Identification of Types of Event-Time Temporal Relations in Portuguese Using a Rule-Based Approach

Dárcio Santos Rocha and **Marlo Souza** and **Daniela Barreiro Claro**
Institute of Computing – Federal University of Bahia
Salvador – BA – Brazil

Abstract

In this article, we present a computational method for identifying types of temporal relations between events and temporal expressions in Portuguese texts. We employ a linguistically-rich approach based on rule learning algorithms, and language-specific manual rules. Experiments on the TimeBankPT corpus demonstrated the effectiveness of our method, outperforming the baseline in terms of accuracy and F1-score. Through the use of explainable rules, our method enables an enhanced understanding of temporal phenomena in texts, allowing further development of resources and linguistic research on the area.

1 Introduction

Temporal understanding in written texts plays a fundamental role in effective communication. By identifying and comprehending the temporal relations present in texts, it is possible to establish the chronological order of events and their interactions, with practical applications such as scene description, story comprehension, document summarization, and more.

In this context, this study aims to develop a method for identifying types of temporal relations between an event and a temporal expression for the Portuguese language, adopting a rule-based, linguistically-rich approach.

The focus of our approach is on interpretable methods, which allow linguistics experts to analyze and discuss the system's decisions. This is an important feature for such a resource-scarce task in the Portuguese language, as such a method can be used to bootstrap the creation of annotated data for temporal relation identification and information extraction. Furthermore, interpretability is a topic of great relevance and interest in the Artificial Intelligence (AI) scientific community, as the ability to understand and explain the decisions made by AI models is crucial not only to ensure transparency and reliability in these systems, but also to enable critical analysis by experts with relevant linguistic knowledge.

To achieve this purpose, we leverage a feature engineering-based approach, exploring features proposed in the literature, and rule learning algorithms to encode the problem of identifying temporal relations as a classification problem. We also investigate different methods for classification based on these rules and methods for combining rules obtained by different algorithms, aiming to investigate whether these algorithms can identify complementary information.

We conduct experiments on the TimeBankPT corpus[1] (Costa and Branco, 2012), which contains annotations of a simplified set of temporal relations in Portuguese, such as BEFORE, AFTER, OVERLAP, etc. The TimeBankPT corpus is a Portuguese translation of the TimeBank corpus (Pustejovsky et al., 2003) and, to our knowledge, constitutes the only annotated corpus for temporal relations available for the Portuguese language.

The results of the experiments demonstrate the effectiveness of the proposed approach in identifying temporal relations in Portuguese. The ruleset generated by the RIPPER algorithm (Cohen, 1995) showed the best performance, resulting in an absolute increase of 3.6 percentage points in the F1-score compared to the baseline - the *LX-TimeAnalyzer* system, proposed by Costa (2012), which was the first published study on the identification of types of temporal relations in Portuguese.

It is important to notice that our work focuses on the identification of temporal relations between events and temporal expressions, as this is still an underdeveloped topic in the literature for the Portuguese language. As such, in our approach, it is assumed that the identification of events and temporal

[1]The TimeBankPT corpus is available at http://nlx-server.di.fc.ul.pt/~fcosta/TimeBankPT/

expressions has already been completed. In other words, our method presupposes those annotations related to events and temporal expressions have been provided beforehand. However, it is worth mentioning some effort has been devoted to the identification of events in the Portuguese language, as evidenced by studies conducted by Cabrita et al. (2014), Mota and Santos (2008), and Sacramento and Souza (2021), or language-independent methods, such as Feng et al. (2018). Regarding the identification and normalization of temporal expressions, contributions can be found in studies by Mota and Santos (2008), Strötgen and Gertz (2013), and Real et al. (2018).

The remainder of this paper is organized as follows. In Section 2, an overview of the main concepts discussed in the article is provided, addressing the theoretical foundations related to understanding temporal relations in texts. Section 3 details the proposed method, describing the steps and procedures used to construct and apply the rules in the process of identifying temporal relations. Next, in Section 4, the conducted experiments and the main results obtained are presented, including performance metrics and comparisons with the baseline. Finally, in Section 5, the study's conclusions are presented, highlighting the contributions and limitations of the proposed method, as well as possible directions for future work.

2 Background

The identification of different types of temporal relations is a very important task in the field of Information Extraction. Verhagen et al. (2007) define temporal relation identification as the automatic identification of all temporal references present in a text, including events, temporal expressions, and temporal relations.

2.1 Temporal Relations

According to UzZaman et al. (2012), a temporal relation connects events or temporal expressions and indicates the order in which they occurred or whether they occurred simultaneously. The temporal ordering between events and temporal expressions is not always explicit, which complicates the identification of the type of temporal relations present. Therefore, even with sophisticated approaches, the identification of types of temporal relations remains a challenge, as stated by Derczynski (2017).

The work by Marsic (2011) underscores that temporal relations are frequently only partially articulated in natural language, employing temporal adverbs, verbal aspects, syntactic dependency relations, and prior knowledge about the world. The author posits that events and temporal expressions constitute fundamental elements in the annotation of temporal discourse.

In accordance with Pustejovsky et al. (2004, p. 4), events are understood as temporal entities that "*can be punctual or last for a period of time*", and they "*are generally expressed by means of tensed or untensed verbs, nominalizations, adjectives, predicative clauses, or prepositional phrases*". The term "event" is utilized broadly to encompass what some literature refers to as events or states. Temporal expressions, on the other hand, are natural language phrases that refer directly to time, giving information on when something happened, how long something lasted, or how often something occurred (Marsic, 2011). A more extensive discussion of these concepts can be found in the work of Rocha (2023).

2.2 Classification Rule Learning

In this study, we investigate the application of rule-learning techniques to identify the types of temporal relations between pairs composed of event and temporal expression (event-time). Association rule learning is a subfield of data mining, popularized by Agrawal et al. (1993), which focuses on extracting patterns or frequent sets from data. An association rule follows the form $A \rightarrow B$, where A and B are sets composed of one or more items. A is the antecedent, and B is the consequent.

To address a classification problem, we impose a syntactic constraint on the consequent of association rules. Specifically, we permit only rules that include a designated item representing the class to be predicted, namely, the type of temporal relation. Once this constraint is defined, the problem transforms into a task of learning classification rules or associative classification, as defined by Liu et al. (1998).

Associative classification rules are considered an effective approach for representing information due to their ease of readability and understanding. In the context of this work, some associative classification algorithms were employed to construct rule-sets capable of identifying the types of event-time temporal relations. The algorithms used were `CBA` (Liu et al., 1998), `CN2` (Clark and Niblett, 1989),

IDS (Lakkaraju et al., 2016), and RIPPER (Cohen, 1995). The choice of these algorithms is primarily driven by their suitability for effectively handling datasets with noise, such as class imbalances and missing data. Additionally, they prioritize rule interpretability and demonstrate robust performance when applied to unknown datasets.

The Classification Based on Associations (CBA) algorithm, developed by Liu et al. (1998), focuses on identifying Class Association Rules (CARs) that meet minimum support and confidence requirements. It employs a variant of the Apriori algorithm and comprises two main steps: CBA-RG, responsible for generating association rules. During this step, iterations over the data are performed to generate frequent rules, with pruning applied to reduce their number. The second step is CBA-CB, which builds a classifier based on CARs. In this phase, rules are organized and selected based on confidence and support metrics, resulting in the creation of a classifier capable of categorizing new cases.

On the other hand, the CN2 algorithm, developed by Clark and Niblett (1989), identifies rules that cover a set of learning instances, removing them and repeating the process until all instances are covered. CN2, designed for noisy or poorly described language environments, incorporates enhancements, including beam-guided search, Laplace estimates, and significance testing of the likelihood ratio, aiming to avoid overfitting. It uses a heuristic based on noise estimates to halt the search during rule construction, resulting in rules that may not cover all training examples but perform well on new data.

The Interpretable Decision Sets (IDS) algorithm, proposed by Lakkaraju et al. (2016), aims to learn non-overlapping rulesets with high accuracy, covering all features and considering minority classes. Learning is guided by an objective function that optimizes interpretability and performance. IDS uses Smooth Local Search (SLS) to find a set of decisions that maximize the objective function, considering samples of rulesets and classes.

Finally, the RIPPER algorithm (Repeated Incremental Pruning to Produce Error Reduction), developed by Cohen (1995), operates in three stages: grow, prune, and optimize. In the growth stage, it employs the "separate-and-conquer" (Pagallo and Haussler, 1990) method to add conditions to a rule until perfectly classifying a subset of data. It then applies an information gain criterion to identify the next splitting attribute. The specificity of a rule is reduced until entropy no longer decreases, at which point the rule is pruned. These steps are repeated until a stopping criterion is reached, at which point the ruleset is optimized using various heuristics. RIPPER effectively addresses overfitting through the Incremental Reduced Error Pruning (IREP) technique, which removes a rule, attempts to relearn it in the context of previous and subsequent rules, avoiding excessive complexity, and improving model generalization.

3 Method

The task addressed in this study was defined based on the work of Verhagen et al. (2007), which deals with the identification of types of temporal relations between an event and a temporal expression (event-time) in the same sentence. To identify the type of temporal relation, we adopted a rule-based approach.

The proposed method involves creating a comprehensive set of features containing relevant linguistic information. These features are used to construct rulesets using rule-learning algorithms. These rulesets are individually applied to the pairs formed by event and temporal expression of the temporal relation, as well as in combination. The application of rules is performed in two ways: by the "first rule triggered" and through "voting". Further details are presented below.

3.1 Survey of Features and Generation of Rulesets

Based on the premise presented by Derczynski (2017) that temporal ordering in texts requires multiple sources of linguistic information, we conducted a literature review to identify sets of features proposed by various authors that are useful in identifying types of temporal relations. Each feature is composed of linguistic information extracted from events and temporal expressions, as well as from words near them, their syntactic governors, and dependents in the sentence. Based on this review, we compiled a set of 70 features, detailed and explained by Rocha (2023) and available in our GitHub[2] repository. In Table 1, we classify the features explored in this work, based on the type of linguistic information encoded. These features served as input for the CBA, CN2, IDS, and RIPPER

[2] https://github.com/temporalrelation/paper

learning algorithms to create our individual rulesets.

Type of Linguistic Information	Quantity
Morphological information	26
Syntactic information	12
Contextual information	11
Temporal signals	10
TimeML annotation	7
Prior knowledge about the world	2
Reichenbachian tenses	1
Lexical information	1
Total of features	**70**

Table 1: Quantity of features by type of linguistic information.

The features that encompass linguistic information annotated in the TimeML format within the corpus serve various functions. Firstly, they indicate the polarity of events, determining whether they are positive or negative. Additionally, these annotations categorize events into different classes, including reporting, perception, aspectual, state, and occurrence. Furthermore, such annotations delineate the type of temporal expression, encompassing `DATE` or `TIME` to denote specific dates or times (e.g., "upcoming Monday"), `DURATION` for temporal intervals (e.g., "two hours" or "three weeks"), and `SET` for recurrent dates or times (e.g., "every third Sunday"). TimeML, as defined by Pustejovsky et al. (2004), is a formal method for describing and processing entities relevant to temporal information extraction.

On the other hand, features related to "prior knowledge about the world" represent information that a speaker possesses about events, individuals, and locations in their environment. This information is useful in inferring temporal relations between events and temporal expressions. The individual meanings of certain words involved in the relation can provide relevant temporal clues for the task of identifying temporal relations, as discussed by Costa (2012). To obtain such information, the author manually mapped the expected temporal relations between specific events and their complements. For instance, events of delaying precede delayed events, events of organizing precede organized events, and reporting events follow reported events. Therefore, this feature records prior knowledge about the world.

The features that encompass contextual information approach various ways of coding relevant elements of temporal relations to make them more advantageous for solving the problem in question. They examine other elements present in the same sentence, in addition to those involved in the temporal relation under consideration, considering the presence, order, and distance of elements such as prepositions, conjunctions, modal verbs, and other events or temporal expressions different from those under classification. For example, it is possible to check for the presence of other events between the pair of event and temporal expression being classified. A feature in this category can indicate the preposition preceding the event or the distance between the entities under classification. Several authors, including Costa (2012), Derczynski (2017), and Mirza and Tonelli (2014), have used this type of linguistic information in their research.

In this study, we avoided the use of word-based features due to the potential data sparsity issues, given the relatively limited size of the corpus, as argued by Costa (2012). However, we included a feature that searches within the content of temporal expressions for lexical information based on a restricted list of words with temporal content that are frequently found in temporal expressions, such as "*ainda*" (yet), "*amanhã*" (tomorrow), "*anterior*" (previous), "*anteriormente*" (previously), etc.

Additionally, features provide information about temporal signals, as investigated by Derczynski (2017), based on words and phrases that explicitly express the nature of a temporal relation. These temporal signals consist of temporal conjunctions and adverbs that often accompany temporal connections, offering explicit information about the type of temporal relation. These features supply information on the temporal signals that precede events or temporal expressions. Examples of such temporal signals include words like "*antes*" (before), "*depois*" (after), "*agora*" (now), "*ontem*" (yesterday), "*hoje*" (today), "*amanhã*" (tomorrow).

The features encompassing morphological information include data on part of speech, tense, and aspect. This information has been widely used by various authors, including Costa (2012), D'Souza (2015), Chambers et al. (2014), and Bethard and Martin (2007). As for the features providing syntactic information, they reveal the relations of government or dependence between the entities involved in the temporal relation based on the syntactic de-

pendency tree. Authors such as Derczynski (2017), and Mirza and Tonelli (2014) have explored this type of linguistic information.

Finally, the feature that encompasses information about Reichenbachian tenses, as explored by Derczynski (2017), utilizes the work of Reichenbach (1947) as a foundation. This work provides a theoretical framework for the analysis of tense and aspect, applicable to predicting the temporal ordering between verbal events and between temporal expressions and verbal events. Intuitively, this feature is relevant for determining the types of temporal relations. We explore in greater detail various aspects of this linguistic information in Rocha (2023).

In addition to our features set, our research involved modifications to the IDS and RIPPER algorithms. These adjustments were aimed at achieving satisfactory data coverage rates, defined as the percentage of instances classified by some rule. Specifically, we established a criterion for satisfactory coverage, with the goal of achieving an average of 90%. This means that approximately 90% of the examples were classified by one or more rules.

The implemented modifications involved conducting training iterations exclusively on unclassified data, meaning those that were not predicted by any existing rule. In each iteration, the newly generated rules were accumulated with the rules obtained in the previous iteration. This includes adding the new rules to the existing ruleset and removing duplicate rules.

In addition to the individual rulesets generated by each algorithm, we also developed a set of manual rules for Portuguese, inspired by the rules proposed by D'Souza (2015) for the English language. These rules were composed of lexical information, part of speech, morphological information, syntactic dependency relationship between the temporal entities and their governors in the sentence, contextual combinations, and attributes annotated in the corpus.

Furthermore, we also investigate the combination of rules learned by different algorithms in two approaches. The first combines all individual rulesets into a single set. The second set is formed by the best combination of two of the individual rulesets. In addition, we chose to designate the OVERLAP class as the default class in each ruleset, due to its predominant frequency.

It is important to highlight that these rules achieved high coverage rates even without the use of the default class, with an average of 90% on the training data and 92.6% on the test data. By adding the default class, the rule system becomes more comprehensive and robust because it can classify unknown instances that were not covered by specific rules. This improves the system's ability to generalize and makes its classification more consistent.

3.2 Rule-based Classification

Once the rulesets were constructed, they were applied to the event-time pairs in the datasets to identify the type of temporal relation. We investigate different methods for the application of rules. In the first approach, called "**first rule triggered**", the class associated with the first rule triggered, given a certain ordering of the rules, is considered as the final class for the event-time pair. The ordering may be obtained either from the learning algorithm or through evaluation metrics, such as accuracy in the training data. After being classified by a triggered rule, the pair is no longer subjected to the remaining rules, and processing proceeds to the next pair to be classified.

In the second approach, called "**voting**", all event-time pairs are subjected to all the rules in the ruleset. Votes are assigned to each class based on the rules triggered, and the most frequent class is assigned as a result.

To illustrate the application of the rules, consider the sentence "***Teremos*** *um ano razoavelmente em baixa este ano.*" ("We will **have** a reasonably flat year this year."), extracted from the TimeBankPT corpus. In this sentence, the event is "***Teremos***" and the temporal expression is "*este ano*". The application of rule (1), generated by the RIPPER[3] algorithm, allowed us to determine the type of temporal relation between the event-time pair ("*Teremos*", "*este ano*") as OVERLAP. This indicates that there is a relationship in which the event occurs during the same temporal period as the temporal expression.

(1) *event-between-order* = False and *reichenbach-direct-modification* = True ⇒ OVERLAP

The feature *event-between-order* checks whether there is another event between the event and the temporal expression of the relation under classification, while the feature *reichenbach-direct-modification* checks whether the temporal expres-

[3]This ruleset is available in our GitHub repository

sion directly modifies the event, meaning it is in the same syntactic dependency path as the event. Therefore, rule (1) classifies the event-time pair as OVERLAP because there is no other event between "*Teremos*" and "*este ano*", and because the expression "*este ano*" directly modifies the event "*Teremos*", in this case, through the oblique dependency relation.

In the next example, when applying rule (2), also generated by the RIPPER algorithm, to the event-time pair under analysis ("falar", "próximo ano") in the sentence "*Portanto, os seus altos executivos estão a* ***falar*** *abertamente da possibilidade de* ***recomprar*** *alguns dos 172,5 milhões de dólares da empresa em obrigações subordinadas convertíveis no próximo ano.*" ("So its senior executives are **talking** openly about possibly **buying** back some of the company's $172.5 million in subordinated convertible debentures next year."), we observe that the temporal relation between the event and the temporal expression of this pair is classified as BEFORE, as explained below.

(2) *timex3-preposition-precede* = 'no' and *event-between-order* = True and *timex3-relevant-lemmas* = 'próximo' ⇒ BEFORE

The rule in question is composed of conjunctions of three conditions. The first condition is satisfied if the preposition-determiner contraction "no" ("in the") precedes the temporal expression under analysis. This is because the feature *timex3-preposition-precede* is designed to track the preposition preceding the temporal expression under analysis, in this case, "próximo ano". In the second condition of the rule, the feature *event-between-order* evaluates the presence of another event between the event and the temporal expression of the event-time pair under analysis. In this context, the event found was "**recomprar**". Finally, the third condition is determined by the feature *timex3-relevant-lemmas*, which checks whether the uninflected form of the temporal expression contains the word "próximo".

When all three conditions are satisfied, the temporal relation established is identified as BEFORE, which indicates that the event "**falar**" occurred before the temporal moment represented by "próximo ano".

4 Experimental Evaluation

To select the experimental parameters, the training documents from the TimeBankPT corpus were divided into two parts. Ninety percent of the documents were allocated for rule development, while the remaining portion was reserved for validation. Based on the results obtained in the experiments using the validation data, the best experimental configurations were selected.

For the development of individual rulesets, several parameters were considered in our experimental setup, including the individual hyperparameters of each algorithm, a rule accuracy cut-off threshold (0%, 40%, 50%, and 60%), the ordering of the rules (order provided by the learning algorithm, or by accuracy on training data), and feature selection. For feature selection, two approaches were adopted: considering all 70 available features and using the Recursive Feature Elimination with Cross-Validation (RFECV) (Pedregosa et al., 2011) technique to select the most relevant features.

For the development of combined rulesets, the following combinations were made with the individual sets. In the first approach, the ruleset was obtained by combining all individual sets. To perform this combination, the individual sets were evaluated in descending order based on their accuracy and the coefficient of variation of the accuracies obtained in the experiments. Additionally, the ascending order by the number of rules was also considered. In the resulting set, different accuracy cut-off thresholds (70%, 80%, and 90%) were explored, and the rules were ordered by accuracy or kept in their original order.

In the second combination approach, the ruleset was formed by combining two individual sets. All possible combinations between the individual sets were considered, and the same accuracy cut-off thresholds used in the first approach (70%, 80%, and 90%) were applied. The rules were ordered by accuracy. The results obtained from the validation data were used for selecting the best hyperparameters.

To evaluate the final performance of our method, we used the previously partitioned training and test sets from TimeBankPT. The training data consists of 89% of the corpus documents and was used to retrain the selected models with the best configurations. The test data corresponds to 11% of the documents and was used for the final evaluation of the method's performance. This allows us to verify the effectiveness and generalization of the method when applied to an unseen dataset.

The evaluation metrics used to measure the performance of our method were accuracy and F1-

score. As a comparison reference, we adopted the *LX-TimeAnalyzer*, proposed by Costa (2012), which represents the first published study for the Portuguese language addressing the identification of types of temporal relations, to our knowledge. The *LX-TimeAnalyzer* achieved an accuracy of 66.9% and an F1-score of 62.5% on the test data when dealing with the task of identifying the type of event-time temporal relation.

4.1 Results

We present the main results of the experiments conducted in the task of identifying types of event-time temporal relation within the same sentence. The results consider two different approaches for rule application: the first rule triggered and the voting system. We will also present the results of selecting the best configurations for each ruleset, as well as the number of rules in each set.

Table 2 displays the optimal configurations employed for each individual ruleset, based on the validation data. The best cut-off threshold based on rule accuracy for most rulesets was 50% accuracy. When it comes to rule order, the original sequence proved to be more effective for the manual, CN2, and RIPPER rulesets. Ordering by accuracy proved to be more effective for the rulesets generated by the CBA and IDS algorithms.

Regarding the number of features used to generate rules, the set generated by the RIPPER algorithm benefited from using all 70 available features. The sets generated by CN2 and IDS performed better when using only the top 52 most relevant features selected by the Recursive Feature Elimination with Cross-Validation technique. However, due to the constraints imposed by computational resources, as the computer used had 32 GB of RAM, only the top 41 most relevant features could be used in generating the ruleset by the CBA algorithm.

As for the size of the individual rulesets, CBA was the largest, totaling 568 rules, followed by IDS with 383 rules, CN2 with 205 rules, RIPPER with 146 rules, and the manually created rules with only 35 rules.

Table 3 presents the optimal configuration for composing the combined rulesets, based on validation data. For the set composed of all individual sets, the best joining order was based on the accuracy of each ruleset in descending order. In the case of the set composed of two individual sets, the best combination was found with the sets generated by the IDS and CBA algorithms. The best cut-off threshold based on the accuracy of the rule was 80% for both sets combined. The ordering of the rules in both sets was based on the accuracy of the rule. The resulting combined sets totaled 980 rules for the set composed of all individual sets and 797 rules for the set composed of the combination of the IDS and CBA algorithms.

Table 4 displays the accuracy and F1-score metrics results for different rulesets, which were evaluated using test data. The evaluation considered two approaches for applying the rules: the first rule triggered and the voting system.

The RIPPER algorithm generated the best-performing ruleset with an accuracy of 69.2% and an F1-score of 66.1%. In second place, the combination of rulesets from the IDS and CBA algorithms achieved an accuracy of 68% and an F1-score of 63.4%. In third place, the combination of all individual sets resulted in an accuracy of 67.5% and an F1-score of 63.3%. These rulesets outperformed the baseline in terms of accuracy and F1-score, demonstrating their effectiveness compared to the reference method.

To confirm the statistical significance of the obtained results, one-way analysis of variance (ANOVA) (Snedecor and Cochran, 1989) and the Tukey multiple comparison test (Tukey, 1953) were employed with a significance level of 0.05. Comparisons among the experiments revealed a statistically significant difference between the means of the ruleset generated by the RIPPER algorithm and all other rulesets, as depicted in Figure 1.

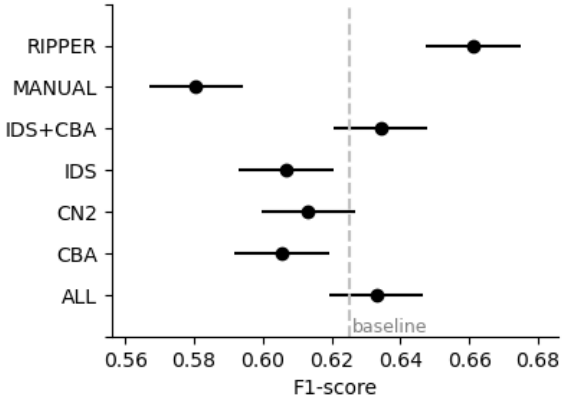

Figure 1: Simultaneous comparison of means by Tukey's test with a significance level of 0.05

To compare our results with the established baseline, a one-sample t-test was conducted, which is a single-sample comparison test to assess whether the experiment's mean is significantly different

	Manual	CBA	CN2	IDS	RIPPER
Cut-off threshold for accuracy	50%	50%	40%	50%	50%
Ordering of the rules	original	accuracy	original	accuracy	original
Number of features	-	41	52	52	70
Number of rules	35	568	205	383	146

Table 2: Better configuration and elements of each individual ruleset

	Combination of all	Combination of two
Order for joining / combining	accuracy of each set	IDS e CBA
Cut-off threshold for accuracy	80%	80%
Ordering of the rules	accuracy	accuracy
Number of rules	980	797

Table 3: Better configuration and elements of combined rulesets

	First Rule		Voting	
Rulesets	**Acc**	**F1**	**Acc**	**F1**
RIPPER	65,1	64,2	**69,2**	**66,1**
Combination of IDS and CBA	65,7	62,0	68,0	63,4
Combination of all	63,3	59,6	67,5	63,3
Baseline	*66,9*	*62,5*	*66,9*	*62,5*
CN2	65,7	61,3	65,1	57,6
IDS	64,5	59,6	66,9	60,7
CBA	62,7	59,9	62,7	60,5
Manual	66,9	58,1	66,9	58,1

Table 4: Results of all rulesets based on the test data, ordered by the highest F1-score

from the reference value. The results indicated that the mean of the ruleset generated by RIPPER differs significantly from the reference value ($p = 0.00041$), suggesting that the differences are highly unlikely to occur by chance and providing strong evidence of the superiority of this ruleset over the baseline.

The analysis of the results provides statistical evidence confirming the overall better performance of the ruleset generated by the RIPPER algorithm, even surpassing the reference method. This finding validates the effectiveness of this ruleset in identifying types of event-time temporal relations.

The approach of applying the rules through the voting system was the most effective for classifying new data. This approach had superior performance compared to the "first rule triggered" approach, except for the ruleset generated by the CN2 algorithm.

We also observed that the combination of rules from different algorithms did not result in superior performance compared to individual rulesets, as the rules generated by the RIPPER algorithm outperformed the combinations of rulesets in terms of performance. Although the combinations achieved second and third places, the fact that an individual ruleset surpassed these combinations indicates that the hypothesis was not confirmed.

All the rulesets are available in our GitHub repository.

5 Conclusions

This study introduced a computational method for identifying types of temporal relations between events and temporal expressions in Portuguese texts. The results demonstrated the effectiveness of our rule-based approach, with superior performance compared to the reference method in terms of accuracy and F1-score. Specifically, the best-performing ruleset generated by the RIPPER algorithm achieved an absolute increase of 2.3 percentage points in accuracy and 3.6 percentage points in the F1-score.

However, a limitation of this study was the scarcity of annotated data in the Portuguese language. In future work, addressing this limitation is crucial to further enhance the performance and generalization of the proposed approach. In this sense, we believe our method may be employed to help producing such resource in a semi-automated annotation strategy, the systems classifications can be evaluated by linguistics experts by means of the relevant linguistically-based rules. Furthermore,

we believe a qualitative approach, accompanied by in-depth linguistic analysis to validate the rules based on data and linguistic knowledge, would represent a significant contribution to enriching our explainable approach to temporal relations.

This research contributes to advancing natural language processing applications by providing an enhanced and explainable understanding of temporal relations. By continuously refining and expanding this research, we aim to uncover new possibilities for temporal understanding in texts.

Acknowledgments

This material is partially based upon work supported by the FAPESB under grant INCITE PIE0002/2022 and by CAPES Finance Code 001.

References

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216.

Steven Bethard and James H Martin. 2007. Cu-tmp: Temporal relation classification using syntactic and semantic features. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 129–132.

Viviana Cabrita, Nuno Mamede, and Jorge Baptista. 2014. *Identificar, ordenar e relacionar eventos*. Ph.D. thesis, Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa ....

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Peter Clark and Tim Niblett. 1989. The cn2 induction algorithm. *Machine learning*, 3(4):261–283.

William W Cohen. 1995. Fast effective rule induction. In *Machine learning proceedings 1995*, pages 115–123. Elsevier.

Francisco Costa and António Branco. 2012. Timebankpt: A timeml annotated corpus of portuguese. In *LREC*, volume 12, pages 3727–3734.

Francisco Nuno Quintiliano Mendonça Carapeto Costa. 2012. *Processing Temporal Information in Unstructured Documents*. Ph.D. thesis, Universidade de Lisboa (Portugal).

Leon RA Derczynski. 2017. *Automatically ordering events and times in text*. Springer.

Jennifer D'Souza. 2015. *Extracting Time and Space Relations from Natural Language Text*. Ph.D. thesis, The University of Texas at Dallas.

Xiaocheng Feng, Bing Qin, and Ting Liu. 2018. A language-independent neural network for event detection. *Science China Information Sciences*, 61:1–12.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684.

Bing Liu, Wynne Hsu, and Yiming Ma. 1998. Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*, pages 80–86.

Georgiana Marsic. 2011. *Temporal Processing of News: Annotation of Temporal Expressions, Verbal Events and Temporal Relations*. Ph.D. thesis, University of Wolverhampton.

Paramita Mirza and Sara Tonelli. 2014. Classifying temporal relations with simple features. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 308–317.

Cristina Mota and Diana Santos. 2008. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo harem.

Giulia Pagallo and David Haussler. 1990. Boolean feature discovery in empirical learning. *Machine learning*, 5:71–99.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Recursive feature elimination with cross-validation example. Accessed on: 2023-12-20.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The timebank corpus. *Proceedings of Corpus Linguistics*.

James Pustejovsky, Robert Ingria, José M Castaño, Jessica Littman, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2004. The specification language timeml.

Livy Real, Alexandre Rademaker, Fabricio Chalub, and Valeria de Paiva. 2018. Towards temporal reasoning in portuguese. In *Proceedings of the LREC2018 Workshop Linked Data in Linguistics*.

Hans Reichenbach. 1947. Elements of symbolic logic.

Dárcio Santos Rocha. 2023. Identificação de tipos de relações temporais event-time em português: Uma abordagem baseada em regras com classificação associativa. Master's thesis, Universidade Federal da Bahia, Salvador, BA, Brasil, Agosto.

Anderson da Silva Brito Sacramento and Marlo Souza. 2021. Joint event extraction with contextualized word embeddings for the portuguese language. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 496–510. Springer.

George W Snedecor and William G Cochran. 1989. Statistical methods, eight edition. *Iowa state University press, Ames, Iowa*, 1191(2).

Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47:269–298.

John Wilder Tukey. 1953. The problem of multiple comparisons. *Multiple comparisons*.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv preprint arXiv:1206.5333*.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic. Association for Computational Linguistics.