TransAlign: An Automated Corpus Generation through Cross-Linguistic Data Alignment for Open Information Extraction

Alan Melo and **Bruno Cabral** and **Daniela Barreiro Claro** and **Rerisson Cavalcante** and **Marlo Souza**
FORMAS - Research Center on Data and Natural Language
Federal University of Bahia - Salvador, Bahia - Brazil
{alan.melo, bruno.cabral,dclaro,msouza,}@ufba.br

Abstract

This paper introduces a comprehensive approach to address the limited availability of training data on Open Information Extraction (OpenIE) for underrepresented languages by leveraging datasets from languages with abundant resources. We present TransAlign, a cross-linguistic data alignment framework for translating and aligning OpenIE datasets to target languages using language-specific grammatical rules. We explore this methodology for the Portuguese language, employing LSOIE, a large-scale dataset for supervised Open Information Extraction, AACTRANS+CLP, and CARB datasets. We employed high-quality translation models and hand-crafted alignment rules, based on grammatical information, to ensure that the triples are correctly aligned according to the grammar of Brazilian Portuguese. This process resulted in the generation of 96.067 high-quality triples, which laid the foundation for our Portuguese-specific OpenIE dataset. We trained two models by utilizing this dataset, which achieved 10.53% improvement in F1 scores compared to the existing state-of-the-art systems for the Portuguese language, such as PortNOIE (Cabral et al., 2022), including LLMs models.

1 Introduction

Open Information Extraction (OpenIE) aims to extract structured information from unstructured text, without the need to previously define the nature of the information to be extracted. It enables the development of a wide range of downstream applications such as knowledge base construction, question-answering systems, and text summarization (Banko et al., 2007; Etzioni et al., 2008). Despite significant progress in developing OpenIE systems for English (Angeli et al., 2015; Stanovsky et al., 2018; Ro et al., 2020), a performance gap persists for underrepresented languages due to the lack of adequate training data and resources (Akbik et al., 2019).

Recently, cross-lingual transfer learning approaches (Conneau and Lample, 2019; Pires et al., 2019) have emerged as promising strategies to overcome the challenge of limited training data in underrepresented languages. With the substantial advances in machine translation models (Vaswani et al., 2018; Lewis et al., 2020), it is now feasible to utilize translations as an intermediate step for creating OpenIE datasets in underrepresented languages. To harness the potential of these advancements, we introduce TransAlign, a cross-linguistic data alignment framework, which translates and aligns OpenIE datasets from resource-rich languages to target languages using language-specific alignment rules based on grammatical information.

In this paper, we demonstrate the effectiveness of our methodology using the Portuguese language, an underrepresented language in terms of available OpenIE resources. We employ LSOIE (Solawetz and Larson, 2019), AACTRANS+CLP (Kolluru et al., 2022b), and Carb (Bhardwaj et al., 2019), all of which are comprehensive datasets for supervised Open Information Extraction in English, as the foundation for our approach. By integrating high-quality translation models and language-specific alignment rules, we generate a new Portuguese dataset comprising 96.067 high-quality triples suitable for training a Portuguese-specific OpenIE system.

By training a new model on this generated dataset, we achieve significant improvements in the performance of OpenIE methods for the Portuguese language. The model we developed is competitive with actual state-of-the-art systems, such as PortNOIE (Cabral et al., 2022), exhibiting a 10.53% increase in F1 scores. This work highlights the potential of large-scale datasets and translation tools in promoting supervised OpenIE research for underrepresented languages, thereby contributing to developing more inclusive and robust NLP applications.

The paper is organized as follows: Section 2 provides an overview of the relevant work in OpenIE, cross-lingual transfer learning, and machine translation. Section 3 elaborates on our proposed TransAlign framework, detailing the process of translating and aligning the English dataset to Portuguese. Section 4 discusses our experimental setup, results, and an analysis of our model's performance. Section 5 concludes and suggests directions for future research.

2 Related Work

Most existing OpenIE systems have been primarily designed for the English language, which can benefit from extensive resources available for English, including annotated corpora and pre-trained models. However, there has been a growing interest in developing OpenIE systems for other languages, especially with the advent of multilingual models like Multilingual BERT (Devlin et al., 2018).

Fariqui et al. (Faruqui and Kumar, 2015) proposed a cross-lingual annotation projection method for language-independent relation extraction. Their approach involves translating a sentence from a source language to English, performing relation extraction in English, and then projecting the relation phrase back to the source language sentence. Zhang et al. (Zhang et al., 2017) introduced a semi-supervised cross-lingual method that takes a Chinese sentence as input and produces predicate-argument structures in English. CrossOIE (B.S. et al., 2020) created a cross-lingual classifier that utilized contextual embeddings to determine the extraction's validity. Multi2OIE (Ro et al., 2020) employed M-BERT for feature embedding and predicate extraction and used multi-head attention blocks for argument extraction, creating extractors for multiple languages, including English, Portuguese, and Spanish.

However, the development of neural OpenIE systems for Portuguese has been relatively slow due to the scarcity of resources for training. The first deep learning extractor for Portuguese, Multi2OIE (Ro et al., 2020), was developed based on an English dataset that was automatically translated into Portuguese. Following this, PortNOIE (Cabral et al., 2022) proposed a neural framework for Portuguese OpenIE combining rich contextual word representation with neural encoders to process OpenIE as a sequence labeling problem. Despite these advancements, the development of neural OpenIE systems for Portuguese and other languages remains a challenging task due to the need for large-scale annotated corpora and pre-trained models.

A significant contribution to multilingual OpenIE is the work of Kolluru et al. (Kolluru et al., 2022b), who introduced the Alignment-Augmented Consistent Translation (AACTrans) model. This model translates English sentences and their corresponding extractions consistently with each other, ensuring no changes to vocabulary or semantic meaning that may result from independent translations. Using the data generated with AACTRANS, they trained a novel two-stage generative OpenIE model, Gen2OIE, which outputs for each sentence 1) relations in the first stage and 2) all extractions containing the relation in the second stage. Their work demonstrated significant improvements in OpenIE performance across five languages, outperforming prior systems by 6-25% in F1 scores. This approach of automated data conversion can handle even low-resource languages, making it a valuable reference for our work. However, such an approach identifies potential inefficiencies in the translation process. For instance, a single word in one language may translate into two words with identical meanings in another language. Moreover, considering the contextual and cultural differences between languages, such issues may increase.

A straightforward option for translating one dataset into another language would involve using a translation system to translate the original sentence and the extractions directly. However, this method has its drawbacks. The translation introduces words in the extractions that are absent in the translated sentence. This word could be incorrect because it is translated without the surrounding context, altering its meaning. Alternatively, it could be a correct translation, but use a word not present in the translated sentence. Figure 1 illustrates this.

A direct translation using a commercial system (Google, 2023) of the original sentence with the extraction creates an extraction where the English word ”dominated” was translated to ”dominado”. In contrast, the complete translated sentence shifted to another tense, ”dominou”. This inconsistency can pose a problem for methods that rely on sequence labeling to generate the extractions, as the OpenIE extraction may contain words not present in the original sentence.

Our proposed technique, TransAlign, deviates from existing methods by concentrating on translating an OpenIE dataset, followed by a data alignment process. TransAlign tackles the challenge of maintaining the annotation features of sentences during translation by translating the complete sentence with a new sentence composed of concatenated extraction parts. The extraction is reconstructed using heuristics based on Part-of-speech and syntactical dependency information to find the best matching extraction.

English Sentence	The Dutch Empire dominated Maldives for four months
English Extraction	ARG0 = The Dutch Empire REL = dominated ARG1 = Maldives
Portuguese Translation	O Império Holandês dominou as Maldivas por quatro meses
Direct Translation	ARG0 = O Império Holandês REL = **dominado** ARG1= Maldivas
TransAlign Extraction	ARG0 = O Império Holandês REL = **dominou** ARG1= **as** Maldivas

Table 1: Example of translation from English to Portuguese

3 cross-linguistic Data Alignment for OpenIE

cross-linguistic data transfer involves converting datasets from a language abundant in resources, such as English, to a language with limited resources, for instance, Portuguese. The primary challenge lies in preserving the subtleties and meanings of the original dataset while accommodating the linguistic and cultural differences between the two languages. The translation process can introduce inconsistencies and data loss, potentially degrading the quality of the translated dataset. Our approach to mitigate these issues combines translation and alignment methods tailored explicitly for the OpenIE task in the target language.

Our approach aligns with heuristics for Portuguese, but the methodology can be transposed to other languages. The goal is to overcome the constraints imposed by the scarcity of training data for OpenIE in most resource-limited languages, thereby expanding the quality and range of Natural Language Processing (NLP) applications for languages beyond English. Portuguese was chosen as the target language due to its underrepresentation (Claro et al., 2019) in Open Information Extraction (OpenIE) research. The lack of resources and training data for Portuguese has inhibited the progress of neural OpenIE systems for this language. To tackle this, we introduced TransAlign, a cross-linguistic data alignment framework that translates and aligns OpenIE datasets from resource-rich languages, like English, to Portuguese.

Our major strengths lie in its versatility. Our approach is not limited to Portuguese but can be transposed to other languages. The prerequisites are a translator and a set of Part-of-Speech and dependency tree rules specific to the target language. The translator converts the dataset from the source language to the target language, while the set of rules assists in accurately aligning the translated data, preserving subtleties and meanings of the source dataset, accommodating the linguistic and cultural differences between the languages.

3.1 TransAlign

TransAlign begins with translating an existing OpenIE dataset from the source language, in this case, English, to the target language, Portuguese, followed by a data alignment process. The translation process can often yield unusable data due to its ineffectiveness. During translation, a single word in one language may be translated into two words with the same meaning in another language. The process may also encounter contextual and cultural incompatibilities between languages.

For example, consider the sentence "models use an idea or numbers." with the argument structure arg0: models, rel: use, arg1: idea or numbers. A direct translation using Google Translate would yield the Portuguese sentence "modelos usam uma ideia ou números." with the argument structure arg0: modelos, rel: usar, arg1: ideia ou números. This example illustrates the potential inconsistencies that can arise during translation.

Our first attempt to create an OpenIE dataset was solely translating the sentences and extractions. We translated the QA-SRL (He et al., 2015) dataset into Portuguese for creating a new OpenIE dataset based on the methodology proposed by Stanovsky et al. (Stanovsky and Dagan, 2016). It resulted in significant noise and data loss, yielding only a small number of high-quality extractions.

Our TransAlign concerns two main steps: translation and alignment. In the translation step, both the original sentence and the extraction parts are translated from the source language to the target language. In the alignment step, the translated extraction parts are reconstructed into a new extraction that aligns with the translated sentence. This reconstruction is guided by a set of heuristics based on Part-of-Speech tags and syntactical dependency information. These heuristics help to identify the best matching extraction in the target language, ensuring the preservation of the original extraction's semantic meaning.

3.1.1 Alignment Process

The alignment process is divided into three stages:

1. The extraction and sentence are tokenized. Then,

all possible subsequences in the extraction are iterated over. For each subsequence, the extraction is divided into *arg0*, *relation*, and *arg1*. The subsequence is then aligned with the sentence tokens.

2. For each alignment, it is checked whether it is valid. If it is, the *POS* and *DEP* tags of the subsequence are gathered. The relation is then divided into start, middle, and end.
3. The beginning of the relation is checked. It is valid if it begins with an adverb. The first token of the middle is a verb or auxiliary, or it begins with a pronoun, and the first token of the middle is a verb or auxiliary, or it begins with an auxiliary, or it begins with a verb, and the dependency tag is ’ROOT’. The middle of the relation is checked. It is considered valid if all its tokens belong to one of the categories: adjective, noun, verb, auxiliary, determiner, pronoun, subordinating conjunction, or proper noun. The end of the relation is checked. It is considered valid if the relation contains only two tokens and the last token is a verb, auxiliary, or adposition, or if the relation contains more than two tokens and the last token is an adposition, verb, or auxiliary. If the start, middle, and end of the relation are all valid, the alignment is added to the list of alignments.

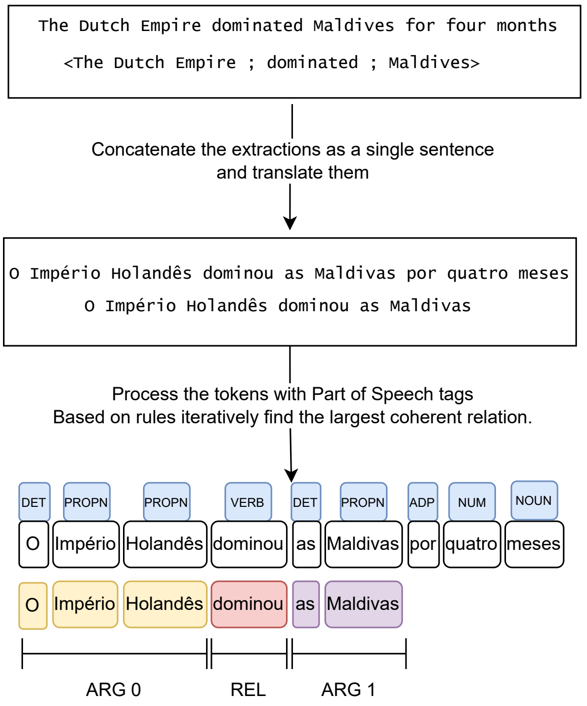

The Dutch Empire dominated Maldives for four months
<The Dutch Empire ; dominated ; Maldives>
Concatenate the extractions as a single sentence and translate them
O Império Holandês dominou as Maldivas por quatro meses
O Império Holandês dominou as Maldivas
Process the tokens with Part of Speech tags
Based on rules iteratively find the largest coherent relation.
DET
PROPN
PROPN
VERB
DET
PROPN
ADP
NUM
NOUN
O
Império
Holandês
dominou
as
Maldivas
por
quatro
meses
O
Império
Holandês
dominou
as
Maldivas
ARG 0
REL
ARG 1

Figure 1: Diagram of the translation and alignment process.

If no valid alignments are found, an empty alignment is added to the list. The function then returns the list of alignments.

Algorithm 1 TransAlign

```
procedure TRANSALIGN($ext$,$sent$)
    Split $ext$ and $sent$ into words
    Process $sent$ using NLP to obtain POS and DEP
    for each subsequence length in $ext$, starting from the longest and decreasing do
        for each subsequence $sub$ in $ext$ do
            Define $arg0$ as words before $sub$ in $ext$
            Define $arg1$ as words after $sub$ in $ext$
            if $sub$, $arg0$, and $arg1$ occur in $sent$ then
                Collect POS and DEP of $arg0$, $sub$, and $arg1$ in $sent$
                Validate the alignment of $sub$ based on POS and DEP
                Initialize flags for start, middle, and end validation
                Analyze Start of $sub$:
                if the first token POS is ‘ADV‘ and next token is ‘VERB‘ or ‘AUX‘ then
                    valid start
                else if the first token POS is ‘ADV‘ and next token is ‘PRON‘ then
                    valid start
                else if the first token POS is ‘PRON‘ and next token is ‘VERB‘ or ‘AUX‘ then
                    valid start
                else if the first token POS is ‘AUX‘ then
                    valid start
                else if the first token POS is ‘VERB‘ and DEP is ‘ROOT‘ then
                    valid start
                else if the first token POS is ‘VERB‘ then
                    valid start
                end if
                Analyze Middle of $sub$:
                for each token in the middle of $sub$ do
                    if token POS is in [‘ADJ‘, ‘NOUN‘, ‘VERB‘, ‘AUX‘, ‘DET‘, ‘PRON‘, ‘SCONJ‘, ‘PROPN‘] then
                        valid middle
                    end if
                end for
                Analyze End of $sub$:
                if the last token POS is ‘VERB‘ and $sub$ has only 2 tokens then
                    valid end
                else if the last token POS is ‘AUX‘ and $sub$ has only 2 tokens then
                    valid end
                else if the last token POS is ‘ADP‘ and $sub$ has only 2 tokens then
                    valid end
                else if $sub$ has more than 2 tokens and last token POS is in [‘ADP‘, ‘VERB‘, ‘AUX‘] and middle is valid then
                    valid end
                end if
                if start, middle, and end are valid then
                    Add $(arg0, sub, arg1)$ to valid alignments
                end if
            end if
        end for
    end for
    if no valid alignment is found then
        Add empty alignment
    end if
    return valid alignments
end procedure
```

This process is implemented in the ***transalign*** function as shown in the Algorithm 1, and illustrated in Figure 1 which takes as input the original extraction and the sentence, and returns a list of valid alignments. The *check_start*, *check_middle*, and *check_end* are omitted for brevity, but it was initially grounded in the principles defining valid relations, as delineated in ReVerb (Fader et al., 2011). The subsequent phase entailed analyzing OpenIE datasets, manually annotated for the Portuguese language, to uncover occurrences and patterns in the relational structure. This examination utilized POS and DEP tagging. Importantly, it is recognized that illustrating all potential patterns for validation is unfeasible, as the alignment does not rely on predefined POS-DEP sequences. Instead, the algorithm dynamically assesses the POS-DEP of tokens within a sequentially generated subsequence. It considers each token about previously validated tokens in the sequence and its position (start, middle, end), employing a permutation-based approach to identify the most viable alignment. This method allows for the identification of an indeterminate array of patterns. Notably, specific POS-DEP configurations such as 'VERB - ROOT' followed by 'ADV - advmod', and POS sequences like 'VERB; VERB; DET' in the relation 'parece estar a', are key to this process. The algorithm particularly focuses on the DEP tag for validating tokens in the 'start' position of a relation. Algorithm refinement was empirically conducted to enhance the encompassment of these detected patterns. This refinement involved aligning the algorithm with the manually annotated datasets and then juxtaposing the resultant and original alignment. Throughout, the emphasis was on manual oversight in the analysis and fine-tuning process, ensuring precision. When multiple candidates match the rules, the most extensive valid alignment is chosen. After selecting the relation, the unified extraction can be realigned, considering all tokens before the first token of the relation as the first argument and all tokens after the last token of the relation as the second argument. Lastly, it is verified whether the first argument is composed of a noun phrase. If so, the triplet is considered valid; otherwise, it is discarded.

3.1.2 Dataset Generation

The datasets employed include LSOIE, CARB, and OIE4. These original datasets in English were translated to Portuguese using translation models. The statistics of the conversion process are summarized in Table 2.

Table 2: TransAlign Conversion Statistics

Dataset	# of Extractions	TransAlign Extractions
LSOIE Train	49.566	15.418
LSOIE Test	10.783	3.365
LSOIE Dev	9.459	2.964
CARB	3.497	745
OIE4 Train	166.032	79.192
OIE4 Valid	1.872	936
Total	231.750	102.620
Total Cleaned	231.750	96.067

The generation of the dataset for our study involved the translation of various existing OpenIE datasets from English to Portuguese. We employed different translation models for this purpose, starting with the Google Translator (Google, 2023), where we translated in the same message the original sentence and the possible extractions. This initial translation process yielded approximately 7,000 valid extractions in the LSOIE dataset, a relatively low number. Most of the errors were because the translated extraction mismatched tokens compared to the translated sentence.

To improve the quality and quantity of valid extractions, we decided to use a larger Language Model. We utilized the GPT-3.5 (OpenAI, 2023). We crafted a prompt designed to guide the GPT-3.5 model in translating not only the sentences but also the specific facts within them. The examples in the prompt served as a blueprint for the model, demonstrating how to accurately translate and adapt the facts to match their representation in the translated sentence. The prompt was iteratively refined based on the model's performance and the quality of the translated extractions. We employed eight examples of translations in the prompt. The sequence from the beginning to the final prompt is described below:

> "Por favor, traduza as seguintes sentenças do inglês para o português. Além disso, identifique e traduza os fatos específicos dentro de cada sentença. Certifique-se de que os fatos traduzidos sejam adaptados para corresponder diretamente à sua representação na sentença traduzida, se baseie nos seguintes exemplos:
>
> EXEMPLOS DE ENTRADA E SAÍDA:
>
> (entrada): SENTENÇA: The dog is walking through the park, it is very happy.
>
> FATO: The dog is very happy.

(saida): SENTENÇA: O cachorro está andando pelo parque, ele está muito feliz.

FATO: O cachorro está muito feliz."

This approach results in a significant increase in the number of valid extractions. Out of the total 69,805 extractions of LSOIE, we obtained 21,747 high-quality valid extractions, significantly more extensive than what we achieved with the Google Translator.

In a nutshell, we started with 231,750 extractions from all datasets. After the translation and alignment process, we obtained 102,620 valid extractions. After a cleaning process to remove duplicates and low-quality extractions, we ended up with a final count of 96,067 high-quality valid extractions. The dataset cleaning process involves assessing the total number of tokens in each extraction, considering the sum of tokens in arg0, rel, and arg1. This sum must be greater than three and less than or equal to 10. Additionally, the POS (Part-of-Speech) of arg0 is scrutinized, where the tokens must strictly possess the POS tags of either 'NOUN' or 'PROPN'. Extractions that do not meet these criteria are categorized as low-quality, while those that conform are deemed high-quality. This dataset represents a significant contribution to the field of OpenIE for Portuguese, providing a valuable resource for future research and development of OpenIE systems for this language. This work, with code and dataset is publicly available at [1].

4 Experiments

4.1 Experimental Design

Our evaluation of quality the generated dataset involved training two distinct models: PortNOIE (Cabral et al., 2022) and Albertina (Rodrigues et al., 2023). PortNOIE, a deep neural network, has purportedly achieved the highest F1 metric result for OpenIE in the Portuguese language. Albertina, on the other hand, is a Large Language Model (LLM) of the BERT family, specifically designed for Portuguese. We also included a comparison with OpenAI GPT-4 (OpenAI, 2023), a commercial LLM. The *temperature* of this model was set to 0.2, while *top_p*, *frequency_penalty*, and *presence_penalty* were all set to 0.

We trained these models using two separate datasets: the dataset created via the TransAlign method, and the Portuguese subset of the AACTRANS+CLP dataset (Kolluru et al., 2022a).

The primary dataset used for performance evaluation was the *PUD 100* dataset(Cabral et al., 2022). This dataset, manually annotated by several academic OpenIE annotators, comprises sentences from news sources and Wikipedia, drawn from the Portuguese section of the Parallel Universal Dependencies corpus (Nivre et al., 2020). It includes 100 sentences and 136 extractions.

To assess the quality of our extractor, we employed precision (P), recall (R), and the F1 measure. We utilized the evaluation code provided by Stanovsky et al.(Stanovsky et al., 2018), which has been widely adopted in subsequent research(Ro et al., 2020; Kolluru et al., 2020). By default, this benchmark uses a scoring method termed **Lexical match**, which deems triples words as a match if they share at least 50% similarity, irrespective of their order.

These metrics were computed by comparing the triples extracted by each model with the gold standard triples in the PUD 100 Dataset. An exact match with a gold standard triple was deemed a match. For partial matches, we adopted a relaxed matching strategy, considering a match if at least two components of the triple (arg1, rel, arg2) corresponded with the gold standard.

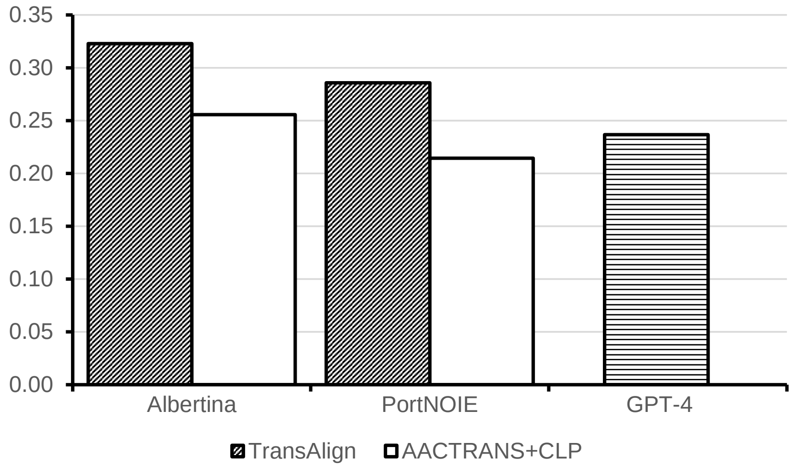
4.2 Experiment Results

The results of our experiments, as presented in Table 3, demonstrate the effectiveness of our proposed TransAlign dataset. The Albertina model, trained on the TransAlign dataset, achieved the highest F1 score of 0.3228, outperforming the same model trained on the AACTRANS+CLP dataset by 6.71 percentage points in the F1 score. This indicates that the TransAlign dataset provides a more effective training ground for the Albertina model, leading to improved performance in OpenIE tasks for the Portuguese language.

Similarly, the PortNOIE model also showed improved performance when trained on the TransAlign dataset, achieving an F1 score of 0.2857, which is 7.15 percentage points higher than when it was trained on the AACTRANS+CLP dataset. This further validates the effectiveness of our TransAlign dataset. However, it's important to note that the PortNOIE model performed slightly better precision when trained on the AACTRANS+CLP dataset. Despite this, it did not result in a higher F1 score due to a lower recall. The most effective dataset for the PortNOIE model remains its original dataset, the PUD 200. When we compared the best performing model in PortNOIE (PUD 200) with the overall best performing model (Albertina with TransAlign), we observed an improvement in the F1 score by 10.53%.

The GPT-4 model, using a 3-shot prompt strategy,

[1] https://github.com/FORMAS/TransAlign



Model	**Dataset**	Precision ↑	Recall ↑	F1 ↑
Albertina	TransAlign	**0.4137**	0.2647	**0.3228**
	AACTRANS+CLP	0.3373	0.2058	0.2557
PortNOIE	TransAlign	0.3783	0.2295	0.2857
	AACTRANS+CLP	0.3913	0.1475	0.2142
	PUD 200	0.3269	0.2615	0.2905
GPT-4	3-shot prompt	0.1980	**0.2941**	0.2366

Table 3: F1 Measures of Different Models for PUD100 dataset

achieved the highest recall of 0.2941 among all models. However, its precision was significantly lower, resulting in an F1 score of 0.2366. This suggests that while the GPT-4 model is capable of identifying a larger number of relevant instances, it also produces a higher number of false positives, thereby reducing its overall effectiveness in OpenIE tasks.

In summary, our experiments demonstrate that the TransAlign dataset generated models more performant than the AACTRANS+CLP for the Portuguese language, as evidenced by the higher F1 scores achieved by both the Albertina and PortNOIE models when trained on this dataset.

4.3 Qualitative experiments

In this section, we dive into a comprehensive qualitative analysis of the TransAlign framework. We will scrutinize examples of both successful and unsuccessful extractions, providing a detailed discussion on each.

4.3.1 Successful Alignments

- **Translated Sentence:** Dr. Smith, por exemplo, é especializado em ecologia.
- **Original Sentence:** Dr. Smith , for example , specializes in ecology .
- **Translated Extraction:** Dr. Smith é especializado em ecologia.
- **Original Extraction:** (ecology; specializes; Dr. Smith)
- **Aligned Extraction:** (Dr. Smith; é especializado em; ecologia)

This extraction is deemed successful due to the accurate identification and alignment of the relation and arguments. The relation ”é especializado em” was correctly translated from ”specializes in”, and the arguments ”Dr. Smith” and ”ecologia” are precisely extracted and translated. It’s noteworthy that despite the original extraction being invalid, we were able to generate a valid extraction, demonstrating the robustness of the TransAlign framework.

- **Translated Sentence:** Ele explica como os seres vivos mudam ao longo do tempo, adaptando-se ao seu ambiente.
- **Original Sentence:** It explains how living things change through time as they adapt to their environment .
- **Translated Extraction:** Os seres vivos mudam ao longo do tempo.
- **Original Extraction:** (living things; change; through time)
- **Aligned Extraction:** (Os seres vivos; mudam a; o longo de o tempo)

This extraction is also deemed successful. The relation “mudam a” was accurately translated from “change”, and the arguments “Os seres vivos” and “o longo de o tempo” are precisely extracted and translated. This example further illustrates the effectiveness of the TransAlign framework in handling complex sentences.

4.3.2 Unsuccessful Alignments

- **Translated Sentence:** O conhecimento científico está sempre mudando porque os cientistas estão sempre fazendo ciência.
- **Original Sentence:** Scientific knowledge keeps changing because scientists are always doing science.
- **Translated Extraction:** O conhecimento científico está mudando porque os cientistas estão sempre fazendo ciência.
- **Original Extraction:** (Scientific knowledge changing; because; scientists are always doing science)
- **Aligned Extraction:** (O conhecimento científico está; mudando porque os cientistas estão; sempre fazendo ciência)

This extraction is unsuccessful due to the incorrect identification and translation of the relation "mudando porque os cientistas estão". The relation should contain ’está’, which is missing in the raw extraction. This example underscores the importance of accurate relation extraction in the overall quality of the alignment.

- **Translated Sentence:** Por exemplo, moinhos de vento eram usados para moer grãos e bombear água.
- **Original Sentence:** For example , windmills were used to grind grain and pump water.
- **Translated Extraction:** Moinhos de vento eram usados para bombear água.
- **Original Extraction:** (windmills; used; pump water)
- **Aligned Extraction:** (Moinhos de vento eram usados para; bombear; água)

This extraction is unsuccessful due to the disproportionate size of *ARG0* compared to *ARG1*. This results in an ’unbalanced’ extraction, which can lead to difficulties in understanding and interpreting the extracted information. This example highlights the need for balanced argument extraction for optimal comprehension and interpretation.

4.4 Trained models comparison

Following the exploration of alignments, we turn our attention to a comparative analysis of the trained models. This section provides a comparison of the performance of the PortNOIE, Albertina, and GPT-4 models, trained on different datasets. The analysis aims to shed light on the strengths and weaknesses of each model, offering insights into their overall effectiveness in OpenIE tasks.

Portuguese Sentence: No início de a semana, Marina, que tinha recentemente retornado de uma conferência em a Suécia, onde conheceu o Dr.

- Albertina(TA) extraction: (Marina; conheceu; o Dr)
- Albertina(ACTRANS+CLP) extraction: (Marina; tinha; de uma conferência em a Suécia)
- PortNOIE(TA) extraction: (Marina; conheceu; o Dr)
- PortNOIE(ACTRANS+CLP) extraction: No Extraction
- GPT-4 extraction: No Extraction

Portuguese Sentence: Mesmo cercada por o burburinho de a cidade moderna, ali, naquele recanto, o tempo parecia ter parado, convidando-a a mergulhar em as páginas de a história.

- Albertina(TA) extraction: (o tempo; parecia ter; parado)
- Albertina(ACTRANS+CLP) extraction: (Mesmo; cercada; por o burburinho de a cidade moderna)
- PortNOIE(TA) extraction: (o tempo; parecia ter; parado)
- PortNOIE(ACTRANS+CLP) extraction: (o tempo; parecia ter; parado)
- GPT-4 extractions: (o tempo; parecia ter parado; naquele recanto) and (o tempo; convidando-a a mergulhar; em as páginas de a história)

When analyzing intricate sentences, it’s important to note certain characteristics of Portuguese grammar that make these sentences complex. For instance, the sentence structure can be complicated by the inclusion of subordinate and adjectival clauses, the use of relative pronouns, and also by the combination of different tenses and moods. These elements can increase the ambiguity and complexity of the sentences.

In the context above, the sentence from the first example contains a subordinate adjectival clause that provides additional information about Marina. In second example, the conjunction "mesmo" (even or although) initiates a concessive adverbial subordinate clause, indicating a contrast or opposing idea.

Considering such grammatical characteristics, it is noticeable that the model trained with the ACTRANS+CLP dataset faces challenges in extracting relationships clearly and accurately in complex sentences. On the other hand, the model trained with the TransAlign dataset demonstrated superior performance, achieving more precise and valid extractions.

In comparison, PortNOIE and GPT-4 models showed varying levels of success. The PortNOIE model was able to extract valid relations in some instances, but failed in others. The GPT-4 model, on the other hand, showed a unique ability to extract multiple valid relations from a single sentence, demonstrating its potential for handling complex sentences. However, it also failed to extract any relations in some cases, indicating areas for improvement.

Limitations

This method has certain limitations due to its annotation rules. It only allows for extractions that include two arguments and a relationship. The samples must strictly follow the *ARG0*, *REL*, *ARG1* annotation sequence, and no elements within each label can be interrupted by tokens with a different label. This requirement limits the variety of extraction structures, excluding formats like *ARG1*, *REL*, *ARG0*, or just *ARG0*, *REL*. The method also doesn't support extractions with more than two arguments, which could improve accuracy in large sentence extractions. For example, it doesn't support the ARG0, REL0, ARG1, REL1, ARG2, REL2, ARG3 label sequence. As a result, many extractions with different label combinations were ignored, mainly because these extraction types were not present in the validation dataset. Another limitation is the potential loss of data, depending on the complexity and quality of the source language data.

5 Conclusion

In this work, we developed a cross-linguistic data alignment methodology, TransAlign, that translates and aligns OpenIE datasets from resource-rich languages to target languages, offering a significant contribution to the field of OpenIE for underrepresented languages. Focusing specifically on the Portuguese language, we successfully converted extensive English OpenIE datasets into high-quality Portuguese OpenIE datasets.

Our approach of employing high-quality translation models in tandem with a set of alignment rules, guided by linguistic and grammatical considerations, has shown promise in managing translation complexities. The methodology has demonstrated its efficacy by generating 96.067 high-quality triples, which substantially enriched our Portuguese-specific OpenIE dataset.

On utilizing this dataset, we trained two models and observed a significant improvement in F1 scores, surpassing the previous state-of-the-art systems by 10.53%. These encouraging results reflect the efficacy and potential of our methodology and have led us to envision its application in other underrepresented languages.

In essence, our study has established that the judicious use of large-scale datasets, efficient translation tools, and well-devised alignment rules can enhance supervised OpenIE for underrepresented languages. As the field progresses, we envisage the potential of our methodology in contributing to more inclusive and effective NLP applications.

Future research directions could aim at refining the alignment rules and optimizing the translation process. Exploring mechanisms to retain more original data and improving alignment heuristics to accommodate varying grammatical structures and constructions in different languages could also be worthy endeavours.

Acknowledgments

This material is partially based upon work supported by the FAPESB under grant INCITE PIE0002/2022. This material is partially supported by the FAPESB TIC 0002/2015. This material is partially based upon work supported by CAPES Financial code 001.

References

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 724–728. Association for Computational Linguistics.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. 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 344–354. Association for Computational Linguistics.

Michele Banko, Michael Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical intelligence*, pages 2670–2676. University of Washington.

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam Mausam. 2019. CaRB: A crowdsourced benchmark for open IE. 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 6262–6267, Hong Kong, China. Association for Computational Linguistics.

Cabral B.S., Glauber R., Souza M., and Claro D.B. 2020. Crossoie: Cross-lingual classifier for open information extraction. In Aluísio S. Quaresma P., Vieira R., editor, *Computational Processing of the Portuguese Language (PROPOR 2020)*, volume 12037 of *Lecture Notes in Computer Science*, pages 201–213. Springer, Cham.

Bruno Cabral, Marlo Souza, and Daniela Barreiro Claro. 2022. Portnoie: A neural framework for open information extraction for the portuguese language. In *Computational Processing of the Portuguese Language*, pages 243–255, Cham. Springer International Publishing.

D.B. Claro, M. Souza, C. Castellã Xavier, and L. Oliveira. 2019. Multilingual open information extraction: Challenges and opportunities. *Information*, 10(7):228.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. Facebook AI Research, Sorbonne Universités, Université Le Mans.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual open relation extraction using cross-lingual projection. *arXiv preprint arXiv:1503.06450*.

Google. 2023. Google translate. Accessed: October 20, 2023.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Soumen Chakrabarti, et al. 2020. Openie6: Iterative grid labeling and coordination analysis for open information extraction. *arXiv preprint arXiv:2010.03147*.

Keshav Kolluru, Muqeeth Mohammed, Shubham Mittal, Soumen Chakrabarti, and Mausam . 2022a. Alignment-augmented consistent translation for multilingual open information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2502–2517, Dublin, Ireland. Association for Computational Linguistics.

Keshav Kolluru, Muqeeth Mohammed, Shubham Mittal, Soumen Chakrabarti, et al. 2022b. Alignment-augmented consistent translation for multilingual open information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2502–2517.

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

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

OpenAI. 2023. Chatgpt. Large language model, accessed on May 1, 2023.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Youngbin Ro, Yukyung Lee, and Pilsung Kang. 2020. Multiˆ2OIE: Multilingual open information extraction based on multi-head attention with BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1107–1117, Online. Association for Computational Linguistics.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt-*.

Jacob Solawetz and Stefan Larson. 2019. LSOIE: A large-scale dataset for supervised open information extraction. *arXiv preprint arXiv:2101.11177*.

Gabriel Stanovsky and Ido Dagan. 2016. Creating a large benchmark for open information extraction. pages 2300–2305.

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, Volume 1 (Long Papers)*, pages 885–895. Association for Computational Linguistics.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416.

Sheng Zhang, Kevin Duh, and Benjamin Van Durme. 2017. Mt/ie: Cross-lingual open information extraction with neural sequence-to-sequence models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 64–70.