Journal of Theoretical and Applied Information Technology
30th June 2012. Vol. 40 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

GLEANING DISASTER RELATED INFORMATION FROM WORLD WIDE WEB USING GATE

[1]IJAZ AHMED, [2]QAZI MUDASSAR ILYAS, [3]JAVERIA AJOON, [4]MEHTAB AFZAL

[1]Asstt Prof., Deptt. of CS, COMSATS Institute of Information Technology, Abbottabad, Pakistan
[2]Asstt. Prof., Deptt. of IS, College of Computer Sciences and IT, King Faisal University, Saudi Arabia
[3]Deptt. of CS, COMSATS Institute of Information Technology, Abbottabad, Pakistan
[4]School of Information and Computer Science, Southwest Jiaotong Univeristy, Chengdu, China

E-mail: [1]Ijazahmad@ciit.net.pk,[2]qilyas@kfu.edu.sa, [3]javariaajoon@ciit.net.pk, [4]mehtabafzal@gmail.com

ABSTRACT

SAHARA is a framework proposed by the authors of this paper to integrate Semantic Web and Natural Language Processing tools, to timely collect and disseminate disaster information to stakeholders to help in disaster management. This paper is related with information extraction component of SAHARA and presents a set of rules developed in GATE to extract disaster related information from online text resources. The developed pattern-action rules can be used to extract disaster entities including disaster location, type, magnitude, date and number of dead, injured, lost, homeless and affected people. A corpora is developed for various types of disasters such as earthquakes, hurricanes, floods, tsunami, forest fires, suicide bombing and military operations. The developed rule set is tested against this corpora. We achieved varying results for overall precision, recall and f-measure of extracted entities. The best results were achieved for disaster magnitude and the worst for date and time.

Keywords: *Information Extraction, NLP (Natural Language Processing), GATE (General Architecture For Text Engineering), Semantic Disaster Management System*

1. INTRODUCTION

Disasters, whether natural or human inflicted, have always affected mankind severely. Timely collection and dissemination of information helps to mitigate the affects of the disaster. Current disaster management solutions [6] [12] are inadequate to help manage disaster due to manual data collection and entry; thus resulting in delayed dissemination of information. The idea is to (semi) automate the data collection and dissemination process. The proposed automated process includes:

1) Automated identification of the WWW sources that contains disaster information, i.e the need to develop a crawler;

2) Automated extraction of disaster information from identified sources (web pages) through IE (Information Extraction)

3) Automated representation of disaster information into concepts, i.e the need to develop disaster ontology and

4) Automated dissemination of the extracted information to the stakeholders.

In a previous work, the authors have proposed SAHARA, a framework that outlines a conceptual architecture of a semi-automated disaster management system, based on state-of-art SW (Semantic We) and NLP (Natural Language Processing) tools and techniques [11]. SAHARA provides details of the major components of a semi automated disaster management system that includes Crawler, Information Extraction, Disaster Ontology and Query Analyzer. Additionally, it highlights challenges of a semi automated disaster management system that ranges from accuracy of extracted information to timely dissemination of information. In another previous work [1], the authors have developed a disaster ontology[1] that covers the important concepts of disaster domain and their relationships. It is pertinent to mention here that the existing disaster ontologies lack the

[1] http://www.yso.fi/onto/disaster/

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

comprehensives that we are looking for, to represent the domain of disaster.

The work presented in this paper is continuation of the previous work. This paper presents the details of an automated Information EXtraction(IE) system, based on NLP tools and techniques. The general idea is to extract the disaster related information from web pages. The paper presents (a subset of) developed rules to extract the information that includes disaster venue, disaster date, number of killed(affected) peoples and loss of infrastructure etc. The main challenge is the accuracy of the extracted information and paper also provides results, to evaluate the accuracy of extracted information. We used the General Architecture for Text Engineering (GATE) tool to extract the disaster information from text. Rest of the paper is organized as follows: Section 2 summarizes the previous work done in disaster management systems and information extraction. Basic concepts of developing rules in JAPE are given in section 3. Rules developed for disaster IE are presented in section 4. Section 5 gives results and conclusions.

2. RELATED WORK

Related work can be divided into two parts. First disaster management systems are discussed, be it very briefly, to emphasize motivation for IE for disaster management. Secondly, related most promising IE systems are discussed to give a measure of their usefulness for developing an IE system for disaster management. Sahana [6] is an open source disaster management system that has been widely used in the world for disaster management. Sahana has a modular structure for effective communication and information sharing among various stakeholders including government, NGOs and affected people. However, being a traditional database management system, it requires manual data entry. Global Disaster Information Network (GDIN)[2] is another conventional web based information system that provides effective communication. However it lacks in the effective management of disaster data. other notable disaster management systems are Queensland Disaster Management System[3], AusDIN [4], and Disaster Management Information System (DMIS)[4]. All these systems, share the common limitation, i.e. no support for automatic information collection. This is very important as time is the most precious entity in disaster management and sometimes even seconds can save lives. Regarding IE systems for disaster entities, the systems and frameworks worth mentioning include the following. TOPO [13] is an IE system that extracts disaster information from natural language text through Text Categorization. The system, however, is limited to Spanish language resources- thus making it unsuitable for dominant language on the world wide web i.e English. T-REX (The RDF Extractor) [2] is another IE system used to extract cultural and violent events information from text. The system has been used to extract violence information about different tribes living in Pakistan-Afghanistan borderland. The system also used RDF (Resource Description Framework) schema to organize and store information. TEXTRUNNER [3], KnowItAll [10] and and KnowItNow [5] are other general purpose IE system. All these system, however, do not offer any flexibility to be used for disaster related IE and an IE system specific for extracting disaster related information needs to be developed.

3. PRELIMINARIES

GATE (General Architecture for Text Engineering) [7] is an NLP tool, widely used in industry and academia to extract information from text. ANNIE (A Nearly New Information Extraction System) – a plug-in of GATE – is used to extract disaster information. ANNIE [9] uses PRs (Processing Resources) and PRs are developed using JAPE (Java Annotation Pattern Engine) language. JAPE [8] is a pattern/action rule language and is based on regular expressions. A JAPE rule has two parts pattern and action, commonly known as left and right part of the rule. Rule 1 shows a simple JAPE rule that describes the disaster type. We identified two different types of disaster in our system, natural disaster and human inflicted disaster.

JAPE uses gazetteer lists which organize dictionary and synonyms. The gazetteer lists are plain text files, with one entry per line. We developed significant gazetteer lists to capture various possible expressions for the same entity related to disaster, as shown in Table 1. We also used the existing gazetteer lists such as built-in gazetteer list, location.

[2] http://www.gdin.org/
[3] http://www.disaster.qld.gov.au/
[4] https://www-secure.ifrc.org/DMISII

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
Rule: NaturalDisaster
Priority: 5
(
{Token.string==~"[Ee]arthquake"} |
{Token.string==~"[Ff]lood"} |
{Token.string==~"[Tt]ornado"} |
{Token.string==~"[Dd]rought"} |
{Token.string==~"[Cc]yclone"} |
{Token.string==~"[Ss]torm"} |
{Token.string==~"[Tt]sunami"} |
{Token.string==~"[Bb]lizzard"}|
{Token.string==~"[Hh]urricane"}|
{Token.string==~"[Ll]andslide"}|
{Token.string==~"[Vv]olcano"} |
{Token.string==~"[Ff]orestfire"})
:N_Disaster
 -->
:N_Disaster.Natural_Disaster=
{kind = "Disaster", rule =
```

Rule 1 Extracting disaster type

Table 1 An excerpt from gazetteer lists developed

Gazetteer List	For extracting	Brief contents
Affected.lst	People affected	Affect affects affected victims …
Buildings.lst	Buildings damaged	houses destroyed houses collapsed schools destroyed schools fallen …
Dead.lst	People died	Dead death killed casualties lost their lives death toll …
...	...	...

4. EXTRACTION OF DISASTER RELATED INFORMATION

This section describes JAPE rules developed to extract information from text. We analyzed the text from blogs, wikiwiki and newspapers to understand the text patterns to describe disaster. The idea was to understand the disaster text patterns and then develop rules to extract the information, as accurate as possible. It is pertinent to mention that only a brief excerpt of the rules is presented in this paper for the sake of brevity.

4.1. Disaster Location and Type

Location is an important property of a disaster. It is noticed that usually, the initial news about disaster carry location information. For instance, a typical initial news on a wiki wiki about disaster has the text "*An earthquake of magnitude 7.2 has hit south-western Pakistan and at least 500 people died*". Rule 2 presents an abstract excerpt of the rule that extracts location of the disaster. The rule successfully extracted the location of disaster from given text i.e. Pakistan. Additionally with Rule 1 successfully extracted disaster type i.e. Natural_Disaster. The statement Lookup.Majortype == location uses the built-in gazetteer list, location. The gazetteer list, location includes the list of locations. Similarly, the presented rule successfully extracted the disaster location and type from the news, "*A suicide car bomber blew up a small clinic in eastern Afghanistan on 25th December,2010*". The system recognize Afghanistan as a disaster location and Human_Inflicted_Disaster as a disaster type. We are not presenting the detail of Human_Inflicted_Disaster rule for the sake of brevity.

```
(({Token.orth=="upperInitial",
Token.kind == "word"})?

({Token.orth=="upperInitial",
Token.kind == "word"} |
{Lookup.majorType == location})?
({Token.string == ","})?

({Lookup.majorType == location}) )
```

Rule 2 Extracting disaster location

Rule 3 presents another rule for the extraction of location. On the left hand side of rule, various patterns for location of disaster are matched with text. We have exploited the list of locations provided by GATE containing all the major countries, areas and cities of the world. The line "{Lookup.majorType == location}" calls the reference to all these listed locations to match with the appearing text. We have defined macros to make effective use of codes that are repeatedly used in rules.

4.2. Disaster Magnitude

The disasters like tsunami and earthquakes have a magnitude property that can be used to measure severity of disaster. Six rules are developed to

Journal of Theoretical and Applied Information Technology
30th June 2012. Vol. 40 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

extract the magnitude information (only two rules are given here).

```
Rule: TempDisasterLocation
 ((ANY_DISASTER)
     (({Token.string == "in"})?
     ({Token.kind == word})*))

(({Token.orth=="upperInitial",
     Token.kind == "word"})?

({Token.orth=="upperInitial",
Token.kind == "word"})?
|({Lookup.majorType == location})?

({Token.string == ","})?

({Lookup.majorType ==
location})):loc -->

:loc.Dis_Location = {kind = "word",
     rule = "TempDisasterLocation"}
```

Rule 3 Exploiting GATE list for extracting disaster location

```
Rule: Magnitude1
(({Token.string ==
"magnitude"}|{Token.string ==
"Magnitude"}|{Token.string == "M"})

 ({Token.kind == word})*
)

(
 (MAGNITUDE_NUMBER)):Mag -->

:Mag.Magnitude = {kind = Number, rule
= "Magnitude1"}
```

Rule 4 Extracting magnitude of a disaster

```
Rule: Magnitude2
 ((MAGNITUDE_NUMBER))
: Mag (({Token})
(MAG_FORMAT)) -->

: Mag.Magnitude = {kind = Number,
rule = "Magnitude2"}
```

Rule 5 Using macros to extract magnitude of a disaster

The first rule given in Rule 4, looks for the word "magnitude" in a sentence and control is transferred to the macro that determines whether the number is a magnitude of disaster or not. The second rule given in Rule 5 uses another macro MAG_FORMAT given in Rules 6 (a) and (b) that matches a particular sentence structure that can appear in text describing the magnitude of disaster. In Rule 7, temporary annotations of magnitude are removed and new "Disaster_Magnitude" annotation is created.

This set of rules successfully extracts the magnitude (7.2) of earthquake from the text, "*An earthquake of magnitude 7.2 has hit south-western Pakistan and at least 500 people died*".

```
Macro: MAGNITUDE_NUMBER
(
({Token.kind == number}
  {Token.string == "."}
  {Token.kind == number})

 ({Token.string == "and"}|
{Token.string == "to"}|
{Token.string == "-"}|
{Token.string == "or"})?

({Token.kind == number}
  {Token.string == "."}
      {Token.kind == number})?
)
```

Rule 6 (a) Macro used in Magnitude rules

```
Macro: MAGNITUDE_FORMAT
(
({Token.kind == punctuation})?
({Token.string == "Mw"}|
{Token.string == "Me"}|
{Token.string == "Ms"}|
{Token.string == "Mb"}|
{Token.string == "ML"}|
{Token.string == "mbLg"})

({Token.kind == punctuation})?
)
```

Rule 6 (b) Macro used in Magnitude rules

4.3. Disaster Epicenter and Disaster Date

Rule 8 and 9 are used to extract epicenter of disaster, if any. The rule is similar to Magnitude rule with the addition of rule Direction. The rule Direction extracts the direction of epicente. For an instance, the rule successfully extract the epicenter of earthquake from the text, "*The earthquake epicenter was at 34.402 degrees North*". We also used the rule location to identify the location of epicenter.

Journal of Theoretical and Applied Information Technology
30th June 2012. Vol. 40 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
Rule: MagnitudeFinal
(
{Magnitude}
)
:Mag -->
 {
//removes Magnitude1 annotation,
gets the rule feature and adds a new
Magnitude annotation

gate.AnnotationSet Mag =
(gate.AnnotationSet)bindings.get("Ma
g");

gate.Annotation MagAnn =
(gate.Annotation)Mag.iterator().next
();

gate.FeatureMap features =
Factory.newFeatureMap();

features.put("rule1",
MagAnn.getFeatures().get("rule"));

features.put("rule2",
"MagnitudeFinal");

annotations.add(Mag.firstNode(),
Mag.lastNode(),
"Disaster_Magnitude",
features);

annotations.removeAll(Mag);
}
```

Rule 7 Converting temporary annotations of magnitude into Disaster_Magnitude

```
Rule: EpicenterRule1

(
({Token.string ==
"epicenter"}|{Token.string ==
"epicenter"})
({Token.kind == word})*

)
 (DEGREES1)
:epicenter
-->
:epicenter.Epicenter = {rule =
"EpicenterRule1"}
```

Rule 8 First rule for extracting epicenter of a disaster

The analysis of the news indicates that some times, the epicenter is not represented as a number, rather as a location. We used GATE built-in macro Date to extract the disaster date.

```
Rule: EpicenterRule2

(
({Token.string ==
"epicenter"}|{Token.string ==
"epicenter"})
({Token.kind == word})*
)
(
{Lookup.majorType == location}
)
:epicenter
-->
:epicenter.Epicenter = {rule =
"EpicenterRule2"}
```

Rule 9 Second rule for extracting epicenter of a disaster

```
Rule: DeathBeforeNumber
Priority: 100
(
{Lookup.majorType == deadpeople}
({Token.kind == word})*
)
(AMOUNT_NUMBER)
(
{Lookup.majorType == people})?
)
:TotalDeaths
-->
:TotalDeaths.No-of-Deaths = {kind =
Number, rule = "DeathBeforeNumber"}
```

Rule 10 Extracting number of deaths

4.4. Number of dead, injured, lost, affected, and homeless people

Rule 10 and 11 present the excerpt of the rules that extract the number of dead and injured people respectively. The rules use the developed gazetteer lists from the table 1. In rul2 10, the first list provides synonyms for the word dead whereas the second list provides synonyms for the word people. The system successfully extracted the number of dead peoples (500) from the text, "*An earthquake of magnitude 7.2 has hit south-western Pakistan and at least 500 people died*". Similarly, the number of injured, lost and homeless people can be extracted using the respective gazetteer lists.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
Rule: InjuryAfterNumber
(({Lookup.majorType ==
time_modifier})?
(AMOUNT_NUMBER)
 ({Lookup.majorType == peoples})? )
:TotalInjury
(({Token.kind == word})*
{Lookup.majorType == injuredpeople}
)
-->
:TotalInjury.No-of-Injuries = {kind
= Number, rule =
"InjuryAfterNumber"}
```

Rule 11 Extracting number of injured people

4.5. Miscellaneous rules

We developed some other rules such as Percent and Dozens_Million_Billion to extract numbers from the text. For an instance, the rule Percent extracts numbers from the text like, "*Fifty percent of the population is affected from the earthquak*e". Similarly, the second rule extracts number from the text such as "*five hundred people died* ". We also developed rules to extract infrastructure losses, caused by the earthquake.

5. RESULTS AND CONCLUSIONS

There is no gold standard available for disaster domain. We developed our own corpora from 50 different sources that include online newspapers, NGOs web sites, discussion forums and blogs etc. The corpora include news about six different disasters, occurred at different times. First, we manually extracted information from corpora and then we provided the same corpora to system to extract information. A comparison was made between manually extracted and system extracted information to measure accuracy of the extracted information. GATE evaluation tools AnnotationDiff and Benchmark were used to evaluate the data. Figure 1 (a-d) provides statistical measures precision, recall and f-measure of various news corpora for earthquakes, hurricanes, suicide attack and floods. The empty columns in the figure indicates either missing information for a particular type of disaster (e.g., missing people in case of suicide attack) or it did not exist in the corpus. The results indicate that accuracy of IE for earthquake was higher than that for floods. The system successfully extracted the magnitude of earthquake with a precision and recall of almost 1.0. The system also extracted disaster location and death caused by a disaster, with a significant accuracy, however the extraction of date and time showed poor results. We learned from experience that the news from informal sources such as discussion forums result in lesser accuracy as compared to formal sources such as newspaper. There were some sources (web pages) from which system was unable to extract any significant information-mainly because of poor English structure.

REFRENCES:

[1] M. Afzal, J. Ajoon, I. Ahmed, and Q. M. Ilyas, "Conceptualization and Extraction of Disaster Related Information from Unstructured World" In International Conference on Artificial Intelligence and Pattern Recognition, AIPR-09, Orlando, Florida,USA, July 13-16, 2009.

[2] M. Albanese and V. Subrahmanian, "T-rex a Domain-Independent System for Automated Cultural Information Extraction", In First International Conference on Computational Cultural Dynamics, (ICCCD-08), Maryland, U.S.A, 2008.

[3] M. Banko, M. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open Information Extraction from the Web", In International Joint Conference on Artificial Intelligence, IJCAI 07, 2007.

[4] M. Bradley, "The Future of the Australian Disaster Information Network (AusDIN)", Research memorandum, Emergency Management, Australia, 2003.

[5] M. J. Cafarella, D. Downey, S. Soderl, and O. Etzioni, "Knowitnow: Fast, Scalable Information Extraction from the Web", In Proceedings of the Human Language Technology Conference (HLT-EMNLP-05), 2005.

[6] M. Careem, C. De Silva, R. De Silva, L. Raschid, and S. Weerawarana, "Sahana: Overview of a Disaster Management System. 2006 International Conference on Information and Automation, 2006.

[7] H. Cunningham, "Information Extraction, Automatic. Encyclopedia of Language and Linguistics", Second Edition, Elsevier, 2005.

[8] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications", In Proceedings of the 40th Anniversary Meeting

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of the Association for Computational Linguistics, 2002.

[9] H. Cunningham, D. Maynard, and V. Tablan, "JAPE: a Java Annotation Patterns Engine", Second Edition, Research Memorandum CS-00-10, Department of Computer Science, University of Sheffield, November 2000.

[10] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates, "Unsupervised Named-Entity Extraction from the Web: An Experimental Study", Artificial Intelligence, 165(1):91134, 2005.

[11] Q. M. Ilyas and I. Ahmed, "A Conceptual Architecture of Sahara - A Semantic Disaster Management System", World Applied Sciences Journal, 10(8), 2010.

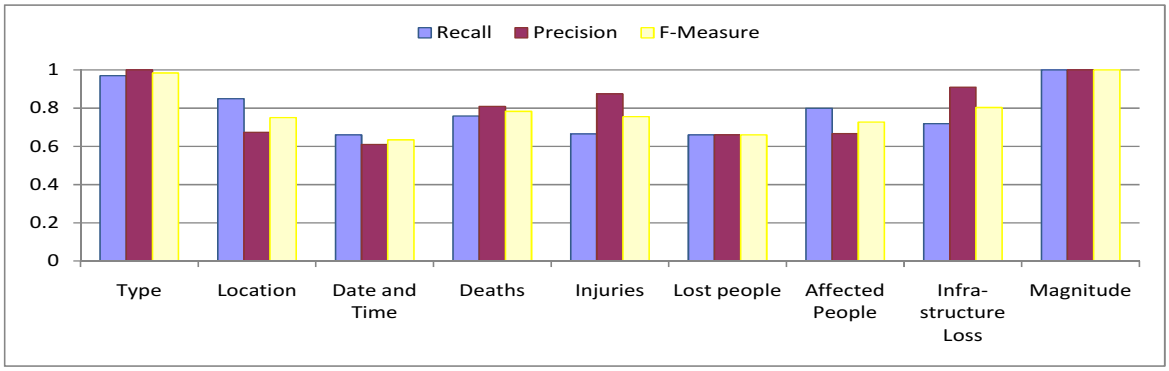
[12] E. Quarantelli, "The Proposed Establishment of A Global Disaster Information Network (GDIN)", In Conference of Hazards and Sustainability: Contemporary Issues in Risk Management, Durham, United Kingdom, 1998.

[13] A. Tllez-Valero, M. M. y Gmez, and L. Villaseor-Pineda, "A Machine Learning Approach to Information Extraction", In International Conference on Intelligent Text Processing and Computational Linguistics, CICLing-2005, Mexico City, USA, 2005.

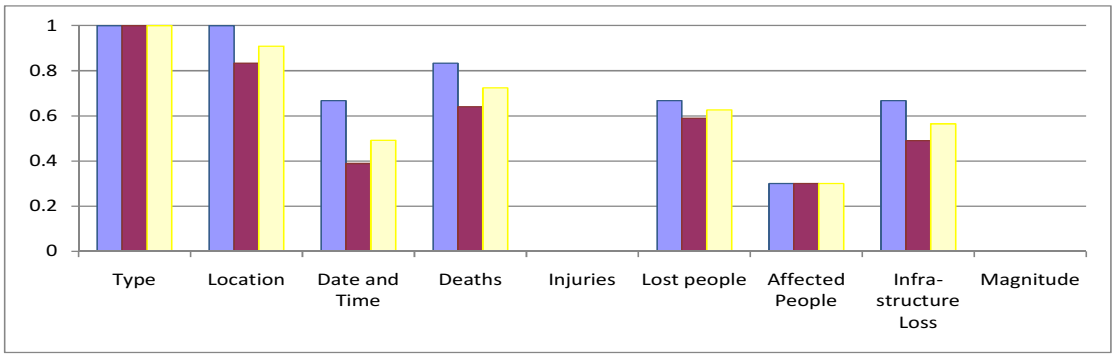
Journal of Theoretical and Applied Information Technology
30th June 2012. Vol. 40 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

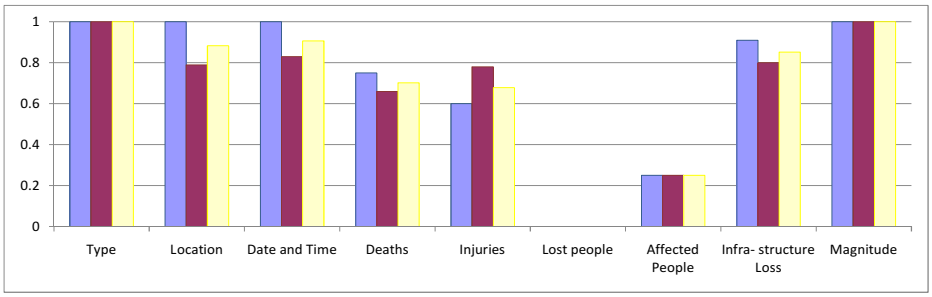
(a)



(b)



(c)



(d)

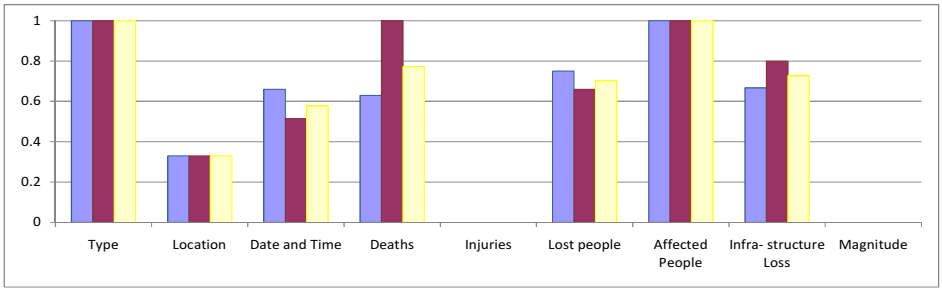


Figure 1 (a – d) Results of IE on corpora related to earthquake, hurricane, suicide attack and flood respectively