IMPROVING CROSS-LINGUAL TRANSFER LEARNING FOR EVENT DETECTION

by

LUIS FERNANDO GUZMAN NATERAS

A DISSERTATION

Presented to the Department of Computer Science
and the Division of Graduate Studies of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2023

DISSERTATION APPROVAL PAGE

Student: Luis Fernando Guzman Nateras

Title: Improving Cross-Lingual Transfer Learning for Event Detection

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Computer Science by:

Thien Huu Nguyen	Chair
Daniel Lowd	Core Member
Thanh Nguyen	Core Member
Kristopher Kyle	Institutional Representative

and

Krista Chronister	Vice Provost for Graduate Studies

Original approval signatures are on file with the University of Oregon Division of Graduate Studies.

Degree awarded June 2023

DISSERTATION ABSTRACT

Luis Fernando Guzman Nateras

Doctor of Philosophy

Department of Computer Science

June 2023

Title: Improving Cross-Lingual Transfer Learning for Event Detection

The widespread adoption of applications powered by Artificial Intelligence (AI) backbones has unquestionably changed the way we interact with the world around us. 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 (NLP) research. Nonetheless, with over 7000 spoken languages in the world, there still remain a considerable number of marginalized communities that are unable to benefit from these technological advancements largely due to the language they speak. Cross-Lingual Learning (CLL) looks to address this issue by transferring the knowledge acquired from a popular, high-resource source language (e.g., English, Chinese, or Spanish) to a less favored, lower-resourced target language (e.g., Urdu or Swahili). This dissertation leverages the Event Detection (ED) sub-task of Information Extraction (IE) as a testbed and presents three novel approaches that improve cross-lingual transfer learning from distinct perspectives: (1) direct knowledge transfer, (2) hybrid knowledge transfer, and (3) few-shot learning.

This dissertation includes both published and unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Luis Fernando Guzman Nateras

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, Oregon, USA
Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Michoacan, Mexico

DEGREES AWARDED:

Master of Science, Computer and Information Science, 2022, University of Oregon
Master of Science, Electrical Engineering, 2014, Universidad Michoacana de San Nicolás de Hidalgo
Bachelor of Science, Computer Engineering, 2008, Universidad Michoacana de San Nicolás de Hidalgo

AREAS OF SPECIAL INTEREST:

Deep Learning
Natural Language Processing
Information Extraction

PROFESSIONAL EXPERIENCE:

Graduate Teaching Assistant, University of Oregon, 2019-2023
Academic Technician, Universidad Michoacana de San Nicolás de Hidalgo, 2008-2018

GRANTS, AWARDS AND HONORS:

Best Graduate Employee Teaching Award, University of Oregon, 2022
Raymund Fellowship Award, University of Oregon, 2018

Promising Scholar Award, University of Oregon, 2018
Graduate Studies Scholarship, Mexican National Science Council, 2012, 2018

PUBLICATIONS:

Guzman-Nateras, L., Dernoncourt, F., Nguyen, T. *"Hybrid Knowledge Transfer for Improved Cross-Lingual Event Detection via Hierarchical Sample Selection" To appear at* ACL, 2023

Guzman-Nateras, L., Nguyen, M., Nguyen, T. *"Cross-Lingual Event Detection via Optimized Adversarial Training"* NAACL, 2022

Guzman-Nateras, L., Lai, V., Dernoncourt, F., Nguyen, T. *"Event Detection for Suicide Understanding"* NAACL, 2022

Guzman-Nateras, L., Lai, V., Dernoncourt, F., Nguyen, T. *"Few-Shot Cross-Lingual Learning for Event Detection"* 2nd Multilingual Representation Learning Workshop @EMNLP, 2022

Ahern, I., Noack, A., Guzman-Nateras, L., Dou, D., Li, B., Huan, J. *"NormLime: A new feature importance metric for explaining deep neural networks"* CoRR, 2019

Guzman-Nateras, L., Camarena-Ibarrola, J. *"On the Use of Locality Sensitive Hashing for Audio Following"* CIARP, 2014

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Thien Nguyen for taking a chance on me when I needed it the most. His clear guidance and thoughtful advice have been fundamental factors in achieving this important milestone in my career. I would also like to thank Dr. Daniel Lowd and Dr. Thanh Nguyen for accompanying me throughout my time at the UO as members of my different committees, and Dr. Kristopher Kyle for agreeing to participate as a member of my dissertation committee.

My lab mates in the NLP research group for all the shared experiences and their selfless willingness to help me get a running start when I began working on a new research area.

All of the amazing professors and departmental staff with whom I have had the chance to work while at the University. In particular, Dr. Kathy Freeman and Phil Colbert for their unconditional support every time I have asked for it; Dr. Juan Flores for his friendship and encouragement; Dr. Hank Childs for his candid advice and guidance; and Dr. Dejing Dou for giving me the opportunity to have this experience.

All my friends and family, both back home in Morelia and locally in Eugene, for always being a source of relief in an otherwise stressful environment. It would take an entire appendix to name everyone but you guys know who you are.

Finalmente, a mi amada esposa Nayeli, porque sin su paciencia, sacrificio, y apoyo incondicional este logro no habría sido posible.

A mis hijos, la motivación absoluta de todo lo que hago.

A mis padres, la causa primigenia de todo éxito obtenido.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Most content for this chapter comes from my dissertation proposal. I was the primary author for this chapter and Thien Nguyen provided editorial suggestions.

1.1 Introduction

Recent years have seen the development and widespread adoption of applications powered by Artificial Intelligence (AI) with Natural Language Processing (NLP) backbones. For example, applications such as automatic question answering, automated personal assistants, fake news identification, and product review sentiment analysis all make use of NLP-based models. These models are usually trained in a supervised manner by leveraging large amounts of labeled data. However, large annotated datasets are a luxury reserved for a handful of widely-spoken popular languages (e.g., English, Chinese, or Spanish) due in large part to the higher availability of potential annotators and the corresponding decrease in annotation costs. In consequence, a vast majority of 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 harness the knowledge acquired from annotated data from a high-resource *source* language and transfer such knowledge into a so-called *target* language. Cross-lingual learning then 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 their communities to benefit from the aforementioned NLP-based tools.

Nonetheless, it should be evident that a cross-lingual setting poses a much more complex scenario than its monolingual counterpart as it must address additional complications. More often than not, there exist substantial differences between the desired source and target languages. Both major differences, such as having distinct word orders or even entirely disjoint alphabets, and more subtle ones, like polysemous or non-existing words, pose significant hurdles for CLL approaches to overcome.

Despite these difficulties, CLL has been successfully applied to several NLP tasks (Pikuliak, Šimko, & Bieliková, 2021a). One of the most prominent among such tasks is Information Extraction (IE). Information extraction, 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 (e.g., social media posts, discussion forums, crowd-maintained archives, etc). Being able to perform computations on the previously-unstructured data is the ultimate goal of IE. Nonetheless, such a final objective is highly complex which is why the IE task has been broken down into the following simpler sub-tasks: Entity Mention Detection (EMD), Co-Reference Resolution (CRR), Relationship Extraction (RE), and Event Extraction (EE) which is subsequently subdivided into Event Detection (ED) and Event Argument Extraction (EAE),

1.1.1 Overarching Dissertation Theme. The holistic objective of this research work is to advance the field of cross-lingual learning. For such purposes, we select the ED sub-task as the main focus of our efforts. Event Detection is a highly challenging problem due to its heavy reliance on contextual information. This characteristic makes ED an ideal testbed for our proposed cross-

lingual approaches. As such, **the central theme for this dissertation is to design strategies for improved event-detection performance tailored specifically for a cross-lingual setting**. We propose to address this overarching objective from three distinct perspectives: (1) Chapter III presents an approach based on the *direct* knowledge transfer paradigm in which cross-lingual models are trained via language-agnostic features; (2) Chapter IV then introduces a *hybrid* method that employs both the aforementioned *direct* transfer paradigm as well as the *data* transfer approach that trains cross-lingual models in the target language directly; (3) Chapter V discusses the entirely novel Few-Shot Cross-Lingual setting for ED and introduces an original Few-Shot Learning (FSL) method customized for cross-lingual learning. The remaining chapters of this dissertation are structured as follows: Chapter II provides an overview of CLL terminology and concepts, as well as a comprehensive review of the state of modern research efforts into Cross-Lingual Information Extraction (CLIE). Finally, Chapter VI includes our conclusions and a discussion on potential future CLIE research directions.

CHAPTER II

BACKGROUND: STATE OF MODERN CROSS-LINGUAL INFORMATION EXTRACTION

The content for this chapter is a modified version of my area exam (candidacy exam). I was the primary author of the original document with editorial suggestions from Thien Nguyen.

2.1 Introduction

This chapter provides a review of modern cross-lingual learning efforts in each of the information extraction sub-tasks, including event detection. We begin by discussing the usual terminology employed in cross-lingual works and providing some background on the resources they leverage. Then, we describe the knowledge-transfer paradigms that characterize them. Finally, the current 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. We discuss their strengths and weaknesses with respect to each other and provide a performance comparison when adequate.

2.2 Cross-Lingual Concepts

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

2.2.1 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.

2.2.1.1 Parallel Corpus. A parallel corpus is one of the most useful, but also the most scarce, bilingual resources. 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.

2.2.1.2 Pseudo-parallel Corpus. Automated machine translation has witnessed great advances in recent years by leveraging encoder-decoder models (Bahdanau, Cho, & Bengio, 2015; Y. Liu et al., 2020) and, of course, Google's translation API (Y. 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. Afterward, words in pseudo-parallel sentences can be aligned using automatic tools such as GIZA++ (Och & Ney, 2003), Fast-align (Dyer, Chahuneau, & Smith, 2013) and Awesome-align (Dou & 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.

2.2.1.3 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, Tsai, & Roth, 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.

2.2.1.4 Multilingual Word Embeddings. Monolingual word embeddings such as Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and Glove (Pennington, Socher, & Manning, 2014) are collections of dense, high-dimensional, 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., 2021a). 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, Vulić, & Søgaard, 2019). In a sense, multilingual embeddings are secondary cross-lingual resources 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 (Artetxe, Labaka, & Agirre, 2018; Bojanowski, Grave, Joulin, & Mikolov, 2017; X. Chen & Cardie, 2018).

2.2.1.5 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 probability. 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, Chang, Lee, and Toutanova (2019) and GPT (Radford & Narasimhan, 2018) and their variations (RoBERTa, GPT-2, GPT-3, GPT-4) are probably the most well-known monolingual LMs. Multilingual Language Models (MLMs) are just extensions of their monolingual counterparts. 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.

2.2.2 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*.

2.2.2.1 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 (Ehrmann, Turchi, & Steinberger, 2011; Fu, Qin, & Liu, 2014; Hwa, Resnik, Weinberg, Cabezas, & Kolak, 2005; Yarowsky, Ngai, & Wicentowski, 2001; Zeman & Resnik, 2008), or neural machine translation systems (Jain, Paranjape, & Lipton, 2019; Shah, Lin, Gershman, Frederking, & Translatortm, 2010; Tiedemann, Agić, & Nivre, 2014). 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 (Dou & Neubig, 2021; Dyer et al., 2013; Och & Ney, 2003) 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, Yang, Neubig, Smith, & Carbonell, 2018b). 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 if the target language is close to the training/source language Tsai, Mayhew, and Roth (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 of 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 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 pre-trained encoders. For the word alignment, they compare using the statistical model Fast-align (Dyer et al., 2013) and Awesome-align (Dou & Neubig, 2021) which computes alignments based on contextualized-embedding similarity. They evaluate their approach in five downstream tasks: event extraction, using ACE05 Walker, Strassel, Medero, and Maeda (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.

2.2.2.2 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 exploit a shared representation for the source and target languages (Bharadwaj, Mortensen, Dyer, & Carbonell, 2016; Chaudhary et al., 2018; Kozhevnikov & Titov, 2014; Täckström, McDonald, & Uszkoreit, 2012). For instance, Ni, Dinu, and Florian (2017) propose to project monolingual word embeddings into a common space as

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

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, especially 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, Labaka, and Agirre (2020) found that the translation process can introduce subtle artifacts that have a notable impact on 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.

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, Ravishankar, Vulić, and Glavaš (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 (Z. Ahmad, Varshney, Ekbal, & Bhattacharyya, 2019; W. Chen, Jiang, Wu, Karlsson, & Guan, 2021; Guzman-Nateras, Nguyen, & Nguyen, 2022; Keung, Lu, & Bhardwaj, 2019; Phung, Tran, Nguyen, & Nguyen, 2021).

*2.2.2.3 **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 2.2.2.1) where they leverage the language-invariant capabilities of pre-trained multilingual encoders and so-called *silver* target-data generated with annotation projection.

Knowledge distillation (W. Chen et al., 2021; Liang et al., 2021; Q. Wu, Lin, Karlsson, Lou, & Huang, 2020; Q. Wu, Lin, Karlsson, Huang, & Lou, 2020) 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, Awasthy, Ni, and Florian (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, McCann, Keskar, Xiong, and Socher (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 second *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.

2.3 Entity Mention Detection

2.3.1 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 of classes defined by a particular ontology. Commonly used categories include names for organizations, locations, persons, companies, and numerical values such as 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, Màrquez, & Padró, 2003). The segmentation sub-task deals with identifying contiguous spans of tokens representing an entity. A common restriction EMD systems assume 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 regarded as an entity itself. As for the classification sub-task, once entity candidates have been identified, they are categorized into ontology-specific types. This means 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 of five tokens instead of two.

2.3.2 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 these 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, Feng, Qin, Feng, and Liu (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 a 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. (2018b) 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 order, 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., 2018b), 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

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

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 a 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. Afterward, 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 computed 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.

2.3.3 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, with 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 applying 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 & 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 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 (Bojanowski et al., 2017; Mikolov, Sutskever, et al., 2013; Pennington et al., 2014) 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). Afterward, 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.

S. 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

[3]www.wikidata.org

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 S. 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, Joty, and Jwalapuram (2020) propose a model that leverages two distinct BiLSTM-based encoders, one for each language. They argue that 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 (Keung et al., 2019; S. Wu & Dredze, 2019) that leverage simpler model architectures.

A meta-learning-based approach for EMD is presented by Q. Wu, Lin, Wang, et al. (2020). 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 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 T_i is defined by a pseudo testing set $D^{T_i}_{test} = x_i$, and a pseudo training set

$D_{train}^{T_i}$ comprised by K of x_i most similar examples in the latent space. Then the model is trained on a randomly-sampled task T_i to minimize the loss computed on $D_{train}^{T_i}$ (`inner update`) to obtain an updated set of parameters θ'. These updated parameters θ' are then evaluated on $D_{test}^{T_i}$ and another update is made (`meta update`). During the adaptation phase, the model is applied to target languages. Here, each target-language test example $x_j \in D_{test}^{T}$ is used as the test set $D_{test}^{T_j}$ for a target task T_j. The task training set $D_{train}^{T_j}$ is again obtained by retrieving the top-k similar examples of x_j from D_{train}^{T}. Once more, the model is first fine-tuned to minimize the error on $D_{train}^{T_j}$ 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 a 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 S. Wu and Dredze (2019): a linear classifier on top of mBERT.

2.3.4 Hybrid Transfer Cross-lingual EMD. Q. Wu, Lin, Karlsson, Lou, and Huang (2020) 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 a 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, Charniak, and Johnson (2010) between the target language and each corresponding source language.

The UniTrans model (Q. Wu, Lin, Karlsson, Huang, & Lou, 2020) 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, Turban, Hamblin, & Hammerla, 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 the 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. Afterward, they leverage a teacher-student learning setup similar to (Q. Wu, Lin, Karlsson, Lou, & Huang, 2020): 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 Q. Wu, Lin, Karlsson, Lou, and Huang (2020) 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 (W. 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.

Appendix B includes a comparison between the performance of the works described in this section.

2.4 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 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 (X. Wang et al., 2020), but there also are domain-specific datasets, such as BRAD (V. Lai, Nguyen, Kaufman, & Nguyen, 2021) for historical events and SuicideED (Guzman-Nateras, Lai, Pouran Ben Veyseh, Dernoncourt, & Nguyen,

2022) 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 sub-tasks: Event Detection (ED) and Event Argument Extraction (EAE).

2.4.1 Event Detection.

2.4.1.1 Task Definition.

Event Detection (ED), EE'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, Vulić, Glavaš, Ponti, & Korhonen, 2021a). 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.

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

Complications such as this one have nudged CLED research efforts to favor direct transfer approaches to take advantage of their language-agnostic training.

2.4.1.2 Data Transfer Cross-lingual ED.

The only recent data-transfer-based method for CLED we could find is the work by J. Liu, Chen, Liu, and Zhao (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, Sutskever, et al. (2013) 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, Conneau, Ranzato, Denoyer, & Jégou, 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 & 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 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.

2.4.1.3 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, Sprugnoli, Speranza, & Monachini, 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, especially 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, Freedman, and May (2019b) 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, Lample, Ranzato, Denoyer, & Jégou, 2017) alongside a bidirectional-LSTM encoder with a CRF (Lafferty, McCallum, & Pereira, 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.

D. 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, Denoyer, & Ranzato, 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, Socher, & Manning, 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. (2021a). 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, Korhonen, Ryant, and Palmer (2006) and FrameNet Baker, Fillmore, and Lowe (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. (2019b) 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. (2019b). 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 re-frames the event extraction task as a question-answering one, the authors of Fincke, Agarwal, Miller, and Boschee (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 pre-trained 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 M. V. Nguyen, Nguyen, Min, and Nguyen (2021) 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 (X. Chen, Sun, Athiwaratkun, Cardie, & Weinberger, 2018; He, Yan, & Xu, 2020; Joty, Nakov, Màrquez, & Jaradat, 2017; Keung et al., 2019), 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 addresses the cross-class alignment issue which translates into improved task performance.

2.4.1.4 Hybrid Transfer Cross-lingual ED.

Similar to the previously-described work by D. Lu et al. (2020). The work by Muis et al. (2018a) does not really address the ED task and focuses instead on 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, Chen, Corrado, & Dean, 2013 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, Kanayama, Ma, Bird, & Cohn, 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 the such bootstrapping approach.

2.4.2 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, Naseem, Ni, Moon, and Florian (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, Khachatrian, and May (2020a). 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.

2.4.3 Event Argument Extraction.

2.4.3.1 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 to cross-lingual EAE as well: different word orders, distinct character sets, non-existing words, polysemous words, etc.

2.4.3.2 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 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 & 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, Bojanowski, Mikolov, Jégou, & Grave, 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 D. 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. (2021a) (section 2.4.1 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 M. V. 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.

W. Ahmad, Peng, and Chang (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 M. V. 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 (H. Tang, Ji, Li, & Zhou, 2020; C. Zhang, Li, & Song, 2019). 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 M. V. 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. Afterward, 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 2.4.1, 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 M. V. Nguyen, Nguyen, et al. (2021) seeks to improve cross-lingual representation learning by conditioning on class information and universal word categories such as POS and dependency relations. Section 2.4.1 provides further details on the model.

K.-H. Huang, Hsu, Natarajan, Chang, and Peng (2022a) 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 W. Ahmad et al. (2021). Their key idea is to fine-tune a pre-trained multilingual generative language model such as mBART (Y. Tang et al., 2020) 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 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).

2.4.3.3 Hybrid Transfer Cross-Lingual EAE.

Z. 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. Afterward, 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.

2.5 Relation Extraction

2.5.1 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 tasks 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 (Kambhatla, 2004; Q. Li & Ji, 2014; Zelenko, Aone, & Richardella, 2003). These methods were later replaced by approaches that make use of deep neural networks trained in a supervised manner (dos Santos, Xiang, & Zhou, 2015; Miwa & Bansal, 2016; T. H. Nguyen & Grishman, 2015a; L. Wang, Cao, de Melo, & Liu, 2016; Zeng, Liu, Lai, Zhou, & Zhao, 2014). Regarding cross-lingual efforts for RE, over the past decade there have been approaches based on active learning (Qian, Hui, Hu, Zhou, & Zhu, 2014), knowledge bases (Verga, Belanger, Strubell, Roth, & McCallum, 2016), and bilingual representations learned through language-independent concepts (Min, Jiang, Freedman, & Weischedel, 2017).

2.5.2 Data Transfer Cross-lingual RE.

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

2.5.3 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, Sutskever, et al., 2013) 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 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 to 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 (T. H. Nguyen & Grishman, 2015a). Their performance on cross-lingual RE also seems to be lacking with respect to the previously released CNN-GAN (Zou, Xu, Hong, & Zhou, 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 2.4.3 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 pre-trained transformer (mBERT Devlin et al., 2019) as its encoder which is then pre-trained 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. Afterward, sentences including entity pairs with Wikidata relations (Vrandečić & 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, FitzGerald, Ling, and Kwiatkowski (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 pre trained 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 (W. Ahmad et al., 2021), described in detail in section 2.4.3 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 M. V. Nguyen, Nguyen, et al. (2021) also test the performance of their CCAR model on the RE task. As mentioned in section 2.4.1, their intuition is to improve the alignment of cross-lingual representations by conditioning on language-invariant information: class information, POS category, and dependency relation.

2.5.4 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, Liu, Liu, & Sun, 2015) alongside

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

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%).

2.6 Co-Reference Resolution

2.6.1 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, and 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 though 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, Rocha, and Cardoso (2018).

2.6.2 Data Transfer Cross-Lingual CRR. Cross-lingual data-transfer-based approaches for CRR are limited to a couple of shared tasks (Ji, Nothman, Hachey, & Florian, 2015) and are primordially based on annotation projection.

2.6.3 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, Sil, Florian, and Hamza (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 & 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, Le, and Sutskever (2013).

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 & 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, Soraluze, and Arregi (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, Arregi, Arregi, & Diaz De Ilarraza, 2017; Soraluze et al., 2016). The authors explore leveraging a large English corpus (OntoNotes) to create a cross-lingual Basque model given that the largest CRR corpora for Basque (Cerberio, Aduriz, Diaz de Ilarraza, & Garcia-Azkoaga, 2018) are 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 are 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, Tran, 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 (B. Yang, Cardie, & Frazier, 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 augments 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 & 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 to 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.

CHAPTER III

OPTIMIZING ADVERSARIAL TRAINING FOR CROSS-LINGUAL EVENT DETECTION

This Chapter contains materials from the published paper "*Luis F. Guzman-Nateras, Minh V. Nguyen, and Thien H. Nguyen. '**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, 2022*"(Guzman-Nateras, Nguyen, & Nguyen, 2022). As the first author of this publication, Luis was responsible for most areas of the project including development, experimentation, and document writing. Minh provided a starting code base and meaningful discussions and insights. Thien had input on the initial project conceptualization and made editorial suggestions for the final document. The original publication contents have undergone some editorial updates to comply with this document's format and purpose.

After the review of modern approaches to cross-lingual information extraction and their associated terminology presented in the previous chapter, this chapter introduces our first contribution to Cross-Lingual Event Detection (CLED). Our proposed methodology follows a direct-transfer-based approach to cross-lingual learning. As discussed in Section 2.2, under such a paradigm, a model is trained using language-invariant features and then directly applied to the target language. Many recent works in CLED have harnessed the language-invariant qualities displayed by pre-trained Multilingual Language Models. Their performance, however, reveals there is room for improvement as they still struggle

with the particular challenges entailed by a cross-lingual setting. As such, we leverage Adversarial Language Adaptation (ALA) to train a language discriminator to discern between the source and target languages using unlabeled data. The discriminator is trained in an adversarial manner so that the encoder learns to produce refined, language-invariant representations that lead to improved performance. More importantly, we propose to optimize the adversarial training process by only presenting the discriminator with the most *informative* samples. We base our intuition about *what* makes a sample informative on two disparate metrics: sample similarity and event presence. Thus, we propose leveraging Optimal Transport (OT) (Villani, 2008) as a solution to naturally combine these two distinct information sources into the selection process. Extensive experiments on 8 different language pairs, using 4 languages from unrelated families, show the flexibility and effectiveness of our model.

3.1 Introduction

Event Detection (ED) is an important sub-task within the broader Information Extraction (IE) task. Event detection consists of being able to identify the words, commonly referred to as *triggers*, that denote the occurrence of events in a sentence, and classify them into a discrete set of event types. For example, in the sentence *"Jamie* ***bought*** *a car yesterday."*, *bought* is considered the trigger of a TRANSACTION:TRANSFER-OWNERSHIP[1] event type. It is a very well-studied task in which there have been lots of previous research efforts that have recently been primarily deep learning-based (Y. Chen, Xu, Liu, Zeng, & Zhao, 2015; J. Liu, Chen, Liu, Bi, & Liu, 2020; T. H. Nguyen, Cho, & Grishman, 2016; T. H. Nguyen, Fu, Cho, & Grishman, 2016; T. H. Nguyen & Grishman, 2015b; T. M. Nguyen

[1]Event type taken from the ACE05 dataset.

& Nguyen, 2019; Sha, Qian, Chang, & Sui, 2018; Wadden, Wennberg, Luan, & Hajishirzi, 2019; S. Yang, Feng, Qiao, Kan, & Li, 2019a; J. Zhang, Qin, Zhang, Liu, & Ji, 2019; Y. Zhang et al., 2020).

Nonetheless, ED remains quite a challenging task as the context in which a trigger occurs can change its corresponding type completely. Furthermore, the same event might also be expressed by entirely different words/phrases. Additionally, the vast majority of the aforementioned efforts are limited to a monolingual setting — performing ED on text belonging to a single language — and usually focused on a small set of popular languages. This is mainly due to the fact that most of the available annotated data belongs to these *high-resource languages*. This problem becomes critical for *low-resource languages* for which the amount of available training data is minimal or non-existent. Consequently, some approaches have proposed taking advantage of the widely available unlabeled data in a semi-supervised manner (Muis et al., 2018b).

Alternatively, CLED proposes the scenario of creating models that effectively perform ED on data belonging to more than one language, which brings about additional challenges. For instance, trigger words present in one language might not exist in another one. Frequent examples of this phenomenon are verb conjugations where some tenses only exist in some languages. Accurate verb handling is of particular importance for the ED task as event triggers are usually related to the verbs in a sentence. Some recent work (Majewska, Vulić, Glavaš, Ponti, & Korhonen, 2021b) has attempted to address this issue by injecting external verb knowledge into the training process. Another similar problematic issue for CLED is triggers with different meanings that are each distinct words in different languages. For instance, the word "*juicio*" in Spanish can either mean

"*judgement*" or "*trial*" in English, depending on the context. These, and other similar, issues make CLED a challenging task.

A compelling approach to creating a cross-lingual model is to use *direct transfer learning* which carries the performance of a model trained on a *source* language over onto a second *target* language. The general idea is leveraging the existing high-quality annotated data available for a high-resource language to train a model in a way that allows it to learn the language-invariant characteristics of the task at hand, ED in this case, so that it also performs effectively on text from a second language. Prior works on direct transfer learning for CLED have relied on pre-trained Multilingual Language Models (MLMs), such as multilingual BERT (mBERT) (Devlin et al., 2019), to take advantage of their innate language-invariant qualities. Yet, their performance still shows room for improvement as they sometimes struggle to handle the difficult instances, unique to cross-lingual settings, mentioned earlier. We identify a significant shortcoming of previous CLED efforts in that they do not exploit the abundant supply of unlabeled data: even though MLMs are trained on immense amounts of it, unlabeled data is not used when fine-tuning for the ED task. It is our intuition that by integrating unlabeled target-language data into the training process, the model is exposed to more language context which should help deal with issues such as verb variation and multiple connotations.

As such, we propose making use of Adversarial Language Adaptation (ALA) (X. Chen et al., 2018; Joty et al., 2017) to train a CLED model. The key idea is to generate language-invariant representations that are not indicative of language but remain informative for the ED task. Unlabeled data from both the source and target languages is used to train a Language Discriminator (LD)

network that learns to discern between the two. The *adversarial* part comes from the fact that the encoder and discriminator are trained with opposing objectives: as the LD becomes better at distinguishing between languages, the encoder learns to generate more language-invariant representations in an attempt to *fool* the LD. An overview of the ALA framework is shown in Figure 1. To the best of our knowledge, our work is the first one proposing the use of ALA for the CLED task.

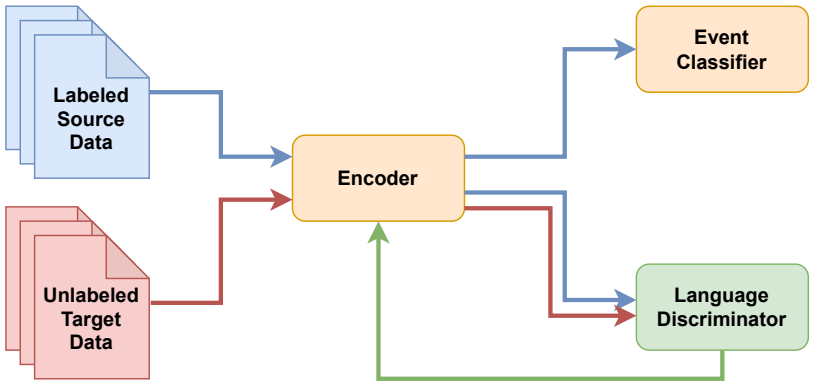

Figure 1. Overview of ALA framework. A multilingual encoder is presented with both labeled data from the source language and unlabeled target data. Then the sentence-level encodings are presented to the language discriminator, whose task is to determine their originating language. The discriminator outputs are then used to train the encoder in an adversarial manner, resulting in language-invariant representations.

Nonetheless, contrary to past uses of ALA where the same importance is given to all unlabeled samples, we recognize that such a course of action is sub-optimal as certain samples are bound to be more informative for the discriminator than others. For example, we would like to present the LD with the samples that allow it to learn the fine-grained distinctions between the source and target languages, instead of relying on syntactic differences. Moreover, in the context of ED, we suggest it would be beneficial for the LD to be trained with examples

containing events, instead of non-event samples, as the presence of an event can then be incorporated into the generated representations.

Hence, we propose refining the adversarial training process by only keeping the most informative examples while disregarding less useful ones. Our intuition as to *what* makes samples more informative for CLED is two-fold: First, we presume that presenting the LD with examples that are too different makes the discrimination task too simple. As mentioned previously, we would like the LD to learn a fine-grained distinction between the source and target languages which, in turn, improves the language-invariance of the encoder's representations. Thus, we suggest presenting the LD with examples that have similar contextual semantics, i.e., similar contextualized representations. Second, we consider that sentences containing events should provide an ED system with additional task-relevant information when compared to non-event samples. Accordingly, we argue that event-containing sentences should have a larger probability of being selected for ALA training.

With these intuitions in mind, we propose Optimal Transport (OT) (Villani, 2008) as a natural solution to simultaneously incorporate both the similarity between sample representations and the likelihood of the samples containing an event into a single framework. Therefore, we cast sample selection as an OT problem in which we attempt to find the best alignment between the samples from the source and target languages.

For our experiments, we focus on the widely used ACE05 Walker et al. (2006) and ERE (Song et al., 2015) datasets which, in conjunction, contain event-annotations in 4 different languages: English, Spanish, Chinese, and Arabic. We work on 8 different language pairs by selecting different languages as the source

and target. Our proposed model obtains considerable performance improvements (+ 2-3% in F1 scores) over competitive baselines and previously published results (M'hamdi et al., 2019b). We believe these results demonstrate our model's efficacy and applicability in creating CLED systems.

The rest of this chapter is organized as follows: section 3.2 provides a thorough description of our proposed model, section 3.3 presents and analyses the results from our experiments, section 3.4 provides a brief review of related work, and section 3.5 presents a summary of our findings.

3.2 Model

3.2.1 Problem Definition. Following prior works (Majewska et al., 2021b; M'hamdi et al., 2019b), we treat ED as a sequence labeling problem. Given a set $\mathcal{D}$ of word sequences $w_i = \{w_{i1}, w_{i2}, ..., w_{in-1}, w_{in}\}$ and their corresponding label sequences $y_i = \{y_{i1}, y_{i2}, ..., y_{in-1}, y_{in}\}$, we use an encoder network E to obtain a contextualized vector representation of the words in the input sequence $\mathbf{h_i} = E(w_i) = \{h_{i1}, h_{i2}, ..., h_{in-1}, h_{in}\}$. Using such representations as input, a prediction network P computes a distribution over the set of possible labels and is trained in a supervised manner using the negative log-likelihood function $\mathcal{L}_P$:

$$\mathcal{L}_P = -\sum_{i=1}^{|D|}\sum_{j=1}^{n} \log P(y_{ij}|h_{ij}) \tag{3.1}$$

In the cross-lingual transfer-learning setting, the data used to train the model and the data on which the model is tested come from different languages known as the *source* and *target*, respectively. As such, we deal with two datasets $\mathcal{D}_{src}$ and $\mathcal{D}_{tgt}$. Furthermore, we assume a zero-shot setting, i.e., we do not have access to the gold labels of the target language y_{tgt}, other than to evaluate our CLED model at testing time.

Our goal is to define a model able to generate language-invariant word representations that are refined enough so that cross-lingual issues, such as the ones described in section 3.1, are properly handled.

3.2.2 Baseline Model. Here, we briefly describe the BERT-CRF model proposed by M'hamdi et al. (2019b) which was the previous state-of-the-art and serves as our main baseline. Using multilingual BERT (mBERT, Devlin et al., 2019) as its encoder, BERT-CRF generates robust, contextualized representations for words from different languages. For words that are split into multiple word-pieces, the average of the representation vectors for all comprising sub-pieces is used as the representation of the full word.

For classification purposes, instead of assigning the labels of each token independently, BERT-CRF uses a Conditional Random Field (CRF) (Lafferty et al., 2001) layer on top of the prediction network to better capture the interactions between the label sequences. In summary, the contextualized representation vectors h_i generated by the mBERT encoder from the words in the sequence are then fed to a CRF layer which finds the optimal label sequence.

3.2.3 Adversarial Language Adaptation. The pre-trained versions of MLMs like mBERT or XLM-RoBERTa (Conneau et al., 2019) generate contextualized representations with a certain degree of language invariance. This can be confirmed by their successful application in cross-lingual settings (Majewska et al., 2021b; M'hamdi et al., 2019b). However, a lingering issue is the difficulty of learning the nuances of the target language such as verb variations that do not exist in the source language used to train them. Majewska et al. (2021b), for instance, propose to address this issue by injecting external verb knowledge into the encoder via task-specific adapter modules (Pfeiffer et al., 2020).

It is our intuition, however, that these issues can be mitigated by achieving a more refined level of language invariance in word representations. As such, we propose using Adversarial Language Adaptation (ALA) (Joty et al., 2017), a technique used to create language-invariant models. The ALA framework consists in including a *Language Discriminator* (LD) whose purpose is to learn language-dependent features and be able to differentiate between the samples from either the source or the target languages.

A fundamental characteristic of the ALA approach is its lack of requirements for annotated data in the target language. As such, we can use data from both $\mathcal{D}_{src}$ and $\mathcal{D}_{tgt}$. An auxiliary dataset $D_{aux} = \{(w_1, l_1), \ldots, (w_{2m}, l_{2m})\}$ is created where w_i is a text sequence from either $\mathcal{D}_{src}$ or $\mathcal{D}_{tgt}$, and l_i is a language label. The cardinality of D_{aux} is $|D_{aux}| = 2m$, where m is equal to the batch size. Text samples $w_1 \ldots w_m \in \mathcal{D}_{src}$, and samples $w_{m+1} \ldots w_{2m} \in \mathcal{D}_{tgt}$. As described earlier, the encoder E receives the text sequences and produces a sequence of contextualized representations $E(w_i) = h_i = \{h_{i0}, h_{i1}, h_{i2}, \ldots, h_{in}\}$ where h_{i0} is the representation of the `[CLS]` token added at the beginning of every input sequence.

In our work, the LD is a simple Multi-Layer Perceptron(MLP) network that takes h_{i0} as input and produces a single sigmoid output. It's trained with the usual *binary cross-entropy* loss function objective:

$$LD_{loss} = \arg\min_{LD} \mathcal{L}(LD(h_{i0}), l_i) \tag{3.2}$$

As the LD learns to distinguish between the source and target languages, we concurrently train the encoder to *"fool"* the discriminator. In other words, the encoder must learn to generate representations that are language-invariant enough that the LD is unable to classify them while still remaining predictive for event-

trigger classification. We optimize the following loss:

$$\arg\min_{E,C} \sum_{j=1}^{n} (\mathcal{L}(C(h_{ij}), y_{ij})) - \lambda \mathcal{L}(LD(h_{i0}, l_i)) \tag{3.3}$$

Where C refers to the CRF-based classifier network and λ is a hyperparameter.

Equation 3.3 is implemented by using a Gradient-Reversal Layer (GRL) (Ganin & Lempitsky, 2015) which acts as the identity during the forward pass, but reverses the direction of the gradients during the backward pass. The first term in Equation 3.3 can, of course, only be applied to annotated data from the source language as target data labels are unavailable.

The GRL is applied to the input vectors, h_{i0}, of the LD. This way, the LD is being trained to differentiate between the two languages while the encoder is trained in the opposite direction, i.e. to generate sequence representations that are harder to discriminate.

3.2.4 Adversarial Training Optimization. ALA has already been shown to be effective at generating language-invariant models (X. Chen et al., 2018; Joty et al., 2017). However, in regular ALA training, all samples in a batch, from both the source and target domains, are treated equally. That is, all samples are used as examples for the discriminator to learn how to better discern between the two domains. We propose that ALA effectiveness can be further improved by carefully selecting the samples with which to train the discriminator. We argue that some samples might be more informative than others and that better adaptation results can be achieved by only using such informative samples during training.

We base our notion as to *what* makes a sample more informative on two factors. First, we argue that presenting the LD with examples from the source and target language that are too dissimilar makes its task easier which, in turn, leads to the LD not learning the fine-grained distinctions between the languages. Instead,

we propose using samples whose vector representations h_{i0} are close to each other in the embedding space. The intuition for this being that, as representations capture the contextual semantics of the samples, closer representations correspond to more similar examples. Second, we suggest that presenting the LD with samples containing events should make the encoder incorporate task-specific information into its representations.

3.2.4.1 Optimal Transport. One challenge of using the two mentioned criteria for the ALA sample selection process is that they come with two different measures which are hard to combine. To address this, we propose using Optimal Transport (OT, Villani, 2008) as a natural way to combine these two metrics into a single framework for sample selection. Optimal transport is, in broad terms, the problem of finding out the cheapest transformation between two discrete probability distributions. It requires a cost function to determine the cost of transforming a data point in one distribution into a data point in the second distribution. When the cost function is based on a valid distance function, the minimum cost is known as the Wasserstein distance. Formally, it solves the following optimization problem:

$$\pi^*(s,t) = \min_{\pi \in \prod(s,t)} \sum_{s \in \mathcal{S}} \sum_{t \in \mathcal{T}} \pi(s,t)\, C(s,t)\, ds\, dt \tag{3.4}$$

$$\textbf{s.t. } s \sim p(s) \text{ and } t \sim q(t)$$

where $\mathcal{S}$ and $\mathcal{T}$ are the two domains to be transformed; $p(s)$ and $q(t)$ are the probability distributions of $\mathcal{S}$ and $\mathcal{T}$, respectively; C is a cost function for mapping $\mathcal{S}$ to $\mathcal{T}$, $C(s,t) : \mathcal{S} \times \mathcal{T} \longrightarrow \mathbb{R}_+$; and finally, $\pi^*(s,t)$ is the optimal joint distribution over the set of all joint distributions $\prod(s,t)$. The problem described by Equation 3.4 is, of course, intractable. Therefore, we use instead the Sinkhorn

algorithm (Cuturi, 2013) which is an entropy-based relaxation of the discrete OT problem.

3.2.4.2 Problem Formulation. We formulate the OT problem as follows: the domains $\mathcal{S}$ and $\mathcal{T}$ are defined as the representation vectors of the text samples in either the source h^s_{i0} or the target h^t_{j0} languages. We use the L2 distance between these representations as the cost function:

$$C(h^s_{i0}, h^t_{j0}) = ||h^s_{i0} - h^t_{j0}||^2_2 \tag{3.5}$$

To define the marginal probability distributions $p(s)$ and $q(t)$ for the $\mathcal{S}$ and $\mathcal{T}$ domains, we propose including an Event-Presence (EP) prediction module and use its normalized likelihood scores as the probability distributions for $\mathcal{S}$ and $\mathcal{T}$. Thus, the auxiliary dataset D_{aux} is augmented to include an event-presence label e_i for each sample. Of course, this can only be done for samples in the source language as the labels for the target-language data are unavailable:

$$D_{aux} = \{(w_1, l_1, e_1), \ldots, (w_m, l_m, e_m),$$
$$(w_{m+1}, l_{m+1}), \ldots, (w_{2m}, l_{2m})\}$$

The EP module is then trained to optimize the following loss:

$$EP_{loss} = \arg\min_{EP} \mathcal{L}(EP(h_{i0}), e_i) \tag{3.6}$$

where $i <= m$, i.e., only using samples from the source language.

The probability distributions $p(s)$ and $p(t)$ are the computed as follows:

$$p(s) = Softmax(EP(h^s_{i0}) \mid l_i == s) \tag{3.7}$$

$$p(t) = Softmax(EP(h^t_{i0}) \mid l_i == t) \tag{3.8}$$

3.2.4.3 Sample Selection. Once the OT optimization problem is solved, we leverage the solution matrix π^*, where an entry $\pi^*(s,t)$ represents the

optimal cost of transforming data point $s \in \mathcal{S}$ into $t \in \mathcal{T}$, to compute an the overall similarity score v_i of a sample $h_{i0} \in \mathcal{S}$ to the samples in the target domain $\mathcal{T}$ by using the average distance:

$$v_i = \frac{\sum_j^m \pi^*(h_{i0}^s, h_{j0}^t)}{m} \tag{3.9}$$

Correspondingly, we compute an overall similarity score v_j of each sample $h_{j0} \in \mathcal{T}$ to the samples in the source domain $\mathcal{S}$:

$$v_j = \frac{\sum_i^m \pi^*(h_{i0}^s, h_{j0}^t)}{m} \tag{3.10}$$

Finally, we select a fraction (determined by hyperparameter γ) of samples with the best similarity scores from both the source and target languages and only utilize these selected samples during ALA training.

3.2.5 OACLED Model. The architecture of our Optimized Adversarial Cross-Lingual Event Detection (OACLED) model is shown in Figure 2. The model is then trained end-to-end with the following loss objective:

$$L_{full} = CRF_{loss} + \alpha LD_{loss} + \beta EP_{loss} \tag{3.11}$$

where α and β are trade-off hyperparameters. Figure 3 visualizes the loss computation.

3.3 Experiments

3.3.1 Datasets. We evaluate our model on the ACE05 (Walker et al., 2006) dataset which includes annotated event-trigger data in 3 languages: English, Chinese and Arabic. To include an additional language in our experiments, we also evaluate on the ACE05-ERE (Song et al., 2015) dataset which has annotated data in English and Spanish. Note that the ACE05 and ERE datasets do not share the same label set: ACE05 involves 33 distinct event types while ERE involves

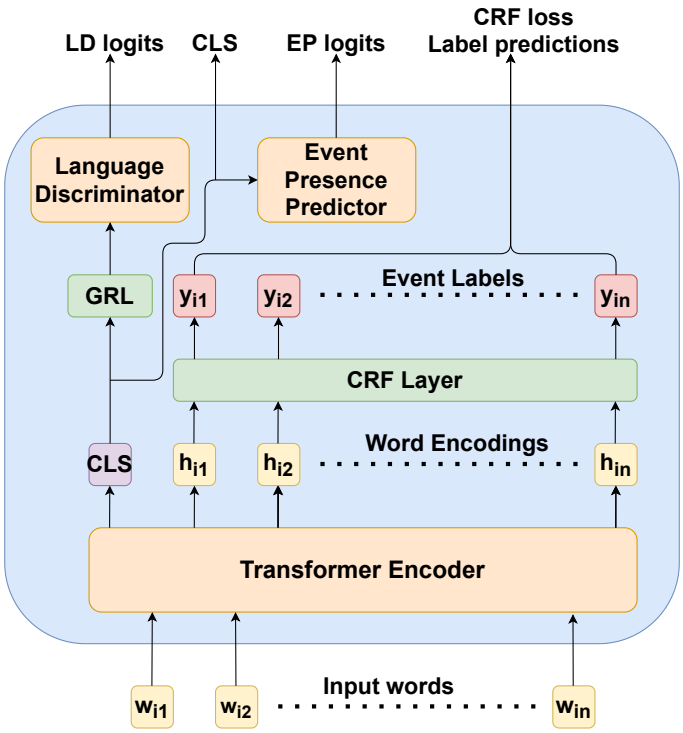

Figure 2. OACLED model architecture. Word representations generated by the encoder are fed to a CRF layer which generates label predictions. The sentence-level representations are fed to the EP predictor and the LD to obtain their corresponding logits outputs.

38 event types. We follow the same data pre-processing and splits as in previous work M'hamdi et al. (2019b) to ensure a fair comparison. Table 1 presents the data statistics.

3.3.2 Main Results. In our experiments, we work with 8 distinct language pairs by selecting each of the available languages as either the source or target language: *English-Chinese, Chinese-English, English-Arabic, Arabic-English, Chinese-Arabic, Arabic-Chinese, English-Spanish, and Spanish-English.* The *Chinese-Spanish, Spanish-Chinese, Arabic-Spanish*, and *Spanish-Arabic* language combinations are unavailable due the previously mentioned incompatibility between the event type sets in ACE05 and ACE05-ERE.

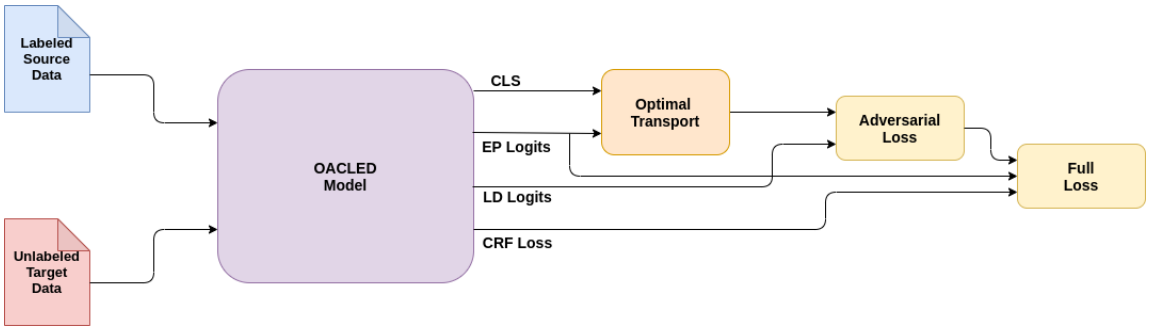

Figure 3. OACLED model loss computation. The sentence-level representations and the EP logits are used as inputs to the OT optimization. Then the LD logits from the selected samples are used to compute the adversarial loss.

We compare our OACLED model against 3 relevant baselines. First, the previous state-of-the-art CLED model BERT-CRF (M'hamdi et al., 2019b) as described in section 3.2.2. Second, the mBERT-2TA model (Majewska et al., 2021b) that aims at improving cross-lingual performance by incorporating language-independent verb knowledge via task-specific adapters. And third, XLM-R-CRF, a model that is equivalent in all regards to BERT-CRF except that it uses XLM-RoBERTa Conneau et al. (2019) as the encoder.

Table 2 and Table 3 show the results of our experiments on the ACE05 and ERE datasets, respectively. In all our experiments, we use the base transformer versions *bert-base-cased* and *xlm-roberta-base* as the encoders, parameters are tuned on the development data of the source language, and all entries are the average of five runs.

From Tables 2 and 3, it should be noted that there is a substantial performance increase by performing the trivial change of replacing mBERT with XLM-RoBERTa as the encoder. Furthermore, our OACLED model clearly, and consistently, outperforms the baselines for all language pairings, with the exception of the *Chinese-Arabic* pair. We attribute this to the impaired performance of

Dataset	Language	Split	Sentences	Events
ACE05	**English**	Train	19,240	4,419
		Dev	902	468
		Test	676	424
	Chinese	Train	6,841	2,926
		Dev	526	217
		Test	547	190
	Arabic	Train	2,555	1,793
		Dev	301	230
		Test	262	247
ERE	**English**	Train	14,219	6,419
		Dev	1,162	552
		Test	1,129	559
	Spanish	Train	7,067	3,272
		Dev	556	210
		Test	546	269

Table 1. Dataset statistics.

XLM-RoBERTa as the encoder for that specific pair as can be confirmed by the poor performance of the XLM-R-CRF baseline on the same configuration. Most importantly, OACLED's improvement over the XLM-R-CRF baseline is present in every configuration, which validates the effectiveness of our optimized approach to ALA training.

The model implementation details can be found in Appendix C.

3.3.3 Ablation Study. We identify 2 main components in our approach: (1) leveraging ALA to create refined language-invariant representations and (2) optimizing the adversarial training process by selecting a subset of samples chosen with OT to incorporate our measures of informativeness into the sample-selection process. As expected, removing ALA training entirely restores the model to the baseline. However, adversarial training optimization via OT has various aspects to it. In order to understand the contribution of these aspects, we explore four different configurations: *OACLED-OT* presents the effects of removing sample

		Target		
Source	**Model**	**English**	**Chinese**	**Arabic**
English	**BERT-2TA**	X	46.9*	29.3*
	BERT-CRF	X	68.5*	30.9*
	XLM-R-CRF	X	70.49±0.85	43.54±2.77
	OACLED	X	**74.64**±0.73	**44.86**±3.1
Chinese	**BERT-CRF**	37.52±1.73	X	**35.05**±2.85
	XLM-R-CRF	41.72±1.4	X	32.76±2.31
	OACLED	**45.77**±1.45	X	34.48±2.43
Arabic	**BERT-CRF**	40.1±3.26	58.78±2.33	X
	XLM-R-CRF	45.22±1.82	61.76±1.57	X
	OACLED	**47.98**±2.07	**63.13** ±1.7	X

Table 2. Results on the ACE05 dataset with standard deviation across random seeds. Entries marked * are taken directly from the original publications.

		Target	
Source	**Model**	**English**	**Spanish**
English	**BERT-CRF**	X	43.28±2.01
	XLM-R-CRF	X	46.79±1.34
	OACLED	X	**47.69**±1.63
Spanish	**BERT-CRF**	39.8±2.27	X
	XLM-R-CRF	45.61±1.76	X
	OACLED	**47.5**±1.89	X

Table 3. Results on ACE05-ERE dataset with standard deviation across random seeds.

selection entirely and using all available samples to train the LD; *OACLED-L2* uses a constant distance between the unlabeled samples instead the standard L2 distance used in the Sinkhorn algorithm; *OACLED-EP* completely removes the EP module and a uniform distribution is used as the probability distributions for both languages; finally, *OACLED-ED-Loss* keeps the EP module, but removes its EP_{loss} term from Equation 3.11. The performance results of these models are presented in Table 4. In this and the following sections (3.3.4, 3.3.5.2), we present the results of experiments using English as the sole source language as it is the source language most ubiquitously used. We, however, found consistency in the displayed effects for different source/target language configurations.

Model version	Target Language		
English	Chinese	Arabic	Spanish
OACLED-OT	70.94	40.55	44.96
OACLED-L2	71.35	41.79	44.39
OACLED-EP	73.08	42.81	46.99
OACLED-EP-Loss	72.93	43.4	46.35
OACLED (*full*)	**74.64**	**44.86**	**47.69**

Table 4. Ablation experiment results

As expected, removing the sample selection through OT leads to the worst performance drop. This highlights the importance of selecting informative examples for the LD. Furthermore, removing the cost function also hurts performance greatly, which shows that a proper distance function is needed for the OT algorithm to work effectively. While the effects of removing the EP module and its corresponding loss term are not of the same magnitude, they are still significant. These results support our claim for the need and utility of all the components in our approach, showing that their inclusion is crucial in achieving state-of-the-art performance.

3.3.4 Language Model Finetuning. A key contribution of our approach is to exploit unlabeled data in the target language, which is usually abundant, by introducing it into the training process to improve our model's language-invariant qualities.

To confirm the utility of our approach, Table 5 contrasts our model's performance against a baseline whose encoder has been finetuned with the same unlabeled data using the standard masked language model objective.

Model Version	Target Language		
English	Chinese	Arabic	Spanish
Finetuned XLM-R	71.06	43.71	**47.82**
OACLED	**74.64**	**44.86**	47.69

Table 5. OACLED performance versus a baseline using an encoder finetuned with unlabeled data.

It can be observed that our model outperforms the finetuned baseline in two out of the three target languages. Additionally, the difference in performance in those two instances is considerably larger (3.58% and 1.15%), than the setting in which the baseline performs better (0.13%).

3.3.5 Analysis. This section analyzes our model's outputs to gain insights into its strengths and weaknesses.

3.3.5.1 Learned Representation Distances. First, we look at the distance between the sentence-level representations h_{i0} generated by the encoder for different source/target language pairs. Figure 4 shows a plot of such distances using cosine distance as the distance function.

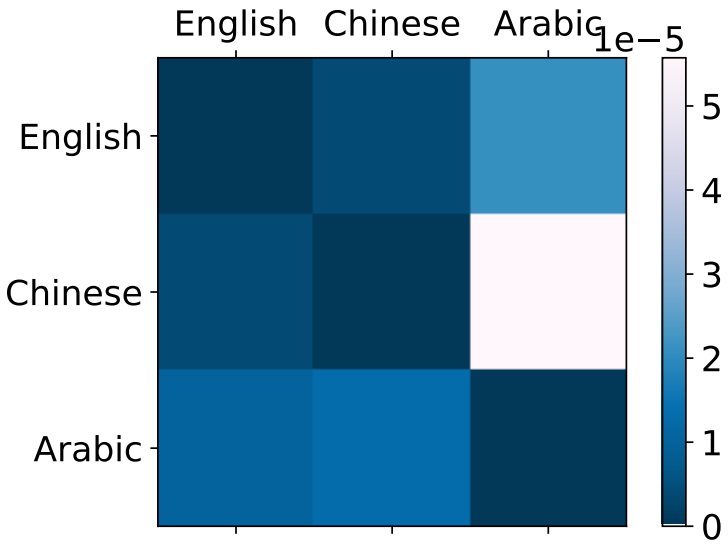

Figure 4. Distance between sentence representations for different language pairs.

When computing the correlation with the performance results in Table 2, we obtain a score $R = -0.6616$, meaning there is moderate negative correlation between the distance of the representations and model performance, i.e. closer representations lead to better performance.

Similarly, Table 6 shows a comparison of the distances between the representations generated by OACLED and those obtained by the XLM-R-CRF baseline.

	Cosine Distance	
Source/Target	**Baseline**	**OACLED**
English/Chinese	3.64e-3	**3.93e-6**
English/Arabic	7.71e-2	**2.08e-5**
English/Spanish	5.4e-3	**5.3e-6**
Chinese/English	3.62e-3	**3.87e-6**
Arabic/English	4.16e-2	**1.02e-5**
Spanish/English	6.87e-3	**1.49e-5**

Table 6. Comparison of representation-vector distances for language pairs between our model and the baseline.

We observe that OACLED representations are closer, by several orders of magnitude, than those obtained by the baseline. This supports our claim that our model's encoder generates more refined language-invariant representations than those obtained by the default version of XLM-RoBERTa.

3.3.5.2 Access to Labeled Target Data. It was previously discussed that a key feature of our approach is that it does not require annotated data in the target language and, instead, leverages the use of unlabeled data which is readily available. Nonetheless, we also explore the performance of our model in the event that there exists a small amount of annotated target data available. Figure 5 shows the results of our experiments when using different amounts of labeled target data during training.

It can be observed that OACLED consistently outperforms the baseline even when there is some availability of annotated data. Additionally, performance steadily increases as more and more data is used. This conforms to expectations,

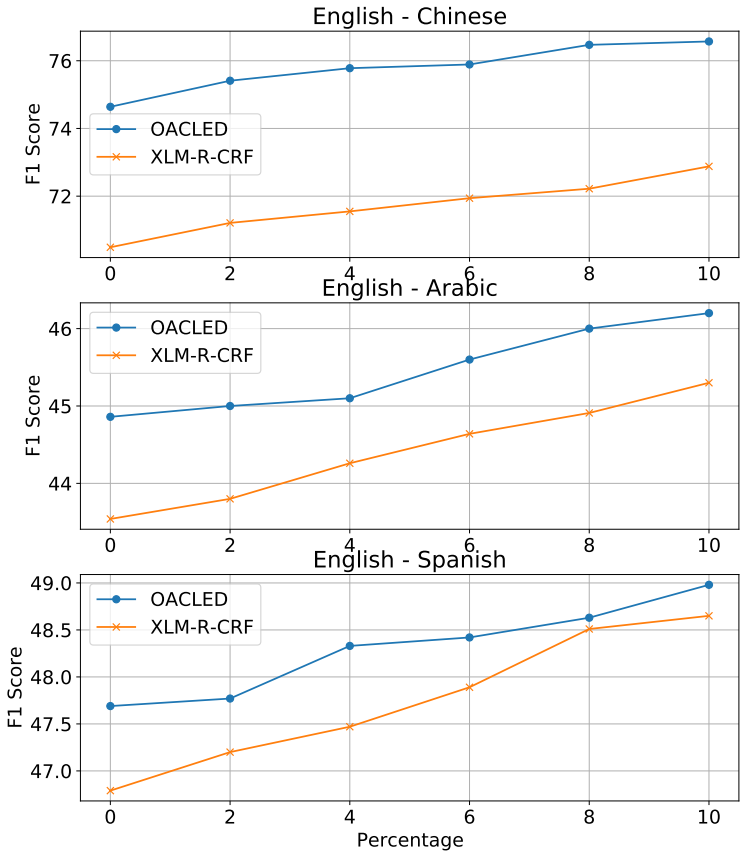

Figure 5. Model performance when training on small quantities of labeled target data. The X axis presents the percentage (0 - 10%) of data used out of the entire training set of the target language.

and confirms that having labeled data in the target language available for training is ultimately beneficial to the model's performance.

3.3.5.3 Case Study. Next, we look into our model's predictions and analyze instances where it outperforms the baseline to exemplify the advantages of dealing with optimized language-invariant representations. We identify two important patterns.

First, our model seems to better classify events in the target language that involve trigger words that have distinct connotations that depend on context. Especially those that are two distinct words in the source language. For example,

the Spanish word *"juicio"* can have two distinct meanings that are different words in English: *"trial"* and *"judgment"*. Our model correctly classifies it as a JUSTICE:TRIAL-HEARING trigger in the sentence *"Dos llamados a* ***juicio*** *fueron hechos por un jurado federal investigador"*. Meanwhile, the baseline fails to even recognize it as a trigger. Another example is the word *"detenido"*, an adjective that can mean both *"detained"*, in a criminal context, and *"stopped"*, as in halted. Our model correctly classifies it in the sentence *"Padilla no debería permanecer* ***detenido*** *durante meses alejado de otros reos"* as a JUSTICE:ARREST-JAIL trigger while the baseline fails to detect the event. We manually identified 23 of these polysemous triggers in the Spanish[2] test set and found that 19 (82.6%) were correctly classified by our OACLED model versus 14 (60.8%) by the baseline (27.8% improvement).

Additionally, we found our model correctly classifies verb conjugation variants that do not exist in the source language. For instance, our model correctly recognizes the words *"venderlos", "vender", "vendes", and "vendedor"* (variants of the verb *"to buy"*) as TRANSACTION:TRANSFER-OWNERSHIP triggers whereas the baseline incorrectly classifies them as being of the TRANSACTION:TRANSFER-MONEY type. As previously mentioned, Majewska et al. (2021b) propose injecting external verb-knowledge into the training to help with verb interpretation for event extraction. Our empirical results, however, outperform their reports which appears to imply that, at least for CLED, holistically learning the language-invariant features shared between the target and source languages works better than injecting language-specific verb knowledge.

[2]We use Spanish for the analysis as it is the mother tongue of the first author.

Another similar example are the trigger words *"matar", "mató", "homicidio", "asesinato"*, all of which are variations that refer to the act of killing or murdering. Our model correctly tags them as LIFE:DIE events while the baseline incorrectly classifies them as CONFLICT:ATTACK.

We believe these findings illustrate how, by introducing additional context in the form of unlabeled data, the model is able to learn fine-grained word representations that better capture the semantics of the words in the target language, and successfully deal with difficult cross-lingual issues.

3.4 Related Work

Research efforts on monolingual ED are extensive and varied. Hand-crafted, feature-based, language-specific methods were the basis of early ED approaches (Ahn, 2006; Hong et al., 2011; Ji & Grishman, 2008; Q. Li, Ji, & Huang, 2013; Liao & Grishman, 2010a, 2010b; McClosky, Surdeanu, & Manning, 2011; Miwa, Thompson, Korkontzelos, & Ananiadou, 2014; Patwardhan & Riloff, 2009; B. Yang & Mitchell, 2016). More recent efforts have primarily made use of deep learning techniques such as convolutional neural networks (Y. Chen et al., 2015; T. H. Nguyen, Fu, et al., 2016; T. H. Nguyen & Grishman, 2015b), recurrent neural networks (V. D. Lai, Nguyen, & Nguyen, 2020; T. H. Nguyen, Cho, & Grishman, 2016; Sha et al., 2018), graph convolutional networks (M. V. Nguyen, Lai, & Nguyen, 2021; T. H. Nguyen & Grishman, 2018; Yan, Jin, Meng, Guo, & Cheng, 2019), adversarial networks (Hong, Zhou, Zhang, Zhou, & Zhu, 2018; T. Zhang, Ji, & Sil, 2019), and pre-trained language models (J. Liu et al., 2020; Pouran Ben Veyseh, Lai, Dernoncourt, & Nguyen, 2021; Pouran Ben Veyseh, Nguyen, Ngo Trung, Min, & Nguyen, 2021; Wadden et al., 2019; S. Yang et al., 2019a; J. Zhang et al., 2019; Y. Zhang et al., 2020).

Works on cross-lingual ED are not as prevalent and generally make use of cross-lingual resources employed to address the differences between languages such as bilingual dictionaries or parallel corpora (J. Liu et al., 2019; Muis et al., 2018b) and, more recently, pre-trained multilingual language models (Hambardzumyan, Khachatrian, & May, 2020b; Majewska et al., 2021b; M'hamdi et al., 2019b). Unlike these previous efforts, our method leverages unlabeled data to further refine the language-invariant qualities of the language models.

Adversarial Language Adaptation, inspired by models in domain adaptation research (Ganin & Lempitsky, 2015; Naik & Rose, 2020; Ngo Trung, Phung, & Nguyen, 2021), has been successfuly applied at generating language-invariant models (X. Chen et al., 2018; Joty et al., 2017; M. V. Nguyen, Nguyen, et al., 2021). Our method improves upon these approaches optimizing the adversarial training process by selecting the most informative examples from the unlabeled data.

Additional examples of downstream applications of cross-lingual learning are document classification (Holger & Xian, 2018), named entity recognition (Xie, Yang, Neubig, Smith, & Carbonell, 2018a) and part-of-speech tagging (Cohen, Das, & Smith, 2011). For a thorough review on cross-lingual learning, we refer the reader to Pikuliak, Šimko, and Bieliková (2021b).

3.5 Summary

In summary, we consider the main contributions of this chapter to be the following:

- We propose a novel deep-learning-based model for CLED that leverages the use of unlabeled data to learn fine-grained language-invariant representations by optimizing the standard ALA training through optimal-transport-based sample selection.

– We perform extensive experiments on 4 different languages from unrelated language families, used both as source and target for a total of 8 language pairings. Our state-of-the-art results confirm our model's effectiveness across languages of diverse characteristics and structures. We believe these results demonstrate our model's robustness and effectiveness at generating refined language-invariant representations that allow for better event detection results.

– An insightful analysis of our model's intermediate outputs and predictions confirms that OACLED's representations are indeed closer to each other and this proximity translates into better handling of difficult cross-lingual instances.

– We also note that, while we focus our experiments on the ED task, our proposed optimization of the adversarial training process is task-independent and can be generalized to other related cross-lingual tasks when leveraging ALA is deemed beneficial.

CHAPTER IV

LEVERAGING HYBRID TRANSFER FOR CROSS-LINGUAL EVENT DETECTION

This Chapter contains materials from the unpublished paper "*Luis F. Guzman-Nateras, Franck Dernoncourt, and Thien H. Nguyen. '**Hybrid Knowledge Transfer for Improved Cross-Lingual Event Detection via Hierarchical Sample Selection.**' To appear in the Proceedings of the 61st annual meeting of the Association for Computational Linguistics, 2023*" (Guzman-Nateras, Dernoncourt, & Nguyen, 2023). As the first author of this publication, Luis was responsible for all areas of the project from initial conceptualization to development, experimentation, and final document writing. Thien and Franck made editorial suggestions for the final document. The original paper contents have undergone some editorial updates to comply with this document's format and purpose.

Most recent CLED efforts, including our approach discussed in Chapter III, follow a direct-transfer approach. However, we argue that these methods fail to take advantage of the benefits of the data-transfer approach where a cross-lingual model is trained on target-language data and is able to learn task-specific information from syntactical features or word-label relations in the target language. As such, in this chapter, we propose a hybrid knowledge-transfer approach that leverages a teacher-student framework where the teacher and student networks are trained following the direct and data transfer approaches, respectively. Our method is complemented by a hierarchical training-sample selection scheme designed to address the issue of noisy labels being generated by the teacher model. We evaluate

our model on 9 morphologically-diverse target languages across 3 distinct datasets, highlighting the importance of exploiting the benefits of hybrid transfer.

4.1 Introduction

Event Detection (ED) is a sub-task of the encompassing Information Extraction (IE) Natural Language Processing (NLP) task. The main objective of ED is to detect and categorize the *event triggers* in a sentence, i.e., the words that most clearly indicate the occurrence of an event. Event triggers are known to be frequently related to the verb in a sentence (Majewska et al., 2021b). However, they can also be other parts of speech such as nouns or adjectives. For instance, in the sentence *"The ceremony was chaired by the* ***former*** *Secretary of State"*, an ED system should recognize *former* as the trigger of a `Personnel:End-Position` event[1].

Generating labeled data for IE tasks such as ED can be a long and expensive endeavor. As such, most labeled ED datasets pertain to a small set of popular languages (e.g., English, Chinese, Spanish). In turn, labeled data is scarce or non-existent for a vast majority of languages. This imbalance in annotated data availability has prompted many research efforts into zero-shot cross-lingual transfer learning which attempts to transfer knowledge obtained from annotated data in a high-resource *source* language to a low-resource *target* language for which no labeled data is available. There are two predominant knowledge-transfer paradigms employed by such cross-lingual methods: *Data transfer* and *Direct transfer*.

Approaches that adhere to the *data transfer* paradigm generate pseudo-labeled data in the target language and then train a model on such data. This pseudo-training data can be constructed by mapping the gold source labels into

[1]Event type taken from ACE05 dataset.

parallel, or translated, versions of the source data, or by leveraging source-trained models to annotate unlabeled target data. Since models in this category are trained on the target language, they can directly exploit word-label relations and other target-language-specific information such as word order and lexical features (Xie et al., 2018b). However, annotated parallel corpora are extremely scarce, and misaligned or incorrect translations introduce noise that affects the model performance.

In contrast, *direct-transfer-based* approaches aim at creating cross-lingual models by training them with delexicalized, language-independent features obtained from the labeled, source-language data. The resulting language-agnostic models can then be applied directly to unlabeled data in the target language.

In recent years, direct transfer has become the favored transfer paradigm as such models have less need for cross-lingual resources and can be applied to a broader range of languages. As such, previous research efforts on Cross-Lingual Event Detection (CLED) have mostly focused on the direct transfer approach (Majewska et al., 2021b; M'hamdi et al., 2019b) and, in consequence, have failed to exploit the aforementioned advantages of training with target-language data.

More recent approaches have attempted to address this issue by incorporating unlabeled target-language data into the training process. For example, M. V. Nguyen, Nguyen, et al. (2021) propose a class-aware, cross-lingual alignment mechanism where they align examples from the source and target languages based on class information. Our OACLED model (Guzman-Nateras, Nguyen, & Nguyen, 2022) discussed in Chapter III proposes instead to improve standard Adversarial Language Adaptation (ALA) (X. Chen et al., 2018; Joty et

al., 2017) by only presenting the language discriminator with *informative* samples. Despite their improved results, these models only learn task-related information from the source language and fail to make use of the potentially useful information contained in word-label relations in the target language. Furthermore, previous studies on similar tasks have shown that, even for direct transfer methods, lexical features are useful if the source and target languages are close to each other (Tsai et al., 2016).

Given that the data transfer and direct transfer paradigms are orthogonal, in this chapter we present a *hybrid transfer* approach for cross-lingual event detection that (1) exploits the desirable features of both and (2) minimizes their respective shortcomings. For this purpose, we propose a *knowledge distillation* framework which has already been proven effective on similar cross-lingual tasks (W. Chen et al., 2021; Liang et al., 2021; Q. Wu, Lin, Karlsson, Lou, & Huang, 2020; Q. Wu, Lin, Karlsson, Huang, & Lou, 2020). In our proposed framework, a teacher model is trained using a direct transfer approach (i.e., with language-invariant features obtained from annotated source data) and applied to unlabeled target-language data. Then, this pseudo-labeled data is utilized to train a student model so that it benefits from the advantages of the data transfer paradigm.

Nonetheless, we recognize that the pseudo-labels obtained from the teacher model are prone to containing noisy predictions which can be hurtful for student training. To address this issue, we argue that the teacher model should produce more dependable predictions on target-language examples that share some similarities with their source-language counterparts. As such, we propose to improve the teacher-student learning process by restricting student

training to samples with such desirable characteristics. We perform our training-sample selection in a hierarchical manner: First, we leverage Optimal Transport (OT, Villani, 2008) to compute similarity scores between batch samples in the source and target languages. Only samples with similarity scores above a certain threshold are selected in this first step. OT has already been shown to be effective at estimating cross-lingual similarities for sample selection (Guzman-Nateras, Nguyen, & Nguyen, 2022; Phung, Minh Tran, Nguyen, & Nguyen, 2021). Then, in the second step, we make use of Cross-domain Similarity Local Scaling (CSLS, Conneau, Lample, Ranzato, Denoyer, & Jégou, 2018) to refine our sample selection. CSLS provides an enhanced measure to obtain reliable matches between samples in the source and target languages by addressing the *hubness* phenomenon that plagues nearest-neighbor-based pair-matching methods. The student model is then trained on the hierarchically-selected target-language samples exclusively.

In order to validate our approach, we compare our model's performance against current state-of-the-art models for CLED. For this purpose, we report our results on the most commonly used CLED benchmarking datasets: ACE05 Walker et al. (2006) and ACE05-ERE (Song et al., 2015). These datasets, in conjunction, contain ED annotations for 3 distinct target languages. Our experimental results show that our approach consistently outperforms such state-of-the-art CLED models. Additionally, we further evaluate the flexibility and applicability of our model by leveraging the recently released MINION dataset (Pouran Ben Veyseh et al., 2022) which contains ED annotations for 8 typologically different languages.

The remainder of this chapter is organized as follows: section 4.2 presents the definition of the ED task and an in-depth description of our model and approach, section 4.3 includes the main results from our experiments and related

analysis, section 4.4 provides a review of previous relevant work, and finally, section 4.5 presents a summary of our conclusions.

4.2 Model

4.2.1 Event Detection: Problem Definition. We follow a similar approach to previous CLED efforts (Guzman-Nateras, Nguyen, & Nguyen, 2022; Majewska et al., 2021b; M'hamdi et al., 2019b) and model the ED task as a sequence labeling problem.

Given a group of sentences $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ where each of such sentences is considered as a sequence of tokens $s_i = \{t_{i1}, t_{i2}, \ldots, t_{im}\}$ accompanied by a corresponding label sequence $y_i = \{y_{i1}, y_{i2}, ..., y_{im}\}$, the main idea is to train a model to generate token-level contextualized representations which can then be used to predict token-level labels.

In broad terms, a sequence labeling model consists of an encoder $\mathcal{E}$ and a classifier $\mathcal{C}$. The encoder consumes a sequence of input tokens t_i and outputs a sequence of contextualized representations h_i (Eq. 4.1). These representations are then fed to the classifier which produces a probability distribution over all of the possible types. A candidate label is selected by choosing the type with the largest probability. The model loss $\mathcal{L}_\mathcal{C}$ is then computed via negative log-likelihood with the classifier-selected labels and the expected *gold* labels (Eq. 4.2).

$$h_{i1}, h_{i2}, \ldots, h_{im} = \mathcal{E}(t_{i1}, t_{i2}, \ldots, t_{im}) \tag{4.1}$$

$$\mathcal{L}_\mathcal{C} = -\frac{1}{n * m} \sum_{i=1}^{n} \sum_{j=1}^{m} \log \mathcal{C}(y_{ij}|h_{ij}) \tag{4.2}$$

4.2.1.1 Zero-shot Cross-lingual Event Detection. In a cross-lingual setting, different languages are utilized during the training and testing

phases. The language utilized during training is referred to as the *source* language. Once training is complete, the model is tested on the so-called *target* language.

A zero-shot setting further assumes that there is no labeled data in the target language to be leveraged during training. Nonetheless, raw, unlabeled target-language text can usually be collected without major difficulties. As such, in our work, we assume the availability of two distinct sets of sentences during training: the labeled source sentences $\mathcal{S}_{src}$ and unlabeled target sentences $\mathcal{S}_{tgt}^{unl}$. For model evaluation purposes, we leverage a set of labeled target-language sentences $\mathcal{S}_{tgt}$.

4.2.2 Hybrid Knowledge Transfer. As mentioned in Section 4.1, we propose to combine the direct transfer and data transfer approaches by leveraging a *Knowledge Distillation* framework. Knowledge distillation was originally proposed as a way to compress models by transferring knowledge from a larger *teacher* model onto a smaller *student* model (Bucilua, Caruana, & Niculescu-Mizil, 2006). However, knowledge distillation has since been applied to several different tasks such as machine translation Weng, Yu, Huang, Cheng, and Luo (2020), automated machine learning (Kang, Mun, & Han, 2020), cross-modal learning (Hu, Xie, Hong, & Tian, 2020), and cross-lingual named entity recognition (W. Chen et al., 2021; Liang et al., 2021; Q. Wu, Lin, Karlsson, Lou, & Huang, 2020; Q. Wu, Lin, Karlsson, Huang, & Lou, 2020).

To the best of our knowledge, our approach is the first effort into leveraging a knowledge-distillation framework for CLED. The following sections present the details of our teacher and student models as well as our hierarchical data-sample selection strategy for student-model training.

4.2.2.1 Teacher Model. Our teacher model architecture follows that of previous direct-transfer-based models for CLED (Guzman-Nateras, Nguyen,

& Nguyen, 2022; Majewska et al., 2021b; M'hamdi et al., 2019b). We leverage a transformer-based pre-trained multilingual language model as the encoder $\mathcal{E}_T$. In particular, we make use of XLM-R Conneau et al. (2019) as it often outperforms multilingual BERT (Devlin et al., 2019) on the CLED task (Pouran Ben Veyseh et al., 2022). For the classifier $\mathcal{C}_T$, we employ a simple Feed-Forward Neural Network (FFNN) with 2 hidden layers (Eq. 4.3). A softmax operation is applied to the resulting predictions to obtain a probability distribution over the event types.

$$\mathcal{C}_T(y_{ij}) = \text{softmax}(W^{C_T 2} \text{ ReLU}(W^{C_T 1} h_{ij})) \tag{4.3}$$

where $W^{C_T 1}$ and $W^{C_T 2}$ are parameter matrices to be learned and $C_T(y_{ij}) \in \mathbb{R}^{|\mathbb{C}|}$ is the probability distribution over the event type set $\mathbb{C}$ for token $t_{ij} \in \mathcal{S}_{src}$.

Some related works use a Conditional Random Field (CRF) layer on top of the FFNN classifier in an attempt to capture the interactions between the label sequences (M'hamdi et al., 2019b). However, we did not find substantial performance differences when using a CRF layer and choose not to include it to keep our model as simple as possible.

4.2.2.2 Teacher Adversarial Training. Pre-trained multilingual language models such as mBERT or XLM-R provide contextualized representations for word sequences in multiple languages by embedding the words into a shared multilingual latent space. However, several studies have shown that, in such multilingual latent space, words from the same language group together, creating language clusters (M. V. Nguyen, Nguyen, et al., 2021; Yarmohammadi et al., 2021). As such, the word representations generated by these encoders are not language invariant. For a cross-lingual model, however, it is beneficial for similar words in the source and target languages to have similar (i.e. close) representations in the latent space. For instance, an English-trained Spanish-tested cross-lingual

model would benefit if the representations for the words *dog* and *perro* were similar to each other as then the model could adequately handle the Spanish sample provided it learns how to handle its English counterpart during training.

A technique that has been frequently used to promote the generation of such language-invariant representations is Adversarial Language Adaptation (ALA) (X. Chen et al., 2018; Joty et al., 2017). ALA introduces a *language discriminator* network $\mathcal{D}$ whose objective is to differentiate between the source and target languages. It learns language-dependent features that allow it to classify word representations as belonging to either the source or target languages. Concurrently, the encoder network is trained in an adversarial manner: it attempts to fool the discriminator by generating language-independent representations that are difficult to classify. A key feature of ALA is that it only requires unlabeled target-language data and, as such, it can be applied in a zero-shot setting using the available $\mathcal{S}_{tgt}^{unl}$ sentence set.

Other works that have leveraged ALA perform adversarial training at the sequence level (Guzman-Nateras, Nguyen, & Nguyen, 2022). That is, they only present the discriminator with sequence-level representations (e.g., the representation for the `[CLS]` token in mBERT). However, in this work we leverage token-level adversarial training which has been found to be more effective at generating language-invariant representations (W. Chen et al., 2021)

We again use a two-layer FFNN for the discriminator network $\mathcal{D}$. Instead of a softmax operation to generate a probability distribution, we employ a sigmoid function σ to predict the associated language l (Eq. 4.4).

$$\mathcal{D}(l_i) = \sigma(W^{D2}\ \mathrm{ReLU}(W^{D1}h_{ij})) \tag{4.4}$$

where W^{D1} and W^{D2} are parameter matrices to be learned and $\mathcal{D}(l_{ij})$ is a scalar $\in [0, 1]$ that indicates how likely it is that the current token representation h_{ij} belongs to the source ($l_i = 0$) or target ($l_i = 1$) languages.

Thus, besides the ED classification loss $\mathcal{L}_\mathcal{C}$ described in Equation 4.2, adversarial training introduces the discriminator loss $\mathcal{L}_\mathcal{D}$ (Eq. 4.5) as an additional training signal.

$$\mathcal{L}_\mathcal{D} = \frac{1}{n * m} \sum_{i=1}^{n} \sum_{j=1}^{m} l_i \cdot \mathcal{D}(h_{ij}) + (1 - l_i) \cdot (1 - \mathcal{D}(h_{ij})) \tag{4.5}$$

Our adversarial training is achieved by minimizing the following term:

$$\arg\min_{\mathcal{E},\mathcal{C}} \sum_{i=1}^{n} \sum_{j=1}^{m} (\mathcal{L}_\mathcal{C}(y_{ij}|h_{ij}) - \lambda \mathcal{L}_\mathcal{D}(l_i|h_{ij})) \tag{4.6}$$

We leverage a Gradient-Reversal Layer (GRL) Ganin and Lempitsky (2015) to implement Equation 4.6 by applying the GRL to the discriminator input vectors h_{ij}. A GRL acts as the identity function during the forward pass and reverses the direction of the gradients during the backward pass. As such, the encoder parameters are trained in the opposite direction to those of the discriminator, effectively learning to generate token representations with language-invariant features.

Figure 6 shows the architecture of the teacher model.

4.2.2.3 Student Model. As described in the previous section, the teacher model is trained using a direct transfer approach: it learns to generate language-independent representations from the labeled source-language data so that it can be directly applied to unlabeled target-language data. However, in our proposed hybrid knowledge transfer approach, we expect the student model to reap the benefits of the data transfer paradigm. Hence, we train the student model using

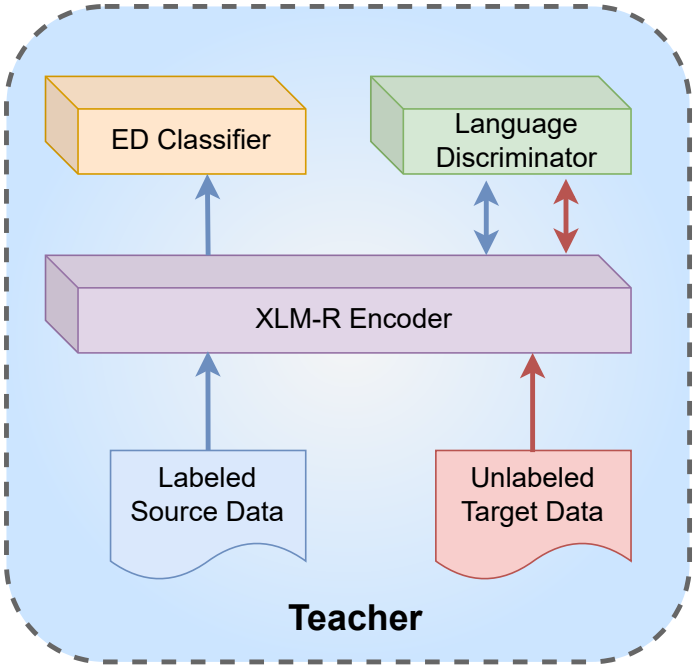

Figure 6. Adversarially-trained Teacher model. Source and target (unlabeled) data is passed through the encoder and fed at a token-level to the language discriminator. The discriminator gradients are then used to update the encoder parameters in an adversarial manner. The ED classifier is trained with the labeled source samples exclusively.

target-language data so that it may learn from syntactical features and word/label relations.

First, we apply the teacher model *Teach* to the unlabeled target dataset $\mathcal{S}_{tgt}^{unl}$ to obtain a pseudo-labeled training set S_{tgt}^{Teach}. Afterward, the student model *Student* is trained in a supervised manner using the obtained pseudo-labels.

The model architecture of our student model mirrors the one of the teacher model: a pre-trained multilingual language model as the encoder $\mathcal{E}_{STU}$ and a two-layer FFNN for a classifier $\mathcal{C}_{STU}$.

$$\mathcal{C}_{STU}(y_{ij}) = \text{softmax}(W^{C_S 2}\ \text{ReLU}(W^{C_S 1} h_{ij})) \tag{4.7}$$

Previous works on knowledge distillation have found that using soft labels (i.e., probability distributions over class types) is beneficial for student learning as they contain richer and more helpful information than hard labels (Hinton, Vinyals, & Dean, 2015). As such, we train the student model to minimize the Mean Squared Error (MSE) between the student-predicted and teacher-generated event-type distributions (Eq. 4.8).

$$\mathcal{L}_{Student} = \frac{1}{n * m} \sum_{i=1}^{n} \sum_{j=1}^{m} (\mathcal{C}_{STU}(\mathcal{E}_{STU}(t_{ij})) - \mathcal{C}_{T}(\mathcal{E}_{T}(t_{ij})))^2 \tag{4.8}$$

4.2.3 Student-Training Sample Selection. An important challenge in our teacher-student framework is that the target pseudo-labels obtained from the teacher model are prone to contain noisy predictions. The teacher model is trained with a direct transfer approach and, even though its word representations are encouraged to be language-independent through adversarial training, it learns task-related information exclusively from the source-language labels. We argue this prevents the teacher from learning task-specific information in the target language as it is unable to exploit the word-label relations specific to such language. Furthermore, even though the student model should be able to benefit from being trained in the target language, any potential benefits can be nullified if the quality of the teacher-generated pseudo-labels is too poor.

To address the aforementioned issue, we argue that the teacher model should produce more reliable pseudo-labels on target-language examples that share some similarities (structural or otherwise) with the source-language examples. Hence, we suggest improving the knowledge-distillation process by restricting student-model training to target-language examples with such desirable characteristics.

We implement this idea by designing a two-step hierarchical sample-selection scheme: First, we leverage Optimal Transport (OT) Villani (2008) to generate an alignment score between source and target samples and select samples above a defined alignment threshold. Then, using the selected source and target samples, we compute their pairwise Cross-domain Similarity Local Scaling scores (CSLS, Conneau et al., 2018) and only keep the pairs with the highest similarities. The following subsections describe each step in further detail.

Figure 7 presents an overview of our teacher-student framework.

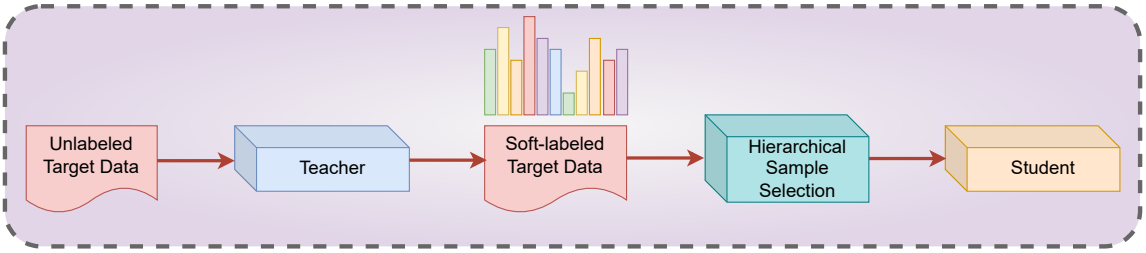

Figure 7. Teacher-student framework. The adversarially trained Teacher is used to annotate unlabeled target samples. Our hierarchical sample selection process picks a subset of samples to be used to train the Student model.

4.2.3.1 Optimal-Transport-based Selection.

Recent research efforts have successfully leveraged OT for cross-lingual language adaptation (Guzman-Nateras, Nguyen, & Nguyen, 2022; Phung, Minh Tran, et al., 2021) and word-label alignment for event detection (Pouran Ben Veyseh & Nguyen, 2022). OT relies on a distance-based cost function to compute the most cost-effective transformation between two discrete probability distributions by solving the following optimization problem:

$$\pi^*(x,z) = \min_{\pi \in \prod(x,z)} \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \pi(x,z)\ D(x,z) \tag{4.9}$$

$$\textbf{s.t.}\ \ x \sim P(x) \text{ and } z \sim P(z)$$

In Eq. 4.9, D is a cost function that maps $\mathcal{X}$ to $\mathcal{Z}$, $D(x, z), \mathcal{X} \times \mathcal{Z} \longrightarrow \mathbb{R}_+$, $P(x)$ and $P(z)$ are probability distributions for the $\mathcal{X}$ and $\mathcal{Z}$ domains, and $\pi^*(x, z)$ is the optimal joint distribution over the set of all joint distributions $\prod(x, z)$ (i.e., the optimal transformation between $\mathcal{X}$ and $\mathcal{Z}$).

For our work, we consider the source and target languages as the $\mathcal{X}$ to $\mathcal{Z}$ domains to be aligned. Each training sample corresponds to a data point in a distribution and is represented by its sentence-level encoding $h_{\_0}$. Following prior work (Pouran Ben Veyseh & Nguyen, 2022), we estimate probability distributions $P(x)$ and $P(z)$ using a single-layer FFNN and use Euclidean distance as the cost function:

$$D(h^x_{i0}, h^z_{j0}) = ||h^x_{i0} - h^z_{j0}||^2_2 \tag{4.10}$$

where h^x_{i0} is the i-th source-language sample and h^z_{j0} is the j-th target-language sample.

Once the OT algorithm converges, we leverage the solution matrix π^* to compute an overall similarity score $k_{\_}$ for each sample $h_{\_0}$ by averaging the optimal cost of transforming it to the other domain:

$$k^x_i = \frac{\sum^m_j \pi^*(h^x_{i0}, h^z_{j0})}{m} \tag{4.11}$$

Finally, a hyperparameter α determines the proportion of samples with the highest similarity scores k to be selected for use in the next step.

4.2.3.2 CSLS-based Selection.

The OT-based similarity score described previously captures the *global* alignment of a sample with the alternate language, e.g., how well a source-language sample aligns with the target language and vice versa. Nonetheless, we propose to further refine our sample selection by considering the *pairwise* similarity between source and target samples.

To this end, we make use of the Cross-domain Similarity Local Scaling (CSLS, Conneau et al., 2018) similarity measure which was originally designed to improve word-matching accuracy in word-to-word translation (Q. Wu, Lin, Karlsson, Huang, & Lou, 2020). CSLS addresses a fundamental issue of pair-matching methods based on Nearest Neighbors (NN): NNs are asymmetric by nature, i.e. if a is a NN of b, b is not necessarily a NN of a. In high-dimensional spaces, this asymmetry leads to *hubness*, a detrimental phenomenon for pair matching: samples in dense areas have high probabilities of being NN to many others, while samples that are isolated will not be a NN to any other sample (Conneau et al., 2018).

As such, when computing the similarity between a pair of samples, CSLS (Eq. 4.12) computes mean similarity $r_{\_}$ of a sample to its neighborhood $\mathcal{N}_{\_}$ (i.e., its K nearest neighbors) in the alternate language and leverages it to increase the similarity scores of isolated samples while decreasing the scores of so-called *hub* samples. For example, the mean similarity r_Z for source sample h_i^x is computed with its target neighborhood $\mathcal{N}_Z$ (Eq. 4.13).

$$\text{CSLS}(h_i^x, h_j^z) = \tag{4.12}$$

$$2\cos(h_i^x, h_j^z) - r_Z(h_i^x) - r_X(h_j^z)$$

$$r_Z(h_i^x) = \frac{1}{|\mathcal{N}_Z|} \sum_{\mathcal{N}_Z} \cos(h_i^x, h_j^z) \tag{4.13}$$

$$r_X(h_j^z) = \frac{1}{|\mathcal{N}_X|} \sum_{\mathcal{N}_X} \cos(h_j^z, h_i^x) \tag{4.14}$$

where *cos* is the cosine similarity. In our work, the source $\mathcal{N}_X$ and target $\mathcal{N}_Z$ neighborhoods are defined as the corresponding sample sets kept by the previous selection step. Again, we keep a proportion of the samples with the best pairwise similarity scores determined by a hyperparameter β.

Figure 8 presents an overview of our proposed hierarchical sample-selection strategy.

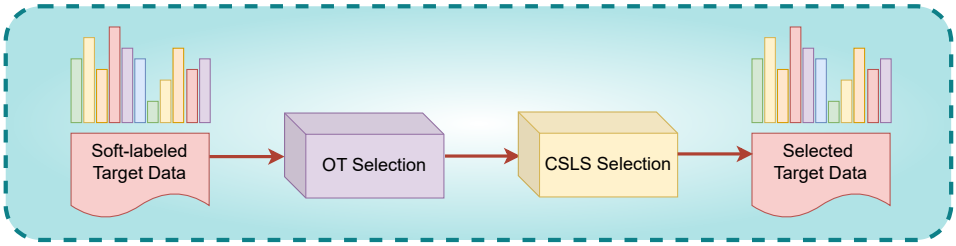

Figure 8. Hierarchical sample selection scheme. The target-language samples annotated by the Teacher model are first filtered by OT-based selection. The remaining samples are then further refined via CSLS. The final subset of samples is used to train the Student model.

4.3 Experiments

4.3.1 Datasets. For our experiments, we leverage the ACE05 Walker et al. (2006) and ACE05-ERE (Song et al., 2015) datasets as they are the most commonly used benchmarking datasets for CLED. ACE05 contains ED annotations in 3 languages: English (En), Chinese (Zh), and Arabic (Ar). ACE05-ERE includes annotations in both English and Spanish (Es).

To further test the applicability of our model, we also make use of the recently released MINION dataset (Pouran Ben Veyseh et al., 2022) which contains annotations for 8 morphologically and syntactically distinct languages: English, Spanish, Hindi (Hi), Japanese (Ja), Korean (Ko), Polish (Pl), Portuguese (Pt), and Turkish (Tr).

Appendix A.2 presents additional details about the aforementioned datasets.

4.3.2 Main results. In order to evaluate our Hybrid Knowledge Transfer for Cross-Lingual Event Detection (HKT-CLED) model, we first present our results on the ACE05 and ACE05-ERE datasets in Table 7. We compare against 6 recent CLED efforts including the current state-of-the-art

model (Guzman-Nateras, Nguyen, & Nguyen, 2022). All the baseline results are taken directly from the original papers and our model's results are the average of 5 runs with different seeds. English is used as the sole source language and Arabic, Chinese, and Spanish are employed as target languages. Following previous works, we report F1 scores.

	Target Language		
Model	**Zh**	**Ar**	**Es**
J. Liu et al. (2019)	27.0	-	-
M'hamdi, Freedman, and May (2019)	68.5	30.9	-
D. Lu et al. (2020)	-	-	41.77
Majewska et al. (2021b)	46.9	29.3	-
M. V. Nguyen, Nguyen, et al. (2021)	72.1	42.7	-
Guzman-Nateras, Nguyen, and Nguyen (2022)	74.64	44.86	47.69
HKT-CLED (Ours)	**75.22**	**46.37**	**48.58**

Table 7. Cross-lingual event detection model performance comparison. English is used as the source language. ACE05 is used for Chinese (Zh) and Arabic (Ar), ACE05-ERE is used for Spanish (Es).

Our proposed approach obtains new state-of-the-art performance across all 3 target languages with improvements of +0.58, +1.51, and +0.89 F1 points for Chinese, Arabic, and Spanish, respectively. We believe these results demonstrate the importance of hybrid knowledge transfer as it gives HKT-CLED an edge over previous works that follow a direct transfer approach (Guzman-Nateras, Nguyen, & Nguyen, 2022; Majewska et al., 2021b; M'hamdi et al., 2019; M. V. Nguyen, Nguyen, et al., 2021).

To validate the effectiveness and general applicability of our approach, Table 8 presents the performance of our HKT-CLED model on the more diverse MINION dataset. Once again, we employ English as the source language and test our model's performance on the remaining 7 languages. For a fair comparison, we use their best XLM-R results. Our model consistently outperforms their reported baseline with an average performance improvement of +7.74 F1 points for all target

languages (+5.25 if the highest and lowest improvements are not considered). In the case of Japanese, HKT-CLED obtains a massive performance improvement of over 25 F1 points. Also of note is that HKT-CLED performance is a lot more uniform across target languages than the baseline. There is a difference of 23.43 F1 points between the best-performing (Pt, 77.28) and the worst-performing (Tr, 53.85) target languages, as opposed to a 37.65 point difference in the baseline case (Pt, 72.77 and Ja, 35.12).

	Target Language						
Model	**Es**	**Hi**	**Ja**	**Ko**	**Pl**	**Pt**	**Tr**
Baseline*	62.83	58.19	35.12	56.78	60.13	72.77	47.21
HKT-CLED	66.03	68.63	61.84	58.24	61.35	77.28	53.85
Improvement	+3.2	+10.44	+26.72	+1.46	+1.22	+4.51	+6.64

Table 8. Cross-lingual ED performance on the MINION dataset. F1 scores are reported. English is used as the source language. Baseline* performance was obtained directly from the original MINION paper (Pouran Ben Veyseh et al., 2022). HKT-CLED results are the average of 3 runs.

The model implementation details can be found in Appendix C.

4.3.3 Analysis.

4.3.3.1 Ablation Study. We first explore the contribution of each model component by performing an ablation study (Table 9). In particular, we evaluate the impact of three aspects: teacher adversarial training, OT-based sample selection, and CSLS-based sample selection. The *Teacher (Vanilla)* results were obtained with a standard sequence-labeling model without any adversarial training. Its performance leaves room for improvement as its word representations do not display any language-invariant qualities. A considerable improvement is achieved when training the teacher model with token-level adversarial training (*Teacher + Adv*). Then, the *Student (Vanilla)* row shows the result of training a student network on the teacher-generated pseudo-labels without any sample selection. We

argue its performance is worse than the adversarially-trained teacher due to the noisy pseudo-labels. By incorporating OT-based selection, *Student + OT* is able to outperform its teacher. However, it is only by performing our hierarchical sample selection that the student model achieves new state-of-the-art performance.

	Target Language		
Model	**Zh**	**Ar**	**Es**
HKT-CLED	75.22	46.37	48.58
Student + OT	74.37	45.53	47.63
Student (Vanilla)	73.48	44.10	46.81
Teacher + Adv	73.85	44.42	47.37
Teacher (Vanilla)	70.51	43.59	46.75

Table 9. Ablation experiment results.

4.3.3.2 Impact of Sample-Selection Ratios. Figure 9 shows the impact of hyperparameter α on model performance. α determines the proportion of student-training samples kept by the OT-based selection step. An $\alpha = 1$ value performs no sample selection and $\alpha = 0.25$ only keeps a fourth of the batch samples with the highest similarity scores.

Best results are obtained when half of the samples are kept ($\alpha = 0.5$) exemplifying the importance of removing training examples with potentially noisy pseudo-labels. However, if too few samples are chosen (e.g., $\alpha = 0.25$) the student performance drops below its *vanilla* version ($\alpha = 1$).

Similarly, Figure 10 presents the effect on performance of hyperparameter β which defines the proportion of samples kept by the CSLS-selection step. A $\beta = 1$ value uses all of the samples selected by the previous step.

Removing about a quarter ($\beta = 0.75$) of the previously-selected samples improves performance across all languages. Of note is the fact that the OT and CSLS similarity scores complement each other. From Figure 9 it would seem that

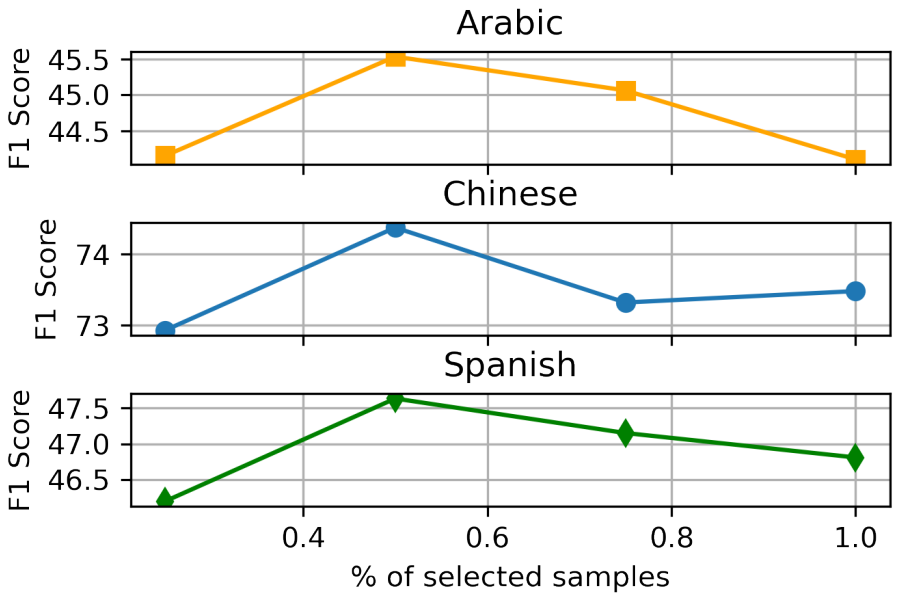

Figure 9. Performance impact of hyperparameter α.

removing more than half of the training samples would only hurt performance. However, given CSLS pairwise focus, it is able to effectively remove some remaining noisy samples and obtain better results.

4.4 Related Work

Cross-lingual event detection has recently gained traction as a research area. The work by J. Liu et al. (2019) presents a data transfer method that learns a mapping between monolingual word embeddings, translates the source training data on a word-by-word basis and uses a graph convolutional network to generate order-independent representations. M'hamdi et al. (2019b) leverage mBERT as an encoder to perform zero-shot transfer learning and a CRF layer to account for label dependency. D. Lu et al. (2020) present a cross-lingual structure transfer approach that represents sentences as language-universal structures (trees, graphs). In their work, Majewska et al. (2021b) argue that event triggers are usually related to the verb in a sentence and propose to incorporate external verb knowledge

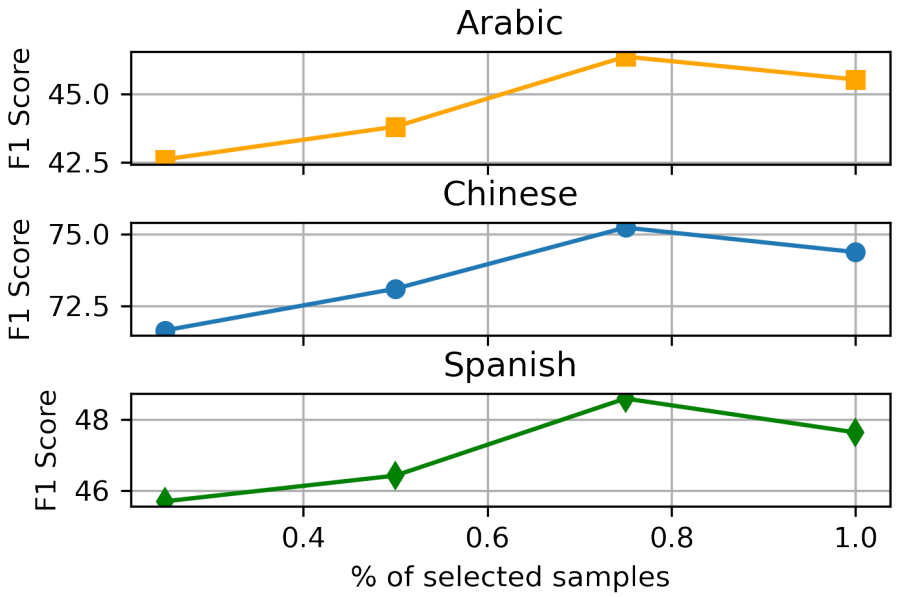

Figure 10. Performance impact of hyperparameter β.

by pre-training their encoder to classify whether two verbs belong to the same class according to two distinct ontologies VerbNet, (Kipper et al., 2006) and FrameNet, (Baker et al., 1998). Model *priming* (Fincke et al., 2021) is a simple, yet effective method that consists in augmenting the encoder inputs by concatenating a candidate trigger to the input sentence so that the encoder learns to generate task-specific representations. M. V. Nguyen, Nguyen, et al. (2021) leverage class information and word categories as language-independent sources of information and condition their encoder to generate representations that are consistent in both the source and target languages. Finally, Guzman-Nateras, Nguyen, and Nguyen (2022) propose to optimize standard adversarial language adaptation by restricting the language discriminator training to *informative* examples.

Our approach is also closely related to knowledge distillation models for cross-lingual Named Entity Recognition (NER). Q. Wu, Lin, Karlsson, Lou, and Huang (2020) were the first to train a NER student model on the label

distributions obtained from a teacher model. Q. Wu, Lin, Karlsson, Huang, and Lou (2020) improved upon this initial approach with a multi-step training method that involves fine-tuning the teacher model with pseudo-labeled data and generating hard labels that are later used for student training. More recent proposals improve the knowledge distillation process with either reinforcement learning (Liang et al., 2021) or adversarial training (W. Chen et al., 2021).

Nonetheless, our approach is the first to leverage a knowledge distillation framework for the CLED task, and our novel hierarchical training-sample selection scheme further differentiates our work from the aforementioned efforts.

4.5 Summary

In summary, we consider the main contributions of this chapter to be the following:

- We present the first effort to leverage a hybrid knowledge-transfer approach for the cross-lingual event detection task which benefits from the advantages of both the direct transfer and the data transfer knowledge transfer paradigms and minimizes their shortcomings.

- We address the issue of noisy pseudo-labels in our teacher-student framework by proposing an entirely novel a hierarchical training-sample selection scheme that effectively constrains the student-training process to pseudo-labeled target-language samples that are similar to their source-language counterparts.

- Our HKT-CLED model sets a new state-of-the-art performance on the most popular benchmarking datasets ACE05 and ACE05-ERE, and obtains substantial performance improvements on the recently-released, and more

diverse, MINION dataset with an average improvement of +7.74 F1 points across 7 distinct target languages.

– We provide an ablation study and complementary analysis to validate the contribution of each of our model's comprising elements and confirm the efficacy of our hierarchical sample selection scheme.

– Our results demonstrate our model's robustness and applicability and validate our claim that combining the benefits of the direct transfer and data transfer approaches is beneficial for cross-lingual learning.

CHAPTER V

EXPLOITING SUPPORT/QUERY SET GLOBAL ALIGNMENT FOR FEW-SHOT CROSS-LINGUAL EVENT DETECTION

This Chapter contains materials from the published paper "*Luis F. Guzman-Nateras, Viet D. Lai, Franck Dernoncourt, and Thien H. Nguyen. '**Few-Shot Cross-Lingual Learning for Event Detection**' In Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL), 2022*"(Guzman-Nateras, Nguyen, & Nguyen, 2022). As the first author of this publication, Luis was responsible for most areas of the project including development, experimentation, and document writing. Viet provided a starting code base and meaningful discussions and insights. Thien had input on the initial project conceptualization, and he and Franck made editorial suggestions for the final document. The original publication contents have undergone some editorial updates to comply with this document's format and purpose.

After exploring a direct-transfer-based approach in Chapter III and a hybrid-transfer approach in Chapter IV, in this chapter we switch our attention to a different learning paradigm. Training of CLED models is usually performed in a standard supervised-learning setting with labeled data available in the source language. The Few-Shot Learning (FSL) paradigm is yet to be explored for CLED despite its inherent advantage of allowing models to better generalize to unseen event types. As such, in this chapter, we study the novel setting of FSL for CLED. Our contribution is threefold: first, we introduce a novel FSL classification method based on Optimal Transport (OT, Villani, 2008); second, we present a novel regularization term to incorporate the global distance between the support and query sets; and third, we adapt our approach to the cross-lingual setting by

exploiting the alignment between source and target data. Our experiments on 3, syntactically-different, target languages show the applicability of our approach and its effectiveness in improving the cross-lingual performance of few-shot models for event detection.

5.1 Introduction

Event Detection (ED) is a significant sub-task within the larger task of Information Extraction (IE) in Natural Language Processing (NLP). Its core purpose is to identify the words, or phrases, that most clearly express the occurrence of an event, known as event *triggers*, and to correctly categorize them into a discrete set of classes. For instance, in the sentence:

*Frank **purchased** his dream house yesterday.*

the word "**purchased**" should be identified by an ED system as the trigger of a `Transaction:Transfer-Ownership` event type[1]. Event detection is a highly active research area which has been lately dominated by deep-learning-based approaches J. Liu et al. (2020); Y. Lu et al. (2021); T. M. Nguyen and Nguyen (2019); Sha et al. (2018); Wadden et al. (2019); S. Yang et al. (2019a); J. Zhang et al. (2019); Y. Zhang et al. (2020). Most of these works use the standard supervised learning paradigm in which lots of labeled data is required during training. However, a significant limitation of models trained in this manner is their inability to properly generalize to new event types that were unobserved during training V. D. Lai, Nguyen, and Dernoncourt (2020).

5.1.1 Few-Shot Learning.

In contrast to the supervised approach, Few-Shot Learning (FSL) proposes a training setting in which a model must quickly learn new concepts from just a few examples, similar to how humans can learn to

[1]Event type example taken from ACE05 dataset.

detect and identify new objects after having observed only a couple of instances. During an FSL training iteration, a model is given a *support* set and a *query* set, each of which contains only a handful of examples for a set of classes. Then, the model is trained to predict the classes for the query samples based on the labeled support samples. Under these constrained training settings, supervised training easily results in model overfitting due to the limited availability of training data. Furthermore, in FSL, a model is evaluated on its ability to generalize to new, unobserved types. To achieve this, during testing an FSL model is provided with new support and query sets whose samples belong to entirely new classes never observed during training.

Typical FSL approaches consist of obtaining a vector representation for each sample and then performing classification based on the distance between such vectors, e.g., Matching Networks Vinyals, Blundell, Lillicrap, Kavukcuoglu, and Wierstra (2016), Relation Networks Sung et al. (2018), and Prototypical Networks Snell, Swersky, and Zemel (2017). The key differences between these approaches often come down to the way the sample representations are generated, and how the distance between such representations is determined.

FSL training allows a model to easily extend to new classes as it only needs to see a few labeled examples in order to successfully classify them. FSL has been applied successfully for many tasks. Recently, there have been several efforts that explore event detection under a few-shot learning setting (FSLED) J. Chen, Lin, Han, and Sun (2021); Cong et al. (2021); Deng et al. (2020); V. Lai, Dernoncourt, and Nguyen (2021); V. D. Lai, Dernoncourt, and Nguyen (2020); V. D. Lai, Nguyen, Nguyen, and Dernoncourt (2021); V. D. Lai, Nguyen, and Dernoncourt (2020); Shen et al. (2021).

5.1.2 Cross-Lingual Event Detection.

Cross-Lingual Learning (CLL) is a paradigm that aims at transferring the knowledge from one language to another (Pikuliak et al., 2021b). CLL can help overcome the lack of data availability that plagues many languages and allow for the creation of NLP-based tools that can benefit their communities.

As such, Cross-lingual Event Detection (CLED) aims at detecting and classifying event triggers with the added complexity of operating on two separate languages. These two languages are referred to as *source* and *target*, respectively. In standard *zero-shot* training, a CLED model is trained using labeled data belonging to the source language exclusively. Then, at testing time, data from the target language is used to evaluate the model's performance Guzman-Nateras, Nguyen, and Nguyen (2022); Majewska et al. (2021b); M'hamdi et al. (2019b); M. V. Nguyen, Nguyen, et al. (2021).

A proper effort on CLED under FSL conditions has yet to be explored despite the potential advantages it could contribute to cross-lingual models. Hence, we recognize this opportunity and propose the novel Few-Shot Cross-Lingual Event Detection (FSCLED) task to integrate these two settings.

The rest of the chapter is organized as follows: Section 5.2 provides a formal definition for FSCLED task, Section 5.3 describes the details our proposed approach, Section 5.4 presents the results of our experiments, and finally, we present our conclusions in Section 5.6.

5.2 Problem Definition

5.2.1 Few-shot Event Detection.

We follow the same problem formulation as in prior work for few-shot ED Deng et al. (2020); V. Lai, Dernoncourt, and Nguyen (2021); V. D. Lai, Nguyen, and Dernoncourt (2020). In

particular, we cast event detection as a token classification task in which a model must learn to correctly classify the trigger tokens. In a standard FSL setting, an iteration involves a support set $\mathcal{S}$ and a query set $\mathcal{Q}$ that cover sample sentences for N distinct classes; each class is represented by $K \in [1, 10]$ examples. Additionally, for event detection, $\mathcal{S}$ and $\mathcal{Q}$ are extended with an additional negative, or non-event, type *NULL* (also with K examples) V. Lai, Dernoncourt, and Nguyen (2021). In this manner, given an input sentence along with an trigger candidate, an FSL model for ED should be able to predict whether the candidate is an event trigger as well as which event type is evoked by the trigger (if any).

Hence, the formal definition of the FSL task is as follows. The $\mathcal{S}$ and $\mathcal{Q}$ sets are defined by:

$$\mathcal{S} = \{(s_i^{j(\mathcal{S})}, t_i^{j(\mathcal{S})}, y_i^{j(\mathcal{S})})\} \tag{5.1}$$

$$\mathcal{Q} = \{(s_i^{j(\mathcal{Q})}, t_i^{j(\mathcal{Q})}, y_i^{j(\mathcal{Q})})\} \tag{5.2}$$

where $i \in [1, K]$[2], $j \in [0, N]$ ($j = 0$ is used for the non-event type), and a single sample $(s_i^{j(\cdot)}, t_i^{j(\cdot)}, y_i^{j(\cdot)})$ contains a sentence $s_i^{j(\cdot)}$, a trigger candidate word $t_i^{j(\cdot)}$ in $s_i^{j(\cdot)}$, and an event label type $y_i^{j(\cdot)}$. As per FSL requirements, the label set used when training the model must be disjoint from those used when evaluating the model to properly assess the model's ability to generalize to unobserved classes.

5.2.2 Few-shot Cross-lingual Event Detection. Cross-Lingual Learning (CLL) methods Pikuliak et al. (2021b) emerged from the need to create NLP models for low-resource *target* languages that lack the required labeled data to perform supervised learning. The core idea is to train models using available labeled data from a high-resource *source* language with techniques that allow them to learn task-specific language-invariant features. The models are then evaluated on

[2]We use the same number of samples for each class in both the support and query sets.

the desired target language without access to target-language labeled data during training. This setting is known as *zero-shot* cross-lingual transfer learning[3].

As such in the zero-shot cross-lingual ED task, the labeled samples used during training $\mathcal{D}_{train}$ and development $\mathcal{D}_{dev}$ belong to the source language while the ones used for testing $\mathcal{D}_{test}$ correspond to the target languages Majewska et al. (2021a); M'hamdi, Freedman, and May (2019a).

In this work, we combine the aforementioned *zero-shot* approach to cross-lingual evaluation with the added intricacy of the standard few-shot setting. During training, the models are presented with a support set $\mathcal{S}^{src}$ and a query set $\mathcal{Q}^{src}$ that belong to the source language. Then, at testing time, the support set $\mathcal{S}^{tgt}$ and query set $\mathcal{Q}^{tgt}$ are taken from the target language for evaluation. Furthermore, given the FSL setting, the label set used during training is disjoint from the label set for development and testing. We designate this novel task as Few-Shot Cross-Lingual Event Detection (FSCLED).

5.3 Model

As done in prior FSL models for ED V. Lai, Dernoncourt, and Nguyen (2021), our model for FSCLED involves two main components: an encoder E and a classifier C.

5.3.1 Encoder. The encoder's purpose is to obtain a representation vector $v_i^{j(\cdot)}$ for each sample in the support $\mathcal{S}$ and query $\mathcal{Q}$ sets:

$$v_i^{j(\cdot)} = E(s_i^{j(\cdot)}, t_i^{j(\cdot)}) \in \mathbb{R}^d \tag{5.3}$$

where d is the vector size, and $\cdot$ can be either $\mathcal{S}$ or $\mathcal{Q}$.

[3]Not to be confused with standard zero-shot learning where zero data for a new class is used by models to perform prediction.

Following recent work on CLED, we leverage the pretrained multilingual language model (mLM) mBERT Devlin et al. (2019) for our encoder to take advantage of its ability to induce language-invariant representations Majewska et al. (2021b). Additionally, we stack a Multi-Layer Perceptron (MLP) layer on top of the transformer outputs to create our multilingual encoder, called BERTMLP S. Yang, Feng, Qiao, Kan, and Li (2019b). Then, we employ the vector representation for $t_i^{j(\cdot)}$ generated by BERTMLP to serve as the representation $v_i^{j(\cdot)}$.

5.3.2 Classifier. For convenience, let v^s and v^q be the representation vectors for the sample $s \in \mathcal{S}$ and $q \in \mathcal{Q}$, and $V^{(\mathcal{S})}$ and $V^{(\mathcal{Q})}$ be the sets of representation vectors for all samples in the support and query sets, respectively.

The classifier C aims to predict a label y^q for each instance q in the query set based on its representation v^q and the representations of the samples in the support set $V^{(\mathcal{S})}$:

$$y^q = C(v^q, V^{(\mathcal{S})}) \tag{5.4}$$

Given the multilingual representations $v_i^{j(\cdot)}$, a feasible approach is to employ existing FSL models (e.g., Matching, Relation, or Prototypical networks) to perform classification in FSCLED. The models can then be trained using the standard cross-entropy loss.

5.3.2.1 Optimal Transport. We recognize, nonetheless, a potential issue with traditional FSL models in that they only consider local distances between individual pairs of samples in the support and query sets. In the case of Prototypical Networks (Snell et al., 2017), for example, the distance is between a query sample and a class prototype. Hence, if the overall global distance between the support and query sets is large, a small difference between the distances of two individual samples becomes less reliable to determine the label assignments. In

turn, we argue that the global distances between $\mathcal{S}$ and $\mathcal{Q}$ should be minimized to improve the reliability of the distances between individual pairs for accurate FSCLED.

To this end, we propose utilizing Optimal Transport (OT) Villani (2008) to estimate the distance between the support $\mathcal{S}$ and query $\mathcal{Q}$ sets for FSCLED. In broad terms, OT aims to find the most cost-effective transformation between two discrete probability distributions. Optimal transport employs a cost function to compute the cost of transforming data points from one distribution to the other. If a distance function (Euclidean, Cosine, etc.) is used as such cost function, the obtained minimum cost is known as the Wasserstein distance. Formally, OT solves the following optimization problem:

$$\pi^*(s,q) = \min_{\pi \in \prod(s,q)} \sum_{s \in \mathcal{S}} \sum_{q \in \mathcal{Q}} \pi(s,q)\ D(s,q) \tag{5.5}$$

$$\textbf{s.t. } s \sim P(\mathcal{S}) \text{ and } q \sim P(\mathcal{Q})$$

where $P(x)$ and $P(z)$ are probability distributions for the $\mathcal{X}$ and $\mathcal{Z}$ domains, and D is a distance-based cost function for mapping $\mathcal{X}$ to $\mathcal{Z}$, $D(x,z) : \mathcal{X} \times \mathcal{Z} \longrightarrow \mathbb{R}_+$. Finally, $\pi^*(x,z)$ is the optimal joint distribution over the set of all joint distributions $\prod(x,z)$ (i.e., the optimal transformation between $\mathcal{Z}$ and $\mathcal{X}$). The described OT optimization problem is, however, intractable as it requires optimizing over the infinite set $\prod(x,z)$. In practice, we instead solve an entropy-based relaxation of the discrete OT problem using the Sinkhorn algorithm Cuturi (2013).

5.3.2.2 Few-Shot Classification via OT.

To adapt FSL classification into an OT formulation we consider the support $\mathcal{S}$ and query $\mathcal{Q}$ sets as the two domains to be transformed. Each sample in $\mathcal{S}$ and $\mathcal{Q}$ represents a data

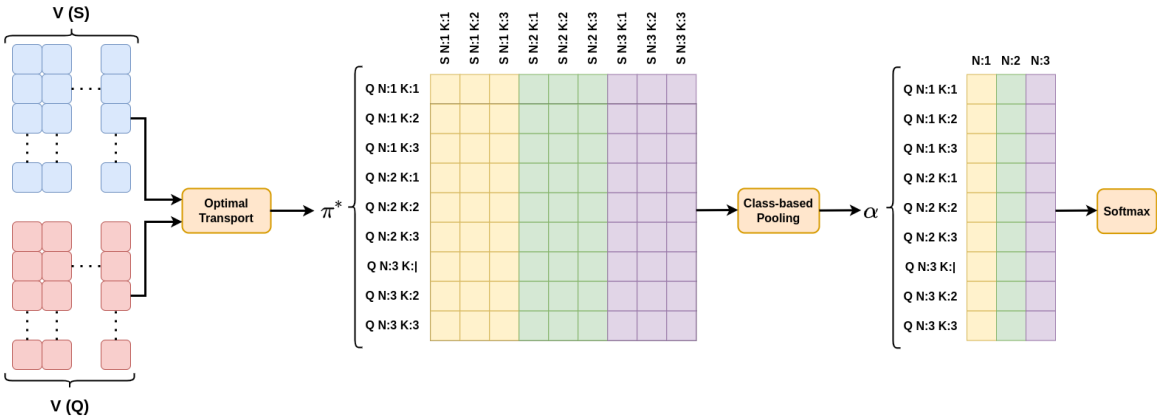

Figure 11. OT-based classification procedure example for a 3-way, 3-shot setting. Optimal Transport is used to obtain a the optimal similarity matrix π^*. Then, the likelihood vectors α are obtained via class-based pooling. Finally, the softmax the similarity vectors is leveraged for training and final class prediction.

point in the corresponding distribution. The probability distributions $P(\mathcal{S})$ and $P(\mathcal{Q})$ are estimated using an *event-presence* module F. In our work, F is a feed-forward neural network (FFNN) with a single output and sigmoid activation that scores the likelihood that a trigger candidate word is actually an event trigger. F receives as input the vector representation of a trigger $v^{(\cdot)}$ from either $\mathcal{S}$ or $\mathcal{Q}$, and outputs a scalar in the range [0-1]. Then, the probability distributions for $\mathcal{S}$ and $\mathcal{Q}$ are obtained by computing the Softmax over F's outputs for the samples in each set:

$$P(\mathcal{S}) = Softmax(F(V^{(\mathcal{S})})) \tag{5.6}$$

$$P(\mathcal{Q}) = Softmax(F(V^{(\mathcal{Q})})) \tag{5.7}$$

To supervise the event-presence module F, we include the cross-entropy loss for event identification into the overall loss function:

$$\mathcal{L}_{ident} = \tag{5.8}$$

$$\sum_{s \in \mathcal{S}} i^s \cdot \sigma(F(v^s)) + (1 - i^s) \cdot \sigma(1 - F(v^s))$$

where i^s is the golden binary variable to indicate if s corresponds to an event trigger or not, and σ is the sigmoid function.

In our model, the distance $D(q, s)$ between a sample in $q \in \mathcal{Q}$ and a sample $s \in \mathcal{S}$ is based on the Euclidean distance between their representation vectors v^s and v^q:

$$D(q, s) = \sqrt{\sum_{i \in d} (v_i^q - v_i^s)^2} \tag{5.9}$$

Once the OT algorithm converges, or the maximum number of iterations is reached, the obtained optimal alignment matrix π^* is a squared matrix with dimensions $((N + 1) * K) \times ((N + 1) * K)$ where each entry $\pi^*_{r,c}$ represents the alignment score between the r-th query sample and c-th support sample.

The conversion from matrix index (r, c) to event type (j) and sample number (i) can be computed in a straightforward manner as all samples from the same class (event type) are contiguous: $j = r//K, i = r \ \% \ K$ where // and % are the integer division and modulo operators.

To perform sample classification and train our FSCLED model, we first use the optimal alignment matrix π^* to compute a likelihood vector α for each query sample (i.e., the r-th) by performing class-based pooling with respect to the $N + 1$ classes:

$$\alpha_r^j = \sum_{i \in [0, K-1]} \pi^*_{r,(j*K)+i} \tag{5.10}$$

where $j \in [0, N]$. As such, the resulting α_r vectors have $N + 1$ dimensions. And the complete α matrix has a $((N + 1) * K) \times (N + 1)$ size. We then apply a Softmax operation over α_r to obtain a class distribution P_r for the r-th query sample: $P_r = \mathit{Softmax}(\alpha_r)$. P_r will then be used for training and inference in our model. In particular, we use the negative log-likelihood loss as the main term of our

overall training loss:

$$\mathcal{L}_{class} = -\sum_{r} P_r(y_r) \tag{5.11}$$

where y_r is the golden class for the r-th query example. Figure 11 shows a visualization of the described procedure for a 3-way, 3-shot setting. As such, a key distinction is that the class distribution P_r in our FSL method is obtained from the support-query alignment scores π^* in optimal transport. This is in contrast to previous FSL models where the class distributions tend to be computed directly from sample representations.

5.3.3 Support-Query Distance. In addition to our optimal-transport-based FSL classifier, we propose computing the Wasserstein distance between $\mathcal{S}$ and $\mathcal{Q}$ and including it into the loss function as a regularization term to minimize the overall distance between the support and query sets for reliable predictions. We obtain the aforementioned Wasserstein distance using the optimal alignment matrix π^*:

$$\mathcal{L}_{dist} = \sum_{s \in \mathcal{S}} \sum_{q \in \mathcal{Q}} \pi^*_{r,c} \, D(q, s) \tag{5.12}$$

where r and c are the matrix indexes for q and s, respectively.

5.3.4 Cross-Lingual Distance. To adapt our approach to the cross-lingual setting, we aim to encourage language-invariant representation learning by regularizing our model so the representation vectors of samples in the source and target languages are closer to each other in the embedding space.

Following our approach discussed in Chapter III Guzman-Nateras, Nguyen, and Nguyen (2022), which leveraged OT to successfully align samples taken from the source and target languages to improve adversarial language adaptation, we propose to further use OT to estimate the distance between samples in the source

and target languages so that it can be included in the overall loss function as an additional regularization term for minimization.

To this end, given the unavailability of labeled data in the target language, we make use of unlabeled data – often readily available for most languages – instead. For convenience, let $\mathcal{R}$ and $\mathcal{T}$ represent the source-language and target-language data set respectively. In any given FSL training iteration, the support $\mathcal{S}$ and the query $\mathcal{Q}$ sets comprise the $\mathcal{R}$ set for the source language. To constitute the set representing the target language $\mathcal{T}$, we collect enough unlabeled samples to match the size of $\mathcal{R}$.

Thus, similarly to the OT formulation described in section 5.3.2.2 that computes the optimal alignment between two domains $\mathcal{S}$ and $\mathcal{Q}$, in this context we consider the source- and target-language data set $\mathcal{R}$ and $\mathcal{T}$ as the domains to be transformed. Subsequently, we employ our BERTMLP multilingual encoder to obtain representation vectors for the samples in both $\mathcal{R}$ and $\mathcal{T}$ that will serve as the inputs for the OT algorithm.

It is important to note that, due to the unavailability of the class information for the target-language samples $\mathcal{T}$ for training, it is less reliable to estimate the probability distribution $P(\mathcal{T})$ for the target language using the event-presence prediction module F as performed for $P(\mathcal{S})$ and $P(\mathcal{Q})$. Hence, we initialize $P(\mathcal{R})$ and $P(\mathcal{T})$ as uniform distributions for the OT computation in this case.

Under this setting, we solve the OT equation to obtain the optimal alignment matrix ρ^* between $\mathcal{R}$ and $\mathcal{T}$. The Wasserstein distance $\mathcal{L}_{cross}$ is then

computed and integrated into the overall loss function for regularization:

$$\mathcal{L}_{cross} = \sum_{r \in \mathcal{R}} \sum_{t \in \mathcal{T}} \rho^*_{n,m} \, D(r,t) \tag{5.13}$$

where n and m are the matrix indexes for r and t, respectively.

5.3.4.1 Full Model. Finally, the overall loss function $\mathcal{L}$ used to train our Optimal-Transport-based Event Detection (OTED) model is:

$$\mathcal{L} = \mathcal{L}_{class} + \alpha \mathcal{L}_{ident} + \beta \mathcal{L}_{dist} + \gamma \mathcal{L}_{cross} \tag{5.14}$$

where α, β, and γ are trade-off hyperparameters.

5.4 Experiments

5.4.1 Datasets. We use the ACE05 Walker et al. (2006) and ACE05-ERE Song et al. (2015) datasets, which are frequently used as the standard benchmarks in cross-lingual event detection efforts (Guzman-Nateras, Nguyen, & Nguyen, 2022; Majewska et al., 2021b; M'hamdi et al., 2019b; M. V. Nguyen, Nguyen, et al., 2021), to evaluate our FSCLED models. In particular, we utilize data in three languages (English, Chinese, and Arabic) from ACE05 and two languages (English and Spanish) from ERE. Both ACE05 and ERE organize their event classes in a hierarchical structure of types and subtypes. For example, in the `Transaction:Transfer-Ownership` class, `Transaction` is the main event type and `Transfer-Ownership` is the subtype. The two datasets have distinct label sets as ACE05 includes 33 event subtypes and ACE05-ERE has 38 event subtypes. Each language in the datasets has its own training/development/test split.

5.4.1.1 FSL Preprocessing. Standard datasets used for supervised learning, such as ACE05 and ERE05, can also be exploited for FSL by simulating a limited-data-availability setting via *episodic training* (V. Lai, Dernoncourt, & Nguyen, 2021). An *episode* is created by sampling a set of K examples from a small

subset of classes N out of the total number of classes in the dataset. This setting is referred to as *N-way, K-shot* and N and K are usually selected in the range of 1 to 10.

Following previous work on FSL for ED V. D. Lai, Nguyen, and Dernoncourt (2020), we further truncate the training, development, and testing portions of the datasets for each language to satisfy the conditions for FSL: (1) the set of event types in the training data must be disjoint from those for the development and test data; (2) the types in each set must contain at least 5 samples (to facilitate 5+1-way 5-shot learning with the additional +1 class being used for non-triggers); and (3) the training set should have as many samples as possible.

Adapting these criteria to cross-lingual FSL, we separate the samples belonging to the `Business, Contact, Conflict`, and `Justice` types to be used for training purposes. Meanwhile, we leave the samples belonging to the `Life, Movement, Personnel`, and `Transaction` event types for development and testing. Furthermore, we remove any subtypes that do not contain enough samples to construct an episode (5 samples minimum). Table 10 shows the total number of remaining classes for each portion of data in different languages for our FSCLED setting. We also list the event subtypes that are removed to meet the criteria in each dataset portion. Note that, while the training label set must be disjoint from the development and testing label sets, there is no requirement for the latter two to be disjoint as done in V. D. Lai, Nguyen, and Dernoncourt (2020).

As the final step in our data preprocessing, we obtain the samples for the non-event type by selecting words, other than the actual triggers, from annotated

sentences similar to the approach taken by V. D. Lai, Nguyen, and Dernoncourt (2020).

Dataset	# Types	Removed Types
ACE05-English (train)	19	Justice:Extradite Justice:Pardon
ACE05-English (dev)	12	
ACE05-Chinese (test)	11	Life:Divorce
ACE05-Arabic (test)	9	Life:Be-Born Life:Divorce Personnel:Nominate
ERE05-English (train)	22	Business:Bankrupcy
ERE05-English (dev)	15	
ERE05-Spanish (test)	14	Personnel:Nominate

Table 10. Dataset preparation for FSCLED. The total number of remaining types is shown for each data section alongside the removed subtypes without a sufficient number of samples for episodic training.

5.4.2 Training Details.

5.4.2.1 Episode Composition. In all our experiments, English is considered the sole source language as it is often used as the benchmark source language in cross-lingual efforts. As such, training and development episodes are constructed from English data. However, given the FSL constraints, their samples must come from disjoint label sets. Hence, in any training iteration, the samples used for both the support $\mathcal{S}$ and query $\mathcal{Q}$ sets are in English and belong to the training subtypes of the `Business, Contact, Conflict`, or `Justice` types. In contrast, during validation, $\mathcal{S}$ and $\mathcal{Q}$ will still be in English but their samples belong to the validation subtypes of the `Life, Movement, Personnel`, or `Transaction` types.

Furthermore, as cross-lingual models are evaluated on the target language, during testing, episodes are created from target-language data and their samples

belong to the same types as the development episodes, i.e., the `Life, Movement, Personnel`, or `Transaction` types.

5.4.3 Results. We compare our Optimal-Transport-based Event Detection (OTED) model, against three typical FSL models adapted to FSCLED as the baselines: Matching networks Vinyals et al. (2016), Prototypical networks Snell et al. (2017), and Relation networks Sung et al. (2018). All models utilize the same mBERT-based encoder for a fair comparison. We use English as the source language during training as it is recurrently utilized the source-language benchmark (Majewska et al., 2021b; M'hamdi et al., 2019b) due to its high-resource availability.

Our main experiment results are presented in Table 11 which shows that our OTED model consistently outperforms the best-performing baselines in every target language: Chinese (+0.21%), Arabic (+0.59%), and Spanish (+1.35%). We believe these results validate OTED as a suitable and effective alternative for FSCLED.

	Target Language								
	Chinese			**Arabic**			**Spanish**		
Model Version	P	R	F1	P	R	F1	P	R	F1
Relation	78.62	79.1	78.86	52.89	53.35	53.12	48.53	48.77	48.65
Matching	85.44	85.79	85.64	66.21	65.92	66.06	56.77	56.95	56.86
Prototypical	85.81	86.12	85.96	70.02	70.44	70.23	60.87	61.17	61.02
OTED (ours)	**86.05**	**86.29**	**86.17**	**70.66**	**70.98**	**70.82**	**62.25**	**62.49**	**62.37**

Table 11. Performance for cross-lingual few-shot event detection. English is the source language used for training. The experiments for Chinese and Arabic are done over ACE05 while ERE05 is used for Spanish.

Furthermore, an additional benefit of OTED's training signals (i.e., the loss terms $\mathcal{L}_{ident}$, $\mathcal{L}_{dist}$, and $\mathcal{L}_{cross}$) is that they can be directly integrated into any existing FSL methods. Thus, we conduct a supplementary set of experiments where we integrate the loss function terms from OTED into Relation, Matching, and Prototypical networks (i.e., combining our training signals in OTED with

the standard cross-entropy losses of such FSL baselines). The performance for these integrated models are presented in Table 12. Comparing the corresponding performance in Tables 11 and 12, it is evident that integrating OTED with traditional FSL methods leads to overall performance improvement across different target languages and FSL models, further demonstrating the benefits and applicability of OTED for FSCLED.

	Target Language		
Model Version	**Chinese**	**Arabic**	**Spanish**
Relation + OTED	79.36	53.41	48.89
Matching + OTED	85.88	66.21	56.97
Prototypical + OTED	**86.42**	**71.11**	**62.43**

Table 12. Model performance for integrating OTED into traditional FSL methods. F1 scores are reported.

The model implementation details can be found in Appendix C.

5.4.4 Ablation study. To evaluate the contribution of the different proposed components (i.e., $\mathcal{L}_{ident}$, $\mathcal{L}_{dist}$, and $\mathcal{L}_{cross}$), we perform an ablation study whose outcomes are presented in Table 13. The left-most column indicates the components being removed from the overall loss $\mathcal{L}$. The first two rows show the performance when either the Wasserstein-distance loss term, i.e., $\mathcal{L}_{dist}$ or $\mathcal{L}_{cross}$ is removed. As expected, removing any of them hurts the performance of OTED across different target languages. This demonstrates the importance of considering the global distances between query and support sets, and the necessity of adapting to the cross-lingual setting by leveraging unlabeled target-language data. Furthermore, the performance of OTED suffers even more when both $\mathcal{L}_{dist}$ and $\mathcal{L}_{cross}$ are excluded.

Similarly, when $\mathcal{L}_{ident}$ is removed in the last row, the performance is also further reduced, dropping significantly by more than 1.5% for Chinese and Arabic

Model	Target Language		
	Chinese	Arabic	Spanish
OTED (full)	86.17	70.82	62.37
-$\mathcal{L}_{dist}$	85.63	70.57	61.85
-$\mathcal{L}_{cross}$	85.45	70.22	61.78
-$\mathcal{L}_{dist}$ -$\mathcal{L}_{cross}$	85.25	69.44	61.19
-$\mathcal{L}_{ident}$ - $\mathcal{L}_{dist}$ - $\mathcal{L}_{cross}$	84.67	68.21	60.65

Table 13. Ablation results over the test data.

compared to the full model. Note that removing $\mathcal{L}_{ident}$ has deeper implications as, in such case, the event-presence module F is not trained. In turn, the $P(\mathcal{S})$ and $P(\mathcal{Q})$ distributions for the support and query sets cannot be estimated reliably and are instead initialized using uniform distributions in the OT computation. These results thus confirm the usefulness of the event identification loss to support the OT computation in our model.

5.5 Related Work

Event detection has been thoroughly studied over the years. Early ED efforts were based on hand-crafted features (Ahn, 2006; Hong et al., 2011; Ji & Grishman, 2008; Q. Li et al., 2013; Liao & Grishman, 2010a, 2010b; McClosky et al., 2011; Miwa et al., 2014; Patwardhan & Riloff, 2009; B. Yang & Mitchell, 2016). More recently, deep learning techniques such as recurrent neural networks (T. H. Nguyen, Cho, & Grishman, 2016; T. M. Nguyen & Nguyen, 2019; Sha et al., 2018), convolutional neural networks (Y. Chen et al., 2015; T. H. Nguyen, Fu, et al., 2016; T. H. Nguyen & Grishman, 2015b), graph convolutional networks (T. H. Nguyen & Grishman, 2018; Yan et al., 2019), adversarial networks (Hong et al., 2018)T. Zhang et al. (2019), pre-trained language models (J. Liu et al., 2020; Wadden et al., 2019; S. Yang et al., 2019a; J. Zhang et al., 2019; Y. Zhang et al., 2020), and generative models (Y. Lu et al., 2021) have

been prevalent. Nevertheless, these works study ED under a supervised or semi-supervised setting.

Alternatively, ED was recently formulated as a few-shot task V. Lai, Dernoncourt, and Nguyen (2021). In a short time, several methods have been proposed using a variety of techniques such as meta-learningDeng et al. (2020); Shen et al. (2021), cross-task prototyping V. Lai, Dernoncourt, and Nguyen (2021), dependency graphs V. D. Lai et al. (2021), causal modeling Cong et al. (2021), and label dependency via conditional random fields J. Chen et al. (2021).

Previous works on cross-lingual ED generally make use of cross-lingual resources such as bilingual dictionaries or parallel corpora (J. Liu et al., 2019; Muis et al., 2018b) to address the differences between languages. More recent approaches exploit the language-invariant characteristics of pre-trained multilingual language models (Hambardzumyan et al., 2020b) along with complementary features such as label dependency (M'hamdi et al., 2019b), verb-class knowledge Majewska et al. (2021b), and class-aware cross-lingual alignment (M. V. Nguyen, Nguyen, et al., 2021).

Optimal transport has also been recently used in cross-lingual settings for information extraction tasks such as event co-reference resolution (Phung, Minh Tran, et al., 2021) and event detection (Guzman-Nateras, Nguyen, & Nguyen, 2022). However, the amalgamation of the few-shot and cross-lingual settings creates unique challenges that have not been tackled by any related work. Consequently, our proposed use of OT differs from related works as it addresses the global alignment between the support and query sets for few-shot learning and between source and target languages for the cross-lingual setting.

5.6 Summary

In summary, we consider the main contributions of this chapter to be the following:

- To the best of our knowledge, this is the first effort at integrating the few-shot and cross-lingual settings for the event detection task. This novel setting combines the limited training-data conditions of FSL with zero-shot cross-lingual transfer learning.
- To provide foundation for future research, we first evaluate the performance of representative FSL methods Snell et al. (2017); Sung et al. (2018); Vinyals et al. (2016) in this task.
- We propose a novel optimal-transport-based method for FSL classification that leverages the optimal alignment between the support and query samples.
- We address a limitation of traditional FSL methods by incorporating a novel regularization term that considers the global distance between the support and query sets.
- To adapt our approach to the cross-lingual setting, we promote language-invariant representation learning by integrating the distance between source and target data into our model.
- Our experiments on three diverse target languages (Arabic, Chinese, and Spanish) show that our approach improves the best-performing FSL methods in the new FSCLED setting and that our proposed training signals can be seamlessly incorporated with other FSL models to improve their performance on the challenging FSCLED task.

CHAPTER VI

CONCLUSIONS AND FUTURE DIRECTIONS

I was the primary author for this chapter and Thien Nguyen provided editorial suggestions.

Finally, in this chapter, we discuss our conclusions and present several suitable directions for future CLED research efforts.

6.1 Conclusions

As stated in Chapter I, the holistic objective of this dissertation was to advance the field of cross-lingual learning by designing strategies that improve event-detection performance under a cross-lingual setting. In consequence, in this work, we explored three novel cross-lingual event detection approaches by addressing the task from diverse perspectives.

First, we proposed a direct-transfer-based approach whose characterizing trait is to refine the standard adversarial language-adaptation scheme via mindful selection of the samples used to train the language discriminator. We perform such sample selection by taking into account each sample's overall similarity with the alternative-language samples and its likelihood to contain an event. This method generates fine-grained language-invariant word representations that result in improved cross-lingual performance, as confirmed by its superior handling of complex cross-lingual complications such as polysemous triggers.

Our second approach expands upon this idea by leveraging a knowledge distillation framework to reap the benefits of the data-transfer paradigm. In this method, a teacher network is trained to generate language-invariant representations via standard adversarial language adaptation. Afterward, such teacher network is used to obtain soft labels (i.e., probability distributions) for unlabeled target-

language samples. Then, these soft-labeled samples go through a hierarchical selection process in which both global and pairwise similarity measures are considered. This selection procedure is meant to filter out potentially noisy labels generated by the teacher network. Then, the selected soft-labeled samples are used to train a student network. Since the student network is trained using target data directly, it is able to learn from language-specific lexical information and word-label relations. At the time of this writing, this method achieves state-of-the-art performance for the cross-lingual event detection task.

Next, in our third approach, we present the entirely novel setting of few-shot cross-lingual event detection. This setting combines the limited training data requirements of few-shot learning with the zero-shot limitation of cross-lingual learning. We begin by proposing an innovative few-shot learning method based on the optimal-transport-obtained similarity between the support and query samples which is further regularized by including the global distance between these sets in the loss function computation. Lastly, we adapt our method to the cross-lingual setting by also incorporating the distance between source and target data samples into the loss, which encourages the learning of language-invariant representations. Our method outperforms traditional few-shot learning methods and our proposed regularization terms can be combined with such traditional methods to improve their performance.

Furthermore, we would like to highlight that, while we leverage the event detection task as a testbed for our proposed cross-lingual methods, they can be naturally applied to other information-extraction tasks, such as entity mention detection or event argument extraction, with minimal or no changes at all. In

summary, we consider to have successfully accomplished the proposed dissertation objective.

6.2 Future Research Directions

We devote this section to discussing a number of promising research directions for future cross-lingual information extraction efforts.

6.2.1 Generative/Prompting Models.

With the recent advancements in generative language models like BART (Lewis et al., 2020), T5 (Raffel et al., 2019), or GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023), 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 (S. Li, Ji, & Han, 2021), and end-to-end event extraction Hsu et al. (2022); Y. Lu et al. (2021). 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, K.-H. Huang, Hsu, Natarajan, Chang, and Peng (2022b) 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. Furthermore, despite the widespread adoption of LLM-powered tools like ChatGPT[1] and their undeniable success on many complex NLP-related tasks, recent studies (V. D. Lai et al., 2023) have revealed that their

[1] https://chat.openai.com/

multilingual performance on is still not on par with task-specific models. These findings suggest that further research into multilingual understanding is needed.

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, Lai, 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.

6.2.2 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, Dhariwal, Nichol, Chu, & Chen, 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 information extraction 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, Eisen, Rekathati, & Sahlgren, 2022).

6.2.3 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 [2] supports 133 languages at various levels and tool-kits such as Trankit (M. V. Nguyen, Lai, Pouran Ben Veyseh, & Nguyen, 2021) 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.

6.2.4 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 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

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

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 (Q. Wu, Lin, Wang, et al., 2020) 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.

6.2.5 Robust Training. Robust training aims at creating models that are not affected by noise or perturbations in the input data (Goodfellow, Shlens, & Szegedy, 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 (S. Wu & 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, K. Huang, Ahmad, Peng, and Chang (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.

APPENDIX A

DATASETS

A.1 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

A.2 Dataset Statistics

Table 14. 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

Table 15. 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

Table 16. ACE05/ERE and MINION dataset ED stats: number of sentences and triggers that the ACE05, ACE05-ERE, and MINION datasets contain for each language.

Dataset	**Lang**	**# Sent**	**# Trig**	**# Tr/St**
ACE05	**En**	20,818	5,311	0.255
	Zh	7,914	3,333	0.421
	Ar	3,118	2,270	0.728
ERE	**En**	16,510	7,530	0.456
	Es	8,169	3,751	0.459
MINION	**En**	65,000	17,644	0.271
	Tr	22,400	8,394	0.374
	Pl	22,395	11,891	0.531
	Es	16,340	6,063	0.371
	Ja	7,500	1,730	0.231
	Ko	7,500	1,526	0.203
	Pt	7,500	1,875	0.25
	Hi	7,495	1,811	0.241

APPENDIX B

MODEL PERFORMANCE COMPARISON

B.1 Entity Mention Detection

Table 17 presents the cross-lingual EMD performance of the works discussed in section 2.3 when tested on the commonly-used CoNLL-2002 (Tjong Kim Sang, 2002) and CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003) datasets. Detailed information about these datasets can be found in Appendix A.2.

	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. (2018b)	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
S. 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
Q. Wu, Lin, Wang, et al. (2020)	76.75	80.44	73.16
Q. Wu, Lin, Karlsson, Lou, and Huang (2020)	76.94	80.89	72.32
Q. Wu, Lin, Karlsson, Huang, and Lou (2020)	77.30	81.20	73.61
Liang et al. (2021)	77.84	82.46	75.48
W. Chen et al. (2021)	79.00	82.90	75.01

Table 17. EMD model performance comparison on the CoNLL-2002 & 2003 datasets. English is used as the source language.

B.2 Event Detection

Table 18 presents the CLED performance of the works discussed in section 2.4.1 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 A.2.

B.2.1 Event Argument Extraction. Table 19 presents a comparison between the cross-lingual EAE performance of the works discussed

	Target		
Model	**ZH**	**AR**	**ES**
J. Liu et al. (2019)	27.0	-	-
M'hamdi et al. (2019)	68.5	30.9	-
D. Lu et al. (2020)	-	-	41.77
Fincke et al. (2021)	-	51.0	-
Majewska et al. (2021a)	46.9	29.3	-
M. V. Nguyen, Nguyen, et al. (2021)	72.1	42.7	-
Guzman-Nateras, Nguyen, and Nguyen (2022)	74.64	44.86	47.69
Guzman-Nateras et al. (2023)	75.22	46.37	48.58

Table 18. Model performance comparison on the ED for the ACE05 dataset. English is used as the source language.

in section 2.4.3when 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 A.2.

	Target		
Model	**ZH**	**AR**	**ES**
Subburathinam et al. (2019)	59.0	61.8	-
D. Lu et al. (2020)	-	-	17.35
Majewska et al. (2021a)	1.9	7.1	-
M. V. Nguyen and Nguyen (2021)	58.4	62.9	-
W. Ahmad et al. (2021)	63.2	68.5	-
Fincke et al. (2021)	-	74.7	-
M. V. Nguyen, Nguyen, et al. (2021)	65.5	69.4	-
K.-H. Huang et al. (2022a)	54.0	44.8	59.7

Table 19. Model performance comparison on the EAE for the ACE05 dataset. English is used as the source language.

B.3 Relation Extraction

Table 20 presents a comparison between the cross-lingual RE performance of the works discussed in section 2.5 when tested on the commonly-used ACE05 Walker et al. (2006) dataset. Detailed information about this dataset can be found in Appendix A.2.

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

Table 20. Model performance on the RE for the ACE05 dataset. English is used as the source language.

B.4 Co-Reference Resolution

The research efforts discussed in section 2.6 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 a common dataset and cannot be directly compared.

APPENDIX C

MODEL IMPLEMENTATION DETEAILS

C.1 OACLED (Chapter III)

We fine-tune the hyper-parameters for our OACLED model using the development data. We apply the following values based on the fine-tuning process:

- AdamW (Loshchilov & Hutter, 2017) as the optimizer.
- 5 warm up epochs.
- A learning rate of $1e^{-5}$ for the transformer parameters and of $1e^{-4}$ for the rest of the parameters.
- A batch size of 16.
- 300 for the dimensionality of the layers in feed-forwards networks.
- A $\gamma = 0.5$ for the percentage of samples used in adversarial training.
- A $\lambda = 0.001$ as the scaling factor of the GRL layer.
- An $\alpha = 1$ and $\beta = 0.001$ as the trade-off parameters of the LD loss and ED loss, respectively.
- A dropout of 10% for added regularization during training.
- We follow the same train/val/test splits as prior works (M'hamdi et al., 2019; Pouran Ben Veyseh et al., 2022). We tune all hyperparameters on the validation sets and report the performance on the test sets.

C.2 HKT-CLED (Chapter IV)

- A single Tesla V100-SXM2 GPU with 32GB memory and PyTorch 1.7.0 was used to implement the models.
- Our full model has 278.5M parameters. The vast majority of these come from XLM-Roberta (278M parameters), the rest of our model accounts for $< 500K$ parameters.
- We use AdamW (Loshchilov & Hutter, 2017) as the optimizer.
- We report label F1 scores computed using seqeval [1].
- We approximate the solution to the intractable problem described by Equation 4.9 by instead solving its entropy-based relaxation using the Sinkhorn iterative algorithm (Cuturi, 2013).
- Following prior works (Q. Wu, Lin, Karlsson, Huang, & Lou, 2020), we freeze the embeddings and first three layers of the XLM-R encoder for student training.
- Representations of words split into multiple word-pieces by the tokenizer are the average of representation vectors for all comprising sub-pieces.
- Learning rate for the transformer parameters is set at $2e^{-5}$ and was found through greedy search over [$2e^{-6}$, $5e^{-6}$, $1e^{-5}$, $2e^{-5}$, $5e^{-5}$, $1e^{-4}$, $2e^{-4}$]
- Learning rate for non-transformer parameters is set at $1e^{-4}$ and was found through greedy search over [$5e^{-5}$, $1e^{-4}$, $2e^{-4}$, $5e^{-4}$, $1e^{-3}$].

[1] https://github.com/chakki-works/seqeval

- The α hyperparameter is set at 0.5 and was found through greedy search over $[0.25, 0.5, 0.75, 1.0]$.
- The β hyperparameter is set at 0.75 and was found through greedy search over $[0.25, 0.5, 0.75, 1.0]$.
- We employ a batch size of 32 for our experiments on the ACE05 and ACE05 datasets and a batch size of 16 on the MINION experiments. Batch size was chosen through greedy search over [8, 16, 24, 32].
- The linear layer sizes where set at 300 and chosen through greedy search over [100, 300, 500, 1000]
- We train the teacher model for 20 epochs and the student model for 100 epochs.
- We use a learning rate linear scheduler with 5 warm-up epochs for teacher models and 10 warm-up epochs for student models.
- We use a parameter weight decay of $1e^{-4}$ for non-transformer parameters chosen greedily from [$1e^{-3}$, $1e^{-4}$, $5e^{-4}$]
- We use a parameter weight decay of 0.5 for transformer parameters chosen greedily from [0.01, 0.05, 0.1, 0.3, 0.5].
- Depending on the dataset, language, and selected batch size our teacher model training takes between 45 min and 2 hours in a single GPU. Student models take between 2 and 4 hours to train. Our overall developing and parameter tuning process took around $\sim$ 600 GPU hours.

– We follow the same train/val/test splits as prior works (M'hamdi et al., 2019; Pouran Ben Veyseh et al., 2022). We tune all hyperparameters on the validation sets and report the performance on the test sets.

C.3 OTED (Chapter V)

– We use a single Tesla V100-SXM2 GPU with 32GB memory operated by Red Hat Enterprise Linux Server 7.8 (Maipo). PyTorch 1.4.0 is used to implement the models.

– We report F1 for trigger token classification complying with previous work V. Lai, Dernoncourt, and Nguyen (2021). The reported results are the average performance of five model runs with different random seeds.

– Our full model has 178.5M parameters. However, the vast majority of these come from the mBERT transformer encoder (178M parameters), the rest of our model accounts for $< 500K$ parameters.

– We utilize a fixed 6-way (5 event types plus the non-event), 5-shot setting for all the experiments.

– Following prior work V. Lai, Dernoncourt, and Nguyen (2021), we use a larger subset of classes during training $(10 + 1)$ as its been found to improve model performance.

– We initialize our encoder E with the pre-trained `bert-base-multilingual-cased` transformer model Devlin et al. (2019) and add a single linear layer followed by a hyperbolic tangent non-linearity on top.

– Our final encoder representations have 512 dimensions.

– AdamW Loshchilov and Hutter (2017) as the optimizer.

– Using 5 warm up epochs.

– Learning rate is set to $3e^{-4}$.

– The α, β and γ hyper-parameters are set to $0.1, 0.01$, and 0.01 respectively.

– The batch size is set to 16.

– 512 for the dimensionality of the layers in the feed-forward networks.

– A dropout of 10% for added regularization during training.

– All hyperparameters were tuned on the development data of the source language, and all reported values are the average obtained from five runs with different random seeds.

REFERENCES CITED

Ahmad, W., Peng, N., & Chang, K.-W. (2021). Gate: Graph attention transformer encoder for cross-lingual relation and event extraction. In *Aaai.* Retrieved from https://arxiv.org/abs/2010.03009

Ahmad, Z., Varshney, D., Ekbal, A., & Bhattacharyya, P. (2019, December). 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* (pp. 135–142). International Institute of Information Technology, Hyderabad, India: NLP Association of India. Retrieved from https://aclanthology.org/2019.icon-1.16

Ahn, D. (2006). The stages of event extraction. In *Proceedings of the workshop on annotating and reasoning about time and events.*

Artetxe, M., Labaka, G., & Agirre, E. (2018, July). 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)* (pp. 789–798). Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P18-1073 doi: 10.18653/v1/P18-1073

Artetxe, M., Labaka, G., & Agirre, E. (2020, November). Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 7674–7684). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.618 doi: 10.18653/v1/2020.emnlp-main.618

Awasthy, P., Naseem, T., Ni, J., Moon, T., & Florian, R. (2020). Event presence prediction helps trigger detection across languages. In *Corr.* Retrieved from https://arxiv.org/pdf/2009.07188.pdf

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Corr.* Retrieved from https://arxiv.org/abs/1409.0473

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet project. In *COLING 1998 volume 1: The 17th international conference on computational linguistics.* Retrieved from https://aclanthology.org/C98-1013

Baldini Soares, L., FitzGerald, N., Ling, J., & Kwiatkowski, T. (2019, July). Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2895–2905). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1279 doi: 10.18653/v1/P19-1279

Bari, M. S., Joty, S., & Jwalapuram, P. (2020). Zero-resource cross-lingual named entity recognition. In *Proceedings of the aaai conference on artificial intelligence.* Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/6237 doi: 10.1609/aaai.v34i05.6237

Bharadwaj, A., Mortensen, D., Dyer, C., & Carbonell, J. (2016, November). 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* (pp. 1462–1472). Austin, Texas: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D16-1153 doi: 10.18653/v1/D16-1153

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. Retrieved from https://aclanthology.org/Q17-1010 doi: 10.1162/tacl_a_00051

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In *Advances in neural information processing systems.* Retrieved from https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Bucilua, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining.* Retrieved from https://doi.org/10.1145/1150402.1150464 doi: 10.1145/1150402.1150464

Carlsson, F., Eisen, P., Rekathati, F., & Sahlgren, M. (2022, June). Cross-lingual and multilingual CLIP. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 6848–6854). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.739

Carreras, X., Màrquez, L., & Padró, L. (2003). A simple named entity extractor using adaboost. In *Proceedings of the seventh conference on natural language learning (conll).* Retrieved from https://doi.org/10.3115/1119176.1119197 doi: 10.3115/1119176.1119197

Caselli, T., Sprugnoli, R., Speranza, M., & Monachini, M. (2014). Eventi evaluation of events and temporal information at evalita 2014.. doi: 10.12871/clicit201425

Caselli, T., & Ustun, A. (2019). There and back again: Cross-lingual transfer learning for event detection. In *Clic-it.*

Cerberio, K., Aduriz, I., Diaz de Ilarraza, A., & Garcia-Azkoaga, I. (2018). Coreferential relations in basque: The annotation process. In *Journal of psycholinguistic research.* Retrieved from https://link.springer.com/article/10.1007/s10936-018-9559-6 doi: https://doi.org/10.1007/s10936-018-9559-6

Chaudhary, A., Zhou, C., Levin, L., Neubig, G., Mortensen, D. R., & Carbonell, J. (2018, October-November). 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* (pp. 3285–3295). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-1366 doi: 10.18653/v1/D18-1366

Chen, J., Lin, H., Han, X., & Sun, L. (2021, November). Honey or poison? solving the trigger curse in few-shot event detection via causal intervention. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 8078–8088). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.637 doi: 10.18653/v1/2021.emnlp-main.637

Chen, W., Jiang, H., Wu, Q., Karlsson, B., & Guan, Y. (2021, August). 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)* (pp. 743–753). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-long.61 doi: 10.18653/v1/2021.acl-long.61

Chen, X., & Cardie, C. (2018, October-November). Unsupervised multilingual word embeddings. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 261–270). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-1024 doi: 10.18653/v1/D18-1024

Chen, X., Sun, Y., Athiwaratkun, B., Cardie, C., & Weinberger, K. (2018). Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification. In *Transactions of the association for computational linguistics.* Retrieved from https://doi.org/10.1162/tacl_a_00039 doi: 10.1162/tacl_a_00039

Chen, Y., Xu, L., Liu, K., Zeng, D., & Zhao, J. (2015). Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the annual meeting of the association for computational linguistics (acl).*

Cohen, S. B., Das, D., & Smith, N. (2011). Unsupervised structure prediction with nonparallel multilingual guidance. In *Proceedings of the conference on empirical methods in natural language processing (emnlp).*

Cong, X., Cui, S., Yu, B., Liu, T., Yubin, W., & Wang, B. (2021, August). Few-Shot Event Detection with Prototypical Amortized Conditional Random Field. In *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 28–40). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-acl.3 doi: 10.18653/v1/2021.findings-acl.3

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. In *Corr.* Retrieved from http://arxiv.org/abs/1911.02116

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics.* Retrieved from http://arxiv.org/abs/1911.02116

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2017). Word translation without parallel data. In *Corr.* Retrieved from http://dblp.uni-trier.de/db/journals/corr/corr1710.html#abs-1710-04087

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. In *Corr.* Retrieved from http://dblp.uni-trier.de/db/journals/corr/corr1710.html#abs-1710-04087

Cruz, A. F., Rocha, G., & Cardoso, H. L. (2018). Exploring spanish corpora for portuguese coreference resolution. In *2018 fifth international conference on social networks analysis, management and security (snams).* Retrieved from https://ieeexplore.ieee.org/document/8554705 doi: 10.1109/SNAMS.2018.8554705

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the 26th international conference on neural information processing systems - volume 2.*

Deng, S., Zhang, N., Kang, J., Zhang, Y., Zhang, W., & Chen, H. (2020). Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. In *Proceedings of the 13th international conference on web search and data mining.* Retrieved from https://doi.org/10.1145/3336191.3371796

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Naacl-hlt.*

dos Santos, C., Xiang, B., & Zhou, B. (2015, July). 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)* (pp. 626–634). Beijing, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P15-1061 doi: 10.3115/v1/P15-1061

Dou, Z.-Y., & Neubig, G. (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.* Retrieved from https://aclanthology.org/2021.eacl-main.181 doi: 10.18653/v1/2021.eacl-main.181

Du, X., & Cardie, C. (2020, November). Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 671–683). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.49 doi: 10.18653/v1/2020.emnlp-main.49

Duong, L., Kanayama, H., Ma, T., Bird, S., & Cohn, T. (2016, November). Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1285–1295). Austin, Texas: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D16-1136 doi: 10.18653/v1/D16-1136

Dyer, C., Chahuneau, V., & Smith, N. A. (2013, June). 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.* Retrieved from https://aclanthology.org/N13-1073

Ehrmann, M., Turchi, M., & Steinberger, R. (2011, September). Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the international conference recent advances in natural language processing 2011* (pp. 118–124). Hissar, Bulgaria: Association for Computational Linguistics. Retrieved from https://aclanthology.org/R11-1017

Faruqui, M., & Kumar, S. (2015, May–June). 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* (pp. 1351–1356). Denver, Colorado: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N15-1151 doi: 10.3115/v1/N15-1151

Feng, X., Feng, X., Qin, B., Feng, Z., & Liu, T. (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. Retrieved from https://doi.org/10.24963/ijcai.2018/566 doi: 10.24963/ijcai.2018/566

Fincke, S., Agarwal, S., Miller, S., & Boschee, E. (2021). Language model priming for cross-lingual event extraction. In *Association for the advancement of artificial intelligence (aaai).*

Fonseca, E., Sesti, V., Collovini, S., Vieira, R., Leal, A., & Quaresma, P. (2017). Collective elaboration of a coreference annotated corpus for portuguese texts.. Retrieved from http://ceur-ws.org/Vol-1881/Overview3.pdf

Fu, R., Qin, B., & Liu, T. (2014). Generating chinese named entity data from parallel corpora. In *Frontiers of computer science* (Vol. 8, p. 629-641).

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd international conference on machine learning.*

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *3rd international conference on learning representations (iclr).* Retrieved from https://arxiv.org/abs/1412.6572

Guzman-Nateras, L., Dernoncourt, F., & Nguyen, T. (2023). Hybrid knowledge transfer for improved cross-lingual event detection via hierarchical sample selection. In *To appear at acl 2023.*

Guzman-Nateras, L., Lai, V., Pouran Ben Veyseh, A., Dernoncourt, F., & Nguyen, T. (2022, July). Event detection for suicide understanding. In *Findings of the association for computational linguistics: Naacl 2022* (pp. 1952–1961). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.findings-naacl.150 doi: 10.18653/v1/2022.findings-naacl.150

Guzman-Nateras, L., Nguyen, M. V., & Nguyen, T. (2022, July). 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* (pp. 5588–5599). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.naacl-main.409 doi: 10.18653/v1/2022.naacl-main.409

Hambardzumyan, K., Khachatrian, H., & May, J. (2020a). 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).* Retrieved from https://www.researchgate.net/profile/Ashot-Harutyunyan-2/publication/344165125_logos_Collaborative_Technologies_and_Data_Science_in_Artificial_Intelligence_Applications/links/5f578f3d458515e96d3962fe/logos-Collaborative-Technologies-and-Data-Science-in-Artificial-Intelligence-Applications.pdf#page=109

Hambardzumyan, K., Khachatrian, H., & May, J. (2020b). The role of alignment of multilingual contextualized embeddings in zero-shot cross-lingual transfer for event extraction. In *Collaborative technologies and data science in artificial intelligence applications.*

He, K., Yan, Y., & Xu, W. (2020). Adversarial cross-lingual transfer learning for slot tagging of low-resource languages. In *2020 international joint conference on neural networks (ijcnn)* (p. 1-8). doi: 10.1109/IJCNN48605.2020.9207607

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. In *Corr.* Retrieved from https://arxiv.org/abs/1503.02531 doi: 10.48550/ARXIV.1503.02531

Holger, S., & Xian, L. (2018). A corpus for multiligual document classification in eight languages. In *Proceedings of the eleventh international conference on language resources and evaluation (lrec).*

Hong, Y., Zhang, J., Ma, B., Yao, J., Zhou, G., & Zhu, Q. (2011). Using cross-entity inference to improve event extraction. In *Proceedings of the annual meeting of the association for computational linguistics (acl).*

Hong, Y., Zhou, W., Zhang, J., Zhou, G., & Zhu, Q. (2018, July). Self-regulation: Employing a generative adversarial network to improve event detection. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 515–526). Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P18-1048 doi: 10.18653/v1/P18-1048

Hsu, I.-H., Huang, K.-H., Boschee, E., Miller, S., Natarajan, P., Chang, K.-W., & Peng, N. (2022, July). 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* (pp. 1890–1908). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.naacl-main.138 doi: 10.18653/v1/2022.naacl-main.138

Hu, H., Xie, L., Hong, R., & Tian, Q. (2020). Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *2020 ieee/cvf conference on computer vision and pattern recognition (cvpr).* Retrieved from https://ieeexplore.ieee.org/document/9156328 doi: 10.1109/CVPR42600.2020.00319

Huang, K., Ahmad, W. U., Peng, N., & Chang, K. (2021). Improving zero-shot cross-lingual transfer learning via robust training.. Retrieved from https://arxiv.org/abs/2104.08645

Huang, K.-H., Hsu, I.-H., Natarajan, P., Chang, K.-W., & Peng, N. (2022a). Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *Acl.* Retrieved from https://arxiv.org/abs/2203.08308 doi: 10.48550/ARXIV.2203.08308

Huang, K.-H., Hsu, I.-H., Natarajan, P., Chang, K.-W., & Peng, N. (2022b, May). 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)* (pp. 4633–4646). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.acl-long.317 doi: 10.18653/v1/2022.acl-long.317

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., & Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. In *Natural language engineering.* Retrieved from https://doi.org/10.1017/S1351324905003840 doi: 10.1017/S1351324905003840

Jain, A., Paranjape, B., & Lipton, Z. C. (2019, November). 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)* (pp. 1083–1092). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1100 doi: 10.18653/v1/D19-1100

Ji, H., & Grishman, R. (2008). Refining event extraction through cross-document inference. In *Proceedings of the annual meeting of the association for computational linguistics (acl).*

Ji, H., Nothman, J., Hachey, B., & Florian, R. (2015). Overview of tac-kbp2015 tri-lingual entity discovery and linking. In *Theory and applications of categories.* Retrieved from https://tac.nist.gov/publications/2015/additional.papers/TAC2015.KBP_Trilingual_Entity_Discovery_and_Linking_overview.proceedings.pdf

Joty, S., Nakov, P., Màrquez, L., & Jaradat, I. (2017). Cross-language learning with adversarial neural networks. In *Proceedings of the 21st conference on computational natural language learning (CoNLL)* (pp. 226–237). Retrieved from https://aclanthology.org/K17-1024 doi: 10.18653/v1/K17-1024

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., & Grave, E. (2018, October-November). Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2979–2984). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-1330 doi: 10.18653/v1/D18-1330

Kambhatla, N. (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.* Retrieved from https://doi.org/10.3115/1219044.1219066 doi: 10.3115/1219044.1219066

Kang, M., Mun, J., & Han, B. (2020). Towards oracle knowledge distillation with neural architecture search. In *Proceedings of the aaai conference on artificial intelligence.* Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/5866 doi: 10.1609/aaai.v34i04.5866

Keung, P., Lu, Y., & Bhardwaj, V. (2019, November). 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)* (pp. 1355–1360). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1138 doi: 10.18653/v1/D19-1138

Kim, S., Jeong, M., Lee, J., & Lee, G. G. (2014). Cross-lingual annotation projection for weakly-supervised relation extraction. In *Acm transactions on asian language information processing.* Retrieved from https://doi.org/10.1145/2529994 doi: 10.1145/2529994

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International conference on learning representations (iclr).* Retrieved from http://arxiv.org/abs/1609.02907

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2006, May). 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). Retrieved from http://www.lrec-conf.org/proceedings/lrec2006/pdf/468_pdf.pdf

Köksal, A., & Özgür, A. (2020, November). The RELX dataset and matching the multilingual blanks for cross-lingual relation classification. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 340–350). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.32 doi: 10.18653/v1/2020.findings-emnlp.32

Kozhevnikov, M., & Titov, I. (2014, June). 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)* (pp. 579–585). Baltimore, Maryland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P14-2095 doi: 10.3115/v1/P14-2095

Kundu, G., Sil, A., Florian, R., & Hamza, W. (2018, July). 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)* (pp. 395–400). Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P18-2063 doi: 10.18653/v1/P18-2063

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning.*

Lai, V., Dernoncourt, F., & Nguyen, T. H. (2021, November). Learning prototype representations across few-shot tasks for event detection. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 5270–5277). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.427 doi: 10.18653/v1/2021.emnlp-main.427

Lai, V., Nguyen, M. V., Kaufman, H., & Nguyen, T. H. (2021). Event extraction from historical texts: A new dataset for black rebellions. In *Findings of the association for computational linguistics: Acl-ijcnlp 2021.* Retrieved from https://aclanthology.org/2021.findings-acl.211 doi: 10.18653/v1/2021.findings-acl.211

Lai, V. D., Dernoncourt, F., & Nguyen, T. H. (2020). Exploiting the matching information in the support set for few shot event classification. In *Advances in knowledge discovery and data mining.* Retrieved from https://arxiv.org/pdf/2002.05295.pdf

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023). Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *Corr.*

Lai, V. D., Nguyen, M. V., Nguyen, T. H., & Dernoncourt, F. (2021). Graph learning regularization and transfer learning for few-shot event detection. In *Proceedings of the 44th international acm sigir conference on research and development in information retrieval.* Retrieved from https://doi.org/10.1145/3404835.3463054

Lai, V. D., Nguyen, T. H., & Dernoncourt, F. (2020, July). Extensively matching for few-shot learning event detection. In *Proceedings of the first joint workshop on narrative understanding, storylines, and events* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.nuse-1.5 doi: 10.18653/v1/2020.nuse-1.5

Lai, V. D., Nguyen, T. N., & Nguyen, T. H. (2020, November). Event detection: Gate diversity and syntactic importance scores for graph convolution neural networks. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 5405–5411). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.435 doi: 10.18653/v1/2020.emnlp-main.435

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. In *International conference on learning representations (iclr).*

Lample, G., Denoyer, L., & Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. In *Corr.* Retrieved from http://arxiv.org/abs/1711.00043

Langedijk, A., Dankers, V., Lippe, P., Bos, S., Cardenas Guevara, B., Yannakoudakis, H., & Shutova, E. (2022, May). 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)* (pp. 8503–8520). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.acl-long.582 doi: 10.18653/v1/2022.acl-long.582

Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020, November). 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)* (pp. 4483–4499). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.363 doi: 10.18653/v1/2020.emnlp-main.363

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... Zettlemoyer, L. (2020, July). 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* (pp. 7871–7880). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.703 doi: 10.18653/v1/2020.acl-main.703

Li, Q., & Ji, H. (2014, June). 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)* (pp. 402–412). Baltimore, Maryland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P14-1038 doi: 10.3115/v1/P14-1038

Li, Q., Ji, H., & Huang, L. (2013). Joint event extraction via structured prediction with global features. In *Proceedings of the annual meeting of the association for computational linguistics (acl).*

Li, S., Ji, H., & Han, J. (2021, June). 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* (pp. 894–908). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.69 doi: 10.18653/v1/2021.naacl-main.69

Liang, S., Gong, M., Pei, J., Shou, L., Zuo, W., Zuo, X., & Jiang, D. (2021). Reinforced iterative knowledge distillation for cross-lingual named entity recognition. In *Corr.* Retrieved from https://arxiv.org/abs/2106.00241

Liao, S., & Grishman, R. (2010a). Filtered ranking for bootstrapping in event extraction. In *Proceedings of the international conference on computational linguistics (coling).*

Liao, S., & Grishman, R. (2010b). Using document level cross-event inference to improve event extraction. In *Proceedings of the annual meeting of the association for computational linguistics (acl).*

Liu, J., Chen, Y., Liu, K., Bi, W., & Liu, X. (2020). Event extraction as machine reading comprehension. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp).* Retrieved from https://aclanthology.org/2020.emnlp-main.128 doi: 10.18653/v1/2020.emnlp-main.128

Liu, J., Chen, Y., Liu, K., & Zhao, J. (2019, November). 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)* (pp. 738–748). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1068 doi: 10.18653/v1/D19-1068

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, *8*, 726–742. Retrieved from https://aclanthology.org/2020.tacl-1.47 doi: 10.1162/tacl_a_00343

Loshchilov, I., & Hutter, F. (2017). Fixing weight decay regularization in adam.. Retrieved from http://arxiv.org/abs/1711.05101

Lu, D., Subburathinam, A., Ji, H., May, J., Chang, S.-F., Sil, A., & Voss, C. (2020, May). Cross-lingual structure transfer for zero-resource event extraction. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 1976–1981). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.243

Lu, Y., Lin, H., Xu, J., Han, X., Tang, J., Li, A., ... Chen, S. (2021, August). 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)* (pp. 2795–2806). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-long.217 doi: 10.18653/v1/2021.acl-long.217

Majewska, O., Vulić, I., Glavaš, G., Ponti, E. M., & Korhonen, A. (2021a). 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.* Retrieved from https://aclanthology.org/2021.acl-long.541 doi: 10.18653/v1/2021.acl-long.541

Majewska, O., Vulić, I., Glavaš, G., Ponti, E. M., & Korhonen, A. (2021b, August). 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 (volume 1: Long papers)* (pp. 6952–6969). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-long.541 doi: 10.18653/v1/2021.acl-long.541

Mayhew, S., Tsai, C.-T., & Roth, D. (2017, September). Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2536–2545). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D17-1269 doi: 10.18653/v1/D17-1269

McClosky, D., Charniak, E., & Johnson, M. (2010, June). Automatic domain adaptation for parsing. In *Human language technologies: The 2010 annual conference of the north American chapter of the association for computational linguistics* (pp. 28–36). Los Angeles, California: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N10-1004

McClosky, D., Surdeanu, M., & Manning, C. (2011). Event extraction as dependency parsing. In *Bionlp shared task workshop.*

M'hamdi, M., Freedman, M., & May, J. (2019). Contextualized cross-lingual event trigger extraction with minimal resources. In *Conference on computational natural language learning (conll).* Retrieved from https://aclanthology.org/K19-1061.pdf

M'hamdi, M., Freedman, M., & May, J. (2019a). Contextualized cross-lingual event trigger extraction with minimal resources. In *Proceedings of the 23rd conference on computational natural language learning (conll).* Retrieved from https://aclanthology.org/K19-1061 doi: 10.18653/v1/K19-1061

M'hamdi, M., Freedman, M., & May, J. (2019b, November). Contextualized cross-lingual event trigger extraction with minimal resources. In *Proceedings of the 23rd conference on computational natural language learning (conll)* (pp. 656–665). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/K19-1061 doi: 10.18653/v1/K19-1061

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *arxiv.* Retrieved from https://arxiv.org/abs/1301.3781 doi: 10.48550/ARXIV.1301.3781

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. In *Corr* (Vol. abs/1309.4168). Retrieved from http://arxiv.org/abs/1309.4168

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 26). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Min, B., Jiang, Z., Freedman, M., & Weischedel, R. (2017, November). 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)* (pp. 674–684). Taipei, Taiwan: Asian Federation of Natural Language Processing. Retrieved from https://aclanthology.org/I17-1068

Miwa, M., & Bansal, M. (2016, August). 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)* (pp. 1105–1116). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P16-1105 doi: 10.18653/v1/P16-1105

Miwa, M., Thompson, P., Korkontzelos, I., & Ananiadou, S. (2014). Comparable study of event extraction in newswire and biomedical domains. In *Proceedings of the international conference on computational linguistics (coling).*

Moon, T., Awasthy, P., Ni, J., & Florian, R. (2019). Towards lingua franca named entity recognition with BERT. In *Corr.* Retrieved from http://arxiv.org/abs/1912.01389

Muis, A. O., Otani, N., Vyas, N., Xu, R., Yang, Y., Mitamura, T., & Hovy, E. (2018a, August). Low-resource cross-lingual event type detection via distant supervision with minimal effort. In *Proceedings of the 27th international conference on computational linguistics* (pp. 70–82). Santa Fe, New Mexico, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/C18-1007

Muis, A. O., Otani, N., Vyas, N., Xu, R., Yang, Y., Mitamura, T., & Hovy, E. (2018b, August). Low-resource cross-lingual event type detection via distant supervision with minimal effort. In *Proceedings of the 27th international conference on computational linguistics.*

Naik, A., & Rose, C. (2020). Towards open domain event trigger identification using adversarial domain adaptation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. Retrieved from https://aclanthology.org/2020.acl-main.681 doi: 10.18653/v1/2020.acl-main.681

Ngo Trung, N., Phung, D., & Nguyen, T. H. (2021, August). Unsupervised domain adaptation for event detection using domain-specific adapters. In *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 4015–4025). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-acl.351 doi: 10.18653/v1/2021.findings-acl.351

Nguyen, M. V., Lai, V. D., & Nguyen, T. H. (2021, June). Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 27–38). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.3 doi: 10.18653/v1/2021.naacl-main.3

Nguyen, M. V., Lai, V. D., Pouran Ben Veyseh, A., & Nguyen, T. H. (2021, April). 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* (pp. 80–90). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.eacl-demos.10 doi: 10.18653/v1/2021.eacl-demos.10

Nguyen, M. V., & Nguyen, T. H. (2021, April). Improving cross-lingual transfer for event argument extraction with language-universal sentence structures. In *Proceedings of the sixth arabic natural language processing workshop* (pp. 237–243). Kyiv, Ukraine (Virtual): Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.wanlp-1.27

Nguyen, M. V., Nguyen, T. N., Min, B., & Nguyen, T. H. (2021, November). 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* (pp. 5414–5426). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.440 doi: 10.18653/v1/2021.emnlp-main.440

Nguyen, T. H., Cho, K., & Grishman, R. (2016). Joint event extraction via recurrent neural networks. In *Proceedings of the conference of the north american chapter of the association for computational linguistics: Human language technologies (naacl-hlt).*

Nguyen, T. H., Fu, L., Cho, K., & Grishman, R. (2016). A two-stage approach for extending event detection to new types via neural networks. In *Proceedings of the 1st acl workshop on representation learning for nlp (repl4nlp).*

Nguyen, T. H., & Grishman, R. (2015a). Combining neural networks and log-linear models to improve relation extraction. In *Corr.* Retrieved from http://arxiv.org/abs/1511.05926

Nguyen, T. H., & Grishman, R. (2015b). Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the annual meeting of the association for computational linguistics (acl).*

Nguyen, T. H., & Grishman, R. (2018). Graph convolutional networks with argument-aware pooling for event detection. In *Association for the advancement of artificial inteligence (aaai).*

Nguyen, T. M., & Nguyen, T. H. (2019). One for all: Neural joint modeling of entities and events. In *Association for the advancement of artificial inteligence (aaai).* Retrieved from https://arxiv.org/pdf/1812.00195.pdf

Ni, J., Dinu, G., & Florian, R. (2017). Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Corr.* Retrieved from http://arxiv.org/abs/1707.02483

Ni, J., & Florian, R. (2019, November). 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)* (pp. 399–409). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1038 doi: 10.18653/v1/D19-1038

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., ... Zeman, D. (2016, May). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 1659–1666). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L16-1262

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. In *Computational linguistics.*

OpenAI. (2023). Gpt-4 technical report. In *Corr.*

Paolini, G., Athiwaratkun, B., Krone, J., Ma, J., Achille, A., Anubhai, R., ... Soatto, S. (2021). Structured prediction as translation between augmented natural languages.. Retrieved from https://arxiv.org/abs/2101.05779

Patwardhan, S., & Riloff, E. (2009). A unified model of phrasal and sentential evidence for information extraction. In *Proceedings of the conference on empirical methods in natural language processing (emnlp).*

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D14-1162 doi: 10.3115/v1/D14-1162

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., ... Gurevych, I. (2020, October). AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 46–54). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-demos.7 doi: 10.18653/v1/2020.emnlp-demos.7

Phung, D., Minh Tran, H., Nguyen, M. V., & Nguyen, T. H. (2021, November). Learning cross-lingual representations for event coreference resolution with multi-view alignment and optimal transport. In *Proceedings of the 1st workshop on multilingual representation learning* (pp. 62–73). Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.mrl-1.6 doi: 10.18653/v1/2021.mrl-1.6

Phung, D., Tran, H. M., Nguyen, M. V., & Nguyen, T. H. (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).* Retrieved from https://aclanthology.org/2021.mrl-1.6

Pikuliak, M., Šimko, M., & Bieliková, M. (2021a). Cross-lingual learning for text processing: A survey. In *Expert systems with applications.* Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417420305893 doi: https://doi.org/10.1016/j.eswa.2020.113765

Pikuliak, M., Šimko, M., & Bieliková, M. (2021b). Cross-lingual learning for text processing: A survey. In *Expert systems with applications.* Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417420305893 doi: https://doi.org/10.1016/j.eswa.2020.113765

Pouran Ben Veyseh, A., Lai, V., Dernoncourt, F., & Nguyen, T. H. (2021, August). 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)* (pp. 6271–6282). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-long.490 doi: 10.18653/v1/2021.acl-long.490

Pouran Ben Veyseh, A., Nguyen, M. V., Dernoncourt, F., & Nguyen, T. (2022, July). MINION: a large-scale and diverse dataset for multilingual event detection. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 2286–2299). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.naacl-main.166 doi: 10.18653/v1/2022.naacl-main.166

Pouran Ben Veyseh, A., Nguyen, M. V., Ngo Trung, N., Min, B., & Nguyen, T. H. (2021, November). Modeling document-level context for event detection via important context selection. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 5403–5413). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.439 doi: 10.18653/v1/2021.emnlp-main.439

Pouran Ben Veyseh, A., & Nguyen, T. (2022, July). Word-label alignment for event detection: A new perspective via optimal transport. In *Proceedings of the 11th joint conference on lexical and computational semantics* (pp. 132–138). Seattle, Washington: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.starsem-1.11 doi: 10.18653/v1/2022.starsem-1.11

Qian, L., Hui, H., Hu, Y., Zhou, G., & Zhu, Q. (2014, June). 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)* (pp. 582–592). Baltimore, Maryland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P14-1055 doi: 10.3115/v1/P14-1055

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Corr.* Retrieved from https://arxiv.org/abs/2103.00020

Radford, A., & Narasimhan, K. (2018). Improving language understanding by generative pre-training.. Retrieved from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. In *Corr.* Retrieved from http://arxiv.org/abs/1910.10683

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. In *Corr.* Retrieved from https://arxiv.org/abs/2204.06125 doi: 10.48550/ARXIV.2204.06125

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... Sutskever, I. (2021). Zero-shot text-to-image generation. In *Corr.* Retrieved from https://arxiv.org/abs/2102.12092

Recasens, M., & Marti, A. (2010). Ancora-co: Coreferentially annotated corpora for spanish and catalan. In *Languages resources and evaluation.* Retrieved from https://link.springer.com/article/10.1007/s10579-009-9108-x

Ruder, S., Vulić, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *CoRR.* Retrieved from https://doi.org/10.1613/jair.1.11640 doi: 10.1613/jair.1.11640

Sha, L., Qian, F., Chang, B., & Sui, Z. (2018). Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction. In *Proceedings of the association for the advancement of artificial intelligence (aaai).*

Shah, R., Lin, B., Gershman, A., Frederking, R., & Translatortm, M. B. (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.*

Shen, S., Wu, T., Qi, G., Li, Y.-F., Haffari, G., & Bi, S. (2021, August). Adaptive knowledge-enhanced Bayesian meta-learning for few-shot event detection. In *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 2417–2429). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-acl.214 doi: 10.18653/v1/2021.findings-acl.214

Shi, T., Liu, Z., Liu, Y., & Sun, M. (2015, July). 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)* (pp. 567–572). Beijing, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P15-2093 doi: 10.3115/v1/P15-2093

Sil, A., & Florian, R. (2016, August). 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)* (pp. 2255–2264). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P16-1213 doi: 10.18653/v1/P16-1213

Singh, J., McCann, B., Keskar, N. S., Xiong, C., & Socher, R. (2019). XLDA: cross-lingual data augmentation for natural language inference and question answering. In *Corr.* Retrieved from http://arxiv.org/abs/1905.11471

Smith, S. L., Turban, D. H. P., Hamblin, S., & Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Corr.* Retrieved from http://arxiv.org/abs/1702.03859

Snell, J., Swersky, K., & Zemel, R. S. (2017). Prototypical networks for few-shot learning. In *31st conference on neural information processing systems (nips).* Retrieved from https://www.cs.toronto.edu/~zemel/documents/prototypical_networks_nips_2017.pdf

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., ... Ma, X. (2015, June). 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* (pp. 89–98). Denver, Colorado: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W15-0812 doi: 10.3115/v1/W15-0812

Soraluze, A., Arregi, O., Arregi, X., & Diaz De Ilarraza, A. (2017). Improving mention detection for basque based on a deep error analysis. In (Vol. 23, p. 351–384). Cambridge University Press. doi: 10.1017/S1351324916000206

Soraluze, A., Arregi, O., Arregi, X., Díaz de Ilarraza, A., Kabadjov, M., & Poesio, M. (2016, June). Coreference resolution for the Basque language with BART. In *Proceedings of the workshop on coreference resolution beyond OntoNotes (CORBON 2016)* (pp. 67–73). San Diego, California: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W16-0710 doi: 10.18653/v1/W16-0710

Subburathinam, A., Lu, D., Ji, H., May, J., Chang, S.-F., Sil, A., & Voss, C. (2019, November). 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)* (pp. 313–325). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1030 doi: 10.18653/v1/D19-1030

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. S., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *2018 ieee/cvf conference on computer vision and pattern recognition.* Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8578229&tag=1

Täckström, O., McDonald, R., & Uszkoreit, J. (2012, June). 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* (pp. 477–487). Montréal, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N12-1052

Tai, K. S., Socher, R., & Manning, C. D. (2015, July). 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)* (pp. 1556–1566). Beijing, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P15-1150 doi: 10.3115/v1/P15-1150

Tang, H., Ji, D., Li, C., & Zhou, Q. (2020, July). Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 6578–6588). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.588 doi: 10.18653/v1/2020.acl-main.588

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., ... Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning.. Retrieved from https://arxiv.org/abs/2008.00401

Tiedemann, J. (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. Retrieved from https://aclanthology.org/2020.wmt-1.139

Tiedemann, J., Agić, Ž., & Nivre, J. (2014, June). Treebank translation for cross-lingual parser induction. In *Proceedings of the eighteenth conference on computational natural language learning* (pp. 130–140). Ann Arbor, Michigan: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W14-1614 doi: 10.3115/v1/W14-1614

Tjong Kim Sang, E. F. (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).* Retrieved from https://aclanthology.org/W02-2024

Tjong Kim Sang, E. F., & De Meulder, F. (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* (pp. 142–147). Retrieved from https://aclanthology.org/W03-0419

Tsai, C.-T., Mayhew, S., & Roth, D. (2016, August). Cross-lingual named entity recognition via wikification. In *Proceedings of the 20th SIGNLL conference on computational natural language learning* (pp. 219–228). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://aclanthology.org/K16-1022 doi: 10.18653/v1/K16-1022

Urbizu, G., Soraluze, A., & Arregi, O. (2019, June). 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* (pp. 35–41). Minneapolis, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W19-2806 doi: 10.18653/v1/W19-2806

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., & Pustejovsky, J. (2013, June). 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)* (pp. 1–9). Atlanta, Georgia, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/S13-2001

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems (nips)*. Retrieved from http://arxiv.org/abs/1706.03762

Verga, P., Belanger, D., Strubell, E., Roth, B., & McCallum, A. (2016, June). 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* (pp. 886–896). San Diego, California: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N16-1103 doi: 10.18653/v1/N16-1103

Villani, C. (2008). *Optimal transport: Old and new*. Springer Berlin Heidelberg. Retrieved from https://books.google.com/books?id=hV8o5R7_5tkC

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching networks for one-shot learning. In *30th conference on neural information processing systems (nips)*. Retrieved from https://proceedings.neurips.cc/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. In *Commun. acm.* Retrieved from https://doi.org/10.1145/2629489 doi: 10.1145/2629489

Wadden, D., Wennberg, U., Luan, Y., & Hajishirzi, H. (2019). Entity, relation, and event extraction with contextualized span representations. 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).* Retrieved from https://aclanthology.org/D19-1585 doi: 10.18653/v1/D19-1585

Walker, C., Strassel, S., Medero, J., & Maeda, K. (2006). Ace 2005 multilingual training corpus. In *Technical report, linguistic data consortium.*

Wang, L., Cao, Z., de Melo, G., & Liu, Z. (2016, August). Relation classification via multi-level attention CNNs. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1298–1307). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P16-1123 doi: 10.18653/v1/P16-1123

Wang, X., Wang, Z., Han, X., Jiang, W., Han, R., Liu, Z., ... Zhou, J. (2020). Maven: A massive general domain event detection dataset. In *Emnlp.*

Weng, R., Yu, H., Huang, S., Cheng, S., & Luo, W. (2020). Acquiring knowledge from pre-trained model to neural machine translation. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34). Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/6465 doi: 10.1609/aaai.v34i05.6465

Wu, Q., Lin, Z., Karlsson, B., Lou, J.-G., & Huang, B. (2020, July). 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* (pp. 6505–6514). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.581 doi: 10.18653/v1/2020.acl-main.581

Wu, Q., Lin, Z., Karlsson, B. F., Huang, B., & Lou, J.-G. (2020). 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.* Retrieved from https://doi.org/10.24963/ijcai.2020/543 doi: 10.24963/ijcai.2020/543

Wu, Q., Lin, Z., Wang, G., Chen, H., Karlsson, B., Huang, B., & Lin, C.-Y. (2020). Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In *Proceedings of the aaai conference on artificial intelligence.* Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/6466/6322 doi: 10.1609/aaai.v34i05.6466

Wu, S., & Dredze, M. (2019, November). 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)* (pp. 833–844). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1077 doi: 10.18653/v1/D19-1077

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. In *Corr.* Retrieved from https://arxiv.org/abs/1609.08144

Xie, J., Yang, Z., Neubig, G., Smith, N. A., & Carbonell, J. (2018b, October-November). Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 369–379). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-1034 doi: 10.18653/v1/D18-1034

Xie, J., Yang, Z., Neubig, G., Smith, N. A., & Carbonell, J. G. (2018a). Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 conference on empirical methods in natural language processing (emnlp).*

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... Raffel, C. (2021, June). 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* (pp. 483–498). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.41 doi: 10.18653/v1/2021.naacl-main.41

Yan, H., Jin, X., Meng, X., Guo, J., & Cheng, X. (2019, November). Event detection with multi-order graph convolution and aggregated attention. 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)* (pp. 5766–5770). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1582 doi: 10.18653/v1/D19-1582

Yang, B., Cardie, C., & Frazier, P. (2015). A hierarchical distance-dependent Bayesian model for event coreference resolution. In *Transactions of the association for computational linguistics.* Retrieved from https://aclanthology.org/Q15-1037 doi: 10.1162/tacl_a_00155

Yang, B., & Mitchell, T. M. (2016). Joint extraction of events and entities within a document context. In *Proceedings of the conference of the north american chapter of the association for computational linguistics: Human language technologies (naacl-hlt).*

Yang, S., Feng, D., Qiao, L., Kan, Z., & Li, D. (2019a). Exploring pre-trained language models for event extraction and generation. In *Proceedings of the annual meeting of the association for computational linguistics (acl).*

Yang, S., Feng, D., Qiao, L., Kan, Z., & Li, D. (2019b, July). Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5284–5294). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1522 doi: 10.18653/v1/P19-1522

Yarmohammadi, M., Wu, S., Marone, M., Xu, H., Ebner, S., Qin, G., ... Durme, B. V. (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).* Retrieved from https://arxiv.org/pdf/2109.06798.pdf

Yarowsky, D., Ngai, G., & Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on human language technology research.* Retrieved from https://aclanthology.org/H01-1035

Zelenko, D., Aone, C., & Richardella, A. (2003). Kernel methods for relation extraction. In *J. mach. learn. res.* Retrieved from https://dl.acm.org/doi/10.5555/944919.944964

Zeman, D., & Resnik, P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 workshop on NLP for less privileged languages.* Retrieved from https://aclanthology.org/I08-3008

Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J. (2014, August). Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 2335–2344). Dublin, Ireland: Dublin City University and Association for Computational Linguistics. Retrieved from https://aclanthology.org/C14-1220

Zhang, C., Li, Q., & Song, D. (2019, November). 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)* (pp. 4568–4578). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1464 doi: 10.18653/v1/D19-1464

Zhang, J., Qin, Y., Zhang, Y., Liu, M., & Ji, D. (2019). Extracting entities and events as a single task using a transition-based neural model. In *Ijcai.*

Zhang, T., Ji, H., & Sil, A. (2019). Joint Entity and Event Extraction with Generative Adversarial Imitation Learning. *Data Intelligence*, *1*(2), 99-120. Retrieved from https://doi.org/10.1162/dint_a_00014 doi: 10.1162/dint_a_00014

Zhang, Y., Xu, G., Wang, Y., Lin, D., Li, F., Wu, C., ... Huang, T. (2020). A question answering-based framework for one-step event argument extraction. In *Ieee access, vol 8, 65420-65431.* Retrieved from https://ieeexplore.ieee.org/document/9055029

Zou, B., Xu, Z., Hong, Y., & Zhou, G. (2018, August). Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of the 27th international conference on computational linguistics* (pp. 437–448). Santa Fe, New Mexico, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/C18-1037