The Title

By

Alice Irankunda (alice.irankunda@aims.ac.rw)

June 2017

AN ESSAY PRESENTED TO AIMS RWANDA IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN MATHEMATICAL SCIENCES

DECLARATION

This work was carried out at AIMS Rwanda in partial fulfilment of the requirements for a Master of Science Degree.

I hereby declare that except where due acknowledgement is made, this work has never been presented wholly or in part for the award of a degree at AIMS Rwanda or any other University.

Scan your signature

Student: Firstname Middlename Surname

https://pdfs.semanticscholar.org/d3b8/22adbf854ead63144b9472f37784a5ce52a3.pdf

Scan your signature

Supervisor: Firstname Middlename Surname

ACKNOWLEDGEMENTS

This is optional and should be at most half a page. Thanks Ma, Thanks Pa. One paragraph in normal language is the most respectful.

Do not use too much bold, any figures, or sign at the bottom.

DEDICATION

This is optional.

Abstract

A short, abstracted description of your essay goes here. It should be about 100 words long. But write it last.

An abstract is not a summary of your essay: it's an abstraction of that. It tells the readers why they should be interested in your essay but summarises all they need to know if they read no further.

The writing style used in an abstract is like the style used in the rest of your essay: concise, clear and direct. In the rest of the essay, however, you will introduce and use technical terms. In the abstract you should avoid them in order to make the result comprehensible to all.

You may like to repeat the abstract in your mother tongue.

Contents

1. Introduction

[Jan: Please read the text again and correct language mistakes. Also pay attention to punctuation.]

1.1. Introduction

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 the large organizations there a huge number of reports, imagine the way it is challenging go through each and every report manually. [Jan: "imagine the way..." This is not good language for a research essay.] 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, There are some solutions to those questions by using combination of statistics and Natural Language Processing (NLP)techniques. Big data and Machine learning is for analysing the huge data by using statistical and computing algorithms. Document modelling by extracting entities is one of the way to deal with natural big data linguistic problems where entity is defined as a single unit of data, it can be classified based on its relationship, Entity can be location , people, organization and so one.

[Jan: Paragraph above has many topics (organizing information, Red Cross, NLP, document modelling) mixed together. Please separate them into paragraphs.]

Let MDRAF003 be IFRC report "Afghanistan MDRAF003 26May2016.pdf", it is composed by 12 pages of texts. [Jan: s/Let MDRAF003.../For example, consider an IFRC report...] To extract entities from MDRAF003 is challenging, what are the key points to be performed?

- The sentences which compose a report must be parsed.
- Entities also must be identified in the report
- Relationship between entities must be modelled.

In this research, there is a clear discussion about powerful techniques to answer the previous questions. Natural Language Processing techniques used to sentence level and content based analysis,Natural Language ToolKit (NLTK) for spliting the sentences into tokens and remove the common words and how to work with corpus.The used reports for the implementation of different language algorithms are from IFRC .

1.2. Motivation

Big data and Machine learning have recently become one of the major and strong solution finder to most difficult problems in heath, statistical prediction, company development and linguistics.Big data is a future for everything. [Jan: “Big data is a future for everything.” This sentence does not really say anything. I would avoid writing sentences like this.] within huge reports,journals or articles ,this work will return significant classified entities which will help the user to not struggle opening the report and get like amount spent in a given activity, the sum of people who participated in an event etc.

2. Literature review

In today's life, many organizations are generating unstructured data while they are communicating.The entities to be extracted from the reports are wealth. [Jan: This is not good english, you can write "There are plenty of entities to be extracted..."] In this research, English is the considered language. [Jan: Be more specific, e.g., "All reports that we consider are written in English." Also, should this be in literature review?] Natural language toolkit(NLTK) is a a tool which deals with natural language, it is also python platform for human linguistic data. The aim of NLTK is to generate a parse tree with a demonstration of relationship between words of a given sentence and the way those words are classified.

Our report sample is called MDRAF003,let us take one sentence from MDRAF003 and call it S :

"Assessment reports indicated 117 deaths, 544 people injured, 12,794 homes damaged and 7,384 houses destroyed"

There are two mains steps which can be performed to this sencence:

- **Tokenizing**: This is a procedure of taking a sentence and extract the composing atomic linguistic elements means 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')]

  The meanings of the used tags for S :

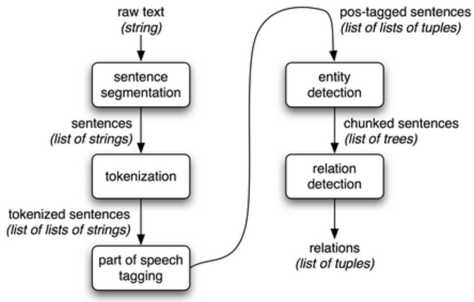
  - JJ : **Adjective** : 'Assessment'
  - NNS :**Noun,plural**:'reports','deaths','people','houses'
  - VBD :**Verbs,past tense**:'indicated','injured','damaged','destroyed'
  - CD :**Cardinal Number**:'117','544','12,794','7,384',
  - CC :**Coordinate Conjugation**:,'and'

  [Jan: Can you give a citation for this method?] The parse tree is formed based on the POS, the classification of word and the way words are arranged in a sentence show a kind of relationship between words. [Jan: The figure is too small to read it.]

Figure 2.1: The parse of the above sentence

The process of classifying entities can be more explained in the following picture [Jan: s/more explained/better explained] [Jan: Is this figure taken from somewhere? If yes, you *have to* provide source. Also, I don't think it is acceptable at all to copy pictures in a research essay. Consult this with Yabebal, maybe he allows it.]

Figure 2.2: information extraction process



To label the boundaries of sentences is one of the important prerequisite steps in Natural language processing(NLP) but the punctuation marks cause some ambiguity (Palmer and Hearst, 1994) for example it is challenging to differentiate the 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 D.Palmer and Marti A.Hearst worked on the problem of punctuations (Palmer and Hearst, 1994), 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 (Baluja et al., 2000).

[Jan: Consider putting text above into separate section (preprocessing for NLP).]

2.1. Named Entity Recognation and Classification NERC

[Jan: Please do not number sections by hand.] The term "Named entity" has been coined in 1996 in "sixth Message understanding Conference"(MUC-6 R. Grishman and Sundheim 1996). [Jan:

No manual citations, use \cite in LaTeX. Also it is not usual to write the conference name, just "has been coined in 1996 by \cite{}"] Entity can be referred as a task, the entity is "named" when it is restricted to one or many rigid designators (Sharnagat, 2014), example: persons,location ,product are the named entities.

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

- ENAMEX: location,product,country ,organization
- NUMEX : percentage,quantity
- TIMEX : time, date

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

2.1.1. 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 needed task to be performed is to find the most likely sequence of tagged names(TN) given a sequence of words(SW).

$$P(TN|SW) = \frac{P(SW|TN)P(TN)}{P(SW)} \tag{2.0.1}$$

The equation (2.0.1) is conditional probability, $P(TN|SW)$ can be called posterior and it is the probability of an event Sequence of word occurring given Tagged names has observed. $P(SW|TN)$ is also called likelihood means it is the probability of obsvering 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 means that it is true even if there is no given evidence at all(masters thesis). Hence,the above sentence is true, there is a permission to say that P(SW) can be ignored. [Jan: "The above sentence is true." Delete that.] [Jan: s/there is a permission to say/we can ignore] the remaining purpose [Jan: s/remaining purpose/remaining objective] is to maximise the probability of getting the sequence of tagged names when sequence of words is given.

$$Max\,[P(TN|SW)] \tag{2.0.2}$$

From the equation (2.0.2) of the maximization , the following estimation can be made

$$P(TN) \approx \prod_{i=1}^{n} P(TN_i|TN_{i-1}) \tag{2.0.3}$$

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

$$P(SW|TN) \approx \prod_{i=1}^{n} P(SW_i|SW_i) \tag{2.0.4}$$

[Jan: There is a typo in formula above. Try explaining better what those estimates mean, something like "we make a simplifying assumption that the tags occur independently from each other".] 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.

$$P(T_i|T_{i-1}) = \frac{K(T_{i-1}, T_i)}{K(T_{i-1})} \tag{2.0.5}$$

Based on the training corpus, $K(T_{i-1}, T_i)$ is referred as a how many times the tag T_i occurs after 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:

$$P(C_i|T_i) = \frac{K(T_i, C_i)}{K(T_i)} \tag{2.0.6}$$

From the equation (2.0.6) , the term $K(T_i, C_i)$ is referred as the sum of the times that a word "C_i" has a tag $T_i$in the training corpus. The process of computing the posterior using the above steps is called Markov model.

[Jan: This is a beginning of a good explanation, but you need to rewrite to correct language mistakes and clarify some points.]

2.1.1.1. Advantages of Hidden Markov Model

[Jan: There is no need for a separate section, just put it in another paragraph.]

It is one of the most powerful statistical and machine learning (ML) techniques in modelling and high qualified in entities extraction.When the researcher is willing to train new data, HMM is very robust and efficient in computations.

2.1.1.2. Disadvantages of Hidden Markov Model

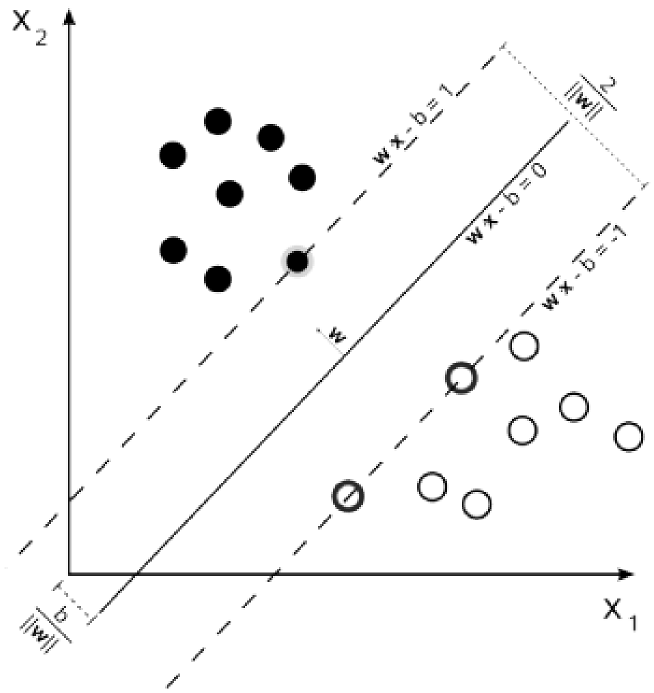
One of the limitations of HMM is that the researcher must have the notion of model topology and statistical techniques on how to deal with large amount of training data.

2.1.2. Supporting Vector Machine based model

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 categorized as either positive or negative and be represented in two dimensional graph. [Jan: You will have many dimensions, not just two] Hyperplane is for separating train documents based on their categories and "w" is a weight vector which is perperndicul to hyperplane is represented by the following equation:

$$w.x - b = 0 \tag{2.0.7}$$

Figure 2.3: SVM in hyperplane representation



From the (2.0.7), 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_{+}. [Jan: I didn't understand this explanation.] Supporting vectors which are demonstrated by the dashed lines on the figure above are formed by :

$$w.x - b = 1 \qquad \text{and also} \tag{2.0.8}$$

$$w.x - b = -1 \tag{2.0.9}$$

There are many algorithms with different approaches to optimization problems but all tends to the same solution says that minimize $\parallel w \parallel$ automatically maximize the margin between h_{-} and h_{+} where the boundary is a half way. [Jan: Don't use \parallel for norm, instead use \|] Now,add another constraint for each document category from the equations (2.0.10) and (2.0.11), in order to hit the target

$$w.x - b \geq 1 \qquad \text{and also} \tag{2.0.10}$$

$$w.x - b \leq -1 \tag{2.0.11}$$

[Jan: Again, I didn't really understand this part.] ......The consideration of non- linear training data[...........still working on it]........

2.1.2.1 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. the machine learning prefers to use the decision tree procedure than SVM. in addition the decision tree has a detailed boolean-like model which is more popular to user.

Overview of rule/patern based systems hand-written rule,decison tree,bootstracpping and

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.(Califf and Mooney, 2003).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 (Patwardhan and Riloff, 2007)

Our research deals with reports generated through a template, compaired to the work of (Patwardhan and Riloff, 2007) 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 (Manning, 2012).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.

$$\theta$$

$$\Updownarrow$$

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.

$$\Updownarrow$$

$$\mathrm{C}$$

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

Information extraction is a combination of segmentation, classification and clustering

3. Third Chapter

References

Shumeet Baluja, Vibhu O Mittal, and Rahul Sukthankar. Applying machine learning for high-performance named-entity extraction. *Computational Intelligence*, 16(4):586–595, 2000.

Mary Elaine Califf and Raymond J Mooney. Bottom-up relational learning of pattern matching rules for information extraction. *Journal of Machine Learning Research*, 4(Jun):177–210, 2003.

Christopher Manning. Information extraction and named entity recognition, 2012.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. Text classification and naive bayes. *Introduction to information retrieval*, 1:6, 2008.

David D Palmer and Marti A Hearst. Adaptive sentence boundary disambiguation. In *Proceedings of the fourth conference on Applied natural language processing*, pages 78–83. Association for Computational Linguistics, 1994.

Siddharth Patwardhan and Ellen Riloff. Effective information extraction with semantic affinity patterns and relevant regions. In *EMNLP-CoNLL*, volume 7, pages 717–727, 2007.

Rahul Sharnagat. Named entity recognition: A literature survey. *Center For Indian Language Technology*, 2014.