Event Identification in STRING

Extended Abstract

José Paulo de Oliveira Rodrigues Marques Dias
IST – Instituto Superior Técnico
L^{2}F – Laboratório de Sistemas de Língua Falada – INESC ID Lisboa
Lisboa, Portugal

ABSTRACT

Event identification in texts is an important task in Natural Language Processing (NLP), as it allows for the extraction of information in a structured way, which can have multiple applications in automatic summarization an event reconnaissance. The work described in this document focused on the identification of seven different types of events, i.e. Crime, Trial, Prision, Location Static, Location Visit, Public and Ephemerid.

These events were incorporated in the processing chain STRING (Statistical and Rule-Based Natural Language Processing), developed at Laboratory for Spoken Language Systems (L2F) of the Institute of Systems Engineering and Computers Research and Development in Lisbon (INESC-ID), in an effort of improving the system's ability of event identification.

KEYWORDS

Portuguese, Natural Language Processing (NLP), Event Identification, Rule-based Identification

1 INTRODUCTION

Ever since the creation of writing, human being have been able to encode information in the texts they write. These texts provide a way to share knowledge among people, taking into account that they have the ability to properly understand the information that is present in the texts.

Proper understanding of a written text is not a trivial matter, as it requires the reader to know all the rules of the language at hand. But even this may not be enough if the intent of the writer is not properly conveyed in the words he has written. Therefore, the analysis texts cannot be considered a trivial matter. It is necessary to determine the both the explicit information and the implicit information.

Nowadays there are methods to automatically extract information [11] from texts, mainly through the use of Natural Language Processing (NLP), which is one area of Artificial Intelligence that specializes in the automatic comprehension of natural languages. The STRING (Statistical and Rule-Based Natural Language Processing) [12] is a text processing chain developed at L2F (Spoken Language System Lab), that has the ability of processing very large corpora and is able to perform several NLP tasks.

This work focuses on the development of task of event identification in STRING, which is performed by its parsing module, the Xerox Incremental Parser [3]. The original iteration of STRING was already able of detecting a few kinds of events, i.e. events of Lifetime, which extracted information relative to major milestones of a persons life, events of Business, and a small portion of events of Crime, Static Location and Public.

The objectives of this work were the expansion of the capabilities of STRING for the identification of events of Crime, Trial, Static and Public, as well as, the implementation of the events of Prision, Visit and Ephemerid.

The work began with the assessment of the the constraints that would determine what situations could be considered as being events in the context of STRING.

Then a phase of formulation of relations, for each type of event, began and the different relations were developed in order to extract the most relevant information related to the event in question. From this formulation, some relations were deemed core for the event they belonged to and had to be extracted in order for the event to be considered valid.

For the purpose of evaluation, sets of sentences were ran through the original system and the modified system, which would then show a overview of how much of an improvement the new implementation is, when compared with the original.

The results were then evaluated based on the systems ability to correctly detect events in sentences and, for the correct events, how well it extracts the relations for each event.

The rest of the document is structured as follows: 2 mentions a few experiments that tackled the issue of event identification, 3 describes the requirements for the identification of each type of event, 4 describes the relations that can be extracted for each type of event, 5 briefly describes the steps taken during implementation, 6 describes the methods and shows the results of the evaluation, 7 discusses the results obtained in the evaluation and 8 conclude this document with an overview of the overall work.

2 STATE OF THE ART

There are a few systems that tackle the issue of event identification.

Knowledge-based Approach for Event Extraction from Arabic Tweets

Al-smadi and Qawasmeh [4] use a Knowledge-based approach in order to extract events from Arabic Tweets.

The authors developed a rule-based system that analyzed the tweeter feed, and extracted events based on the event agent, the location, the target, the event trigger and the event time. The Twitter Streaming API [2] was used to collect the data by searching the tweets for temporal expressions, which resulted in a total of 3000 tweets.

After the tasks of prepossessing, 1000 of these tweets were tagged as being instant type events or interval time events, which resulted on 878 tweets of instant type and 122 tweets of interval type.

A rule-based approach based of the Arabic Annotation Guidelines for Events [9], provided by the Linguistic Data Consortium (LDC). Three different sets of rules were used in order to extract three key aspects of the events: the event trigger, the event time and the event type.

One of the rule set focused on the evaluation of the verbs present in the tweets, while another focused on the nouns and the last set focused on cardinal numbers.

The first two rule sets extracted the event trigger, while the last set extracted the event time. The event type was extracted by evaluating the time expressions present in the tweet.

Word Level Event Identification

March and Baldwin [13] made studies that produced an Event-Based Summary system that was later turned into a component of the DUC 07 summarization system of [17].

The original system used sentence level classification and was effective at answering questions centered on specified topics but had difficulties at answering questions that required the listing of events. Support Vector Machines (SVM) are used in this study to find the most effective representation for the event references in the given newswire and newspaper text documents.

In the new system component, a question is classified as either being an event-based question or a general question, and the corresponding summary is generated by the original system, if the question is a general question, or by a dedicated event-based system, if it is an event-based question.

This work focuses on the task of identification of the event references and, for this, March and Baldwin [13] opted to do the identification at the word level instead of at sentence level.

The experiment was carried out using the TimeBank 1.1 corpus, which consists of 186 news articles, extracted from the Wall Street Journal and the Associate Press newswire, which was marked using TimeML specification language standard 1.1.[16].

Sentence level classification is first used to remove non-event sentences so that only sentences containing events are passed onto the word level classifier.

Five types of text representation techniques are explored in the experiment: the use of a Context Window (0 to 3 preceding words) to identify the events, the Feature Representation of the context words as a bag-of-words or as a ordered list of word positions, the removal of Stop Words, the use of POS tagging instead of word feature for context words and Feature Generalization.

The evaluation of the results revealed that maintaining word order had the biggest impact on the classification, the usage of POS increased the overall performance of the system and the removal of stop words improved it just very slightly.

Streaming First Story Detection with application to Twitter

Petrović et al. [15] address the problem of detecting new events from a stream of Twitter posts by developing a system, based on First Story Detection (FSD), which is one of the subtasks of topic detection and tracking [5], that is capable of detecting events on a web-scale corpus through the use of a locality-sensitive hashing algorithm.

Detection of events from tweets is difficult because of the large volume of data, the large level of noise present in it and because of the speed required to efficiently process such large amount of data. [15] dealt with these issues by developing a FSD system that processes a new document in a constant time, while maintaining a constant space. The constant processing time was achieved with the use of Local Sensitive Hashing (LSH) which dramatically reduces the time needed to find a closest neighbor in the vector space. The constant space was achieved by limiting the amount of documents that can be in memory at any given time.

The FSD system assigns a novelty score to each tweet and finds the most similar tweets, which are used to generate threads with a close level of similarity. The smaller the threads the more specific they become.

To measure the efficiency of the system, the tweets returned from the system were manually labeled has either Event, Neutral and Spam. A tweet was labeled as an event if the information in it was sufficiently explicit and had a significant importance. Neutral events were considered to be all other tweets. All other cases were considered Spam.

Evaluation was done by computing the average precision on the 820 tweets and these were sorted according two criteria: only event tweets are taken as relevant, and event tweets and neutral tweets are taken as relevant.

The results showed that, for the relevant events, any of the used ranking methods performed better than the baseline. For events and neutral taken as relevant, the results were much better than the previous criteria.

Automatic Event Classification Using Surface Text Features

Hardy et al. [10] presents a data driven way of discovering events and their attributes that is part of a question answering system.

[10] use a Text Framing strategy to deal with analytical questions. The strategy imposes a partial structure to the text excerpt so that comparison, with other excerpts and with the question, is possible. General frames represent the topic of the text (e.g. pollution, trade, etc.), which is done by verifying the core verb, or noun, and named entities (e.g. person, date, etc.) of the excerpt. Typed frames represent specific events (e.g. transfer, assist, etc.) and their roles (e.g. destination, source, etc.).

Events were extracted from a corpus of 37444 documents from the Center for Non-Proliferation Studies and from a 178015 documents that were extracted from the web and covered topics related to weapons. A data driven approach was used in order to determine the most interesting events and their structure and then to extract them from the texts.

Using the BBN's IdentiFinder [8], entities were tagged depending on the context they were integrated in. A total of 3996 event instances were identified and split into 10 event types (i.e. Agree, Assist, Attack, Develop, etc.) plus a None event type.

Through the use of the Weka toolkit [18], three classifiers were selected to evaluate the system, as well as a baseline that chose the majority class. The first algorithm, called Logistic, builds a logistic regression model.The second, called Vote, is a meta algorithm that combines the results from four different base classifiers, including Naive Bayes, REP tree, Random Forest and Part. The third is called Bagging, that reduces variance of a classifier (REP tree).

From the 3996 annotated instances that comprise the 11 event types, 20 to 100 most frequent words were collected, excluding stop words, for the parts of speech that were considered to be most useful in the prediction of event types (i.e. noun, verb, adjective).

Tests were made using different combinations of features, totaling in 170, 124, 114 and 94 features, comprised of different amounts of nouns, verbs, adjectives, pronouns and named entities, and Logistic was determined to have the best performance, of all the classifiers, at 124 features.

3 EVENT RECOGNITION

In order to proceed with event identification it is necessary to determine the criteria that determines when a situation can be considered as an event. These criteria change depending on the type of the event.

Crime

A situation can be considered a crime when it involves some sort of violation that is legally condemned by the law of a country/community. The cases that can be considered to have a criminal nature may differ from country to country. This work focuses solely on the list of crimes that are defined in the Portuguese Penal Code (Código Penal Português) [1]. For example, gambling is considered to be a crime in Taiwan, not in Portugal, hence it is not considered a crime in the implementation of crime events.

Different methods can be used in order to identify if a sentence contains a crime. One of the ways is to check for the existence of nouns or verbs that identify the crimes themselves:

- O Pedro cometeu o homicídio de 2 pessoas.
  Pedro committed the murder of 2 people.
- O Pedro conspirou.
  Pedro conspired.

Some verbs, i.e. `roubar` (to steal), `assassinar` (to murder), etc., can directly showcase an intent to practice a crime.

In some situations, verbs alone cannot indicate a crime, requiring further context in order for a crime to manifest, for example:

- O Pedro foi atacado com uma faca.
  Pedro was attacked with a knife.

Some situations can only be considered as crimes when there is a clear human victim, for example:

- O Pedro matou o Jorge.
  Pedro killed Jorge.

One last way to recognize a crime is through the adjectives that characterize the subject of a crime, for example:

- O Pedro é o homicída do João.
  Pedro is the murderer of João.

A situation will only be considered a Crime event if it contains textual elements that denote a criminal action and has either a culprit or a human victim to the action. Meaning that the following situation:

- Multiplos carros foram roubados na passada terça-feira.
  Several cars were robbed in the past Monday.

is not considered a Crime event in STRING.

Trial and Prision

A trial is a session where a responsible entity carefully examines facts related to an infraction of the law in order to determine the innocence or guilt of a person. A situation can be considered part of the Trial event if it fits the previous description and it directly relates to a crime.

In the same way, a situation is considered to be part of the Prision event if it relates to the detention of a person, for the practice of a crime, or if it represents a case where the person is already serving a sentence for a crime he/she has committed. The first case is called `detenção` and the second is called `prisã`.

For each of these types of events, the criminal action that triggered the event must be explicitly described in the sentence, in order to be considered as an event of that type. For example:

- Um homem vai a julgamento por branqueamento de capitais.
  A man will go on trial for money laundering.
- O Pedro está preso pelas sua acções.
  Pedro was judge for his actions.

The first example clearly indicates the criminal action that triggered the event of Trial, while the second example fails to describe the reasoning that lead to the imprisonment of the person, hence it is not considered as being part of the Prision event.

Similarly to the Crime event, both of these events require the identification of the culprit that performed the criminal action.

Static

The Static is the first of two events related to the Location category of events.

A situation can be considered to be part of the Static event if there is an entity, human or an entity metonymicaly being treated as a human [6] [14] or a human construct, i.e. a building or monument. whose location is being defined, such as that it invokes a state of “permanence" from that entity, i.e. the entity is fixed (not mobile) to that location at the time the event was identified. For example:

- A Estátua da Liberdade está em Nova Iorque.
  The Statue of Liberty is situated in New York.

The identification of this event is done through single verbs, e.g. “`estar`" (to be) or “`ficar`" (to stay), or through the combination of verbs, e.g. “`estar situada`" (is situated) or event through the combination of verbs and pronouns, e.g. “`encontra-se`" (is in), for example:

- A Torre Pisa fica situada em Itália.
  Tower of Pisa is situated in Italy.
- O António encontra-se na Microsoft.
  António is at Microsoft.

This event requires the presence of both the location of the entity and the entity itself.

Visit

The Visit event is the last related to the Location category of events.

A situation can be considered to be part of the Visit event if there is an human entity whose location is being defined, such as that the human entity, or an entity being metonymicaly treated as a human [7] [14], is “`visiting`" that location.

A simple way of identifying a Visit event is through the presence of verbs that indicate movement, e.g. “`viajar`" (to travel), “`visitar`" (to visit). For example:

- Maria Fernandes viajou para o Japão.
  Maria Fernandes travelled to Japan.

Nouns that indicate that the entity is going, passing through or acting as a passerby, e.g. “`férias`" (vacation), “`viajem`" (travel), also indicate a Visit event. For example:

- O Pedro está de férias na Escócia.
  Pedro is on vacation in Scotland.

Public

A situation is considered part of the Public event if it relates a human entity to public activity. These public events are divided into five types, “`festa`" (party), “`culto`" (cult), “`reunião` (meeting), “`desfile`" (parade) and `art-session`. “`Festa`" relates to festivities, “`culto`" relates to religious activities, “`reunião` relates to situations where a topic is discussed, “`desfile`" relates to any parade-like activity and `art-session` relates to art related situations.

Other situations that may be, initially, considered as a Public event, but cannot be integrated into these categories, are not considered as part of the Public event.

These events can be identified through the combination of a verb and noun that identifies the events, such has:

- O Pedro foi á festa no Barreiro.
  Pedro went to the party in Barreiro.

It is also possible to identify some of these events through a verb that describes the event, for example:

- Os senhores reuniram para discutir a situação.
  The gentleman have gathered to discuss the situation.

Sentences that do not explicitly indicate the participating entity are not considered as part of the Public event.

Ephemerid

The Ephemerid event is a type of Public event that celebrates the anniversary of relevant past events. For example:

- Celebra-se os 100 anos da abertura do IST.
  One celebrates the 100th anniversary of the opening of IST.

This event can be divided into three types, i.e. "nascimento" (birth), "morte" (death) and "efem(é)ride (ephemerid). The first two events are related to the birth and the death of an individual of renown, while the last event deals with all other relevant celebrations.

This is the only event that does not require the identification of a participating entity, instead, the only requirement is the event that is being celebrated.

The identification of this event is done through the identification of verbs that indicate the a celebration or festivity in the name of another event, e.g. "celebrar" (to celebrate), "festejar" (to party), followed by the event in question.

4 RELATION FORMULATION

The key goal of this work is to obtain information relative to the events through the analysis of each event occurrence.

Each event can extract different kinds of information exclusive to that event that reveals important and useful information. But there are some types of information that can be extracted by other events. These usually relate to the location or time when the event took place, and are identified by the PLACE and DATE relations, respectfully.

Crime

Table 1 displays the relations that were identified as being relevant for the purpose of identification of Crime events.

Relation	Description
AGENT	The perpetrator
VICTIM	The victim
INSTRUMENT	The instrument used
OBJECT	The object taken without consent
DATE	The time of the event
PLACE	The place of the event

Table 1: Relations for the Crime event type

The AGENT relation identifies the person that has committed a crime or is suspected of having committed the crime, while the VICTIM relation identifies the person that as suffered the crime.

The event requires that either the relations off AGENT or VICTIM to be properly identified, otherwise the identification of the event fails.

The INSTRUMENT relation identifies the tool used to aid the practice of the crime, and the OBJECT relation identifies an item that was taken from the rightful owner, during the course of the crime. This last relation can only be extracted if the crime that was practice involved robbing, stealing or smuggling.

Trial

Table 2 displays the relations that were identified as being relevant for the purpose of identification of Trial events.

Relation	Description
DEFENDANT	The person under trial
COURT	The court where the trial is held
JURISDICTION	The jurisdiction of the court
OUTCOME	The outcome of the trial
SENTENCE	The sentence to be served
FINE	The fine to be payed
DATE	The time of the event
PLACE	The place of the event

Table 2: Relations for the Trial event type

The DEFENDANT relation identifies the person being judged for a crime. This is the only required relation for this event.

The COURT relation identifies the court where the trial takes place and the JURISDICTION relation identifies the operational radius of that court. The JURISDICTION can only be extracted if the COURT relation is extracted.

The OUTCOME relation identifies the outcome of the trial, which can take the form of either "absovição" (absolution) or "condenação" (conviction). If the outcome is "condenação, then the SENTENCE and FINE relations can be extracted.

The SENTENCE relation identifies the sentence that the defendant will have to comply, and the FINE relation identifies the monetary amount that he will have to pay.

Prision

Table 3 displays the relations that were identified as being relevant for the purpose of identification of Prision events.

Relation	Description
DETAINEE	The person being arrested
PRISONER	The person being imprisoned
AGENT	The entity that made the arrest
SENTENCE	The sentence to be served
DATE	The time of the event
PLACE	The place of the event

Table 3: Relations for the Prision event type

The DETAINEE relation identifies the person that as been arrested for a crime, while the PRISONER relation identifies the person that is serving a sentence for a crime he/she has

committed. At least one of these relations are necessary to be extracted for the event to be considered valid.

The AGENT relation identifies the authority that as arrested the suspected criminal.

The SENTENCE relation identifies the sentence that the PRISONER is serving.

Static

Table 4 displays the relations that were identified as being relevant for the purpose of identification of Static events.

Relation	Description
PARTICIPANT	The person being located
OBJ	The object being located
MET-PLACE	The person being treated as a place
PLACE	The place of the event
DATE	The time of the event
DATE-START	The time when the event starts
DATE-END	The time when the event ends
DURATION	The duration of the event

Table 4: Relations for the Static location event type

The PARTICIPANT relation identifies a person or an entity, metonymicaly being treated as a person [7][14], whose location is being identified. While the OBJ relation identifies the object, which can be a human construction or a natural site, whose location is being identified.

At least one of these relations must be extracted for the event to be considered valid. Additionally, either the PLACE or the MET-PLACE relation must also be extracted in for the event to be valid.

The MET-PLACE relation identifies an entity that is metonymicaly treated as a place[7].

This relation also possesses the DATE-START, DATE-END and DURATION relations, that identify the beginning, the end and the duration of the event.

The relations of MET-PLACE, DATE-START, DATE-END and DURATION are shared with the Visit, Public and Ephemerid events.

Visit and Public

Table 5 displays the relations that were identified as being relevant for the purpose of identification of Visit and Public events.

The PARTICIPANT relation is required to be extracted, for both events, in order for the event to be considered valid.

Additionally, for the Visit event, either the PLACE or the MET-PLACE relation must be extracted for the event to be considered valid.

Relation	Description
PARTICIPANT	The participant of the event
MET-PLACE	The person treated as a place
PLACE	The place of the event
DATE	The time of the event
DATE-START	The time when the event starts
DATE-END	The time when the event ends
DURATION	The duration of the event

Table 5: Relations for the Public and Visit location event type

Ephemerid

Table 6 displays the relations that were identified as being relevant for the purpose of identification of Ephemerid events.

Relation	Description
ARG	The event being celebrated
ELAPSED-TIME	The time since first occurrence
PARTICIPANT	The participant of the event
PLACE	The place of the event
DATE	The time of the event
DATE-START	The time when the event starts
DATE-END	The time when the event ends
DURATION	The duration of the event

Table 6: Relations for the Ephemerid event type

The ARG relation identifies the event that is being celebrated. This relation is required if the event is identifying an event of birth or an event of death.

The ELAPSE-TIME relation identifies the time that has passed since the first time the event in question was celebrated.

5 DEVELOPMENT

The process of development began with an assessment of the state of event identification in STRING. From this assessment, it was determined that the current existing event types, mainly the events from the Location category and the Public event, had to be migrated to their own specialized files. This process required reviewing of some of the rules rules that were shared with other event types, i.e. PLACE and DATE related rules.

Also, a step of assessment of the quality of the existing rules was made and a new structure for writing of the new rules was developed, which reduced the overall number of comparisons required to identify an event by using the newly generated rules as anchors that prevents the access of those rules outside the event type in question. For example:

if(EVENT[CRIME](#1)

SUBJ[PRE](#1,#2[human]))
CDIR[POST](#1,#3[human])
EVENT[AGENT-GENERIC=+](#1,#2)

In total, the dependency construction effort resulted in the creation of 608 rules, spread over 329 rules for the Crime event category, 99 rules for the location event category and 180 for the public event category.

Additionally, the development yielded over 450 new entries to the lexicon pool.

6 EVALUATION

In order to evaluate the performance of the implementation, a non-annotated corpus, extracted from the recordings of the Portuguese Parliament, was used. From this corpus, seven sets of 50 sentences, one for each type of event, were extracted, with a total of 350 sentences. These sentences were ran through the original and modified system in order to generate the required outputs for evaluation. Sets of corrected outputs were, also, manually developed, in order to allow for a better understanding of the results. This yielded the results shown in Table 7.

	Crime	Trial	Prision	Static
Total	25	23	19	8
Baseline	13	4	0	0
Implement.	14	16	9	3
Improv. (%)	2 %	24%	18%	6%
	Visit	Public	Ephemerid	
Total	18	11	37	
Baseline	0	0	3	
Implement.	1	1	19	
Improv. (%)	2%	2%	29%	

Table 7: Comparison of correctly identified events.

The results of each set of the implementation were evaluated based how well the system was able to correctly extract the existing events. This was done by counting the number of events that were correctly identified, the ones it failed to identify and the events that were wrongfully identified. In a similar way, a evaluation of the extraction of relations was performed for the previously correctly evaluated events.

The results from these evaluations were then compiled using the metric of `precision`, `recall` and `f-measure`. The results for the evaluation of event identification are shown in Table 8 and the results for relation identification are present in Table 9.

	Precision	Recall	F-measure
Crime	0.78	0.56	0.65
Trial	1.0	0.7	0.82
Prision	0.9	0.47	0.62
Static	1.0	0.38	0.55
Visit	1.0	0.06	0.11
Public	0.33	0.09	0.14
Ephemerid	0.9	0.51	0.66
Total	0.84	0.39	0.51

Table 8: Measure results for the event identification.

	Precision	Recall	F-measure
Crime	0.93	0.5	0.65
Trial	0.86	0.93	0.89
Prision	1.0	0.3	0.46
Static	0.83	0.45	0.59
Visit	1.0	1.0	1.0
Public	1.0	1.0	1.0
Ephemerid	0.54	0.72	0.62
Total	0.88	0.7	0.74

Table 9: Measure results for the relation identification.

7 DISCUSSION

Through the analysis of Table 7, it is possible to see that, overall, the system improved in the detection of events. Despite this, it is difficult to assess the degree of improvement due to the constraints originated from the small set used for evaluation.

Through the analysis of the Table 8, it is possible to discern that the overall precision of the implementation is of 84%, which indicates high ability of the system to correctly identifying an event in a sentence without wrongfully identifying that a sentence has an event. However, the high values for precision for most of the events can be attributed to the rather low number of sentences that were evaluated. This low number of sentences implied an even lower number of actual events that could be identified for each type which, in turn, does not reflect a proper evaluation of the systems precision.

The recall of the system takes a values of 39%, which implies that the system fails at detecting a large portion of events. The values of recall are specially low for the events of Visit (6%) and Public (9%), indicating that they are not detecting most of the events in the sentences. On the other hand, Trial has a recall of 70%, indicating a very good detection rate of actual events.

Both the precision and the recall produced an overall f-score of 0.51, revealing an average performance by the system. The individual performances of most of the events were over 0.5 while the events of Visit and Public had a very low, sub 0.15, f-score, indicating very poor performance.

By analyzing the results of the evaluation of relation extraction in Table 9, it is possible to discern issues with the events of Visit and Public. The high values of their evaluation can be attributed to the very low number of events whose relations were evaluated in this phase. Yet, for the other events it is possible to discern that Trial had 100% precision while having the lowest recall of 30%, meaning that, although it failed to extract most of the relations, it correctly identified the ones it did. This resulted on this event having the lowest f-measure score of 46%. The Crime event, also had a very high precision of 93% while failing in extracting half of its relations. The Trial event, on the other hand, was able of extracting most of its relations, having a recall of 93%, failing just on the correct identification of a few (14%).

8 CONCLUSIONS

The objective of this work was to develop the capabilities of the STRING chain to identify and classify new event types. The process of development began with the review of what can be considered to be an event in the context of STRING. Then, the process of reviewing what kinds of information could be considered to be relevant enough to be extracted began. This resulted in a list of well defined relations that can be extracted, for each event type. Some of these relations became core for the event they were part of, which made it so that the event could only be extracted if it was possible to extract that relation.

In total, three different categories of events were developed, i.e. the Crime, which included the Crime event, Trial event and Prision event, the Location Category, which included the Static event and the Visit event, and the Public category which included the Public event and the Ephemerid event.

The development of the system resulted in a more coordinated set of rules which lead to methods of simplification of the methods of writing new rules.

In the end, the work resulted in a set of 608 new rules was created split among the three dedicated dependency files and the addition of over 450 new entries to the existing lexicon.

The evaluation of the system revealed an average f-score measure of 51%, indicating an average performance on event identification while, for relation identification, it yielded a overall f-score measure of 74%, which was influenced by the performance values of two under-evaluated events, which increased the performance value.

ACKNOWLEDGMENTS

First I would like to thank my supervisor, Professor Nuno Mamede, for the guidance provided throughout this work, whose experience and help made this work possible.

I would, also, like to thank my co-supervisor, Professor Jorge Baptista, for the insight on several topics and the attention to detail that helped the improvement of this work.

REFERENCES

[1] [n. d.]. CÃşdigo Penal. http://codigopenal.pt/. Accessed: 2018-05-29.
[2] [n. d.]. Twitter Streaming API. = https://developer.twitter.com/en/docs.
[3] S. Ait-Mokhtar, J. P. Chanod, and C. Roux. 2002. Robustness Beyond Shallowness: Incremental deep parsing. *Natural Language Engineering* (2002).
[4] Mohammad Al-smadi and Omar Qawasmeh. 2016. Knowledge-based Approach for Event Extraction from Arabic Tweets. (2016).
[5] James Allan (Ed.). 2002. *Topic Detection and Tracking: Event-based Information Organization.* Kluwer Academic Publishers.
[6] Jorge Baptista, Nuno Mamede, Caroline Hagège, and Andreia Maurício. 2011. *Time Expressions in Portuguese Guidelines for Identification, Classification and Normalization.* Technical Report. L2F-Spoken Language Laboratory.
[7] Jorge Baptista, Diogo Oliveira, Daniel Santos, and Nuno João Mamede. 2011. *Classification directives for named entities in Portuguese texts.* Technical Report. L2F-Spoken Language Laboratory.
[8] Daniel M Bikei, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1991. Nymble : a High-Performance Learning Name-finder. (1991).
[9] Linguistic Data Consortium. 2008. ACE Arabic Annotation Guidelines for Entities. *Facilities* (2008).
[10] Hilda Hardy, Vika Kanchakouskaya, and Tomek Strzalkowski. 2006. Automatic Event Classification Using Surface Text Features. *Proceedings of AAAI06 Workshop on Event Extraction and Synthesis* (2006).
[11] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. 2011. An Overview of Event Extraction from Text. *CEUR Workshop Proceedings* (2011).
[12] Nuno J. Mamede, Jorge Baptista, Cláudio Diniz, and Vera Cabarrão. 2012. STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. (April 2012).
[13] Olivia March and Timothy Baldwin. 2008. Automatic Event Reference Identification. *Proceedings of the Australasian Language Technology Association Workshop 2008* (2008).
[14] Diogo Oliveira. 2010. *Extraction and Classification of Named Entities.* Master's thesis. Instituto Superior Técnico, Universidade Técnica de Lisboa. MSc Dissertation.
[15] Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming First Story Detection with Application to Twitter. *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference* June (2010).
[16] James Pustejovsky, B. Ingria, and R. Sauri. 2005. The Specification Language TimeML. *The language of time: A reader* (2005).
[17] N Stokes, J Rong, B Laugher, Y Li, and L Cavedon. 2007. NICTA's Update and Question-based Summarisation Systems at DUC 2007. *of the Document Understanding Conference Workshop* (2007).
[18] Ian H. Witten, Eibe Frank, and Mark Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques.*