Wu et al., 2016). Google Translate alone supports 103 languages[4], many of which have either no, or only small, NER datasets.

We address the setting where annotated corpora exist in the source (resource-rich) language—English in our experiments—but for the target (medium-resource) language, we can only afford to label a small validation set. We tackle this problem by first creating an unlabeled dataset in the target language by translating each sentence in the source dataset to the target language. For MT, we use Google Translate, motivated by its large coverage. Next, we annotate this dataset via *entity projection*—first aligning every entity in a source sentence with its counterpart in the corresponding target sentence (*entity alignment*) and then projecting the tags from source to target in the aligned entity pairs (*tag projection*). One consequence of relying on MT as opposed to word-by-word or phrase-by-phrase translation is that the entity projection step can be difficult, owing to the frequency with which original sentences and their translated counterparts are not word-for-word aligned.

Our proposed solution to this problem consists of (a) leveraging MT again for translating entities; (b) matching entities based on orthographic and

[1] Code for our paper can be found at: https://github.com/alankarj/cross_lingual_ner

[2] https://cloud.google.com/translate/

[3] http://damien.nouvels.net/resourcesen/corpora.html

[4] https://cloud.google.com/translate/docs/languages

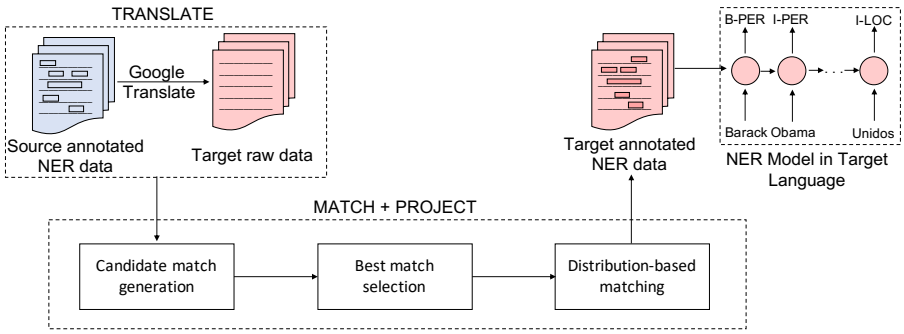

Figure 1: A schematic diagram representing the chief steps in our method.

phonetic similarity; and (c) identifying matches based on distributional statistics derived from the dataset. *Importantly, while our method depends on several matching heuristics, these techniques are remarkably portable across target languages, requiring the tuning of only two hyperparameters.* Our method achieves state-of-the art F_1 scores for cross-lingual NER for Spanish ($+1.1$ points), German ($+1.4$ points), and Chinese ($+5$ points) and beats state-of-the-art baselines on Hindi ($+2.1$ points) and Tamil ($+5$ points). Further, it achieves state-of-the-art F_1 scores for Armenian, a medium-resource language, beating a monolingual model trained on Armenian source data by 0.4 points.

2 Related Work

Cross-lingual approaches have been applied to many NLP tasks, including part-of-speech tagging (Yarowsky et al., 2001; Xi and Hwa, 2005; Das and Petrov, 2011; Täckström et al., 2013), parsing (Hwa et al., 2005; Zeman and Resnik, 2008; Smith and Eisner, 2009; Ganchev et al., 2009), and semantic role labeling (Tonelli and Pianta, 2008; Padó and Lapata, 2009; Kozhevnikov and Titov, 2013, 2014). Prior cross-lingual NLP papers cleave roughly into two distinct approaches: *direct model transfer* and *annotation projection*.

2.1 Direct model transfer

These approaches apply models trained on the source language absent modification (to the model) to data from the target language by exploiting a shared representation for the two languages (Täckström et al., 2012; Bharadwaj et al., 2016; Chaudhary et al., 2018; Kozhevnikov and Titov, 2014; Ni et al., 2017). However, direct model transfer techniques face a problem when applied to markedly dissimilar languages: they lack of lexicalized (especially character-based) features, which are known to have predictive power for tasks such as NER. Xie et al. (2018) provide evidence for this in the cross-lingual setting, comparing otherwise similar annotation projection approaches that differ in their use of lexicalized features.

2.2 Annotation projection

These approaches to cross-lingual NLP train a model in the target language. This requires first projecting annotations from the source data to the (unlabeled) target data. Many approaches in this category rely upon parallel corpora (Yarowsky et al., 2001; Hwa et al., 2005; Zeman and Resnik, 2008; Ehrmann et al., 2011; Fu et al., 2011; Ni et al., 2017), first annotating the source data using a trained model and then projecting the annotations. Only a few works explore the use of MT to first translate a gold annotated corpus to obtain a *synthetic* parallel corpus and then project annotations (Tiedemann et al., 2014). Shah et al. (2010) go in the opposite direction, translating target to source using Google Translate, annotating the translated source sentences using a trained NER system and then projecting annotations back.

When projecting annotations, one encounters the problem of word alignment. Most of the existing works (Yarowsky et al., 2001; Shah et al., 2010; Ni et al., 2017) rely upon unsupervised alignment models from statistical MT literature, such as IBM Models 1-6 (Brown et al., 1993; Och and Ney, 2003). Other works focus on low-resource settings (Mayhew et al., 2017; Xie et al., 2018) perform translation word-by-word or phrase-by-phrase, and thus do not need to perform word alignment. Several papers explore heuristics such as using Wikipedia links across languages to align entities (Richman and Schone, 2008; Nothman et al., 2013; Al-Rfou et al., 2015), matching tokens based on their surface forms and transliterations either in an unsupervised

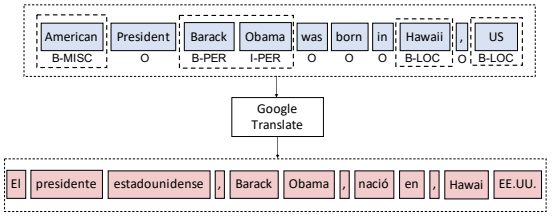

Figure 2: The *Translate* step of our method.

manner (Samy et al., 2005; Ehrmann et al., 2011) or as features in a supervised model trained on a small *seed* dataset (Feng et al., 2004). Many of these papers often rely on language-specific features (Feng et al., 2004) and evaluate their alignment methods on only a few languages.

To our knowledge, few works effectively use translation for annotation projection for NER, especially for medium-resource languages for which strong MT systems exist. Motivated by this research gap, we explore the use of MT systems for translating the dataset and for annotation projection and thus do not rely on parallel corpora. However, we demonstrate the efficacy of our projection method in all three settings: (a) translation from source to target, (b) using parallel corpora and (c) translation from target to source.

3 The Translate-Match-Project Method

In our formulation, we are given an annotated NER corpus in the source language: $\mathcal{D}_A^S = \{(x^{Si}, y^{Si}) : i = 1, 2, ..., N\}$, where $x^{Si} = (x_1^{Si}, ..., x_{L^{Si}}^{Si})$ is the ith source sentence with L^{Si} tokens and $y^{Si} = (y_1^{Si}, ..., y_{L^{Si}}^{Si})$ are the NER tags from a fixed tag set. We work with four tags: PER (person), ORG (organisation), LOC (location), and MISC (miscellaneous). We follow the commonly-used IOB (Inside Outside Beginning) tagging format (Ramshaw and Marcus, 1999). In our experiments, we work with English as the source language due to the availability of high-quality annotated corpora, e.g., CoNLL 2002 (Sang, 2002) and OntoNotes 4.0 (Weischedel et al., 2011). However, our method can easily be applied to any other resource-rich source language as well. See Figure 2 for an example of an annotated source sentence (blue), (x^{Si}, y^{Si}).

We are given a small labeled development set data (but no training data) in the target language T for tuning hyperparameters. Our method, denoted *Translate-Match-Project (TMP)*, proceeds in three steps: First, we translate the annotated corpus in S to T using an off-the-shelf MT system (Google Translate). This results in an un-labeled dataset in the target language, $\mathcal{D}^T = \{x^{Ti} : i = 1, 2, ..., N\}$ (Figure 2); Second, we identify and tag all named entities in the translated target sentences by entity projection, which involves entity alignment and tag projection. We perform entity alignment by first constructing a set of potential matches in the target sentence for every entity in the source sentence (*candidate match generation*, Section 3.1) and then by selecting the best matching pairs of source and target entities (*best match selection*, Section 3.2); Third, after alignment, we project the tag type (PER, LOC, etc.) from the source to the target entity in every pair of aligned entities by adhering to the IOB tagging scheme in target. In Figure 1, we depict the complete pipeline.

3.1 Candidate match generation

To generate candidate matches for an entity in a source sentence, we construct a set of its potential translations and then find matches for each in the corresponding target sentence. We find these matches by token-level matching and then *concatenate matched tokens to obtain multi-token matches*. We drop the index i below for ease of notation.

Token-level matching Consider a source entity $e^S = (x_j^S, ..., x_k^S)$. For every $e^S \in \mathcal{E}^S$, where $\mathcal{E}^S$ is the set of all entities in a source sentence, we obtain a set of potential translations in the target language, $\mathcal{T}(e^S)$, via MT. However, in some cases, translating a standalone entity produces a different translation from that which emerges when translating a full sentence. For example, Google Translate maps the source entity "UAE" to "Emiratos Árabes Unidos" in most sentences but the word-by-word translation is "EAU". Similarly, the (person) name "Tang" (e.g., "Mr. Tang") remains "Tang" in translated sentences, but is translated to "Espiga" (Spanish for "spike", synonymous with the English word "tang"). We address these problems by augmenting $\mathcal{T}(\cdot)$ with translations from *publicly available* bilingual lexicons ("UAE" translates to "Emiratos Árabes Unidos" in one of the lexicons we use) and retain a copy of the source entity ("Tang" will now find a match in the target sentence). Finally, $\mathcal{T}$("UAE") looks roughly like: {"EAU" [Google Translate], "UAE" [copy], "Emiratos Árabes Unidos" [lexicon]}. We note that lexicons exist for a large number of languages to-

day[5]. However, we demonstrate that our method also works in absence of such lexicons in our case study for Armenian (Section 4.3).

Next, we tokenize each candidate translation in $\mathcal{T}(e^S)$ to obtain a set of translation tokens for e^S, $\mathcal{T}^w(e^S)$. For example, $\mathcal{T}^w$("UAE") = {"EAU", "UAE", "Emiratos", "Árabes", "Unidos"}. We do this to allow for soft token-level matches because we observed empirically that matching exact entity phrases might result in few matches. Next, we obtain a match for each hypothesis token $h \in \mathcal{T}^w(e^S)$ by matching it with each reference token $x_l^S \ \forall \ l \in \{1, ..., L^T\}$ in the target sentence $x^T = (x_1^T, ..., x_{L^T}^T)$ of length L^T. This match is carried out at (a) the orthographic (surface form) level; and (b) the phonetic level, by matching transliterations in the International Phonetic Alphabet (IPA) of the two tokens. In either case, we look for the longest sequence of characters in h that are an affix (prefix or suffix) of x_l^T. This soft *affix-matching heuristic* allows for inflection in morphologically-rich target languages. The (token-level) score for the match is given as follows:

$$s^w(h, x_l^T) = \min\left\{\frac{n_l}{L_h}, \frac{n_l}{L_{x_l^T}}\right\}$$

Here, L_h and $L_{x_l^T}$ are the lengths (in characters) of the hypothesis and reference tokens and n_l is the number of matching characters. We take minimum in order to enforce a stricter notion of fractional (soft) match. For example, the phrase "German first-time registrations..." [English] gets translated to "Los registros Alemanes por primera..." [Spanish]. Using our matching heuristic, "Alemán" $\in \mathcal{T}^w$("German") matches to the reference token "Alemanes" with a score of 0.5, since $n_l = 4$ ("Alem") and $L_{x_l^T} = 8 > L_h = 6$. Next, we define the matching (entity-level) score between a source entity e^S and any target token x_l^T as follows:

$$s^e(e^S, x_l^T) = \max_{h \in \mathcal{T}^w(e^S)} s^w(h, x_l^T)$$

Note that the token-level scores, s^w, include scores based on both orthographic and phonetic match, and thus, the entity-level scores, s^e, correspond to the best token-level match (orthographic or phonetic) between any hypothesis token $h \in \mathcal{T}^w(e^S)$ and a target token x_l^T. In Figure 3, we depict the token-level matching procedure. The

[5]Panlex (Kamholz et al., 2014) has lexicons for 10k different languages.

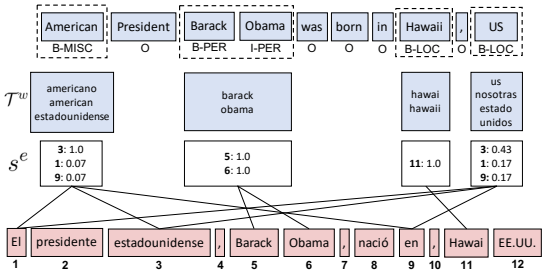

Figure 3: Token-level matching: Blue boxes in row-2 ($\mathcal{T}^w$) show sets of potential translations and white boxes in row-3 (s^e) show target tokens (numbered) each source entity can match with, with their scores.

score between the entity "American" and the target "estadounidense"(labeled 3 in the figure) is the maximum over matching scores between any token in {"americano", "american", "estadounidense"} and the target token ("estadounidense"), i.e., 1.0 (exact match) since the scores for the first two tokens in $\mathcal{T}^w(\cdot)$ are 0. Note some artifacts of token-level matching: (a) our matching heuristic currently handles prefixes or suffixes, but can potentially be extended with character edit distance for other types of affixes (e.g., circumfix) (b) every target token can match with multiple source entities (for e.g., "El" matches with both "American" and "US") and (c) some source entities might fail to find their true match ("US" fails to match with "EE.UU.", a possibly erroneous translation of "US" provided by Google Translate). Further, many matches are of very poor quality (especially those with stop words such as "El" and "en"). We address these issues and describe how to convert these token-level matches into spans to get multi-token target entities next.

Span match generation After token-level matching, we construct a list of potential entity spans in the target sentence that match with a given source entity e^S by grouping adjacent target tokens for which a token-level matching score $s^e(e^S, \cdot)$ is above a threshold δ (to remove spurious matches). In other words, we construct the following set:

$$\mathcal{M}(e^S) = \{\text{span}(q, r) : s^e(e^S, x_u^T) \geq \delta \ \forall \ q \leq u \leq r\}$$

Here, $\text{span}(q, r) = (x_q^T, ..., x_r^T)$ is the phrase spanning tokens indexed from q to r ($1 \leq q, r \leq L^T$) in the target sentence. Further, we require $\text{span}(q, r)$ to be maximal in the sense that $\forall \ q' < q$ and $\forall \ r' > r$, $\text{span}(q', r') \notin \mathcal{M}(\cdot)$, i.e., any target token before or after the span have at best a weak match ($s^e(e^S, \cdot) < \delta$) with e^S.

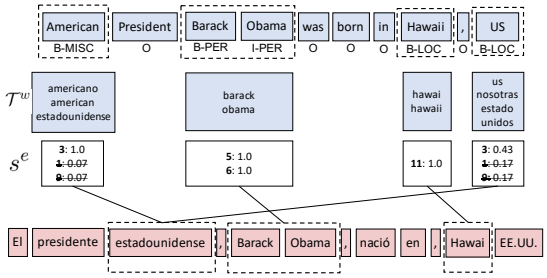

Figure 4: Span match generation: Adjacent target tokens with matching scores higher than a threshold (0.25 here) are concatenated to form span-level matches.

In our running example, choosing $\delta = 0.25$ results in the spans shown in Figure 4, eliminating spurious matches (with "El" and "en") and concatenating "Barack" and "Obama" in the target sentence. However, the token "estadounidense" is still matched with two different source entities. We solve this problem in the next step.

3.2 Best match selection

For selecting the best matching pair of entities, we first expand the set of potential translations $\mathcal{T}(\cdot)$ to include all possible token-level permutations of the translations. We call this set $\mathcal{T}^p(\cdot)$. For example, $\mathcal{T}^p$("UAE") = {"EAU", "UAE", "Emiratos Árabes Unidos", "Emiratos Unidos Árabes", "Árabes Emiratos Unidos", ...}. Then, we greedily align e^S with the target entity span from the set $\mathcal{M}(e^S)$, with the least character edit distance $d_E(\cdot,\cdot)$ from any translation in $T^p(e^S)$, i.e.,

$$e^T = \underset{\text{span}(\cdot,\cdot)\in\mathcal{M}(e^S)}{\text{argmin}} \; d_E(e^S, \text{span}(\cdot,\cdot))$$

In this manner, we form aligned entity pairs (e^S, e^T), along which tags can then be projected. In our running example, since the edit distance between "estadounidense" (in $\mathcal{T}^p(\cdot)$) and the target single-token span "estadounidense" is 0, and is lower than that with "estado", we match "estadounidense" with "American", tagging it B-MISC.

3.3 Distribution-based matching

After selecting best matching pairs, there still remain some source entities that do not find any matching target entity ("US" in our example). These arise either due to significant differences between word-by-word and contextual sentence-level translations either due to literal (e.g., "West Bank" gets translated to "Cisjordania" in sentences and "Banco Oeste" otherwise) or possibly incorrect translations (e.g., "U.S." gets translated to "EE.UU." in sentences and "NOSOTRAS" otherwise).

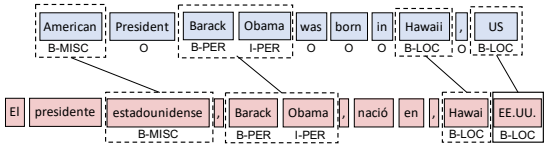

Figure 5: Final aligned pairs and projected tags.

We remedy this by exploiting corpus-level consistency in such discrepancies. For every unmatched source entity, we construct a set of top-k potential matches ordered by their tf-idf (term frequency—inverse document frequency) scores, where tf is calculated over all sentences containing at least one unmatched entity and the idf score is calculated over the entire dataset to severely penalize commonly occurring tokens. Finally, we match each unmatched source entity with an unmatched span in its top-k list with the highest tf-idf score. Figure 5 shows the final matches and tags.

4 Experimental Evaluation

Data In order to compare our method against benchmarks reported for prior approaches, we evaluate its performance on three European languages: Spanish (*es*), Dutch (*nl*) and German (*de*). Further, for a more extensive evaluation, we conduct additional experiments in an Indo-Aryan language (Hindi (*hi*)), a Dravidian language (Tamil (*ta*)), and Simplified Chinese (*zh*).

For all languages except Chinese, we use English NER training data from the CoNLL 2003 shared task (Sang and Meulder, 2003) to translate into the target language. For Chinese, we sample the same number of sentences as in the CoNLL 2003 corpus (14,041) from the OntoNotes 4.0 (2012) dataset for English (Weischedel et al., 2011) to minimize distribution shift from Chinese development data. The development and test datasets for Spanish, Dutch and German are obtained from CoNLL 2002 (Sang, 2002) and CoNLL 2003 shared tasks. For Hindi and Tamil, we obtain the NER corpus from FIRE 2013 shared task[6]. Since this task doesn't provide the test dataset, we create our own splits: two-thirds for training and one-sixth each for development and test (to match with the proportions in the CoNLL dataset). For Chinese, we use the OntoNotes 4.0 development and test datasets. English, Spanish, Dutch and German contain PER,

[6]http://au-kbc.org/nlp/NER-FIRE2013/

Method	Spanish	German	Dutch	Chinese	Hindi	Tamil	Average
TMP	**73.5 ± 0.4**	**61.5 ± 0.4**	69.9 ± 0.4	**50.1 ± 0.2**	**41.7 ± 1.3**	**33.8 ± 2.2**	**55.1 ± 0.8**
fast-align	65.0 ± 1.2	60.1 ± 0.9	67.6 ± 0.7	45.1 ± 0.8	39.6 ± 1.1	28.8 ± 1.8	51.0 ± 1.1
BWET	72.4 ± 0.6	57.8 ± 0.1	**70.4 ± 1.2**	3.51 ± 0.8	26.6 ± 0.8	15.6 ± 0.9	48.5 ± 0.7
Co-decoding	65.1	58.5	65.4	-	-	-	-
Polyglot-NER	63.0	-	59.6	-	-	-	-
Monolingual	86.3 ± 0.4	78.2 ± 0.4	86.4 ± 0.2	68.59 ± 0.3	65.8 ± 1.2	51.8 ± 1.0	73.7 ± 0.6

Table 1: Test F_1 scores for our method (TMP), 4 cross-lingual baselines and a model trained on monolingual data.

ORG, LOC and MISC tags, while Hindi, Tamil and Chinese were preprocessed to contain only PER, LOC and ORG tags. We use MUSE ground-truth bilingual lexicons[7] (gold lexicon) for augmenting the set of potential entity translations and use Epitran (Mortensen et al., 2018) for obtaining IPA transliterations.

Baselines We compare against four other annotation projection approaches that have achieved state-of-the-art results on some of our datasets. Xie et al. (2018) (BWET) use a bilingual lexicon induced using monolingual corpora (Conneau et al., 2017) to translate each source sentence word-by-word and then copy the corresponding NER tags using gold lexicons. As a ceiling for their method, Mayhew et al. (2017) used Google Translate with fast-align (Dyer et al., 2013) (fast-align), an unsupervised expectation maximization based algorithm, for entity alignment. Since this algorithm can produce multiple matches for a given source entity, we post-process the alignments produced by this algorithm and select the longest match and then project tags in the same way as our method. Our third baseline is Ni et al. (2017) (Co-decoding), who use a co-decoding scheme on two different NER models. We also compare our method with Polyglot-NER (Al-Rfou et al., 2015) who use Wikipedia links to project entities. Finally, we also compare our performance with a model trained on annotated data in target language (Monolingual).

NER Model We use the state-of-the-art neural NER tagging model from (Xie et al., 2018) to train TMP and fast-align baseline for all languages. This model adds a self-attention layer to the character- and word-based BiLSTM + CRF model due to Lample et al. (2016). For each experiment, we run our models 5 times using different seeds and report the mean and standard deviation (as recommended by Reimers and Gurevych (2017)) of F_1 measure.

[7] https://github.com/facebookresearch/MUSE

Hyperparameters For the fast-align baseline, we tune their λ parameter, which controls how much the model deviates from perfectly diagonal alignments, for each language separately. For TMP, we tune δ, the score threshold and k, the number of top candidates selected in distribution-based matching. We use the same hyperparameters for the NER model as Xie et al. (2018) for all our experiments.

4.1 Results

Our technique outperforms previous state-of-the-art cross-lingual methods on Spanish, German, Chinese, Hindi and Tamil and performs competitively on Dutch (Table 1). In particular, our method shows marked improvements over BWET, a word-by-word translation baseline, for languages such as German, Hindi, Tamil and Chinese that differ markedly in word ordering (with respect to English), demonstrating the impact of improved machine translation quality on final NER tagging accuracy. For more distant languages, word ordering can drastically affect the position of entities in a sentence, which can hurt performance on a test set in the target language. For instance, consider the Hindi word-by-word translation in Figure 6 (c), which is incoherent and violates the Subj-Obj-Verb ordering of Hindi. On languages that are closer to English, like Spanish and Dutch, the gains are comparatively modest, indicating that word order and quality MT is not critical for such languages.

We also show improvements over the fast-align baseline, which performs unsupervised word-level alignment over the full sequence. This can lead to alignment errors for named entities, which tend to be low-frequency words. Moreover, since fast-align allows for multiple target words to be aligned to a given source word, several noisy tags are added to the target sentence (see Figures 6 (a) and (b)).

4.2 Comparison of projection settings

Having established the performance of TMP as a method for cross-lingual NER, in this section, we

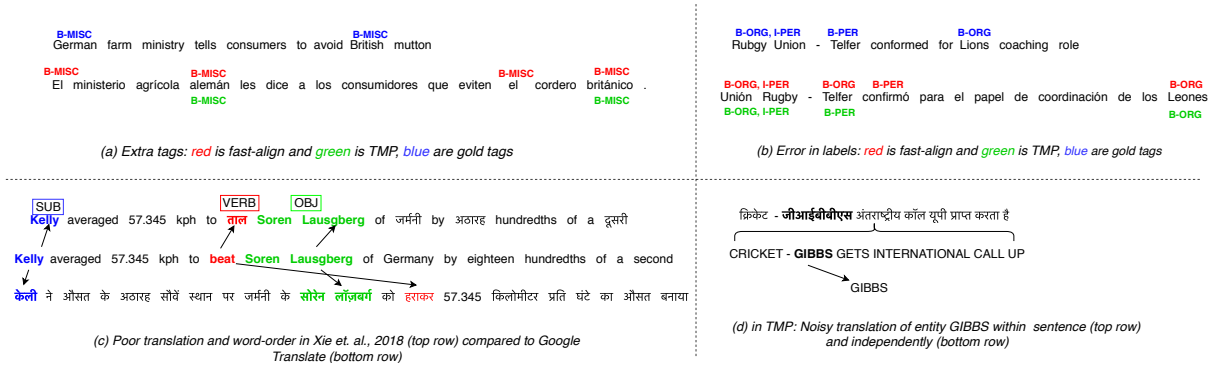

Figure 6: Examples of different errors (details in individual captions).

conduct deeper experiments to evaluate the effectiveness of the matching (M) and projection (P) steps of TMP over the other projection baseline, fast-align. As mentioned in Section 2, there are variants of the annotation projection paradigm for cross-lingual NER that require an entity projection step, namely (i) reversing the direction of machine translation and (ii) using parallel corpora. We compare MP with fast-align for Spanish and Hindi languages under both these settings.

Lang.	Method	Forward	Reverse	Parallel
es	MP	**73.5 ± 0.4**	65.3	61.2 ± 1.2
	fast-align	65.0 ± 1.2	57.8	39.3 ± 0.5
hi	MP	41.7 ± 1.3	47.7	**52.8 ± 1.4**
	fast-align	39.6 ± 1.1	34.3	51.8 ± 1.5

Table 2: Performance of MP and fast-align on Forward, Reverse and Parallel settings in terms of F_1.

Reversing the direction of translation In this setting, we translate the target test set into the source language using Google Translate and then use the NER tagger with state-of-the-art results *Flair*[8] to tag entities in the translated English sentences. Finally, we employ MP/fast-align to project the tagged entities back to the target sentence. As shown in table 2, MP outperforms fast-align for both Spanish and Hindi and performs better than the forward direction translation for Hindi. This can be attributed to (a) the inherent difficulty of NER tagging in Hindi, which is morphologically richer than English and (b) the superior quality of the English NER model.

Parallel corpora In order to remove translation errors while evaluating TMP and fast-align, we experiment with parallel corpora. For English-Spanish, we use the Europarl corpus (Koehn, 2005) and for English-Hindi, the IIT Bombay parallel corpus (Kunchukuttan et al., 2017). We again use *Flair* to obtain NER tags in English, which are then projected to their corresponding target sentences to generate a training dataset, which is used to train an NER model in the target language. To minimize confounding variables, we sample 14k (same as CoNLL) high quality tagged sentences (average confidence score > 0.9). Results in Table 2 show that MP once again outperforms fast-align. Further, it performs better than Forward for Hindi by a significant margin possibly because the chosen parallel corpus is closer in time period to the test set, thereby reducing distribution shift.

4.3 Case study: Armenian

So far, we have only evaluated the performance of our method on languages for which large or moderately-sized gold annotated corpora already exist that provide an upper-bound for cross-lingual NER methods. Here, we evaluate our method on a true medium-resource language, Armenian. Recently, Ghukasyan et al. (2018) introduced a ground truth test corpus for Armenian along with a train corpus with silver annotations extracted from Wikipedia. This test dataset is comprised of 2566 sentences (53k tokens) from political, sports, local and world news between August 2012 and July 2018. Since the English CoNLL 2003 dataset contains sentences nearly two decades older, we expect to see significant distribution shift if we follow TMP (Forward approach). Further, we are not aware of any large English-Armenian parallel corpora. So, we choose the Reverse paradigm for this problem. We achieve an F_1 score of 62.6, which is significantly higher than that achieved by fast-

[8] https://github.com/zalandoresearch/flair

align (44.8). Further, this is 0.4 points higher than the current state-of-the-art model trained on over 160k tokens of Armenian. Note that our model does not make use of any external resources for Armenian (gold lexicons, Epitran, etc.) other than an MT system. This provides evidence towards our proposed approach being an effective and generalizable cross-lingual NER method that can be used for rapid deployment to new languages.

5 Analysis

Measuring alignment accuracy Since we do not possess ground truth word alignments for the "synthetic" parallel corpus generated through translation, we rely on heuristics to measure the accuracy of alignments. We measure the annotation *miss rate* among target sentences with equal or fewer tagged entities as compared to source. We also calculate the *excess rate*, representing the fraction of excess entities among sentences with more tagged entities. Both methods perform similarly in terms of miss rate, 0.79 % (MP) vs 0.83 % (fast-align) on Spanish and 3.96 % (MP) vs 3.48 % (fast-align) on Hindi. However, fast-align seems to add more noisy annotations as compared to MP, with higher excess rates for both Spanish (8.29 % vs 0.49 %) and Hindi (6.35 % vs 2.20 %). A representative illustration of these noisy tags is shown in Figure 6 (a). where fast-align tags frequent words like "El", "de", "en" as entities. To offer a more fine-grained evaluation of alignment performance, we manually annotate 100 examples from the translated Spanish and Hindi training data and calculate precision, recall and F_1 score. MP outperforms fast-align for both the languages (Table 3).

Lang.	Method	Precision	Recall	F_1
es	MP	**96.2**	**96.7**	**96.4**
	fast-align	84.6	87.4	85.9
hi	MP	**87.4**	**77.6**	**82.2**
	fast-align	82.0	76.8	79.3

Table 3: Alignment performance on 100 sentences.

Ablation of features for alignment We also conduct an ablation study (Table 4) to understand the sources of our gains beyond a base model that uses translations only from Google Translate and orthographic affix matching. To this base model, we successively add various features of our method: phonetic matching, exact copy translations, gold lexicons and finally distribution-based alignment (dist) of remaining entities. For both languages, we observe that every additional feature improves the performance of tagging, with the most important features being phonetic matching for Spanish and use of gold lexicons for Hindi. Interestingly, addition of phonetic matching hurts Hindi because of the low value of the threshold ($\delta = 0.25$), which results in spurious matches due to phonetic matching. In Table 4, we also see that the number of entities tagged (as a fraction of total entities) increase with the introduction of almost every feature (however, all matches might not be correct). This underscores the correlation between quality of entity alignment and performance on the downstream tagging task.

Model	es		hi	
	% **Entities**	F_1	% **Entities**	F_1
Base model	91.8	67.8	77.9	37.7
+phonetic	93.7	71.4	83.0	35.0
+copy	97.2	72.2	85.4	37.6
+gold	98.3	73.3	88.5	42.0
+dist	99.9	74.2	94.9	43.4

Table 4: Ablation study for Spanish and Hindi

Sources of errors in TMP We also analyze mistakes made by TMP in aligning entities. Many false negative errors can be traced back to a high threshold δ, resulting in an empty set of candidate matches. Errors also arise due to noise and variation introduced in the contextual sentence level translation of a word (Figure 6 (c) where GIBBS is interpreted as an acronym, (d) where MEDVEDEV is mistranslated). This causes discrepancies between translations of standalone entities and those in context, thereby, causing TMP to not find a match. However, these errors can be reduced as off-the-shelf MT systems continue to improve.

6 Conclusion

In this paper, we tackled the problem of entity projection for cross-lingual NER. Our proposed method leverages MT for translating entities, matches entities based on orthographic and phonetic similarity, and identifies matches based on distributional statistics derived from the dataset to achieve state-of-the-art results for cross-lingual NER on a diverse set of languages. Further, our method beats state-of-the-art monolingual baseline for Armenian, an actual medium-resource language (off-the-shelf translation systems exist, but large-scale NER corpora do not). In the future, we would

like to explore ways to extend our method to languages not supported by Google Translate through the use of pivot languages.

While dependence on MT restricts our approach to languages covered by off-the-shelf MT systems, these systems continue to improve in coverage and quality, outpacing the availability of large-scale corpora for a variety of other tasks. Moreover as translation quality improves, approaches like ours are poised to benefit. Finally, although our method beats state-of-the-art baselines, not surprisingly, it falls short of NER models trained on large monolingual corpora. We suspect that a significant portion of this degradation is due to distribution shift (as evidenced by improvement in Hindi F_1 in Parallel regime). Thus one promising route to improving our models might be to incorporate domain adaptation techniques, which aim to build classifiers robust to various forms of distribution shift.

References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *International Conference on Data Mining (ICDM)*.

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R Mortensen, and Jaime G Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. *arXiv preprint arXiv:1808.09500*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Association for Computational Linguistics (ACL)*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

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.

Donghui Feng, Yajuan Lv, and Ming Zhou. 2004. A new approach for english-chinese named entity alignment. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Ruiji Fu, Bing Qin, and Ting Liu. 2011. Generating chinese named entity data from a parallel corpus. In *International Joint Conference on Natural Language Processing (IJCNLP)*.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 369–377. Association for Computational Linguistics.

Tsolak Ghukasyan, Garnik Davtyan, Karen Avetisyan, and Ivan Andrianov. 2018. pioner: Datasets and baselines for armenian named entity recognition. *arXiv preprint arXiv:1810.08699*.

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

David Kamholz, Jonathan Pool, and Susan M Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *International Conference on Language Resources and Evaluation (LREC)*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Mikhail Kozhevnikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1190–1200.

Mikhail Kozhevnikov and Ivan Titov. 2014. Cross-lingual model transfer using feature representation projection. In *Association for Computational Linguistics (ACL)*.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Association for Computational Linguistics (ACL)*.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Empirical Methods in Natural Language Processing (EMNLP)*.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Association for Computational Linguistics (ACL)*.

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.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research (JAIR)*, 36:307–340.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. *arXiv preprint arXiv:1707.09861*.

Alexander E Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *Association for Computational Linguistics (ACL)*.

Doaa Samy, Antonio Moreno, and Jose M Guirao. 2005. A proposal for an arabic named entity tagger leveraging a parallel corpus. In *International Conference RANLP, Borovets, Bulgaria*, pages 459–465.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning (CoNLL)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Conference on Natural Language Learning at HLT-NAACL*.

Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. 2010. Synergy: a named entity recognition system for resource-scarce languages such as swahili using online machine translation. In *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, pages 21–26.

David A Smith and Jason Eisner. 2009. Parser adaptation and projection with quasi-synchronous grammar features. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics (TACL)*, 1:1–12.

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. Association for Computational Linguistics.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Eighteenth Conference on Computational Natural Language Learning (CoNLL 2014)*.

Sara Tonelli and Emanuele Pianta. 2008. Frame information transfer from english to italian. In *International Conference on Language Resources and Evaluation (LREC)*.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Chenhai Xi and Rebecca Hwa. 2005. A backoff model for bootstrapping resources for non-english languages. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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. Association for Computational Linguistics.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP-08 Workshop on NLP for Less Privileged Languages*.