Cross-lingual named-entity lexicon are an important resource to multilingual NLP tasks such as machine translation and cross-lingual wikification. While knowledge bases contain a large number of entities in high-resource languages such as English and French, corresponding entities for lower-resource languages are often missing. To address this, we propose Lexical-Semantic-Phonetic Align (LSP-Align), a technique to automatically mining cross-lingual entity lexicon from the web. We demonstrate LSP-Align outperforms baselines at extracting cross-lingual entity pairs and mine 164 million entity pairs from 120 different languages aligned with English which we freely release as a resource to the NLP community.

1 Introduction

Named entities are references in natural text to real-world objects such as persons, locations, or organizations that can be denoted with a proper name. Recognizing and handling these named entities in many languages is a difficult, yet crucial, step to language-agnostic text understanding and multilingual natural language processing (NLP).

As such, cross-lingual named entity lexicon can be an invaluable resource for multilingual natural language processing, however, the coverage of many such dictionaries (e.g., Wikipedia titles) is less complete for lower-resource languages. Approaches to automatically generate such dictionaries need to identify mentions from raw text. However, the quality of low-resource taggers can be unreliable making the creation of these dictionaries for low-resource languages a difficult task.

To perform low-resource NER, previous efforts have applied word alignment techniques to project available labels to other languages. Kim et al. (2010) applies heuristic approaches along with alignment correction using an alignment dictionary of entity mentions. Das and Petrov (2011) introduced a novel label propagation technique that creates a tag lexicon for the target language while Wang and Manning (2014) instead projected model expectation rather than labels allowing for the transfer of word boundary uncertainty. Additional work jointly performs word alignment while training bilingual name tagging (Wang et al., 2013); however this method assumes the availability of named entity taggers in both languages. Other methods have leveraged bilingual embeddings for projection (Ni et al., 2017; Xie et al., 2018).

In this work, we propose using named-entity projection to automatically curate a large cross-lingual entity lexicons for many language pairs. As shown Figure 1, we construct this resource by building on the accepted standard of performing NER in a higher-resource language, then projecting the entities onto text in a lower-resource language using word-alignment models.

Our main contribution is that in addition to relying on lexical co-occurrence techniques such as FastAlign (Dyer et al., 2013), we also introduce semantic and phonetic alignment signals to better project named entities. Our final alignment model, LSP-Align, principally combines the Lexical, Semantic, and Phonetic signals to extract higher-quality cross-lingual entity pairs as verified on ground-truth entity pairs.

With LSP-Align, we mine over $164M$ distinct cross-lingual entity pairs spanning 120 languages.
pairs and freely release the dataset in hope it spurs further work in cross-lingual NLP.

2 Preliminaries

We formally define an entity collection as a collection of extracted text spans tied to named entity mentions. These named entity mentions $M = \{ne_i\}_{i=1}^n$, where $ne_i$ is the $i_{th}$ named entity in the mention collection $M$ and $n$ is the size of $M$.

Cross-lingual entity lexicon creation seeks to create two entity collections $M_1$ and $M_2$ in a source and target language respectively. These two collections should be generated such that for each entity mention in $ne_i \in M_1$ in the source language, there is a corresponding named entity $ne_j \in M_2$ in the target language such that $ne_i$ and $ne_j$ refer to the same named entity in their respective language.

3 Mining Cross-lingual Entities

We introduce our approach to automatically extract cross-lingual entity pairs from large mined corpora.

3.1 High-Resource NER

We begin with large collections of comparable bitexts mined from large multilingual web corpora (CCAligned (El-Kishky et al., 2020), WiKiMatrix (Schwenk et al., 2019a), and CCMATRIX (Schwenk et al., 2019b)) due to the wide diversity of language pairs available. We select language pairs of the form English-Target and tag each English sentence with named entity tags (Ramshaw and Marcus, 1999) using a pretrained NER tagger provided in the Stanza NLP toolkit\(^1\) (Qi et al., 2020). This NER model adopts a contextualized string representation-based tagger in (Akbik et al., 2018) and utilizes a forward and backward character-level LSTM language model. At tagging time, the representation at the end of each word position from both language models with word embeddings is fed into a standard Bi-LSTM sequence tagger with a conditional-random-field decoder.

3.2 Entity Projection via Word Alignment

We introduce three approaches for projecting entity mentions using lexical-aligned word alignments. We apply FastAlign (Dyer et al., 2013), a fast loglinear re-parameterization of IBM Model 2 (Brown et al., 1993) and symmetrize alignments using the grow-diagonal-final-and (G DFA) heuristic.

FastAlign performs unsupervised word alignment over the full collection of mined bitexts using an expectation maximization based algorithm. While FastAlign is state-of-the-art in word alignment, due to its reliance on lexical co-occurrences, it may suffer from alignment errors for named entities, which may be low-frequency words.

3.2.2 Semantic Alignment

We leverage multilingual representations (embeddings) from the LASER toolkit (Artetxe and Schwenk, 2019) to align words that are semantically close. We propose a simple greedy word alignment algorithm guided by a distance function between words:

$$sem(w_s, w_t) = 1 - \frac{v_s \cdot v_t}{||v_s|| ||v_t||} \quad (1)$$

**Algorithm 1: Distance Word Alignment**

**Input:** $P = \{(w_s, w_t) \mid \forall w_s \in S_s, w_t \in S_t\}$

**Output:** $P' = \{(w_{x,i}, w_{y,i}) \}_i \subset P$

1. $\text{word-pairs} \leftarrow \{(p, \text{dist}(p)) \text{ for } p \in P\}$
2. $\text{sorted} \leftarrow \text{sort(word-pairs)}$ in ascending order
3. $\text{aligned, } S_x, S_t \leftarrow \emptyset, \emptyset, \emptyset$
4. $\text{free} \leftarrow ||S_x|| - ||S_t||$
5. $\text{for } w_s, w_t \in \text{sorted do}$
6. $\text{if } w_s \notin S_x \land w_t \notin S_t \text{ then}$
7. $\text{aligned} \leftarrow \text{aligned} \cup \{(w_s, w_t)\}$
8. $S_x \leftarrow S_x \cup w_s$
9. $S_t \leftarrow S_t \cup w_t$
10. $\text{else if } free > 0 \land |S_x| < |S_t| \land w_x \in S_x \text{ then}$
11. $\text{aligned} \leftarrow \text{aligned} \cup \{(w_x, w_t)\}$
12. $S_t \leftarrow S_t \cup w_t$
13. $free \leftarrow free - 1$
14. $\text{else if } free > 0 \land |S_t| > |S_x| \land w_t \in S_t \text{ then}$
15. $\text{aligned} \leftarrow \text{aligned} \cup \{(w_s, w_t)\}$
16. $S_w \leftarrow S_w \cup w_w$
17. $free \leftarrow free - 1$
18. $\text{return } \text{aligned}$

where Equation 1 shows that the semantic distances between a source word ($w_s$) and target word ($w_t$) is simply $1$ minus the cosine similarity between $v_s$ and $v_t$, the LASER vector representations of $w_s$ and $w_t$ respectively. As shown in Algorithm 1, we take each source-target sentence pair and perform alignment between their tokens guided by the semantic distances between words. Of course, as source and target sentences, may be of different sizes, tokens in the shorter sentence may be aligned with multiple target tokens. Unlike lexical alignment with FastAlign, our distance-based alignment...
is deterministic and only needs a single pass through the bitexts.

3.2.3 Phonetic Alignment

Recognizing that in many cases, phonetic transliterations are the avenue by which proper names travel between languages (e.g., Alexander in English is pronounced al-Iskandar in Arabic), we propose using phonetic signals to perform alignment and match named entities.

To align words based on their phonetic similarity, we leverage the distances between their transliterations and align words between the source and target that are “close” in this phonetic space. We adopt an unsupervised transliteration system developed by (Chen and Skiena, 2016) to transliterate between source and target languages and utilize Levenshtein distance (aka edit distance) (Wagner and Fischer, 1974) to calculate distances between transliterated words:

\[ \text{phon}(w_s, w_t) = \min \left\{ \frac{\text{LD}(T_{w_s}, w_t)}{\max(|T_{w_s}|, |w_t|)}, \frac{\text{LD}(w_s, T_{w_t})}{\max(|w_s|, |T_{w_t}|)}, \frac{\text{LD}(w_s, w_t)}{\max(|w_s|, |w_t|)} \right\} \]

(2)

where \( \text{LD}(\cdot, \cdot) \) is the Levenshtein distance between two strings and \( T_a \) is the transliteration of word \( a \) into word \( b \)’s language. Equation 2 selects the minimum normalized distance between a source transliteration, target transliteration, and no transliteration to guide Algorithm 1 for a greedy word alignment. Once again, only a single pass over the data is required for alignment.

3.2.4 Estimating Translation Probabilities

Leveraging lexical alignment (i.e., FastAlign) alongside semantic and phonetic alignment results in three potential word alignments for a bitext collection. For alignment method \( k \), we can iterate through the alignments and compute the counts of source-to-target \((s, t)\) word pairings; we denote this count \( \text{cnt}(s, t) \). We can estimate the maximum likelihood translation probability from \( s \) to \( t \) given by alignment method \( k \) as follows:

\[ \theta_{k, s, t} = \frac{\text{cnt}(s, t)}{\sum_{t'} \text{cnt}(s, t')} \]

(3)

Using Equation 3, we can compute the translation probabilities for lexical, semantic, and phonetic alignments which we use in our LSP-Align model.

3.3 LSP Named-entity Projection

We describe LSP-Align, which combines the three alignment signals for better entity-pair mining.

![Diagram of LSP-Align](image)

**Algorithm 2: LSP-Align Generative Model**

Input: \( S = \{s_1 \ldots s_m\} \) // source sentence

Output: \( T = \{t_1 \ldots t_n\} \) // translated sentence

1. let \( \theta_k; k \in \{1, 2, 3\} \) be the translation distributions

2. draw length \( n \) for translation \( T \) using \( |S| = m \)

3. for each \( j \in \{1 \ldots n\} \) do

4. draw \( a_j \in \{1, \ldots, m\} \) \sim \( U(0, m) \)

5. draw \( k_j \sim U(1, 3) \)

6. draw \( t_j \sim \theta_{k_j, a_j, t_j} \)

7. end

8. return \( T \)

As described in Algorithm 2, the generative process takes in a source sentence \( S \) and translates this sentence into the target sentence by drawing an alignment variable and translation mechanism (lexical, semantic, or phonetic) for each position in the target sentence and drawing a translated word from the corresponding translation distribution.

The graphical model for LSP-Align depicted in Figure 2, is similar to IBM-1 (Brown et al., 1993). The main difference is that, in addition to latent alignment variables \( A \), we introduce latent translation mechanisms \( K \). The translation distributions \( \theta_{K, s} \) is chosen based on the latent alignment and mechanism variables. As we previously demonstrate in Equation 3, we can leverage the alignments for each alignment signal to estimate \( \theta_{K, s} \) for each translation distribution. Using these estimated distributions in our model, we can infer the alignment variables as follows:

\[ P(a_j = i|S, T, \theta) = \sum_{k_j=1}^{3} P(a_j = i|S, T, k_j, \theta) \cdot P(k_j) \]

\[ = \sum_{k_j=1}^{3} \theta_{k_j, s, t_j} \cdot \frac{1}{3} \]

(4)

where we assign the most probable alignment variable to each target word after marginalizing over
Table 1: Fuzzy-F1 scores of mined cross-lingual entity pairs evaluated against gold-standard pairs.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Language</th>
<th>Num Bitexts</th>
<th>Distinct Ents</th>
<th>Lexical</th>
<th>Semantic</th>
<th>Phonetic</th>
<th>LSP-Align</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Russian</td>
<td>3.2M</td>
<td>40.4K</td>
<td>0.84</td>
<td>0.81</td>
<td>0.83</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>5.2M</td>
<td>28.4K</td>
<td></td>
<td>0.85</td>
<td>0.78</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Turkish</td>
<td>2.5M</td>
<td>27.4K</td>
<td>0.88</td>
<td>0.89</td>
<td>0.87</td>
<td>0.90</td>
</tr>
<tr>
<td>Mid</td>
<td>Arabic</td>
<td>4.9M</td>
<td>26.4K</td>
<td>0.88</td>
<td>0.80</td>
<td>0.81</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>Hindi</td>
<td>1.2M</td>
<td>7.60K</td>
<td>0.89</td>
<td>0.73</td>
<td>0.87</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>Romanian</td>
<td>2.1M</td>
<td>26.2K</td>
<td>0.93</td>
<td>0.94</td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td>Low</td>
<td>Estonian</td>
<td>1.3M</td>
<td>15.2K</td>
<td>0.87</td>
<td>0.89</td>
<td>0.87</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>Armenian</td>
<td>52K</td>
<td>2.30K</td>
<td>0.78</td>
<td>0.44</td>
<td>0.83</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>Tamil</td>
<td>45K</td>
<td>2.50K</td>
<td>0.67</td>
<td>0.50</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>Avg</td>
<td></td>
<td></td>
<td></td>
<td>0.84</td>
<td>0.75</td>
<td>0.83</td>
<td>0.86</td>
</tr>
</tbody>
</table>

seen in Table 1, while lexical alignment outperforms semantic alignment, it displays similar performance to phonetic with phonetic performing better on low-resource languages and lexical performing better on high-resource. However, LSP-Align outperforms or matches lexical alignment consistently showing that using all signals yields superior NE projection.

Figure 3, separates the evaluated entities by frequency in the web-data bitexts (low=0-3, mid=4-10, high=11+), and shows LSP-Align outperforming FastAlign when the entity is infrequent in the corpus. However, as entity frequency follows a long-tailed distribution, most entity mentions are infrequent.

5 Conclusion

We propose a technique that combines lexical alignment, semantic alignment, and phonetic alignment into a unified alignment model. We demonstrate this unified model better extracts cross-lingual entity pairs over any single alignment. Leveraging this model, we automatically curate a large, cross-lingual entity resource covering 100 languages paired with English which we freely release to the community.
References


Chen-Tse Tsai and Dan Roth. 2018. Learning better name translation for cross-lingual wikification. In *AAAI*.


