RelEx-PT: A Portuguese Sentence-Level Relation Extraction Dataset

Tomás Pinto, Catarina Silva, Hugo Gonçalo Oliveira
University of Coimbra, CISUC/LASI, DEI, Coimbra, Portugal
{tomaspinto, catarina, hroliv}@dei.uc.pt

Abstract

We introduce RelEx-PT, a new sentence-level Relation Extraction dataset for Portuguese. Addressing the scarcity of high-quality, controlled resources for the language, RelEx-PT provides a balanced benchmark comprising 18 Wikidata-derived relation types across diverse domains. The dataset is built through a distant supervision pipeline that links Wikidata triples with Portuguese Wikipedia sentences and enhanced by a Natural Language Inference (NLI)-based filtering process, combining scalability with quality assurance. Additionally, we conduct baseline experiments to evaluate the dataset's applicability across diverse extraction settings, including Relation Classification (RC), Relation Triple Extraction, and Open Information Extraction. These experiments leverage both prompting and fine-tuning strategies using Large Language Models. The results show that RelEx-PT effectively supports a range of extraction paradigms, yielding high performance in RC and competitive results in structured triple generation, while also highlighting key challenges in open-ended extraction.

Keywords: Relation Extraction, Portuguese Language, Dataset Creation, Language Resources

1. Introduction

Relation Extraction (RE) is a fundamental task in Information Extraction (IE) and, more broadly, in Natural Language Processing (NLP), aimed at identifying semantic relationships between entities mentioned in text (Zhao et al., 2024). By transforming unstructured text into structured relational knowledge, RE supports a wide range of downstream applications, including Knowledge Graph Construction, Question Answering, Semantic Search, and Information Retrieval (Detroja et al., 2023). The success of these applications, however, depends heavily on the availability of high-quality annotated datasets, which serve as the foundation for training and evaluating extraction models.

Most advances in RE have focused on English, driven by the abundance of large, well-annotated resources such as TACRED (Zhang et al., 2017) and DocRED (Yao et al., 2019a). In contrast, Portuguese remains less resourced: existing datasets are scarce, often domain-specific (Pavanelli et al., 2023), imbalanced (Seganti et al., 2021), or designed for cross-sentence contexts (Cabot et al., 2023), limiting their use for controlled experimentation. Moreover, many of these datasets rely heavily on manual annotation efforts (Freitas et al., 2009), and while it remains the gold standard for dataset quality, it quickly becomes time-consuming, costly, and impractical to scale (Diaz-Garcia and Lopez, 2025). Automated methods like distant supervision and translation offer scalability but often introduce noise, highlighting the need for robust quality-control mechanisms (Hedderich et al., 2021).

A gap in Portuguese RE is the absence of sentence-level, controlled datasets, similar to widely adopted English benchmarks such as CoNLL04 (Roth and Yih, 2004) or SemEval-2010 Task 8 (Hendrickx et al., 2010). Such datasets play a crucial role as simpler, cleaner environments for studying RE methods before scaling to more complex, cross-sentence or document-level settings.

To address this, we present RelEx-PT, a balanced, sentence-level RE dataset for Portuguese. It covers 18 Wikidata[1]-derived relation types, created through a distant supervision pipeline aligning Wikidata triples with Portuguese Wikipedia[2] sentences and refined via a Natural Language Inference (NLI)-based filtering stage to remove noisy alignments. This design combines automation and precision, ensuring both scalability and quality.

To demonstrate the dataset's applicability, we conduct experiments across three complementary tasks: Relation Classification (RC), Relation Triple Extraction (RTE), and Open Information Extraction (OpenIE). In RC, models classify the relation between two predefined entities within a sentence into one of the 18 relation types defined in the dataset. In RTE, the task is extended to the generation of full triples, while in OpenIE, we remove the schema constraint, allowing models to extract triples with freely generated relations that are later mapped back to the schema for evaluation. Although the dataset was not originally designed for OpenIE, this setting serves to show its broader applicability. We utilize Large Language Models (LLMs) via both prompting and fine-tuning, establishing baselines for future research.

Our main contributions are as follows:

- A controlled sentence-level dataset covering Wikidata-derived relation types, with balanced

[1] https://www.wikidata.org/
[2] https://pt.wikipedia.org/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7681–7691
11–16 May 2026.

class distribution and explicit entities, providing a benchmark for Portuguese RE research.

- A publicly released pipeline[3] combining distant supervision with NLI–based filtering to ensure automation and reduce noise, which can be used to further extend the dataset or even create new ones.
- Baseline experiments illustrating RelEx-PT's use across RC, structured triple extraction (RTE), and open-ended extraction (OpenIE) tasks with diverse LLMs.

The remainder of this paper is organised as follows: Section 2 reviews related work on RE datasets and construction methodologies. Section 3 presents the dataset design and specifications. Section 4 details the creation pipeline. Section 5 reports the experiments, and Section 6 concludes with future directions.

2. Related Work

RE has long relied on high-quality datasets to support the development and evaluation of extraction approaches. Early efforts focused on manually annotated datasets, such as ACE-2005[4] and SemEval-2010 Task 8 (Hendrickx et al., 2010), which offered high precision and carefully controlled annotation guidelines. These resources provided strong benchmarks but were limited in scalability due to the time and cost required for manual annotation. Later, distant supervision approaches (Mintz et al., 2009) emerged as a practical alternative, automatically aligning knowledge base triples (e.g., from Freebase or Wikidata) with textual mentions in large corpora such as Wikipedia or news collections. This method enabled large-scale data creation but introduced substantial label noise, since the co-occurrence of entities does not guarantee the entailment of the intended relation (Diaz-Garcia and Lopez, 2025).

In recent years, the field has expanded toward LLM-assisted dataset generation where LLMs are used to synthesize relation-labeled data (Pandya et al., 2024; Jiang et al., 2024). While this strategy aims to bridge the gap between quality and scalability, hallucinations generated by LLMs still pose a significant challenge to the reliability of the resulting data.

From a structural perspective, existing RE datasets differ in their unit of annotation, which strongly influences the amount of context and noise involved (Zhao et al., 2024). Sentence-level datasets, like the widely used CoNLL04 (Roth and Yih, 2004), focus on identifying relations within a single, self-contained sentence, typically expressing one, or at most a few, triples. This setup provides a more controlled evaluation scenario, with limited contextual ambiguity. In contrast, cross-sentence datasets, such as DocRED (Yao et al., 2019b) or WikiReading (Hewlett et al., 2016), operate on multi-sentence texts, where relations often span across sentences or depend on broader discourse context. While this richer setup better reflects real-world complexity, it also introduces higher variability, annotation uncertainty, and noise, making the task more challenging for both annotation and automatic extraction.

While English remains the dominant language for RE research, there has been increasing interest in developing multilingual and non-English datasets (Diaz-Garcia and Lopez, 2025). Efforts such as SMiLER (Seganti et al., 2021) and SRED^{FM} (Cabot et al., 2023) are multilingual datasets that also include Portuguese among their covered languages, facilitating research beyond English. SMiLER contains a large Portuguese corpus covering 22 relation types, with an uneven distribution across them. It was created semi-automatically from Wikipedia and DBpedia[5] using distant supervision, with part of the data manually verified by linguists to correct alignment errors. SRED^{FM} is even larger, covering 400 relation, and is also highly unbalanced. It follows a distant supervision approach between Wikipedia and Wikidata but adopts a cross-sentence structure with multiple triples. Unlike SMiLER, SRED^{FM} relies on automatic filtering based on an NLI model to remove noisy annotations, a strategy that we too adopt in our dataset construction pipeline. Translation-based approaches have also been explored, such as in mLAMA (Kassner et al., 2021), where English RE datasets are automatically translated into other languages, although this often introduces alignment errors and semantic drift.

Focusing solely on Portuguese RE, a few dedicated resources have been developed. ReRelEM (Freitas et al., 2009) provides manually annotated document-level relations between named entities, initially covering 12 texts and 4 main relation types, which were later refined and expanded into 20 subtypes. Bete (Pavanelli et al., 2023) was created for the diabetes healthcare domain through manual annotation, comprising five relation types defined within a specialized medical context. In contrast, DBpediaRelations-PT (Batista et al., 2013) targets a general-domain, sentence-level setting, covering ten relation types obtained via distant supervision between Portuguese Wikipedia and DBpedia in-

[3] The dataset and construction pipeline are available at https://github.com/TomasCCPinto/RelEx-PT.

[4] https://catalog.ldc.upenn.edu/LDC2006T06

[5] https://www.dbpedia.org/

foboxes. A manual review was later conducted for a small subset of examples to assess alignment quality and reduce noise.

OpenIE, a related but distinct task that extracts relations without predefined schemas, has seen more work done for Portuguese. Resources such as TransAlign PTOIE (Rios et al., 2024), derived from English OpenIE datasets through translations, and manually annotated corpora like PUD100, PUD200 (Cabral et al., 2022) and Pragmatic (Sena and Claro, 2020), have advanced open extraction in Portuguese. However, these datasets focus on unconstrained relation patterns, differing from schema-based RE resources.

In this context, RelEx-PT fills a key gap by introducing a balanced, sentence-level RE dataset for Portuguese, offering a controlled resource comparable in design and purpose to widely used English benchmarks such as CoNLL04 and TACRED. It combines the scalability of distant supervision with the precision of NLI-based filtering, yielding cleaner instances.

3. Dataset Design and Specification

Each instance in RelEx-PT corresponds to a single Portuguese sentence, annotated with two mentioned entities and one relation between them. In other words, in this sentence-level scope, every instance maps directly to one ‹subject, relation, object› triple. Figure 1 shows example instances from RelEx-PT, presented exactly in the format in which it is made available. By deliberately restricting the context to a single sentence, the dataset maintains low contextual complexity, making it a controlled setting to test and analyze RE methods.

The dataset is fully Portuguese-native, with sentences, entities, and relation labels in Portuguese, filling a gap in resources that often rely on English or translated material. In terms of coverage, RelEx-PT includes 18 relation types derived from Wikidata, spanning themes such as biographical, cultural, organizational, locational, and taxonomic relations. A full list of these relation types (in English) is presented in Table 1. Their selection reflects both empirical considerations, as these were among the most frequent relations in Wikidata and thus more accessible for retrieval through the pipeline, and deliberate design choices to ensure diversity and broad semantic coverage. Further details on these aspects are discussed as we progress through the subsequent sections.

A distinctive feature of RelEx-PT is its perfect class balance. Each relation type is equally represented, which avoids the skewness that commonly affects RE datasets and allows fairer comparisons across relations. In total, the dataset comprises 1,800 instances, divided into a training set of 1,260 examples (70 per relation type) and a test set of 540 examples (30 per relation type). This split provides consistent training conditions while ensuring that the evaluation set is sufficiently representative.

Relation Types	
author (P50)	located in or next to body of water (P206)
composer (P86)	occupation (P106)
country (P17)	place of birth (P19)
director (P57)	production company (P272)
developer (P178)	religion or worldview (P140)
father (P22)	sport (P641)
genre (P136)	subclass of (P279)
taxon rank (P105)	headquarters location (P159)
instance of (P31)	main subject (P921)

Table 1: Relation types covered on the RelEx-PT dataset (based on Wikidata properties).

Wikipedia textual content is distributed under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license. Wikidata, in contrast, is released under the Creative Commons CC0 1.0 Universal Public Domain Dedication, which permits unrestricted reuse without attribution requirements. Because RelEx-PT incorporates Wikipedia, the dataset as a whole is released under the CC BY-SA 4.0 license, allowing it to be freely used, shared, and adapted for research and other purposes, provided that appropriate attribution is given and that any derivative works are distributed under the same license terms. The construction pipeline and source code are publicly available on GitHub[3], and the dataset is also released on Hugging Face[6].

4. Creation Methodology

This section details the methodology followed to construct the RelEx-PT dataset, outlining each major phase of the pipeline, which is summarized in Figure 2. We describe the design decisions made at each step and conclude with a discussion of the main challenges and limitations encountered during the creation process.

4.1. Triple-Sentence Alignment

The first step in the construction process was to align Wikidata triples with sentences from Portuguese Wikipedia. We began by downloading a full dump of Wikidata triples, which amounts to several gigabytes of data. To make this manageable, we trimmed the file by discarding descriptions and other metadata, keeping only the entity and relation identifiers (Qids and Pids). An optional filtering step was also applied at this stage to retain only a subset of properties from the hundreds that the triples dump contains.

[6] https://huggingface.co/datasets/NLP-CISUC/RelEx-PT

```jsonl
{"sentence": "Francesinha é uma sanduíche originária da cidade do Porto, em Portugal.", "subject": "Francesinha", "relation": "país", "object": "Portugal", "nli_prediction": 1, "nli_probabilities": [0.00010203252168139443, 0.9998979568481445]}
{"sentence": "Divinity II é um jogo de RPG de ação desenvolvido pela Larian Studios, utilizando o motor Gamebryo, para Windows e Xbox 360.", "subject": "Divinity II", "relation": "desenvolvedor", "object": "Larian Studios", "nli_prediction": 1, "nli_probabilities": [7.409470981656341e-06, 0.9999926090240479]}
{"sentence": "A cláusula arbitral ou cláusula compromissória é um mecanismo utilizado para submeter um contrato à arbitragem.", "subject": "Cláusula arbitral", "relation": "tema(s) principal(is)", "object": "arbitragem", "nli_prediction": 1, "nli_probabilities": [0.0003820341662503779, 0.9996179342269897]}
{"sentence": "Atamante ou Atamas, na mitologia grega, foi um rei de Orcomeno na Beócia, filho de Éolo e Enarete.", "subject": "Atamante", "relation": "pai", "object": "Éolo", "nli_prediction": 1, "nli_probabilities": [1.513753340987023e-05, 0.999984860420227]}
{"sentence": "O Clube Desportivo de Mafra é um clube português, com sede na vila de Mafra, no distrito de Lisboa.", "subject": "Clube Desportivo de Mafra", "relation": "sede", "object": "Mafra", "nli_prediction": 1, "nli_probabilities": [5.0173664931207895e-05, 0.9999498128890991]}
{"sentence": "Odontoforídeos (Odontophoridae) é uma família de aves pertencente à ordem Galliformes, que inclui os urus e a perdiz-da-califórnia, Callipepla californica.", "subject": "Odontoforídeos", "relation": "categoria taxonómica", "object": "família", "nli_prediction": 1, "nli_probabilities": [2.4322429453604855e-05, 0.9999756813049316]}
{"sentence": "Didactica magna (título em latim) ou Didática Magna (título em português), também conhecido por Tratado da Arte Universal de Ensinar Tudo a Todos, é um livro de Comenius publicado em 1649.", "subject": "Didactica magna", "relation": "autor", "object": "Comenius", "nli_prediction": 1, "nli_probabilities": [6.352933996822685e-05, 0.9999364614486694]}
```


Figure 1: Example instances from RelEx-PT represented in the original JSONL format.

With this reduced triple set, we proceeded to link it to textual evidence. For each triple, we used the Wikidata API to translate the identifiers into human-readable labels in Portuguese. These labels were then queried through the Wikipedia API to retrieve the text of the corresponding Portuguese Wikipedia pages. Whenever this lookup failed, either because the IDs could not be translated, because no corresponding page was available, or the page had no text, the triple was discarded.

Having obtained the relevant Wikipedia pages, we searched their content for sentences that contained both entities of a given triple. Whenever such a sentence was found, we extracted it and assumed, following the principle of distant supervision, that the sentence expressed the relation indicated by the triple, moving to another triple. Each successful match therefore became a candidate instance consisting of the original Portuguese sentence, the two entity mentions, and the associated relation label, as shown in Figure 1.

4.2. NLI-Based Filtering

While distant supervision provides a scalable way of aligning triples with sentences, it inevitably introduces noise. Many sentences mention the two entities of a triple without explicitly expressing the intended relation. To mitigate this issue and improve dataset precision, we applied a filtering stage based on NLI, inspired by Cabot et al. (2023).

We employed a DeBERTa-based model fine-tuned on manually annotated Portuguese data for recognizing textual entailment[7] (Chaves Rodrigues, 2023). The model takes a premise and a hypothesis as input, outputting a score indicating the likelihood that the premise entails the hypothesis. In our setup, each sentence extracted in the first step serves as the premise, while the hypothesis is crafted from the corresponding triple. To generate hypotheses, we designed handmade templates for each of the 18 relation types. These templates reformulate a <subject, relation, object> triple into a natural language statement that expresses the relation. For example, a triple with relation *author* would be reformulated into a hypothesis such as “X is the author of Y”, where X and Y are replaced with the entities from the triple. The templates for each relation can be found in the study’s repository.

Each candidate instance was then evaluated by the NLI model, producing an entailment score. This score allowed us to verify whether the sentence truly conveyed the relation expressed by the triple. Instances with high entailment scores were kept as valid examples, while those with low scores were filtered out. We applied a threshold of 0.95, meaning that only instances with scores close to 1 (indicating strong entailment) were retained. Lower scores signaled weaker or absent entailment, and such cases were removed from the dataset. Table 2 presents two examples of instances that did not pass this filtering stage, as no entailment was verified between the sentence and the candidate triple. In this way, NLI-based filtering served as an additional validation layer, confirming the quality of alignments and discarding mismatches introduced during the previous stage.

The filtering process had already eliminated the majority of malformed sentences, including those affected by residual HTML tags or incomplete phras-

[7] https://huggingface.co/ruanchaves/mdeberta-v3-base-assin2-entailment

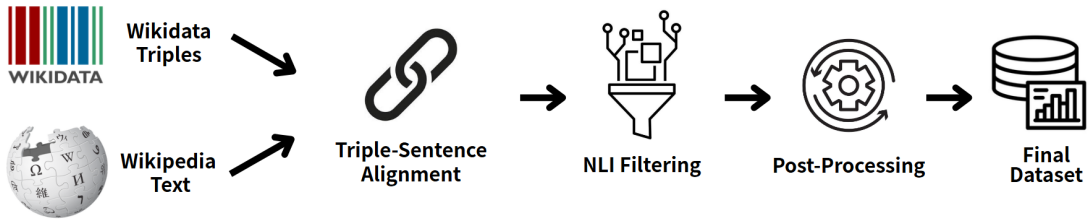

Figure 2: Overview of the RelEx-PT dataset construction pipeline.

ing. The subsequent cleaning stage further refined the data by stripping extra spaces, removing newlines, and discarding duplicates arising from overlapping or redundant triples in the Wikidata dump. Consequently, each relation type is represented by a compact and reliable set of examples.

Sentence	Triple
Apart from these brief references at the end of the Book of Job, Jemimah is not mentioned elsewhere in the Bible.	(Jemimah, father, Job)
EastEnders is a British soap opera that has been broadcast on BBC One since February 19, 1985.	(EastEnders, production company, BBC)

Table 2: Two instances filtered out by the NLI-based Filtering due to non-entailment. Original sentences were translated to English for presentation.

4.3. Quality Control

The creation of RelEx-PT involved multiple filtering and cleaning stages, each substantially reducing the number of candidate instances while improving overall quality. For each relation, 15,000 Wikidata triples were sampled sequentially from the Wikidata dump. On average, only about 11.9% ± 7.7% of these triples could be aligned with a Portuguese Wikipedia sentence containing both entities, making this stage responsible for the largest data reduction. For example, the relation *father* exhibited the sharpest decrease, with only 1.2% of its original triples surviving this step, while *birth place* retained roughly 22.5%. This demonstrates how difficult it is to find Portuguese Wikipedia sentences that explicitly mention both entities from a given triple.

The subsequent NLI-based filtering and cleaning introduced an additional reduction of about 26.2% ± 21.4% on average over the aligned pairs. Although smaller in magnitude, this stage showed considerable variation between relations. For instance, *subclass of* lost about 76.7% of its candidate instances, whereas *director* retained nearly all of them (only 0.37% removed). This indicates that in some relations, triple–sentence alignments exhibit strong entailment almost immediately, while others contain a higher proportion of non-entailing matches, further highlighting the importance of semantic validation through NLI.

Despite the substantial data loss across stages, the combined pipeline proved essential for improving precision and eliminating noisy alignments. Each relation was normalised to an equal number of instances in the final version (100), ensuring a fully balanced experimentation.

To estimate the final precision of RelEx-PT, we manually inspected 100 randomly sampled instances from the final dataset, corresponding to 5–6 instances per relation type. Each instance was evaluated for semantic entailment between the sentence and the associated triple. We identified 2 incorrect instances, resulting in an estimated precision of 98%. The first error corresponded to an entailment mismatch: although the sentence explicitly stated that Anarchist communism is a branch of anarchism, the associated triple encoded it as a subclass of communism, which is not supported by the sentence. The second error originated from the Wikidata triple itself, which encoded marriage as an instance of institution rather than a subclass. While the sentence discusses the general concept of marriage, the *instance of* relation would only be correct for a specific, concrete marriage event. While these errors reveal potential sources of residual noise, their minimal presence (2%) demonstrates that the multi-stage filtering and validation pipeline is highly effective in producing a high-precision dataset.

4.4. Challenges and Limitations

During the creation process, we initially considered a larger set of relation types, with twice the number included in the final dataset. However, exploratory tests revealed that many Wikidata properties were too sparsely represented due to the skewed distribution of properties in Wikidata. As a result, our triple dump was heavily dominated by a small subset of properties, while many others appeared only rarely. This imbalance made it difficult to collect enough triples for a broader range of relations. One possible way to address this limitation in future work is to leverage a much larger portion of the Wikidata dump, which contains billions of triples, far beyond the subset processed in this study.

A limitation of RelEx-PT concerns its non-exhaustive annotation. Each sentence in the dataset is paired with a single ground-truth triple,

even though, in some cases, the sentence could validly express additional relations. As a result, in RE, models may extract other correct triples that are not present in the reference data, which are therefore not considered during evaluation.

Another limitation arises from the dataset's dependence on external APIs. Both Wikidata and Wikipedia were accessed programmatically, which introduces potential fragility to the creation process. Temporary service interruptions, query-rate limits, or excessive API traffic can occasionally affect data retrieval, although these issues are operational rather than methodological and can be mitigated through repeated queries.

5. Experiments

This section presents experiments on RelEx-PT, covering three key settings: RC, RTE, and OpenIE. We showcase tasks for which RelEx-PT can be used, defining baselines and evaluating the performance of LLMs.

5.1. Experimental Setup

In RC, the goal is to classify the semantic relation between two marked entities in a sentence into one of the 18 predefined relation types of RelEx-PT. The task is approached in two ways:

1. Text generation task via prompting using the decoder-based LLMs listed in Table 3, where models directly generate the relation label;

2. Supervised classification task using encoder-only models (BERTimbau Base (Souza et al., 2020) and Albertina PTPT (Santos et al., 2024)) fine-tuned on RelEx-PT's training set. We use a learning rate of 2e-5, 5 epochs, batch sizes of 8 (training) and 16 (evaluation).

Two input variants are considered for the prompting: one including the full sentence (context-aware) and another using only the entity pair (context-free), allowing the evaluation of how model performance depends on contextual information.

Building on the RC experiments, the evaluation is extended to triple extraction, a more complex task that requires generating complete ‹subject, relation, object› triples from text. Two complementary settings are explored: a schema-based formulation (RTE) and a schema-free formulation (OpenIE). In RTE, the goal is to extract full triples where the relation is one of the 18 predefined types of RelEx-PT, extending classification into structured generation. In contrast, OpenIE removes this constraint, allowing free-form relation generation. These open relations are later mapped to the closest schema types for evaluation, reflecting real-world scenarios where relations may not align with fixed ontologies.

Both settings are treated as text generation tasks using prompt-based LLMs. Table 3 lists the models used, along with their number of parameters and versions. Models are prompted to generate triples directly from input sentences, and outputs are evaluated against the dataset's ground-truth triples. To control costs and ensure efficiency, the maximum token limit for generation was set to 100. Given that we are working at the sentence level with relatively limited context, this token limit is considered sufficient for capturing the necessary information.

Model	Parameters	Version
gemma-2-2b-it (Team et al., 2024)	2B	Full
Mistral-7B-Instruct-v0.3 (Jiang et al., 2023)	7B	Full
gemma-2-9b-it (Team et al., 2024)	9B	Full
gemma-3-27b-it (Team et al., 2025)	27B	Q4_K_M
Llama-3.3-70B-Instruct (Meta, 2024)	70B	Q4_K_M
DeepSeek-V3-0324 (Liu et al., 2024)	681B	Full
GPT-4o (Hurst et al., 2024)	N/A	Full
GPT-4o-mini (Hurst et al., 2024)	N/A	Full
GPT-4.1 (OpenAI, 2025)	N/A	Full

Table 3: Decoder-only LLMs for prompt-based inference. Q4_K_M denotes quantized variants.

Additionally, fine-tuning for RTE is performed using the PTT5-Base model, an encoder–decoder model well-suited for sequence-to-sequence generation. The model is fine-tuned using the Hugging Face Transformers library, where the input consists of a simple instruction prompting the extraction of the triple from the sentence, and the output corresponds to the target triple. We employ the same hyperparameters used in RC (learning rate 2e-5, 5 epochs), with smaller batch sizes of 4 (training) and 8 (evaluation), and a generation length of 128.

Since RelEx-PT is not exhaustively annotated, for the triple extraction, the objective is to extract all valid triples expressed in each sentence, avoiding unfair penalties when the model generates alternative but semantically correct outputs.

5.2. Relation Classification

In the RC setting, entities are assumed to be pre-identified using the ground-truth mentions in the dataset. The task focuses solely on classifying their relation among the 18 predefined types, isolating relation reasoning from entity recognition. This setting is also explored in our previous work (Pinto et al., 2025), which focuses specifically on RC.

Two zero-shot prompt formulations were evaluated and can be found in the project's repository. These were inspired by successful designs in prior RC and LLM prompting literature (Zhang et al., 2023; Wei et al., 2024):

- Descriptive: provides the sentence together with the associated subject and object, followed by the list of 18 relation types, each accompanied by a short description. This helps

the model understand the semantic scope of each label and select the relation that best matches the entities in context.

- QA-style: frames the task as a question–answer interaction, asking the model which of the 18 relations best describes the connection between the entities in the given sentence.

To examine the influence of contextual information, these prompts were also adapted to a context-free setting, in which the input sentence is removed and only the two entities are provided. This design probes whether models rely primarily on the sentence context or can infer relations based on pretraining knowledge and the label space information itself.

Given the balanced nature of the data, model performance is evaluated using standard macro-averaged F1-scores, ensuring equal weighting across relation types. Table 4 presents the best macro F1 by model, under both prompting and fine-tuning setups. The fine-tuned encoder models achieved the strongest overall performance, with BERTimbau Base reaching 0.96 F1, confirming the effectiveness of supervised adaptation when sufficient high-quality labeled data is available. On the prompting side, performance was more varied across models, with GPT-4o obtaining the highest results (0.92 F1), closely approaching fine-tuned performance.

Prompting		
Model	**Context-Free**	**Context-Aware**
Gemma2-2B	0.39	0.64
Mistral-7B	0.82	0.80
Gemma2-9B	0.65	0.85
Gemma3-27B	0.85	0.89
Llama3.3-70B	0.81	0.84
DeepSeekV3	0.80	0.86
GPT-4o	**0.92**	**0.92**
GPT-4o-mini	0.76	0.90
GPT-4.1	0.87	**0.92**
Fine-tuning		
BERTimbau Base		**0.96**
Albertina 100M PTPT		0.94

Table 4: Best macro F1-scores achieved by each model on the RelEx-PT dataset for the RC task. Results are shown for both prompting (context-free and context-aware) and fine-tuning setups.

The impact of context proved moderate. Smaller models, such as Gemma2-2B, GPT-4o-mini, and Gemma2-9B, exhibited more pronounced drops in the context-free setting, whereas larger and more capable LLMs like GPT-4o, DeepSeekV3, and Llama3.3 maintained nearly identical scores across both configurations. This suggests that robust LLMs can often infer the correct relation from the entities alone, leveraging internalized world knowledge and learned expectations about how entities are typically related. They also appear to reason more effectively over the label space provided in the prompt, using the semantic cues in the options to identify the most plausible link between the entities even without explicit context.

Performance also varied substantially across relation types. Relations such as *occupation*, *director*, and *country* tended to yield perfect scores, reflecting their clearer lexical and conceptual associations. In contrast, more abstract or hierarchical relations, including *instance of* and *subclass of*, remained notably harder (around 0.70 F1), requiring reasoning over subtler semantic distinctions that are not always explicit in text or easily captured through direct linguistic cues. Relations with well-defined linguistic or factual anchors are more readily detected, while those relying on deeper semantic inference remain challenging.

5.3. Relation Triple Extraction

For the RTE experiments, two prompting strategies were employed, both operating under a schema-based setup constrained to the 18 predefined relation types of RelEx-PT:

- Instruct: A direct prompt asking the model to extract all relations between entities of the sentence, listing triples in the format "(Subject, Relation, Object)".

- Descriptive: Extends the first by adding brief definitions for each relation type, aiming to reduce ambiguity and improve label selection. These definitions clarify the types of entities expected for a triple of each relation.

Performance is evaluated using macro-averaged Precision, Recall, and F1-score. Evaluation operates at the triple level, where a true positive requires an exact match of subject, relation, and object between predicted and ground-truth triples. Predictions with correct subject and object but an incorrect relation are false positives, indicating confusion among relation labels. A false negative occurs when a ground-truth triple is missed, due to incorrect relation, misidentified entities, or no output.

Table 5 presents the scores obtained by each model using the two RTE prompting strategies. Overall, larger and more robust models achieved substantially better performance, reflecting their superior capacity to capture relational structure and produce coherent triples. The best results were obtained with GPT-4.1, reaching 0.77 F1 under the *Descriptive* prompt, followed by the open Gemma3-27B and DeepSeek-V3, both scoring above 0.69 F1. In contrast, smaller models such as Gemma2-2B and Mistral-7B struggled to generate accurate

triples, with low recall and F1 below 0.25, indicating frequent omissions or incomplete extractions.

Prompting						
	Instruct			**Descriptive**		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**
Gemma2-2B	0.59	0.16	0.21	0.52	0.13	0.18
Mistral-7B	0.68	0.11	0.16	0.57	0.11	0.17
Gemma2-9B	0.81	0.22	0.33	0.79	0.30	0.42
Gemma3-27B	0.88	0.52	0.63	0.91	0.62	0.71
Llama3.3-70B	0.90	0.56	0.66	0.89	0.58	0.68
DeepSeekV3	0.88	0.44	0.55	0.93	0.59	0.69
GPT-4o	0.91	0.50	0.63	0.94	0.56	0.67
GPT-4o-mini	0.87	0.42	0.54	0.88	0.53	0.63
GPT-4.1	**0.91**	**0.61**	**0.70**	**0.94**	**0.67**	**0.77**
Fine-tuning						
PTT5-base	0.35	0.30	0.32			

Table 5: Macro-averaged Precision (P), Recall (R), and F1 scores by model and method for RTE on the RelEx-PT dataset.

Across models, the *Descriptive* prompt consistently improved performance over the *Instruct* version, particularly in recall. This suggests that giving models minimal semantic grounding about the relations helps them interpret the schema more accurately and produce better-aligned extractions.

The fine-tuned PTT5 model reached 0.32 F1 with better recall than small LLMs but way below the strongest prompting setups. This may be due to the training data's focus on identifying a single triple per sentence in the RelEx-PT dataset, even if other valid triples can exist. This means that the model sometimes identifies a triple that is stated in the sentence but is not the one in the ground truth annotation, which is then counted as a miss. This contrast highlights the relative advantage of large generative models in complex extraction scenarios, where their capacity for open-ended reasoning and flexible generation allows them to capture multiple relational triples beyond the annotated ground truth.

A similar pattern to the RC task emerges in the RTE results. Relations such as *director*, *country*, and *place of birth* tend to push the overall scores upward (around 0.9 F1), likely due to their clearer lexical cues and well-defined linguistic patterns. In contrast, more abstract relations like *subclass of*, *taxon rank*, and *instance of* remain challenging (around 0.5 F1), as they often depend on implicit contextual or hierarchical reasoning that is not directly expressed in the text.

5.4. Open Information Extraction

A single prompt was used for the OpenIE setting, following the same structure as the *Instruct* RTE prompt. This time, it does not provide the list of relation types, thereby encouraging the model to produce open-ended relational expressions based on its own linguistic and semantic understanding.

As the generated relations may differ lexically from the dataset schema, we implement a mapping procedure to align them with the 18 predefined relation types of RelEx-PT for evaluation. To achieve this, we employ Serafim-900M (Gomes et al., 2024), a transformer-based sentence encoder for Portuguese. The relation component of each extracted triple is encoded into an embedding vector, which is then compared against embeddings of the 18 relation labels from RelEx-PT using cosine similarity. Each relation is assigned to the closest label by highest similarity score, and the resulting mapped triples are then evaluated using the same metrics as in the RTE setting.

Looking at the results in Table 6, we obtain interesting yet unexpected first impressions: despite large differences in model size, performance remains similar across systems, with F1-scores around 0.25–0.30. This consistency highlights significant limitations in the current approach to OpenIE under this setup.

Model	**Precision**	**Recall**	**F1 Score**
Gemma2-2B	0.28	0.07	0.11
Mistral-7B	0.46	0.15	0.21
Gemma2-9B	0.47	0.26	0.30
Gemma3-27B	0.40	0.24	0.27
Llama3.3-70B	0.43	0.27	0.29
DeepSeekV3	0.52	0.24	0.29
GPT-4o	0.60	0.21	0.26
GPT-4o-mini	0.52	0.25	0.29
GPT-4.1	0.45	0.25	0.30

Table 6: Macro-averaged Precision, Recall and F1 scores achieved by each model on the RelEx-PT dataset for the OpenIE setting.

Two main factors appear to underlie these limitations. The first concerns the mapping procedure: many generated relation expressions are valid but differ lexically or semantically from the fixed schema labels, and the embedding alone, without more contextual grounding, often fails to capture the intended correspondence, leading to frequent misclassifications. For instance, for the relation *occupation*, models may generate plausible expressions such as "works as" or "is a" instead of the exact label "occupation", which can align but are not recognized as a match by the embedding-based mapping. The second factor relates to the absence of an explicit label space in the prompt. Without predefined relation options, models lack guidance on which types of connections to prioritize, causing them to overlook relations that are less salient in the sentence and miss relevant entities. For example, if *taxon rank* were included as a possible label, models would be more likely to extract entities such as "species" or "family" linked by that relation. In its absence, they tend to focus on more "visible" entities, like the names of the species or families themselves, missing the intended relation

and, consequently, the complete triple.

Some relations, such as *birth place*, *father*, or *developer*, are typically the main focus of the sentence and have clear lexical realizations that often match Wikidata labels directly, resulting in comparatively higher performance (F1 scores above 0.50). Others, such as *taxon rank*, *main subject*, or *occupation*, are more abstract or expressed in diverse ways, resulting in ineffective mappings and F1 scores often close to zero across models.

6. Conclusion

While English has long benefited from high-quality IE benchmarks, Portuguese remains comparatively less-resourced, with existing datasets often focusing on open or cross-sentence extraction and lacking the controlled, schema-based structure of standard RE benchmarks. To fill this gap, we present RelEx-PT, a new sentence-level RE dataset for Portuguese that offers a quality benchmark covering 18 Wikidata-derived relation types across diverse domains. It is constructed through a pipeline that aligns Wikidata triples with Wikipedia sentences and is refined with an NLI-based filtering stage that ensures semantic validity and minimizes noise.

To demonstrate its applicability, we perform experiments in RC, RTE, and OpenIE using both prompt-based and fine-tuned LLMs, illustrating how the dataset supports different extraction approaches. Results show that the dataset is particularly well-suited for RC, where the impact of context proved moderate, but it can also effectively support the extraction of complete triples, where robust LLMs achieved strong performances. In contrast, the OpenIE setting proved more challenging, mainly due to the relation mapping stage, which limited model performance and highlighted the complexity of open-ended evaluation.

The dataset, creation pipeline, and experimental prompts are available at the aforementioned repository, promoting replicability and community use. The presented methodology enables users to extend the dataset or develop new resources.

For future work, we plan to expand the dataset by adding more relations and increasing the number of instances per class. We intend to extend the current pipeline to also handle cross-sentence data and multiple triple coverage, going beyond the single-sentence setup explored here. On the OpenIE side, we aim to investigate alternative mapping strategies, such as embedding-based refinement and hybrid lexical-semantic alignment, to possibly enhance overall extraction accuracy.

7. Acknowledgements

This work was financed by the Portuguese Recovery and Resilience Plan (PRR), through project C645008882-00000055 – Center for Responsible AI.

This work was also supported by FCT – Foundation for Science and Technology, I.P., under the projects UIDB/00326/2025 and UIDP/00326/2025; and produced as part of the N-GenERP project with reference COMPETE2030-FEDER-02219400 (operation no. 21343) supported by the European Regional Development Fund (FEDER) through the Innovation and Digital Transition Programme (COMPETE 2030) of Portugal 2030 and the European Union.

8. Bibliographical References

David Soares Batista, David Forte, Rui Silva, Bruno Martins, and Mário Silva. 2013. Extracção de relações semânticas de textos em Português explorando a Dbpédia e a Wikipédia. *Linguamatica*, 5(1):41–57.

Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. 2023. RED FM: a filtered and multilingual relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4326–4343, Toronto, Canada. Association for Computational Linguistics.

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.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. 2020. PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. *arXiv preprint arXiv:2008.09144*.

Ruan Chaves Rodrigues. 2023. Lessons learned from the evaluation of portuguese language models. Master's thesis, University of Malta.

Kartik Detroja, CK Bhensdadia, and Brijesh S Bhatt. 2023. A survey on relation extraction. *Intelligent Systems with Applications*, 19:200244.

Jose A Diaz-Garcia and Julio Amador Diaz Lopez. 2025. A survey on cutting-edge relation extraction techniques based on language models. *Artificial Intelligence Review*, 58(9):287.

Cláudia Freitas, Diana Santos, Cristina Mota, Hugo Gonçalo Oliveira, and Paula Carvalho. 2009. Detection of relations between named entities: report of a shared task. In *Proceedings of NAACL-HLT, Semantic Evaluations: Recent Achievements and Future Directions Workshop*, SEW 2009, Boulder, Colorado. ACL Press.

Luís Gomes, António Branco, João Silva, João Rodrigues, and Rodrigo Santos. 2024. Open sentence embeddings for portuguese with the Serafim PT* encoders family. In *Proceedings of 23rd EPIA Conference on Artificial Intelligence, EPIA 2024*, pages 267–279. Springer.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. Wikireading: A novel large-scale language understanding task over wikipedia. *arXiv preprint arXiv:1608.03542*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Yizhi Jiang, Jinlong Li, and Huanhuan Chen. 2024. Relation classification via bidirectional prompt learning with data augmentation by large language model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13885–13897.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Meta. 2024. Llama 3.3 Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

OpenAI. 2025. Introducing GPT-4.1 in the API.

Pranshu Pandya, Mahek Bhavesh Vora, Soumya Bharadwaj, Ashish Anand, et al. 2024. Amalrec: A dataset for relation extraction and classification leveraging amalgamation of large language models. *arXiv preprint arXiv:2412.20427*.

Lucas Pavanelli, Yohan Bonescki Gumiel, Thiago Ferreira, Adriana Pagano, and Eduardo Laber. 2023. Bete: A brazilian portuguese dataset for named entity recognition and relation extraction in the diabetes healthcare domain. In *Brazilian Conference on Intelligent Systems*, pages 256–267. Springer.

Tomás Pinto, Bruno Ferreira, Catarina Silva, and Hugo Gonçalo Oliveira. 2025. Prompting LLMs for relation classification in portuguese: Is it worth it? In *Proceedings of 24th EPIA Conference on Artificial Intelligence, EPIA 2025*, pages 249–261, Cham. Springer Nature Switzerland.

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*, pages 196–206.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the eighth conference on computational natural language learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8.

Rodrigo Santos, João Rodrigues, Luís Gomes, João Silva, António Branco, Henrique Lopes Cardoso, Tomás Freitas Osório, and Bernardo Leite. 2024. Fostering the ecosystem of open neural encoders for portuguese with Albertina PT* family. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, pages 105–114.

Alessandro Seganti, Klaudia Firląg, Helena Skowronska, Michał Satława, and Piotr Andruszkiewicz. 2021. Multilingual entity and relation extraction dataset and model. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume*, pages 1946–1955.

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.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2024. Chatie: Zero-shot information extraction via chatting with chatgpt.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019a. DocRED: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019b. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Kai Zhang, Bernal Jiménez Gutiérrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. *arXiv preprint arXiv:2305.11159*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 35–45.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11):1–39.