XLEnt: Mining a Large Cross-lingual Entity Dataset with Lexical-Semantic-Phonetic Word Alignment

Ahmed El-Kishky[1] **Adi Renduchintala**[1] **James Cross**[1] **Francisco Guzmán**[1] **Philipp Koehn**[2]
[1]Facebook AI [2]Johns Hopkins University
{ahelk,adirendu,jcross,fguzman}@fb.com, phi@jhu.edu

Abstract

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).

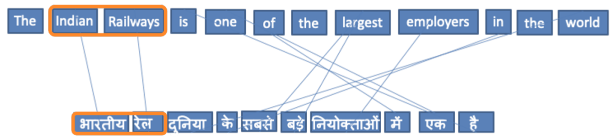


Figure 1: Identify entity pairs by projecting English entities onto lower-resource languages via word-alignment.

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* language 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 entities and LSP-Align which combines all three.

3.2.1 Lexical Alignment

To perform word alignment using lexical-cooccurences, 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 (GDFA) 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-occurences, 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:

$$sem(w_s, w_t) = 1 - \frac{\mathbf{v}_s \cdot \mathbf{v}_t}{||\mathbf{v}_s||\ ||\mathbf{v}_t||} \tag{1}$$

Algorithm 1: Distance Word Alignment

Input: $P = \{(w_s, w_t) \mid w_s \in S_s, w_t \in S_t\}$
Output: $P' = \{(w_{s,i}, w_{t,i}), \ldots\} \subset P$

$word{-}pairs \leftarrow \{(p, dist(p))$ for $p \in P\}$
$sorted \leftarrow sort(word{-}pairs)$ in ascending order
aligned, S_s, $S_t \leftarrow \varnothing, \varnothing, \varnothing$
$free \leftarrow ||S_s| - |S_t||$
for $w_s, w_t \in sorted$ **do**
  if $w_s \notin S_s \wedge w_t \notin S_t$ **then**
    $aligned \leftarrow aligned \cup \{(w_s, w_t)\}$
    $S_s \leftarrow S_s \cup w_s$
    $S_t \leftarrow S_t \cup w_t$
  else if $free > 0 \wedge |S_s| < |S_t| \wedge w_s \in S_s$ **then**
    $aligned \leftarrow aligned \cup \{(w_s, w_t)\}$
    $S_t \leftarrow S_t \cup w_t$
    $free \leftarrow free{-}1$
  else if $free > 0 \wedge |S_s| > |S_t| \wedge w_t \in S_t$ **then**
    $aligned \leftarrow aligned \cup \{(w_s, w_t)\}$
    $S_w \leftarrow S_w \cup w_w$
    $free \leftarrow free{-}1$
end
return 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

[1] https://stanfordnlp.github.io/stanza/

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:

$$phon(w_s, w_t) = \min \left\{ \begin{array}{l} \mathrm{LD}(T_{w_s}, w_t))/\max(|T_{w_s}|, |w_t|) \\ \mathrm{LD}(w_s, T_{w_t}))/\max(|w_s|, |T_{w_t}|) \\ \mathrm{LD}(w_s, w_t)/\max(|w_s|, |w_t|) \end{array} \right\} \tag{2}$$

where $LD(\cdot, \cdot)$ 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 phoentic 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 $cnt(s, t)$. We can estimate the maximum likelihood translation probability from s to t given by alignment method k as follows:

$$\theta_{k,s,t} = \frac{cnt(s,t)}{\sum_{t'} cnt(s,t')} \tag{3}$$

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

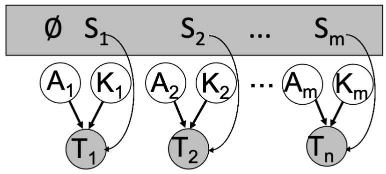


Figure 2: S and T are source/target sentences; target words are drawn from a distribution determined by (1) alignment, (2) source word, and (3) translation method

3.3 LSP Named-entity Projection

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

Algorithm 2: LSP-Align Generative Model

Input: $S = \{s_1 \dots s_m\}$ `// source sentence`
Output: $T = \{t_1 \dots t_n\}$ `// translated sentence`

let θ_k: $k \in \{1, 2, 3\}$ be the translation distributions
    `// 1=lexical, 2=semantic, 3=phonetic`
draw length n for translation T using $|S| = m$
for *each* $j \in 1 \dots n$ **do**
    draw $a_j \in \{0, 1, \dots, m\} \sim \mathcal{U}(0, m)$
    draw $k_j \sim \mathcal{U}(1, 3)$
    draw $t_j \sim \theta_{k_j, s_{a_j}, t_j}$
end
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:

$$\begin{aligned} P(a_j = i|S, T, \theta) &= \sum_{k_j=1}^{3} P(a_j{=}i|S, T, k_j, \theta) \cdot P(k_j) \\ &= \sum_{k_j=1}^{3} \theta_{k_j, s_i, t_j} \cdot \frac{1}{3} \end{aligned} \tag{4}$$

where we assign the most probable alignment variable to each target word after marginalizing over

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

Table 1: Fuzzy-F1 scores of mined cross-lingual entity pairs evaluated against gold-standard pairs.

the latent translation mechanisms (lexical, semantic, phonetic) which have equal probability.

4 Experiments & Results

We evaluate the quality of our mined entity pairs.

4.1 Experimental Setup

Datasets We create a gold standard evaluation lexicon following (Pan et al., 2017) by compiling eight named parallel entity corpora[2] creating a gold standard cross-lingual entity lexicon. We select nine languages from a diverse set of resource availability, language families, and scripts for evaluation.

Evaluation Protocol We evaluated the performance of the methods using the commonly used fuzzy-f1 score (Tsai and Roth, 2018) which is defined as the harmonic mean of the fuzzy precision and fuzzy recall scores. This metric is based on the longest common subsequence between a gold and mined entity, and has been used for several years in the NEWS transliteration workshops (Li et al., 2009; Banchs et al., 2015). The fuzzy precision and recall between a predicted string p and the correct string t is computed as follows:

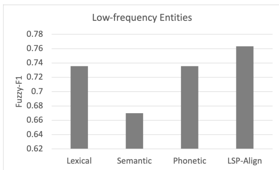
$$\text{fuzzy-precision}(p,t) = {}^{|\mathrm{LCS}(p,t)|}/_{|p|},$$
$$\text{fuzzy-recall}(p,t) = {}^{|\mathrm{LCS}(p,t)|}/_{|t|},$$

where $LCS(\cdot,\cdot)$ is the *longest common subsequence* between two strings.

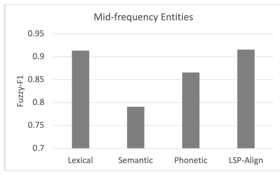
4.2 Cross-lingual Entity Extraction

We take mined entity pairs from each projection technique and compute Fuzzy-F1 of the mined entities using the gold-standard as a reference. As

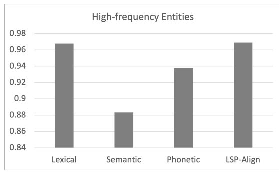
[2]Chinese-English Wikinames, Geonames, JRC names, LORELEI LRLP, NEWS 2015 task, Wikipedia names, Wikipedia places, and Wikipedia titles



(a) low-frequency



(b) mid-frequency



(c) high-frequency

Figure 3: Fuzzy-F1 by entity-frequency

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

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Rafael E Banchs, Min Zhang, Xiangyu Duan, Haizhou Li, and A Kumaran. 2015. Report of news 2015 machine transliteration shared task. In *Proceedings of the Fifth Named Entity Workshop*, pages 10–23.

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

Yanqing Chen and Steven Skiena. 2016. False-friend detection and entity matching via unsupervised transliteration. *arXiv preprint arXiv:1611.06722*.

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.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzman, and Philipp Koehn. 2020. Ccaligned: A massive collection of cross-lingual web-document pairs. In *EMNLP*.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2010. A cross-lingual annotation projection approach for relation detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 564–571.

Haizhou Li, A Kumaran, Vladimir Pervouchine, and Min Zhang. 2009. Report of news 2009 machine transliteration shared task. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 1–18.

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 (Volume 1: Long Papers)*, pages 1470–1480.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

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.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019a. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019b. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

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

Robert A Wagner and Michael J Fischer. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173.

Mengqiu Wang, Wanxiang Che, and Christopher D Manning. 2013. Joint word alignment and bilingual named entity recognition using dual decomposition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1082.

Mengqiu Wang and Christopher D Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association for Computational Linguistics*, 2:55–66.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime G 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.