Benchmarking Portuguese Open Information Extraction

Gabriel Silva[1234], **Mário Rodrigues**[1345], **António Teixeira**[1234], **Marlene Amorim**[467]
IEETA[1], DETI[2], LASI[3], UA[4], ESTGA[5], DEGEIT[6], GOVCOPP[7]
University of Aveiro
grsilva@ua.pt, mjfr@ua.pt, ajst@ua.pt, mamorim@ua.pt

Abstract

Open Information Extraction (OIE) has seen significant advancements for English, but progress in Portuguese has been hindered by a lack of resources such as Datasets and standardized evaluation benchmarks. This work addresses this critical gap by establishing a systematic and reproducible benchmark for Portuguese OIE systems. We conduct a comprehensive evaluation of eight systems, spanning a decade of research and encompassing both rule-based and neural architectures. The performance of these systems is measured against three distinct Portuguese corpora (WIKI200, CETEN200, and Gamalho) using the established CaRB methodology. Our results reveal that no single system excels across all three datasets. Rule-based models perform strongly on general text (WIKI200, CETEN200) but falter on specialized corpora (Gamalho), while neural systems demonstrate more consistent but not superior performance. With overall F1 scores averaging around 40%, our findings confirm that Portuguese OIE remains a largely unsolved task. This benchmark provides a baseline for future research and highlights the need for a high-quality, manually annotated gold-standard dataset to drive meaningful progress in the field. The evaluation benchmark/framework is made publicly available at https://github.com/gabrielrsilva11/PT-OIE-Benchmark.

Keywords: Open Information Extraction, Portuguese Open Information Extraction, Evaluation Benchmark

1. Introduction

Open Information Extraction (OIE) has become a foundational task in Natural Language Processing, providing a framework for converting large volumes of unstructured text into a structured, machine-readable format. By extracting relational tuples, typically in the format of (subject; relation; object), and without depending on a predefined schema, OIE functions as a critical enabling technology for various downstream applications, including knowledge base construction, question answering, and semantic search (Pei et al., 2023; Mausam, 2016). However, the effective application and advancement of OIE technologies are closely associated with the linguistic characteristics of the target language and the resources available for each language. While significant progress has been made for English, non-English languages such as Portuguese are often left behind due to having fewer resources and funding making it a challenging task to create new solutions (Souza et al., 2025) and compare them with the already existing solutions as there is no standard baseline. For example, the English language has, at least, two well established datasets and evaluation benchmarks that allow researchers to compare themselves with the state-of-the-art: OIE2016 (Stanovsky and Dagan, 2016) and CaRB (Bhardwaj et al., 2019). However, no such thing exists for the Portuguese language.

This work aims to address this existing gap by conducting a systematic comparison of OIE systems and standardizing the evaluation process of Portuguese OIE. It should foster reproducibility and provide the community with a comprehensive understanding of the strengths and weaknesses of current systems thereby facilitating coordinated progress in the field. The primary contribution is not the introduction of a new dataset but rather the development of a robust and reproducible methodology for benchmarking OIE systems utilizing three established corpora: WIKI200 (de Oliveira et al., 2017), CETEN200 (de Oliveira et al., 2017), and Gamalho (Gamallo and Garcia, 2015) using the established CaRB approach. We consider thirteen different systems spanning the last decade, of which we evaluate eight. The evaluation framework is available at: https://github.com/gabrielrsilva11/PT-OIE-Benchmark

2. Related Work

Open Information Extraction (OpenIE) was first introduced in 2007 with the TextRunner (Yates et al., 2007) system, designed to extract a wide variety of relations from text without relying on a predefined set of categories. The evolution of this field can be categorized into three main eras: pre-neural models, neural networks, and Large Language Models (LLMs) with progress often measured using standard benchmarks.

The initial era of pre-neural models emphasized rule-based and statistical methods utilizing linguistic techniques. Notable systems from this period include ReVerb (Fader et al., 2011), which used hand-written patterns to identify relations, and ClausIE (Del Corro and Gemulla, 2013), which first identified clauses within a sentence to extract in-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7692–7700
11–16 May 2026.

formation. Although these approaches were innovative, they often yielded incoherent results due to their dependence on surface-level text patterns and a lack of adaptability to different contexts.

A significant paradigm shift occurred with deep learning and foundational models moving the field toward supervised learning. This new era reframed OIE as a sequence labeling task, where each word is tagged with its role in a potential relation. The introduction of pre-trained models such as BERT (Devlin et al., 2019) substantially improved this approach, resulting in more sophisticated systems. Notable examples include OpenIE6 (Kolluru et al., 2020), in 2020, which utilized BERT and self-attention mechanisms.

The most recent developments are driven by LLMs. These models have shifted the approach from traditional training to prompting, where the model can be instructed to perform extractions with zero-shot (no examples) or few-shot (a few examples) learning (Pai et al., 2024). This has also introduced the broader concept of Universal Information Extraction (Lu et al., 2022), which aims to handle various extraction tasks using flexible, natural language-based schemas.

The development of OIE has largely followed global trends, albeit at a slower pace. The progression began with rule-based systems adapted from their English counterparts, such as RePort (Pereira and Pinheiro, 2015), InferPortOIE (Sena and Claro, 2019), and the dependency-based DptOIE (Oliveira et al., 2023). Subsequently, the creation of neural models like PortNOIE (Cabral et al., 2022). Recently, initial research has explored the application of LLMs for Portuguese OIE. However, literature reviews and studies highlight significant challenges, including a limited availability of large, high-quality, manually annotated datasets compared to English (Souza et al., 2025). This scarcity of resources, combined with the linguistic complexity of Portuguese, makes the creation of robust tools more difficult.

To evaluate and compare the performance of these evolving systems, the field relies on several benchmarks (both datasets and evaluation methods). For English we have the widely used ones such as OIE2016 (Stanovsky and Dagan, 2016), Re-OIE2016 (Zhan and Zhao, 2020), CaRB (Bhardwaj et al., 2019) or even WiRe-57 (Lechelle et al., 2019). For non-English languages we have BenchIE (Gashteovski et al., 2022) which provides a benchmark for German and Chinese in addition to English. Recent works include, for example, a benchmark from 2023 (Mishra et al., 2023) aimed at evaluating the Open Information Extraction performance of several different systems on Indic languages, ScandEval (Nielsen, 2023) which evaluates the performance of over 100 models on a variety of Natural Language Processing (NLP) tasks for the Nordic languages.

Dataset	Size	Annotation Type	Variant
Gamalho	103 Sentences	Human Confirmation	Pt-Br
CETEN200	200 Sentences	Human Confirmation	Pt-Br
WIKI200	200 Sentences	Human Confirmation	Pt-Br
IBERLEF	25 Sentences	Manual	Mixed
OIEC-PT	300 Sentences	Manual	Pt-Br
TransAlign	2000 Sentences	Automatic	Pt-Br

Table 1: Datasets considered and their respective annotations and sizes.

3. Datasets and Systems

In this section, we present the datasets and systems evaluated, along with the rationale for their inclusion or exclusion. For each dataset and system considered, we provide a summary of their approach to the OIE problem.

3.1. Datasets

To perform a comprehensive evaluation of Portuguese OIE systems it is important to select datasets from different domains with different sentence structures as it is important to assess the generalization of systems across both. We started off by surveying the existing datasets. The considered datasets can be seen in Table 1.

The primary criteria for selecting datasets were:

Annotation type: Prefer datasets which were either manually annotated (described as *Manual*) or manual verification/validation of automatic annotations by a certain system (referred to as *Human*).

Availability: The datasets had to be free and available to download

Domain: A selection of datasets from different domains to assess the generalization across different styles

Variant: Ideally the benchmark should cover the European (Pt-Pt) and Brazilian (Pt-Br) variants of Portuguese.

Size: The dataset should have a minimum of 100 sentences.

From the six datasets reviewed, we selected three for our final evaluation framework: Gamalho (Gamallo and Garcia, 2015), CETEN200 (de Oliveira et al., 2017), and WIKI200 (de Oliveira et al., 2017). This selection was based on their alignment with our key criteria. All three datasets involve human participation in their annotation processes, as each dataset incorporates human validation of the extracted triples,

Dataset	Domain	Sentences	Extractions
Gamalho	Scientific	103	358
CETEN200	News	202	366
WIKI200	Wikipedia	201	427

Table 2: Sentences and extractions per dataset selected for the benchmark

thereby enhancing their reliability for evaluation. Moreover, these datasets are publicly available and have different domains, showcasing diverse writting styles and sentence structures. WIKI200 consists of random sentences from Wikipedia articles, CETEN200 is derived from a news corpus, and the Gamalho dataset focuses on ecological issues, featuring domain-specific sentences.

The decision to exclude three datasets was based on their partial alignment with our established criteria. The IBERLEF dataset (Collovini et al., 2019), while manually annotated, was excluded due to its small size (25 sentences) and significant domain overlap with the larger WIKI200 and CETEN200 corpora. The TransAlign dataset (Rios et al., 2024), despite its large size, was automatically generated, extracting only one triple per sentence, and lacked human validation of the extractions. The OIEC-PT (Souza et al., 2025) dataset was inaccessible, as the link provided in the associated paper was inactive. Although we attempted to contact the authors, we did not receive the dataset, preventing us from including it in this benchmark.

The final benchmark consists of 506 sentences and a total of 1151 extractions. Table 2 presents a breakdown by dataset.

3.2. Systems

Similarly to the datasets, we began our search for systems to test by surveying the available options. Two criteria were established for the system search:

- The systems were freely available (an executable, GitHub repository, etc..).
- The systems were made for Portuguese or multi-lingual but were tested on Portuguese.

In total, our search resulted in 13 systems up for evaluation, these systems are shown on Table 3. These systems can be further divided into two distinct subsets: rule-based and neural-based.

Starting with rule-based systems, these systems primarily rely on hand-crafted linguistic rules, patterns, and constraints to extract relational triples from text. Among these is RePort (Pereira and Pinheiro, 2015), a system modeled after the English system, (Fader et al., 2011). Its process involves detailed morphological and syntactic analysis, followed by the application of syntactic and lexical constraints to identify and filter relational

System	Year	Language
ArgOE	2015	Multi
RePort	2015	Portuguese
InferPortOIE	2017	Portuguese
DependenteIE	2017	Portuguese
PragmaticOIE	2018	Portuguese
DptOIE	2018	Portuguese
CrossOIE	2020	Multi
Multi^{2}Oie	2020	Multi
PortNOIE	2022	Multi
DetIE	2022	Multi
TransAlign	2023	Multi
PortOIE-Llama	2024	Portuguese*
GraphIE	2024	Multi

Table 3: Systems that were considered for evaluation, their respective language and release year.

phrases before a final set of rules identifies the arguments. Another system, ArgOE (Gamallo and Garcia, 2015), takes a language-independent approach by using dependency parsing. Its core is a small set of universal, hand-crafted rules that map common dependency relations to the components of a relational triple, allowing for easy adaptation to new languages. This system was later incorporated into Linguakit (Gamallo et al., 2018) which is the version considered for this Benchmark.

Similarly, DependenteIE (de Oliveira et al., 2017) uses a set of manually crafted, linguistically-inspired rules that map dependency relations to a relational triple's components. It traverses the dependency tree using a Depth-First Search algorithm, with rules specifically tailored to handle the grammatical nuances of the Portuguese language. DptOIE (Oliveira et al., 2023) also begins with a preprocessing step involving a tokenizer, a Part-of-Speech (POS) tagger, and a Dependency Parser. At its core are hand-crafted rules designed specifically for Portuguese grammar, which are applied as the system traverses the dependency tree with a Depth-First Search algorithm to identify the constituents of a triple.

Other systems in this category focus on inference and context. InferPortOIE (Sena and Claro, 2019) uses a baseline OpenIE system to extract explicit facts and then employs a Support Vector Machine classifier to determine if a sentence contains a transitive or symmetric relationship. Based on the classification, it applies one of two sets of hand-crafted rules to infer new relationships. Following a similar theme, PragmaticOIE (Sena and Claro, 2020) also builds upon the ReVerb (Fader et al., 2011) system with three main modules: an inferential module with transitivity and symmetry rules, a contextual module to extract conditional facts, and an intentional module to analyze condi-

tional verb tenses and infer outcomes.

Moving to neural-based systems, PortNOIE (Cabral et al., 2022) uses a modular architecture that gathers rich linguistic information, including BERT embeddings. It generates multiple extraction candidates from a single sentence, which are then processed by an encoder-decoder model to form the final relational triple. Another system, CrossOIE (Cabral et al., 2020), uses pretrained multilingual word embedding models like M-BERT and XLM to create vector representations of a sentence and a candidate triple. These are then fed into a neural classifier that scores the validity of the extraction, enabling cross-lingual transfer learning. Multi^{2}OIE (Ro et al., 2020) combines BERT with multi-head attention blocks in a two-step process. It first identifies all possible predicates and then extracts arguments for each one. By using multilingual BERT, this method can be applied to languages without specific training data. DetIE (Vasilkovsky et al., 2022) utilizes a Transformer-based, encoder-only architecture that treats OIE as a direct set prediction problem. It predicts a set of token masks for the subject, relation, and object, using an order-agnostic loss function that allows for significantly faster inference times.

Several systems leverage the power of Large Language Models (LLMs). TransAlign (Rios et al., 2024) approach involves translating an English OpenIE dataset to Portuguese using models like GPT-3.5. A crucial alignment step then uses language-specific rules to ensure the translated triples are grammatically correct, and the resulting dataset is used to train new neural OpenIE models. In a more direct application of LLMs, PortOIE-Llama (Cabral et al., 2024) was developed by fine-tuning the Llama 2 7B model specifically for the task using a Portuguese OpenIE dataset. Finally, GraphOIE (Silva et al., 2024) takes a novel approach by first converting text into a knowledge graph enriched with syntactic information. It uses a GraphSAGE model to identify candidate words for the triple's components and then queries a Large Language Model (LLM) with these candidates to form the most suitable final extractions.

Of these thirteen systems, RePort, DependenteIE and InferPortOIE could not be tested as there was no way provided to run them. Of the remaining ten, there were an additional two (PortNOIE and CrossOIE) which had a repository and instructions provided, however, could not be run either due to dependency or hardware problems. The remaining eight models were tested on the datasets that were previously defined.

Table 4 lists the final systems selected for evaluation and the resources they provide. PortOIE-LLama is tagged as Portuguese due to the fine-tune being made for Portuguese OIE, however, as it is an LLM it should be capable of extractions in other languages.

System	Instructions	Training Required	Language
PragmaticOIE	No	No (Rule-Based)	Portuguese
DptOIE	Yes	No (Rule-Based)	Portuguese
ArgOE	Yes	No (Rule-Based)	Multi
PortOIE-LLama	No	No (Weights Provided)	Portuguese*
Multi^{2}OIE	Yes	No (Weights Provided)	Multi
DetIE	Yes	No (Weights Provided)	Multi
GraphIE	Yes	No (Weights Provided)	Multi
TransAlign	Yes	Yes	Multi

Table 4: Systems to be evaluated and their respective out-of-the-box ease-of-use.

4. Evaluation Methodology

Our benchmark approach follows the widely recognized CaRB (Bhardwaj et al., 2019) methodology, which utilizes a token-level, tuple-based matching system, in contrast to the OIE2016 (Stanovsky and Dagan, 2016) approach that focuses solely on lexical matching.

This methodology ensures that relations are matched with corresponding relations and arguments with their respective arguments, avoiding any conflation between the two. A table is constructed to represent all possible pairings between gold extractions and predicted extractions, from which we compute recall and precision for each pair.

Recall is calculated based on multi-match criteria, where the highest recall score for each gold extraction is utilized. In contrast, Precision is derived from single-match criteria, wherein gold extractions and system extractions are matched from best to worst. The advantages of this approach are highlighted by the original authors. In terms of recall, it prevents penalizing systems that produce longer extractions which might match several gold extractions. For precision, it penalizes systems that generate multiple redundant extractions.

However, the CaRB approach has limitations, as it assumes that gold extractions are as atomic as possible, which is not always the case, and it relies on lexical matching, penalizing relations that may have equivalent meanings expressed in different wording.

For systems requiring training (TransAlign) or prompting (PortOIE-LLama, GraphIE) we utilized the default configurations provided, or the settings specified in their respective publications. The evaluation involved executing each system once for each dataset and collecting the produced extractions. Most systems supported file input, in these cases, the input consisted of a file in which each line represented a sentence. For systems that did not permit file input and instead operated on a single sentence basis, we developed a script that pro-

cessed the file line by line, executing the command for each line and saving the output.

Since each system produced a different output format, we decided to standardize all outputs into a consistent format: a CSV file with the format `Sentence, ARG0, V, ARG1`. Additionally, we converted the datasets into this same format to facilitate the comparison between gold standard data and predictions.

For each of the 8 systems we measured the number of triples extracted and their performance (Recall, Precision, F1).

5. Results

All systems successfully processed all three datasets of the benchmark, with the exception of DetIE, which failed to produce results for the Gamalho dataset.

The number of triples extracted per system, for the 3 datasets, is illustrated in Figure 1 while the benchmark results are presented in Table 5.

		WIKI200	CETEN200	Gamalho	Average
PragmaticOIE	**Precision**	74.2	72.7	37.0	61.3
	Recall	**82.0**	**80.4**	42.6	**68.3**
	F1	**77.9**	**76.3**	39.6	**64.6**
DptOIE	**Precision**	23.5	19.7	0	14.4
	Recall	47.6	39.5	0	29.0
	F1	31.4	26.3	0	19.2
TransAlign	**Precision**	**77.6**	**65.1**	**71.5**	**71.4**
	Recall	40.2	37.1	35.2	37.5
	F1	53.0	47.3	47.2	49.1
PortOIE-Llama	**Precision**	9.4	27.7	39.5	25.6
	Recall	53.3	35.9	25.3	38.2
	F1	16.0	31.3	30.8	26.1
Multi^{2}Oie	**Precision**	52.5	46.2	55.1	51.3
	Recall	50.7	49.7	**52.3**	50.9
	F1	51.6	47.9	**53.7**	51.0
DetIE	**Precision**	53.1	44	50.4	49.1
	Recall	51.0	47.6	50.5	49.7
	F1	52.0	45.7	50.4	49.4
GraphIE	**Precision**	35.6	35.1	44.9	38.5
	Recall	37.0	33.6	37.4	36.0
	F1	36.3	34.3	40.8	37.1
ArgOE	**Precision**	26.5	34.4	34.6	31.8
	Recall	19.6	40.0	18.1	25.8
	F1	22.5	37.0	23.7	27.7
Average	**Precision**	44.0	43.1	41.6	42.9
	Recall	47.7	45.4	32.7	41.9
	F1	47.7	43.3	35.8	40.5

Table 5: Evaluation results for all the systems on the Benchmark.

On average, the number of extractions closely approximates the gold standard for the WIKI200 and CETEN200 datasets. The systems achieve an average of 3 and 1.9 extractions per sentence, respectively, while the gold standard reports 2.1 and 1.8 extractions. However, in the Gamalho dataset there is quite a disparity in the number of extractions made between the systems and the gold, with the systems having 1.6 extractions per

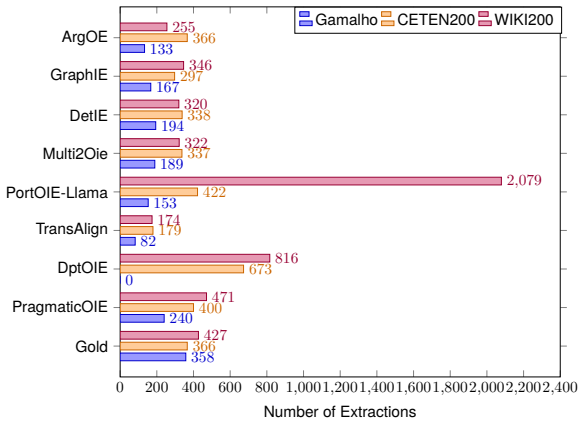

Figure 1: Number of triples extracted per system across the benchmark datasets

sentence while the gold has 3.4. In Table 5 we can also see that PragmaticOIE, with scores of 77.9% and 76.3%, outperformed other systems on the WIKI200 and CETEN200 datasets by a considerable margin. Conversely, on the Gamalho dataset, Multi^{2}OIE achieves the best performance with an F1 score of 53.7%. Overall the best performing model is PragmaticOIE with an F1 of 64.6%. The worst systems were DptOIE (which did not process the Gamalho dataset) and ArgOE with average F1 scores of 19.2% and 27.7%.

A more comprehensive view of the results for each system is presented in Figure 2. Although the best results might indicate that WIKI200 and CETEN200 are easier datasets compared to Gamalho, this figure indicates that the median F1 and Precision scores for each dataset are relatively comparable. However, the Recall metric highlights the challenges associated with the Gamalho dataset. This implies that Gamalho is a more difficult dataset than the other two, potentially due to its more specific (Scientific versus News/Wikipedia corpus).

Taking a further look at the results and the systems themselves, the top-performing system employs a rule-based approach developed using two data sources: Wikipedia and CetenFOLHA [1]. These are the two datasets where this system exhibits the highest performance. However, on the Gamalho dataset, its performance declines significantly (from 77.9% and 76.3% to 39.6%), indicating the necessity of transitioning from rigid rule-based methods to more flexible neural approaches. TransAlign achieved the highest precision among the systems evaluated; however, it is restricted to performing only one extraction per sentence. PortOIE-LLama was an outlier on the number of extractions it performed on the WIKI200 dataset, looking at the output the reason was clear: on one

[1] https://www.linguateca.pt/cetenfolha/

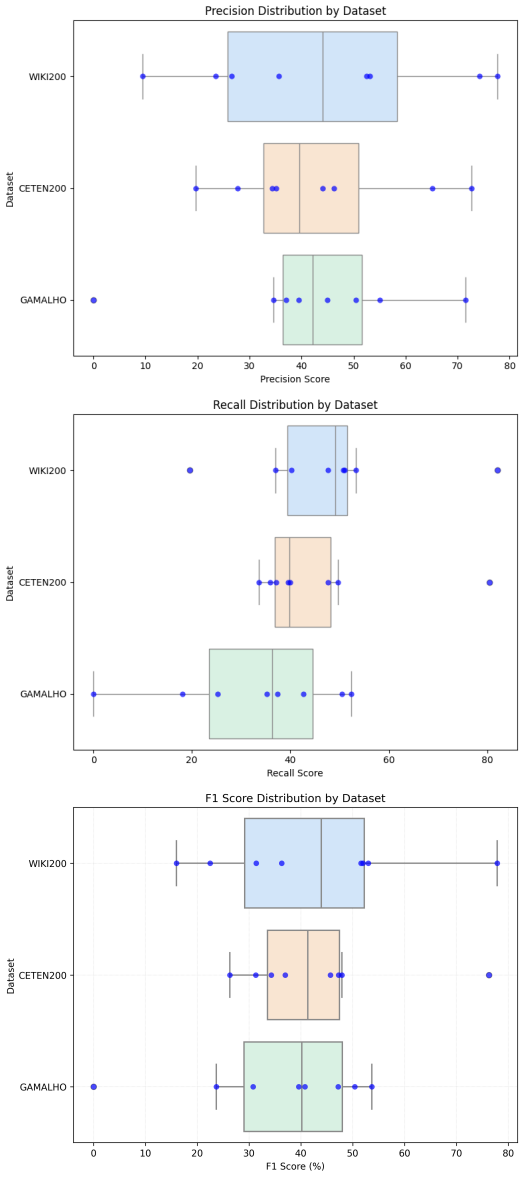

Figure 2: Performance metrics on each dataset

of the sentences the system got stuck in a loop outputting repeated extractions only stopping when the maximum context length was reached.

Overall, these results establish a baseline for each dataset, with overall metrics of Precision, Recall, and F1 consistently around 40%. To surpass the best systems, the F1 scores to achieve are 77.9% on WIKI200 and 76.3% on CETEN200 by PragmaticOIE, and 53.7% on the Gamalho dataset by Multi^{2}OIE.

6. Discussion

The goal of creating a standard, comprehensible and reproducible framework for benchmarking Portuguese OIE systems was achieved. We present a benchmark for eight different systems evaluated on three datasets from different domains using this framework while making it freely available for adoption. Having this standardized evaluation is an important step in comparing different systems and understanding the current situation with Portuguese OIE. Researchers and developers can now adopt this framework, much like the English systems evaluate on CaRB and OIE2016, to compare themselves with the state-of-the-art.

Original Sentence	Escrevi meu livro contra essa idéia de que o cânone morreu, de que Shakespeare não interessa mais porque é um europeu macho, branco e morto .
Translated Sentence	I wrote my book against this idea that the canon is dead, that Shakespeare is no longer interesting because he is a dead, white, male European.

Extracted Triples	(Shakespeare, não interessa mais porque, é) (mais porque, é, europeu) (um europeu, é, mais porque)
Translated Triples	(Shakespeare, is no longer interesting because, he is) (no longer, he is, European) (European, he is, because)

Table 6: Example of Sentence with wrong Gold extractions from the CETEN200 dataset.

However, there are still limitations to the framework. The datasets used in this benchmark are not perfect themselves. In Table 6 we can see an example from the CETEN200 dataset that is invalid. This example would penalize valid extractions from systems that managed to correctly identify the correct triples for this sentence. For example, for this sentence, the GraphIE system extracted the following triples:

- Original: `(Shakespeare, é, europeu macho) (Shakespeare, é, branco) (Shakespeare, é, morto)`
- Translated: `(Shakespeare, is, male European) (Shakespeare, is, white) (Shakespeare, is, dead)`

While the majority of the sentences are correct, having examples like these can hurt the training and validation of these systems. The need for a human-annotated, or more carefully selected, dataset is evident.

Another limitation is the evaluation metrics, as the CaRB dataset is designed for atomic triple extractions, where each extracted fact is indivisible. This design influences the evaluation metrics, particularly the calculation of Precision. However, for the datasets used here, that is not the case. Several extractions in these datasets could be further refined to achieve indivisibility. Furthermore, relying on exact lexical matching is arguably more punishing for Portuguese. English is characterized by a relatively rigid syntactic structure and minimal inflection, making surface-level token matching a reliable proxy for accuracy. In contrast, Portuguese exhibits the high linguistic complexity of a morphologically rich language, featuring flexible word order and extensive inflectional morphology (Oliveira

et al., 2023). Because Portuguese relies heavily on complex verb conjugations, gender and number agreement across noun phrases, and the intricate placement of clitics and reflexive pronouns, there are vastly more valid ways to express the exact same relational tuple than in English. Moreover, Portuguese is a language where subjects are frequently omitted from the surface text. Systems capable of deep semantic analysis might successfully infer and reconstruct these implicit subjects, yet they would be penalized by lexical evaluation metrics simply because the recovered tokens do not strictly match the annotated surface text (Sena and Claro, 2020). Consequently, a system might extract a semantically perfect tuple that is scored as a failure due to minor morphological or syntactical variations. This intrinsic linguistic variance exacerbates the limitations of token-based evaluation, highlighting the urgent need for semantically-aware metrics tailored to the unique complexities of the Portuguese language (Souza et al., 2025).

In general, for future work, new evaluation methods for OIE should appear with a focus on the meaning of the triple itself instead of token matching to ensure correctness. There are several ways a triple can be formulated while having the same meaning which is not reflected in the usual metrics used. The adaptation of these datasets to align with a CaRB-style format is also a focus for our future work. However, this task demands significant time and the involvement of multiple annotators to ensure the accuracy of the corpus.

7. Conclusion

This paper addresses a significant gap in Portuguese NLP by establishing a systematic and reproducible benchmark for OIE systems. By evaluating eight distinct systems, both rule and neural-based, across three different established datasets (WIKI200, CETEN200 and Gamalho) using the CaRB methodology we provide a baseline that will, hopefully, foster standardized comparisons and guide future research.

Our findings show that the field of Portuguese OIE is far from solved with no single system demonstrating superior performance across all domains. While the systems have evolved throughout the years, it remains an unsolved problem. There is a lack of annotated data on which most of these systems need to be trained, as well as a reliance on rules despite the trend being neural-based models and, more recently, language models. The rule-based system PragmaticOIE achieved the highest scores on the WIKI200 and CETEN200 datasets, however, its performance declined significantly on the more specialized Gamalho dataset. Neural-based systems exhibited a more stable performance across different datasets which highlights their flexibility and adaptability to different contexts. The average F1 scores are around 40% leaving room for substantial improvement. Improving the efficiency of these systems would be beneficial for several different downstream tasks. Given this lack of resources the direction which these systems should take seems to be a multi-lingual approach but with a focus on tuning them for Portuguese and its different variants (Pt-Pt, Pt-Br).

This benchmarking process also showed limitations within the Portuguese OIE resources. We identified inconsistencies in the datasets and their extractions, which complicates evaluation and penalizes some systems. The lack of datasets for the European Portuguese variant is also very present as five out of the six datasets were based on the Brazilian variant and the remaining one was mixed. The reliance on token-based metrics, while valuable for standardization, fail to capture meaning a key aspect of information extractions.

For future work we identify two important paths moving forward: the creation of a high-quality, manually annotated gold standard corpus for Portuguese OIE designed with the CaRB principles and the need for the broader OIE community to create new evaluation metrics that move beyond lexical/token/tuple matching and into the broader concept of meaning.

8. Acknowledgements

This work was funded by the PhD grant UI/BD/153571/2022
This work was supported by the Foundation for Science and Technology (FCT) through contract doi.org/10.54499/UID/00127/2025.

9. Bibliographical References

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 6263–6268, Hong Kong, China. Association for Computational Linguistics.

Bruno Cabral, Daniela Claro, and Marlo Souza. 2024. Exploring open information extraction for portuguese using large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 127–136.

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

Bruno Souza Cabral, Rafael Glauber, Marlo Souza, and Daniela Barreiro Claro. 2020. Crossoie: Cross-lingual classifier for open information extraction. In *International conference on computational processing of the Portuguese language*, pages 368–378. Springer.

Sandra Collovini, Joaquim Francisco Santos Neto, Bernardo Scapini Consoli, Juliano Terra, Renata Vieira, Paulo Quaresma, Marlo Souza, Daniela Barreiro Claro, and Rafael Glauber. 2019. Iberlef 2019 portuguese named entity recognition and relation extraction tasks. In *IberLEF@ SEPLN*, pages 390–410.

Leandro Souza de Oliveira, Rafael Glauber, and Daniela Barreiro Claro. 2017. Dependentie: An open information extraction system on portuguese by a dependence analysis. *Encontro Nacional de Inteligência Artificial e Computacional*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '11)*, Edinburgh, Scotland, UK.

Pablo Gamallo and Marcos Garcia. 2015. Multilingual open information extraction. In *Progress in Artificial Intelligence*, pages 711–722, Cham. Springer International Publishing.

Pablo Gamallo, Marcos Garcia, César Piñeiro, Rodrigo Martinez-Castaño, and Juan C. Pichel. 2018. Linguakit: A big data-based multilingual tool for linguistic analysis and information extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244.

Kiril Gashteovski, Mingying Yu, Bhushan Kotnis, Carolin Lawrence, Mathias Niepert, and Goran Glavaš. 2022. BenchIE: A framework for multi-faceted fact-based open information extraction evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4472–4490, Dublin, Ireland. Association for Computational Linguistics.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020. 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)*, pages 3748–3761, Online. Association for Computational Linguistics.

William Lechelle, Fabrizio Gotti, and Phillippe Langlais. 2019. WiRe57 : A fine-grained benchmark for open information extraction. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 6–15, Florence, Italy. Association for Computational Linguistics.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Mausam Mausam. 2016. Open information extraction systems and downstream applications. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 4074–4077. AAAI Press.

Ritwik Mishra, Simranjeet Singh, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Pushpak Bhattacharyya. 2023. IndIE: A multilingual open information extraction tool for Indic languages. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 312–326, Nusa Dua, Bali. Association for Computational Linguistics.

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Leandro Oliveira, Daniela Barreiro Claro, and Marlo Souza. 2023. Dptoie: a portuguese open information extraction based on dependency analysis. *Artificial Intelligence Review*, 56(7):7015–7046.

Liu Pai, Wenyang Gao, Wenjie Dong, Lin Ai, Ziwei Gong, Songfang Huang, Li Zongsheng, Ehsan Hoque, Julia Hirschberg, and Yue Zhang. 2024.

A survey on open information extraction from rule-based model to large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9586–9608, Miami, Florida, USA. Association for Computational Linguistics.

Kevin Pei, Ishan Jindal, Kevin Chen-Chuan Chang, ChengXiang Zhai, and Yunyao Li. 2023. When to use what: An in-depth comparative empirical analysis of OpenIE systems for downstream applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 929–949, Toronto, Canada. Association for Computational Linguistics.

Victor Pereira and Vládia Pinheiro. 2015. Reportum sistema de extração de informações aberta para língua portuguesa. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 191–200. SBC.

Alan Rios, Bruno Cabral, Daniela Claro, Rerisson Cavalcante, and Marlo Souza. 2024. Transalign: an automated corpus generation through cross-linguistic data alignment for open information extraction. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 196–206.

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*, pages 1107–1117, Online. Association for Computational Linguistics.

Cleiton Fernando Lima Sena and Daniela Barreiro Claro. 2019. Inferportoie: A portuguese open information extraction system with inferences. *Natural Language Engineering*, 25(2):287–306.

Cleiton Fernando Lima Sena and Daniela Barreiro Claro. 2020. Pragmaticoie: a pragmatic open information extraction for portuguese language. *Knowledge and Information Systems*, 62(9):3811–3836.

Gabriel Silva, Mário Rodrigues, António Teixeira, and Marlene Amorim. 2024. Advancing open information extraction for portuguese by leveraging graph structures and large language models. In *Proc. IberSPEECH 2024*, pages 61–65.

Marlo Souza, Bruno Cabral, Daniela Claro, and Lais Salvador. 2025. Challenges in expanding portuguese resources: A view from open information extraction.

Gabriel Stanovsky and Ido Dagan. 2016. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2300–2305.

Michael Vasilkovsky, Anton Alekseev, Valentin Malykh, Ilya Shenbin, Elena Tutubalina, Dmitriy Salikhov, Mikhail Stepnov, Andrey Chertok, and Sergey Nikolenko. 2022. Detie: Multilingual open information extraction inspired by object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11412–11420.

Chenguang Wang, Xiao Liu, and Dawn Song. 2022. IELM: An open information extraction benchmark for pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8417–8437, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexander Yates, Michele Banko, Matthew Broadhead, Michael Cafarella, Oren Etzioni, and Stephen Soderland. 2007. TextRunner: Open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 25–26, Rochester, New York, USA. Association for Computational Linguistics.

Junlang Zhan and Hai Zhao. 2020. Span model for open information extraction on accurate corpus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9523–9530.