Contextual Label Projection for Cross-Lingual Structured Prediction

Tanmay Parekh[†] **I-Hung Hsu**[‡] **Kuan-Hao Huang**[§]
Kai-Wei Chang[†] **Nanyun Peng**[†]

[†]Computer Science Department, University of California, Los Angeles
[‡]Information Science Institute, University of Southern California
[§]Department of Computer Science, University of Illinois Urbana-Champaign
{tparekh, kwchang, violetpeng}@cs.ucla.edu
{ihunghsu}@isi.edu, {khhuang}@illinois.edu

Abstract

Label projection, which involves obtaining translated labels and texts jointly, is essential for leveraging machine translation to facilitate cross-lingual transfer in structured prediction tasks. Prior research exploring label projection often compromise translation accuracy by favoring simplified label translation or relying solely on word-level alignments. In this paper, we introduce a novel label projection approach, CLaP, which translates text to the target language and performs *contextual translation* on the labels using the translated text as the context, ensuring better accuracy for the translated labels. We leverage instruction-tuned language models with multilingual capabilities as our contextual translator, imposing the constraint of the presence of translated labels in the translated text via instructions. We benchmark CLaP with other label projection techniques on zero-shot cross-lingual transfer across 39 languages on two representative structured prediction tasks — event argument extraction (EAE) and named entity recognition (NER), showing over 2.4 F1 improvement for EAE and 1.4 F1 improvement for NER. We further explore the applicability of CLaP on ten extremely low-resource languages to showcase its potential for cross-lingual structured prediction.

1 Introduction

Cross-lingual transfer for structured prediction tasks such as named entity recognition, relation extraction, and event extraction, has gained considerable attention recently (Huang et al., 2022; Cao et al., 2023; Tedeschi and Navigli, 2022; Cabot et al., 2023; Fincke et al., 2022; Jenkins et al., 2023; Ahmad et al., 2021b). It generalizes models trained in source languages to applications on other target languages (Chen and Ritter, 2021; Subburathinam et al., 2019; Pouran Ben Veyseh et al., 2022).

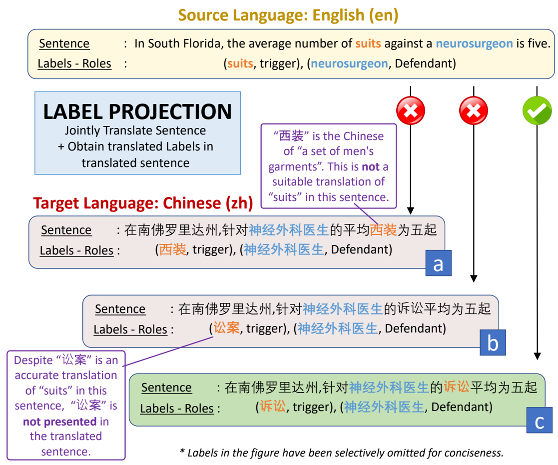


Figure 1: Illustration of the task of *label projection* from English to Chinese. Label projection converts sentences from a source to a target language while translating the associated labels jointly. Failures in this process occur when (a) labels are either inaccurately translated or (b) missing in the translated sentence in the target language.

One effective and simple way to improve cross-lingual transfer performance is translate-train, which leverages machine translation techniques to generate pseudo-training data in the target languages by translating source language training data (Xue et al., 2021; Ruder et al., 2021; Yu et al., 2023). However, adopting translate-train to structured prediction necessitates a *label projection* step, which involves jointly translating input sentences and labels (Chen et al., 2023). Label projection requires not only *accurate translation* of the labels but also *maintaining the association* between the translated texts and labels. As illustrated in Figure 1, while "*suits*" can have multiple valid translations, only "诉讼" is present in the translated sentence and a proper translation at the same time.

Prior works have dealt with label projection through two primary frameworks. The first one, illustrated in Figure 2(a), performs machine trans-

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5738–5757

lation on modified source sentences that incorporate label annotations using special markers (Chen et al., 2023; Hennig et al., 2023). Translated labels can be extracted if special markers are retained in the translations. In this approach, the quality of the translation is *inherently compromised* due to the inclusion of special markers (Chen et al., 2023). The other framework uses word similarity to procure word alignments between the source and translated sentences. Label translations are further constructed by combining mapped tokens in the translated sentence (Stengel-Eskin et al., 2019; Akbik et al., 2015; Aminian et al., 2019), as shown in Figure 2(b). However, it is hard for this framework to ensure *accurate* label translation by merely using word alignments, as we will show in § 4.4.

In this work, we introduce CLaP (**C**ontextual **La**bel **P**rojection), which obtains projected label annotations by conducting contextual machine translation for the labels. We first acquire the translation of the input sentence by any plug-and-play machine translator. Then, inspired by the idea of contextual machine translation (Wong et al., 2020; Voita et al., 2018), we use the translated input text as context to perform label translation, as shown in Figure 2(c). Exploiting contextual machine translation strongly enhances the *accuracy* of the translated labels while preserving their *association* to the translated sentence. Furthermore, translating the input sentence in an unmodified manner better leverages machine translators and assures the quality of translated sentence. To implement contextual machine translation, we utilize a small instruction-tuned language model with multilingual capabilities, Llama-2-13B (Touvron et al., 2023). [1] We encode the translated input sentence and the constraint for the presence of labels in the form of instruction prompts and ask the language model to perform the label translation task.

Extensive experiments conducted on two representative tasks, event argument extraction (EAE) and named entity recognition (NER), reveal the following insights:

- Compared to existing label projection methods, CLaP performs the best on intrinsic evaluation by achieving the best label translation accuracy (§ 4.4). Through extrinsic evaluation on downstream tasks, CLaP yields an average improvement of 2.4 and 1.4 F1 scores over the best baseline across 39 languages for EAE on ACE and NER on WikiANN datasets respectively (§ 4.5).
- In comparison to directly prompting LLMs for the downstream task, we show that CLaP's LLM usage for contextual machine translation provides significantly larger gains (§ 4.5).
- Focusing on low-resource languages, CLaP demonstrates strong applicability by generalizing to ten extremely low-resourced African and American languages (§ 6). Using larger LLMs for CLaP yields further improvements for low-resource languages, underlining CLaP's future potential to improve continually (§ 5.2).

Our code can be found at `https://github.com/PlusLabNLP/CLaP`.

2 Background

2.1 Structure Prediction Tasks

Given an input sentence $\mathbf{x}$, structure prediction models aim to predict structured output $\mathbf{y} = [\mathbf{x}[i_1 : j_1], \mathbf{x}[i_2 : j_2], \ldots, \mathbf{x}[i_n : j_n]]$ (where $\mathbf{x}[i_1 : j_1]$ is an input sentence span from token i_1 to j_1) corresponding to a set of roles $\mathbf{r} = [r_1, r_2, \ldots, r_n]$ (where $r_i \in \mathcal{R}$, a pre-defined set of roles). This vastly differs from standard classification-based tasks wherein the output prediction y is a singular value from a fixed set of classes independent of the input sentence $\mathbf{x}$.

2.2 Zero-shot Cross-Lingual Transfer

Zero-shot cross-lingual transfer (Hu et al., 2020; Ahmad et al., 2019; Huang et al., 2021) aims to train a downstream model for the target language l_{tgt} using supervised data $\mathcal{D}_{src}$ from a source language l_{src} without using any data in the target language (i.e. $\mathcal{D}_{tgt} = \phi$). The paradigm has effectively advanced language technologies for under-resourced languages.

2.3 Translate-Train

Translate-train (Hu et al., 2020; Ruder et al., 2021) is a popular and powerful zero-shot cross-lingual transfer technique that leverages machine translators $\mathcal{T}$ to boost downstream model performance. Specifically, in translate-train, $\mathcal{D}_{src}$ is translated into the target language as pseudo training data $\mathcal{D}_{src}^{tgt}$ and the downstream model is trained using a combination of $\{\mathcal{D}_{src}, \mathcal{D}_{src}^{tgt}\}$.

Utilizing translate-train for structured prediction tasks requires *Label Projection*, which includes two sets of translations: (1) Sentence translation ($\mathbf{x}^{src} \rightarrow \mathbf{x}^{tgt}$), where we use $\rightarrow$ to denote that $\mathbf{x}^{tgt}$

[1] We also explore using GPT-3.5-Turbo in § 5.2.

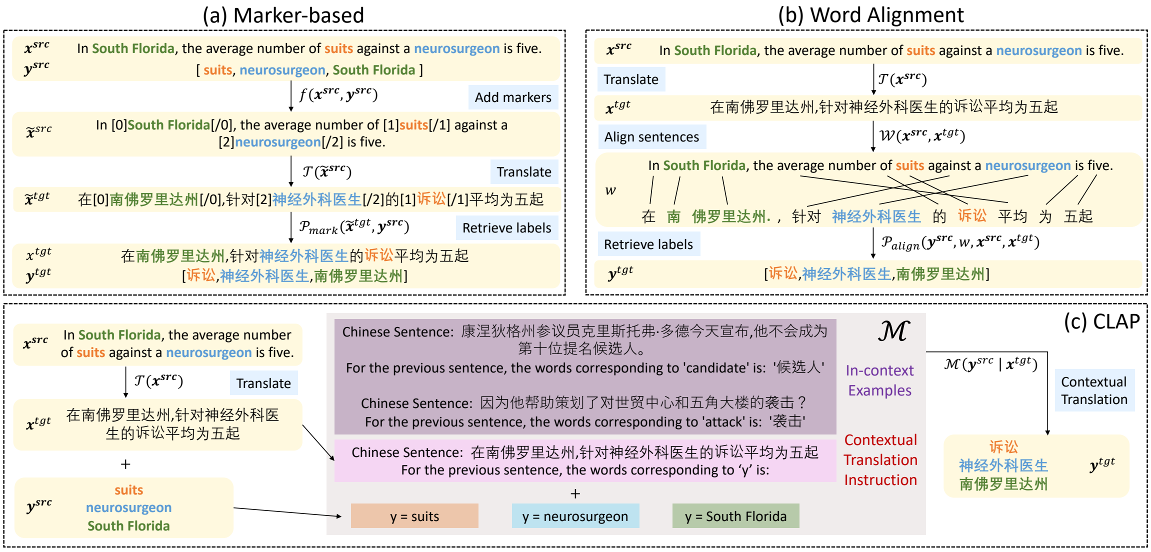


Figure 2: Illustration of the various techniques to conduct label projection: (a) **Marker-based** methods use markers to transform the sentence and translate the transformed sentence with label markers jointly, (b) **Word Alignment** methods use external word alignment tools to locate the translated labels in the translated sentence, and (c) **CLaP** (ours) performs contextual translation on labels using $\mathcal{M}$ (Here, we demonstrate the use of an instruction-tuned language model as $\mathcal{M}$ to identify translated labels within a translated sentence.).

is the transformation of $\mathbf{x}^{src}$; and (2) Label translation ($\mathbf{y}^{src} \rightarrow \mathbf{y}^{tgt}$), such that the translated label $\mathbf{y}^{tgt}$ is appropriately *associated with* $\mathbf{x}^{tgt}$. This demand makes translate-train for structure prediction tasks more complex than that for classification tasks, as the latter only requires sentence translation (since y is independent of $\mathbf{x}$). [2]

Translate-Test Besides translate-train, translate-test is another commonly used technique in zero-shot cross-lingual transfer. During inference, models trained on $\mathcal{D}_{src}$ are used to predict on translated test sentences ($\mathbf{x}^{tgt} \rightarrow \mathbf{x}^{src}$), and the predictions on $\mathbf{x}^{src}$ are later mapped back to $\mathbf{x}^{tgt}$. We mainly focus on translate-train in this work but discuss CLaP's effectiveness for translate-test in § 5.5.

2.4 Label Projection

We hereby technically define the problem of *label projection* (Akbik et al., 2015; Chen et al., 2023):

$$
\begin{array}{lll}
 & \mathbf{x}^{src} \rightarrow \mathbf{x}^{tgt} & \\
\& & y_m^{src} \rightarrow y_m^{tgt} & \forall y_m^{src} \in \mathbf{y}^{src} \\
s.t. & y_m^{tgt} \in \mathbf{x}^{tgt} & \forall y_m^{tgt} \in \mathbf{y}^{tgt}.
\end{array}
$$

This problem requires optimizing two properties of accuracy and faithfulness in the translations:

- **Accuracy** ensures that $[\mathbf{x}^{tgt}, y_1^{tgt}, \ldots, y_n^{tgt}]$ are accurate translations of $[\mathbf{x}^{src}, y_1^{src}, \ldots, y_n^{src}]$.
- **Faithfulness** ensures that each y_m^{tgt} is associated with $\mathbf{x}^{tgt}$ (the constraint of $y_m^{tgt} \in \mathbf{x}^{tgt}$).

How to do this joint translation is non-trivial as standard translation models $\mathcal{T}$ cannot simply impose the additional faithfulness constraint, as shown in the failure cases in Figure 1(b). This demonstrates the challenge of label projection.

3 Methodology

In this section, we first formally define the previous attempts at label projection and later introduce CLaP, which provides a new perspective of using contextual machine translation for label projection.

3.1 Baseline Methods

The primary frameworks used in prior works include Marker-based and Word-alignment methods.

Marker-based methods (Lewis et al., 2020; Hu et al., 2020; Chen et al., 2023) solve the label projection by first marking labels to the input sentence $\mathbf{x}^{src}$, forming $\tilde{\mathbf{x}}^{src}$, and then use the translation model to obtain the potential translation of input sentence and labels jointly. For example, in Figure 2(a), "South Florida" is delineated by markers [0] and [\0]. Assuming the preservation of markers after translation of $\tilde{\mathbf{x}}^{src}$, a post-processing step, $\mathcal{P}_{mark}$, is performed to retain the translated labels $\mathbf{y}^{tgt}$ and translated sentence $\mathbf{x}^{tgt}$. Putting every step

[2] For certain structure prediction tasks like relation classification (Ahmad et al., 2021b; Hsu et al., 2021) (determining the relationship between two entities in $\mathbf{x}$), even if the output y is scalar, translate-train necessitates label projection step due to the required projection of the two given entities into the translated sentence.

together, we have

$$\tilde{\mathbf{x}}^{src} = f(\mathbf{x}^{src}, \mathbf{y}^{src}), \quad \tilde{\mathbf{x}}^{tgt} = \mathcal{T}(\tilde{\mathbf{x}}^{src})$$
$$\mathbf{x}^{tgt}, \mathbf{y}^{tgt} = \mathcal{P}_{mark}(\tilde{\mathbf{x}}^{tgt}, \mathbf{y}^{src}),$$

where f denotes the marker addition step and $\tilde{\mathbf{x}}^{tgt}$ is the translation of $\tilde{\mathbf{x}}^{src}$ using translator $\mathcal{T}$.

Despite their simplicity, these methods suffer from poor translation quality and reduced robustness to different translation models owing to their input sentence transformations and strong assumptions about the retention of markers in $\tilde{\mathbf{x}}^{tgt}$.

Word Alignment approaches (Akbik et al., 2015; Yarmohammadi et al., 2021) first translate the input sentence and acquire word alignments (Dyer et al., 2013; Dou and Neubig, 2021) between the translation pairs. Each translated label y_m^{tgt} is then procured by merging the aligned words of y_m^{src} in the translated sentence using the word mappings w. For example, in Figure 2(b), the translated label for "South Florida" is obtained by merging two aligned words, which is done by a heuristic post-processing algorithm $\mathcal{P}_{align}$. Formally, we have

$$\mathbf{x}^{tgt} = \mathcal{T}(\mathbf{x}^{src}), \; w = \mathcal{W}(\mathbf{x}^{src}, \mathbf{x}^{tgt})$$
$$y_m^{tgt} = \mathcal{P}_{align}(y_m^{src}, w, \mathbf{x}^{src}, \mathbf{x}^{tgt}) \qquad \forall y_m^{src} \in \mathbf{y}^{src}$$

Although these approaches deliver high-quality sentence translations, the accuracy of their translated labels is compromised. This is because the translated labels are reconstructed from word-level translations, lacking joint consideration of the entire span (Akbik et al., 2015; Chen et al., 2023).

3.2 CLaP

We tackle the task of label projection through a new perspective — performing actual translation on labels instead of recovering them from translated text $\mathbf{x}^{tgt}$. This better ensures the accuracy of the translated labels $\mathbf{y}^{tgt}$. To accomplish this, we leverage the idea of *contextual machine translation* on the label translation with $\mathbf{x}^{tgt}$ as context.

Contextual machine translation, which aims to perform phrase-level translations conditional on the context of the translated sentence, is tangentially explored for applications like anaphora resolution (Voita et al., 2018) and pronoun translations (Wong et al., 2020). The main goal of this task is to maintain the consistency of phrasal translations in the given context. In our work, we develop a novel model CLaP to extend the idea of contextual translation to the application of label projection.

As illustrated in Figure 2(c), CLaP first utilizes machine translation model $\mathcal{T}$ to translate input sentence $\mathbf{x}^{src}$ to $\mathbf{x}^{tgt}$. Treating $\mathbf{x}^{tgt}$ as the context, the contextual translation model $\mathcal{M}$ translates the labels $\mathbf{y}^{src}$ to $\mathbf{y}^{tgt}$. Contextual translation implicitly imposes the *faithfulness constraint* which requires $y_m^{tgt} \in \mathbf{x}^{tgt} \quad , \forall y_m^{tgt} \in \mathbf{y}^{tgt}$, hence, slackly satisfying the requirement of label projection. These two steps can be formally described as:

$$\mathbf{x}^{tgt} = \mathcal{T}(\mathbf{x}^{src})$$
$$y_m^{tgt} = \mathcal{M}(y_m^{src} | \mathbf{x}^{tgt}) \qquad \forall y_m^{src} \in \mathbf{y}^{src}$$

where y_m^{tgt} is generated from $\mathcal{M}(y_m^{src}|\mathbf{x}^{tgt})$, drawing the difference from the previous works.

Compared to word alignment approaches using simple word-similarity aligners $\mathcal{W}$, we use models with translation capabilities $\mathcal{M}$, to improve the accuracy of translated labels. Furthermore, the independence of $\mathcal{T}$ and $\mathcal{M}$ for translating $\mathbf{x}^{src}$ and $\mathbf{y}^{src}$ respectively assures that CLaP has better translation quality for $\mathbf{x}^{tgt}$ and is more robust than the marker-based baselines. We empirically back these intuitions in § 4.4.

3.3 Implementing CLaP

To implement our concept, we first configure $\mathcal{T}$ to be a modular component that can be replaced by any third-party translation model. For $\mathcal{M}$, we use an instruction-tuned language model (LM) with multilingual capabilities (Wei et al., 2021; Scao et al., 2022). Instruction-tuned LMs can accept conditional information in their natural language prompt. Specifically, we encode the translated target sentence $\mathbf{x}^{tgt}$ as well as the faithfulness constraint $y_m^{tgt} \in \mathbf{x}^{tgt}$ implicitly in the form of natural language instructions (highlighted as "Contextual Translation Instruction" in Figure 2(c)). Following Brown et al. (2020), we also provide n randomly chosen in-context examples (highlighted as "In-context examples" in Figure 2(c)) to improve the instruction-understanding capability of the model.[3] Instruction-tuned LMs sacrifice some translation ability compared to supervised machine translation models (Zhu et al., 2023), however, they provide better control of contextual constraints.

After obtaining label translations, we employ simple string-matching algorithms to get the exact span index of y_m^{tgt} in $\mathbf{x}^{tgt}$. Though this may not be

[3]The in-context examples are generated using Google translation and initial prediction from instruction-tuned LMs. The label predictions are further verified by back-translation.

	ACE	WikiANN
# Train Instances	4,202	20,000
# Dev Instances	450	10,000
# Avg. Test Instances	194	6,469
# Test Languages	2	39

Table 1: High-level data statistics for ACE and WikiANN datasets for EAE and NER tasks respectively. # = 'number of' and Avg. = average.

the optimal solution when duplicated strings exist in $\mathbf{x}^{tgt}$, it works well in practice as stated in prior word-alignment methods (Dou and Neubig, 2021).

4 Experiments and Results

This section outlines our experimental settings, which includes the datasets, baselines, and implementation details. Subsequently, we provide an in-depth analysis of CLaP through both intrinsic and extrinsic evaluations.

4.1 Task and Dataset

We choose two structure prediction tasks, event argument extraction (EAE) (Sundheim, 1992; Hsu et al., 2023a) and named entity recognition (NER) (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) for evaluating our label projection method. EAE requires the extraction of text segments serving as arguments corresponding to an event and mapping them to their corresponding argument roles. NER aims to identify and categorize named entities from the input sentence. For EAE, we use multilingual ACE dataset (Doddington et al., 2004) and follow the pre-processing by Huang et al. (2022) to retain 33 event types and 22 argument roles. For NER, we consider the WikiANN (Pan et al., 2017; Rahimi et al., 2019) with pre-processing by Hu et al. (2020). We list the basic statistics for these datasets in Table 1 and more details in § A. For experiment, we consider the zero-shot cross-lingual transfer using English (*en*) as the source language.

4.2 Baselines

We select two label projection models as baselines, each representing the two baseline frameworks we covered in Section 3.1, respectively: (1) **EasyProject** (Chen et al., 2023), a recent marker-based label-projection method, utilizes numbered square braces (e.g. [0] and [/0]) to mark the labels in the input sentence. (2) **Awesome-Align** (Dou and Neubig, 2021), a neural bilingual word alignment model, uses multilingual language models to find word similarities to derive word alignments, which are later used for label projection.

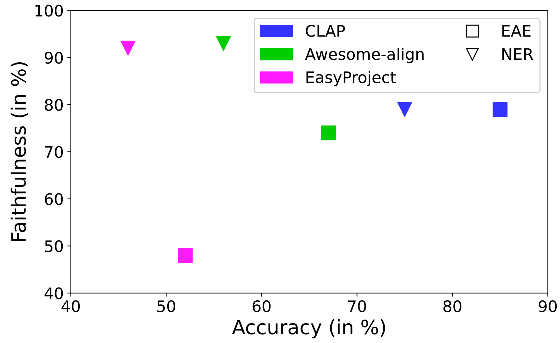


Figure 3: Reporting faithfulness and accuracy (in %) for the different label projection models on EAE and NER. The closer the model is to the top-right, the better it is.

	ar	zh	Avg
LLM-Infer	16.9	24.0	20.5
Zero-shot*	40.3	51.9	46.1
Awesome-align	48.6	54.5	51.6
EasyProject	38.5	56.3	47.4
CLaP (ours)	**49.3**	**58.6**	**54.0**

Table 2: Extrinsic evaluation of the different label projection techniques regarding downstream model performance using translate-train and the LLM-Infer baseline for EAE. Avg = Average. * indicates the reproduced results of X-Gear (Huang et al., 2022).

4.3 Implementation Details

For the translation model $\mathcal{T}$, we experiment with the Google Machine Translation (GMT) (Wu et al., 2016).[4] For CLaP, we use the text-completion version of Llama-2 (Touvron et al., 2023) with 13B parameters as $\mathcal{M}$. We use $n = 2$ in-context examples for CLaP prompts. For Awesome-align, we use the unsupervised version of their model utilizing multilingual BERT (Devlin et al., 2019) as it provides better results (Chen et al., 2023). [5] Additional details are provided in Appendix C.

4.4 Intrinsic Evaluation

We first evaluate CLaP by directly evaluating the label projection quality, mainly focusing on evaluating the **accuracy** and **faithfulness** of the translated

[4] `https://cloud.google.com/translate`. We use the free scraping tool to reduce the translation costs.

[5] We utilize the non fine-tuned version of EasyProject since we experiment using GMT. The original work also explores finetuning the machine translation model but it requires open-source access for finetuning.

Lang	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv	ka
LLM-Infer	50.9	24.8	66.9	12.0	44.2	42.2	59.5	41.6	36.7	19.5	46.7	53.5	15.6	18.9	20.6	30.3	56.0	35.7	28.7	21.7
Zero-shot	77.4	48.1	82.8	77.0	78.8	80.6	74.5	78.7	61.4	69.2	79.3	79.4	57.3	70.6	80.8	53.1	79.4	19.1	58.5	72.3
Awesome-align	**77.9**	46.0	**81.0**	**81.2**	**78.8**	71.7	65.3	**78.0**	66.8	46.4	**77.4**	78.2	55.3	**73.9**	**77.4**	52.8	79.3	20.3	56.3	70.4
EasyProject	76.1	34.4	**81.0**	78.6	**78.8**	69.3	70.5	73.9	54.8	49.1	**77.8**	78.8	**61.1**	73.0	75.6	51.0	79.0	41.3	62.4	66.4
CLaP	74.4	**48.7**	**81.0**	78.1	78.4	**75.9**	**74.7**	77.4	**68.8**	**59.0**	75.9	**79.4**	58.4	73.1	72.4	**56.1**	**80.1**	**45.3**	**64.8**	**70.5**
	kk	**ko**	**ml**	**mr**	**ms**	**my**	**nl**	**pt**	**ru**	**sw**	**ta**	**te**	**th**	**tl**	**tr**	**ur**	**vi**	**yo**	**zh**	**Avg**
LLM-Infer	20.9	18.5	11.1	16.5	46.5	10.1	64.3	46.4	22.7	33.4	12.8	9.2	19.8	46.1	31.0	11.6	37.3	28.6	41.0	32.1
Zero-shot	51.9	57.5	66.4	65.3	53.4	65.8	83.0	80.0	74.2	68.4	60.3	62.1	0.4	74.5	65.6	62.2	75.0	34.1	24.6	64.2
Awesome-align	**47.7**	57.7	**63.4**	**62.4**	70.7	54.1	**83.0**	75.8	64.8	70.1	**62.4**	**55.4**	2.4	**80.9**	62.8	53.7	66.4	**61.5**	45.4	63.5
EasyProject	31.7	48.2	56.5	59.8	71.7	60.3	81.9	**79.6**	66.3	**71.5**	53.2	54.2	11.4	78.2	**66.8**	**63.8**	65.6	68.8	42.0	63.2
CLaP	42.8	**60.1**	60.3	61.4	**73.5**	**61.5**	82.2	78.2	**68.3**	70.6	59.6	53.1	**13.2**	74.6	62.9	32.9	**75.8**	59.6	**49.7**	**64.9**

Table 3: Extrinsic evaluation of the different label projection techniques in terms of downstream model performance using translate-train and the LLM-Infer baseline for NER. Avg = Average.

labels, with the definition stated in § 2.4.

We employ native speakers to assess the accuracy of label translations. The evaluation is carried out using a ranking framework, in which the label translations from each model are ranked, including the option for ties. The final accuracy score represents the average percentage at which the model outperformed all other competitors. We conduct this evaluation on 50 data samples for Chinese, Arabic, Hindi, and Spanish, respectively.

Faithfulness measures the fulfillment of the label projection constraint. It is measured as a percentage of projected data points when all the translated labels are present in the translated input sentence $\left(y_m^{tgt} \in \mathbf{x}^{tgt} \quad, \forall y_m^{tgt} \in \mathbf{y}^{tgt}\right)$. The statistics use the complete test set on ACE and WikiANN.

Results: The accuracy and faithfulness of the models are plotted together in Figure 3. An ideal model should optimize both these metrics and thus, the closer the models are to the top-right, the better they are deemed. Overall, this figure shows how CLaP performs the best intrinsically as it is the closest to the top-right for both the tasks. For EAE, CLaP is better than all models in both the metrics, while for NER, CLaP compromises faithfulness slightly for stronger accuracy. Awesome-align and EasyProject are both great at attaining higher projection rates but produce less accurate label translations. Overall, intrinsic evaluation demonstrates that CLaP offers the optimal balance between accuracy and faithfulness on a qualitative basis.

4.5 Extrinsic Evaluation

Extrinsic evaluation implicitly assesses the effectiveness of various label projection methods in generating pseudo-training data for downstream tasks. The projected data is filtered based on the faithfulness constraint as $\mathcal{D}_{src}^{tgt}$ and used along with the original English data $\mathcal{D}_{src}$ for downstream training.

For **EAE**, we use X-Gear (Huang et al., 2022), the current state-of-the-art model for zero-shot cross-lingual EAE, as the downstream model. For **NER**, we use XLM-RoBERTa$_{\text{large}}$ (Conneau et al., 2020) as our downstream model and follow XTREME (Hu et al., 2020) setup for implementations. All results are the average over five runs.

Results: We present the EAE results in terms of argument classification F1 scores in Table 2. For reference, we also include the zero-shot baseline (training only on $\mathcal{D}_{src}$). Evidently, CLaP performs the best providing an average gain of 2.4 F1 points over the next best baseline of Awesome-align and a net gain of 7.9 F1 points over the zero-shot baseline. This result is in sync with our intrinsic evaluation wherein CLaP performed the best for EAE.

The primary findings for the F1 scores of entity classification are shown in Table 3. Overall, CLaP outperforms all benchmarks, achieving an absolute enhancement of 0.7 F1 points compared to the zero-shot baseline, and surpassing previous studies by 1.4-1.7 F1 points. The superior performance of the downstream model powered by CLaP, highlights CLaP's efficacy in improving downstream tasks.

LLM usage comparison - Direct Inference v/s Contextual Translation: We compare the fine-tuned models with **LLM-Infer**, a large language model (LLM) baseline directly inferring on the downstream task in the target language. We utilize the chat version of Llama2-13B model (Touvron

Source Sentence	Source Label	Target Lang	Technique	Translated Label	Explanation
Born in Castelvetrano , Trapani and raised in Catania , he moved to Madrid to keep up his busy career .	Castelvetrano	hi	Awesome-align	कैस्टेलवेट्रानो ट्रापानी (Castelvetrano Trapani)	Extra word
			EasyProject	Castelvetrano	No translation
			CLaP	कैस्टेलवेट्रानो (Castelvetrano)	Perfect
Unilaterally leading a coalition featuring tyrannies, effect such change remains a bad idea, Iraq's elections notwithstanding.	Iraq	zh	Awesome-align	伊拉 (Ira-)	Incomplete
			EasyProject	尽管伊拉克 (although Iraq)	Extra word
			CLaP	伊拉克 (Iraq)	Perfect

Table 4: Qualitative examples highlighting the error-cases of the baseline models along with explanations for Hindi (hi) and Chinese (zh). We also show how CLaP performs better and fixes the errors. Blue text is English translation.

et al., 2023) for the baseline. [6] We explore various cross-lingual prompting strategies, following Ahuja et al. (2023) (complete experiments in Appendix E), and report the performance for the best prompt here. From results in Table 2 & 3, we can assert how LLM-infer performs significantly poorer than any fine-tuned model, indicating how LLMs can't infer well on cross-lingual structured prediction. On the other hand, we demonstrate that LLMs can be better utilized to do contextual translation, as used in CLaP, which leads to the best performance for both the downstream tasks. Additional experiments with ChatGPT (Brown et al., 2020) are also provided in Appendix E.

5 Analysis

5.1 Qualitative Analysis

Diving deeper, we qualitatively study typical error cases for the translated labels in four languages by different label projection techniques. In 200 examples of our study, we found that 18% of the time, EasyProject predicts nothing due to markers dropped in the translated sentence, and for 19%, EasyProject simply copies the English label failing to translate it to the target language. For Awesome-align, the majority of errors are due to additional words or incomplete label translations, similar to the observation presented in (Chen et al., 2023). This could be because it is hard for the word-alignment module to decide alignments between sub-words, leading to over-alignment or under-alignment. We show two selected examples of our study from Hindi (hi) and Chinese (zh) in Table 4, where we show how Awesome-align predicts extra words or incomplete words owing to misalignments, and EasyProject fails to translate the word for Hindi while producing extra tokens for Chinese. In both cases, we show how CLaP makes accurate predictions and is more robust in maintaining accurate label translations.

	Model Size	EAE ar	EAE zh	NER yo	NER ur	NER kk
CLaP (w/ Llama-2-13B)	13B	**49.3**	**58.6**	59.6	32.9	42.8
CLaP (w/ GPT-3.5-Turbo)	175B	49.1	58.4	**62.3**	**60.1**	**46.6**

Table 5: Extrinsic evaluation of CLaP using Llama-2-13B and GPT-3.5-Turbo for five languages.

5.2 CLAP with Larger LLMs

We utilize a relatively small LLM Llama-2 (Touvron et al., 2023) with 13B parameters as $\mathcal{M}$ for our experiments with CLaP. Here, we analyze the impact of utilizing a larger LLM for CLaP. More specifically, we compare Llama-2-13B based CLaP with a larger GPT-3.5-Turbo (Brown et al., 2020) based CLaP for five languages for EAE and NER in Table 5. [7] We notice that using GPT-3.5-Turbo in CLaP is at par with the Llama-2 variant for medium to high-resource languages like Arabic (ar) and Chinese (zh). On the other side, for lower-resourced languages like Yoruba (yo), Urdu (ur), and Kazakh (kk), GPT-3.5-Turbo introduces significantly larger improvements of 3 to 30 F1 points. Thus, we hypothesize that larger multilingual LLMs can further improve CLaP, especially for low-resource languages, also evidenced in Bandarkar et al. (2023).

[6]Compared to the text version, the chat version of Llama2 provided better results.

[7]GPT-3.5-Turbo costs $20-$30 per language. Thus, owing to budget constraints, we restrict ourselves to 5 languages.

	ar	zh	**Avg**
Zero-shot	40.3	51.9	43.9
Awesome-align	47.1	53.8	48.4
EasyProject	36.5	55.6	45.4
CLaP (ours)	**48.2**	**56.9**	**50.4**

Table 6: Extrinsic evaluation of the different label projection techniques using translate-train for EAE using the mBART-50 many-to-many translation model.

	ar	zh	**Avg**
Zero-shot	40.3	51.9	43.9
Independent	44.8	54.3	47.6
Constrained	45.6	55.6	48.8
CLaP (ours)	**48.2**	**56.9**	**50.4**
Supervised	63.2	69.7	65.0

Table 7: Ablation study comparing different contextual translation techniques for label projection. Performance is measured by downstream EAE performance.

5.3 Generalization to other translation models

To verify the generalizability of our approach to other translation models, we perform an extrinsic evaluation of the label projection techniques on the EAE task using the mBART-50 many-to-many (MMT) (Kong et al., 2021) translation model. We show the results for this evaluation in Table 6. We see that CLaP performs the best with an average improvement of 2 F1 points over the next best baseline of Awesome-align and 6.5 F1 points over the zero-shot baseline. This result shows our CLaP is a generalizable label projection technique and agnostic to the underlying translation model.

5.4 Ablation Study for CLaP

To study the impact of using instruction-tuned models for *contextual translation*, we conduct an ablation study comparing CLaP with the following baselines which put extra focus on accuracy or faithfulness for contextual machine translation: (1) **Independent** translation uses the translation model $\mathcal{T}$ to independently (without any context of the input sentence) translate the source text labels to the target language (i.e. $\mathbf{y}^{tgt} = \mathcal{T}(\mathbf{y}^{src})$), (2) **Constrained** translation which uses a decoding constraint to carry out the faithfulness requirements. More specifically, during translation, it limits the generation vocabulary to the tokens in the translated sentence x^{tgt}. We follow De Cao et al. (2022); Lu et al. (2022) for implementing these constraints.

We extrinsically evaluate the model performances of the techniques on the task of EAE using the MMT translation model [8] and show the results in Table 7. The independent model compromises faithfulness while the constrained model sacrifices accuracy - but both models outperform the zero-shot baseline. CLaP provides high accuracy and faithfulness and achieves the best performance improving by 1.6 to 2.8 F1 over the ablation baselines.

	EAE		**NER**			**Avg**
	ar	zh	it	es	id	
Zero-shot	36.3	47.3	79.4	74.5	53.1	58.1
Awesome-align	32.8	30.1	**77.5**	69.6	51.4	52.3
EasyProject	17.0	11.5	65.9	62.6	51.8	41.8
CLaP (ours)	**34.3**	**39.5**	73.4	**75.0**	**57.4**	**55.9**

Table 8: Extrinsic evaluation of the different label projection techniques using translate-test using GMT for EAE and NER. Avg = Average

5.5 CLaP for Translate-Test

Another popular technique for cross-lingual transfer is translate-test (Hu et al., 2020; Ruder et al., 2021) which was discussed in § 2.3. As part of this analysis, we study the applicability of CLaP for translate-test using extrinsic evaluation on Arabic (ar) and Chinese (zh) for EAE and Italian (it), Spanish (es), and Indonesian (id) for NER. We show the results in Table 8. Overall, we see how CLaP outperforms both the other methods significantly achieving the best scores for 4 out of the 5 languages. EasyProject performs the worst as it uses the translation model twice causing higher error propagation. We also note how translate-test doesn't yield improvements over the zero-shot baseline, especially for EAE as it requires using label projection twice (once for trigger and once for arguments), thus leading to error propagation.

6 CLaP for Low-Resource Languages

To cater our model to a wide range of languages, we study the applicability of CLaP for low-resource languages. Specifically, we consider the task of NER for 10 low-resource languages from Africa and South America. For the test datasets, we utilize MasakhaNER (Adelani et al., 2022) for 9 African languages: Hausa (ha), Igbo (ig), Chichewa (ny), Kinyarwanda (rw), chShona (sn), Kiswahili (sw), isiXhosa (xh), Yorùbá (yo), isiZulu (zu), and refer to Zevallos et al. (2022) for the South American

[8]Since decoding-time constraints for the Constrained model can't be applied to GMT

Lang	ha	ig	ny	rw	sn
Zero-shot	**72.9**	46.4	49.0	45.0	50.2
Awesome-align	72.2	**64.1**	**64.9**	**55.9**	55.4
EasyProject	72.0	54.6	50.5	54.5	42.5
CLaP (ours)	69.9	60.5	58.7	53.6	**59.7**
	sw	**xh**	**yo**	**zu**	**qu**
Zero-shot	**88.6**	61.0	**33.6**	**67.1**	37.9
Awesome-align	82.9	52.4	30.8	57.9	46.1
EasyProject	81.3	50.6	25.2	44.3	44.1
CLaP (ours)	80.7	**61.3**	30.6	54.4	**48.7**

Table 9: Extrinsic evaluation of the different label projection techniques using translate-train using GMT for NER for 10 low-resource languages.

language Quechua (qu). We conduct extrinsic evaluation of translate-train models transferring from the English CoNLL training data[9] using the GMT model and present the results in Table 9. We observe that this is a particularly challenging setting as all the label projection techniques fail to improve over the zero-shot model for 4 languages. Our model CLaP improves for 6 languages and performs the best for 3 languages. This result is particularly encouraging as our model uses a small and English-centric 13B Llama-2 model and utilizing larger multilingual LLMs will amplify these improvements further (as shown in § 5.2). [10]

7 Related Works

Zero-shot Cross-lingual Structure Extraction Since the emergence of strong multilingual models (Devlin et al., 2019; Conneau et al., 2020), various works have focused on zero-shot cross-lingual learning (Hu et al., 2020; Ruder et al., 2021) and code-switching (Garg et al., 2018; Hsu et al., 2023b) for various structure extraction tasks like named entity recognition (Li et al., 2021; Yang et al., 2022), relation extraction (Ni and Florian, 2019; Subburathinam et al., 2019), slot filling (Krishnan et al., 2021), and semantic parsing (Nicosia et al., 2021; Sherborne and Lapata, 2022). Recent works have focussed on building datasets (Pouran Ben Veyseh et al., 2022; Parekh et al., 2023), benchmarking (Huang et al., 2023) as well as developing novel modeling designs exploring the usage of parse trees (Subburathinam et al., 2019; Ahmad et al., 2021a; Hsu et al., 2023c), data projection (Yarmohammadi et al., 2021), pooling strategies (Agarwal et al., 2023) and generative models (Hsu et al., 2022; Huang et al., 2022) to improve cross-lingual transfer. We utilize the state-of-the-art model X-Gear (Huang et al., 2022) and XLM-R (Conneau et al., 2020) as the downstream models for EAE and NER respectively, and improve them further using CLaP-guided translate-train.

Label Projection Techniques Several works have attempted to solve label projection for various structure extraction tasks such as semantic role labeling (Aminian et al., 2017; Fei et al., 2020), slot filling (Xu et al., 2020), semantic parsing (Moradshahi et al., 2020; Awasthi et al., 2023), NER (Ni et al., 2017; Stengel-Eskin et al., 2019), and question-answering (Lee et al., 2018; Lewis et al., 2020; Bornea et al., 2021). The earliest works (Yarowsky et al., 2001; Akbik et al., 2015) utilized statistical word-alignment techniques like GIZA++ (Och and Ney, 2003) or fast-align (Dyer et al., 2013) for locating the labels in the translated sentence. Recent works (Chen et al., 2023) have also explored the usage of neural word aligners like QA-align (Nagata et al., 2020) and Awesome-align (Dou and Neubig, 2021). Another set of works has explored the paradigm of mark-then-translate using special markers like quote characters ("") (Lewis et al., 2020), XML tags (<a>) (Hu et al., 2020), and square braces ([0]) (Chen et al., 2023) to locate the translated labels. Overall, both these techniques can be error-prone and have poorer translation quality (Akbik et al., 2015), as shown in § 4.4 and 5.1. A recent concurrent work CODEC (Le et al., 2024) improves the translation quality of text with markers by constrained decoding and data augmentation.

8 Conclusion and Future Work

In our work, we propose a novel approach CLaP for label projection, which utilizes contextual machine translation using instruction-tuned language models. Experiments on two structure prediction tasks of EAE and NER across 39 languages demonstrate the effectiveness of CLaP compared to other label projection techniques. Intrinsic evaluation provides deeper insights that justify our model improvements. Additional experiments using larger LLMs, various translation models, translate-test paradigm, and 10 extremely low-resource languages demonstrate the generalizability and future potential of CLaP for cross-lingual structured prediction.

[9] For qu, we only use 3,000 CoNLL training data points due to budget constraints.

[10] Owing to budget constraints, we left the exploration as future work.

Acknowledgements

We thank Xueqing Wu, Yang Chen, Kareem Ahmed, Syed Shahriar, Tao Meng, and Sidi Lu for their valuable insights, experimental setups, intrinsic evaluation, paper reviews, and constructive comments. We thank the anonymous reviewers for their feedback. This work was partially supported by NSF 2200274, AFOSR MURI via Grant #FA9550- 22-1-0380, Defense Advanced Research Project Agency (DARPA) grant #HR00112290103/HR0011260656, and a Cisco Sponsored Research Award.

Limitations

In our work, we show the effectiveness of our model CLaP on two representative structure prediction tasks of EAE and NER. Its effectiveness for other structure prediction tasks remains unknown and can be extended in future works. For CLaP, we utilized the 13B version of the Llama-2 model as the base instruction-tuned language model as a proof-of-concept for the effectiveness of CLaP. Future works can explore the usage of other stronger LLMs to enhance the model performance. Lastly, we would like to point out that our model doesn't improve over the zero-shot model for several languages, mainly owing to the limited language understanding and poor translation quality. However, the focus of our work has been to show the effectiveness of our model with other used label projection techniques. With growing model sizes and enhanced coverage of languages, we posit that our model will eventually be able to provide significant improvements for all languages.

Ethical Concerns

We use an instruction-tuned language model (specifically LLama-2) as the base model for CLaP. Since these instruction-tuned models are not trained equitably in all languages, the model generation quality may vary drastically for each language. Furthermore, since these models are not trained on filtered safe content data, the model may potentially generate harmful content.

References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shantanu Agarwal, Steven Fincke, Chris Jenkins, Scott Miller, and Elizabeth Boschee. 2023. Impact of subword pooling strategy on cross-lingual event detection. *CoRR*, abs/2302.11365.

Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021a. Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554, Online. Association for Computational Linguistics.

Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021b. GATE: graph attention transformer encoder for cross-lingual relation and event extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12462–12470. AAAI Press.

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng. 2019. Cross-lingual dependency parsing with unlabeled auxiliary languages. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 372–382, Hong Kong, China. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition Banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China. Association for Computational Linguistics.

Maryam Aminian, Mohammad Sadegh Rasooli, and Mona Diab. 2017. Transferring semantic roles using translation and syntactic information. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 13–19, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Maryam Aminian, Mohammad Sadegh Rasooli, and Mona Diab. 2019. Cross-lingual transfer of semantic roles: From raw text to semantic roles. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 200–210, Gothenburg, Sweden. Association for Computational Linguistics.

Abhijeet Awasthi, Nitish Gupta, Bidisha Samanta, Shachi Dave, Sunita Sarawagi, and Partha Talukdar. 2023. Bootstrapping multilingual semantic parsers using large language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2455–2467, Dubrovnik, Croatia. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *CoRR*, abs/2308.16884.

Mihaela A. Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. 2021. Multilingual transfer learning for QA using translation as data augmentation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12583–12591. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. 2023. Red$^{\text{fm}}$: a filtered and multilingual relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4326–4343, Toronto, Canada. Association for Computational Linguistics.

Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Zero-shot cross-lingual event argument extraction with language-oriented prefix-tuning. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12589–12597. AAAI Press.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.

Yang Chen and Alan Ritter. 2021. Model selection for cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5675–5687, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference* *on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

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, Atlanta, Georgia. Association for Computational Linguistics.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, Online. Association for Computational Linguistics.

Steven Fincke, Shantanu Agarwal, Scott Miller, and Elizabeth Boschee. 2022. Language model priming for cross-lingual event extraction. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10627–10635. AAAI Press.

Saurabh Garg, Tanmay Parekh, and Preethi Jyothi. 2018. Code-switched language models using dual RNNs and same-source pretraining. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3083, Brussels, Belgium. Association for Computational Linguistics.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *CoRR*, abs/2301.07597.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *CoRR*, abs/2305.14450.

Leonhard Hennig, Philippe Thomas, and Sebastian Möller. 2023. MultiTACRED: A multilingual version of the TAC relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Toronto, Canada. Association for Computational Linguistics.

I-Hung Hsu, Xiao Guo, Premkumar Natarajan, and Nanyun Peng. 2021. Discourse-level relation extraction via graph pooling. *CoRR*, abs/2101.00124.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

I-Hung Hsu, Kuan-Hao Huang, Shuning Zhang, Wenxin Cheng, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2023a. TAGPRIME: A unified framework for relational structure extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12917–12932, Toronto, Canada. Association for Computational Linguistics.

I-Hung Hsu, Avik Ray, Shubham Garg, Nanyun Peng, and Jing Huang. 2023b. Code-switched text synthesis in unseen language pairs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5137–5151, Toronto, Canada. Association for Computational Linguistics.

I-Hung Hsu, Zhiyu Xie, Kuan-Hao Huang, Prem Natarajan, and Nanyun Peng. 2023c. AMPERE: AMR-aware prefix for generation-based event argument extraction model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10976–10993, Toronto, Canada. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1684–1697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kuan-Hao Huang, I-Hung Hsu, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4633–4646, Dublin, Ireland. Association for Computational Linguistics.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Premkumar Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2023. TextEE: Benchmark, reevaluation, reflections, and future challenges in event extraction. *CoRR*, abs/2311.09562.

Chris Jenkins, Shantanu Agarwal, Joel Barry, Steven Fincke, and Elizabeth Boschee. 2023. Massively multi-lingual event understanding: Extraction, visualization, and search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 247–256, Toronto, Canada. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt A good translator? A preliminary study. *CoRR*, abs/2301.08745.

Xiang Kong, Adithya Renduchintala, James Cross, Yuqing Tang, Jiatao Gu, and Xian Li. 2021. Multilingual neural machine translation with deep encoder and multiple shallow decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1613–1624, Online. Association for Computational Linguistics.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 211–223, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Duong Minh Le, Yang Chen, Alan Ritter, and Wei Xu. 2024. Constrained decoding for cross-lingual label projection. *CoRR*, abs/2402.03131.

Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. 2018. Semi-supervised training data generation for multilingual question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Bing Li, Yujie He, and Wenjin Xu. 2021. Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment. *CoRR*, abs/2101.11112.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *CoRR*, abs/2304.11633.

Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. 2022. Summarization as indirect supervision for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6575–6594, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mehrad Moradshahi, Giovanni Campagna, Sina Semnani, Silei Xu, and Monica Lam. 2020. Localizing open-ontology QA semantic parsers in a day using machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5970–5983, Online. Association for Computational Linguistics.

Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. 2020. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565, Online. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Jian Ni and Radu Florian. 2019. Neural cross-lingual relation extraction based on bilingual word embedding mapping. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 399–409, Hong Kong, China. Association for Computational Linguistics.

Massimo Nicosia, Zhongdi Qu, and Yasemin Altun. 2021. Translate & Fill: Improving zero-shot multilingual semantic parsing with synthetic data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3272–3284, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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, Vancouver, Canada. Association for Computational Linguistics.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023. GENEVA: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3664–3686, Toronto, Canada. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Nguyen. 2022. MEE: A novel multilingual event extraction dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9603–9613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Tom Sherborne and Mirella Lapata. 2022. Zero-shot cross-lingual semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153, Dublin, Ireland. Association for Computational Linguistics.

Elias Stengel-Eskin, Tzu-ray Su, Matt Post, and Benjamin Van Durme. 2019. A discriminative neural model for cross-lingual word alignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 910–920, Hong Kong, China. Association for Computational Linguistics.

Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. 2019. Cross-lingual structure transfer for relation and event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 313–325, Hong Kong, China. Association for Computational Linguistics.

Beth M. Sundheim. 1992. Overview of the fourth Message Understanding Evaluation and Conference. In *Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Simone Tedeschi and Roberto Navigli. 2022. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.

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 2002 (CoNLL-2002)*.

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*, pages 142–147.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652.

KayYen Wong, Sameen Maruf, and Gholamreza Haffari. 2020. Contextual neural machine translation improves translation of cataphoric pronouns. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5971–5978, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jian Yang, Shaohan Huang, Shuming Ma, Yuwei Yin, Li Dong, Dongdong Zhang, Hongcheng Guo, Zhoujun Li, and Furu Wei. 2022. CROP: Zero-shot cross-lingual named entity recognition with multilingual labeled sequence translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 486–496, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton Murray, Aaron Steven White, Mark Dredze, and Benjamin Van Durme. 2021. Everything is all it takes: A multi-pronged strategy for zero-shot cross-lingual information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1967, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

Pengfei Yu, Jonathan May, and Heng Ji. 2023. Bridging the gap between native text and translated text through adversarial learning: A case study on cross-lingual event extraction. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 754–769, Dubrovnik, Croatia. Association for Computational Linguistics.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022. Introducing QuBERT: A large monolingual corpus and BERT model for Southern Quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13, Hybrid. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *CoRR*, abs/2304.04675.

A Data Statistics

We present the extensive data statistics for the ACE and WikiANN datasets used for downstream model evaluation on EAE and NER respectively. For ACE, we follow the pre-processing by Huang et al. (2022) to retain 33 event types and 22 argument roles. For WikiAnn, we follow the pre-processing steps described in Rahimi et al. (2019); Hu et al. (2020). For ACE, Table 10 provides statistics about the number of events and arguments for each language. For WikiANN, we present the statistics in Table 11.

Language	Train English	Dev English	Test Arabic	Test Chinese
# Events	4,202	450	198	190
# Arguments	4,859	605	287	336

Table 10: Data Statistics in terms of events and arguments of the ACE dataset for the downstream task of EAE. # indicates 'number of'.

B Complete Results for Intrinsic Evaluation

B.1 Accuracy Evaluation

Accuracy evaluation is done by 5 native bilingual speakers for Chinese, Arabic, Hindi, and Spanish by ranking the translation quality of the translated labels. The native speakers were undergraduate and graduate students who were well-versed in their respective native languages. We present the interface of the google sheets along with the instructions shown to the annotators for Chinese in Figure 4. Similarly, annotation was performed for the other languages as well. We present the complete results as an A/B comparison of the different techniques in terms of their win rates (i.e. percentage when A is better than B) in Table 12. We note how CLaP is more accurate than previous baselines of Awesome-align and EasyProject while at par with the Independent baseline.

B.2 Faithfulness Evaluation

We present the complete results for the faithfulness evaluation per language in Tables 13 and 14 for EAE and NER tasks respectively. For EAE, CLaP has the best faithfulness followed by Awesome-align. For NER, Awesome-align and EasyProject have the highest faithfulness.

Split	Language	# Sentences	# Entities
Train	English (en)	20,000	27,931
Dev	English (en)	10,000	14,146
Test	Afrikaans (af)	1,000	1,487
	Arabic (ar)	10,000	11,259
	Bulgarian (bg)	10,000	14,060
	Bengali (bn)	1,000	1,089
	German (de)	10,000	13,868
	Greek (el)	10,000	12,163
	Spanish (es)	10,000	12,260
	Estonian (et)	10,000	13,892
	Basque (eu)	10,000	13,459
	Farsi (fa)	10,000	10,742
	Finnish (fi)	10,000	14,554
	French (fr)	10,000	13,369
	Hebrew (he)	10,000	13,698
	Hindi (hi)	1,000	1,228
	Hungarian (hu)	10,000	14,163
	Indonesian (id)	10,000	11,447
	Italian (it)	10,000	13,749
	Japanese (ja)	10,000	13,446
	Javanese (jv)	100	117
	Georgian (ka)	10,000	13,057
	Kazakh (kk)	1,000	1,115
	Korean (ko)	10,000	14,423
	Malayalam (ml)	1,000	1,204
	Marathi (mr)	1,000	1,264
	Malay (ms)	1,000	1,115
	Burmese (my)	100	119
	Dutch (nl)	10,000	13,725
	Portuguese (pt)	10,000	12,823
	Russian (ru)	10,000	12,177
	Swahili (sw)	1,000	1,194
	Tamil (ta)	1,000	1,241
	Telugu (te)	1,000	1,171
	Thai (th)	10,000	16,970
	Tagalog (tl)	1,000	1,034
	Turkish (tr)	10,000	13,587
	Urdu (ur)	1,000	1,020
	Vietnamese (vi)	10,000	11,305
	Yoruba (yo)	100	111
	Chinese (zh)	10,000	12,049

Table 11: Data Statistics in terms of sentences and entities of the WikiANN dataset for the downstream task of NER. # indicates 'number of'.

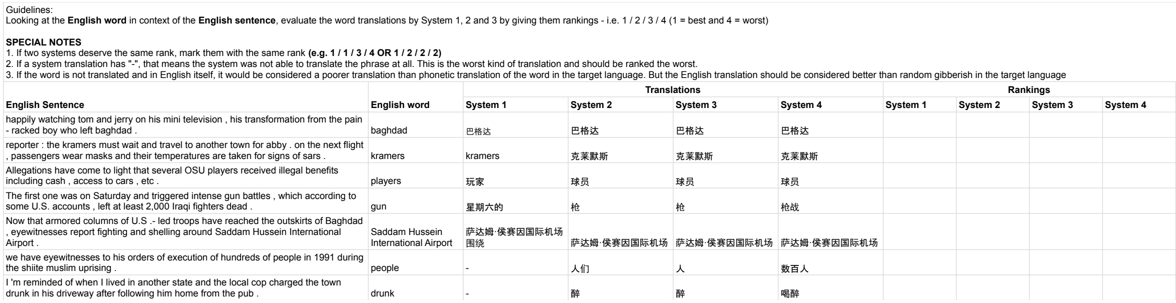
C Additional Implementation Details

C.1 X-Gear

X-Gear is used as the downstream model for EAE for extrinsic evaluation of the label projection techniques. The original X-Gear work (Huang et al., 2022) explored two base multilingual models: mBART-50-large (mBART) (Kong et al., 2021) and the mT5-base (mT5) (Xue et al., 2021). They also explored the usage of copy mechanism (See et al., 2017) to prompt the models to predict strings from the input sentence. In our work, we utilized mBART without copy (mBART), mT5 without copy (mT5), and mT5 with copy mechanism

System 1	v/s	System 2	Arabic			Chinese			Hindi			Spanish		
			S1	Tie	S2	S1	Tie	S2	S1	Tie	S2	S1	Tie	S2
CLaP		Awesome-align	**36%**	58%	6%	**45%**	50%	5%	**20%**	74%	6%	**12%**	84%	4%
CLaP		EasyProject	**52%**	32%	16%	**56%**	39%	5%	**42%**	48%	10%	**30%**	66%	4%
CLaP		Independent	18%	60%	**22%**	12%	71%	**17%**	**18%**	64%	**18%**	**24%**	68%	8%
Independent		Awesome-align	**44%**	42%	14%	**39%**	57%	4%	**28%**	60%	12%	**20%**	64%	16%
Independent		EasyProject	**50%**	44%	6%	**50%**	46%	4%	**52%**	36%	12%	**32%**	52%	16%
Awesome-align		EasyProject	**42%**	26%	32%	**34%**	50%	16%	**42%**	42%	16%	**26%**	64%	10%

Table 12: A/B comparison of the various label projection techniques for accuracy evaluation for the Google Translation model. Accuracy is measured as the label translation quality by native human speakers. Here, **S1** = System 1 is better, **S2** = System 2 is better, and **Tie** = similar quality. The better systems are highlighted in **bold**.


Guidelines:
Looking at the **English word** in context of the **English sentence**, evaluate the word translations by System 1, 2 and 3 by giving them rankings - i.e. 1 / 2 / 3 / 4 (1 = best and 4 = worst)

SPECIAL NOTES
1. If two systems deserve the same rank, mark them with the same rank **(e.g. 1 / 1 / 3 / 4 OR 1 / 2 / 2 / 2)**
2. If a system translation has "-", that means the system was not able to translate the phrase at all. This is the worst kind of translation and should be ranked the worst.
3. If the word is not translated and in English itself, it would be considered a poorer translation than phonetic translation of the word in the target language. But the English translation should be considered better than random gibberish in the target language

		Translations				Rankings			
English Sentence	**English word**	**System 1**	**System 2**	**System 3**	**System 4**	**System 1**	**System 2**	**System 3**	**System 4**
happily watching tom and jerry on his mini television , his transformation from the pain - racked boy who left baghdad .	baghdad	巴格达	巴格达	巴格达	巴格达				
reporter : the kramers must wait and travel to another town for abby . on the next flight , passengers wear masks and their temperatures are taken for signs of sars .	kramers	kramers	克莱默斯	克莱默斯	克莱默斯				
Allegations have come to light that several OSU players received illegal benefits including cash , access to cars , etc .	players	玩家	球员	球员	球员				
The first one was on Saturday and triggered intense gun battles , which according to some U.S. accounts , left at least 2,000 Iraqi fighters dead .	gun	星期六的	枪	枪	枪战				
Now that armored columns of U.S .- led troops have reached the outskirts of Baghdad , eyewitnesses report fighting and shelling around Saddam Hussein International Airport .	Saddam Hussein International Airport	萨达姆·侯赛因国际机场围绕	萨达姆·侯赛因国际机场	萨达姆·侯赛因国际机场	萨达姆·侯赛因国际机场				
we have eyewitnesses to his orders of execution of hundreds of people in 1991 during the shiite muslim uprising .	people	-	人们	人	数百人				
I 'm reminded of when I lived in another state and the local cop charged the town drunk in his driveway after following him home from the pub .	drunk	-	醉	醉	喝醉				

Figure 4: Annotation Interface for conducting the intrinsic evaluation for Accuracy. The shown examples are for Chinese, while the study was done for Hindi, Spanish, and Arabic as well.

Techniques	ar	zh	Avg.
Independent	33	38	35
Awesome-align	66	83	74
EasyProject	31	66	48
CLaP	74	85	**79**

Table 13: Faithfulness evaluation of the various label projection techniques for EAE as a percentage of the times the translated labels were present in the translated input sentence. Numbers are in percentage (%). Higher faithfulness is better and the best techniques are highlighted in **bold**.

(mT5+Copy) as the downstream models. We present details about the hyperparameter settings for these models in Table 16. We run experiments for CLaP on a NVIDIA GeForce RTX 2080 Ti machine with support for 8 GPUs.

C.2 XLM-R

XLM-R (Conneau et al., 2020) is used as the downstream model for NER for extrinsic evaluation of the label projection techniques. We mainly follow the XTREME (Hu et al., 2020) framework for setting up the task and model. We present details about the hyperparameter settings for this model in Table 15. We run experiments for CLaP on a NVIDIA GeForce RTX 2080 Ti machine with support for 8 GPUs.

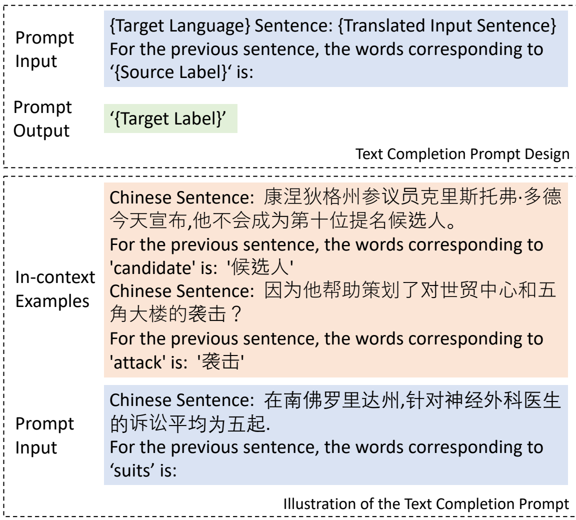


Figure 5: Illustration of the text-completion prompt used for contextual machine translation for our CLaP model.

C.3 CLaP

We provide a couple of prompt designs we used for our model in Figure 5 along with an illustration for Chinese. We additionally provide a similar template for chat version of the model (which is used for experiments with GPT3.5-turbo as reported in

Techniques	af	ar	bg	bn	de	el	es
Independent	78	66	67	74	79	57	70
Awesome-align	99	95	98	92	99	98	99
EasyProject	100	98	83	98	97	89	99
CLaP	94	75	63	93	79	46	84
	et	**eu**	**fa**	**fi**	**fr**	**he**	**hi**
Independent	70	64	61	71	71	71	65
Awesome-align	98	97	96	99	98	95	93
EasyProject	97	94	99	98	99	94	36
CLaP	92	91	72	92	74	80	90
	hu	**id**	**it**	**ja**	**jv**	**ka**	**kk**
Independent	68	77	74	68	66	64	56
Awesome-align	98	99	99	58	98	95	94
EasyProject	97	99	98	95	94	99	77
CLaP	93	84	78	67	53	70	85
	ko	**ml**	**mr**	**ms**	**my**	**nl**	**pt**
Independent	63	57	73	80	53	76	76
Awesome-align	96	88	92	99	90	99	97
EasyProject	93	87	73	98	62	100	99
CLaP	64	88	95	82	55	85	89
	ru	**sw**	**ta**	**te**	**th**	**tl**	**tr**
Independent	59	79	72	76	66	81	76
Awesome-align	97	96	91	91	51	99	98
EasyProject	99	97	91	87	99	99	98
CLaP	66	94	96	90	57	58	94
	vi	**ur**	**yo**	**zh**	**Avg.**		
Independent	74	74	45	66	69		
Awesome-align	83	97	92	92	**93**		
EasyProject	98	94	77	92	92		
CLaP	89	91	88	60	79		

Table 14: Faithfulness evaluation of the various label projection techniques for NER as a percentage of the times the translated labels were present in the translated input sentence. Numbers are in percentage (%). Higher faithfulness is better and the best techniques are highlighted in **bold**.

§ 5.2) in Figure 6. We report the hyperparameter settings for our model in Table 17. We run experiments for CLaP on a NVIDIA GeForce RTX 2080 Ti machine with support for 8 GPUs.

C.4 EasyProject

Compared to the original EasyProject work, we made certain changes in the re-implementation for our work to provide a fair comparison. First, we use square-indexed markers (e.g. [0] and [/0]) compared to XML markers (e.g. <LOC> and </LOC>) used by EasyProject. This is mainly because we obtained much higher retention rates using square-indexed markers (88.2%) compared to XML markers (6.2%) in our initial studies. Secondly, the original EasyProject model uses a finetuned NLLB-200-3.3B model as the translation model. Since we don't finetune CLaP or Awesome-align, we use the non-finetuned Google Machine Translation (GMT) model as the translation model.

Base Model	XLM - Roberta - Large
# Training Epochs	5
Training Batch Size	32
Evaluation Batch Size	32
Learning Rate	2×10^{-5}
Weight Decay	0
Max Sequence Length	128
# Accumulation Steps	1
# Saving Steps	1000

Table 15: Hyperparameter details for the NER downstream XLM-R model.

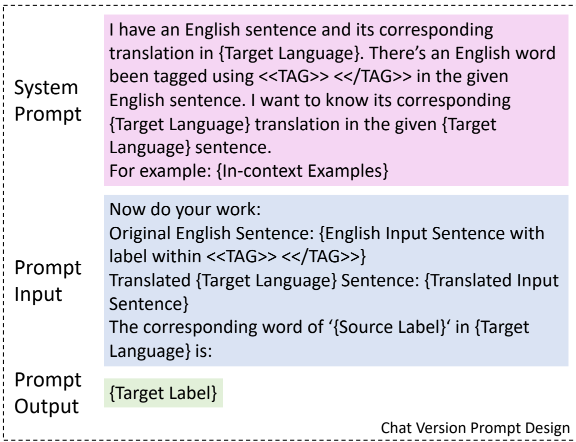


Figure 6: Illustration of the chat version prompt used for contextual machine translation for our CLaP model.

D Complete results for Extrinsic Evaluation

D.1 Event Argument Extraction

Here, we explore three versions of the X-Gear (Huang et al., 2022) model: mBART without copy (mBART), mT5 without copy (mT5), and mT5 with copy mechanism (mT5+Copy). We present the extrinsic evaluation for EAE by training these three models with the label projection techniques for translate-train in Table 18. Results indicate how CLaP performs the best across all the three variations of the model.

E Large Language Model Direct Inference Analysis

Large language models (LLMs) have shown great zero-shot and few-shot capabilities for several tasks like sentiment analysis, machine translation, and question-answering (Guo et al., 2023; Jiao et al., 2023). However, employing a directly prompted

	mBART	mT5	mT5+Copy
Base Model	multilingual BART-Large	multilingual T5-Large	multilingual T5-Large
Usage of copy	No	No	Yes
Training Batch Size	16	16	16
Eval Batch Size	32	32	32
Learning Rate	2×10^{-5}	1×10^{-4}	2×10^{-5}
Weight Decay	1×10^{-5}	1×10^{-5}	1×10^{-5}
# Warmup Epochs	5	5	5
Gradient Clipping	5	5	5
Max Training Epochs	60	60	60
# Accumulation Steps	1	1	1
Beam Size	4	4	4
Max Sequence Length	350	350	350
Max Output Length	100	100	100

Table 16: Hyperparameter details for the EAE downstream X-Gear model.

Base Model	llama-2-13b
Temperature	0.6
Top-p	0.9
Maximum Generation Length	64-128
# In-context examples	2

Table 17: Hyperparameter details for the CLaP model.

	mBART		mT5		mT5+Copy		Avg
	ar	zh	ar	zh	ar	zh	
LLM-Infer	-	-	-	-	16.9^{+}	24.0^{+}	20.5
Zero-shot*	36.3	47.3	36.7	51.0	40.3	51.9	43.9
Awesome-align	45.2	49.4	**46.8**	53.7	48.6	54.5	49.7
EasyProject	37.9	52.3	34.5	54.6	38.5	56.3	45.7
CLaP (ours)	**46.0**	**53.4**	44.3	**56.5**	**49.3**	**58.6**	**51.4**

Table 18: Extrinsic evaluation of the different label projection techniques regarding downstream model performance using translate-train for EAE. Avg = Average. * indicates the reproduced results of X-Gear (Huang et al., 2022). Results for LLM-Infer (marked with $^{+}$) are independent of the XGear base model.

LLM for information extraction and structured prediction tasks in cross-lingual settings is an understudied area. Current evidence, including recent studies by Han et al. (2023) and Li et al. (2023), indicates that LLM performance for these tasks, even for English, lags behind best fine-tuned models. To this end in our work, we evaluate LLMs for direct inference on non-English structured prediction through our baseline **LLM-Infer**.

We utilize two LLMs of varying sizes for LLM-Infer: Llama-2-chat (13B version) (Touvron et al., 2023) and GPT-3.5-Turbo (Brown et al., 2020). We illustrate the prompts used for this baseline in Figure 7. Our LLM prompts involve 2-shot and 4-shot in-context examples, and we meticulously explore three distinct prompting strategies, specifi-

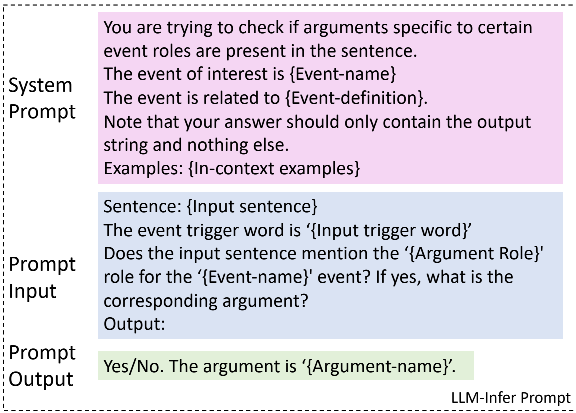


Figure 7: Illustration of the prompt used for the LLM-infer baseline to directly utilize LLMs for downstream structured prediction tasks.

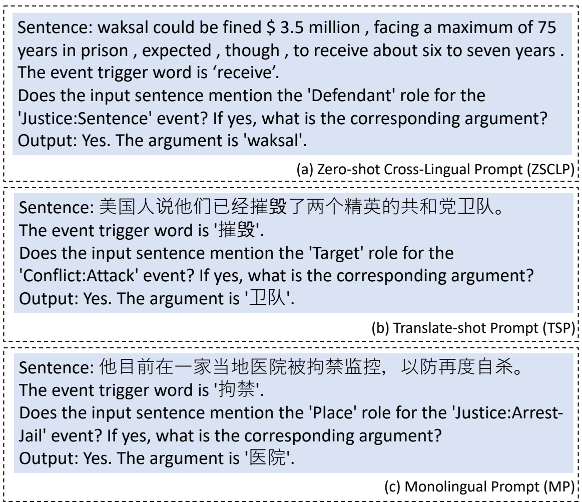


Figure 8: Illustration of the in-context examples used for the three different prompting strategies for LLM-Infer baseline.

Base Model	Prompting Strategy	k-shot	EAE ar	EAE zh	NER hi	NER ms	NER yo	Avg
Llama2-13b-chat	ZSCLP	2	13.4	20.0	21.7	30.1	26.4	22.3
Llama2-13b-chat	ZSCLP	4	14.2	17.9	39.5	38.3	31.9	28.4
Llama2-13b-chat	TSP	2	16.9	24.0	18.9	46.5	28.6	27.0
Llama2-13b-chat	TSP	4	8.7	22.8	17.5	43.5	36.2	25.7
Llama2-13b-chat	MP	2	18.9	28.1	13.7	49.2	17.6	25.5
Llama2-13b-chat	MP	4	11.9	26.0	13.7	61.5	17.4	26.1
GPT-3.5-turbo	ZSCLP	2	15.8	22.3	64.4	50.7	39.7	38.6
GPT-3.5-turbo	ZSCLP	4	15.9	23.6	65.0	53.0	39.0	39.3
GPT-3.5-turbo	TSP	2	17.1	22.3	59.3	54.6	53.3	41.3
GPT-3.5-turbo	TSP	4	17.2	24.5	52.3	57.2	48.8	40.0
GPT-3.5-turbo	MP	2	15.3	25.2	59.5	64.1	51.0	44.7
GPT-3.5-turbo	MP	4	19.5	28.8	58.5	65.4	48.5	44.1
Zero-shot Model			40.3	51.9	70.6	53.4	34.1	50.1
CLaP Translate-train (Ours)			**49.3**	**58.6**	**73.1**	**73.5**	**59.6**	**62.8**

Table 19: Evaluation of LLM-based inference and their comparison with our label projected translate-train model CLaP. This study is done on Event Argument Extraction (EAE) for two languages - Arabic (ar) and Chinese (zh) - and on Named Entity Recognition (NER) for three languages: Hindi (hi), Malay (ms), and Yoruba (yo).

cally for the cross-lingual setting, following Ahuja et al. (2023) (also illustrated in Figure 8). These strategies are listed as follows:

1. **Zero-shot Cross-Lingual Prompt (ZSCLP)**: This strategy involves using k-shot examples from a pivot language (English in our study), which differs from the language of the test example, as shown in Figure 8(a).

2. **Translate-shot Prompt (TSP)**: In this strategy, we first obtain k-shot examples from the pivot language and subsequently perform label projection (using CLaP) to the target language on these examples. These label-projected examples are used as in-context examples in the final prompt (Figure 8(b)).

3. **Monolingual Prompt (MP)**: This method uses k-shot human-labeled examples directly from the target language (Figure 8(c)).

While the first two strategies align with the zero-shot cross-lingual transfer setting, where the availability of data is limited to English, the third strategy offers a slight variation. It presupposes the availability of a few examples in the target languages. For a fair comparison, only the first two strategies are used to compare with CLaP, while the third strategy serves as a comparison datapoint for elucidating the difference between label-projected and human-labeled data as in-context examples.

We conduct this analysis on EAE across two languages and NER across three languages (as it's expensive to conduct this study for all the languages). The selection of languages for NER is to consider both resource diversity (hi: medium-high resource; ms: medium resource; yo: low resource) and script diversity. We compare these models with the zero-shot baseline and our proposed CLaP translate-train model. We show the model performance results in terms of F1 scores for this study in Table 19.

This study reveals several insights: (1) We observe that GPT-3.5-turbo significantly performs better than the Llama-2-13B model - signifying the importance of a larger model size. (2) Comparing different prompting strategies, we observe little variation in model performance for the Llama-2-13B model, while a larger variation for GPT-3.5-turbo. Majorly, we observe that the label-projected in-context examples are better than the English examples, while human-labeled examples provide further gains of 3-4 F1 points. (3) We observe that on average, the LLM-Infer models perform poorer than the zero-shot fine-tuned model. These differences are massive for EAE, while for NER, LLM-Infer performs better for low-resource languages (ms and yo) using our label projected examples. (4) Finally, we observe that CLaP performs the best across all tasks and all languages, even in cases where few-shot examples in target languages are used (MP prompting strategy). All these insights validate CLaP's manner of leveraging LLMs to solve zero-shot cross-lingual structured prediction tasks i.e. CLaP is better than direct LLM prompting.