Event Extraction in Basque: Typologically motivated Cross-Lingual Transfer-Learning Analysis

Mikel Zubillaga Oscar Sainz Ainara Estarrona
Oier Lopez de Lacalle Eneko Agirre

HiTZ Basque Center for Language Technology - Ixa NLP Group
University of the Basque Country UPV/EHU
{mikel.zubillaga, oscar.sainz, ainara.estarrona, oier.lopezdelacalle, e.agirre}@ehu.eus

Abstract

Cross-lingual transfer-learning is widely used in Event Extraction for low-resource languages and involves a Multilingual Language Model that is trained in a source language and applied to the target language. This paper studies whether the typological similarity between source and target languages impacts the performance of cross-lingual transfer, an under-explored topic. We first focus on Basque as the target language, which is an ideal target language because it is typologically different from surrounding languages. Our experiments on three Event Extraction tasks show that the shared linguistic characteristic between source and target languages does have an impact on transfer quality. Further analysis of 72 language pairs reveals that for tasks that involve token classification such as entity and event trigger identification, common writing script and morphological features produce higher quality cross-lingual transfer. In contrast, for tasks involving structural prediction like argument extraction, common word order is the most relevant feature. In addition, we show that when increasing the training size, not all the languages scale in the same way in the cross-lingual setting. To perform the experiments we introduce EusIE, an event extraction dataset for Basque, which follows the Multilingual Event Extraction dataset (MEE). The dataset and code are publicly available.

Keywords: Event Extraction, Cross-lingual Transfer-Learning, Basque language, Typology based-analysis

1. Introduction

Event Extraction (EE) is one of the fundamental tasks in Information Extraction (IE) and aims to extract event mentions and their arguments (i.e., participants) from text. Typically, EE involves the identification of trigger words (e.g. *married*, *the attack*) that denote a mention of an event or action. In parallel, the entities in the sentence are extracted. In a final step, once the event is known, the participants that take part in the event are identified in the context. Table 1 shows a complete example of an event extraction pipeline, in which we show the entities, event triggers, and the arguments of the corresponding event mention. Due to the complexity involved, and the high interest in the task, EE has been historically one of the most relevant tasks in the field of Information Extraction.

Information Extraction tasks in general and EE in particular pose significant challenges (Grishman, 2019), as it is a complex task that demands humans to meticulously follow complicated guidelines, often riddled with numerous exceptions. To tackle this challenge, the conventional approach is to train computational models with large amounts of annotated examples. Obtaining the examples entails extensive manual effort by domain experts, making it impractical for situations with limited resources, especially for low-resource languages.

Due to the recent advances in Natural Language Processing (Min et al., 2023), Large Language Models (LLMs) are capable of transferring knowledge across languages, i.e. training in one language and performing inferences in another. This is referred to as Cross-Lingual Transfer Learning and has represented a significant advancement for languages other than English, as it allows to obtain EE models using data from high-resource languages (primarily English). The success of this approach allowed to develop ambitious programs, such as BETTER (Mckinnon and Rubino, 2022), where English data is provided for training while the models are tested on other languages. While the proposed task was interesting and really challenging, it only uses English data for training. In recent work, Pouran Ben Veyseh et al. (2022) developed MEE (Multilingual Event Extraction), an extension of the well-known ACE 2005 (Christopher Walker and Stephanie Strassel and Julie Medero and Kazuaki Maeda, 2006) dataset to 8 languages.

One limitation of current EE research is the under-exploration of non-English languages, due mainly to the lack of high-quality multilingual dataset. MEE allows for such kind of research, and we thus use MEE to explore whether the typology of target and source languages impacts cross-lingual transfer capabilities. In particular, we study what would be the best transfer choice to develop

LREC-COLING 2024, pages 6607–6621
20-25 May, 2024.

Sentence	Peter, the CEO of XYZ company, got married in Brazil.
Entities	[Peter]$_{PER}$, the [CEO]$_{JOB}$ of [XYZ]$_{ORG}$ company, got married in [Brazil]$_{LOC}$.
Events	Peter, the CEO of XYZ company, [got married]$_{Life.Marry}$ in Brazil.
Arguments	[Peter]$_{Person\text{-}arg}$, the CEO of XYZ company, $$$got married$$$ in [Brazil]$_{Place\text{-}arg}$.

Table 1: A sample sentence annotated according to the three tasks related to Event Extraction.

an IE system for a language with no training data.

In order to increase the typological diversity of languages in MEE, we added Basque, a language isolate with no known related language. The Basque language has a particularly interesting set of features, very different from the surrounding languages, making it an interesting candidate to study which features of languages affect the quality of cross-lingual transfer. We thus annotated an evaluation set for Basque, **the first Event Extraction benchmark for the language**[1]. We follow the same procedure as in (Pouran Ben Veyseh et al., 2022) when collecting and annotating examples, with the exception that we used expert annotators in contrast to crowd-working services. Due to the high cost, we only annotated evaluation data.

In our experiments, we first explored which linguistic characteristics affect positively and negatively when we evaluate in Basque. We thus trained several models on each MEE language under the same conditions and evaluated them in Basque. Further, we extend this analysis to all the languages and measure systematically how the language typology affects cross-lingual transfer. The code of the experiments is publicly available[2] The results show that the transfer quality depends on the shared linguistic characteristic between source and target language, but varies across each of the tasks. Further analysis reveals that for tasks involving token classification (i.e. entity and trigger identification) sharing writing script shows higher cross-lingual transfer benefit. In contrast, when structural understanding is involved (e.g. argument extraction) word order matters the most.

2. Related Work

Event Extraction. Early methods addressed the task by defining human-crafted features and applying rules (Ji and Grishman, 2008; Gupta and Ji, 2009; Hong et al., 2011; Li et al., 2013). These methods were replaced by deep learning approaches (Chen et al., 2015; Feng et al., 2016; Liu et al., 2018) in the last decade. Soon, sequence labeling became the standard approach for EE (Nguyen et al., 2016; Chen et al., 2018; Araki and Mitamura, 2018; Ding et al., 2019; Lin et al., 2020; Guzman-Nateras et al., 2022). With the development of pre-trained Large Language Models, several works reformulated the task into language understanding tasks such as Question Answering (Du and Cardie, 2020; Li et al., 2020; Liu et al., 2020; Wei et al., 2021; Sheng et al., 2021; Zhou et al., 2022) and Textual Entailment (Sainz et al., 2022a,b) to benefit from the implicit knowledge and capabilities encoded the model. Recently, with the increasing popularity of generative models, works based on conditional generation have also been proposed (Xiangyu et al., 2021; Lu et al., 2021; Li et al., 2021b; Hsu et al., 2022; Zeng et al., 2022; Li et al., 2022; Liu et al., 2022; Huang et al., 2022; Du et al., 2022). Lastly, multi-task instruction-based models have been applied to perform several tasks together, including event extraction (Wang et al., 2023; Sainz et al., 2023).

Cross-lingual approaches for IE Pre-trained LLMs allowed a simplified approach to cross-lingual IE with state-of-the-art performance (Conneau et al., 2019; Lample and Conneau, 2019; Conneau et al., 2020). Previously, state-of-the-art consisted of using parallel data to project labels from one language to the other (Agerri et al., 2018). Related to this, the improvement of machine translation and alignment models allowed effective augmentation of training examples (García-Ferrero et al., 2022; Li et al., 2021a; Lou et al., 2022). In this paper, we choose to use a multilingual sequence labeling approach to efficiently analyze the characteristics of cross-lingual transfer learning.

Existing datasets for Event Extraction are mostly available only for English, such as CySecED (Man Duc Trong et al., 2020), CASIE (Satyapanich et al., 2020), LitBank (Sims et al., 2019), MAVEN (Wang et al., 2020), RAMS (Ebner et al., 2020) and, WikiEvents (Li et al., 2021b) among others. Additionally, there are a few multilingual EE datasets like ACE 2005 (Christopher Walker and Stephanie Strassel and Julie Medero and Kazuaki Maeda, 2006) and more recently BETTER (Mckinnon and Rubino, 2022) and MEE (Pouran Ben Veyseh et al., 2022). ACE 2005 and BETTER include only English training data. MEE contains annotated train and

[1] https://huggingface.co/datasets/HiTZ/EusIE
[2] https://github.com/MikelZubi/GrAL

Entities	PER, ORG, GPE, LOC, FAC, VEH, WEA, CRIME, TIME, MON, POS, *OBJ*	
Events	Life:Be-Born	Person, Time, Place
	Life:Marry	Person, Time, Place
	Life:Divorce	Person, Time, Place
	Life:Injure	Agent, Victim, Instrument, Time, Place
	Life:Die	Agent, Victim, Instrument, Time, Place
	Movement:Transport	Agent, Artifact, Vehicle, Price, Origin, Destination, Time
	Transaction:Transfer-Ownership	Buyer, Seller, Beneficiary, Price, Artifact, Time, Place
	Transaction:Transfer-Money	Giver, Recipient, Beneficiary, Money, Time, Place
	Business:Start-Organization	Agent, Organization, Time, Place
	Conflict:Attack	Attacker, Target, Instrument, Time, Place
	Conflict:Demonstrate	Entity, Time, Place
	Contact:Meet	Entity, Time, Place
	Contact:Phone-Write	Entity, Time
	Personnel:Start-Position	Person, Entity, Position, Time, Place
	Personnel:End-Position	Person, Entity, Position, Time, Place
	Justice:Arrest-Jail	Person, Agent, Crime, Time, Place

Table 2: Annotation schema used to annotate EusIE. The schema is the same as the one used by MEE and is based on ACE 2005. Except for the label *OBJ* that does not exist on MEE, and therefore, it is not used for evaluation.

evaluation datasets in eight languages. In this work, we follow ACE 2005 and MEE guidelines, and annotate a Basque Event Extraction dataset to perform our experiments.

3. EusIE: Basque Event Extraction

In this section, we present EusIE (**Eus**karazko **I**nformazio-**E**rauzketa)[3], which is the first EE dataset for Basque. We decided to extend the Multilingual Event Extraction (MEE) dataset ((Pouran Ben Veyseh et al., 2022)) by following the well-known ACE05 (Christopher Walker and Stephanie Strassel and Julie Medero and Kazuaki Maeda, 2006) ontology. The MEE dataset covers 8 diverse languages that we use in our experiments in conjunction with Basque.

Although the dataset creation process followed similar steps to MEE, few modifications were implemented. On one hand, due to the difficulty of finding Basque-speaking crowd workers, two native experts annotated the dataset. On the other hand, due to our small budget, we limit the annotation to the development and test splits. This way we provide quality over quantity. In the following sections, we describe the data collection, filtering, and annotation process.

3.1. Data collection

We collect the initial set of documents from a snapshot of Basque Wikipedia[4]. From the initial set, we select the documents related to events (*Gertaerak* category) that were labeled as part of the following topics: Economy (*Ekonomia*), Politics (*Politika*), Technology (*Teknologia*), Natural Disasters (*Hondamen Naturalak*), Military (*Militarrak*) and Crimes (*Krimenak*). We keep the same topics as the original MEE to avoid domain shifts.

After collecting the documents, we removed the markup from the documents using WikiExtractor (Attardi, 2015). Additionally, section titles and other structural information were removed too. We split the documents into sentences, and, we tokenized them using IXA-pipes (Agerri et al., 2014) an NLP toolkit designed for Basque. Similar to MEE and RAMS (Ebner et al., 2020), we grouped 5 sentences to form an annotation **segment**. The segment is our annotation boundary, and thus the relations between the events and arguments can occur within the sentence as well as cross sentences, but always inside the segments.

3.2. Annotation schema

The annotation schema used to annotate the dataset is shown in Table 2. We adapted the schema used in MEE to include entity types that could potentially be argument candidates for the defined events. We included `OBJ` to categorize entities that are candidates for the `Artifact` role. We do not have examples for the `NUM` entity type, as all numerical mentions could be labeled either with `DATE` or `MON`.

[3] *Basque Information-Extraction* in Basque.

[4] The downloaded snapshot was from October 10th, 2022. https://dumps.wikimedia.org/other/cirrussearch/current/

Language	Tokens	Entities	Events	Arguments
English	123	14.66	1.36	1.04
Spanish	112	14.69	1.85	0.24
Portuguese	102	16.98	1.30	8.21
Polish	108	14.06	2.42	0.76
Turkish	117	8.59	1.87	0.31
Hindi	98	12.53	1.21	1.41
Japanese	99	12.78	1.44	2.27
Korean	103	8.34	0.75	1.16
Mean	108	12.83	1.53	1.93
Basque	94	16.58	2.17	4.49

Table 3: Average statistic per segment for each language.

3.3. Annotation

We annotated a total of 300 segments (1500 sentences) and divided them into 150 for development and the rest of 150 for testing. That is, we annotated a similar amount of segments provided in the evaluation partition of the MEE dataset. The annotator was provided with the Inception (Klie et al., 2018) annotation tool and the ACE 2005 guidelines[5]. Statistics of the annotation process are shown in Table 3. Overall, the annotated segments contain substantially more annotations than the average.

To ensure annotation quality, we asked a second expert to annotate a portion of the data, 35 segments and computed the Cohen's κ between both annotators. For span annotations which include entities and event triggers, the annotators obtained an agreement of 0.94, and, for the arguments, the annotators obtained an agreement of 0.92. The obtained agreement is indicative of the quality of the expert annotators and the annotation guidelines.

The annotated data was converted from WebAnno 3.0 to a JSON format introduced by Lin et al. (2020) for simplicity. Once converted, the data was split into development and test ensuring that segments from the same original document remain on the same partition.

4. Experimental Setup

In this work, we explore the cross-lingual capabilities of the multilingual Language Models in EE for the Basque language. We deploy the aforementioned MEE and EusIE datasets. Typically, EE is a sequence of three tasks that are evaluated as a pipeline, reporting the final F1 results. As we want to compare the transfer qualities of each language empirically in each of the pipelined steps, we reported the F1 scores for each task (entity, event, and argument extraction). All the tasks are evaluated independently using gold annotations from the previous step in the pipeline [6]. Additionally, three different runs are executed for each configuration in order to provide average and deviation scores.

We organize the experiments in three main parts. First, we try to replicate the in-language experiments performed in (Pouran Ben Veyseh et al., 2022), where models are evaluated and trained in the same language. We use their baseline approach (called *Pipeline* in the paper), trained and evaluated in the original data splits provided by the authors. Next, we run the cross-lingual experiments, where we train models in one language and evaluate them in Basque (EusIE). For fair comparisons across languages, all languages have the same number of train examples, i.e. we discard examples in the languages with most training data. More details in Section 4.2. Finally, we run an analysis of linguistic features to gain insight into what makes a language good for cross-lingual transfer-learning. We will categorize each language based on its typological features, and perform all-vs-all experiments to analyze the impact of those features.

4.1. Model

As mentioned above, EE is typically divided into entity detection, event detection, and, event argument extraction. Therefore, we train three models, one per each task. As shown in Table 1, we formulated all the tasks as sequence labeling problems. Both, entity and event detection tasks are simply formulated as predicting the label for each token in the input text. For the event argument-extraction task, however, the output must be conditioned on the event to analyze. As we indicate in the Table 1, we surround the event trigger with markers, "$$$" in our case, and label only the corresponding arguments. The backbone language model is the base version of XLM-RoBERTa (Conneau et al., 2019).

Regarding the hyperparameters, we set their values based on a few preliminary experiments. The overall best performing hyperparameters were a learning-rate of $5e^{-5}$, 32 for the batch-size and a weight decay of $1e^{-3}$. We run the models for 64 epochs, as we found out that the F1 score was increasing in the development even if the loss was increasing too.

[5] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf

[6] E.g. we used gold event triggers when detecting the arguments.

Languages	Entities	Events	Arguments
English	80.48±0.19	**78.47**±0.33	63.60±0.09
Spanish	**84.56**±0.38	63.86±1.20	40.45±1.91
Portuguese	80.42±0.32	61.79±0.63	68.18±0.99
Polish	81.00±0.40	69.09±0.91	**76.25**±1.26
Turkish	70.83±0.06	56.62±0.43	24.08±1.66
Hindi	76.00±0.41	48.21±2.54	45.31±1.55
Japanese	47.43±0.20	35.74±1.92	52.93±1.26
Korean	71.31±0.78	45.30±0.49	34.71±3.06
All	78.63±0.17	68.01±0.27	59.74±0.99

Table 4: Results obtained by our model for each task and language. *All* reports results obtained after training and testing with all languages together.

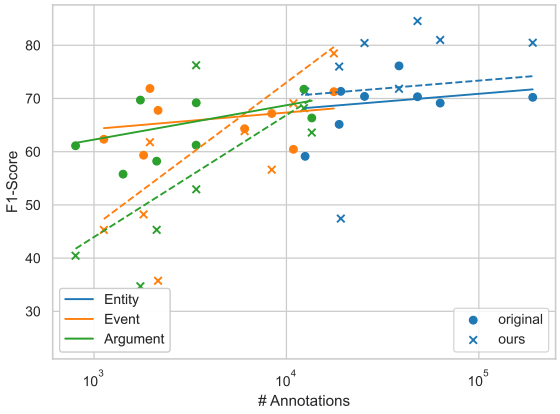

Figure 1: Comparison between our model and the original reported results based on the number of total annotations and F1-Score. Dashed lines show the linear relationship of our system and the number of annotation in training. The continuous line refers to (Pouran Ben Veyseh et al., 2022).

4.2. Comparable Training Size

In order to compare the cross-lingual transfer capability of the languages in a comparable manner (results reported in Section 5.2), we try to control the size of the training data. We thus equalize the amount of training data for all the languages. That is, we remove training examples from larger languages until all the languages contain the same amount of annotations. Note that we performed the under-sampling by counting annotations and not the number of segments, as the latter could lead to a different number of annotations per language. We also add examples from development and testing in order to increase the size of the training set (note that this training is uniquely used in the cross-lingual experiments where models are evaluated in Basque). As a result, for each task, each language's data was reduced to the amount of data for the language with the least amount of annotations: 12508 annotated entities, 1125 event mentions, and 1416 arguments.

5. Results

In this section, we discuss the results obtained in the experiments. First, similar to the MEE authors, we report the results using the original splits. Second, we report the results obtained on the EusIE benchmark. Finally, we explore the effect of scaling training data.

5.1. Result on In-language Scenario

Table 4 shows the F1 scores obtained in each language. Additionally, we included the *All* row, which represents a model trained and tested using all the languages available in the dataset. We repeated each experiment 3 times and reported the average F1 score and the standard deviation for each setting. Language-wise comparisons do not show any clear pattern in which we can distinguish a particular language that overperforms the rest of the languages across all the tasks. A language that performs strongly in a specific task, shows poor performance in the other task (e.g. Spanish shows an outstanding 84.5 of F1 in entity detection, whereas it is far from top results in Argument identification).

It is important to note that our results substantially deviate from the ones reported in the original paper (Pouran Ben Veyseh et al., 2022)[7]. Discussion with the authors did not reveal any reason for this difference apart from the use of a different model. However, we found that the results obtained by our system correlate better with the number of annotations in the training set, as shown in Figure 1. The linear correlation of our model is plotted with dashed lines, the original ones with continuous lines. In particular, for event and argument detection, in which the number of annotations is much smaller compared to entity detection, our system linearly improved when we increased the number of annotations. In the case of entities, as we have more annotations, it does not show a significant positive relationship with the number of annotations (both systems show similar behavior in this case).

5.2. Results on EusIE

Table 5 shows the results for the models trained on each language when evaluated in Basque. Note that, in this experiment, all the languages have the same amount of training examples (cf. Section 4.2). The best-performing language varies across tasks. We had hypothesized that the best

[7] See Table 7 in Appendix A for comparing the results with the original ones

Languages	Entities	Events	Arguments
English	**59.65**±1.19	42.65±6.57	13.93±2.06
Spanish	56.56±0.45	43.71±2.63	2.88±0.79
Portuguese	59.62±1.67	24.30±2.14	14.02±0.96
Polish	59.48±1.35	**46.37**±1.94	10.28±0.29
Turkish	55.72±2.49	44.46±2.73	14.84±3.82
Hindi	56.97±1.44	34.70±4.71	10.62±0.70
Japanese	47.17±1.92	5.7±4.03	10.96±0.91
Korean	46.67±0.92	21.56±7.62	**15.03**±0.81
All	56.92±1.12	55.58±0.80	28.05±1.52

Table 5: Results on the EusIE dataset.

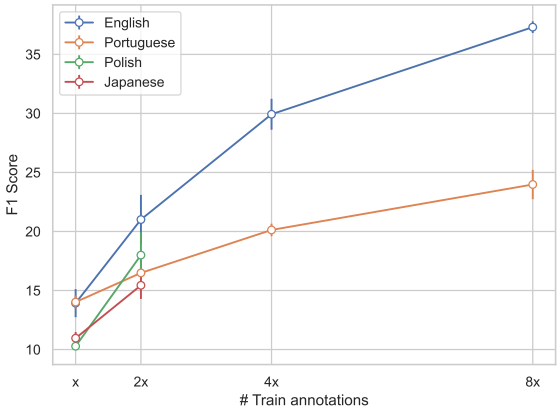

Figure 2: Train data scaling experiments on the event argument extraction task. $x = 1,416$ instances.

results would be for Spanish, as it is an official language in the Basque Country and the contact between the two languages has been happening since Spanish became a language on its own, but that is not the case.

Regarding entity detection, results and the standard deviation show that English, Portuguese, and Polish obtain very similar results and outperform the rest of the languages, whereas Spanish, Turkish, and Hindi are close to the best results. The results obtained with Japanese and Korean, lag significantly behind the rest by a large margin.

We observe a similar pattern for event detection but with larger differences between languages. The best-performing language, in this case, is Polish followed by Turkish, Spanish, and English. The rest of the languages fall behind, Japanese in particular. It is important to note that the standard deviations are very high for some of the languages.

The results for argument extraction are significantly lower than for the other two tasks. The task is harder to transfer from language to language and might be severely affected by the undersampling of the training set. Language-wise, the best-performing language is Korean, followed very closely by Turkish. English and Portuguese are also significantly better than the average. Spanish is a special case due to its marginal amount of training data. We had to consider the next smaller language because Spanish was too small for argument extraction.

Finally, we can see that the *All* model outperforms the rest of the languages on event detection and event argument extraction, but not on entity detection. We hypothesize that adding several languages is helpful when training datasets are small, but when there is enough data, it introduces noise.

5.3. Data scaling results

We showed that the results on the event argument extraction are significantly lower. We hypothesized that this is due to the insufficient amount of the training data to learn properly the task. The fact that using all languages for training nearly doubles the best results for argument extraction in Table 5 is some evidence in this direction.

To study the effects of the training size in the cross-lingual setting, we trained new models with different amounts of data as shown in Figure 2. Here, the x denotes the initial training amount (1,416 instances), the same as used in Table 5. We scaled the x by 2, 4, and, 8 when possible.

Results in Figure 2 confirm our hypothesis: argument extraction is more challenging and requires more data to properly model the task. The figure shows how fast languages scale the performance with more data. Despite using fewer languages, we can see that languages scale at a different pace with the training data. For example, a comparison of English and Portuguese reveals that while both perform very similarly on the initial values, Portuguese scales much slower in the long term.

Significantly English x8 outperforms the results attained with *All* (in Table 5) by a large margin. Note that the amount of training data used is the same as the *All* results but using only monolingual data. Together with the results for entities shown in the previous section, we can conclude that training size is a relevant factor in cross-lingual transfer and that mixing the data from all languages is beneficial only for smaller training sizes.

6. Analysis according to Language Typology

From the results on EusIE, we can draw two main conclusions: (1) There is no dominant language across tasks, and, (2) A language that is effective in a particular cross-lingual task does not guarantee that it will be good for another. For further understanding, we run an analysis on the following hypotheses:

Language	Morphology	Morphosyntactic Align.	Word Order	Script	Geographical Location
English	Fusional	Nominative-Accusative	SVO	Latin	West Europe*
Spanish	Fusional	Nominative-Accusative	SVO	Latin	West Europe*
Portuguese	Fusional	Nominative-Accusative	SVO	Latin	West Europe*
Polish	Fusional	Nominative-Accusative	SVO	Latin	East Europe
Turkish	Agglutinative	Nominative-Accusative	SOV	Latin	East Europe
Hindi	Fusional	Split Ergative	SOV	Devanagari	India
Japanese	Agglutinative	Nominative-Accusative	SOV	Kanji & Kana	East Asia
Korean	Agglutinative	Nominative-Accusative	SOV	Hangul	East Asia
Basque	Agglutinative	Ergative-Absolutive	SOV*	Latin	West Europe

Table 6: Language typology features. * indicates that the values are simplified, see main text for details.

1. Similar languages should benefit more from cross-lingual transfer.

2. Different tasks require different skills. The tasks of detecting entities and events require more lexical knowledge, whereas extracting arguments requires a more structural understanding of the text.

Focusing only on Basque as the target language would be limiting. So we decided to carry out the same experiments we did for Basque, but in this case, running the cross-lingual experiments for all the possible language pairs. We exclude the combinations that include the same source and target languages.

6.1. Language categorization

To validate our hypotheses, we first defined a set of linguistic typology features that help categorize the languages in our dataset. We selected the features that could be relevant for cross-lingual transfer. Table 6 summarizes our categorization[8].

Morphology refers to the study of words and how they are formed. We categorized each language into *Agglutinative* or *Fusional* categories. Our initial guess is that languages with similar morphology should perform better on tasks requiring more lexical knowledge (Entities and Events).

Morphosyntactic alignment refers to the study of the relationship between different arguments of verbs. We categorized the languages into *Nominative-Accusative*, *Ergative-Absolutive*, and *Split-Ergative*. As it is directly related to how the event arguments are marked in the sentence, we guess that languages with similar morphosyntactic alignment will perform better. Note that most languages are categorized as *Nominative-Accusative*, making it difficult to measure the effects of this feature.

[8]The categorization was based on Wikipedia.

Word order refers to the order of the syntactic constituents of a language. We categorized the languages into *Subject-Object-Verb* (SOV) or *Subject-Verb-Object* (SVO) categories. An important consideration is that Basque usually follows the SOV order (de Rijk, 1969), however, but also allows other orders depending on pragmatic factors (Laka, 1996). The word order has a significant impact when defining the different roles each part of the sentence has with respect to the verb, and therefore to an event. We guess that it would positively affect the event argument extraction task.

Script refers to how the language is written. This is particularly important because it affects directly the tokenizer of the model, and therefore, how the token is defined and how the information is stored in the model we define very broad categories, in which we distinguish languages into two main groups: *Latin* based and *non-Latin* based. Our guess is that the script should affect all the tasks, as it is an important feature that impacts directly the model architecture. Although, with the same script, words from different languages might be better or worse represented depending on how the tokenizer was constructed.

Geographical location refers to where the language is spoken. It is impossible to determine a concrete geographical location for a language, as nowadays a lot of languages are spoken all around the world, particularly English, Spanish, and Portuguese. However, for simplification purposes, we will consider them as languages spoken mainly in the west of Europe. The location of a language has a great cultural impact, and therefore, different entities appear more frequently in text on one language than another. This feature, however, is highly correlated with the script feature, as geographically adjacent languages tend to have similar scripts. We think this feature may have a greater impact on lexical tasks.

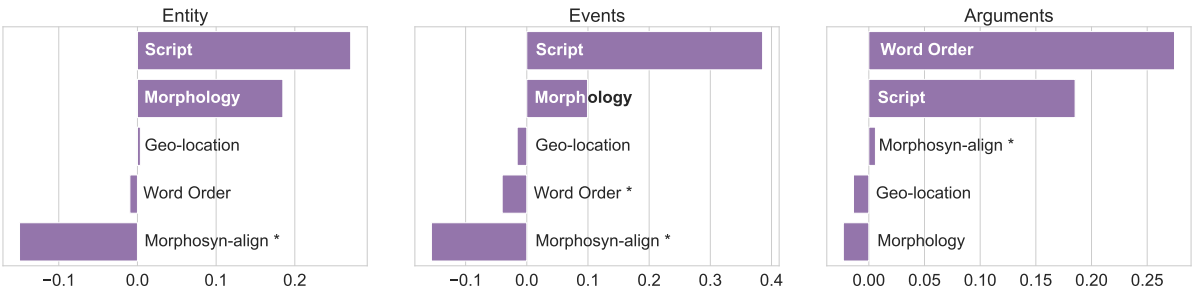

Figure 3: Estimated feature impact for each task. * indicates the result is not statistically significant (`p_value >= 0.01`).

6.2. Results of the analysis

To analyze the effect of each typological feature, we ran the same cross-lingual experiments as we did with Basque, but in this case, running all possible source-target language combinations. The results are shown in Table 8 (in Appendix A). Based on these results we run multiple regression analysis for each task separately and use the resulting coefficients to measure the relative contribution of each linguistic feature described above.

We prepare the data in order to correctly perform the analysis. First, we normalized the results across target languages as we noticed that the values on each target language ranged very differently (even across the same task). That is, we transform using Min-Max scaling the F1 scores of each task and target language into values between 0 and 1. Regarding input variables, we generate features that indicate whether the value of the given source f_i^{source} and target f_i^{target} languages are the same:

$$f_i = \begin{cases} 1 & f_i^{source} = f_i^{target} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

We formalize the linear regression analysis as shown in Equation 2, where s is the normalized F1 score for each language and task, and w_i is the coefficient that measures the relative contribution of feature i. As additional information, Figure 4 (in Appendix A) shows the distribution of normalized F1 scores for each linguistic feature in isolation.

$$s = w_o + \sum_{i}^{|F|} w_i \cdot f_i \quad (2)$$

Figure 3 shows the contribution of each linguistic feature to performance in each task. *Script* turns out to be the most relevant feature across tasks. *Morphology* is an important feature in entity and event detection, where lexical information could play a significant role, as we hypothesized. The importance of these two features might be due to the fact that languages with the same script and morphology share more tokens in the language model vocabulary. On the other hand, *Word Order*, as we hypothesized, affects significantly the argument extraction task. Surprisingly, *Geo-location* is not relevant when transferring knowledge from one language to another, it is well-known that geographically close languages share many lexical entries. Therefore, the relevance of geo-location is low probably because most of the correlation is explained with *Script*. As expected we did not find any correlation for *Morphosyntactic alignment*, probably because all languages but one shared the same feature.

All in all the analysis shows that sharing the script is a key factor in all three tasks and that sharing the script might overshadow the relevance of sharing the geographical location. As we hypothesized, morphology is relevant for the two tasks which are more lexical (entity and event extraction), while word order turns out to be relevant for the more syntactic argument extraction.

7. Conclusions

In this paper we explore the contribution of different training languages when transferring into other languages, presenting a set of exhaustive experiments on three Event Extraction tasks and eight languages with different language typological features. In a first experiment with Basque as a target, we see that there is no clear pattern. In a subsequent experiment, we performed a typologically motivated correlation analysis over all the language combinations and concluded that transfer quality does correlate with some linguistic features, which change depending on the task. For entity and trigger identification sharing the script and morphological typology between source and target languages are the two most relevant features. In argument extraction sharing word order and script play the most relevant roles. In addition, we show that source languages scale differently as we increase training data. Finally, we present the first Basque Event Extraction evalu-

ation benchmark, which was used alongside the MEE dataset Pouran Ben Veyseh et al. (2022) in all experiments.

For the future, we would like to extend the analysis to more tasks and languages, as well as taking into account other alternative typological features. A better understanding of the interactions between typology and cross-lingual transfer opens a new research avenue that can be beneficial for low-resource languages.

Acknowledgements

This work has been partially supported by the Basque Government (Research group funding IT-1805-22, IKER-GAITU and ICL4LANG Grant no. KK-2023/00094). We are also thankful to a MCIN/AEI/10.13039/501100011033 DeepKnowledge (PID2021-127777OB-C21) and FEDER, EU. This work has been also funded by The EFA104/01 LINGUATEC-IA project "Cross-border network of technological cooperation in artificial intelligence applied to language for the construction of a trans-Pyrenean linguistic infrastructure" is 65% co-financed by the European Regional Development Fund (ERDF) through the Interreg V-A Program Spain-France-Andorra (POCTEFA 2021-2027).

Bibliographical References

Rodrigo Agerri, Josu Bermudez, and German Rigau. 2014. IXA pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3823–3828, Reykjavik, Iceland. European Language Resources Association (ELRA).

Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, and German Rigau. 2018. Building named entity recognition taggers via parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jun Araki and Teruko Mitamura. 2018. Open-domain event detection using distant supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 878–891, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Giusepppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.

Yubo Chen, Hang Yang, Kang Liu, Jun Zhao, and Yantao Jia. 2018. Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1267–1276, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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

Rudolf P.G. de Rijk. 1969. Is basque an s. o. v. language? *Fontes Linguae Vasconum*, 1(3):319–352.

Ning Ding, Ziran Li, Zhiyuan Liu, Haitao Zheng, and Zibo Lin. 2019. Event detection with trigger-aware lattice neural network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 347–356, Hong Kong, China. Association for Computational Linguistics.

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

Xinya Du, Sha Li, and Heng Ji. 2022. Dynamic global memory for document-level argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguis-*

tics (Volume 1: Long Papers), pages 5264–5275, Dublin, Ireland. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Xiaocheng Feng, Lifu Huang, Duyu Tang, Heng Ji, Bing Qin, and Ting Liu. 2016. A language-independent neural network for event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 66–71, Berlin, Germany. Association for Computational Linguistics.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ralph Grishman. 2019. Twenty-five years of information extraction. *Natural Language Engineering*, 25(6):677–692.

Prashant Gupta and Heng Ji. 2009. Predicting unknown time arguments based on cross-event propagation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 369–372, Suntec, Singapore. Association for Computational Linguistics.

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

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1127–1136, Portland, Oregon, USA. Association for Computational Linguistics.

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

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

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, pages 254–262, Columbus, Ohio. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Itziar Laka. 1996. *A brief grammar of Euskara, the Basque language*. Euskal Herriko Unibertsitatea, Leioa-Donostia.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

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

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838, Online. Association for Computational Linguistics.

Haochen Li, Tong Mo, Hongcheng Fan, Jingkun Wang, Jiaxi Wang, Fuhao Zhang, and Weiping Li. 2022. KiPT: Knowledge-injected prompt tuning for event detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1943–1952, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, Sofia, Bulgaria. Association for Computational Linguistics.

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

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

Xiao Liu, Heyan Huang, Ge Shi, and Bo Wang. 2022. Dynamic prefix-tuning for generative template-based event extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5216–5228, Dublin, Ireland. Association for Computational Linguistics.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256, Brussels, Belgium. Association for Computational Linguistics.

Chenwei Lou, Jun Gao, Changlong Yu, Wei Wang, Huan Zhao, Weiwei Tu, and Ruifeng Xu. 2022. Translation-based implicit annotation projection for zero-shot cross-lingual event argument extraction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2076–2081, New York, NY, USA. Association for Computing Machinery.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

Hieu Man Duc Trong, Duc Trong Le, Amir Pouran Ben Veyseh, Thuat Nguyen, and Thien Huu Nguyen. 2020. Introducing a new dataset for event detection in cybersecurity texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5381–5390, Online. Association for Computational Linguistics.

Timothy Mckinnon and Carl Rubino. 2022. The IARPA BETTER program abstract task four new semantically annotated corpora from IARPA's BETTER program. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3595–3600, Marseille, France. European Language Resources Association.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2).

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.

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

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. 2022a.

Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States. Association for Computational Linguistics.

Oscar Sainz, Haoling Qiu, Oier Lopez de Lacalle, Eneko Agirre, and Bonan Min. 2022b. ZS4IE: A toolkit for zero-shot information extraction with simple verbalizations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 27–38, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. Casie: Extracting cybersecurity event information from text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8749–8757.

Jiawei Sheng, Shu Guo, Bowen Yu, Qian Li, Yiming Hei, Lihong Wang, Tingwen Liu, and Hongbo Xu. 2021. CasEE: A joint learning framework with cascade decoding for overlapping event extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 164–174, Online. Association for Computational Linguistics.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *CoRR*, abs/2304.08085.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Kaiwen Wei, Xian Sun, Zequn Zhang, Jingyuan Zhang, Guo Zhi, and Li Jin. 2021. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4672–4682, Online. Association for Computational Linguistics.

Xi Xiangyu, Wei Ye, Shikun Zhang, Quanxiu Wang, Huixing Jiang, and Wei Wu. 2021. Capturing event argument interaction via a bi-directional entity-level recurrent decoder. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 210–219, Online. Association for Computational Linguistics.

Qi Zeng, Qiusi Zhan, and Heng Ji. 2022. EA^{2}E: Improving consistency with event awareness for document-level argument extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2649–2655, Seattle, United States. Association for Computational Linguistics.

Jie Zhou, Qi Zhang, Qin Chen, Qi Zhang, Liang He, and Xuanjing Huang. 2022. A multi-format transfer learning model for event argument extraction via variational information bottleneck. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1990–2000, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Language Resource References

Christopher Walker and Stephanie Strassel and Julie Medero and Kazuaki Maeda. 2006. *ACE 2005 Multilingual Training Corpus*. Web Download. Philadelphia: Linguistic Data Consortium, ISLRN 458-031-085-383-4.

A. Additional results

Languages	Ours Entities	Ours Events	Ours Arguments	MEE Entities	MEE Events	MEE Arguments
English	80.48±0.19	78.47±0.33	63.60±0.09	70.22	71.28	66.34
Spanish	84.56±0.38	63.86±1.20	40.45±1.91	70.33	64.32	61.12
Portuguese	80.42±0.32	61.79±0.63	68.18±0.99	70.39	71.88	71.75
Polish	81.00±0.40	69.09±0.91	76.25±1.26	69.14	60.45	61.23
Turkish	70.83±0.06	56.62±0.43	24.08±1.66	76.13	67.18	55.78
Hindi	76.00±0.41	48.21±2.54	45.31±1.55	65.14	59.34	58.22
Japanese	47.43±0.20	35.74±1.92	52.93±1.26	71.34	67.77	69.19
Korean	71.31±0.78	45.30±0.49	34.71±3.06	59.13	62.34	69.70
All	78.63±0.17	68.01±0.27	59.74±0.99	-	-	-

Table 7: Comparison between our system and that of the original MEE paper.

Entities							
Source lang.	English	Spanish	Portuguese	Polish	Turkish	Hindi	Japanese	Korean
English	-	**76.00**±0.00	68.67±0.58	66.67±0.58	57.33±1.53	67.00±1.00	29.33±1.15	46.00±1.00
Spanish	**72.33**±1.15	-	**72.67**±0.58	**69.67**±0.58	59.67±0.58	**68.67**±0.58	27.67±1.53	47.33±0.58
Portuguese	66.33±0.58	71.00±0.00	-	58.33±0.58	50.67±0.58	60.00±1.00	27.00±2.00	40.33±0.58
Polish	69.00±0.00	73.33±0.58	67.00±1.73	-	59.33±0.58	68.33±0.58	31.00±2.00	48.00±1.73
Turkish	71.33±0.58	70.33±1.53	63.67±1.53	66.33±0.58	-	70.00±0.00	33.00±2.65	**51.33**±1.15
Hindi	67.33±1.53	70.33±1.53	62.67±2.08	65.67±1.15	58.33±0.58	-	35.00±0.00	48.00±0.00
Japanese	40.33±5.69	43.33±6.51	41.00±2.00	48.00±5.29	48.00±2.65	44.67±5.03	-	41.67±1.53
Korean	60.00±1.00	58.00±1.73	53.67±0.58	60.67±0.58	**63.33**±0.58	59.33±0.58	**37.00**±1.00	-
Events								
Source lang.	English	Spanish	Portuguese	Polish	Turkish	Hindi	Japanese	Korean
English	-	**52.33**±0.58	37.67±1.15	**50.67**±5.77	**48.67**±3.21	43.33±1.53	1.33±2.31	**35.00**±2.65
Spanish	53.00±2.65	-	**39.00**±1.00	40.67±3.21	48.00±1.00	45.33±2.52	0.33±0.58	32.67±2.89
Portuguese	50.00±0.00	37.00±4.36	-	46.33±3.51	37.00±2.00	34.67±0.58	0.00±0.00	27.67±1.15
Polish	**61.67**±0.58	49.00±0.00	29.33±6.11	-	45.00±2.00	38.00±5.57	0.00±0.00	29.33±4.04
Turkish	51.67±11.1	46.67±4.04	36.00±1.73	39.00±14.9	-	**47.00**±1.73	1.33±1.53	33.33±2.31
Hindi	56.00±4.58	43.33±6.66	18.33±5.51	31.33±11.5	42.00±1.00	-	3.67±2.08	30.00±3.46
Japanese	9.00±11.3	7.33±10.9	7.00±7.55	23.67±8.39	19.33±6.35	7.67±8.08	-	24.33±8.50
Korean	41.00±5.00	21.67±12.6	29.33±5.51	28.67±7.02	30.67±1.15	31.33±6.11	**6.00**±3.00	-
Arguments								
Source lang.	English	Spanish	Portuguese	Polish	Turkish	Hindi	Japanese	Korean
English	-	**20.33**±2.08	**43.33**±1.15	40.67±1.53	**14.67**±3.21	15.33±3.06	14.33±1.53	14.00±2.65
Spanish	13.00±2.65	-	25.67±2.31	14.00±3.46	3.00±3.46	8.00±2.00	0.00±0.00	2.00±0.00
Portuguese	**37.00**±1.73	17.00±2.65	-	**42.00**±2.00	13.67±2.52	16.00±1.73	19.33±1.15	12.00±1.73
Polish	31.00±1.73	15.00±3.00	36.00±2.00	-	14.33±1.53	14.33±2.89	16.67±2.08	14.00±4.36
Turkish	22.33±0.58	18.33±2.52	31.67±0.58	33.67±1.15	-	**29.33**±1.15	**20.00**±2.65	13.33±2.31
Hindi	14.33±1.15	9.33±4.04	19.00±1.00	23.67±1.53	12.33±0.58	-	17.33±0.58	10.33±2.31
Japanese	12.67±0.58	3.67±0.58	14.00±2.65	12.00±1.00	11.33±3.21	13.67±1.53	-	**20.33**±3.06
Korean	14.00±3.46	4.33±1.15	16.00±1.00	18.00±3.61	7.67±2.31	17.00±1.00	17.33±2.31	-

Table 8: Results for every combination of source-target languages excluding Basque. Rows indicate the source language and columns the target. For every combination 3 different runs were done in order to compute the mean and standard deviation.

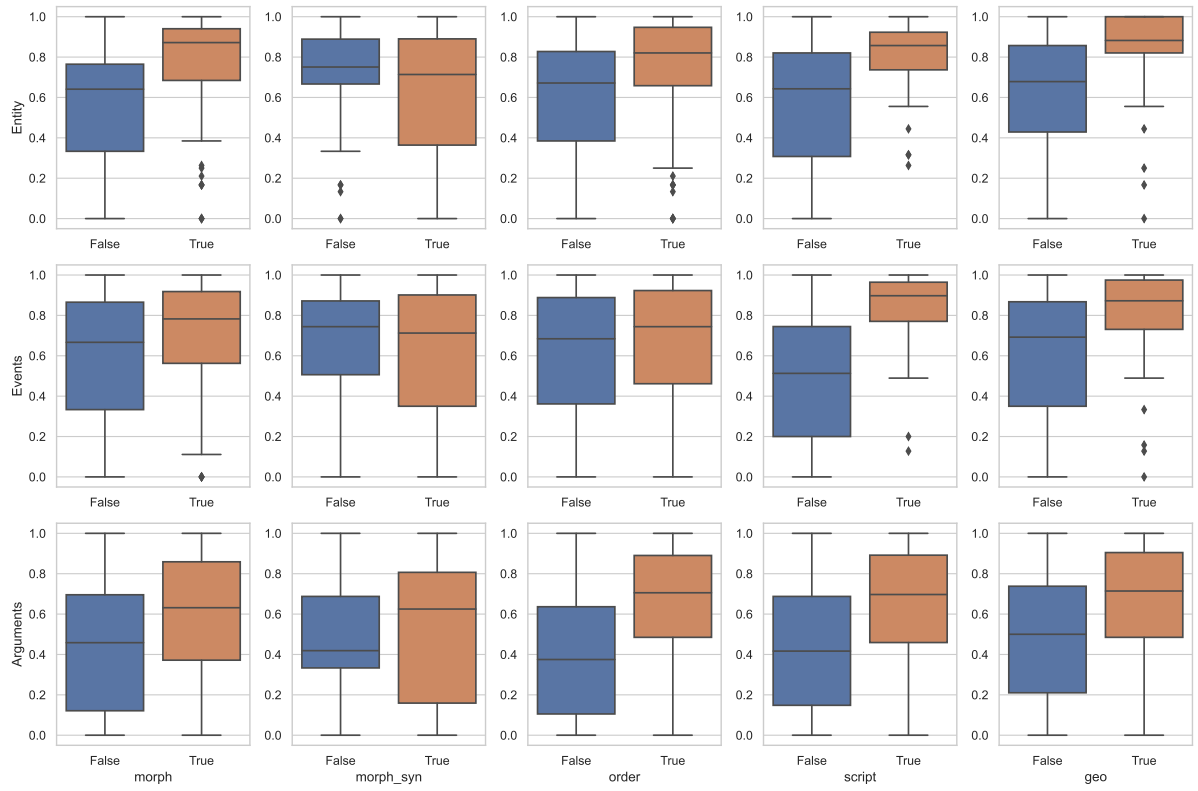

Figure 4: Boxplots of each task and features. The label "True" indicates that source and target languages sharing the same value for a given feature.