Modern Cross-Lingual Information Extraction

Luis F. Guzman-Nateras
Department of Computer Science
University of Oregon
lfguzman@cs.uoregon.edu

Abstract

Applications such as automated personal assistants, automatic question answering, and machine-based translation systems have become mainstays of modern culture thanks to the recent considerable advances in Natural Language Processing research. However, a vast majority of such efforts remain limited to a small set of languages. With 7000+ languages spoken around the world, this unbalanced focus leaves marginalized communities unable to take advantage of such technological innovations. Cross-Lingual Learning looks to address this inequality by transferring knowledge from a high-resource *source* language into a low-resource *target* language. This paper provides a survey of recent Cross-Lingual efforts for Information Extraction (CLIE). We first provide some background on the resources leveraged by cross-lingual methods and the knowledge-transfer paradigms that characterize them. Then, state-of-the-art methods are organized into a taxonomy based on the information extraction sub-task they tackle and the knowledge-transfer archetype they employ. Finally, we discuss several suitable directions for future CLIE research efforts.

1 Introduction

Recent years have seen development and widespread adoption of Machine Learning (ML) based applications with Natural Language Processing (NLP) backbones. For instance, applications such as automatic question answering, automated personal assistants, fake news identification, and product review sentiment analysis make use of NLP-based models which are usually trained on a supervised manner by leveraging large amounts of labeled data. Large annotated datasets are, however, remain a luxury reserved for a handful of widely-spoken languages, e.g., English, Chinese, Spanish. As such, the vast majority of NLP research efforts focus on these, so-called, *high-resource* languages. This biased focus marginalizes communities where *low-resource* languages are primarily spoken as they are unable to take advantage of the aforementioned technological innovations.

Cross-Lingual Learning (CLL) provides an alternative to address the lack of labeled data in low-resource languages. The main idea behind CLL is to utilize annotated data from a high-resource *source* language to create models that work effectively in a low-resource *target* language. As such, CLL opens up the possibility of creating entirely new NLP models for languages suffering from data scarcity, or increasing the performance of already existing ones, allowing them to benefit from the aforementioned NLP-based tools. Additionally, CLL efforts usually work under the assumption that no annotated data is available for the target language. This setting is referred to as *zero-shot* cross-lingual learning.

A prominent NLP task to which CLL can be applied is Information Extraction (IE). The information extraction task, as a whole, can be thought of as taking raw, unstructured texts and producing structured versions. It has acquired great significance in the past couple of decades due to the increasing amount of unstructured information available from online platforms: social media posts, discussion forums, crowd-maintained archives, etc. Being able to perform computations on the previously unstructured data is the ultimate goal of information extraction. Nonetheless, such objective is complicated which is why the IE task has been broken down into several, simpler sub-tasks: Entity Mention Detection (EMD), Event Extraction (EE), Relationship Extraction (RE), and Co-Reference Resolution (CRR). We refer to this area of NLP research as Cross-Lingual Information Extraction (CLIE).

The objective of this survey paper is to provide a taxonomy of recent CLIE research works. Nonetheless, CLL efforts can be categorized by a number

of distinct factors: the languages they address, the cross-lingual resources they leverage, the tasks they tackle, or the knowledge-transfer paradigm they utilize. For instance, Pikuliak et al. (2021) provide a comprehensive CLL survey and base their categorization on *what* is being transferred between languages: labels, features, parameters, or representations. Alternatively, given its focus on CLIE in particular, in this survey we choose to employ the addressed IE sub-task as the main categorization feature for the surveyed works. Then, we utilize knowledge-transfer paradigm as a secondary characteristic to create our proposed taxonomy.

We organize the rest of the document as follows: Section 2 describes general cross-lingual terminology used throughout this work. Sections 3, 4, 5, 6 delve into each of IE's composing sub-tasks EMD, EE, RE, and CRR, respectively. Each one includes a definition and an overview of recent CLIE works for the corresponding task. Finally, Section 7 discusses several research directions for future CLIE efforts.

2 Cross-Lingual Concepts

Before discussing the details of current CLIE approaches, this section presents a brief description of relevant cross-lingual concepts that are used throughout this work.

2.1 Cross-Lingual Resources

Depending on the chosen pair, the differences between the source and target languages can be quite significant. For example, the languages could have different word orders, vocabularies, syntax, or even use completely distinct sets of characters. As such, when creating cross-lingual models, it is necessary to have resources that show how the two languages relate to one another. This section describes the most commonly used of such *cross-lingual* resources.

2.1.1 Parallel Corpus

A parallel corpus is one of the most useful, but also the most scarce, bilingual resource. Creating a parallel corpus can, in some cases, be even more expensive than creating a labeled dataset for a specific task (Langedijk et al., 2022). Though parallel corpora have been created for specific domains, e.g., the Bible has been translated for multiple languages, this domain-specificity limits their general application.

2.1.2 Pseudo-parallel Corpus

Automated machine translation has witnessed great advances in recent years by leveraging encoder-decoder models (Bahdanau et al., 2015; Liu et al., 2020) and, of course, Google's translation API (Wu et al., 2016) continues to make state-of-the-art translation available for the general public. As such, machine-translation systems can be leveraged to obtain pseudo-parallel text. Afterwards, words in pseudo-parallel sentences can be aligned using automatic tools such as GIZA++ (Och and Ney, 2003), Fast-align (Dyer et al., 2013) and Awesome-align (Dou and Neubig, 2021). A pseudo-parallel corpus via machine translation is an attractive option for cross-lingual models. However, it is limited by the availability of a translation system for the required target language. Furthermore, the quality of the translations plays a crucial role in cross-lingual model performance.

2.1.3 Bilingual Dictionaries/Gazetteers

Bilingual dictionaries, also called lexicons, are collections of pairs of matching words from two different languages. They provide a very natural way of linking the source and target languages and are commonly used to guide the training process of other cross-lingual resources such as bilingual embeddings. Though they are readily available for many language pairs (Mayhew et al., 2017), they also have significant drawbacks as they are frequently incomplete or are plagued with incorrect translations which can lead to noisy cross-lingual results.

2.1.4 Multilingual Word Embeddings

Monolingual word embeddings such as Word2Vec (Mikolov et al., 2013c) and Glove (Pennington et al., 2014) are collections of dense, high-dimentional, real-valued vectors that capture the semantic of words in a language by training them on large amounts of unlabeled monolingual text. These embeddings were the *de facto* standard for word representations in machine learning models for several years (Pikuliak et al., 2021). Multilingual word embeddings, also called bilingual word embeddings, are obtained by having the representations of multiple languages share the same semantic vector space. This is usually achieved by either (1) training monolingual embeddings individually for each language and then learning a projection into a single shared space, or (2) by jointly training using unlabeled

data from multiple languages directly (Ruder et al., 2019). In a sense, multilingual embeddings are secondary cross-lingual resource since they need additional cross-lingual resources, e.g., a bilingual dictionary, to guide the alignment process. There have been, however, proposals for entirely unsupervised multilingual embeddings (Chen and Cardie, 2018; Artetxe et al., 2018; Bojanowski et al., 2017).

2.1.5 Multilingual Language Models

A Language Model (LM) is a probability distribution over sequences of words in a particular language. Language models are trained so that word sequences that appear more frequently in a language will have a higher probabilities. In recent years, large transformer-based (Vaswani et al., 2017) language models trained on large amounts of unlabeled data have obtained state-of-the-art results in several NLP tasks. BERT (Devlin et al., 2019) and GPT (Radford and Narasimhan, 2018) and their variations (RoBERTa, GPT-2, GPT-3) are probably the most well-known monolingual LMs. Multilingual Language Models (MLMs) are just extensions of their monolingual counter parts. They are trained using unlabeled data from multiple languages, e.g, multilingual BERT was trained on Wikipedia content from 104 different languages, and can be leveraged to obtain contextualized multilingual representations that display language-independent features to an extent. Multilingual BERT (mBERT, Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) are two of the most popular pre-trained MLMs.

2.2 Cross-Lingual Transfer Paradigms

With some exceptions, cross-lingual learning methods can be broadly classified into two categories based on the approach to transfer knowledge from source to target: *Data transfer* and *Direct transfer*.

2.2.1 Data Transfer

Cross-lingual learning data transfer methods train a model directly in the target language. Given the unavailability of labeled target-language data under the usual zero-shot setting, this requires projecting the labels from the annotated source data to unlabeled target data. Many approaches in this category rely on the availability of either sentence-aligned parallel corpora (Yarowsky et al., 2001; Hwa et al., 2005; Zeman and Resnik, 2008; Ehrmann et al., 2011; Fu et al., 2014), or neural machine translation systems (Shah et al., 2010; Tiedemann et al., 2014; Jain et al., 2019). In both cases, obtaining good word alignments is key for successful annotation projection as method performance is highly correlated with the quality of the generated data. As such, they usually make use of state-of-the-art automated alignment methods (Och and Ney, 2003; Dyer et al., 2013; Dou and Neubig, 2021) or employ manually-crafted alignments (Jain et al., 2019). An alternative to get around the need for word alignments is to instead do word-by-word, or phrase-to-phrase, translations (Mayhew et al., 2017; Xie et al., 2018). However, these methods do not consider factors such as different word orders in the source and target languages which can introduce noisy training signals.

Data transfer methods can have several advantages over direct transfer methods. In particular, they can directly exploit the lexical features, and other language-specific information, of the target language. Lexical features are very important for several tasks and can be particularly useful useful if the target language is close to the training/source language (Tsai et al., 2016). However, model performance will ultimately depend on how well these language-specific features are explored.

Yarmohammadi et al. (2021) present an in-depth analysis on the benefits of data projection for zero-shot cross-lingual learning on several tasks. They point out that, even though using multilingual pre-trained encoders, e.g., mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020), leads to a strong cross-lingual results, their performance on target languages is usually below that of source languages. The core idea in their work is to augment the training data with so-called *"silver"* data generated by (1) translating the source sentences into the target language, (2) aligning the words between the original and translated parallel sentences, and (3) projecting the labels using the obtained alignments. Then, the obtained silver data is used alongside the original *gold* (source) data to train a cross-lingual model. To evaluate the usefulness of their data projection scheme, they compare against a self-training approach in which a zero-shot cross-lingual model trained solely on source data is used to obtain the labels of the translated sentences. For machine translation, they compare a publicly available one (Tiedemann, 2020) with several of their own models that incorporate using pretrained encoders. For the word alignment, they com-

pare using the statistical model Fast-align (Dyer et al., 2013) and Awesome-align (Dou and Neubig, 2021) which computes alignments based on contextualized-embedding similarity. They evaluate their approach in five downstream tasks: event extraction, using ACE05 (Walker et al., 2006) and BETTER[1], Named Entity Recognition (NER), Part-of-Speech (POS) tagging, and dependency parsing. Their results show that the best-performing model is task dependent given that none of the configurations clearly outperformed the rest. An important finding is that the *large* versions of multi-lingual encoders do not seem to benefit from the additional training data as it is the case for their *base* counterparts.

2.2.2 Direct Transfer

In contrast to data transfer, direct transfer methods train models exclusively on labeled source-language data and rely on developing delexicalized language-independent features so that the task knowledge acquired from the training data can be directly applied to unlabeled target data.

A common approach for direct transfer cross-lingual models is to exploiting a shared representation for the source and target languages (Täckström et al., 2012; Bharadwaj et al., 2016; Kozhevnikov and Titov, 2014; Chaudhary et al., 2018). For instance, Ni et al. (2017) propose to project mono-lingual word embeddings into a common space as language independent features. More recently, it is usual to leverage the encoding capabilities of pre-trained multilingual language models such as mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020).

The greater appeal of direct transfer models is evident: they do not require any labeled data for the target language which is a highly-desirable characteristic, specially for low-resource languages. Furthermore, by not relying on translations or word alignments, they avoid introducing noise into the training signals which can deteriorate model performance. In their work, Artetxe et al. (2020) found that the translation process can introduce subtle artifacts that have a notable impact for cross-lingual transfer learning. For example, for the Natural Language Inference (NLI) task, they found that translating the premise and hypothesis independently reduces the lexical overlap between them which devolves into lower classification performance.

[1]https://www.iarpa.gov/index.php/research-programs/better

Nonetheless, direct transfer techniques have disadvantages as well. Mainly that they cannot leverage target-language lexical features or learn from word-label relations. This puts them at a clear disadvantage when applied to markedly dissimilar languages. Lauscher et al. (2020) found that zero-shot transfer is most successful when applied among typologically similar languages, and less so for languages distant from each other.

To address this limitation, some direct transfer methods have started leveraging unlabeled target data as a means to integrate target-language specific information into the training process via using adversarial learning for instance (Ahmad et al., 2019; Keung et al., 2019; Chen et al., 2021; Phung et al., 2021; Guzman-Nateras et al., 2022c).

2.2.3 Hybrid Transfer

The *data transfer* and *direct transfer* paradigms are orthogonal and can be used in tandem (Tsai et al., 2016). That is, a cross-lingual model can benefit from training with language-agnostic features and also exploit target-language-specific lexical features via annotation projection.

An example of such *hybrid* training is the work by Yarmohammadi et al. (2021) described above (Section 2.2.1) where they leverage the language-invariant capabilities of pre-trained multilingual encoders and so-called *silver* target-data generated with annotation projection.

Knowledge distillation (Wu et al., 2020a,b; Liang et al., 2021; Chen et al., 2021) has also been leveraged for hybrid cross-lingual training: a source-trained multilingual teacher model (direct transfer) annotates unlabeled target data which is then used to train a student model (data transfer).

A direct transfer model can still benefit from data transfer even if a translation system for the target language does not exist. Some studies have shown that learning from multiple source languages can be ultimately beneficial for cross-lingual models (Moon et al., 2019). As such, the original source data can be projected into a second source language (ideally a language close to the desired target) and the cross-lingual model can be trained on both sets of data.

The work by Singh et al. (2019) exemplifies this approach. They propose XLDA: a simple but effective approach to improve the performance of cross-lingual NLP models by using bilingual training samples. Such bilingual examples are created by translating mono-lingual training data into a sec-

ond *augmentor* language and combining both the original text and its translation into a single sample. They evaluate their approach on the Question Answering (QA) and NLI tasks. In NLI, for example, they create the inputs to the model by either translating the premise or the hypothesis. Their experiments use language pairs created from 14 different languages ranging from high (English, Chinese) to very low-resource (Urdu, Swahili). Some interesting findings from their work are: (1) for every language they tested, there is an augmentor language that improves performance over the mono-lingual setting; (2) most languages, other than very low-resource ones, work as suitable augmentors; and (3) low-resource languages benefit the most from XLDA.

3 Entity Mention Detection

3.1 Task Definition

Entity Mention Detection (EMD), also referred to as entity extraction or recognition, is an NLP task for detecting entities in unstructured text and classifying them into a discrete set classes defined by a particular ontology. Commonly used categories include names for organizations, locations, persons, companies, and numerical values such monetary amounts, percentages, time expressions, and codes. For example, in the sentence:

John bought a Dell computer in 2018.

an EMD system would recognize *John* as a ***Person*** entity, *Dell* as an ***Organization/Company*** entity, and *2018* as a ***Time*** entity type.

EMD is a complex task that is usually decomposed as two distinct sub-tasks: segmentation and classification (Carreras et al., 2003). The segmentation sub-task deals with identifying contiguous spans of tokens representing an entity. A common restriction assumed by EMD systems is that there can be no nesting. For instance, in the sentence:

Bank of America closed its doors permanently.

the tokens ***Bank of America*** should be considered as a single entity, disregarding that the token ***America*** could be considered an entity itself. As for the classification sub-task, once entity candidates have been identified, they are categorized into ontology-specific types. This means that the same entity can be designated to a different type when another ontology is used.

A cross-lingual setting implies additional complexity for the EMD task. While some entities such as proper names can remain unchanged in different languages, other, more nuanced, entities can have significant differences. For example, in the English sentence:

Mark Zuckerberg testified before the US Senate.

Mark Zuckerberg should be identified as a *Person* entity and ***US Senate*** should be identified as an *Organization* entity. However, the same sentence in Spanish becomes:

Mark Zuckerberg testificó ante el Senado de los Estados Unidos.

and while the *Person* entity remains the same, the *Organization* entity is very different: it is composed by five tokens instead of two.

3.2 Data Transfer Cross-lingual EMD

Mayhew et al. (2017) refer to their approach as *"Cheap Translation"* as it is not based on large parallel corpora. Instead, they leverage smaller bilingual dictionaries called *lexicons* which contain word-to-word translations as well as word-to-phrase, phrase-to-word, and phrase-to-phrase translations. Using these lexicons they create target-language training data by doing one-to-one word translations from the labeled source-language data. The limited size of the lexicons (not every word from the source language is covered) and the simplicity of their approach (their translations do not account for word re-ordering) means that the translated data contains several issues: some words are not translated or translated incorrectly. Nonetheless, the authors argue that despite this problems, most of the context around entities is reasonably preserved which still leads to good entity detection performance. In their experiments, they also notice that their approach works better when the source and target languages have similar properties (e.g., word order, alphabets) or belong to the same language family.

In semi-concurrent work, Feng et al. (2018) propose to enrich the representations of target-language words by incorporating information from their corresponding source-language translations. Their intuition is that different languages provide complementary information about entities and that these cues can be transferred via bilingual dictionaries. Thus, they generate a *translation memory unit* for each target-language word by stacking together the embeddings of all suitable translation candidates obtained from a bilingual dictionary (a single word usually has several translation candidates). Additionally, the embeddings in these translation

units are weighted by an *attention network* that estimates the semantic relatedness of each translation candidate with the target word. To deal with out-of-lexicon words, they introduce lexicon extension strategy in which they learn a linear transformation between the target-word embeddings and the translation-unit embeddings. Finally, to perform entity detection, the target-word embeddings are concatenated with their corresponding translation units and fed into a Bi-LSTM with a CRF layer on top.

Following on the work by Ni et al. (2017), Xie et al. (2018) present an approach that combines the use of Bilingual Word Embeddings (BWE) with word-by-word translation. They assert that, while BWE-based approaches have small cross-lingual resource requirements, approaches that attempt to model such shared space directly fail to obtain better results due to the differences in each language's linguistic properties. These differences lead to an imperfect alignment between the two embedding spaces which results in reduced model performance. Furthermore, they also state that translation-based approaches can leverage lexical information from the target language which complements the BWE approach. Thus, in their Bilingual-Word-Embedding-based Translation (BWET) model, they obtain BWE for the source and target languages but then use this shared space to perform word-by-word sentence translations via nearest neighbor search. Their EMD model is then trained on the translated target-language data. Furthermore, in order to account for word ordering, they propose incorporating self-attention (Vaswani et al., 2017) which allows their model to consider the most relevant context for each word in the sentence. Their architecture consists of a hierarchical Bi-LSTM-CRF model. A character-level Bi-LSTM is followed by a word-level Bi-LSTM that incorporates self-attention. Finally, a CRF layer makes the label predictions.

Another translation based approach is presented by Jain et al. (2019). They focus on so-called *medium-resource* languages that do not have large task-specific annotated datasets (EMD in their case) but for which there are off-the-shelf machine translation systems. As such, instead of performing word-to-word translations like previous approaches (Mayhew et al., 2017; Xie et al., 2018), they leverage Google Translate[2] to generate a target-language version of the annotated source-language data. Then labels are projected onto the translated data by matching the annotating entities with their corresponding translations. The matching process consists of several steps. First, they translate each annotated entity into the target language by itself. This is done because translation results vary depending on the context and there are instances in which the translation for an instance by itself is different from its translation within a full sentence. They also augment each entity's translation set using publicly-available bilingual dictionaries. In the next step, they perform token-level matching where each token in an entity's translation set is matched with a token in the translated target-language sentence. This matching is performed using an heuristic that incorporates orthographic (character affixes) and phonetic features (transliterations using the International Phonetic Alphabet). After token-level matching, they generate a list of potential entity spans by grouping adjacent tokens in the target sentence above a certain threshold. Afterwards, the best matching pair of entities is selected by greedily aligning each source entity with the span that has the least character edit distance. Source entities that are not aligned after the first three steps are annotated by constructing a set of top-k potential matches using their tf-idf scores where term frequency is calculated over all sentences that contain at least one unmatched entity and the inverse document frequency is computer over the entire dataset. The unmatched entity is aligned with the candidate with the highest score. Finally, a self-attention-assisted BiLSTM-CRF tagger is trained on the annotated target data.

3.3 Direct Transfer Cross-lingual EMD

Tsai et al. (2016) present an interesting approach in which they leverage Wikipedia as their sole multilingual resource. Their model depends on the existence of a cross-lingual wikifier. However, the wikifier only requires a multilingual Wikipedia section for the target language, no sentence or word alignments at all. Their core contribution is to make use of wikification (i.e., the process of linking an entity to its corresponding Wikipedia page) and entity linking and apply them to EMD. They use wikification to obtain language-independent features that provide useful information for EMD classification such as FreeBase (a now-deprecated knowledge base, succeeded by Wikidata (Vrandečić and

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

Krötzsch, 2014)[3]) types and Wikipedia categories. Their model also makes use of both non-lexical (e.g., previous tags) and lexical features (word form, capitalization, affixes, word type). Their approach obtained state-of-the-art performance at the time and did so without the requirement for parallel texts or interactions with a target-language native speaker. They also show that the obtained language-independent features are beneficial in for monolingual training as they improve the performance of monolingual models. Moreover, their approach is particularly interesting as wikification is traditionally considered a downstream task of EMD, i.e., entities are first identified and then linked to their respective Wikipedia pages.

Ni et al. (2017) instead propose a transfer-learning approach based on bilingual word embeddings (BWE). Their core idea is to project the monolingual embeddings (Mikolov et al., 2013c; Pennington et al., 2014; Bojanowski et al., 2017) from the source and target languages into a shared space to create a universal representation of the words. Such projection is guided by relatively small bilingual dictionaries (5K entries). Afterwards, their EMD model is trained using the labeled data from the source language and can be directly applied to the target language without having to re-train the model.

Wu and Dredze (2019) present one of the first efforts addressing the zero-shot, cross-lingual capabilities of pre-trained multilingual language models. They evaluate the performance of multilingual BERT (Devlin et al., 2019) in five different NLP tasks, including entity detection, under cross-lingual settings. They find that using mBERT as the encoder alongside simple, task-specific, neural-network architectures displays strong cross-lingual performance across all five tasks, in some cases even state-of-the-art performance for the time, without additional cross-lingual training signals. For entity detection in particular, they use a simple linear classification layer with softmax. Given that mBERT splits words into multiple sub-words, to perform the word-level predictions they utilize the representation of the first sub-word.

An extension of the previous work is proposed by Keung et al. (2019) where they introduce adversarial training which encourages the model to generate language-independent embeddings. The authors leverage unlabeled data in the target language by introducing a *language discriminator* which is trained to predict whether a sample sentence belongs to the source or the target languages. To force the encoder to generate embeddings that do not contain language-specific information, the authors include a *generator loss* that is only applied to the encoder parameters and works in the opposite direction of the *discriminator loss*. In their implementation, their EMD model follows Wu and Dredze (2019) and the language discriminator is a simple linear binary classifier.

In their work, Moon et al. (2019) do not propose a novel model architecture. Instead, their effort focuses on testing different training schemes for the usual mBERT + classifier model. Their experiments show that training a model with data from multiple source languages can be beneficial even if the languages used are not from the same language family or use the same script. They also experiment with multi-task learning, i.e., training the model with additional objectives to solve different tasks. However, their results with additional tasks, such as Language Identification or the Cloze task, do not show generalized improvements for every tested target language. Instead, some task/target-language pairs seem to be beneficial while others deteriorate the baseline performance.

Bari et al. (2020) propose a model that leverages two distinct BiLSTM-based encoders, one for each language. They argue that a separate encoders allow them to explicitly model specific characteristics, such as word order or morphology, of each language. These encoders are linked together by sharing character-level embeddings. They then learn a mapping between the source and target embedding spaces through word-level adversarial training. Furthermore, since the adversarially-learned mapping does not provide task-specific information, they propose a fine-tuning method where they jointly train the source and target encoders. This approach seems somewhat out of place as its method is fairly complex but reports lower performance than other previous efforts (Wu and Dredze, 2019; Keung et al., 2019) that leverage simpler model architectures.

A meta-learning-based approach for EMD is presented by Wu et al. (2020c). Though it can still be classified as a direct transfer method, the authors argue that source-trained models can be further improved if meta learning is used to learn good parameter initializations. Meta learning is split

[3] www.wikidata.org

into two phases: 1) meta training and 2) adaptation. During the meta-training phase the model is trained on a set of tasks so that it can quickly adapt to new tasks with only a small number of training examples. They simulate these tasks by leveraging the fact that, in the mBERT (Devlin et al., 2019) generated latent space, sentence representations that are close to each other display similar structural and/or semantic properties. Thus, for each source training example $x_i \in D^T_{train}$ a task $\mathcal{T}_i$ is defined by a pseudo testing set $D^{\mathcal{T}_i}_{test} = x_i$, and a pseudo training set $D^{\mathcal{T}_i}_{train}$ comprised by K of x_i most similar examples in the latent space. Then the model is trained on a randomly-sampled task $\mathcal{T}_i$ to minimize the loss computed on $D^{\mathcal{T}_i}_{train}$ (`inner update`) to obtain an updated set of parameters θ'. These updated parameters θ' are then evaluated on $D^{\mathcal{T}_i}_{test}$ and another update is made (`meta update`). During the adaptation phase, the model is applied on target languages. Here, each target-language test example $x_j \in D^T_{test}$ is used as the test set $D^{\mathcal{T}_j}_{test}$ for a target task $\mathcal{T}_j$. The task training set $D^{\mathcal{T}_j}_{train}$ is again obtained by retrieving the top-k similar examples of x_j from D^T_{train}. Once more, the model is first fine-tuned to minimize the error on $D^{\mathcal{T}_j}_{train}$ using a single gradient update and then used to predict the labels for x_j. A noteworthy observation from the authors is that, as entity-related words have considerably lower frequency than common words in the training corpus, their representations are not well-aligned across languages in the shared space. Thus, to address this issue they propose to randomly mask some entity tokens during the meta-training phase to encourage the model to make predictions using context information instead of relying on their representations. As for their tagging model, they use the same architecture as Wu and Dredze (2019): a linear classifier on top of mBERT.

3.4 Hybrid Transfer Cross-lingual EMD

Wu et al. (2020a) propose a teacher-student learning model to distill knowledge directly from single and multiple language sources. They propose to address the limitations of previous EMD approaches, both entity projection and direct transfer models. Mainly, they argue that (1) entity projection efforts require labeled data in the source language which may not be readily available and (2) direct transfer models do not leverage unlabeled data in the target language which is cheap to obtain and contains useful language information. As such, they propose to leverage previously trained EMD models for the source language as the teacher model. These teacher models must, nonetheless, be able to generate multilingual representations as they are then used to predict the label distributions (soft labels) for unlabeled data in the target language. Such distributions are then used to train a student model in the target language using the pseudo-labeled data obtained from the teacher model. They claim that their method does not rely on annotated data in the source language, however, it does indirectly depend on it as a core requirement is the existence of a previously trained EMD model to use as teacher. They also experiment with multi-source learning by leveraging several teacher models (trained on distinct source languages) at once. In order to do so, they propose a weighting scheme in which they leverage the language similarity (McClosky et al., 2010) between the target language and each corresponding source language.

The UniTrans model (Wu et al., 2020b) attempts to unify the model transfer and projection approaches. The authors argue that both approaches provide complementary information as the language-independent features used by direct-transfer models allow making predictions through contextual information while data-projection models benefit from word-label relations in the target language. Their approach consists of several steps. First, they create a pseudo training set in the target language by performing word-to-word translations and then projecting the labels directly from the annotated source data, similar to Mayhew et al. (2017). However, unlike Mayhew et al. (2017), their translations are not guided by a bilingual dictionary. Instead, they generate a dataset-specific seed dictionary by leveraging identical "character strings" (Smith et al., 2017) in both languages. Then, they learn a linear mapping between the multilingual embeddings of such identical character strings. To perform word-to-word translations, a source-word embedding is mapped into the target-language embedding space and its corresponding translation is obtained by nearest-neighbor search. A teacher EMD model is then trained using the annotated source data (Θ_{src}) and fine-tuned on the translated target data. In this manner, the teacher model (Θ_{teach}) is expected to obtain the advantages of both model transfer and data projection. Afterwards, they leverage a teacher-student learning setup similar to (Wu et al., 2020a): the teacher

model is applied to unlabeled target-language data and the generated label distributions are used to train a student model. This allows the student model(Θ_{stu}) to capture target-language-specific information and improve upon the teacher model. Additionally, the student model training is complemented by incorporating hard-label training. Since no ground-truth labels are available for the target-language data, the authors propose a voting scheme to generate pseudo hard labels. First, a new model (Θ_{trans}) is trained exclusively on the translated target data. Then, its predictions are compared with the predictions from (Θ_{src}) and (Θ_{teach}) models. A "hard label" is only generated if the predictions of such three models coincide. Finally, the student model (Θ_{stu}) is trained using the generated hard labels.

RIKD (Liang et al., 2021) introduces a reinforcement-learning-based approach that *smartly* selects instances to improve teacher-student knowledge transfer. Their teacher-student framework has a similar structure as Wu et al. (2020a) where the initial EMD teacher model leverages a multilingual encoder and is trained using annotated source-language data. A student model, with the same architecture, is then trained to mimic the probability distributions (soft labels) generated by the teacher model on unlabeled target-language data. The distinctive feature of their approach is that not all pseudo-labeled target-language examples are used to train the student model. Instead, they first perform a reinforcement-learning-guided selection of target-language examples to filter out noisy predictions from the teacher model. States, actions, and rewards for their reinforcement learning approach are modeled as follows: (1) The state of each target-language instance is modeled by a continuous real-valued vector. These *state vectors* are created from the concatenation of features such as the number of predicted entities, the length of the instance, and the inference loss of the source model on the instance. (2) Their action space is binary $a_i \in \{0, 1\}$ (to either select the example for training or discard it) and the policy network π is implemented by a two-layer linear network. (3) Delayed rewards are assigned using the improvement, or deterioration, between the training loss reported by the current and previous step models. Furthermore, as the student model outperforms the teacher thanks to the smart selection of training examples, the authors propose a bootstrapping-inspired scheme in which the student becomes a new teacher and the whole process is repeated for K iterations.

AdvPicker (Chen et al., 2021) improves upon the approach presented by Keung et al. (2019) by leveraging adversarial training and knowledge distillation in complementary ways. First, a teacher EMD model is trained on the source-language annotated data with adversarial training so as to encourage the encoder to produce language-independent token representations. It is relevant to point out that, while the approach proposed by (Keung et al., 2019) deals with sentence-level adversarial training (i.e., sentence-level representations are presented to the discriminator), the AdvPicker model deals with token-level adversarial training. Once the teacher model is trained, it is used on unlabeled target-language data to produce pseudo-labels. However, not all of these pseudo-labeled examples are utilized to train the student model. Instead, they are first passed through the language discriminator and only the most *language-independent* samples are selected. An example's *language independence* is measured by the discriminator's confidence in classifying it as coming from either the source or target languages. Examples that are hard for the discriminator to classify contain less language-specific information which is helpful for cross-lingual learning. Finally, the student model is trained on the selected target-language data using the soft-labels produced by the teacher model as ground-truth.

3.5 Performance Comparison

Table 1 presents the EMD performance of the works discussed in this section when tested on the commonly-used CoNLL-2002 (Tjong Kim Sang, 2002) and CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) datasets. Detailed information about these datasets can be found in Appendix B.

4 Event Extraction

Event Extraction (EE) task aims to obtain structure from text by answering *WH* questions related to events that are present in it, i.e. *What* happened? *Who* did it? *When* did it happen? *Where* did it happen? *Why* did it happen?, etc.

An *event* can be described as the occurrence of an activity or, in more general terms, as a change of state. Nonetheless, the concept of *what* is considered an event is domain dependent and context dependent as something that is admissible in one

	Target Language		
Model	**ES**	**NL**	**DE**
Tsai et al. (2016)	60.55	61.56	48.12
Ni et al. (2017)	65.10	65.40	58.50
Mayhew et al. (2017)	65.95	66.50	59.11
Xie et al. (2018)	72.37	71.25	57.76
Jain et al. (2019)	73.5	69.9	61.5
Bari et al. (2020)	75.93	74.61	65.24
Wu and Dredze (2019)	74.96	77.57	69.56
Keung et al. (2019)	74.3	77.6	71.9
Moon et al. (2019)	75.67	80.38	71.42
Wu et al. (2020c)	76.75	80.44	73.16
Wu et al. (2020a)	76.94	80.89	72.32
Wu et al. (2020b)	77.30	81.20	73.61
Liang et al. (2021)	77.84	82.46	75.48
Chen et al. (2021)	79.00	82.90	75.01

Table 1: EMD model performance comparison on the CoNLL-2002 & 2003 datasets. English is used as the source language.

domain might not be pertinent in a different one. As such, there are general domain datasets, e.g., ACE05 (Walker et al., 2006) and MAVEN (Wang et al., 2020), but there also are domain-specific datasets, such as BRAD (Lai et al., 2021) for historical events and SuicideED (Guzman-Nateras et al., 2022a) for suicide-related events, each with its own event definition and event-type categories.

Altogether, event extraction is a complex task which is why it is further divided into two main subtasks: Event Detection (ED) and Event Argument Extraction (EAE).

4.1 Event Detection

4.1.1 Task Definition

Event Detection (ED), IE's first main sub-task, consists in, first, selecting the words or phrases, commonly referred to as *triggers*, that denote the occurrence of events in a sentence. This first step is often referred to as *trigger identification*. In a second step, known as *trigger classification*, the event triggers are allocated into a discrete set of categories called *event types*. In the literature, the term event detection refers to performing both the identification and classification of the trigger words simultaneously (e.g. using sequence labeling). For example, in the sentence:

John recently bought a house.

an ED system should first identify the word ***bought*** as a candidate event trigger and then classify it as a `Transaction:Transfer-Ownership` event type[4].

As is the case for EMD, the cross-lingual setting brings with it additional complexities for a CLED model to tackle. For instance, event triggers are known to be frequently related to the verb a sentence (Majewska et al., 2021). In a cross-lingual setting, the target language could have verb tenses/conjugations that do not exist in the source language, or vice versa. Spanish, for example, has 18 distinct verb tenses while English only has 12 of them. Complications such as this one have nudged CLED research efforts to favor direct transfer approaches to take advantage of their language-agnostic training.

4.1.2 Data Transfer Cross-lingual ED

The only recent data-transfer-based method for CLED we could find is the work by Liu et al. (2019). They present an approach that aims at addressing the different-order problem of cross-lingual ED. Languages such as English and Chinese can have rather different word orders, however, they share similar syntactic structures. As such, in their approach, they first train monolingual word embeddings via the skip-gram model (Mikolov et al., 2013c) and then compute a context-depending lexical mapping for the source and target languages. In order to create such mapping, they first learn a multilingual alignment leveraging a small seed bilingual dictionary. Notably, the alignment parameters are not learned through training, instead a closed-form solution is computed using singular value decomposition (SVD). Next, a set of translation candidates is retrieved for each token in the sentence via the cross-domain similarity local scaling (CSLS) metric (Lample et al., 2018). Finally, a translation candidate is selected via a contextual self-attention mechanism (Vaswani et al., 2017). With this procedure, the authors obtain a translated version of the original sentence. The last step in their approach is to generate order-independent token representations which they achieve by feeding the syntactic tree of the sentence to a Graph Convolutional Neural Network (GCN, Kipf and Welling, 2017) where the initial node representations are set as the translated-word embeddings. Then their model is co-trained on both the source and target languages at the same time via cross-entropy loss. Their results show

[4]This type example is taken from the ACE05 dataset event types.

that their approach outperforms monolingual state-of-the-art models at the time by training on both the translated data from the source language and labeled data in the target language. Furthermore, the authors acknowledge that their approach depends on the availability of syntactic parsers for each language which could potentially affect its applicability.

4.1.3 Direct Transfer Cross-lingual ED

The work by Caselli and Ustun (2019) is probably the first to evaluate the generalization abilities of Multilingual BERT (mBERT, Devlin et al., 2019) for the ED task in a zero-shot cross-lingual setting. They do not report their performance on the ACE datasets and instead make use of the TempEval-3 corpus (UzZaman et al., 2013) for English, and the EVENTI dataset (Caselli et al., 2014) in Italian. Both of these datasets share the same annotation scheme and are annotated with only 7 event categories. Their straightforward approach consists of an mBERT-based encoder and a softmax classifier over each token. For multi-token words, they take the first token of each word to make the predictions. In their experiments, they found that their simple multilingual approach lagged behind its state-of-the-art monolingual counterparts but still achieved acceptable performance, specially when a minimal amount of target-language labeled data was introduced.

In a concurrent approach for the Cross-Lingual Event Detection (CLED) task, M'hamdi et al. (2019) also cast the task as a sequence-labeling problem and compare the performance of two different neural architectures harnessing distinct multilingual resources. In their first approach, they make use of the MUSE bilingual word embeddings (Conneau et al., 2017) alongside a bidirectional-LSTM encoder with a CRF (Lafferty et al., 2001) layer on top of a classifier linear layer. The second model shares the same classifier/CRF setup but instead leverages a pre-trained multilingual language model (mBERT) as the encoder. Their experiments exemplify the advantages of using contextualized word representations versus static word embeddings as the representations from mBERT greatly outperform the ones generated by the bi-LSTM on the CLED task.

Lu et al. (2020) present a cross-lingual structure transfer approach in which sentences are represented by language-universal structures: either dependency trees or fully connected graphs. The nodes of these structures are the multilingual embeddings (Lample et al., 2017) of the words in each sentence. The structure is then fed to an encoder which produces contextualized representations for the words in the sentence. They do not really do ED and instead tackle the simpler Event Trigger Labeling (ETL) task in which triggers are already identified and must only be classified. Each candidate trigger representation is passed through a linear layer followed by a Softmax transformation to predict its class. The dependency parsers for each language are manually trained using Treebanks (Nivre et al., 2016). They experiment with both Tree-LSTM (Tai et al., 2015) and Transformer-based (Vaswani et al., 2017) encoders and find the best model performance using a transformer encoder and a fully-connected graph structure. Their results also show that their model, trained exclusively on English data, achieves comparable performance on the target languages (Spanish, Russian, Ukrainian) as a supervised model trained on about 1,500 annotated sentences.

Another effort that addresses the CLED task via direct transfer is the work by Majewska et al. (2021). The key contribution of their work is incorporating external, language-specific, verb knowledge into the training process. The intuition behind their proposal is that, as verbs are prominently related to events in sentences, incorporating specific verb-processing information should be beneficial for event-related tasks. As such, they use VerbNet (Kipper et al., 2006) and FrameNet (Baker et al., 1998) as external knowledge sources and utilize dedicated adapter modules (Pfeiffer et al., 2020) to seamlessly incorporate the new knowledge while avoiding catastrophic forgetting during training. Verb-knowledge injection is performed through an intermediate binary classification task: using verb pairs, their model predicts if they belong to the same class (according to either VerbNet of FrameNet). They follow a similar architecture to M'hamdi et al. (2019) using an mBERT encoder with a CRF layer on top. They experiment with two training settings: full-model training, where the encoder's parameters are trained alongside the adapters; and adapter-only training, where they freeze the encoder's parameters. Their results show that their approach does improve performance over a zero-shot mBERT/CRF setting. However, their results on trigger detection and classification are below those reported by M'hamdi et al. (2019). This

could be due to the fact that their model concurrently performs both the ED and EAE tasks instead of following a training objective specifically designed for event detection.

Inspired by Du and Cardie (2020), which reframes the event extraction task as a question answering one, the authors of Fincke et al. (2021) present a language-agnostic method of incorporating task-specific information for cross-lingual event extraction. Their IE-PRIME approach consists in including augmented inputs for a pretrained multilingual transformer encoder so that it learns to generate task-specific word representations. For event detection, the priming is performed by concatenating each token from the sentence to the input as a candidate trigger. Their model then targets two training objectives from two different modules: (1) a BIO-label-based span prediction performed by a bi-LSTM with a CRF layer on top, and (2) an event type classification objective performed with a linear layer that takes as input the concatenation of the representations of the candidate trigger and `[CLS]` token. An important drawback of their approach is its efficiency as it must perform a forward pass for each word in the sentence.

The work by Nguyen et al. (2021b) proposes to refine the alignment of cross-lingual word representations by conditioning on class information and language-universal word categories. They argue that previous cross-lingual approaches suffer from monolingual bias as they are trained exclusively on source language data and that, even when leveraging unlabeled target data with adversarial training (Joty et al., 2017; Chen et al., 2018; Keung et al., 2019; He et al., 2020), a target language example from a class can be incorrectly aligned with source examples from a different class, thus hindering the model performance on downstream tasks. Their core intuition is that class information can be used to bridge the representation vectors between languages. As such, they obtain two representation vectors for each class: one from the source and one from the target language. The source class representations are computed as the average of the source examples belonging to each class. However, as the class information for target examples is unknown, the target class representations are obtained via a weighted aggregation of examples by estimating the probability that each example belongs to any of the classes. Then, during training, they encourage these two representations to be closer to each other which serves as a class-aware cross-lingual alignment mechanism. Additionally, they also propose to exploit dependency relations and universal parts of speech as language independent information that can further improve the learned representations. Similar to the class-aware alignment, they encourage the representations from words in the source and target languages that belong to the same part-of-speech category, or dependency relation, to be closer to each other. They test the performance of their CCAR model on three downstream tasks: ED, EAE, and RE. Their experiments show that their approach effectively effectively addresses the cross-class alignment issue which translates into improved task performance.

Extending on the work by M'hamdi et al. (2019), the authors of Guzman-Nateras et al. (2022b) use a similar architecture, a multilingual transformer encoder with a CRF-based classifier, but improve upon it in two main ways. First, make use Adversarial Language Adaptation (ALA) (Joty et al., 2017; Chen et al., 2018), a technique based on domain adaptation research (Ganin and Lempitsky, 2015) training, to generate language-invariant word representations that are not indicative of the language used for training but remain informative for the task at hand. ALA leverages unlabeled target-language data by introducing a language discriminator module that learns to discern between the source and target languages. Concurrently, the model encoder is trained in the *opposite* direction in an attempt to fool the discriminator. This *adversarial* training is what allows the encoder to generate language-invariant representations. Their main contribution is optimizing the adversarial training process by only presenting the discriminator with *informative* examples that force it to learn the fine-grained distinction between the languages which, in turn, improves the language-invariance of the encoder's representation. They base their notion of *what* makes an example informative on contextual semantics similarity and event presence likelihood. They propose Optimal Transport (OT) (Villani, 2008) as a natural solution to incorporate these two metrics into a single framework. Even though most of their performance improvements over previous work come from using a different encoder (XLM-RoBERTa, Conneau et al., 2020), their results show that their approach successfully generates refined language-invariant word representations that lead

to better CLED results and better handling of difficult cross-lingual instances.

4.1.4 Hybrid Transfer Cross-lingual ED

Similar to the previously-described work by Lu et al. (2020). The work by Muis et al. (2018) does not really address the ED task and focuses instead the simpler ETL task. Thus, they tackle event-type classification task with 11 categories that are referred to as *Situation Frames* (SF): issues or needs being described in text extracts. They compare two distinct approaches: (1) a keyword-matching system that leverages a small bilingual dictionary and (2) a neural-network-based model that generates bilingual word representations. In their keyword-based approach, they first build a list of keywords for each SF using the source language, and then translate such words into the target language with the bilingual dictionary. The keyword lists are generated in a two-step process: an initial candidate list is created by taking the top-100 words with the highest tf-idf scores for each class, and for each of these candidate words the 30 most similar words (based on word2vec Mikolov et al., 2013a cosine similarity) are added to the list. Then, for each candidate in the extended list, they compute a label-affinity score with the labels of each SF class using the cosine similarity between their embeddings. The final keyword set contains only those words whose label-affinity scores are above a threshold. For their neural-network-based approach, they first train bilingual word embeddings for the words in the source and target languages using XlingualEmb (Duong et al., 2016): a cross-lingual extension of word2vec. Then, they use a CNN encoder to generate contextualized word embeddings. These contextualized representations are then fed to a classifier that performs the prediction. However, they note that the bilingual word embeddings fail to capture the ground-truth mapping between the source and target languages, and propose to minimize this issue via standard ALA training. As these two approaches show similar performances, the authors also propose a data augmentation approach in which the keyword-based system is used to generate new training data to be used by the neural-network system. They found that they could considerably improve the performance of their neural network model using such bootstrapping approach.

4.1.5 Performance Comparison

Table 2 presents the CLED performance of the works discussed in this section when tested on the commonly-used ACE05 (Walker et al., 2006) and ACE05-ERE (Song et al., 2015) datasets. Detailed information about these datasets can be found in Appendix B.

	Target		
Model	**ZH**	**AR**	**ES**
Liu et al. (2019)	27.0	X	X
M'hamdi et al. (2019)	68.5	30.9	X
Lu et al. (2020)	X	X	41.77
Fincke et al. (2021)	X	51.0	X
Majewska et al. (2021)	46.9	29.3	X
Nguyen et al. (2021b)	72.1	42.7	X
Guzman-Nateras et al. (2022b)	74.64	46.86	47.69

Table 2: Model performance comparison on the ED for the ACE05 dataset. English is used as the source language.

4.2 Event Identification

Event Identification (EI), not to be confused with the aforementioned *trigger identification* step in the ED task, is a binary classification task for predicting whether or not an event is present in a text sample. As such, it is sometimes also referred to as Event Presence Prediction (EPP). EI is usually performed at the sentence level. For instance, the sentence:

John recently bought a house.

should be classified as containing an event (positive instance). Meanwhile, the prediction for the sentence:

John likes to eat pizza.

should be that it does not contain an event (negative instance).

Event identification is a simple, low-level task which is why there are not many research efforts that focus solely on it. Instead, EPP can be useful for other, higher-level tasks. Awasthy et al. (2020) show, for instance, that including an additional EI-based training signal can improve the performance of an event detection system. Although their work does not present a cross-lingual setting, they report monolingual settings for three languages showing that their approach is language agnostic.

A cross-lingual data-transfer effort focused on EI is presented by Hambardzumyan et al. (2020). The authors leverage Google's translation API to translate English and Arabic sections of the

ACE05 dataset into German to obtain a parallel corpus. They then train their encoder (multilingual BERT Devlin et al., 2019) to generate representations that are aligned (i.e., close to each other in the embedding space) for pairs of parallel sentences. Their intuition is that training the encoder in such a manner can help with zero-shot cross-lingual transfer for event presence prediction. Their results, however, show that while their approach does generate aligned sentence-level representations, using such aligned representations does not provide significant performance improvements.

4.3 Event Argument Extraction

4.3.1 Task Definition

The Event Argument Extraction (EAE) task consists in identifying the participants of an event (argument *identification*) and classifying them into a discrete set of categories called roles (Argument Role Labeling (ARL)). For example in the sentence:

John recently ***bought*** *a house.*

an EAE system should recognize the word *John* as a ***Buyer*** argument and the word *house* as the ***Object*** argument for the event denoted by the *bought* trigger.

The cross-lingual-associated adversities mentioned for EMD and ED apply for cross-lingual EAE as well: different word orders, distinct character sets, non-existing words, polysemous words, etc.

4.3.2 Direct Transfer Cross-Lingual EAE

Though not exclusive to the EAE task, Subburathinam et al. (2019) present an approach based on cross-lingual structure transfer. The key idea behind their work is to take advantage of the observation that some relational facts, such as the relationship between an event and its arguments, are expressed through identifiable patterns that display some consistency across languages. Hence, these patterns can be considered as language-universal features. They propose dependency trees as one of such language-independent features as similar event-argument relations in different languages share common dependency paths. As such, the first step in their approach is to convert sentences in both the source and target languages into language-universal dependency tree structures. Each node in the tree is represented by a vector made from the concatenation of each word's multilingual word embedding, POS embedding, entity-type embedding, and dependency-role embedding. Then, they leverage a Graph Convolutional Network (GCN, Kipf and Welling, 2017) encoder to obtain a contextualized representation for each node that takes into account information from the node's neighbors in the dependency tree. They train their EAE system using these language-independent representations using labeled data from the source language which can then be seamlessly applied to target-language data that has been encoded in a similar manner. For the EAE task, a full-sentence representation h^s is obtained by max-pooling over the representations of all nodes in the tree. Then, argument h^a and trigger h^t representations are obtained by max pooling over the representations of the nodes comprising the candidate argument a and the corresponding event trigger t. Their classifier is trained using the concatenation of these three vectors ($[h^t; h^s; h^a]$). In their experiments, they use the MUSE (Joulin et al., 2018) multilingual embeddings which are, in turn, obtained by aligning monolingual embeddings learned with FastText (Bojanowski et al., 2017) from Wikipedia; 17 universal POS tags and 27 dependency relations defined by the Universal Dependencies program (Nivre et al., 2016); and the seven entity types defined in the ACE05 dataset.

A very similar, though more straightforward, work is presented by Lu et al. (2020) who also propose to leverage language-universal structures such as dependency trees and fully connected graphs. In their approach, they feed these structures into a Tree-LSTM (Tai et al., 2015) or a Transformer (Vaswani et al., 2017) encoder to obtain contextualized representations for each word in a sentence. Then a concatenation of the representations of the event trigger and a candidate argument are passed through a linear layer and a softmax transformation to predict the argument's role.

The work by Majewska et al. (2021) (section 4.1 also addresses the EAE task. As a reminder, their approach integrates verb lexical knowledge into the training process as verbs and their arguments are commonly related to the events in a sentence. They do so by training dedicated adapters (Pfeiffer et al., 2020) to predict whether two verbs belong to the same class according to an external knowledge base. Then, these pre-trained *verb adapters* are integrated into their model when fine-tuning for the downstream EAE task. Though their experiments show an improvement when the verb adapters are

used, their reported results for EAE are well below other contemporary efforts.

In Nguyen and Nguyen (2021), the authors propose to incorporate language-independent knowledge to improve transfer learning for cross-lingual EAE. They utilize 3 distinct sources of information: syntax based, semantic based, and relation based. For syntax information, they use the adjacency matrix obtained from the sentence dependency tree. The semantic information is a similar matrix whose values are obtained by learning a semantic-similarity score between the multilingual representation vectors of pairs of words in the sentence. Such multilingual representation vectors are obtained through the concatenation of a word's MUSE embedding, POS tag embedding, entity type embedding, and dependency-relation embedding. Finally, relation-based information is incorporated by creating another matrix whose values are learned using embedding vectors for each dependency relation between a word and its governor. These three matrices are then linearly combined and passed through a GCN to obtain the final representation for each word in the sentence which is then used to predict the distribution over all possible argument roles. Their results show that incorporating these additional sources of information leads to better cross-lingual EAE performance as it allows their model to assign more nuanced importance scores to each word in the sentence with respect to the event trigger.

Ahmad et al. (2021) present the Graph Attention Transformer Encoder (GATE) model that, similar to previous works, leverages universal dependency parses to capture long-rage dependencies and mitigate the word-order difference issue in cross-lingual transfer. However, unlike the efforts by Subburathinam et al. (2019) and Nguyen and Nguyen (2021), they use self-attention mechanisms (Vaswani et al., 2017), instead of GCNs, to encode the dependency trees as GCNs tend to perform poorly in capturing long-distance dependencies and disconnected words in the tree (Zhang et al., 2019; Tang et al., 2020a). Their key idea is to allow attention between inter-connected words in the dependency tree and aggregate information across layers. Furthermore, they propose a revision of the self-attention mechanism in order to incorporate syntactic structure and distances into the computation. They use a non-parameterized function to modify the attention weights that, in essence, divides each of them by the syntactic distance between the related tokens as computed from the universal dependency parse. When encoding the input sentences, they first utilize multilingual pre-trained language models (mBERT, XLM-RoBERTa) to obtain contextualized word embeddings which are then concatenated with POS tag embeddings, dependency-relation embeddings, and entity-type embeddings, similar to the approach by Nguyen and Nguyen (2021). To perform classification, they generate fixed-length vectors for the candidate argument e_a, the event trigger e_t and the full sentence s, each of which is obtained by max-pooling over their respective set of contextual representations. Afterwards, the concatenation of these three vectors $[e_t; e_a; s]$ is fed to a linear classifier that predicts the role label.

As discussed in detail in section 4.1, the IE-PRIME model (Fincke et al., 2021) leverages *model priming*: augmenting a model's input with task-specific information. For argument extraction, IE-PRIME augments the input in two distinct ways: (1) by pre-pending the trigger to the input sentence and (2) by also pre-pending one of the argument roles associated with the trigger event type. The argument roles are codified as integer numbers to keep their system language agnostic. This second approach obtains better EAE performance, however, it has the considerable drawback of requiring one forward pass for each possible argument role.

The CCCAR model (Nguyen et al., 2021b) seeks to improve cross-lingual representation learning by conditioning on class information and universal word categories such as POS and dependency relations. Section 4.1 provides further details on the model.

Huang et al. (2022b) present their X-GEAR model that leverages generative models to perform cross-lingual EAE, instead of the more commonly used classification-based models such as CL-GCN (Subburathinam et al., 2019) and GATE (Ahmad et al., 2021). Their key idea is to fine-tune a pretrained multilingual generative language model such as mBART (Tang et al., 2020b) or mT5 (Xue et al., 2021) with training samples where the input has been augmented with a template. Their proposal entails two main challenges: (1) in the cross-lingual setting, the input language changes during training and testing, and (2) the generated outputs must be parsed into final predictions. To address these challenges they design *language-agnostic*

templates. A template includes the event trigger and all possible argument roles associated with the corresponding event type, encoded as special tokens, with the appropriate arguments. By formatting the templates in such a manner, the event type information does not need to be explicitly included as such information is implicitly included. Furthermore, by using special tokens to represent the argument roles, the templates are completely language agnostic. Their model is then trained to generate output strings that conform to the template format. The inputs to their model are composed by the original passages and a *prompt* that includes the event trigger and the type-specific template. In these input templates, each argument role is filled with a special `[None]` token that is to be replaced by the generative model. For their experiments on the ACE05 and ACE05-ERE datasets, they compare against their own implementations of CL-GCN and GATE and found that that their approach outperforms these classification-based cross-lingual EAE models, and even other generative models that use language-dependent templates such as TANL (Paolini et al., 2021).

4.3.3 Hybrid Transfer Cross-Lingual EAE

Ahmad et al. (2019) present a hybrid multilingual effort for EAE. The core of their approach is to learn a mapping between monolingual word embeddings obtained with fastText (Bojanowski et al., 2017) via adversarial language adaptation. Then, they use a hybrid CNN-LSTM encoder to obtain the representation of each word in a sentence. These representations are then passed to a feed-forward network to obtain a shared representation for the EAE task. Afterwards, they propose adding a separate language layer for each language they consider (English, Hindi, and Bengali). Each of these language layers is only trained when the input data matches their corresponding language. Finally, after each language layer, they use six independent fully connected layers, one for each argument type, for a total of 18. The reasoning behind this decision is that argument types are not mutually exclusive and, consequently, a single word could display multiple roles simultaneously. For their experiments, they use their own human-annotated dataset crawled from popular news websites in each language. Their results show that multi-lingual training generally improves their model's performance for argument types with fewer training examples. However, they also notice that it can deteriorate the performance of types with lots of training examples in which the monolingual models perform better. Though they focus their experiments on a domain-specific dataset, their approach can be readily applied to any domain.

4.3.4 Performance Comparison

Table 3 presents a comparison between the cross-lingual EAE performance of the works discussed in this section when tested on the commonly-used ACE05 (Walker et al., 2006) and ACE05-ERE (Song et al., 2015) datasets. Detailed information about these datasets can be found in Appendix B.

	Target		
Model	**ZH**	**AR**	**ES**
Subburathinam et al. (2019)	59.0	61.8	X
Lu et al. (2020)	X	X	17.35
Majewska et al. (2021)	1.9	7.1	X
Nguyen and Nguyen (2021)	58.4	62.9	X
Ahmad et al. (2021)	63.2	68.5	X
Fincke et al. (2021)	X	74.7	X
Nguyen et al. (2021b)	65.5	69.4	X
Huang et al. (2022b)	54.0	44.8	59.7

Table 3: Model performance comparison on the EAE for the ACE05 dataset. English is used as the source language.

5 Relation Extraction

5.1 Task Definition

Relation Extraction (RE) is the task of identifying and classifying the semantic relations that exist between entities (organizations, persons, locations, events) in a text sample. For example, in the sentence:

John was born in Eugene, Oregon.

an RE system would predict that the entities *John* and *Eugene* participate in a ***bornInCity*** type relation and that *Eugene* and *Oregon* participate in a ***locatedIn*** type relation. Relation extraction is a useful task for other higher-level task such as question answering, text summarization, text mining, and knowledge base population.

As is the case with other tasks, traditional RE models relied on feature engineering by combining syntactic, lexical, and semantic features (Zelenko et al., 2003; Kambhatla, 2004; Li and Ji, 2014). These methods were later replaced by approaches that make use of deep neural networks trained in a

supervised manner (Zeng et al., 2014; dos Santos et al., 2015; Nguyen and Grishman, 2015; Miwa and Bansal, 2016; Wang et al., 2016). Regarding cross-lingual efforts for RE, over the past decade the have been approaches based on active learning (Qian et al., 2014), knowledge bases (Verga et al., 2016), and bilingual representations learned through language independent concepts (Min et al., 2017).

5.2 Data Transfer Cross-lingual RE

Earlier methods for cross-lingual RE relied the data transfer paradigm and were based on annotation projection using either parallel corpora (Kim et al., 2014) or pseudo-parallel corpora obtained via machine translation (Faruqui and Kumar, 2015).

5.3 Direct Transfer Cross-lingual RE

Ni and Florian (2019) propose an approach that relies on embedding projections instead of parallel corpora or machine translation. Their approach consists in, first, generating monolingual Word2Vec (Mikolov et al., 2013c) word embeddings for both the source and target languages and, then, learning a linear mapping between the two latent spaces by minimizing the mean squared error between the representation vectors of aligned word pairs obtained from a small (1K words) bilingual dictionary. Their model has four main layers. An embedding layer maps every word in an input sentence to its corresponding monolingual vector representation. They also make use of entity-label embeddings: randomly initialized, real-valued vectors to represent entity types. Next, a context layer whose purpose is to create context-aware representations for each word in the sentence. Here, they experiment with both LSTM-based and CNN-based context encoders. A summarization layer that generates a single fixed-length vector to be used for classification purposes. They perform element-wise max pooling among the context-aware vectors of all words that appear before the first entity, the words that comprise the first entity, the words in-between the first and second entity, the words comprising the second entity, and the words appearing after the second entity. Then, these five vectors are concatenated into a single vector that is used as the input for the output layer. Finally, the output layer returns a probability distribution over the set of relation types. To perform cross-lingual classification, the sentence word embeddings in the target language are projected into the source language embedding space using the learned linear mapping and the model is applied normally on the projected embeddings. The authors mention that they specifically do not use language-specific resources such as dependency parsers as their availability cannot be guaranteed for low-resource target languages. They experiment with both an *in-house* dataset with six languages and 56 entity types and ACE05 dataset that has seven entity types. Their monolingual results on source data (English) lag behind the state-of-the-art ensemble model VOTE-BW (Nguyen and Grishman, 2015). Their performance on cross-lingual RE also seems to be lacking with respect to the previously released CNN-GAN (Zou et al., 2018) as their reported F1 scores on the En-Zh language pair 20% lower. However, this might be due to the use of a distinct data split from previous works. From their experiments they also recognize that their approach works best with languages that share the same syntactic structure as the source language. In the case of English, for example, languages such as German, Spanish, Italian, and Portuguese that follow the same SVO (subject, verb, object) structure perform considerably better than, for instance, Japanese which has an SOV convention. While the performance reported in this work seems to be lacking, it has several characteristics that work in its favor such as its simplicity and its general applicability due to its low requirements of cross-lingual resources.

The work by Subburathinam et al. (2019), described in greater detail in section 4.3 for the EAE task, also addresses the RE task. For relation extraction, the authors train a classification layer using a concatenation of the representations of each entity in the relation pair under consideration, h^{m_1} and h^{m_2}, with the full sentence representation h^s. Recall that, in their approach, these representations are obtained by max-pooling over the language-universal representations obtained by a GCN-based encoder of the nodes in a dependency tree.

The authors of Köksal and Özgür (2020) present the first transformer-based approach for the cross-lingual RE task. Their model leverages a multilingual pretrained transformer (mBERT Devlin et al., 2019) as its encoder which is then pretrained on a proxy task via distant supervision. To this end, they collect a large number of sentences from Wikipedia in several languages with entities marked by hyperlinks. Afterwards, sentences including entity pairs with Wikidata relations (Vrandečić and Krötzsch,

2014) are selected. They generate positive samples by selecting pairs of sentences that share the same entities and relation type in two distinct languages. Negative examples are created by selecting sentences that share one entity but that do not belong to the same relation type. Then, mBERT is trained on the binary classification task of predicting whether the two sentences in a pair show the same relation or not. Furthermore, in the collected sentence pairs, the entities are replaced by a special token `[BLANK]` with a fixed probability, so that mBERT learns to capture text patterns instead of memorizing the entities. In essence, they finetune mBERT using the standard masked-language modeling objective and their matching the multilingual blanks (MTMB) objective – a multilingual version of the approach proposed by Baldini Soares et al. (2019). They publicly release the two new cross-lingual RE datasets used in their experiments: RELX and RELX-Distant. In their experiments, the authors compare a baseline model – a standard mBERT encoder with a classification layer on top – with their proposed with their version that pretrained using MTMB and find that the pre-training improves cross-lingual RE performance by as much as 4.5% in some languages (Spanish). In additional experiments, they show that their approach greatly outperforms the baseline in low-resource settings. In Spanish, for instance, the MTMB-trained model achieves the same performance as a vanilla mBERT model using only around 20% of the training data. Unfortunately, their results are not directly comparable with other previous efforts as they only report their performance on their proposed datasets.

The GATE model (Ahmad et al., 2021), described in detail in section 4.3 also addressed the RE task. Similar to their EAE approach, for RE they obtain fixed-length representations for the full sentence s, and each entity in an entity pair (e_s, e_o) by performing a max-pooling over their contextualized word representations. Then, a concatenation of these vectors $[e_s; e_o; s]$ is passed through a linear classifier that outputs the predicted relation types (if any). Their RE classifier is trained with the standard cross-entropy loss.

The authors of Nguyen et al. (2021b) also test the performance of their CCAR model on the RE task. As mentioned in section 4.1, their intuition is to improve the alignment of cross-lingual representations by conditioning on language-invariant information: class information, POS category, and dependency relation.

5.4 Hybrid Transfer Cross-lingual RE

In their work, Zou et al. (2018) propose utilizing two twin encoder networks – for source and target languages – that learn to extract language-invariant features that remain indicative of relation information but not of originating language. They obtain pseudo-parallel target-language sentences by leveraging Google's machine translation API[5]. Then they transform both the original and translated sentences into vector sequences by utilizing bilingual word embeddings (Shi et al., 2015) alongside randomly-initialized positional and entity-type embeddings. These sequences are then used as the input for the twin encoder networks. Their encoders output a single vector which is then fed into a discriminator network tasked with identifying the originating language. During training the source-language representations are also fed to a classifier network that predicts the relation contained in the sample. The target encoder is trained in an adversarial manner in an attempt to fool the discriminator. As such, as the source encoder learns to generate representations that are informative for the relation extraction task, the target encoder learns to generate similar features stemming from target-language samples which should share the aforementioned informative qualities. At testing time, target-language samples are fed into the corresponding encoder and its output is passed to the classifier. In their experiments, they explore both CNN-based and LSTM-based encoder networks with CNNs coming slightly on top. They compare their model performance against the state-of-the-art model at the time BI-AL (Qian et al., 2014) which they substantially improve upon ($\sim 4\%$improvement). Supplemental experiments also show that their unsupervised approach outperforms a supervised model when the size of the available labeled training data is small (< 700 samples) and that their model is able to make effective use of the available source training data as training with limited amounts – only 10% of the data, for instance – led to small performance declines ($\sim 6\%$) compared to the BI-AL baseline ($\sim 20\%$).

5.5 Performance Comparison

Table 4 presents a comparison between the cross-lingual RE performance of the works discussed in

[5]https://translate.google.com/

	Target	
Model	**ZH**	**AR**
Zou et al. (2018)	68.4	X
Subburathinam et al. (2019)	42.5	58.7
Ni and Florian (2019)	46.8	36.4
Ahmad et al. (2021)	55.1	66.8
Nguyen et al. (2021b)	58.1	67.9

Table 4: Model performance on the RE for the ACE05 dataset. English is used as the source language.

this section when tested on the commonly-used ACE05 (Walker et al., 2006) dataset. Detailed information about this dataset can be found in Appendix B.

6 Co-Reference Resolution

6.1 Task Definition

A *co-reference* occurs when there are several expressions (*mentions*) in a text sample that mention the same entity. For example, in the sentence:

John said he did not got to the party.

the words "***John***" and "***he***" refer to the same person.

The definition of an entity in the context of this task is different from that of the EMD task as it has a broader interpretation: it includes persons, things, organizations, but it can also involve events, concepts, or other intangible abstractions. For example, in the sentence:

This year there wasn't much inflation, but it will get much worse.

the words "***inflation***" and "***it***" should be identified as referring to the same entity even thought such entity is just a concept. A Co-Reference Resolution (CRR) system should then be able to identify any co-references that occur in a text sample.

Systems that use entity-related features to make mention-wise linking decisions are called *entity-mention*. Whereas, *mention-pair* models use only local information to determine mention co-reference (Cruz et al., 2018).

6.2 Data Transfer Cross-Lingual CRR

Cross-lingual data-transfer-based approaches for CRR are limited to a couple of shared tasks (Ji et al., 2015) and are primordially based on annotation projection.

6.3 Direct Transfer Cross-Lingual CRR

For the purposes of this survey, we focus on direct-transfer-based CRR efforts which has been the favored approach in recent years.

Kundu et al. (2018) propose an entity-mention approach that gradually merges the mentions in a document to produce entities leveraging a zero-shot Entity Linking system (Sil and Florian, 2016). They train their own monolingual word-embeddings for the source and target languages and then build a cross-lingual embedding space following Mikolov et al. (2013b). Their system receives entity pairs (not mention pairs) as inputs. Since each entity represents a set of mentions, the entity-pair embedding is obtained from the embeddings of mention pairs produced using the cross product of the entity pairs. Then, for each mention pair in the cross product a set of features is computed and embedded as a real-numbered vector. Among the features they use are: string matching, word/sentence distance between mentions, mention types, entity types, and whether one mention is an acronym of the other. The embedded features are concatenated with the average of the mentions' word embeddings and passed through an attention layer before the classifier.

Cruz et al. (2018) present instead a mention-pair approach in which they leverage a large coreferentially-annotated Spanish corpora (Recasens and Marti, 2010) to create a cross-lingual model for the lower-resourced Portuguese (Fonseca et al., 2017) language. In their approach, they leverage FastText (Bojanowski et al., 2017) multilingual embeddings along with language-agnostic features such as the sentence/word distance between mentions. The mentions' word-level embeddings are combined by either non-parametric methods (e.g., summation) or using neural encoders (CNNs, LSTMs, dense layers) and then concatenated with the distance features before being passed to a dense-layer-based binary classifier network.

Urbizu et al. (2019) work on a CRR for the Basque language. Being a language spoken only on specific regions of Spain and France, Basque is a low-resource language for which not many monolingual CRR efforts exist (Soraluze et al., 2016, 2017). The authors explore leveraging a large English corpora (OntoNotes) to create a cross-lingual Basque model given that the largest CRR corpora for Basque (Cerberio et al., 2018) is insufficient to effectively train a monolingual neural model. They

use a straightforward neural model comprised of three dense layers (500, 300, and 100 neurons, respectively) with ReLU activations. As inputs, they utilize FastText multilingual embeddings and complemented by a few independent features such as the distance in words between mentions, the distance in mentions between the mentions, whether or not the mentions are in the same sentence, and string matching. They report improved CRR results from their cross-lingual model compared to those of a monolingual model trained in a supervised manner with the Basque corpus. These results assert the usefulness of a CLL approach when target language resources are limited resources. In cases such as Basque, the smaller-sized annotated Basque corpora can be leveraged to fine-tune the cross-lingually trained model.

Phung et al. (2021) present the first cross-lingual effort focused on Event Co-Reference Resolution (ECR). Event co-reference resolution is considered a more challenging task than entity co-reference resolution because of the more complex structures of event mentions (Yang et al., 2015). They cast the ECR problem as a binary classification task where their model receives as input a sequence of words that contains two event mentions and aims at determining whether the two mentions refer to the same event or not. Being the first work on this problem, they first establish a baseline model that uses a multilingual transformer (XLM-RoBERTa, Conneau et al., 2020) as the encoder and augment the input sequence with two special tokens (<e></e>) that are used to identify the location of event triggers. To predict the co-reference, they use the concatenated representations of the special tokens surrounding both triggers as the input for their classifier. Then they propose three main improvements upon their baseline. First, the use of adversarial training (Ganin and Lempitsky, 2015) to improve the language-invariance properties of the representations generated by the encoder. Second, they argue that, given the lack of co-reference labels for pairs of event mentions in the target languages, the discriminator can potentially align co-referential with non co-referential examples. To address this issue, they propose to generate two separate representation vectors for each example for both the source and target languages via two independent neural networks. Then, the target-language representations are regularized to be similar with each other while the source-language representations are regularized to be different from each other. These two opposing regularizations help penalize unexpected alignments as they implicitly inject into the loss function the difference between source and target examples with different co-reference labels.

6.4 Performance Comparison

The research efforts discussed in for this sub-task address different languages or even have distinct focus (e.g., entity co-reference vs event co-reference). As such, they do not evaluate their results using common dataset and cannot be directly compared.

7 Future Research Directions

This section presents a number of promising research directions for future cross-lingual information extraction efforts.

7.1 Lexical/syntactic target-language information integration

The motivation behind the vast majority of cross-lingual works is to provide low-resource target languages with NLP tools that could not be created otherwise due to the lack of annotated data. In turn, cross-lingual approaches usually refrain from leveraging potentially-useful information from lower-level tasks, such as Part-of-Speech (POS) tagging or dependency parsing, under the assumption that these tools are not available for the target language.

However, as cross-lingual research gains traction and public interest, there are more tools available for an increasing amount target languages. For instance, Google's translation API [6] supports 133 languages at various levels and tool-kits such as Trankit (Nguyen et al., 2021a) provide fundamental NLP tasks for over 100 languages. Thus, research efforts focusing on these *medium resource* languages (Jain et al., 2019) can benefit from incorporating target-language lexical/syntactic information derived from such lower-level features.

7.2 Meta-learning/Few-shot learning

In standard supervised training tasks, models are trained on large quantities of data with the expectation that they will learn to generalize and work adequately on unseen samples. On the contrary, Few-Shot Learning (FSL) is a setting where a model is trained using very limited amounts of data. For this reason, FSL models cannot be trained in the traditional supervised setting as the limited availability

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

of training data leads to poor generalization. This training-data limitation is something FSL shares with CLL where target-language data is scarce.

Few-shot training is performed via *episodes* (Vinyals et al., 2016). An episode is constructed by sampling a subset out of the entire set of training classes and selecting a few examples belonging to such classes. In this sense, training is performed in *N-way, K-shot* settings where N refers to the number of classes and K refers to the number of examples for each class (K is usually low in the $[1-10]$ range). The $N \times K$ samples that compose an episode are called the ***support set***. Additionally, there are further examples belonging to the same classes that are used to evaluate the performance of the model while training, these are called the ***query set***. When the model is done training, at testing time, new episodes are constructed using samples from entirely different classes never seen during training. The model is then evaluated on its performance on the episode's query set based on the knowledge of its support set.

As such, FSL can be thought of as a type of *Meta Learning* where the purpose is to teach a model to *learn how to learn*. Meta-learning-based approaches have already been proven successful in cross-lingual IE tasks like EMD (Wu et al., 2020c) and could make a significant impact in CLL given their capability of learning from just a few labeled examples which can be easily obtained, even for the most obscure target languages.

7.3 Generative/Prompting Models

With the recent advancements in generative language models like BART (Lewis et al., 2020), GPT-3 (Brown et al., 2020), or T5 (Raffel et al., 2019), several NLP tasks have been formulated as text-generation tasks in monolingual settings. Information extraction tasks have not been the exception and generative-based approaches have been proposed for relation extraction (Paolini et al., 2021), argument extraction (Li et al., 2021), and end-to-end event extraction (Lu et al., 2021; Hsu et al., 2022). These approaches have since shown remarkable performances that are competitive or even better than the state-of-the-art traditional efforts. Given that some of these models already have multilingual versions (e.g., mBART, mT5), cross-lingual variants of such approaches have already started to appear. For instance, Huang et al. (2022a) formulate EAE as a generative prompt-filling task. They design *language-agnostic templates* that represent the event argument structures and leverage pre-trained multilingual generative language models to generate sentences that fill such templates.

Another way in which generative models can be exploited for IE tasks is to generate, or augment, the existing annotated datasets. Efforts like the one by Pouran Ben Veyseh et al. (2021) have already shown the value of this approach for tasks like event detection. This approach could be particularly useful in cross-lingual settings where annotated target-language data scarcity is usually assumed.

7.4 Multimodality

Leveraging non-textual sources of information could help improve the performance of zero-shot cross-lingual models. Images can be regarded as language-independent so, for instance, visual features extracted from pictures of recognizable entities could be integrated into a cross-lingual model and be beneficial for entity mention detection.

Furthermore, the recently released Contrastive Language-Image Pre-training model (CLIP Radford et al., 2021) from OpenAI provides a bridge between text and images and offers an unprecedented opportunity to link these two, usually separate, domains. Image-generation models that make use of CLIP's capabilities such as Dall-E (Ramesh et al., 2021) and Dall-E2 (Ramesh et al., 2022) are already being used by artists, researchers, and the general public to generate high-quality realistic images from textual descriptions. Their public release and widespread use could foster the creation of hybrid text-image datasets for cross-lingual IE tasks such as event extraction or coreference resolution. There already have been efforts at creating a multilingual version of CLIP by re-training its textual encoder for various non-English languages (Carlsson et al., 2022).

7.5 Robust Training

Robust training aims at creating models that are not affected by noise or perturbations in the input data (Goodfellow et al., 2015). Robust models are created as a means to defend against adversarial attacks which are input samples with small perturbations designed to fool classifiers into making wrong predictions:

$$c(\tilde{x}) = c(x + \Delta) \neq c(x)$$

where c is a classifier, Δ is a small perturbation, and $\tilde{x}$ is a perturbed sample.

Multilingual encoders such as mBERT or XLM-R have a shared embedding space for words in different languages (Wu and Dredze, 2019). In such space, the representations of similar words are close to each other, e.g., the representations for the word *cat* and its Spanish equivalent (*gato*) should be similar. These representations, however, are not completely aligned. In this sense, the differences between the representations of the same word in the source and target languages can be considered as perturbations, similar to that of an adversarial example. Thus, cross-lingual learning can be approached as a robustness perspective.

For instance, Huang et al. (2021) propose the idea of treating cross-lingual transfer as a representation-alignment issue. It is their intuition that by training a cross-lingual model to be robust against such perturbations, the model becomes able to better transfer the learned knowledge from one language to the other. They explore two robust training methods: *adversarial training* and *randomized smoothing*. In this context, adversarial training means considering the most effective adversarial perturbation at each iteration, i.e., the perturbation that is most likely to change the prediction, while at the same time ensuring the model remains able to classify it correctly. On the other hand, randomized smoothing focuses on expectation and uses random perturbations instead. They evaluate their training scheme on two cross-lingual classification tasks: paraphrase detection and Natural Language Inference (NLI). In their experiments, they found that randomized smoothing usually leads to better performance than adversarial training. They argue that the reason behind such behavior is that, even though adversarial training is more suitable to defend against examples specifically designed to attack the classifier, for cross-lingual knowledge transfer the average of randomized perturbations better reflects the difference between languages.

References

Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Gate: Graph attention transformer encoder for cross-lingual relation and event extraction. In *AAAI*.

Zishan Ahmad, Deeksha Varshney, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-linguality helps: Event-argument extraction for disaster domain in cross-lingual and multi-lingual setting. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 135–142, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Parul Awasthy, Tahira Naseem, Jian Ni, Taesun Moon, and Radu Florian. 2020. Event presence prediction helps trigger detection across languages. In *CoRR*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *CoRR*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

M Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. 2020. Zero-resource cross-lingual named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2003. A simple named entity extractor using adaboost. In *Proceedings of the Seventh Conference on Natural Language Learning (CONLL)*.

Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. Eventi evaluation of events and temporal information at evalita 2014.

Tommaso Caselli and A. Ustun. 2019. There and back again: Cross-lingual transfer learning for event detection. In *CLiC-it*.

Klara Cerberio, Itziar Aduriz, Arantza Diaz de Ilarraza, and Ines Garcia-Azkoaga. 2018. Coreferential relations in basque: The annotation process. In *Journal of Psycholinguistic Research*.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295, Brussels, Belgium. Association for Computational Linguistics.

Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje Karlsson, and Yi Guan. 2021. AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER. 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 743–753, Online. Association for Computational Linguistics.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270, Brussels, Belgium. Association for Computational Linguistics.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification. In *Transactions of the 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*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. In *CoRR*.

André Ferreira Cruz, Gil Rocha, and Henrique Lopes Cardoso. 2018. Exploring spanish corpora for portuguese coreference resolution. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Cícero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. 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 626–634, Beijing, China. Association for Computational Linguistics.

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*.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

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

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

pages 118–124, Hissar, Bulgaria. Association for Computational Linguistics.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual open relation extraction using cross-lingual projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1356, Denver, Colorado. Association for Computational Linguistics.

Xiaocheng Feng, Xiachong Feng, Bing Qin, Zhangyin Feng, and Ting Liu. 2018. Improving low resource named entity recognition using cross-lingual knowledge transfer. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization.

Steven Fincke, Shantanu Agarwal, Scott Miller, and Elizabeth Boschee. 2021. Language model priming for cross-lingual event extraction. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Evandro Fonseca, Vinicius Sesti, Sandra Collovini, Renata Vieira, Ana Leal, and Paulo Quaresma. 2017. Collective elaboration of a coreference annotated corpus for portuguese texts.

Ruiji Fu, Bing Qin, and Ting Liu. 2014. Generating chinese named entity data from parallel corpora. In *Frontiers of Computer Science*, volume 8, pages 629–641.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations (ICLR)*.

Luis Guzman-Nateras, Viet Lai, Amir Pouran Ben Veyseh, Franck Dernoncourt, and Thien Huu Nguyen. 2022a. Event detection for suicide understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Luis Guzman-Nateras, Minh Nguyen, and Thien Nguyen. 2022b. Cross-lingual event detection via optimized adversarial learning. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Luis Guzman-Nateras, Minh Van Nguyen, and Thien Nguyen. 2022c. Cross-lingual event detection via optimized adversarial training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5588–5599, Seattle, United States. Association for Computational Linguistics.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2020. The role of alignment of multilingual contextualized embeddings in zero-shot cross-lingual transfer for event extraction. In *Collaborative Technologies and Data Science in Smart City Applications (CODASSCA)*.

Keqing He, Yuanmeng Yan, and Weiran Xu. 2020. Adversarial cross-lingual transfer learning for slot tagging of low-resource languages. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

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.

Kuan-Hao Huang, Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training.

Kuan-Hao Huang, I-Hung Hsu, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022a. 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, Premkumar Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022b. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *ACL*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. In *Natural Language Engineering*.

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity projection via machine translation for cross-lingual NER. 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 1083–1092, Hong Kong, China. Association for Computational Linguistics.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of tac-kbp2015 tri-lingual entity discovery and linking. In *Theory and Applications of Categories*.

Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language learning with adversarial neural networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL)*, pages 226–237.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. 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 1355–1360, Hong Kong, China. Association for Computational Linguistics.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2014. Cross-lingual annotation projection for weakly-supervised relation extraction. In *ACM Transactions on Asian Language Information Processing*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Abdullatif Köksal and Arzucan Özgür. 2020. The RELX dataset and matching the multilingual blanks for cross-lingual relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 340–350, Online. Association for Computational Linguistics.

Mikhail Kozhevnikov and Ivan Titov. 2014. Cross-lingual model transfer using feature representation projection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–585, Baltimore, Maryland. Association for Computational Linguistics.

Gourab Kundu, Avi Sil, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual coreference resolution and its application to entity linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 395–400, Melbourne, Australia. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

Viet Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021. Event extraction from historical texts: A new dataset for black rebellions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

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

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. In *CoRR*.

Anna Langedijk, Verna Dankers, Phillip Lippe, Sander Bos, Bryan Cardenas Guevara, Helen Yannakoudakis, and Ekaterina Shutova. 2022. Meta-learning for fast cross-lingual adaptation in dependency parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8503–8520, Dublin, Ireland. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–412, Baltimore, Maryland. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Shining Liang, Ming Gong, Jian Pei, Linjun Shou, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. 2021. Reinforced iterative knowledge distillation for cross-lingual named entity recognition. In *CoRR*.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2019. Neural cross-lingual event detection with minimal parallel resources. 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 738–748, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Di Lu, Ananya Subburathinam, Heng Ji, Jonathan May, Shih-Fu Chang, Avi Sil, and Clare Voss. 2020. Cross-lingual structure transfer for zero-resource event extraction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1976–1981, Marseille, France. European Language Resources Association.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. 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 2795–2806, Online. Association for Computational Linguistics.

Olga Majewska, Ivan Vulić, Goran Glavaš, Edoardo Maria Ponti, and Anna Korhonen. 2021. Verb knowledge injection for multilingual event processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

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

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36, Los Angeles, California. Association for Computational Linguistics.

Meryem M'hamdi, Marjorie Freedman, and Jonathan May. 2019. Contextualized cross-lingual event trigger extraction with minimal resources. In *Conference on Computational Natural Language Learning (CoNLL)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *arXiv*.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. In *CoRR*, volume abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Bonan Min, Zhuolin Jiang, Marjorie Freedman, and Ralph Weischedel. 2017. Learning transferable representation for bilingual relation extraction via convolutional neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 674–684, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.

Taesun Moon, Parul Awasthy, Jian Ni, and Radu Florian. 2019. Towards lingua franca named entity recognition with BERT. In *CoRR*.

Aldrian Obaja Muis, Naoki Otani, Nidhi Vyas, Ruochen Xu, Yiming Yang, Teruko Mitamura, and Eduard Hovy. 2018. Low-resource cross-lingual event type detection via distant supervision with minimal effort. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 70–82, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021a. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Minh Van Nguyen and Thien Huu Nguyen. 2021. Improving cross-lingual transfer for event argument extraction with language-universal sentence structures. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 237–243, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Minh Van Nguyen, Tuan Ngo Nguyen, Bonan Min, and Thien Huu Nguyen. 2021b. Crosslingual transfer learning for relation and event extraction via word category and class alignments. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Thien Huu Nguyen and Ralph Grishman. 2015. Combining neural networks and log-linear models to improve relation extraction. In *CoRR*.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *CoRR*.

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.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. In *Computational Linguistics*.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Duy Phung, Hieu Minh Tran, Minh Van Nguyen, and Thien Huu Nguyen. 2021. Learning cross-lingual representations for event coreference resolution with multi-view alignment and optimal transport. In *Proceedings of the first Workshop on Multilingual Representation Learning (MRL)*.

Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2021. Cross-lingual learning for text processing: A survey. In *Expert Systems with Applications*.

Amir Pouran Ben Veyseh, Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021. Unleash GPT-2 power for event detection. 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 6271–6282, Online. Association for Computational Linguistics.

Longhua Qian, Haotian Hui, Ya'nan Hu, Guodong Zhou, and Qiaoming Zhu. 2014. Bilingual active learning for relation classification via pseudo parallel corpora. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592, Baltimore, Maryland. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *CoRR*.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. In *CoRR*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. In *CoRR*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *CoRR*.

Marta Recasens and Antonia Marti. 2010. Ancora-co: Coreferentially annotated corpora for spanish and catalan. In *Languages Resources and Evaluation*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models.

Rushin Shah, Bo Lin, Anatole Gershman, Robert Frederking, and Microsoft Bing Translatortm. 2010. Synergy: A named entity recognition system for

resource-scarce languages such as swahili using online machine translation. In *In Proceedings of International Conference on Language Resource and Evaluation Workshop on African Language Technology*.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 567–572, Beijing, China. Association for Computational Linguistics.

Avirup Sil and Radu Florian. 2016. One for all: Towards language independent named entity linking. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2255–2264, Berlin, Germany. Association for Computational Linguistics.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. XLDA: cross-lingual data augmentation for natural language inference and question answering. In *CoRR*.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *CoRR*.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Ander Soraluze, Olatz Arregi, Xabier Arregi, and Arantza Diaz De Ilarraza. 2017. Improving mention detection for basque based on a deep error analysis. volume 23, page 351–384. Cambridge University Press.

Ander Soraluze, Olatz Arregi, Xabier Arregi, Arantza Díaz de Ilarraza, Mijail Kabadjov, and Massimo Poesio. 2016. Coreference resolution for the Basque language with BART. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 67–73, San Diego, California. 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.

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

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. 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 1556–1566, Beijing, China. Association for Computational Linguistics.

Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou. 2020a. Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6578–6588, Online. Association for Computational Linguistics.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020b. Multilingual translation with extensible multilingual pretraining and finetuning.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 130–140, Ann Arbor, Michigan. 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.

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

Gorka Urbizu, Ander Soraluze, and Olatz Arregi. 2019. Deep cross-lingual coreference resolution for less-resourced languages: The case of Basque. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 35–41, Minneapolis, USA. Association for Computational Linguistics.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual relation extraction using compositional universal schema. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 886–896, San Diego, California. Association for Computational Linguistics.

C. Villani. 2008. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one-shot learning. In *30th Conference on Neural Information Processing Systems (NIPS)*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. In *Commun. ACM*.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. In *Technical report, Linguistic Data Consortium*.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307, Berlin, Germany. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juan-Zi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. Maven: A massive general domain event detection dataset. In *EMNLP*.

Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 2020a. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6505–6514, Online. Association for Computational Linguistics.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jian-Guang Lou. 2020b. Unitrans : Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje Karlsson, Biqing Huang, and Chin-Yew Lin. 2020c. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Z. 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 R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. In *CoRR*.

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

Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent Bayesian model for event coreference resolution. In *Transactions of the Association for Computational Linguistics*.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton W. Murray, Aaron Steven White, Mark Dredze, and Benjamin Van Durme. 2021. Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction.

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

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. 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 4568–4578, Hong Kong, China. Association for Computational Linguistics.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 437–448, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A Language Key

am - Armenian, ar - Arabic, bn - Bengali, de - German, en - English, es - Spanish, eu - Basque, hi - Hindi, it - Italian, ja - Japanese, ko - Korean, nl - Dutch, no - Norwegian, or - Oromo, pt - Portuguese, ru - Russian, ta - Tamil, ti - Tigrinya, tl - Tagalog, tr - Turkish, yr - Yoruba, zh - Chinese

B Dataset Statistics

B.1 CoNLL

Table 5: Number of entity instances in the CoNLL-2002 and CoNLL 2003 datasets.

Language	Train	Dev	Test
German-de (CoNLL-2003)	11,851	4,833	3,673
English-en (CoNLL-2003)	23,499	5,942	5,648
Spanish-es (CoNLL-2002)	18,798	4,351	3,558
Dutch-nl (CoNLL-2002)	13,344	2,616	3,941

B.2 ACE

Table 6: Number of instances for ED, RE, and EAE in the ACE05 and ACE05-ERE datasets.

Language	Data	RE (#rels)	ED (#trgs)	EAE (#args)
Arabic-ar	Train	2,918	1,986	3,959
	Dev	357	112	495
	Test	378	169	495
English-en	Train	4,974	4,420	7,018
	Dev	626	505	877
	Test	620	424	878
Chinese-zh	Train	4,767	2,213	5,931
	Dev	572	111	741
	Test	605	197	742
English-en (ERE)	Train	5,045	6,419	X
	Dev	424	552	X
	Test	477	559	X
Spanish-es (ERE)	Train	1,698	3,272	X
	Dev	120	210	X
	Test	108	269	X

Reference	Task	Tgt Langs	CL Resources	CL Approach	Technique
Tsai et al. (2016)	EMD	bn, es, de, nl ta, tl, tr, yr	Wikipedia	Direct transfer	Wikification
Mayhew et al. (2017)	EMD	bn, es, de hi, nl, ta, yr	Bilingual dictionary	Data transfer	Lexicon-based translation
Ni et al. (2017)	EMD	es, de, ja ko, nl, pt	Bilingual dictionary Bilingual embeddings	Direct transfer	Zero-shot transfer learning
Cruz et al. (2018)	CRR	pt	Bilingual embeddings	Direct transfer	Language-independent features
Feng et al. (2018)	EMD	en, nl, zh	Bilingual dictionary	Data transfer	Translation-enriched representations
Kundu et al. (2018)	CRR	es, zh	Bilingual embeddings	Direct transfer	Language-independent features
Muis et al. (2018)	ETL	ti, or	Bilingual dictionary Bilingual embeddings	Hybrid transfer	Adversarial learning
Xie et al. (2018)	EMD	bn, es, de hi, nl, ta, yr	Bilingual dictionary Bilingual embeddings	Data transfer	Self-attention-aided translation
Zou et al. (2018)	RE	en, zh	Machine translation Multilingual embeddings	Hybrid transfer	Adversarial learning
Ahmad et al. (2019)	EAE	bn, hi	Multilingual embeddings	Hybrid transfer	Adversarial learning Independent language layers
Caselli and Ustun (2019)	ED	it	Multilingual LM	Direct Transfer	Zero-shot learning
Jain et al. (2019)	EMD	am, de, es hi, tm, zh	Machine translation Bilingual dictionary	Data transfer	Annotation projection
Keung et al. (2019)	EMD	de, es, fr, it ja, nl, ru, zh	Multilingual LM	Direct transfer	Adversarial learning
Liu et al. (2019)	ED	es, zh	Bilingual dictionary	Data transfer	Context-dependent translation Syntactic-order detector
M'hamdi et al. (2019)	ED	ar, zh	Multilingual LM Multilingual embeddings	Direct transfer	Zero-shot learning
Moon et al. (2019)	EMD	ar, de es, nl, zh	Multilingual LM	Direct transfer	Multi-source training
Ni and Florian (2019)	RE	ar, de, es it, ja, pt, zh	Bilingual dictionary	Direct transfer	Embedding projection
Subburathinam et al. (2019)	EAE RE	ar, zh	Multilingual embeddings	Direct transfer	Dependency parsing
Urbizu et al. (2019)	CRR	eu	Bilingual embeddings	Direct transfer	Language-independent features
Wu and Dredze (2019)	EMD	de, es, nl, zh	Multilingual LM	Direct transfer	Zero-shot transfer learning

Table 7: Summary of surveyed works for cross-lingual information extraction. Language key can be found in appendix A.

Reference	Task	Tgt Langs	CL Resources	CL Approach	Technique
Bari et al. (2020)	EMD	ar, de es, fi, nl	Bilingual embeddings	Direct transfer	Adversarial learning
Hambardzumyan et al. (2020)	EI	ar, de	Machine translation	Data transfer	Representation alignment
Köksal and Özgür (2020)	RE	ar, de, es, zh	Multilingual LM	Direct transfer	Distantly supervised pretraining
Lu et al. (2020)	ETL ARL	es, ru, uk	Multilingual embeddings	Direct transfer	Structure transfer
Wu et al. (2020a)	EMD	de, es, nl	Multilingual LM	Hybrid transfer transfer	Knowledge distillation
Wu et al. (2020b)	EMD	de, es, nl, no	Multilingual LM	Hybrid transfer	Embedding-based translation Knowledge distillation
Wu et al. (2020c)	EMD	de, es fr, nl, zh	Multilingual LM	Direct transfer	Meta-Learning
Ahmad et al. (2021)	EAE RE	ar, zh	Multilingual LM	Direct transfer	Weighted self-attention
Chen et al. (2021)	EMD	de, es, nl	Multilingual LM	Hybrid transfer	Adversarial learning Knowledge distillation
Fincke et al. (2021)	ED EAE	ar	Multilingual LM	Direct transfer	Model priming
Liang et al. (2021)	EMD	ar, de es, hi zh	Multilingual LM	Hybrid transfer	Reinforcement learning Knowledge distillation
Majewska et al. (2021)	ED EAE	ar, zh	Multilingual LM	Direct transfer	Verb lexical knowledge
Nguyen and Nguyen (2021)	EAE	ar, zh, ar	Multilingual Embeddings	Direct transfer	Language-universal structures
Nguyen et al. (2021b)	ED EAE RE	ar, zh	Multilingual LM	Direct transfer	Word categories and classes
Phung et al. (2021)	ECR	es, zh	Multilingual LM	Direct transfer	Adversarial learning Multi-view alignment
Yarmohammadi et al. (2021)	EMD EE RE	ar, de, es, fr hi, ru, vi, zh	Multilingual LM	Hybrid transfer	Machine Translation
Guzman-Nateras et al. (2022b)	ED	ar, es, zh	Multilingual LM	Direct transfer	Adversarial learning
Huang et al. (2022a)	EAE	ar, es, zh	Multilingual Generative LM	Direct transfer	Language-agnostic templates

Table 8: Summary of surveyed works for cross-lingual information extraction (Cont...) Language key can be found in appendix A.

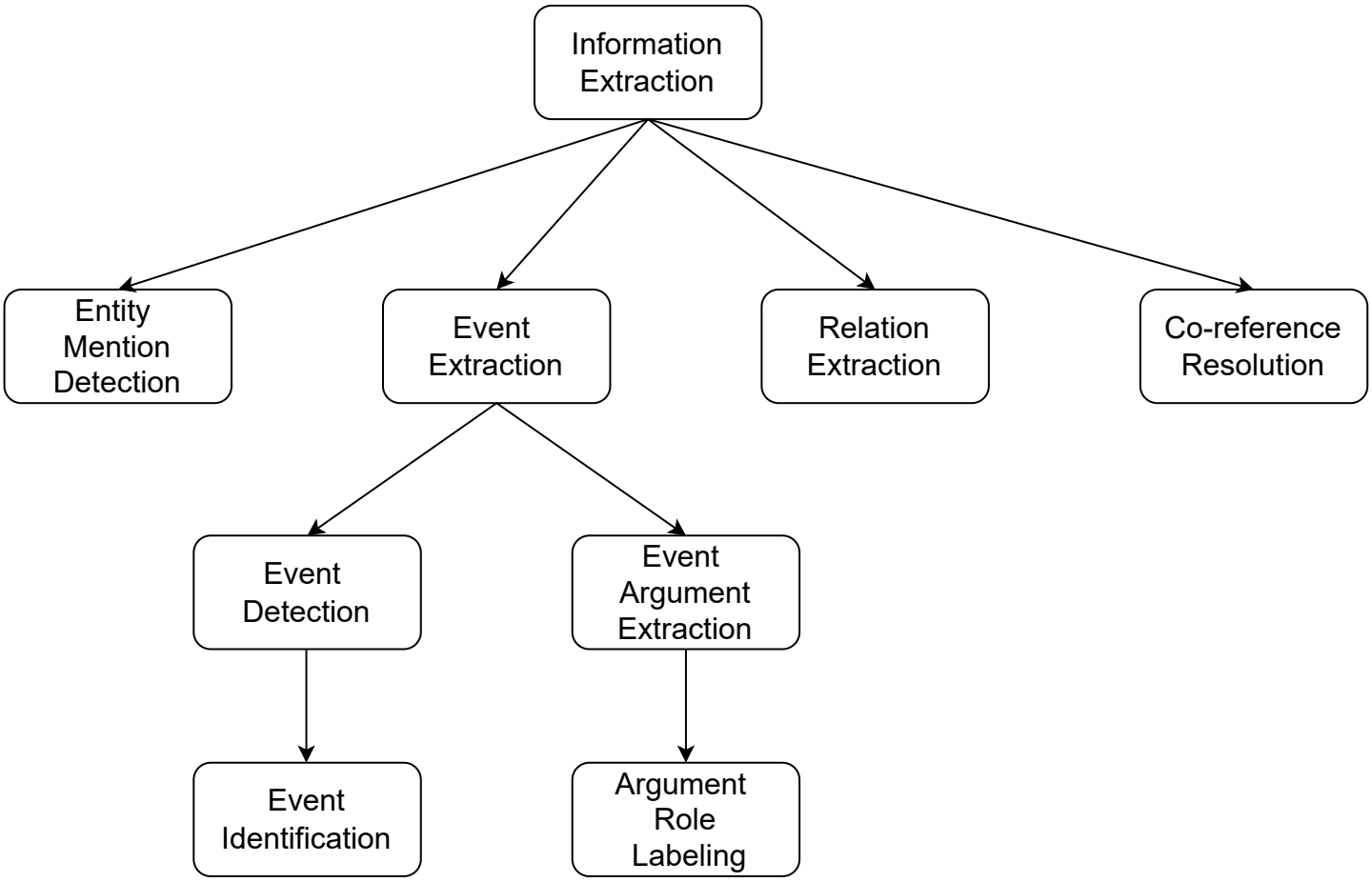

Figure 1: Information Extraction Conceptual Map