What, When, Where and How Much (Who and How?): Extracting datasets from online news articles

Syed Fahad Sultan, Hamza Humayun

Technology for People Initiative

{sultan.fahad, hamza.hamayun}@tpilums.org

ABSTRACT

<Briefly discuss how the motivation for this work is to make data available to policy makers, journalists, academics and general citizens and how availability of data in Pakistan and other developing countries..>

This paper tries to solve this problem by extracting data from a low cost and abundant resource: online news articles. We present a system that extracts What (event), When (time), Where (location) and How Much (intensity). Using online news articles from the most popular english newspapers of Pakistan, we evaluate our results against available datasets and report on accuracy, recall and F-measure scores.

Keywords

Information Extraction, Events Extraction

1. INTRODUCTION

<A paragraph or two on the problem of unavailability of data in Pakistan and developing countries>

In this paper, we try and solve this problem by making use of a vastly abundant and low cost source of data: online news articles. Extractions are essentially tuples of the form (What, Where, When, How Much, Who and How) where What is verb phrase representing the event, Where is the location of the event, When is the time of the event and How Much is the intensity Who is the actor(s) of the event and How is the Context. It is important to highlight here that we do not pre specify events or location or time and so in that sense our system is domain independent. However, the design of the system does limit it to only news article datasets and not any text corpus.

Our extraction works at the sentence level and makes use of standard NLP techniques of dependency parsing, part of speech tagging and named entity recognition.

Furthermore, to each tuple we assign a confidence score based on the intuition that important facts would be reported redundant times and in multiple sources [1,2]

<Contribution of the paper>

<This paper is structured as follows paragraph>

2. RELATED WORK

While analysis of news articles dates back to start of the field of natural language processing itself, open domain information extraction is a fairly new area of research. Some of the pioneering work in this area is done by Oren Etzioni et al. in 2007 paper [3]

<Discuss how none of the other papers focus on numbers>

<Oren Etzioni's open info extraction and Brendan O Connor's paper>

3. DATA

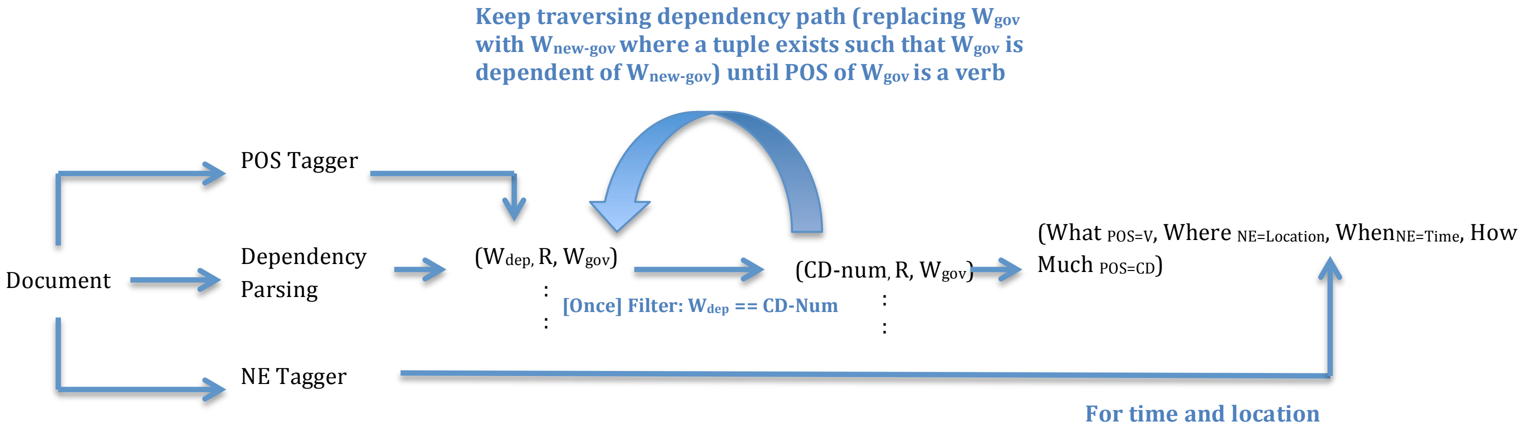
3.1. News articles corpora

We use news articles from online archives of Pakistan's X most popular [citation] english newspapers: Dawn, Express Tribune, The News, The Nation and The FridayTimes.

	Publication	Number of articles	Duration
1	Dawn	X	Jan 2002-Dec 2014
2	Express Tribune		
3	The News		

3.2. Evaluation Datasets

	Data	Source	Duration
1	People injured in bomb attacks	South Asian Terrorism Portal [citation]	1988-2015
2	Change in fuel prices	Oil and Gas Regulatory Authority [citation]	X-Y


Keep traversing dependency path (replacing W_gov with W_new-gov where a tuple exists such that W_gov is dependent of W_new-gov) until POS of W_gov is a verb
POS Tagger
Document
Dependency Parsing
NE Tagger
(W_dep, R, W_gov)
:
:
[Once] Filter: W_dep == CD-Num
(CD-num, R, W_gov)
:
:
(What POS=V, Where NE=Location, When NE=Time, How Much POS=CD)
For time and location

3. METHODOLOGY

4.1. Extraction

For each document in the corpus, the system performs dependency parsing, POS tagging and NER tagging. Currently, it uses Stanford CoreNLP to perform all three of these tasks.

Dependency parsing is performed at the sentence level. For each sentence, the dependency parser returns a list of three part tuples of the form ($W_{dependent}$, relation, $W_{governor}$) where $W_{dependent}$ and $W_{governor}$ are words and relation is one of the dependencies pre-specified in the Stanford Dependency Grammar [citation]. Each of these tuples represents edge of the dependency tree for that sentence, going from $W_{governor}$ to $W_{dependent}$

From this list, we separate out tuples that have dependents with the POS tag: 'cardinal number'. For these tuples, we look at the governor, lets call it g, and its POS tag. If the POS tag is not a verb, we look at tuples T_2 for where g is the dependent. If we come across a verb governor in T_2 we use that as the 'what' for the cardinal number 'how much'. If not, we keep repeating the process until we do. To put it simply, starting from the cardinal number, we traverse the dependency tree up until we come across a verb node.

To extract 'who', the actor for the verb (or the object for the subject 'what'), we use the same technique except for looking for the first noun governor instead of the first verb governor.

4.2. Confidence Score

To each tuple we assign a confidence score based on the fact that important events are generally reported multiple times and in multiple sources.

4.2. Query Processing

For query processing, we using synonyms and stemming of the 'what' verb and the query input. Currently, we are using the Lancaster Stemmer [citation] that comes with the Stanford CoreNLP package for stemming and Wordnet for synonyms.

3. RESULTS

<Extraction results without evaluation>

4. EVALUATION

	Query						Evaluation Dataset	Accuracy	Recall	F-Measure
	What	Who	How	Where	When	How Much				
1	Injured	People	Bomb blasts				South Asian Terrorism Portal			
2	Increased	Fuel prices					Oil and Gas Regulatory Authority, Pakistan			
3	Died	People	Dengue							
4		Energy shortfall					LSE Report			

5. CONCLUSIONS AND FUTURE WORK

ACKNOWLEDGEMENTS

8. REFERENCES

1. Assuming Facts are Expressed more than Once. Justin Betteridge, Alan Ritter and Tom Mitchell.

2. A Probabilistic Model of Redundancy in Information Extraction. D. Downey, O. Etzioni, and S. Soder- land. In *Proc. of IJCAI*, 2005