The Title

By

Alice Irankunda (alice.irankunda@aims.ac.rw)
Supervised by Dr.Yabebal FANTAYE

June 2017

AIMS | African Institute for Mathematical Sciences RWANDA

Contents

1. Introduction of the Research

In every organization there is a way to communicate. One of the most popular way to transmit the information is to produce a written report which explains how different activities of the organization are going. For large organizations there are huge number of reports and it is so challenging to go through each and every report manually. This research has an aim of providing an easy way of visualizing and extracting the important information locked in reports from NGO and large organisations.

In 1919, the International Federation of Red Cross and Red Crescent societies (IFRC) has been founded, it has some millions of reports related to humanitarian support, How to know automatically the number of people who suffered from a disease? How to know the fraction of fund spent on shelter?

In this research, we tried to use a combination of statistics formulae and techniques of Natural Language Processing (NLP) to find the solution for the extracting entities, Big data and Machine learning for analysing the huge data by using statistical and computing algorithms. Entity can be defined as an instance of existence of something, for example what is the activity done on what place when and how ?

Document modelling by extracting entities is one of the way to deal with natural big data linguistic problems where entity can be considered as a single unit of data like location, people, organization and so on. Entities can be classified based on their relationship.

These are key procedures to be performed for extracting entities:

- The sentences which compose a report must be parsed.

- Entities must be identified in the report and classified.

- Relationship between entities must be modelled.

A report is composed by paragraphs, each paragraph is made by sentences. Natural Language Processing techniques deal with sentences and content based analysis by splitting the sentences into tokens then remove the common words and work with corpus to get entities. The meaning of a word can depend on its surroundings as well as it can be independent. For extracting significant entities, the context of a word is one of the points to be considered carefully.

1

2. Literature review

In today's life, many organizations are generating unstructured data while they are communicating. There are plenty of entities to be extracted. In this research, all reports we considered are written in English.

To label the boundaries of sentences is one of the important prerequisite steps in Natural Language Processing. The punctuation marks cause some ambiguity (**?**) for example it is challenging to differentiate the point in abbreviations and a full stop. To handle this ambiguity some systems use the special purpose-regular expression grammar, exception rule method etc.

David Palmer and Marti A. Hearst worked on the problem of punctuation marks. (**?**). They developed an efficient system with high accuracy in automatic labelling the boundaries of the sentence by using the feed forwarding neural-networks where the input was the POS probabilities of all tokens which are surrounding the punctuation and output was found as the label to be assigned to the token. This work was able to correct up to 98.5% for punctuation of sentence-boundaries. A proposed new approach was how to represent the context of punctuation marks without ambiguities.

This research will also look at how neural networks can be used to label different tokens.

Capitalization can be used in different ways such as the beginning of the proper noun, the abbreviation, the post of high level profile people etc. Considering the English language text, if we are given a particular token it is not by chance to determine whether it is a name or not. Some of the approaches to indicate a name are to use capitalization, detection of sentence boundaries and dictionaries (**?**).

2.1 Parse Tree

One of the sentences that compose our sample report says: "Assessment reports indicated 117 deaths, 544 people injured, 12,794 homes damaged and 7,384 houses destroyed", Suppose that this sentence is called "S"

There are two mains steps which can be performed to get the entities from this sentence:

- **Tokenizing**: This is a procedure of taking a sentence and extract the composing atomic linguistic elements e.i. words, verbs, punctuations, adjectives etc . S has the following tokens: ['Assessment', 'reports', 'indicated', '117', 'deaths', ',', '544', 'people', 'injured', ',', '12,794', 'homes', 'damaged', 'and', '7,384', 'houses', 'destroyed']

- **POS**: part-of-speech is a process of attaching to every linguistic element of the sentence a corresponding tagg based on grammar rules. The POS of S are: [('Assessment', 'JJ'), ('reports', 'NNS'), ('indicated', 'VBD'), ('117', 'CD'), ('deaths', 'NNS'), (',', ','), ('544', 'CD'), ('people', 'NNS'), ('injured', 'VBN'), (',', ','), ('12,794', 'CD'), ('homes', 'NNS'), ('damaged', 'VBN'), ('and', 'CC'), ('7,384', 'CD'), ('houses', 'NNS'), ('destroyed', 'VBD')]

71   The meanings of the used tags for S:

72   - JJ: **Adjective**: 'Assessment'

73   - NNS: **Noun, plural**: 'reports', 'deaths', 'people','houses'

74   - VBD: **Verbs, past tense**: 'indicated', 'injured', 'damaged', 'destroyed'

75   - CD: **Cardinal Number**: '117', '544' '12,794','7,384',

76   - CC: **Coordinate Conjugation**: 'and'

77 The parse tree is formed based on the POS, the classification of word and the way words are
78 arranged in a sentence show a kind of relationship between words.

Figure 2.1: Relations extraction from
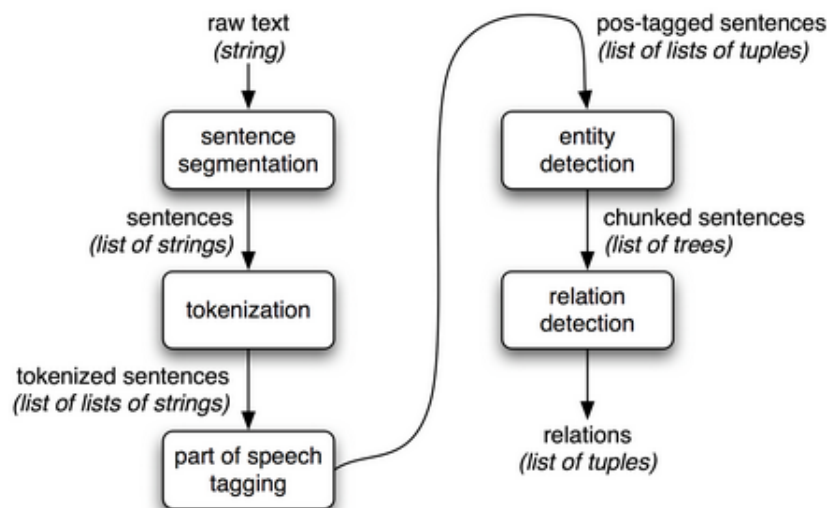
79 # 2.2   Named Entity Recognation and Classification NERC

80 The term "Named entity" has been coined in 1996 in "sixth Message understanding Conference"(MUC-
81 6 R. Grishman and Sundheim 1996). Entity can be referred as a task, the entity is "named"
82 when it is restricted to one or many rigid designators (**?**), example: persons, location, product
83 are the named entities.

84 Based on the classification of Standard Generalizes Markup Language (SGML) a task can be
85 divided into three subtasks:

86   1. ENAMEX: location,product,country ,organization

87   2. NUMEX : percentage,quantity

88   3. TIMEX : time, date

89   The entities from different reports. For extracting entities in a report there are different models
90   which can be used:

2.3   Hidden Markov Model

This model is based on Bayesian probability inference which has been initiated in 18th century.
HMM is the earliest applied model for Natural Entities Recognition for English language.The way
to perform these tasks is to find the most likely sequence of tagged names(TN) given a sequence
of words(SW).

$$P(TN|SW) = \frac{P(SW|TN)P(TN)}{P(SW)} \tag{2.3.1}$$

The equation (**??**) is conditional probability, $P(TN|SW)$ can be called posterior and it is the prob-
ability of an event Sequence of word occurring given Tagged names has observed. $P(SW|TN)$
is also called likelihood e.i. it is the probability of observing the sequence of words(SW) when
the given hypothesis tagged name (TN) is true. On another hand P(TN) doesn't depend on the
evidences, P(TN) is called prior e.i. that it is true even if there is no given evidence at all(masters
thesis). We can be ignored P(SW) and the remaining objective is to maximise the probability of
getting the sequence of tagged names when sequence of words is given.

$$Max\left[P(TN|SW)\right] \tag{2.3.2}$$

From the equation (**??**) of the maximization, the following estimation can be made.

$$P(TN) \approx \prod_{i=1}^{n} P(TN_i|TN_{i-1}) \tag{2.3.3}$$

Where TN_i is a tag in the sequence of names (TN), for the likelihood probability can be estimated
as :

$$P(SW|TN) \approx \prod_{i=1}^{n} P(SW_i|TN_i) \tag{2.3.4}$$

The above estimations was for a small sequence where TN_i is a tag in the sequence of names
(TN) and SW_i is a tag at index i in a sequence words (SW). For the large training corpus, the
needed step is estimate based on the number of times the tag occurs and the position of the tag
in a given corpus.

$$P(T_i|T_{i-1}) = \frac{K(T_{i-1}, T_i)}{K(T_{i-1})} \tag{2.3.5}$$

92 Based on the training corpus, $K(T_{i-1}, T_i)$ is referred as a how many times the tag T_i occurs after
93 the tag T_{i-1}. In the corpus, $K(T_{i-1})$ is considered as the number of occurrences for the tag T_{i-1}.

Therefore the estimation can be performed as follow:

$$P(C_i|T_i) = \frac{K(T_i, C_i)}{K(T_i)} \tag{2.3.6}$$

94 From the equation (**??**), the term $K(T_i, C_i)$ is referred as the sum of the times that a word "C_i"
95 has a tag T_i in the training corpus. The process of computing the posterior using the above steps
96 is called Markov model.

97 It is one of the most powerful statistical and machine learning (ML) techniques in modelling and
98 high qualified in entities extraction. When the researcher is willing to train new data, HMM is
99 very robust and efficient in computations. One of the limitations of HMM is that the researcher
100 must have the notion of model topology and statistical techniques on how to deal with large
101 amount of training data.

102 2.4   Supporting Vector Machine based model(To be edited
103   for non linear data)

This model has an aims of classifying the named entities by using the linear support vector
machine which separate input train documents into two categories, a document must be cate-
gorized as either positive or negative and be represented in two dimensional graph. Hyperplane
is for separating train documents based on their categories and "w" is a weight vector which is
perpendicular to hyperplane is represented by the following equation:

$$w.x - b = 0 \tag{2.4.1}$$

104

105 From the (**??**), the offset of the hyperplane is $\frac{b}{\|w\|}$

The target is to maximize the the margin between the the points which represent two categories.
Remember that the vectors which pass through each of the point representative is perpendicular
to the w, suppose that there will be an imaginary line which join two borders points h_- and h_+.
Supporting vectors which are demonstrated by the dashed lines on the figure above are formed
by :

$$w.x - b = 1 \qquad \text{and also} \tag{2.4.2}$$
$$w.x - b = -1 \tag{2.4.3}$$

106 There are many algorithms with different approaches to optimization problems but all tends to
107 the same solution says that minimize $njwn$ automatically maximize the margin between h_- and
108 h_+ where the boundary is a half way.

Figure 2.2: SVM in hyperplane representation
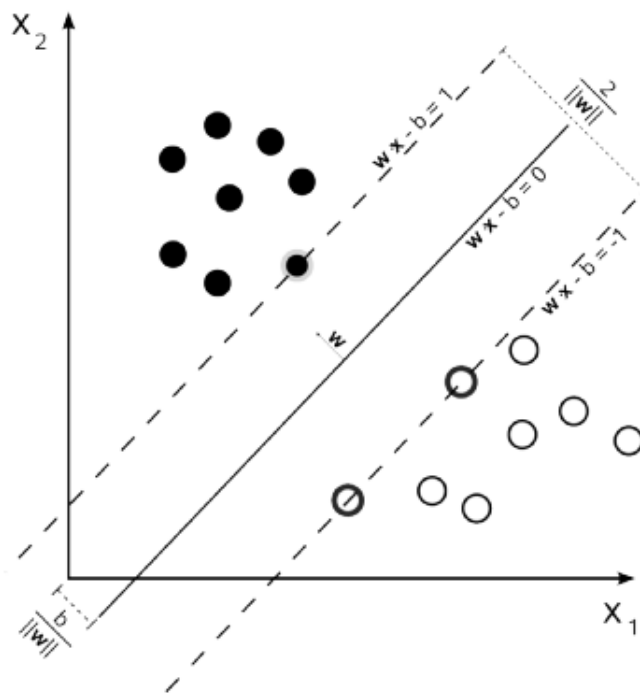
Now,add another constraint for each document category from the equations (**??**) and (**??**), in order to hit the target

$$w.x - b \geq 1 \qquad \text{and also} \qquad (2.4.4)$$
$$w.x - b \leq -1 \qquad (2.4.5)$$

Disadvantages of SVM

The classification of particular documents is not easy to be performed by SVM without destroying the constructed weights but with hand-written rule model. Machine learning uses decision tree procedure than SVM. In addition the decision tree has a detailed boolean-like model which is more popular to user.

2.5   Some Terminologies

Hand-written rule

It is one of the standard approaches of NER and IE, it has been used for extracting the patterns from automated pages such as amazon, NLP is so useful for unstructured humman-written text by delivering part-of-speech (POS), syntactic parsing and categories of semantic words.

Rule /pattern based extraction

Many IE systems uses rule/pattern to extract words and also phrases by looking to the context of those words or based on the their surroundings.(**?**). Some system decided if the procedure of extracting the words should rely on the meaning of each word independently or on the context of their surroundings in a phrase. The limitation of this method is that some words do not have a closer mining to their surroundings that is why Patwardhan Siddharth with help of Ellen Rilo in workshop called "ACL 2006" presented another approach which was generating an automated IE system to learn patterns from a large fixed data set within a specific domain (**?**)

Bag of Words

It is referred to the multi-set of words represented by Natural Language Processing and Information Retrieval(IR). They are used for classifying the documents.

Our research deals with reports generated through a template, compared to the work of (**?**) templates usages is a limitation.

2.1.3. Text classification and Naive Bayes

It is one of the most important algorithm in text classification by using base rule and bag of words to classify the entities (**?**).The user instead of going through the report and start posing many queries, text classification algorithm transient the need information. Its aims is to build a function θ which takes the bag of words and returns the class of sentiment C either positive or negative.

138 θ

139 $\Updownarrow$

> ARCS initiated its response immediately after the earthquake struck to address the immediate needs. The National Society (NS) regional branches were at the forefront of the response and worked with Disaster Response Units (DRU). ARCS staff and volunteers were deployed promptly to support rescue efforts, provide first aid to the injured and distribute immediate relief supplies to affected people alongside undertaking initial assessments. A total of 900 volunteers were mobilised to support this response operation. ARCS also supported to transport critically injured people to hospital and mobilized community members for voluntary non-remunerated blood donations.

140 $\Updownarrow$

141   C

142 The procedure is to look for all words and retrieve those which form the subsets. Bag of words
143 are formed after throwing away all words except the subsets.The use of the function θ is for
144 attributing to each item of the bag of words a sentiment.

145 2.6   Machine learning for Named Entities

146 The natural language processing is not enough to handle the sophistication and ubiquity of textual
147 data. Deep learning using machine learning techniques has been introduced to solve this problems.
148 The advantages of machine learning for Named Entities:

149   • Manual extraction of entities is too expensive.

150   • Fast processes of extraction.

151   • extraction done by learning algorithms and Natural language tools.

152   • No limitation of languages.

3. Research Methodology

3.1 Data and tools

World wide non governmental organizations publish some of their reports on their official websites. Web scrapping is one of the ways to extract data from website to the local machines. We downloaded the reports about appeals in pdf format from IFRC website. We used R-scripts for web scraping form our co-supervisor professor Xavier.

We downloaded 1262 reports which have been submitted between 1st January 2015 and 31st December 2016. To differentiate the reports, each report has a report Id but different reports can refer to the same appeal Id. As an international organization which insights on the largest humanitarian activities in the world, IFRC reports we have talk about disasters and cash transfer program. Cash Transfer Program (CTP) describes the money used by IFRC to buy food, shelter, etc.

3.2 Data Analysis and Filtering

Portable document format (PDF) has content which can not be extracted and manipulated easily. The data we have has to be changed into another format in order to pull the information we need. We managed to transform 1260 reports. Our folder has the 1260 txt files which is considered as dataset. For analysing the data, we used python programming language.

Our data reported on different areas of the World such as continents, countries and cities. For example "Europe IB23102015 23Oct2015.txt" covers European continent , "Afghanistan MDRAF003 05Nov2015.txt" and "Japan 0 16Apr2016.txt" reported on specific countries and "Port au Prince country cluster 0 04Oct2016.txt" reported on the most popular city of Haiti. -

3.3 Supervised vs Unsupervised Machine Learning

- **Supervised** is a machine learning part which deals with "labelled data", data are categorized and classified. We have a csv document which summarize the appeals what we have. The shape of this document is 25 columns and 3997 rows. The "CTP" feature indicate if the appeal is classified as a Cash Transfer Program document or not. Among 3997 appeals, 404 are CTP.

- **Unsupervised** can be defined as a way machine learning processes "unlabelled data". the data are unstructured, uncategorised and unclassified. The reports we have are good example of unlabelled data.

    - clustering is a technique for analysing data by identifying hidden groups in a data set.

The hidden groups helps the machine to classify them data into small groups called "cluster" based on similarities or relationship found in data (**?**).

Unsupervised Machine Learning is very important, the analysis of data require the machine to use its brain Supervised learning.

3.4   NLP Corpus

Corpus is a set of large data which are semi-structured. To extract entities from corpus is simple than to deal with unstructured data. To get the corpus we filtered the data by using unicode of utf-8.

To get compatible data, we have to filter using the Unicode provides canonical and compatible equivalence.

Regular Expressions: in Python, regular expression has operations and modules like "re.py" as so on. they are used to manipulate characters in strings. regular expressions use a backslash ("\") to indicate a special form without invoking the meaning of the special form. There are many regular expressions functions but some of what we used the most are :

- re.split(): this function split par pattern and return the list of string.
- re.search(): it returns match objects.
- the match object ".end()": in a search string, it returns the end position of the match.
- the match object: in a search string, it returns the start position of the match.

Python string Strip() method: strip method helped us to remove unwanted characters from the beginning and end of the string. to indicate the position of the character to be stripped we use left(l.strip()) which removes the character at the beginning of a string or right(r.strip()) to remove the character at the end of the string.

ASCII stands for American Standard Code for Information Interchange. it is uses numbers to represent text by using 128 characters. Computer uses ASCII exist within unicode for storing texts easily. All ASCII uses unicode. ASCII characters are used to send and receive the e-mails, for text files and data conversions.

- ASCII-encoder: transform text to numbers.
- ASCII-decoder: transform numbers to text.

Figure 3.1: ASCII TABLE

Figure **??** demonstrates Unicode standard. It provides a unique number to each character composing a text regardless the language, program or platform. UTF stands for Unicode Transformation Format. Unicode characters are set into binary values 0 and 1. UTF-8 for encoding 8 bytes, UTF-16 for encoding 16 bytes and UTF-32 which is a standard for encoding 32-bytes are three current standards. For our corpus we used UTF-8.

After getting semi-structured documents, we removed the StopWords which are defined as unnecessary words for extraction of entities from corpus.

Normally the StopWords return vast amount of unwanted information. Some example of English Stop Words: almost, are, or, details, during, upon and so on.

Now we can check how for all of 1260 documents and count the Stop Words to be removed from vocabularies of corpus. We trained corpus by nltk package called FreqDist which uses frequency distribution of each word occurs in corpus, then the module of nltk technique called "nltk.corpus.PlaintextCorpusReader" helped us to get total 58104 stop words over the whole 7796263 vocabularies.

3.5   Extraction of Entities

To extract entities we used default dictionary built in collection package of nltk. Our dataset now is a folder containing 1260 corpus files, we used nltk chruncker to get sets of sentences of corpus. let have a look for our sample document the way sentences are split.

Figure **??** shows the 45 first lines of the sample document. each each line is ended by $'\backslash n'$.

Figure 3.2: Set of sentences

```
['DREF operation n MDRAF003 Glide n EQ-2015-000147-AFG\n', 'Date of Issue: 26 May 2016 Date of disaster: 26 October
2015\n', 'Operation start date: 3 November 2015 Operation end date: 2 March 2016\n', 'Operation budget: CHF 465,684
Current expenditure: CHF 379,353\n', 'Number of people affected: 65,653\n', '1\n', 'Number of people assisted: 14,0
00 people (2,000 families)\n', 'Host National Society(ies) present (n of volunteers, staff, branches):\n', 'The Afg
han Red Crescent Society (ARCS) has at least 1,800 staff, 25,000 volunteers and 34 provincial branches and\n', 'sev
en regional offices nationwide. A total of 13 branches of ARCS are involved in the earthquake response, with some
\n', '700 volunteers mobilized to support activities\n', 'to the benefit of affected people.\n', 'N of National Soc
ieties involved in the operation:\n', 'The International Federation of Red Cross and Red Crescent Societies (IFRC)
 with the Movement partner actively\n', 'involved in supporting the ARCS response. IFRC and ARCS also maintained go
od coordination with other movement\n', 'partners, the International Committee of the Red Cross (ICRC), partners wi
th present in Afghanistan that include the\n', 'Canadian Red Cross Society, Danish Red Cross, Norwegian Red Cross,
 and Qatar Red Crescent Society. However,\n', 'Red Crescent Society of the Islamic Republic of Iran, Red Cross Soci
ety of China and Turkish Red Crescent Society\n', 'do not have offices in Afghanistan but have supported the earthq
uake response through bilateral arrangements with\n', 'ARCS.\n', 'N of other partner organizations involved in the
 operation:\n', 'Afghanistan National and provincial Disaster Management Authorities, Ministry of Rural Rehabilitat
ion and\n', 'Development (MRRD), UN agencies (WFP, UNICEF, WHO), International Organization for Migration (IO
M),\n', 'International Rescue Committee (IRC), People in Need (PIN), Care International and Oxfam.\n', 'Partners wh
o have contributed to the replenishment of this DREF include Canadian Red Cross Society/\n', 'Canadian Government
 (DFATD), DG ECHO, and Netherland Red Cross/ Netherlands Government (SEF). The\n', 'unspent balance of CHF 86,331 w
ill be returned to the DREF pot.\n', 'A. Situation analysis\n', 'Description of the disaster:\n', 'Around 13:40 loca
l time (UTC +4:30) on 26 October 2015, a magnitude 7.5 earthquake struck Badakhshan province\n', 'in the north-east
region of Afghanistan. Badakhshan, Nangarhar, Baghlan and Kunar provinces were ranked the most\n', 'affected provin
ces. The Afghanistan National Disaster Management Authority (ANDMA) coordinated the initial\n', 'assessments in par
tnership with in-country humanitarian partners.\n', 'Assessment reports indicated 117 deaths, 544 people injured, 1
2,794 homes damaged and 7,384 houses destroyed.\n', 'In Badakhshan province alone, more than 51,000 people were aff
ected. The province also reported to have the most\n', 'extensive damages to properties. Kunar and Nangarhar provin
ces were recorded to have the highest number of\n', 'deaths and casualties as a result of the earthquake. Food and
 non-food items (NFIs), emergency shelter, and\n', 'psychosocial support services were identified to be among the i
mmediate needs. As the country moved into winter\n', 'season, winterization materials were being prioritized in the
response plan. Access to the affected population\n', '1\n', 'Afghanistan Earthquake, OCHA Situation Report No. 3 (a
s of 12 November 2015)\n', 'DREF Final Report\n', 'Afghanistan: Earthquake\n', '\x0cremained the most significant c
hallenge in delivering humanitarian assistance in a timely and effective manner. With\n', 'the support of the gover
nment, roads were cleared to pave way for humanitarian actors to reach the earthquake\n']
```

3.6 Twenty Five First Lines Datasemt

From the analysis of IFRC pdf reports, most of them have a small table on the top. this table gives the image of what the report is talking about. This table summarizes what the document is talking about. For example the total amount of money spent in recovering a disease, the number of people who participated in a given activity, the location and so on.

While we were transforming the pdf data into txt format, this table occupied almost 25 first lines. Due to the limited time of the research, We decided to split those twenty five first lines of each document. the collection of those first twenty five documents has been considered as our new corpus.

Now we can use one of the algorithms to extract entities and for classification.

3.7 Stanford Named Entities Recognition

The data to be trained is unlabelled. Named Entities Recognizer labels the data to be extracted easily. it recognises sequence of words and its classification is mainly to name of persons, localization and organization.

Stanford Named Entities Recognition is an extractor implemented in java. It takes the sequence of words and label them Stanford named entities recognition is a able to identify correctly the named recogniser which labels sequences of words in a text. The next step is to split the sentences into set of words called tokens. By using the Stanford NER tokenizer where token can be tagged.

- **Stanford NER Tagger** is a package which has modules for classifying tokens with the taggs. A tagg can be defined as one of classes of significant words like nouns, adjectives etc. we used the package Stanford POS Tagger to classify the words.

- **Stanford NER Models** are many Stanford has different models such as "stanford-corenlp-full-2016-10-31", "stanford-ner-2014-01-04" which is the version we used.

- **Stanford Classifier** is a package which classify the entities into defined categories. It has four specific classes such as "Locations", "Persons", "Organizations" and "Others".

We specified the named entities that we wanted to extract. We classified them into the four categories by Stanford classifier. The last category called "others" combined all numerical entities such as time, amount of money, number of people, percentage, etc.

The reports from our corps are order by appeal numbers, the entities are in classified by nltk algorithm.

Figure 3.3: IFRC entities from Stanford NER

	- Global - MAA00001 22Jul2015.txt	- Global - MAA00006 24Apr2015.txt	- Global - MAA00010 10Nov2015.txt	- Global - MAA00021 02Jun2015.txt	- Global - MAA00028 01May2015.txt	- Global - MAA00029 21Jun2016.txt	- Global - MAA00040 02Jun2015.txt	- Global - MAA00040 10Nov2015.txt
locations	[Neonatal]	[Geneva, Geneva]	[Bolivia]	[Sendai, Japan, Geneva, Cali, Colombia]	[Geneva, Panama, Kuala Lumpur, Nairobi, Dubai,...	[Syria, Iraq, Afghanistan, Libya, Ukraine, Yem...	NaN	NaN
organizations	[Global Health Report Health Department 2014 T...	[National Societies, NSKD, International Feder...	[DREF 2013 Number Amount, Red Cross Red Cresce...	[Preparatory Committee of WCDRR, DRR, Fourth G...	[IFRC Global Logistics Service, IFRC, National...	[Red Cross, Red Crescent, IFRC, IFRC, Middle E...	[Federation of Red Cross, Rules for Disaster R...	[IFRC, Crisis Management Department (DCM) Gl...
other	NaN	[2014, 2014, 2014 The Difference Overview The ...	[31 %, 69 %, 2014, 2 per cent, 2013, April 201...	[2015, March 2015, July, November 2014, June, ...	[2015, Strategy 2020, 2015, 2014, 2015, 2014, ...	[2014, 2014, 1990, 7 %, 2014]	NaN	[January June 2015, January 2015 12 months 72]
persons	NaN	NaN	NaN	NaN	[Sierra Leone]	[Jaime Sepulveda, Christopher Murray]	NaN	[Pankaj Mishra, Hakan Karay]

From Figure **??**, Consider the for the report "-Global MAA00029 21Jun2016.txt", locations row shows that the report covered Syria,Irak, Afganistan, Libia,Ukraine, Yemen,etc. The extraction of entities separates clearly the categories.

3.8 Natural Language ToolKit (NLTK)

Natural language toolkit is one of the algorithm to extract named entities. It has different modules which are used to process the data alongside the extraction. NLTK chunkparser is a one of nltk

268 module which uses Regular expressions. NLTK tokenize which splits the sentences into small
269 units called tokens. This module helps the NLTK tagger to identify words independently.

270 Generally NLTK classify the entities into four categories which are known as Location, Organi-
271 zation, Persons and Others.

Figure 3.4: IFRC entities from NLTK

	- Global - MAA00001 22Jul2015.txt	- Global - MAA00006 24Apr2015.txt	- Global - MAA00010 10Nov2015.txt	- Global - MAA00021 02Jun2015.txt	- Global - MAA00028 01May2015.txt	- Global - MAA00029 21Jun2016.txt	- Global - MAA00040 02Jun2015.txt	- Global - MAA00040 10Nov2015.txt
locations	NaN	NaN	NaN	NaN	NaN	[West Africa, West Africa, Caribbean]	NaN	NaN
organizations	[Global Health Report Health, Contents, CBHFA,...	[Global, Difference, National Society, NSKD, N...	[Overview Statistics, DREF, CHF Total, DREF, D...	[DRR, HFA2, DRR, WCDRR, WCDRR, HFA2, UNISDR, W...	[IFRC Global Logistics Service, GLS, IFRC, Nat...	[oPt, Ebola Virus Disease, EVD, Sahel, Horn, R...	[Red Cross, Red Crescent, MAA00040, DCMs, DCM,...	[DEVELOPMENT, UPDATE, INTERVENTION, DCM, CHF, ...
other	[Disease, Maternal, Neonatal, Child, Sanitation]	[Geneva, Long, Geneva]	[Bolivia, Bolivian]	[Sendai, Japan, Geneva, Cali, Colombia]	[Geneva, Panama, Dubai, Las Palmas, Ebola, Ira...	[Iraq, Afghanistan, Libya, Palestinian, Yemen,...	NaN	NaN
persons	[Annual, Annexes Annex, Health, First Aid]	[Knowledge Development Division, Term Planning...	[Start, Red Cross Red Crescent, Emergency Fund...	[Billion Coalition, Climate, Climate]	[Overview, Kuala Lumpur, Nairobi, Guinea, Arab]	[Latin America, Global Health, Jaime Sepulveda...	[Crisis Management, Rules, Disaster Relief, Gl...	[Disaster, Crisis Management Department, Globa...

272 From Figure **??**, Consider organizations extracted from the report "-Global-MAA00021 02 Jun
273 2015.txt", NLTK entities classifier was able to extract DRR, HFAR, WCDRR, HFAR2, UNISDR,
274 etc. The classifier uses nltk tagger and default dictionary which help it to identify the names,
275 verbs and adjectives.

3.9 Polyglot Named classifier

277 Compared to previous entities extractor, Polyglot has only three categories which are "Persons",
278 "Locations" and "Organizations". For nltk, any entity which is classified into those three cate-
279 gories is not considered as named.

280 Let us take an example report "-Global-MAA000029 21 Jun 2016.txt" from Figure **??**, the entities
281 which are classified as "Persons" Jaime, Sepulveda and Christoper Murray.

Figure 3.5: IFRC entities from Polyglot

	- Global - MAA00001 22Jul2015.txt	- Global - MAA00006 24Apr2015.txt	- Global - MAA00010 10Nov2015.txt	- Global - MAA00021 02Jun2015.txt	- Global - MAA00028 01May2015.txt	- Global - MAA00029 21Jun2016.txt	- Global - MAA00040 02Jun2015.txt	- Global - MAA00040 10Nov2015.txt
locations	NaN	[Geneva, Geneva]	[Bolivia, Bolivia]	[Sendai, Japan, Geneva, Cali , Colombia, Cali]	[Geneva, Panama, Kuala Lumpur, Nairobi, Dubai,...	[Syria, Iraq, Afghanistan, Libya, Ukraine, Yem...	NaN	NaN
organizations	[Health, First, Adolescent, Sanitation, Cross,...	[Global, National Society and Knowledge Develo...	[Crescent, Cross, Red Crescent Societies, Disa...	[World Conference, UNISDR, WCDRR, Community Re...	[Logistics, Global Logistics Service, National...	[Global Health]	[International, of Red, Red Crescent, Managemen...	[Crisis Management Department, Crisis Management]
persons	NaN	NaN	NaN	[DRR]	[GLS, GLS, GLS]	[Jaime Sepulveda, Christopher Murray]	NaN	[Simon Eccleshall, Pankaj Mishra, Hakan Karay,...

3.10 Sample Files

The type of data we have can be considered into two different ways. There are some reports which are classifed as CTP documents. This documents cover the overview of how IFRC money was invested in humanitarian activities.

Non CTP reports are focused on other activities which didnt require IFRC to invest money.

The next step is to take example report document for CTP and Non-CTP to have a comparison on the extracted entities.