# Consistent Crosslingual Data Transfer for Open Information Extraction

Anonymous ACL submission


## Abstract

Progress with supervised Open Information Extraction (OpenIE) has been primarily limited to English due to the scarcity of training data in other languages. In this paper, we explore techniques to automatically convert English text for training OpenIE systems in other languages. We introduce the Alignment Augmented Consistent Translation (AACTrans) model to translate English sentences and their corresponding extractions *consistently* with each other — with no changes to vocabulary or semantic meaning which may result from independent translations. Using the data generated with AACTrans, we train a novel two-stage generative OpenIE model, which we call Gen2OIE, that outputs for each sentence: 1) relations in the first stage and 2) all extractions containing the relation in the second stage. Gen2OIE increases relation coverage using a training data transformation technique that is generalizable to multiple languages, in contrast to existing models that use an English-specific training loss. Evaluations on 5 languages — Spanish, Portuguese, Chinese, Hindi and Telugu — show that the Gen2OIE-AACTrans outperforms prior systems by a margin of 6-40% F1.[1]


## 1 Introduction

With widespread adoption of Deep Learning in NLP, Open Information Extraction (OpenIE) systems have gone through a paradigm shift from using rule-based, statistical systems to supervised neural models. However, both types of systems have been available in only a few languages – earlier because they required language-specific OpenIE insights, and now, because the requirement of annotated training corpus poses a major barrier, particularly for those languages that are considered low-resource.

Related tasks such as Semantic Role Labeling face similar challenges in extending to multiple languages. X-SRL (Daza and Frank, 2020) addresses this by automatic translation of English sentences to the target language followed by label projection to infer the semantic role labels in the translated sentence. However, translating the sentence alone may be insufficient for OpenIE because extractions can include additional words absent in the sentence or require some changes to the word morphology used in the sentence. Although less prevalent in English, these characteristics need to be addressed in other languages.

X-SRL approach may be extended such that each extraction can also be automatically translated and subject, relation, object labels projected from English extractions. However, *independent* translation of sentence and extraction may introduce unwanted *lexical* (e.g. synonyms) or *semantic* (e.g., change in gender) variations between the translations, as shown in Table 1. Such translation inconsistencies in the training data lead to invalid OpenIE examples.

To maintain consistency between translations of a sentence and its extractions, both the translations must use same words or their morphological variants as much as possible. Hence, we propose Alignment Augmented Consistent Translation (AACTrans), a seq2seq model that translates the given input text in a way that is consistent with a *reference translation* by biasing the translation to use words similar to the reference. To ensure that translations of sentence and extractions are consistent with each other, we use AACTrans model to translate each of them with the same reference. In Section 4.1, we describe the reference used in training and inference.

Both generation based (Kolluru et al., 2020b) and labeling based (Ro et al., 2020) architectures have shown competitive performance on English OpenIE. However, labeling based models cannot

[1] Code and data will be released at github.com:xxx

| | |
|---|---|
| **Lexical Inconsistency** | |
| English Sentence | *The shield of Athena Parthenos, sculpted by Phideas, depicts a fallen Amazon* |
| English Extraction | *<s> The shield of Athena Parthenos </s> <r> depicts </r> <o> a fallen Amazon </o>* |
| Spanish Sentence | El escudo de Atena Parthenos, sculptado por Phideas, representa un Amazonas fallecido |
| Spanish Extraction (Indp) | <s> El escudo de Atena Parthenos </s> <r> representa </r> <o> un Amazonas caído </o> |
| Spanish Extraction (Const) | <s> El escudo de Atena Parthenos </s> <r> representa </r> <o> un Amazonas fallecido </o> |
| **Semantic Inconsistency** | |
| English Sentence | *The discovery was remarkable as the skeleton was almost identical to a modern Kuvasz* |
| English Extraction | *<s> skeleton </s> <r> was </r> <o> almost identical to a modern Kuvasz </o>* |
| Spanish Sentence | Un descubrimiento notable porque fósil era casi idéntica a un Kuvasz moderno |
| Spanish Extraction (Indp) | <s> skeleto </s> <r> era </r> <o> casi idéntica a una Kuvasz moderna </o> |
| Spanish Extraction (Const) | <s> fósil </s> <r> era </r> <o> casi idéntica a un Kuvasz moderno </o> |

Table 1: OpenIE examples transferred from English to Spanish, using both Independent (Indp) and Consistent (Const) translations. Independent translation results in inconsistencies which may have the same meaning (by using synonyms, fallecido vs. caído) or may change the meaning (changing gender from male to female, moderno to moderna). Consistent translation avoids these issues, resulting in better quality of training data.

naturally introduce new words or change morphology of sentence words required in some languages. Therefore, we use a new generative model, GEN2OIE, that contains two stages: the first stage produces all the relations in the sentence and the second stage generates the extractions containing the given relation. We also use a training heuristic specific to two stage models that increases relation coverage across multiple languages.

Our major contributions are that we:

1. introduce a novel technique for transferring data from English to other languages using the AACTRANS model and label projection,
2. propose two-stage generative model, GEN2OIE, for training OpenIE system in multiple languages,
3. release OpenIE evaluation datasets for two Indian languages, Hindi and Telugu, and
4. outperform prior systems by 6-40% in F1 over five languages.

## 2 Related Work

The task of OpenIE involves extracting a set of <*subject; relation; object*> tuples from a sentence in an ontology-independent manner where all the fields in the tuple are phrases extracted from the original sentence. Each tuple is referred to as an *extraction*. We focus primarily on these three fields but our method can be easily extended to extract other fields such as location and time, generated by prior systems such as OpenIE-4 (Pal and Mausam, 2016; Christensen et al., 2011).

Many of the prior OpenIE systems, both non-neural (OpenIE-5 (Saha et al., 2017; Saha and Mausam, 2018), ClausIE (Del Corro and Gemulla, 2013)) and neural (RnnOIE (Stanovsky et al., 2018), OpenIE-6 (Kolluru et al., 2020a)) only work for English language. Moreover, OpenIE systems built for other languages often work only for a single language due to their reliance on language-specific resources. For example, (Bassa et al., 2018; Rahat and Talebpour, 2018; Romadhony et al., 2018; Guarasci et al., 2020; Papadopoulos et al., 2021) focus on German, Persian, Indonesian, Italian, and Greek. Claro et al. (2019) presents the importance and various challenges involved with building multilingual OpenIE systems. Neural models like Logician (Sun et al., 2018) and CrossOIE (Cabral et al., 2020) use language specific training data. Reliance on manually-annotated data or language-specific resources makes it infeasible to develop systems for the plurality of languages in the world, due to the cost and effort involved. However, our automated data conversion method can handle even low-resource languages like Telugu.

Non-neural systems such as PredPatt (White et al., 2016) and ArgOE (Gamallo and Garcia, 2015) work for multiple languages by using CoNLL-X and Universal Dependency parses respectively to extract the predicate-argument structures. Due to their pipelined nature, their performance is below that of neural systems like Multi$^2$OIE (Ro et al., 2020). Multi$^2$OIE is a two-stage labeling model that works for English, Spanish and Portuguese. GEN2OIE model extends this 2-stage design to the generative paradigm which allows for better modeling of the OpenIE task. The underlying mBERT encoder in Multi$^2$OIE allows for cross-lingual generalization across various languages even after training with only English su-

pervised data. However, dependence on zero-shot generalization limits the performance of the model. Using this model as the baseline, we find that we are able to achieve better performance by using training data generated for a specific language.

Two types of methods have been proposed for constraining the outputs of the machine translation systems: 1) altering the decoding algorithm (Hasler et al., 2018), or 2) modifying the training methodology (Chen et al., 2020; Dinu et al., 2019). We follow the second approach for constraining translations in AACTrans model to be consistent to that of a reference sentence. Unlike prior work which focuses on constraining translations of few words our task requires constraining the entire translation. We make of awesome-align (Dou and Neubig, 2021), an unsupervised word alignment technique (Och and Ney, 2003), that outputs the alignment between words in sentences of two languages. Awesome-align is trained using only parallel set of sentences in the two languages and generates corresponding aligned target words for each source word.

## 3 Notation

For the transfer of OpenIE data from one language to another, we represent the source language[2] as $E$ and the target language as $F$. Further, we use $sent_E$ and $ext_E$ to represent a sentence and extraction in the source language and $aact\text{-}sent_F$ and $aact\text{-}ext_F$ to represent the transferred sentence and extraction in the target language.

To aid in the translation of extractions, we create a sub-sentence from each extraction by concatenating the phrases in all the fields of the extraction. The order of concatenation is such that the formed sub-sentence is grammatically valid. We refer to this sub-sentence as an *ext-sentence* and represent it as $es_L$, where the subscript $L$ represents its language. For most English extractions, the ext-sentence corresponds to concatenating the fields in the order of subject, relation and object. However, other languages may follow a different order or allow for multiple orders. We rely on the output of system that translates the English ext-sentence to determine the ext-sentence in other languages. Moreover, each extraction can be seen as a labeling over the words of ext-sentence with either the **S**ubject, **R**elation or **O**bject tags. Tags for each word in the ext-sentence can also be regarded as the extraction.

[2] In the current work, we always use English as source

## 4 Crosslingual Data Transfer

In this section we describe the technique used to convert OpenIE training data from source language $E$ to a target language $F$. The source sentence, $sent_E$, and all its corresponding ext-sentences, $es_E$, are consistently translated to language $F$ (Section 4.1), and then, for each extraction in language $E$, $ext_E$, the **S**, **R** or **O** labels are projected to the translated ext-sentence, $es_F$, to form the extraction, $ext_F$, in language $F$ (Section 4.2). Figure 1 describes the pipeline with the help of an example.

### 4.1 Consistent Translation

We introduce a new Seq2Seq-based translation model called Alignment Augmented Consistent Translation (AACTrans) to ensure that sentences and ext-sentences are translated consistently from languages $E$ to $F$. We define two translations as consistent if similar phrases have same grammatical structure, vocabulary and morphology while allowing for minimal changes necessary to ensure fluency.

To ensure consistency among translations of multiple pieces of text (both the sentence and respective ext-sentences present in an English OpenIE instance), we make use of a reference text in language $F$ to guide all of their translations. By individually maintaining consistency with the reference, their respective translations end up being consistent to one another as well.

To generate a translation $\mathbf{f}$ (language $F$) of text $\mathbf{e}$ (language $E$), consistent with a reference $\mathbf{r}$ (language $F$), we use the following procedure.

Firstly, given $\mathbf{e} = e_1 e_2 \dots e_N$ and $\mathbf{r} = r_1 r_2 \dots r_M$, we find the set of aligned words $A_{e_i}$=$\{r_j\}$ for each word $e_i$ in $\mathbf{e}$, using a word alignment model.

Secondly, the aligned text $\mathbf{e}'$ is constructed by concatenating each of the words $e_i$ in $\mathbf{e}$, with their aligned words $A_{e_i}$, using ## as a separator (shown as <1>, <3> $\rightarrow$ <4> and <2>, <3> $\rightarrow$ <5> in Figure 1). If $e_i$ is aligned to the words $r_j$, $r_k$ ($j < k$), then $\mathbf{e}'$ contains $e_i$ ## $r_j$ $r_k$ #. If $e_i$ has no aligned words, then $\mathbf{e}'$ contains $e_i$ #.

Thirdly, the AACTrans model takes $\mathbf{e}'$ as input and produces the sequence $\mathbf{f}$ as output, which represents a translation of $\mathbf{e}$ that is biased to use the aligned reference words (shown as <4> $\rightarrow$ <7> and <5> $\rightarrow$ <8> in Figure 1).

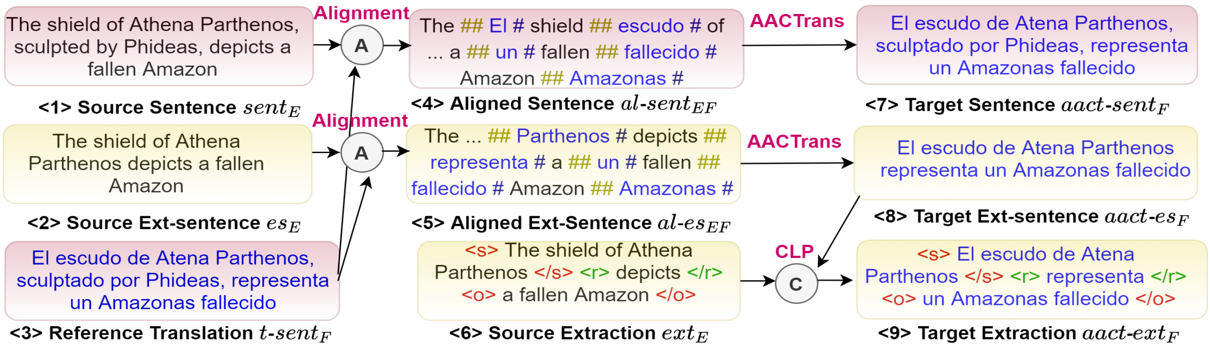


Figure 1: Crosslingual Data Transfer pipeline from English to Spanish. The sentence and ext-sentence in English are aligned with a translation of the sentence. The AACTrans model uses the aligned text to generate the final consistent translations. Cross Lingual Projection (CLP) introduces **S**, **R**, **O** tags in the extraction.

Next we discuss the training and inference of AACTrans model.

**Training**: We use parallel sentences of languages $E$ and $F$ that are available in existing translation corpora for training the AACTrans model. For each parallel sentence pair $\mathbf{e}$ and $\mathbf{f}$, we use the sentence $\mathbf{f}$ itself as the reference $\mathbf{r}$. Using the alignments between the words of $\mathbf{e}$ and $\mathbf{f}$, we form the input $\mathbf{e}'$, as discussed. The AACTrans Seq2Seq model is trained with $\mathbf{e}'$ as input and $\mathbf{f}$ as output. Since $\mathbf{e}'$ has words from $\mathbf{f}$, the model learns to use them during training.

Note that this is an easier task than standard translation as the aligned words provide a strong clue about the output sequence. However, the task still remains non-trivial as not all words in $\mathbf{f}$ may be aligned and the order of words in $\mathbf{e}$ may differ from the order in which they are present in $\mathbf{e}'$.

**Inference**: Here, we consistently translate English sentence $sent_E$ and each of its ext-sentences $es_E$. We use an off-the-shelf translation system to translate $sent_E$ to language $F$, represented as $t\text{-}sent_F$. $t\text{-}sent_F$ is used as the common reference $\mathbf{r}$ for constructing aligned sentence $al\text{-}sent_{EF}$ and aligned ext-sentence $al\text{-}sent_{EF}$ from sentence $sent_E$ and ext-sentence $es_E$, respectively. We then apply the trained AACTrans model on $al\text{-}sent_{EF}$ and $al\text{-}sent_{EF}$ to generate target sentence $aact\text{-}sent_F$ and target ext-sentence $aact\text{-}es_F$ respectively.

### 4.2 Crosslingual Label Projection (CLP)

Each word in the target ext-sentence, $aact\text{-}es_F$, must be labeled with either the Subject, Relation, or Object tag to form the completed extraction in language $F$. The tags from the corresponding $ext_E$ are projected onto $aact\text{-}es_F$ using the Crosslingual Projection algorithm (Faruqui, 2015), which uses word alignments between $es_E$ and $aact\text{-}es_F$ and produces as output, the tags over $aact\text{-}es_F$, giving extraction $aact\text{-}ext_F$. The final set of sentence, extractions pairs constitute the data for training OpenIE system in language $F$. In Appendix A, we describe the label projection with the help of an example.

## 5 Gen2OIE Model

To train OpenIE systems in multiple languages, we use a novel Gen2OIE model that extends the 2-stage design of Multi$^2$OIE (Ro et al., 2020) to a generative paradigm. The first stage generates all possible relations and the second stage generates all the OpenIE extractions that contain a given relation.

Gen2OIE can produce overlapping relations and multiple extractions containing the same relation, thus overcoming the limitations of Multi$^2$OIE model. Moreover, due to its generative nature, Gen2OIE can add new words or introduce changes in morphology that may be necessary for producing correct extractions, which cannot be achieved by labeling models.

Both the stages of the Gen2OIE (shown in Figure 2) use Seq2Seq models as follows:

**Stage-1 Seq2Seq**: The input sentence is passed to the encoder and decoder generates a string formed by concatenating the set of relations from all the extractions, separated by an *[SEP]* token. During training, the target relations are concatenated in the order in which they occur in the sentence. We find that a deterministic order is important for adding stability to the model training.

**Stage-2 Seq2Seq**: To produce extractions corre-

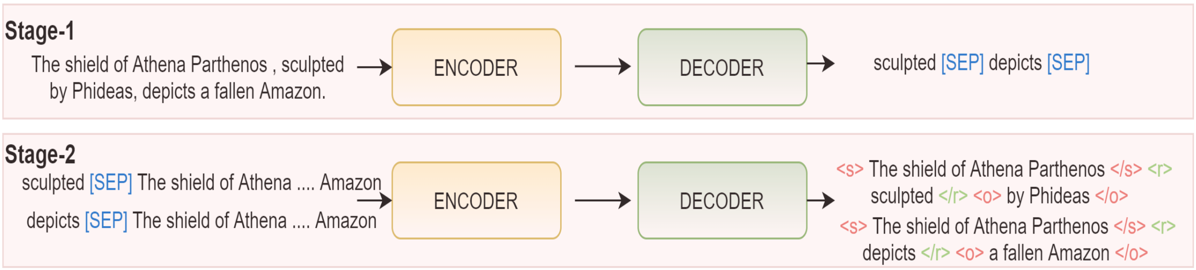


Figure 2: GEN2OIE model contains two Seq2Seq models. In Stage-1, it generates all relations in the sentence, separated by an [SEP] token. For each detected relation in Stage-2, it generates extractions containing the relation.

sponding to each relation generated in Stage-1, the relation $r$ is concatenated with the input sentence $s$ and passed to the encoder as "$r$ *[SEP]* $s$". The decoder is trained to generate all the extractions containing the relation $r$. Multiple extractions are separated by an <*e*> token and each extraction contains delimiters tokens to identify the various parts of the extraction. The surrounding <*s*>...</*s*>, <*r*>...</*r*> and <*o*>...</*o*> tokens are used to identify the subject, relation and object phrases.

Labeling models like OpenIE-6 (Kolluru et al., 2020a) have used constrained training to increase the relation coverage. However, the constraints are limited to English and specific to labeling architectures. We introduce a simple parts-of-speech based heuristic during Stage-1 training of GEN2OIE that increases the relation coverage in the generative paradigm while being applicable across languages.

**Relation Coverage Heuristic**: We observe that for generating all possible extractions, all the verbs in the sentence must be contained in some relation. However, the extractions of training data may be incomplete and not satisfy this property. Therefore, during the training phase, we modify the input to the Stage-1 model by removing the verbs in the sentence which are not present in relation of any extraction. Thus the model learns that every verb must be included in some relation and applies the same during inference as well. Note that this heuristic does not effect Stage-2 model training.

## 6 Confidence Scoring

The word log probabilities assigned by the Stage-2 decoder can be summed up to be used as confidence score for the generated extraction. Similar to Kolluru et al. (2020a), we also experiment with using separate models for obtaining the confidence scores of extractions generated by GEN2OIE model. We use Seq2Seq models and sequence labeling models.

Both the models are trained with sentence, extraction as the input, output pairs. Given the input sentence, they compute the confidence scores by summing up the word log probabilities assigned to the extraction generated from GEN2OIE.

## 7 Experimental Setting

We train OpenIE systems in 5 languages, Spanish (ES), Portuguese (PT), Chinese (ZH), Hindi (HI) and Telugu (TE), by using the training data transferred from English to the respective language. For training the Seq2Seq models used in the data generation pipeline and the OpenIE systems based on the GEN2OIE architecture, we choose either the mBART (Liu et al., 2020) or mT5 (Xue et al., 2020) model depending on the particular language. Both of them are pre-trained multilingual Seq2Seq models that are trained with a span denoising objective on a large corpus of text containing many languages. mBART is pre-trained on CC25 and mT5 is pre-trained on mC4 corpus which contain text in 25 and 101 languages, respectively. Since mBART does not support Portuguese and Telugu, we use mT5 for these two languages and mBART for the remaining 3 languages. Although mT5 supports all the 5 languages, we use it for only the two languages as it depends on TPUs[3] for efficient training, which are expensive to use, while mBART can be trained using the fairseq library[4] on GPUs. We use the default hyperparameters recommended for these models and they are reported in Appendix D.

**Training Datasets**: For training the AACTRANS model, we make use of parallel English, language $F$ sentences available in standard translation corpora using the method described in Section 4. For Spanish we use parallel sentences from EuroParl corpus (Koehn et al., 2005), and for Portuguese we use a subset of the ParaCrawl corpus (Bañón

[3] https://cloud.google.com/tpu
[4] github:pytorch/fairseq

| | EN | ES | PT | ZH | HI | TE |
|---|---|---|---|---|---|---|
| Translation | | | | | | |
| Train | - | 1.9M | 5M | 1M | 1.6M | 4.8M |
| Test | - | 38473 | 99,087 | 2001 | 2507 | 2390 |
| OpenIE | | | | | | |
| Train | 91K | 91K | 91K | 91K | 91K | 91K |
| Test | 641 | 594 | 594 | 3833 | 100 | 100 |

Table 2: Data statistics for OpenIE examples and (English, language *F*) parallel sentences.

et al., 2019), as chosen by Lopes et al. (2020). For Hindi we use the IIT-B corpus (Kunchukuttan et al., 2018), and for Telugu we use the Samanantar corpus (Ramesh et al., 2021). For Chinese we use the data released for WMT19 (Barrault et al., 2019). We list the BLEU scores of the various systems in Appendix C.

We use the OIE4 training corpus from Kolluru et al. (2020b) and transfer it to other languages for training OpenIE systems.

**Evaluation Datasets and Metrics**: For evaluating translation systems we use the test sets available in the respective corpora and use SacreBLEU (Post, 2018) as the metric.[5] For evaluating different OpenIE systems we use the Optimal F1 and Area Under Curve (AUC) as computed by the CaRB (Bhardwaj et al., 2019) scoring function. For Spanish, Portuguese OpenIE we use test sets provided in Ro et al. (2020). For Chinese OpenIE, we randomly choose 10% of the SAOKE dataset (Sun et al., 2018).

In order to evaluate our method on medium and low resource languages, we annotate new OpenIE test sets in Hindi and Telugu. Human annotators (two for each language) who are fluent in both the language and are knowledgeable about the OpenIE task translated 100 randomly chosen sentences and their corresponding extractions from CaRB test set.

Table 2 lists the number of examples in different languages used for training and evaluating translation and OpenIE systems.

## 8 Experiments

We perform experiments that address the following questions:

1. How effective is the GEN2OIE model for the task of OpenIE?

[5] BLEU+case.mixed+numrefs.1+smooth.none+tok.intl+version.1.5.1

| Model | EN | |
|---|---|---|
| | F1 | AUC |
| IMoJIE | 53.6 | 33.3 |
| IGL | 52.5 | 33.8 |
| CIGL | 54 | 36 |
| OpenIE6 | 52.7 | 33.7 |
| Multi$^2$OIE | 52.5 | 31.6 |
| GENOIE | 52.1 | 30.3 |
| GEN2OIE w/o RC | 51.9 | 29.7 |
| GEN2OIE | 54.0 | 32.0 |
| (seq2seq-rescore) | **54.4** | 32.1 |
| (label-rescore) | 54.2 | **38.5** |

Table 3: Performance of OpenIE systems in English, evaluated with the CaRB metric. GEN2OIE along with Label Rescoring produces the best performance.

2. What is the quality of data generated with the AACTRANS+CLP pipeline, assessed by performance of systems trained with it, and how does it compare with alternative methods of data generation?

3. What are the roles of different components in the GEN2OIE and AACTRANS+CLP data?

### 8.1 Effectiveness of GEN2OIE

To study the baseline monolingual effectiveness of GEN2OIE, we first train and evaluate the system on English data. The results are shown in Table 3. We compare with previously proposed English OpenIE models such as Multi$^2$OIE (Ro et al., 2020), OpenIE6 (Kolluru et al., 2020a) and IMoJIE (Kolluru et al., 2020b). We also consider individual components in OpenIE6, the IGL and Constrained-IGL (CIGL) architectures. CIGL achieves the highest performance among all prior models but makes use of English specific constraints in the loss function.

We find that GEN2OIE, which uses the proposed language-agnostic relation coverage (RC), matches the performance of CIGL with 54% F1. However, its AUC remains lower. Therefore, we rescore the generated extractions with seq2seq and labeling-based rescoring models (Section 6). This results in a new state of the art for English in F1 and AUC with the labeling-based rescoring resulting in a 2.5% gain over CIGL in AUC.

To further analyze the effectiveness of our 2-stage architecture, we introduce another model called GENOIE that outputs all extractions for a sentence as a single string, separated by an <e> token. We find that using GENOIE results in (1.9, 1.7)% drop in F1, AUC compared to GEN2OIE

| Model | Training Data | ES | | PT | | ZH | | HI | | TE | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | AUC | F1 | AUC | F1 | AUC | F1 | AUC | F1 | AUC |
| (Faruqui, 2015) | English | 45.5 | 28.6 | 48.5 | 31.5 | 13.7 | 3.3 | 42.4 | 22.3 | 40.6 | 22.2 |
| Multi$^2$OIE | English | 60.0 | 41.5 | 60.2 | 41.1 | 23.7 | 8.1 | 24.5 | 7.2 | 19.5 | 5.5 |
| Multi$^2$OIE | SentTrans+CLP | 62.0 | 42.8 | 60.9 | 41.3 | 21.2 | 6.5 | 56.7 | 37.6 | 38.0 | 21.4 |
| | SentTrans+CLP | 60.4 | 40.6 | 63.5 | 43.7 | 20.9 | 4.9 | 57.4 | 32.9 | 47.2 | 21.0 |
| GenOIE | SentExtTrans+CLP | 58.3 | 39.7 | 57.3 | 36.5 | 20.8 | 5.6 | 60.0 | 35.2 | 45.7 | 23.5 |
| | AACTrans+CLP | 60.8 | 41.3 | 63.9 | 44.8 | 23.1 | 5.9 | 58.3 | 34.3 | 47.9 | 23.3 |
| | SentTrans+CLP | 64.2 | 44.6 | 65.6 | 50.0 | 29.0 | 8.9 | 61.6 | 37.7 | 49.7 | 21.7 |
| Gen2OIE | SentExtTrans+CLP | 64.7 | 46.1 | 63.7 | 45.5 | 29.3 | 10.2 | 62.0 | 38.7 | 50.4 | 24.9 |
| | AACTrans+CLP | 65.9 | 47.2 | 66.4 | 50.5 | 29.8 | 10.3 | 62.1 | 38.8 | **52.2** | 26.6 |
| (seq2seq-rescore) | AACTrans+CLP | **66.0** | 46.8 | **66.5** | 50.8 | **30.5** | 9.8 | 62.1 | 37.6 | 52.0 | 26.7 |
| (label-rescore) | AACTrans+CLP | 65.9 | **51.5** | 66.3 | **54.1** | 29.8 | **13.8** | **62.3** | **48.2** | 52.1 | **35.3** |

Table 4: F1 and AUC performance of OpenIE systems in Spanish (ES), Portuguese (PT), Chinese (ZH), Hindi (HI) and Telugu (TE). Training with AACTrans+CLP data shows strong performance with both GenOIE and Gen2OIE models. Labeling-based rescoring improves AUC in all languages.

which leverages RC heuristic. We note that this heuristic cannot be applied to GenOIE as it involves removal of words in the input sentence that may appear in other fields of the extraction. RC heuristic is applicable to Stage-1 training as it only predicts the relations. We also report Gen2OIE performance without the RC heuristic.

### 8.2 Quality of AACTrans+CLP data

In order to test the quality of the OpenIE examples generated using the AACTrans+CLP pipeline, we train both the GenOIE and Gen2OIE models over the data generated for different languages. In Table 4, we compare it with examples generated from two other methods, SentTrans+CLP and SentExtTrans+CLP.

SentTrans+CLP represents an adaptation of X-SRL (Daza and Frank, 2020) for OpenIE where only the sentence is translated and each extraction, which is expressed as labeling over the words in the sentence, are projected onto the translated sentence using the CLP algorithm described in Section 4.2. The projected extraction is now a labeling over the translated sentence and hence it uses the same morphology as the sentence and cannot add new words. SentExtTrans+CLP uses independent translation of English sentence and ext-sentences followed by CLP algorithm between the English and translated ext-sentences to transfer the labels. Although this allows for adding new words and changing morphology, it can result in a lack of consistency between the translations.

We find that both GenOIE and Gen2OIE show consistent gains with AACTrans+CLP data across various languages, when compared with SentExtTrans+CLP and SentTrans+CLP data. The highest gain for consistent translation over independent translation is (2.7, 5)% in F1, AUC for Portuguese Gen2OIE.

We further use rescoring models that are trained on the same AACTrans+CLP data. Seq2seq-based rescoring improves the F1 marginally but labeling-based rescoring achieves significantly higher AUC, with as much as 9.2% gain in Hindi.

We experiment with two versions of Multi$^2$OIE: 1) trained only on English OpenIE data and applied to other languages in a zero-shot manner and 2) using language-specific training data generated from SentTrans+CLP. We specifically choose SentTrans+CLP data as all the extractions can be expressed as labels over the sentence, which is a requirement for training Multi$^2$OIE which is itself a labeling model. We find that Multi$^2$OIE model trained with SentTrans+CLP data improves over the zero-shot setting in all languages other than Chinese (discussed below). However, it performs significantly worse than Gen2OIE by (6.5, 2.7)% in (F1, AUC) on average, even on training with the same SentTrans+CLP data. This can be attributed to Multi$^2$OIE's lack of capability to handle: 1) overlapping relations, 2) multiple extractions per relation, 3) adding auxiliary words or 4) changing inflectional forms, as shown in Table 5.

We additionally compare with Faruqui (2015), where the test sentence is translated into English, extractions are generated using OpenIE6 and they are projected back onto the test sentence. We find that the system results in poor performance due to lack of language-specific training.

We observe that all systems have low perfor-

| | |
|---|---|
| Sentence<br>Extractions | George Bluth Sr., patriarch of the Bluth family, is the founder and former CEO of the Bluth Company.<br><s> George Bluth Sr. </s> <r> is patriarch of </r> <o> the Bluth family </o><br><s> George Bluth Sr. </s> <r> is </r> <o> the founder and former CEO of the Bluth Company </o><br><s> George Bluth Sr. </s> <r> is </r> <o> patriarch of the Bluth family </o> |
| Telugu<br>English<br>Extraction | షరోన్ యొక్క దీర్ఘకాల ప్రత్యర్థి బెంజమిన్ నెతన్యాహును లికుడ్ నాయకుడిగా ఎన్నుకున్నారు<br>*Sharon's longtime rival Benjamin Netanyahu was elected as leader of Likud*<br><s> షరోన్ యొక్క దీర్ఘకాల ప్రత్యర్థిని </s> <o> లికుడ్ నాయకుడిగా </o> <r> ఎన్నుకున్నారు </r> |
| Hindi<br>English<br>Extraction | जॉन लैंबर्ट ने सरकार के साधन के रूप में जाना जाने वाला एक नया संविधान सामने रखा<br>*John Lambert put forward a new constitution known as the Instrument of Government*<br><s> एक नया संविधान </s> <o> सरकार के साधन के रूप में </o> <r> जाना जाता है </r> |

Table 5: Sentence and OpenIE predictions of GEN2OIE in English, Telugu and Hindi. It is capable of generating overlapping relations (*is*, *is patriarch of*), multiple extractions per relation (*is*), add auxiliary words (जाने -> जाता है ) or change inflection forms (ప్రత్యర్థి -> ప్రత్యర్థిని ) as necessary.

| Model (Data) | ES | | ZH | | HI | |
|---|---|---|---|---|---|---|
| | F1 | AUC | F1 | AUC | F1 | AUC |
| GEN2OIE (AACTRANS+CLP) | **65.7** | **47.2** | **29.8** | **10.3** | **62.1** | 38.8 |
| GEN2OIE (AACTRANS w/o Sentence Consistency+CLP) | 64.0 | 44.3 | 29.6 | 10.3 | 60.9 | 37.5 |
| GEN2OIE w/o Relation Ordering (AACTRANS+CLP) | 65.2 | 45.6 | 29.6 | 9.8 | 61.6 | 38.4 |
| GEN2OIE w/o Relation Coverage (AACTRANS+CLP) | 60.6 | 40.3 | 23.9 | 6.6 | **62.1** | **39.8** |

Table 6: Ablations of GEN2OIE model trained with AACTRANS+CLP data on ES, ZH and HI. We analyze the effect of removing 3 components and re-training the model: 1. Sentence Consistency used in AACTRANS data generation, and 2. Relation Ordering used and 3. Relation Coverage heuristic used in Stage-1 model training.

mance on Chinese. We attribute this to the various artifacts present in the SAOKE test set, that include special relations such *DESC*, *TIME*, *ISA*, etc. Since these extractions cannot be generated in our pipeline, we observe performance of only 30.5% F1 with our best model, when compared to training GEN2OIE with SAOKE training data, which gives 52.5% F1 and 32% AUC.

### 8.3 Ablation Study

We choose three representative languages to conduct the ablation study — Spanish, Chinese, and Hindi. Portuguese and Telugu belong to the same language family as Spanish and Hindi, respectively. In Table 6, we show the results of individually removing components from the GEN2OIE trained on AACTRANS+CLP data.

In AACTRANS w/o Sentence Consistency, we use regular translation of sentence while continuing to use consistent translation of extraction. This leads to a drop of (1.7, 0.2, 1.2)% pts in F1 for the three languages, respectively, and shows the importance of using consistent translation on both the sentence and extraction.

In GEN2OIE w/o Relation Ordering, we train Stage-1 GEN2OIE with randomly shuffled relations. This reduces the performance as our model uses auto-regressive training which benefits from following a fixed order, which we choose as the order of occurrence of the relations in the sentence.

In GEN2OIE w/o Relation Coverage, we find that performance decreases in Spanish and Chinese by 5.1% and 5.9% in F1, respectively, but remains the same in Hindi, possibly due to the smaller number of examples in the test set.

**Error Analysis**: We find that the AACTRANS+CLP suffers from: 1) missing or 2) wrong word alignments and 3) inability to label discontinuous S, R, O phrases. We show examples of these cases in Appendix B.

## 9 Conclusion

We develop a novel AACTRANS+CLP pipeline for consistently transferring English OpenIE examples to other languages. We also present a novel two-stage generative model, GEN2OIE, for training OpenIE systems in various languages. We show strong improvements over the existing baseline of Multi$^2$OIE, with an average improvement of 19.5% in F1 and 10.6% in AUC. We show its effectiveness in over five languages, which is the largest number of languages covered by a single OpenIE technique known to us. To encourage research in medium and low-resource languages, we additionally release new OpenIE evaluation examples in Hindi and Telugu.

## References


Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2019. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*.

Akim Bassa, Mark Kröll, and Roman Kern. 2018. Gerie-an open information extraction system for the german language. *J. Univers. Comput. Sci.*, 24(1):2–24.

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam. 2019. CaRB: A Crowdsourced Benchmark for OpenIE. 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), 2019*, pages 6263–6268.

Bruno Souza Cabral, Rafael Glauber, Marlo Souza, and Daniela Barreiro Claro. 2020. Crossoie: Cross-lingual classifier for open information extraction. In *PROPOR*, pages 368–378.

Guanhua Chen, Yun Chen, Yong Wang, and V. Li. 2020. Lexical-constraint-aware neural machine translation via data augmentation. In *IJCAI*.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the sixth international conference on Knowledge capture*, pages 113–120. ACM.

Daniela Barreiro Claro, Marlo Souza, Clarissa Castellã Xavier, and Leandro Oliveira. 2019. Multilingual open information extraction: Challenges and opportunities. *Information*.

Angel Daza and Anette Frank. 2020. X-srl: A parallel cross-lingual semantic role labeling dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Luciano Del Corro and Rainer Gemulla. 2013. ClausIE: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web (WWW), 2013*, pages 355–366. ACM.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Y. Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *ACL*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Manaal Faruqui. 2015. Multilingual open relation extraction using cross-lingual projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Pablo Gamallo and Marcos Garcia. 2015. Multilingual open information extraction. In *Portuguese Conference on Artificial Intelligence*, pages 711–722. Springer.

Raffaele Guarasci, Emanuele Damiano, Aniello Minutolo, Massimo Esposito, and Giuseppe De Pietro. 2020. Lexicon-grammar based open information extraction from natural language sentences in italian. *Expert Systems with Applications*, 143:112954.

E. Hasler, A. Gispert, Gonzalo Iglesias, and B. Byrne. 2018. Neural machine translation decoding with terminology constraints. In *NAACL*.

Philipp Koehn et al. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020a. OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam, and Soumen Chakrabarti. 2020b. IMoJIE: Iterative Memory-Based Joint Open Information Extraction. In *The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Seattle, U.S.A.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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

Alexandre Lopes, Rodrigo Nogueira, Roberto Lotufo, and Helio Pedrini. 2020. Lite training strategies for Portuguese-English and English-Portuguese translation. In *Proceedings of the Fifth Conference on Machine Translation*.

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

Harinder Pal and Mausam. 2016. Demonyms and compound relational nouns in nominal OpenIE. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, pages 35–39.

Dimitris Papadopoulos, Nikolaos Papadakis, and Nikolaos Matsatsinis. 2021. PENELOPIE: Enabling open information extraction for the Greek language through machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium. Association for Computational Linguistics.

Mahmoud Rahat and Alireza Talebpour. 2018. Parsa: An open information extraction system for persian. *Digital Scholarship in the Humanities*, 33(4):874–893.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Youngbin Ro, Yukyung Lee, and Pilsung Kang. 2020. Multi^2OIE: Multilingual open information extraction based on multi-head attention with BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Ade Romadhony, Ayu Purwarianti, and Dwi H Widyantoro. 2018. Rule-based indonesian open information extraction. In *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, pages 107–112. IEEE.

Swarnadeep Saha and Mausam. 2018. Open information extraction from conjunctive sentences. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2288–2299.

Swarnadeep Saha, Harinder Pal, and Mausam. 2017. Bootstrapping for numerical OpenIE. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 317–323. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised Open Information Extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long Papers)*, pages 885–895.

Mingming Sun, Xu Li, Xin Wang, Miao Fan, Yue Feng, and Ping Li. 2018. Logician: A unified end-to-end neural approach for open-domain information extraction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 556–564.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal Decompositional Semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. 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*.

# Consistent Crosslingual Data Transfer for Open Information Extraction (Appendix)

## A Crosslingual Label Projection (CLP)

In this section, we discuss CLP algorithm for projecting labels from English extraction to other language. Consider English sentence, E: *Dutil - Dumas experiment was promoted by an organization called Encounter 2001 denotes* and Spanish sentence, S: *Experimento Dutil - Dumas fue promovido por una organización llamada Encounter 2001*. The word alignments between these sentences are listed in Figure 3 and equivalent phrases from the phrase extract algorithm are shown in Table 7. Consider the English extraction, *(Dumas experiment; was promoted; by an organization)*. For each phrase in the tuple, CLP algorithm looks for the highest BLEU match phrase from Table 7. The subject phrase *Dumas experiment* has best BLEU match to *Dutil - Dumas experiment* and so the corresponding Spanish phrase *Experimento Dutil - Dumas* will be marked as subject. Note that the phrase *Dumas experiment* is not present in Table 7 because its aligned phrase is not continuous in Spanish sentence as can be seen in Figure 3. Similarly for the relation phrase *was promoted*, we find *fue promovido* from Table 7. Continuing the same algorithm, we get *(Experimento Dutil - Dumas; fue promovido; por una organización)* as the final Spanish extraction.

## B Error Analysis

We list three cases that decrease the quality of transferred data using the AACTrans+CLP pipeline.

**Missing word alignments**: For example, English extraction, *A couple of trojans have also been found orbiting with Mars* translates to *También se han encontrado un par de trojas en órbita con Mars* in Spanish. The verb *orbiting* changes to the form *en órbita (in orbit)* (nominalization). The word *en* in Spanish does not align with any word in the English extraction as can be seen in Figure 4. So, projection of *(A couple of trojans; have also been found; orbiting with Mars)* leads to *(un par de trojas; También se han encontrado; órbita con Mars)* which is not fluent because of missing word *en* in the object phrase.

In languages like Spanish and Portuguese, we found alignments to be of high precision but often miss some alignments, as shown above. Next, we see how wrong alignments can affect projection quality.

**Wrong word alignments**: Consider the following English (E) and Hindi (H) ext-sentences, E: *Many organizations like the Samskrita Bharati are conducting Speak Sanskrit workshops to popularize Sanskrit* and H: संस्कृता भारती जैसे कई संगठन संस्कृति को लोकप्रिय बनाने के लिए बोल संस्कृति कार्यशालाएं आयोजित कर रहे हैं . We find that the word *the* is wrongly aligned to the hindi word, कर . So, the subject phrase *Many organizations like the Samskrita Bharati* does not have a continuous phrase in Hindi sentence because it has many words till कर that do not map to the subject phrase in English sentence. Therefore, the CLP algorithm matches a partial phrase *Many organizations like* which is the best BLEU match to the given subject phrase and its equivalent continuous phrase जैसे कई संगठन संस्कृति को gets tagged as subject in Hindi. Whereas संस्कृता भारती जैसे कई संगठन संस्कृति को would be an ideal subject phrase.

**Discontinuous phrases**: Pharse extract in the CLP algorithm assumes continuous phrases in English map to continuous phrase in other language. This assumption would lead to incomplete extractions in the other languages. For example, consider English extraction E: *(Winston Churchill; twice suggested; naming a British battleship)* and its Telugu extraction sentence T: విన్స్టన్ చర్చిల్ రెండుసార్లు బ్రిటిష్ యుద్ధనౌకకు పేరు పెట్టాలని సూచించారు . The relation phrase *twice suggested* is mapped as follows in Telugu: The word *twice* is mapped to రెండుసార్లు and *suggested* is mapped to సూచించారు . The equivalent phrase *twice suggested* is no longer continuous in Telugu language. CLP algorithm looks for best BLEU match that results in matching to the phrase *twice* and its equivalent రెండుసార్లు is tagged as relation. The ideal relation in this example would be రెండుసార్లు సూచించారు

## C BLEU scores

Table 8 contains the BLEU scores of both the normal as well as consistent translations. We find that the performance remains nearly the same, indicating that the improved OpenIE performance stems from the consistency in the translations.

## D Reproducibility

**Compute Infrastructure**: We use V100 (32 GB) GPU for training the mBERT models and use TPU v3-8 for training the mT5 models.

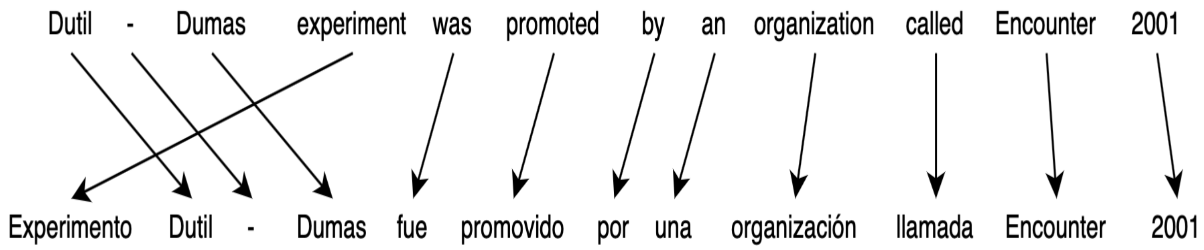


Figure 3: Equivalent English and Spanish sentence with corresponding word alignments between them

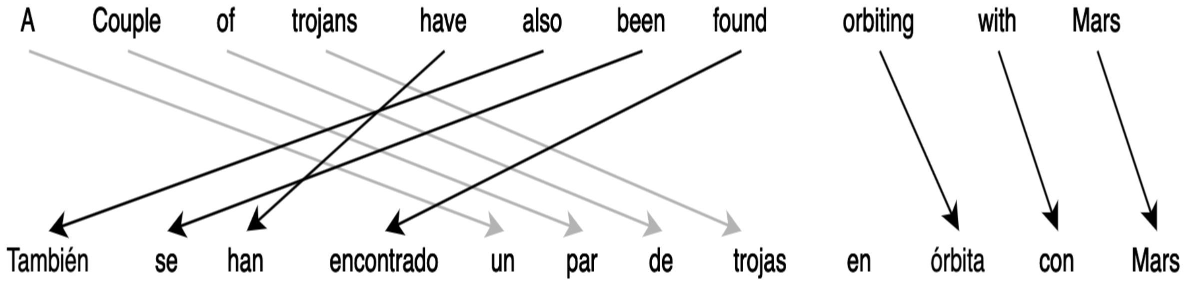


Figure 4: Equivalent English and Spanish sentence with corresponding word alignments between them

| English Phrases | Spanish Phrases |
|---|---|
| Dutil - Dumas experiment | Experimento Dutil - Dumas |
| Dumas | Dumas |
| experiment | Experimento |
| was promoted | fue promovido |
| .... | .... |

Table 7: Mapped continuous phrases between English (E) and Spanish (S) language sentences from the phrase extract algorithm

| BLEU | ES | PT | ZH | HI | TE |
|---|---|---|---|---|---|
| Translation | 45.2 | 48.4 | 26.8 | 20.5 | 7.0 |
| AAC-Translation | 43.7 | 47.8 | 28.2 | 20.1 | 7.5 |

Table 8: BLEU scores of translation and AAC-translation are similar showing that the performance improvement is because of the added consistency.

**Hyper-parameters**: We list the final hyper-parameters used for training mBART model in Table 9 and mT5 model in Table 10. We don't conduct any grid search and use the default hyperparameters suggested in the respective systems.

**Number of parameters**: mBART has 610 million parameters and mT5-base has 580 million parameters.

| Hyper-parameter | Value |
|---|---|
| Maximum tokens per batch | 1024 |
| Learning Rate | 3e-5 |
| LR Scheduler | Polynomial Decay |
| Warmup Updates | 2500 |
| Dropout | 0.3 |
| Max Updates | 40,000 (for OpenIE) and 1,00,000 (for translation) |

Table 9: mBART hyperparameters

| Hyper-parameter | Value |
|---|---|
| Maximum tokens per batch | 24576 |
| Learning Rate | 0.001 |
| LR Scheduler | Constant |
| Warmup Updates | 0 |
| Dropout | 0.1 |
| Max Updates | 20,000 (for OpenIE) and 1,00,000 (for translation) |

Table 10: mT5 hyperparameters