Information Extraction from Business News Articles

Madhura Dole
Department of Computer Science
The University of Texas at Dallas
Email: mxd1624430@utdallas.edu

Abstract

Detecting Named Entities(NEs) and discovering relations among them from given text is both a challenging as well as useful task in the domain of Natural Language Processing, with applications in Information Retrieval (IR), Summarization (SUM), Question Answering (QA) and Textual Entailment (TE).
The work presented here is the result of an attempt to solve practical issues faced while extracting various relationships among different business entities appearing in Business News articles. The approach consists of applying regular expression patterns on a large corpus – 'ieer' – for the extraction of relations between two named entities.

1. Introduction

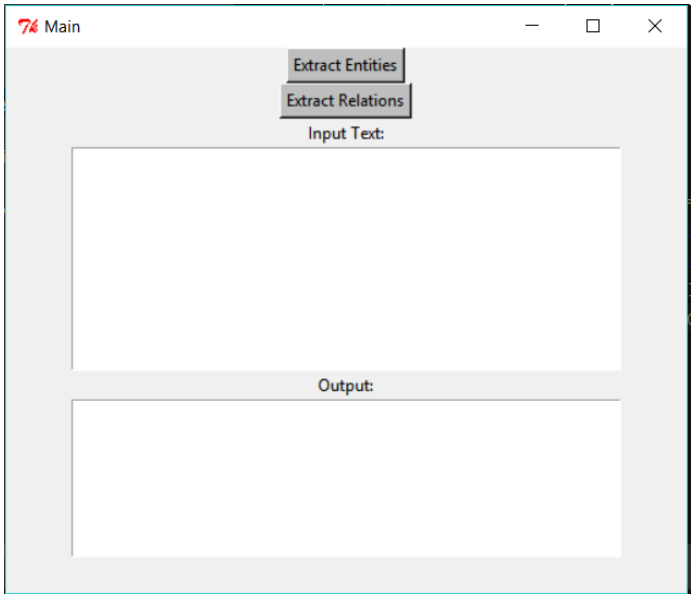
Business articles are published every minute around the world announcing achievements of multinational organizations, startup initiations, business unit take-overs etc. It is very important for companies and their employees to be able to easily access and interpret various happenings in business industry and use this information to develop new business strategies for keeping their heads in the competition. These articles are a great source of information and the work presented here aims at developing a software capable of performing Named Entity Recognition and Relation Detection. In this project, we attempt to develop an IE engine capable of performing Named Entity Recognition(NER) and Relation Extraction(RE) using Python language. Entities focused in this project include business leaders, cofounders, employees, organizations, locations and relations considered consist of organization-location, person-organization etc.

2. Related Work

There have been many successful high accuracy attempts of solving the problem of Named Entity Recognition and Relation Detection in the field of Information Extraction. Branimir T. Todorovic and Svetozar R. Rancic proposed a system for Named Entity Recognition and Classification using Context Hidden Markov Model. Mohamed Hashem proposed A Supervised Named-Entity Extraction System for Medical Text. Andreea Bodnari. Louise Deleger proposed system for Effective Adaptation of a Hidden Markov Model-based Named Entity Extraction from medical text [1].
In this project, we aim to extract Named Entities from a given text and their relations building on the algorithms and methods provided in previous work of NER and Relation Extraction systems.

3. Proposed Work

The proposed system, built on MVC framework, consists of two main modules NER and Relation Detection. Relation Detection module is programmed to extract relations of type ORGANIZATION-LOCATION and PERSON-ORGANIZATION.
This IE engine processes input text provided by the user and allows user to choose between two options as shown in the below image.



3.1Flow of the Proposed System

This system makes use of Natural Language Toolkit(NLTK) to perform NER and extract relations among entities and present the results in user understandable format. Following shows a high-level architecture.

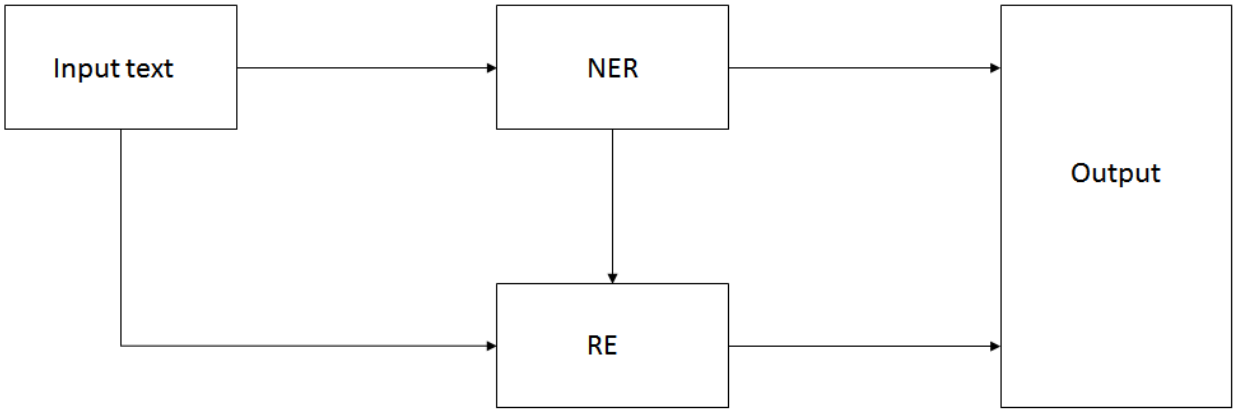


Fig. 3.1 High-level architecture of IE engine

From a high-level perspective, this IE engine will pass raw input text to two main modules, NER and RE and output the results on screen. NER and RE modules, explained further in the report, undergo numerous other steps of text processing to produce expected results.

3.2Named Entity Recognition

A named entity can be a word, a number or a phrase that is used to identify persons, locations, organizations or even numerical entities such as dates, time, money and numbers. The project uses NLTK for performing NER on business articles, mainly categorizing entities in person, location, organization and date classes.

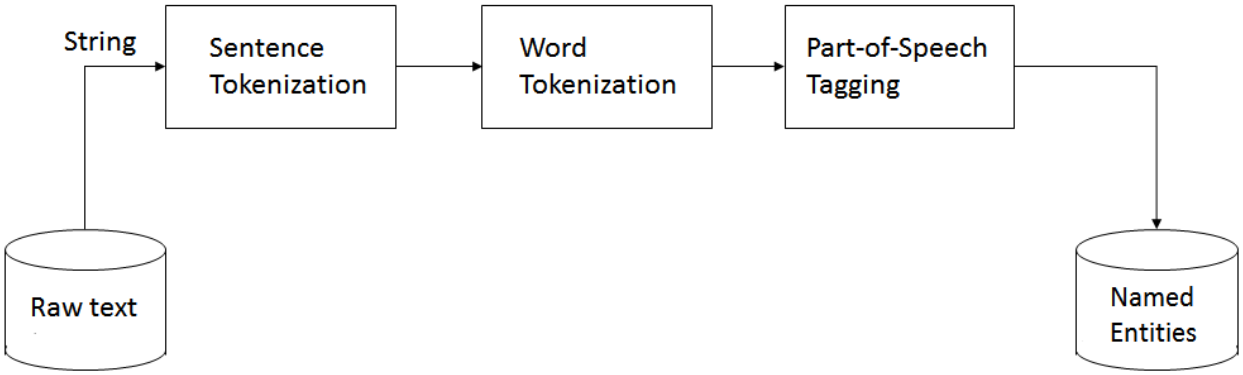
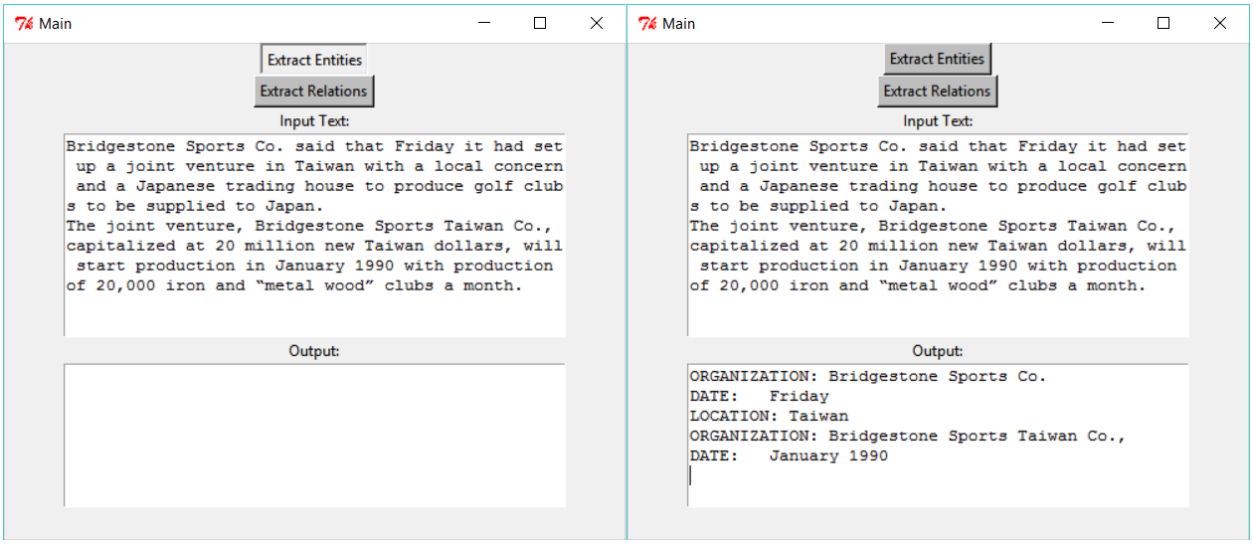


Fig. 3.2. Named Entity Recognition

Below images represent the output of NER system designed in this project. This NER system achieves a good accuracy for most of the input text.



3.3Relation Extraction between NEs

Relationship mining or Relation Extraction(RE) seeks to identify a relation of interest that might exist between two or more entities, however for this project we focus on relation between two entities only.
In this RE module, NLTK methods are overloaded to extract relations from chunked sentences of the text. This project focuses on extracting two types of relationships between two NEs:

1. A PERSON belonging to a certain ORGANIZATION.
2. LOCATION of an ORGANIZATION.

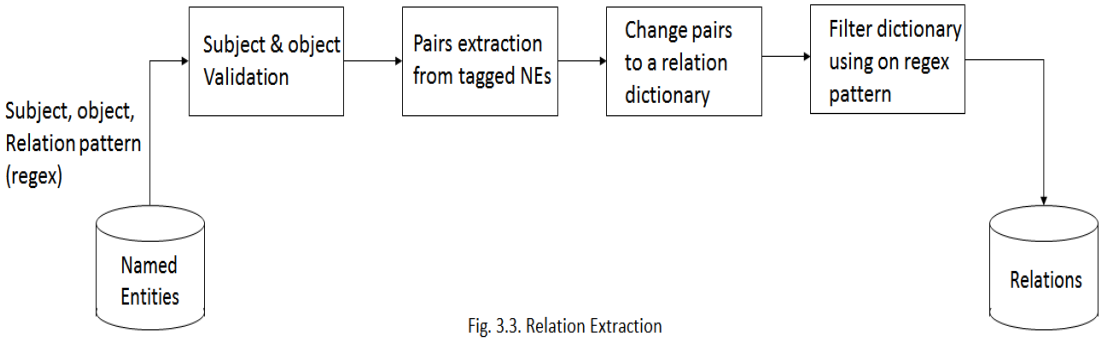
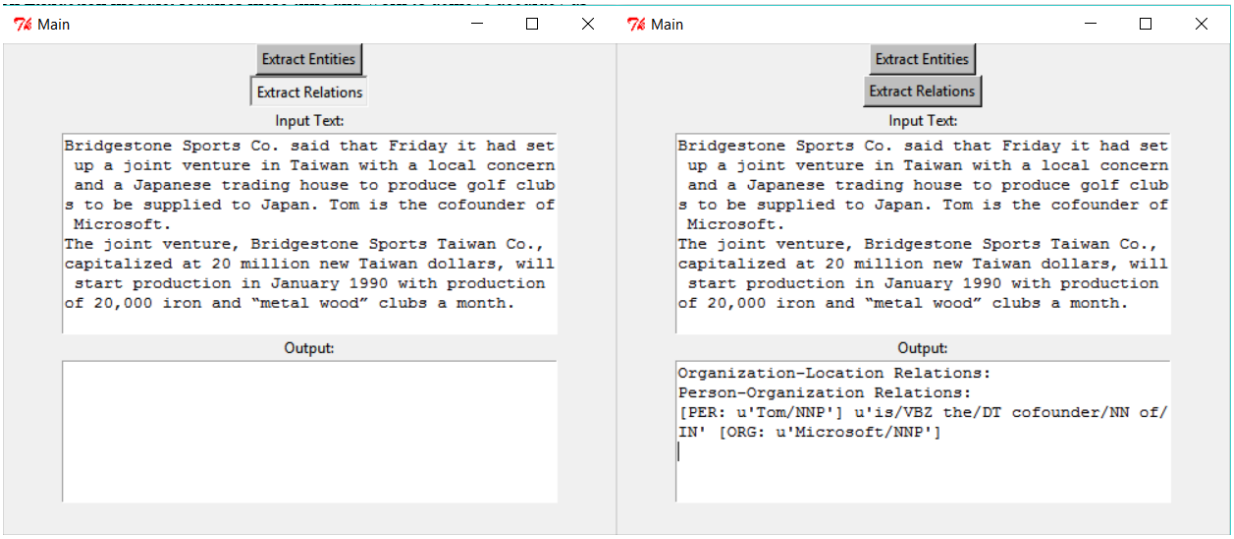


Fig. 3.3. Relation Extraction

- The process involves checking validity of ‘subject class’ and ‘object class’ as the first step. Next, we extract pairs from NE tagged inputs.
- We use a tree structure to determine the relation between specified class of entities and return a chunked sentence in the form of list containing, tree of subject class, possible relation in the middle items of the list, and a tree of object class.
- Ex. - [[[], Tree ('PERSON', [('Tom', 'NNP')])], [[('is', 'VBZ'), ('the', 'DT'), ('cofounder', 'NN'), ('of', 'IN')], Tree ('ORGANIZATION', [('Microsoft', 'NNP')])]]
- This tree structure and the regex (Regular Expression) pattern is then provided as an input to a filtering function, that processes each sentence from the text and return a list output containing of sentences matching with the pattern.



4. Resources Used

- Here, we make use of Natural Language Toolkit(NLTK) available for Python language.
- For the NER module, we use sentence, word tokenizer of NLTK followed by Part-of-Speech tagger. The result of POS tagging is passed to a chunking function to tag entities into different classes ORG, LOC, PER, DATE etc. The chunked result is parsed to produce in a user understandable format. The results are verified using Stanford NER Tagger available in NLTK package.
- For the Relation Extraction module, we overload methods of "nltk.sem.relextract" to extract relations of type 'location of an organization' and 'person of an organization'. The overloading of methods allows to write our own implementation of filtering the regex pattern of a relationship thereby allowing more accuracy in results.

5. Summary

- To summarize, the NER module of the project provides nearly perfect accuracy towards expected output except for some unforeseen exceptions that may occur for some ambiguous words.
- The Relation Extraction module, requires more time and work to achieve accuracy as good as NER. However, not entirely unsuccessful, the RE module can extract, 'location of an organization' and 'person of an organization' relationships hidden in input text. The output shown to the user on the screen is a set of tagged sentences preceded by the type of the relation they have.
- Following tables explain the results achieved for NER and RE modules.
  Totally correct: The extracted information was exactly the desired information.
  Nearly correct: The extracted information was quite close to what we were looking for, but sometimes had unrelated text in the output.

Field	Totally Correct	Nearly Correct
PERSON	70%	80%
LOCATION	75%	85%
ORGANIZATION	70%	85%
DATE	60%	70%

Table 1: Results achieved using POS tagging and chunking for NER

Relations	Totally Correct	Nearly Correct
PERSON of an ORGANIZATION	55%	70%
LOCATION of an ORGANIZATION	20%	50%

Table 2: Results achieved using pattern matching for Relation Extraction

6. Conclusion

As the research for achieving higher level of accuracy for NER and Relation Extraction still goes on, this project attempts to extract information to highest accuracy possible. In case of relation extraction, this project tries to extract maximum number of relations observed in the text/corpus. Also, this method does not require annotated data and therefore can be applied to different domains and corpora. However, the project will require more insight and work to extract relations and entities not captured through method proposed here. To extract information not found by the pattern matcher we can use a Hidden Markov Model trained on pre-tagged announcements. To evaluate accuracy of results we need use more training data and make use of Machine Learning algorithms to train and classify the data and calculate results.

7. References

[1] Andreea Bodnari, Louise Deleger, Thomas Lavergne, *"A Supervised Named-Entity Extraction System for Medical Text"*.

[2] Félix López , Víctor Romero *"Mastering Python Regular Expressions"*, Packt Publishing, February 21, 2014.

[3] Nitin Hardeniya, *"NLTK Essentials*", Packt Publishing, July 27, 2015.

[4] Gunjan Dhole, Dr. Nilesh Uke, *"Medical Information Extraction Using Natural Language Interpretation"*, Advances in Vision Computing: An International Journal (AVC), Vol.1, No.1, March 2014.

[5] Califf, M.E., & Mooney, R. (1997). *"Relational learning of pattern-match rules for information extraction"*. Working Papers of ACL-97 Workshop on Natural Language Learning.

[6] Adrian Iftene, Alexandra Balahur-Dobrescu, *"Named Entity Relation Mining Using Wikipedia"*.

[7] *"Novel algorithms for relationship mining*", European Seventh Framework Programme FP7-218086-Collaborative Project.

[8] Jacob Perkins, *"Python 3 Text Processing with NLTK 3 Cookbook"*, Packt Publishing, August 26, 2014.