Using Machine Learning for Extracting Information from Natural Disaster News Reports

Usando Aprendizaje Automático para Extraer Información de Noticias de Desastres Naturales

Alberto Téllez Valero, Manuel Montes y Gómez and Luis Villaseñor Pineda
Laboratorio de Tecnologías del Lenguaje, Coordinación de Ciencias Computacionales,
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE).
Luis Enrique Erro #1, Tonantzintla, Puebla, México.
{albertotellezv, mmontesg, villasen}@ccc.inaoep.mx

Article received on July 17, 2008; accepted on April 03, 2009

Abstract.
The disasters caused by natural phenomena have been present all along human history; nevertheless, their consequences are greater each time. This tendency will not be reverted in the coming years; on the contrary, it is expected that natural phenomena will increase in number and intensity due to the global warming. Because of this situation it is of great interest to have sufficient data related to natural disasters, since these data are absolutely necessary to analyze their impact as well as to establish links between their occurrence and their effects. In accordance to this necessity, in this paper we describe a system based on Machine Learning methods that improves the acquisition of natural disaster data. This system automatically populates a natural disaster database by extracting information from online news reports. In particular, it allows extracting information about five different types of natural disasters: hurricanes, earthquakes, forest fires, inundations, and droughts. Experimental results on a collection of Spanish news show the effectiveness of the proposed system for detecting relevant documents about natural disasters (reaching an F-measure of 98%), as well as for extracting relevant facts to be inserted into a given database (reaching an F-measure of 76%).
Keywords: Machine Learning, Information Extraction, Text Categorization, Natural Disasters, Databases.

Resumen.
Los desastres causados por fenómenos naturales han estado presentes desde el principio de la historia del hombre; sin embargo, sus consecuencias son cada vez mayores. Esta tendencia podría no ser revertida en los próximos años; al contrario, se espera que los fenómenos naturales puedan incrementar en número e intensidad debido al calentamiento global. A causa de esta situación es de gran interés tener suficientes datos relacionados a los desastres naturales, ya que estos datos son absolutamente necesarios para analizar su impacto así como para establecer conexiones entre su ocurrencia y sus efectos. En correspondencia con esta necesidad, en este artículo describimos un sistema basado en métodos de Aprendizaje Automático que mejora la adquisición de datos de desastres naturales. Este sistema automáticamente llena una base de datos de desastres naturales con la información extraída de noticias de periódicos en línea. En particular, este sistema permite extraer información acerca de cinco tipos de desastres naturales: huracanes, temblores, incendios forestales, inundaciones y sequías. Los resultados experimentales en una colección de noticias en Español muestran la eficacia del sistema propuesto tanto para detectar documentos relevantes sobre desastres naturales (alcanzando una medida-F de 98%), así como para extraer hechos relevantes para ser insertados en una base de datos dada (alcanzando una medida-F de 76%).
Palabras claves: Aprendizaje Automático, Extracción de Información, Clasificación Temática de Textos, Desastres Naturales, Bases de Datos.

1 Introduction

The disasters caused by natural phenomena have been present all along human history. The impact of the damages caused by nature is greater each time, due to the world's population growth and the current climatic changes. According to world statistics provided by CRED[1], the increase in the number of world disasters between the decades

[1] Centre for Research on the Epidemiology of Disasters, url: http://www.cred.be/

of 1987-1996 and 1997-2006 was 60% (going from 4,241 to 6,806), whereas, the number of dead people during these periods increased a hundred percent (it went from more than 600,000 to more than 1,200,000). In the coming years, this tendency will not be reverted, it is foreseen that natural phenomena will increase in number and intensity due to the global warming, such as it was reported in the Fourth Assessment Report: Climate Change 2007[2].

The World Disasters Report 2005[3] points out that "the data about disasters establish the basis to reduce their own risk". More specifically, this report stresses the importance of having sufficient data related to natural disasters, since these data are absolutely necessary to analyze their impact as well as to create links between their occurrence and their effects. This way, *disaster databases* fulfill an important role in detecting tendencies and reducing future risks. These databases are very helpful, because they can be interpreted by analytical tools, as well as by early warning alert systems, where both contribute to define the order of priorities of the international action to reduce the risk of disasters. The above mentioned shows the importance of a systematic acquisition of data related to the impact of disasters, as recognized in the United Nations Hyogo Framework for Action 2005-2015[4].

The acquisition of this type of data is not trivial. In many occasions, it is not possible to quantify the impact of a natural phenomenon after several days or weeks. This is due to the dimensions of the catastrophe, or because it occurred in a remote area. Therefore, the work of the organizations in charge of monitoring and registering this information is very difficult. This is precisely the subject of the present work: *the enhancement of the acquisition process of natural disaster data*.

Due to technological progress, nowadays it is very easy to access different sources of information. For instance, the information about the characteristics and effects of natural disasters can be found in several online newspapers. This would allow an automatic system to monitor certain sites and obtain the information for populating a natural disaster database. For this purpose, the system should: (i) select the news belonging to the domain of interest; and (ii) identify from the selected news the data that have to be registered into the database. These tasks have been widely studied in the Artificial Intelligence field; specifically, they are known as Text Categorization (TC) and Information Extraction (IE), respectively. The objective of TC [Jackson & Moulinier, 2007] is to automatically determine the category (or topic) of a document within a previously defined set of possible categories; whereas, on the other hand, the goal of an IE system [Moens, 2006] is to identify and extract specific facts from natural language text.

In this paper, we present the architecture and evaluation of *TOPO*, a system that automatically extracts information from natural disaster news reports. This system is entirely based on a *machine learning approach* and its architecture consists of a set of components that, firstly, identify the texts related to natural disasters, and subsequently, extract the relevant data for populating a determined database. Its evaluation shows its effectiveness for detecting the relevant documents about natural disasters (reaching an F-measure of 98%), and for extracting relevant facts to be inserted into a given database (reaching an F-measure of 76%).

The rest of the paper is organized as follows. Section 2 summarizes the different approaches used for information extraction. Section 3 describes the *TOPO* system. Section 4 presents the evaluation results. Finally, Section 5 exposes our conclusions and some ideas for future work.

2 Related Work

The construction of databases about disasters caused by natural phenomena is not a recent issue. For instance, the previously cited World Disasters Report describes the four most important databases: EM-DAT[5], NatCat[6], Sigma[7], and Desinventar[8]. In the same report, it is mentioned that the acquisition of disaster data has significantly improved

[2] Report of the Intergovernmental Panel on Climate Change (IPCC), url: http://ipcc-wg1.ucar.edu/wg1/wg1-report.html

[3] Published by the International Federation of Red Cross and Red Crescent Societies, url: http://www.ifrc.org/publicat/wdr2005/

[4] From the final report of the World Conference on Disaster Reduction, url: http://www.unisdr.org/wcdr/intergover/official-doc/L-docs/Hyogo-framework-for-action-english.pdf

[5] Database maintained by the CRED, url: http://www.emdat.be/

[6] Database maintained by Munich Reinsurance Company, url: http://mrnathan.munichre.com/

[7] Database maintained by Swiss Reinsurance Company, url: http://www.swissre.com/

[8] Database maintained by the Social Studies Network for Disaster Prevention in Latin America, url: http://www.desinventar.org/

in the last 20 years. Nevertheless, this report also stresses the fact that there are still challenges to sort out in the *systematization of data acquisition*. One of these challenges is the cost of the acquisition by itself, where human experts must fill and check local databases. Other problem is the transmission of the information from local databases to form part of general repositories. In this phase, the people in charge of transmitting the information could be tempted to exaggerate or omit data in order to take professional, political or economical advantage of the situation. These problems evidence the great necessity to have automatic or semi-automatic methods for the acquisition of this kind of data.

In particular, the Information Extraction (IE) task focuses on the automatic acquisition of data from free text documents. In other words, the purpose of an IE system is to structure the information contained in documents that are relevant for the study of a particular scenario, called extraction domain [Riloff & Jeffrey, 1999]. The idea is to identify from a text the relevant information units for a determined application. In this case, where the extraction domain is natural disasters, the relevant units would be data like: the place of impact of the natural phenomenon, the number of people affected, the material damage, etc. Once these data have been identified, they will be used to populate a database.

As supposed, the major challenge of an IE system is to have the ability to identify, in an automatic way, the relevant information units. Since we are still far from achieving the automatic understating of language, current solutions for this problem rely on the application of extraction patterns that capture its common regularities. Basically, an extraction pattern denotes the context used by an author to introduce or describe a piece of information. Once it has been decided which pieces of information will be extracted, it is possible to locate the contexts that were used to express them. In this way, each time an extraction pattern matches a text fragment, it is possible to obtain a fact of interest. For instance, when the following lexical pattern "<*NUMBER people to evacuate*>" matches the fragment "<*The hurricane Isidore lashed the Yucatan Peninsula with driving rain and huge waves, forcing 70,000 people to evacuate*>", it is possible to establish the number of people that were evacuated due to the natural phenomenon.

There are two tasks that are essential for building an IE system [Bouckaert, 2002; Hobbs, 1992]: first, to determine the set of information units that will be extracted; and second, to establish the set of extraction patterns. In both cases, human intervention is necessary. In particular, it is required that an expert in the application domain determines the units of information that have to be extracted. Besides, human intervention is also necessary to recognize the extraction patterns. In this case, human intervention will somehow be significant depending on the level of abstraction used to express the extraction patterns. There are two levels of abstraction: a poor level, where only words are used to describe the patterns (as shown in the former example), and a rich level, where syntactic information is used. The more abstract a pattern is, the more applicable it will be. On the other hand, at a lower abstraction level, patterns will be more specific, and will be only applicable in particular cases. Unfortunately, the intervention of a trained staff in linguistics is necessary to establish a syntactic pattern. On the contrary, any person is able to determine a lexical pattern.

It is important to mention that IE systems are highly specific to the application domain, and therefore, that the contribution of the experts cannot be reutilized in new scenarios. Because of this situation, research in IE is mainly focused on the *automatic discovery of the extraction patterns* [Muslea, 1999; Peng, 1999; Stevenson & Greenwood, 2006; Turno, 2003]. In particular, modern IE approaches are supported on machine leaning (ML) techniques [Ireson *et al.*, 2005]. This kind of systems can be classified in the following three categories:

Rule Learning. This approach relies on a symbolic inductive learning process, where extraction patterns represent training examples by means of relations among textual elements. In this case, it is possible to distinguish two different kinds of IE systems: those using propositional learning [Riloff 1996; Sonderland, *et al.*, 1995] and those based on relational learning [Freitag, 1998; Sonderland, 1999]. Due to the generality of this approach, it has been used to learn from structured, semi-structured and free-text documents.

Linear Separators. In this approach the classifiers are learned as sparse networks of linear functions (i.e., linear separators of positive and negative examples). It has been commonly used to extract information from semi-structured documents [Roth & Yih, 2001], and has been applied in problems such as: affiliation identification and citation parsing [Bouckaert, 2002], extraction of data from job ads [Zavrel, *et al.*, 2000], and detection of an e-mail address change [Kushmerick, *et al.*, 2001].

Statistical Learning. This approach is focused on learning Hidden Markov Models (HMMs) as useful knowledge to extract relevant fragments from documents [Scheffer, *et al.*, 2001]. For instance, [Seymore, *et al.*, 1999] presents a method for learning model structure from data in order to extract a set of fields from semi-structured texts.

In particular, *TOPO*, our IE system, follows the *linear separator approach*. It is mainly based on the hypothesis that looking at the word combinations around the relevant data is enough for learning the required extraction patterns, or strictly speaking, a function that maps input data to a predefined set of categories. Its main advantage, compared to other methods from the same approach, is that it does not require applying a deep linguistic analysis of texts in order to generate the classification features. Instead, it exclusively relies on the use of features expressed at lexical level. This characteristic allows our method to be easily adapted to specific domains and languages, where current language processing tools tend to show a relative poor performance.

In addition, it is also important to comment that, different from previous linear-separator methods that have been only applied to semi-structured domains, our method is specially suited to extract information from free text documents. Hence, in some degree, our research attempts to empirically determine the limits of this approach when dealing with a complex domain and free texts instead of semi-structured documents.

3 The proposed IE System

This section presents *TOPO*, a system specially suited to extract information related to natural disasters from Spanish newspaper articles. This system considers five different types of disasters[9]: hurricanes, forest fires, inundations, droughts, and earthquakes; and allows extracting the following kind of information:

- Information related to the disaster itself such as the date, place and magnitude of the phenomenon.
- Information related to the people, for instance: the number of dead, wounded, missing, damaged and affected persons.
- Information related to the infrastructure: the number of destroyed and affected houses.
- Information related to the economic impact, such as the number of affected hectares, monetary lost, among others.

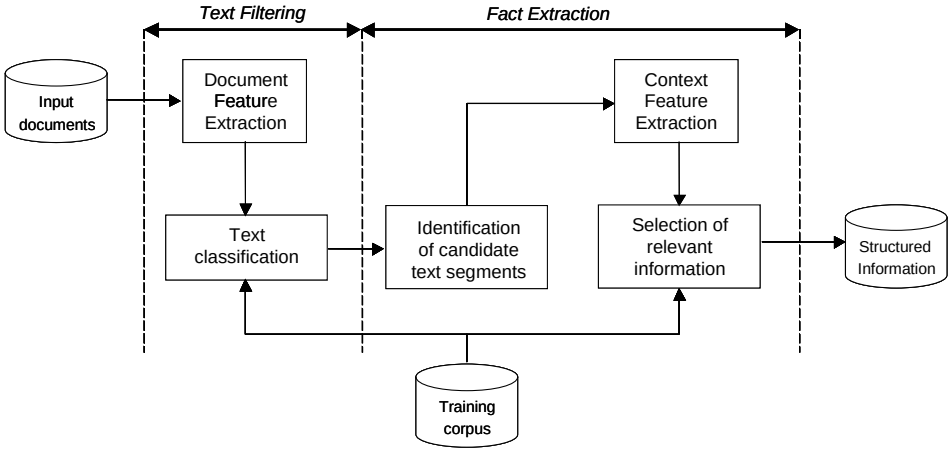

Fig. 1. General architecture of our information extraction system

Figure 1 shows the general architecture of the proposed system. It includes two main modules: text filtering and fact extraction. The first module focuses on the selection of news reports about natural disasters, whereas, the second

[9] The most frequent natural disasters in Mexico, according to the DesInventar database.

considers the extraction of relevant data from the selected reports. The most important characteristic of this architecture is the application of the same kind of solution in both modules. In particular, both of them apply a *supervised classification approach* based on the use of lexical features.

Supervised classification is a ML technique that generates a function from training data that allows mapping input objects to a set of predefined categories [Mitchell, 1997]. Three issues are very important in the construction of any classifier: first, the definition of the input objects and their possible categories; second, the definition of the number and kind of features that will be used to represent the objects; and third, the selection of the learning algorithm that will be used to generate the mapping function. The following sections describe the characteristics of the classifiers involved in the two modules of the proposed architecture.

3.1 Text Filtering

As we previously mentioned, the purpose of this module is to select the documents about natural disasters from a set of news reports. In particular, this module considers the classification of news reports in six different categories, one for each type of natural disaster (hurricanes, forest fires, inundations, droughts, and earthquakes) and other one for non relevant documents. Figure 2 illustrates the function of this module.

This module performs a *typical classification process* [Joachins, 2002; Sebastiani, 2002]. First, each input document is represented using a predefined set of features, and then, a classifier, which was previously trained for this particular task, decides the category of the input document. In particular, in our experiments (refer to Section 4), we considered the whole vocabulary from the training set as document features, we used two different weighting schemes (Boolean and term frequency) and applied three different leaning algorithms, namely, Support Vector Machines (SVM), Naïve Bayes (NB), and C4.5.

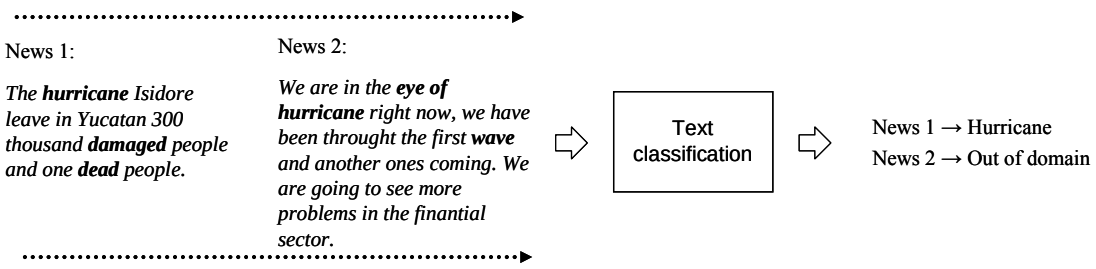

Fig. 2. Filtering of relevant documents

3.2 Fact Extraction

The purpose of this second module is to extract the relevant data from the selected news reports. The design of this module is supported on the idea that looking at the word combinations around the interesting data is enough to decide about its category. Based on this hypothesis, this module considers the following two processes:

1. The identification of all text segments having some possibility of being part of the output template, and
2. The selection of the relevant segments based on their context information.

Figure 3 illustrates the function of this module by means of a short example.

Identification of candidate text segments

The objective of this process is to extract all text segments that have some possibility for being part of the extraction template (i.e., the output database). Given that most relevant data about natural disaster are factoid, this process exclusively focuses on the identification of proper names, quantities and dates.

In particular, the identification of the candidate text segments is done by means of a *regular expression analysis*. This kind of analysis is very robust and simple, but allows achieving very high recall rates.

The first part of Figure 3 shows this process. In this case, the uppercase words correspond to the candidate text segments of the input text, and the rest of the words indicate their lexical context. That is, "ISIDORE" is a candidate text segment (i.e., a possible relevant data) and the set of words "the, hurricane, leave, in" represents its context.

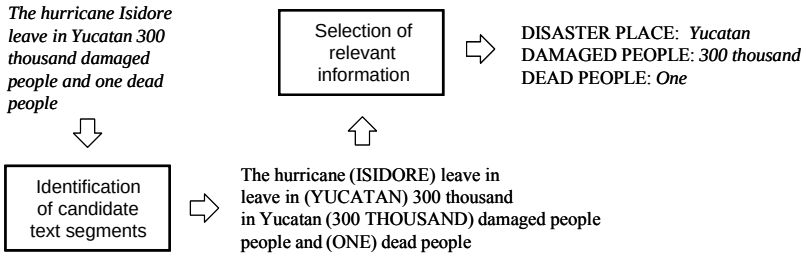

Fig. 3. Extracting fact information from natural disasters news reports

Selection of relevant information

The purpose of this process is to extract the text segments relevant for a predefined output template; in other words, it is responsible of classifying the candidate text segments into one of the following thirteen categories: disaster date, disaster place, disaster magnitude, number of dead people, number of wounded people, number of damaged people, number of affected people, number of missing people, number of destroyed houses, number of affected houses, number of affected hectares, economic lost, and non-relevant.

This process, similar to the text filtering module, is also defined as a *classification task*. However, in this case, the input objects are not documents but candidate text segments (the set of proper names, quantities and dates that were previously identified), and the features correspond to the words from their immediate neighborhoods and not to the whole collection vocabulary. The second part of Figure 3 illustrates this process, where the text segment (ISIDORE) was discarded and (YUCATAN PENINSULA) and (70,000) were classified as the disaster place and the number of affected people respectively.

In our experiments we used three different learning algorithms (SVM, Naïve Bayes and C4.5) and seven different context sizes. In particular, we searched for the configuration that allows obtaining the highest recall rate, and therefore, that allows implementing a useful support tool for a human analyst.

4 Experimental Evaluation

4.1 The Disaster News Corpus

In order to build and evaluate our IE system we assembled a corpus of Spanish news reports about natural disasters. This corpus was obtained from several online Mexican newspapers; it gathers news reports from 1996 to 2004.

Given that our IE system is based on a supervised learning approach, we manually tagged all collected news. The tagging was done from two perspectives. On the one hand, each news report was thematically tagged in accordance to the five categories of interest[10]. This tagging was necessary for building the text filtering module. On the other hand, each proper name, quantity and date happening in the disaster reports was tagged based on the thirteen categories from the extraction template[11]. This information was useful for the construction of the fact extraction module.

Table 1 shows some general numbers from the collected corpus. It is interesting to notice the existence of irrelevant instances. In the case of text filtering, these instances correspond to documents that are out of the domain of natural disasters but that use similar words to the relevant news. Whereas, in the case of fact extraction, these instances correspond to the set of named entities –proper names, quantities and dates– that occur in the disaster reports but that are irrelevant for the extraction template.

[10] Hurricanes, earthquakes, forest fires, inundations, and droughts.

[11] Disaster date, disaster place, disaster magnitude, number of dead people, number of wounded people, number of damaged people, number of affected people, number of missing people, number of destroyed houses, number of affected houses, number of affected hectares, economic lost, and non-relevant.

Other relevant fact about these data is that they maintain the natural distribution of the categories. For instance, we have more instances about earthquakes than about droughts, and we also have much more irrelevant named entities than relevant ones.

Table 1. The disaster news corpus

(a) Tagging for text filtering

	Relevant instances	Irrelevant instances
Hurricane	76	23
Forest fire	92	53
Inundation	87	58
Drought	41	32
Earthquake	143	63
Total	*439*	*229*

(b) Tagging for fact extraction

	Relevant instances	Irrelevant instances
Name facts	966	4853
Date facts	86	147
Quantity facts	973	2926
Total	*2025*	*7926*

This corpus, as we previously mentioned, was specially gathered for building and evaluating our IE system. In order to do that it was divided in two equal parts. The first one was used for training and tuning the system, and the second for its evaluation. The following section presents some results from these two phases.

We applied this evaluation procedure based on Salzberg's suggestions [Salzberg, 1999], which advised to do not use a fold-cross validation strategy over the complete corpus in order to avoid corrupting the validity of results by including a repeated tuning. Nevertheless, it is important to mention that we certainly applied a 10-fold-cross validation strategy during the experiments regarding the system tuning.

4.2 System Tuning

In order to determine the most appropriate configuration of our system for the task at hand, we performed some experiments using a 10-fold-cross validation over the given training set.

The first set of experiments focused on the text filtering module. In this case, we considered the traditional representation of the documents (vectors indicating the presence or absence of certain words), but evaluated two different weighting schemes as well as three different learning algorithms. Specifically, we evaluated the Boolean and Term Frequency weightings, and the SVM, Naïve Bayes and C4.5 learning algorithms. In all cases we applied the Information Gain technique for feature selection, obtaining a set of only 648 features (with IG > 0) from the whole vocabulary of the training corpus. Table 2 shows the results from these experiments. They indicate that the combination of the Boolean weighting and the SVM algorithm allowed obtaining the best classification accuracy (0.92), and that this result significantly outperformed[12] the results from the other two learning approaches.

Table 2. Classification accuracy in the text filtering module

	Weighting Scheme	
	Term Frequency	Boolean
Naïve Bayes	0.86	0.87
C4.5	0.87	0.86
SVM	0.90	**0.92**

The second set of experiments focused on the term extraction module. The purpose of these experiments was to determine the most appropriate context size as well as the best classification algorithm for the selection of the relevant text segments. Mainly, we attempted to obtain some information for constructing a specific classifier for each different type of output data (i.e., one for the proper name facts, one for the quantity facts and other for the date facts). Table 3 shows the results from these experiments. These results indicate that the SVM algorithm is the best

[12] We evaluated the statistical significance of the results using the *t-test* and a $p = 0.005$.

option for all kinds of data, and that its best result for each data type significantly outperformed[13] the results from the other two learning algorithms. In addition, these results also indicate that a context of eight words (four to the right and four to the left) is the best selection for the classification of proper names and dates, whereas, a context of only six words was better for the case of quantities.

4.3 Evaluation Results

For the evaluation of our system we used the test segment of the collected corpus. For this last experiment, we configured the system as detailed in the previous section, that is: for text filtering we applied a SVM classifier using Boolean weights, and for term extraction we used a SVM classifier considering eight context words for names and dates and six context words for quantities.

The evaluation of the system relied on traditional metrics for text classification and information extraction: precision, recall and F-measure [Jackson & Moulinier, 2007]. Precision is a measure of correctness, and recall is a measure of completeness. That is, precision is defined as the fraction of correct instances among those instances that are extracted by the model, whereas, recall is defined as the fraction of correct instances extracted by the model over the total number of instances that exist in the data set. Finally, F-measure is defined as the weighted harmonic mean of precision and recall.

Table 3. Classification accuracy in the fact extraction module

	Context Window Size (in number of words)						
	2	4	6	8	10	12	14
	Proper name facts						
Naïve Bayes	0.69	0.70	0.70	0.71	0.72	0.72	0.71
C4.5	0.50	0.69	0.69	0.69	0.69	0.69	0.65
SVM	0.68	0.72	0.74	**0.75**	0.74	0.73	0.69
	Date facts						
Naïve Bayes	0.97	0.97	0.97	0.96	0.96	0.95	0.94
C4.5	0.97	0.97	0.97	0.97	0.97	0.97	0.97
SVM	0.98	0.98	0.98	**0.99**	0.99	0.98	0.98
	Quantity facts						
Naïve Bayes	0.64	0.66	0.66	0.66	0.65	0.65	0.64
C4.5	0.74	0.76	0.76	0.77	0.77	0.77	0.77
SVM	0.76	0.79	**0.81**	0.80	0.79	0.79	0.79

The following paragraphs describe the results of our system in the tasks of text filtering and fact extraction.

Text Filtering. As it can be noticed in Table 4, the results for text filtering are very accurate. We consider that two main reasons allowed achieving this high performance. On the one hand, the given problem only consists of five categories, and on the other hand, these categories are easily differentiable since their keywords are considerably different. For instance, the most relevant words for the category of forest fires were fireman, fire and hectares, whereas, the most relevant words for earthquakes were Richter, hit, and epicenter.

From this table, it is also possible to notice that the results from all categories are almost equally satisfactory. The worst result corresponds to Droughts (an F-measure of 92%). In this case, we attribute this inferior performance to the lack of sufficient training examples.

[13] For these experiments we also evaluated the statistical significance of the results using the *t-test* and a $p = 0.005$

Table 4. Evaluation results from the text filtering module

	Precision	Recall	F-measure
Hurricane	1.00	1.00	1.00
Forest fire	0.98	1.00	0.99
Inundation	1.00	1.00	1.00
Drought	0.88	0.97	0.92
Earthquake	0.96	0.98	0.97
Average	0.96	0.99	0.98

Fact extraction. Table 5 resumes the results from the second module of the system. These results are clearly not as good as those from the text filtering module. We presume that this performance is because this second module includes a greater number of categories, and also because these categories are much more similar among each other. For instance, the contexts about the number of destroyed and affected buildings are almost the same. In order to improve this performance we consider that it will be necessary to collect a greater training corpus.

It is interesting to notice that for all categories the recall rates are better than the precision scores. This fact indicates that our system could extract most of the relevant information from the news reports, but that it also extracts several irrelevant data. However, despite the modest results achieved in this module, we consider that this first prototype of *TOPO* can be a useful tool for extracting information from natural disaster news reports. In particular, it can be of great utility for a human analyst, since given its high recall it can be used as a first extraction step.

Finally, it is important to mention that, in spite of not having any comparable system (since each one is focused on a different application domain), our results are equivalent to other recently reported. For instance, at the "Pascal Challenge on Evaluating Machine Learning for Information Extraction" [Ireson, *et al.*, 2005], there were evaluated 20 different IE systems in the task of extracting eleven different facts from the call for papers of 200 workshops. In this evaluation exercise, the best system achieved an F-measure of 73%, whereas, the average F-measure was of 53%. On the other hand, our results are also similar to those reported at [Li, *et al.*, 2005], where there were evaluated different supervised IE approaches on the task of extracting information from job offers. In this task, the reported F-measures ranged from 76% to 81%. Nevertheless, it is important to point out that all considered approaches used more sophisticated language processing tools than our system.

Table 5. Evaluation results from the term extraction module

	Precision	Recall	F-measure
Disaster date	0.95	0.99	0.97
Disaster place	0.50	0.60	0.55
Disaster magnitude	0.73	0.96	0.83
Number of dead people	0.66	0.73	0.69
Number of wounded people	0.91	0.84	0.87
Number of missing people	0.78	0.93	0.85
Number of damaged people	0.62	0.84	0.71
Number of affected people	0.60	0.82	0.69
Number of destroyed buildings	0.70	0.83	0.76
Number of affected buildings	0.72	0.83	0.77
Number of affected hectares	0.66	0.92	0.77
Economic lost	0.49	0.95	0.65
Average	0.69	0.85	0.76

5 Conclusions and Future Work

This paper presented some ideas for enhancing the *acquisition process of natural disaster data*. In particular, it proposed a system that automatically populates a natural disaster database by extracting information from online news reports. The construction of this kind of systems is of great relevance, since these data are absolutely necessary to analyze the impact of natural disasters as well as to establish links between their occurrence and their effects, and

therefore, to reduce their future risks. Regarding this issue, it is important to mention that, to our knowledge, our proposal is the first effort to automate this task.

The proposed system is entirely based on a *machine learning approach* and its architecture includes two main modules, one for text filtering and other for fact extraction. The first module focuses on the selection of news reports about natural disasters, whereas, the second considers the extraction of relevant data from the selected reports. The experimental results demonstrated the pertinence and potential of this solution; using a very small training set it was possible to achieve an *F*-measure of 98% in the detection of documents about natural disasters, and an *F*-measure of 76% in the extraction of relevant data from these documents.

It is also important to comment that although our method applies traditional machine learning techniques, it differs from other previous IE systems in that *it does not depend on sophisticated resources* for natural language processing. In particular, it only uses lexical features and avoids the usage of complex syntactic attributes. In consequence, our system can be easily adapted to specific domains and languages, where current language processing tools tend to show a poor performance.

The proposed system has two main disadvantages. On the one hand, it does not correctly extract information from documents describing more than one disaster, and on the other hand, it does not allow integrating the data extracted from different documents related to the same disaster. Our future work will be mainly focused on the solution of these two inconveniences. In addition, it will consider the collection of a bigger training set and the construction of a set of binary classifiers, one for each kind of desired data.

Acknowledgments

This work was partially supported by Conacyt through research grants (CB-61335, CB-82050 and CB-83459) and scholarship (171610).

References

1. **Bouckaert, R. (2002).** "Low level information extraction". In *Proceedings of the workshop on Text Learning* (TextML-2002), Sydney, Australia.
2. **Cowie, J. and Lehnert, W. (1998).** "Information Extraction". *Communications of the ACM*, Vol. 39, No. 1, pp. 80-91
3. **Freitag, D. (1998).** "Machine Learning for Information Extraction in Informal Domains". *Ph.d. thesis*, Computer Science Department, Carnegie Mellon University.
4. **Hobbs, J. R. (1992).** "The Generic Information Extraction System". In B. Sundheim, editor. *Fourth Message Understanding Conference (MUC-4)*, Mc Lean, Virginia, June. Distributed by Morgan Kauffman Publishers, Inc., San Mateo, California.
5. **Ireson, N., Ciravega, F., Califf, M. E., Freitag, D., Kushmerick, N., and Labelli, A. (2005).** "Evaluating Machine Learning for Information Extraction", In *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany.
6. **Jackson, P. & Moulinier, I. (2007).** "Natural Language Processing for Online applications: text retrieval, extraction and categorization". John Benjamins Publishing Co, second edition, June.
7. **Joachins, T. (2002).** "Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms". Kluwer Academic Publishers, May.
8. **Kushmerick, N., Johnston, E., and McGuinness, S. (2001).** "Information Extraction by Text Classification". *Seventeenth International Join Conference on Artificial Intelligence (IJCAI-2001)*, N. Kushmerick Ed. Adaptive Text Extraction and Mining (Working Notes), Seattle, Washington , pp. 44-50.
9. **Li, Y., Bontcheva, K., and Cunningham, H. (2005).** "SVM Based Learning System for Information Extraction". In *Proceedings of Sheffield Machine Learning Workshop,* Lecture Notes in Computer Science. Springer Verlag.
10. **Mitchell, T. (1997).** "Machine Learning". McGraw Hill.

11. **Moens M. (2006).** "Information Extraction: Algorithms and Prospects in a Retrieval Context". Springer (Information retrieval series, edited by W. Bruce Croft), October.
12. **Muslea, I. (1999).** "Extraction Patterns for Information Extractions Tasks: A Survey". In *Proceedings of the AAAI Workshop on Machine Learning for Information Extraction*, July, Orlando, Florida.
13. **Peng, F. (1999).** "Models Development in IE Tasks - A survey". CS685 (Intelligent Computer Interface) course project, Computer Science Department, University of Waterloo.
14. **Riloff, E. (1996).** "Automatically Generating Extraction Patterns from untagged text". In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI)*, pp. 1044-1049.
15. **Riloff, E. & Jeffrey L. (1999).** "Extraction-based text categorization: Generating domain-specific role relationships automatically". In Tomek Strzalkowski (Ed.), *Natural Language Information Retrieval* (pp. 167-196). Dordrecht, The Netherlands: Kluwer Academic Publishers.
16. **Roth, D. & Yih, W. (2001).** "Relational Learning Via Propositional Algorithms: An Information Extraction Case Study". In *Proceedings of the 15th International Conference on Artificial Intelligence (IJCA-01I)*, Morgan Kauffman Publisher, Inc., San Francisco, California, pp. 1257-1263.
17. **Salzberg, S. L. (1999).** "On Comparing Classifiers: A Critique of Current Research and Methods". *Data Mining and Knowledge Discovery*, 1:1-12.
18. **Scheffer T., Decomain C., & Wrobel S. (2001).** "Active hidden Markov models for information extraction". *Lecture Notes in Computer Science*, Vol. 2189, Springer, pp. 309-318.
19. **Sebastiani, F. (2002).** "Machine Learning in Automated Text Categorization". *ACM Computing Surveys.* 34(1):1-47.
20. **Seymore, K., McCallum, A., & Rosenfeld, R. (1999).** "Learning Hidden Markov Model structure for Information Extraction". In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI)*, pp. 37-42.
21. **Sonderland, S., Fisher, D., Aseltine, J., & Lehnert, W. (1995).** "CRYSTAL: Inducing a Conceptual Dictionary". In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1314-1321.
22. **Sonderland, S. (1999).** "Learning Information Extraction Rules for Semi-Structured and Free Text". *Machine Learning*, No. 34, pp. 233-272.
23. **Stevenson M. & Greenwood M. A. (2006).** "Comparing Information Extraction Pattern Models", In *Proceedings of the Workshop on Information Extraction Beyond The Document*, Association for Computational Linguistics, Sydney, pp.12-19.
24. **Turno, J. (2003).** "Information Extraction, Multilinguality and Portability". *Revista Iberoamericana de Inteligencia Artificial*, No. 22, pp. 57-78.
25. **Zavrel, J., Berck, P., & Lavrijssen, W. (2000).** "Information Extraction by Text Classification: Corpus Mining for Features". In *Proceedings of the workshop Information Extraction meets Corpus Linguistics*, Athens, Greece.

Alberto Téllez Valero *PhD student at the Computational Sciences Department of the National Institute of Astrophysics, Optics and Electronics, located in Puebla, Mexico. His areas of interest include information extraction and retrieval, text classification, question answering, answer validation, and textual entailment recognition.*

Manuel Montes y Góme *He is a full-time lecturer at the Computational Sciences Department of the National Institute of Astrophysics, Optics and Electronics, located in Puebla, Mexico. He obtained his PhD in Computer Science from the Computing Research Center of the National Polytechnic Institute, Mexico. His research interests include information extraction, information retrieval, text classification, question answering, and text mining.*

Luis Villaseñor Pineda *He obtained his PhD in Computer Science from the Université Joseph Fourier, Grenoble, France. Since 2001, he is professor of the Department of Computational Sciences of the National Institute of Astrophysics, Optics and Electronics, located in Puebla, Mexico. His areas of interest include thematic and non-thematic text classification, lexical semantics, and information retrieval.*