Expert Systems for Natural Disaster Event Extraction  
CSCI 6962: Natural Language Processing  
Final Project Report  
by Hongtan Sun

Introduction

Natural disasters are the natural processes of the Earth that results in serious disruptions of the functioning of the society. They can cause widespread human, material, economic or environmental loses and are one of the biggest threat to people’s life and property. For example, in the year of 2012, there were 905 natural disasters worldwide. The overall costs were US$170 billion and that is only a moderate year on a global scale [1]. As a result, using modern technology to help disaster relief and post-disaster restoration has always been on of the main topic in academia.

In this project, we propose a framework of domain-specific natural language processing (NLP) system that extracts events and information about natural disaster from online news articles. The motivation of this topic is from my PhD dissertation project. My PhD dissertation project is on Integrated Network Design and Scheduling problems arising from supply chain and infrastructure restoration after extreme events. A key source of references and informations of natural disasters and their impacts are news articles (e.g. [2]). Sometimes, hundreds of articles are searched and read manually in order to get valuable informations [3]. So it would significantly improve the efficiency of our research work and would extend our searching horizon if there is some NLP machinery that could help pre-screen the online news articles and extract useful informations automatically.

Domain-specific NLP information extraction systems is one of the main research interests around 1990’s, and become popular again recently in the domain of biomolecular event extraction from scientific documents [5]. One of the well developed field is knowledge-based NLP system that rely on a domain-specific dictionary of concepts. Some of them (e.g. AutoSlog) could achieve 98% of performance of a hand-crafted dictionary [4]. There are domain-specific NLP systems developed for terrorism domain, joint ventures domain and microelectronics domain, because of the Message Understanding Conference (MUC) [6, 7] from 1987 to 1997 that is initiated and financed by Defense Advanced Research Project Agency (DARPA).

Recently the biomolecular event extraction from scientific documents gets lots of interest because of the development of machine learning (ML) techniques. The type of the systems has been moving from purely rule-based and dictionary-based towards ML-based. The event extraction in biomolecular domain is a challenging task due to the ambiguity and variability of scientific documents, and the complexity of the biological process. And the state-of-art methods (BioNLP) uses a multi-stage approach which combines basic pattern matching, parsing techniques and sophisticated machine learning methods. In particular, most event extraction strategies split the into two steps: first identify trigger words that indicate a event and second fully specifies the event by filling the corresponding arguments. And some of the
Figure 1: Framework flow chart

Text article collection from internet

Extract event elements

Information normalization and visualization

state-of-art systems can achieve a precision ranging from 55% to 73%, for a recall varying between 19% and 22% [5].

Another popular domain-specific event extraction system is on the terrorist activities with many of the system developed based on the data set from MUC-4 [8]. Previous systems developed in the 1990’s are mainly knowledge-based or keyphrase-based [9]. While recently machine learning methods are developed to automatically learn the relevant features that requires limited prior knowledge. In 2012, the state-of-art systems can achieve a 59% F-score on the MUC-4 data set, and later model based on word embeddings can achieve 73% F-score.

The Automatic Content Extraction (ACE) program leads the open-domain event extraction tasks, which provides the identification of 8 event types with 33 subtypes [10]. These event types focus on social activities including people’s life, business, conflicts and justice and does not include any natural events.

In conclusion, to the best of my knowledge, the how to develop a domain-specific event extraction for natural disasters is still an open problem. However, the recent developments for natural language processing, machine learning and information technology have built up the essential platform that people can explore a good NLP system for the natural disaster events. In this project, we propose a framework for domain specific event extraction system for natural disaster events. I then use Hurricane Sandy as a case study to explain the architecture of the framework. One of the key block of the system will be tested using the open-domain event extraction system developed based on the ACE event types. As a result, a system development along the lines of the ACE systems will be proposed.

**Framework outline**

In this section, I describe the architecture of the domain specific natural language event extraction system framework for natural disaster event. The objective of the event extraction system is to automatically obtain natural disaster related information from the online news
articles, extract key events and output a user-friendly representation of the events. To achieve this goal, the framework will consist of three key blocks: data collection and pre-processing, event extraction and result presentation, as it is shown in Figure 1.

The first block is data collection and pre-processing. As the objective of the system is to automatically extract information from online news articles, the first thing it does is to search news articles from website using a web crawler. The data downloaded from internet or news agent website are usually HTML or HTM files, so the next step is to clean up the format and transform the main text of the article into a text document or a XML document that can be accepted by a NLP program. Figure 2 shows the flow chart of data collection step, with screen shots of the expected output or target in each step. The third step is to check if the text document contains information that is related to natural disasters. A binary classifier between natural disaster related article and unrelated article is suffices to achieve this task. The method could be either a trigger word classifier or a term-based machine-learning classifier. In case of trigger word classifier, the classifier might need hand-crafted data as training data. While for term-based machine-learning classifier, it could be supervised learning model trained with pre-classified articles or unsupervised learning model based on term similarities.

The second block is to extract event elements for the events in the news articles that are classified as natural disaster related. In particular, the time, location and event type are three key elements that need to identify. In some of the events, for example, the events about disaster relief, the participants is also one of the element that could be extracted. The level or severity of the damage is an element that are relatively hard to extract. So I will focus on the three key elements first. Another reason to work with time, location and event type is because it is compatible with current event extraction systems. Event element extraction is the core part of the system and it will also be the bottle neck of the system, as we might see in the introduction. It determines the accuracy of the system and are usually taking the longest time to process. We will use a NLP system developed under the ACE framework in
the next section to illustrate how it works.

The third block is to normalize the information and present the extracted event in an user-friendly way. The first step is information normalization, which contains two layers: identification and wikification. Figure 3 shows two examples from the news article *Flooded Tunnels May Keep Citys Subway Network Closed for Several Days* on the December 31, 2012 New York Times [2]. The first one is the time of the event. Using the event extraction model, it will extract the time entity ‘Tuesday’ from the first paragraph. Then referencing the date of the article ‘OCT, 30, 2012’, the time of the event should be identified as ‘Tuesday, 10/31/2012’ and can be linked to a calendar, which is a desired normalized information. By the ‘wikification’ layer, it could only not be achieved for some the event element entities that has a wiki page. For instance, in the second example shown in Figure 3, the first paragraph contains an facility entity ‘subway network’. Referencing the second paragraph shown in the figure, it could be identified that the ‘subway network’ refers to ‘the New York City subway system’ which has a wiki link https://en.wikipedia.org/wiki/New_York_City_Subway. In that case, this piece of information can be normalized as ‘the New York City subway system’ and linked to the wiki link and even the Google map.

Information visualization and presentation focuses on output the extracted event information in a user-friendly format, which in fact depends on the user for the system. In case of my PhD research study, a table listing all the events and multidimensional matrices representing the infrastructure networks, as it is needed for the integer programming set-up in [11]. For general users, a modified map demo with markers and informations of each events happened on each location could be a useful output. Similar output has been implemented for crime record website, for example the crimemapping website [12]. Figure 4 shows an example of information visualization. The sentence ‘The South Ferry station at the southern tip of Manhattan was filled track to ceiling with water, the authority said.’ from the previously mentioned New York Times news article [2] contains facility failure event happened to the South Ferry subway station in New York City on Tuesday, 10/30/2012. It can be visualized on a map with a mark at the South Ferry subway station and a table pop-up containing the event elements extracted from the text.

**Event element extraction under ACE framework**

In this section, we test the event element extraction block using an NLP system developed
The ACE framework defined 8 types of events with 33 subtypes, but most of them concerned about the social activities and none of them corresponds to the events happened around natural disasters. On the other hand, the system has a solid and flexible theoretical structure that has the potential of modified to work for natural disaster event extraction. For example, some of the events happened around a natural disaster has a similar structure as the event type ‘conflict’ or ‘movement’ in ACE. ACE also has blocks that can annotate geographical entities including locations and facilities, which fits well into our natural disaster event extraction framework.

We used a NLP system developed under the ACE framework and trained using newswire data to annotate the article *Flooded Tunnels May Keep City’s Subway Network Closed for Several Days* on the December 31, 2012 New York Times [2] and we found some useful information is extracted.

For example, Figure 5 shows a name entity ‘flooded tunnels’ is identified as a generic person from the following sentence.

*What they found was an unprecedented assault: flooded tunnels, battered stations and switches and signals likely damaged.*

As in the natural disaster event extraction system, the name entity ‘flooded tunnels’ is supposed to annotated as a non-operational facility. The annotation is triggered by the word ‘tunnels’, which could also be the trigger word when identifying the broken facilities and their locations in natural disaster events.

Another interesting name entity annotation I got is the ‘the New York City subway system’ from the following paragraph, which is one of the annotations that are very close to what we wanted in a natural disaster event extraction system.
"The New York City subway system is 108 years old," Joseph J. Lhota, the chairman of the Metropolitan Transportation Authority, said. "It has never faced a disaster as devastating as what we experienced last night."

It has been recognized as a generic geographic name entity due to the trigger word 'New York City'. There are actually some ambiguity in this name entity, as well as for many other geographic name entities appear in a natural disaster related news article. It could refer to both the location and the agent of the event 'face a disaster' that is mentioned later in this paragraph. The New York City subway system are in fact mentioned several times in the article, and each of them could be a different arguments. And our system need to distinguish the roles of the name entities in order to improve the accuracy of event extraction.

Because of the system we are using is not developed for natural disaster events, there are only a few events identified from the article. On of the reason is that, the event extraction takes a multi-step approach: first identify a trigger word, and then fill in the arguments. The list of trigger words are build-in to the system when it is being trained, and determines the types of events it could recognize. For each type of events, there is a list of trigger words for that. If the system is not developed and trained for natural disaster events, it would not contain the essential trigger words, and would not trigger an event extraction. Meanwhile, the syntactic and dependency structures of the sentences are very similar. So the ACE system with many models implementing it is expected to be modified for the natural disaster event extractions.

Figure 6 shows an extracted event which is identified as 'Movement/Transport' in the ACE categories. The original sentence is

As the remnants of Hurricane Sandy left the city on Tuesday, transit officials surveyed the damage to the system, which they shut down on Sunday night as a precaution.
And the event identified is 'Hurricane Sandy left the city’. The trigger word is 'left' and the hurricane 'Hurricane Sandy' is recognized as a person. In the natural disaster event extraction system, we would like to identify the landmarks of the natural process, in particular, the beginning and end of the hurricane, earthquake, flood, etc. As for this sentence, the event is correctly identified, but we want it to be classified as the 'end of natural process’. This is related to choose the correct meaning of the ambiguous trigger word 'left' and provide a good interpretation.

Modify ACE-based systems for event element extraction

The examples in the previous section shows that the ACE-based systems could be modified and become the event element extraction block in the natural disaster event extraction framework. In particular, new event types need to be defined and trigger words dictionary need to be set up for natural disaster events. Table 1 shows the event types and subtypes needed for our system and table 2 shows some of the trigger words.

Conclusion

In this project, I reviewed some previous works on domain-specific and open-domain natural language processing systems. It turns out that the field of natural disaster event extraction expert system has not been well-studied. So I proposed a framework for the expert system with three blocks including data collection and preprocessing, event element extraction, and information normalization and visualization. A ACE-based system is used to analyse data from the 2012 Hurricane Sandy. And it turns out that the event element extraction block can be developed based on the extension of the ACE framework.

The next step of this project is to develop the programs that implement the framework of natural disaster event extraction system. As I suggested in this report, one of the approach is to develop the event element extraction block from the ACE framework. The information normalization and visualization block is also an interesting topic to explore, which is related to coreference and wikification in event extraction.
### Table 1: Event types for natural disaster events

<table>
<thead>
<tr>
<th>Event type</th>
<th>Subtypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Disaster</td>
<td>Forecast</td>
</tr>
<tr>
<td></td>
<td>Start - natural process</td>
</tr>
<tr>
<td></td>
<td>End - natural process</td>
</tr>
<tr>
<td></td>
<td>Damage - facility</td>
</tr>
<tr>
<td></td>
<td>Damage - personnel</td>
</tr>
</tbody>
</table>

### Table 2: Some trigger words for natural disaster events

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Ambiguous words</th>
</tr>
</thead>
<tbody>
<tr>
<td>hurricane</td>
<td>restore</td>
<td>lose</td>
</tr>
<tr>
<td>restoration</td>
<td>flood</td>
<td>suspension</td>
</tr>
<tr>
<td>outrage</td>
<td>damage</td>
<td>return</td>
</tr>
<tr>
<td>debris</td>
<td>sweep across</td>
<td>leave</td>
</tr>
<tr>
<td>storm</td>
<td>cancel</td>
<td>shut down</td>
</tr>
<tr>
<td>corrosion</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### References


