SNLP-MinProj
Group name : SmartBV^2

Burhan Otour

Vineet Vasishta Eranki

Sandeep Varma Ganaraju

GitHub repo: https://github.com/BurhanOtour/SNLP-MinProj.git

Corpus creation

- Data Set : New York Times 2013
- Number of Articles: Approx. 10,000 Articles
- Source Format: JSON
- After cleaning of JSON tags , the file is converted to plain text.

Corpus Analysis

1. Named Entity Extraction

- FOX (Federated knOwledge eXtraction framework)

  provides users with accurately disambiguated and linked named entities in several RDF serialization formats

- Fox implements below 4 NER Tools:

  OpenNLPEN

  StanfordEN

  IllinoisExtendedEN

  BalieEN

  Stanford NER Model used is

  3 class: Location, Person, Organization

NER 7 class: Location, Person, Organization, Money, Percent, Date, Time

Normal Text

His death was reported in Huntsville, Ala., by Hugh McInnish, a friend and retired military specialist who said Rudolph died in an ambulance on the way to a Hamburg hospital.

As project manager, the sheer size of the moon rocket was the source of his nightmares during its development and construction. ``It was all harder than I ever expected,'' he told an interviewer in 1969. ``Since it is so large and has so many components, you get more people involved, and that makes it difficult. You have to get them all to sing from the same sheet of music. My task has been as choir director.''

After Named Entity Extraction

His death was reported in <ENAMEX TYPE="LOCATION">Huntsville</ENAMEX>, <ENAMEX TYPE="LOCATION">Ala.</ENAMEX>,
by <ENAMEX TYPE="PERSON">Hugh McInnish</ENAMEX>, a
friend and retired military specialist who said <ENAMEX TYPE="PERSON">Rudolph</ENAMEX> died in an
ambulance on the way to a <ENAMEX TYPE="LOCATION">Hamburg</ENAMEX> hospital.
As project manager, the sheer size of the <ENAMEX TYPE="LOCATION">moon</ENAMEX> rocket was the
source of his nightmares during its development and construction.
``It was all harder than I ever expected,'' he told an interviewer
in <TIMEX TYPE="DATE">1969</TIMEX>. ``Since it is so large and has so many components, you get
more people involved, and that makes it difficult. You have to get
them all to sing from the same sheet of music. My task has been as
choir director.''

FOX also disambiguates and links named entities against DBpedia by relying on the AGDISTIS framework

2.Entity Disambiguation

- Link entities against every Linked Data knowledge base.

3.Relation Extraction

- find relations between two entities.
- Used for extracting relation triples from Text.
- **triple** is a data entity composed of subject-predicate-object
- Example : [Leipzig University] located in [Leipzig]
- RDF triples are stored in Turtle document in .TTL file extension

Fact Checking and Benchmarking

- The truth value τ (e) ∈ [0, 1] of a new statement e = (s, p, o) is derived from a transitive closure of the Wikipedia Knowledge Graph [1]. More specifically, the truth value is obtained via a path evaluation function:

$$\tau(e) = \max \mathcal{W}(P_{s,o}).$$

- Based on this true statements are assigned higher truth values than false ones with probability.

[1] Computational Fact Checking from Knowledge Networks, Giovanni Luca Ciampaglia , Prashant Shiralkar, Luis M. Rocha, Johan Bollen,Filippo Menczer,Alessandro Flammini