ENHANCING NAMED-ENTITY RECOGNITION WITH DBPEDIA

Ramya Kalathingal and Maurani Saha

Guided By : Dr. Richard Scherl

MONMOUTH UNIVERSITY
STUDENT SCHOLARSHIP WEEK

Introduction

An important component of processing natural language texts is to identify the entities named in the text. For example, news articles talk about politicians, companies, cities, dates and so on. We have been using off-the-shelf NLP (natural language processing) tools, including parsers and a named entity recognizer. Even though the named entity recognizer is state-of-the-art, there is significant room for improvement as it makes some errors and identifies named entities as belonging to a limited set of categories. One of our aims is to improve the depth of the information about named entities.

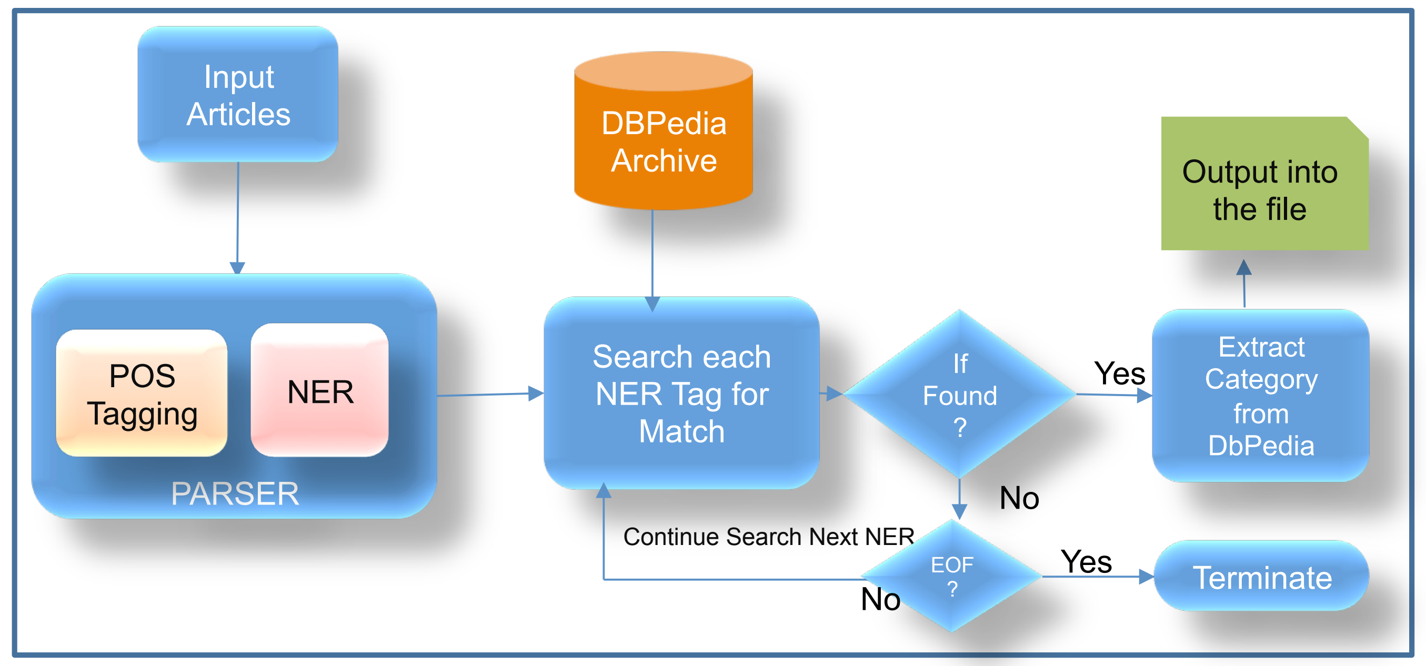
Research Question

Can DBpedia be used to enhance named-entity recognition?

Methods

- We are using a state-of-the-art parser to process natural language texts such as news articles. The parser uses a named-entity recognizer (NER) to identify phrases as names of people, organizations, and locations.
- DBpedia is a semantically organized database that is automatically created from the structured content of Wikipedia. In accordance with the principles of the Semantic Web, DBpedia has a large ontology that further categorizes people, organizations and locations.
- By accessing the latest archives of DBpedia, we are experimenting with matching the named entities found in the text with those categorized by DBpedia to obtain a finer-grained categorization.
- For example, this technique allows us to identify people as scientists or politicians, and organizations as companies or universities. Our goal is to use this information to improve the performance of text clustering and classification methods.

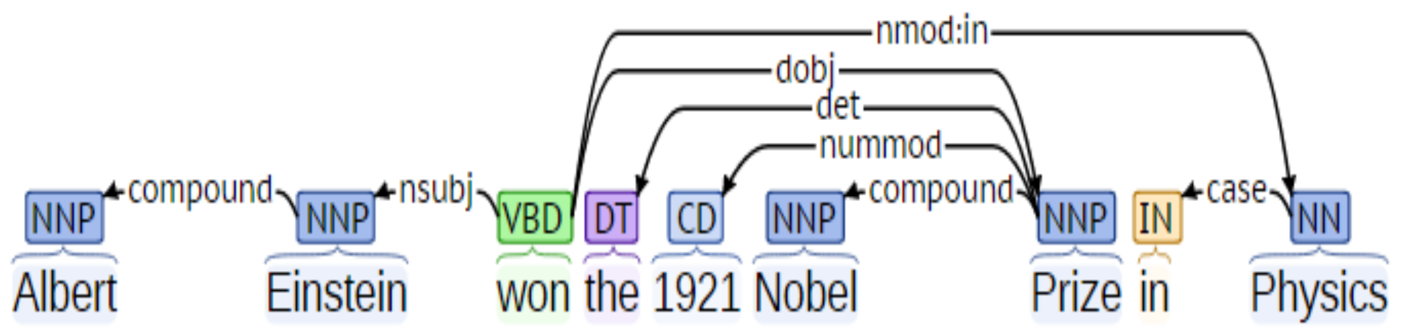
Work Flow



Dependency Parsing

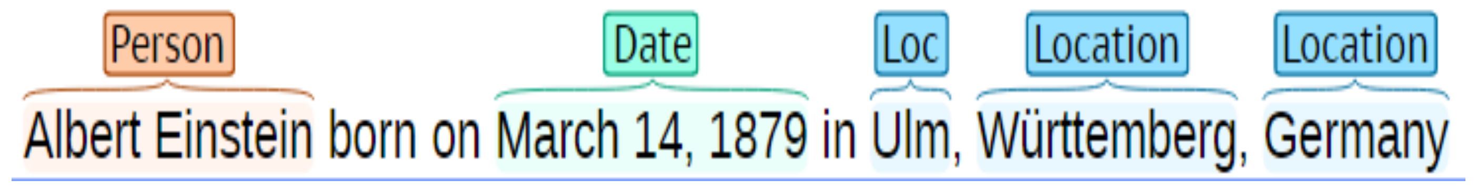
- A dependency parser builds a parse incrementally as it scans the words in a sentence. The approach is relatively fast as the time taken is linear in the length of the sentence.
- A dependency parse is an analysis of the grammatical structure of a sentence, establishing relationships between "head" words and words which modify those heads. The parse is represented as a set of triples (the name of the relation, the governing word and the dependent or modifying word). It also has a natural graphical representation.
- We have experimented with two different dependency parsers, MaltParser and DependencyParseAnnotator, which is part of the Stanford CoreNLP suite of tools. Both are written in Java. We have used them from Java and also with a Python wrapper provided by the NLTK toolkit.
- Given below is a graphical representation of the dependency parse for the text:

  "Albert Einstein won the 1921 Nobel Prize in Physics."



Stanford NERClassifierCombiner

- The NERClassifierCombiner from the Stanford CoreNLP suite of tools identifies named entities in the text and classifies them as Person, Location, and Organization. It also identifies numerical entities as Money, Number, and Percent, and temporal entities as Date and Time. Below is the visual representation of a the output on the text:
  "Albert Einstein born on March 14, 1879 in Ulm, Württemberg, Germany."


Person
Date
Loc
Location
Location
Albert Einstein born on March 14, 1879 in Ulm, Württemberg, Germany

Tools & Technology

Python 3, Java 1.8, JDOM Parser, Stanford core NLP Parsers, NLTK parsers, Malt Parser.

DBpedia

- The data stored in DBpedia is written in RDF (Resource Description Framework). Statements in RDF are triples of subjects (a thing represented by a URI, a predicate, and a value).
- The current version of DBpedia describes 4.58 million things. Most of these have a URI which is resolvable to a web page. Additionally most of the things have a *type* relation with a value in the DBpedia ontology. Most have the *primaryTopic* predicate relating it to the relevant Wikipedia page.
- There are many other predicates describing different entities, e.g., *birthdate*, *title* (position held for elected officials).

A Portion of the DBpedia Ontology for Person

- Ambassador
- Archaeologist
- Architect
- Aristocrat
- Artist (Actor, Comedian, ComicsCreator, Dancer, FashionDesigner, Humorist, MusicalArtist, Painter, Photographer, Sculptor).
- Astronaut
- Athlete (ArcherPlayer, AthleticsPlayer, AustralianRulesFootballPlayer, BadmintonPlayer, BaseballPlayer….)
- BeautyQueen
- BusinessPerson
- Celebrity,
- Cleric
- .
- .
- .

Sample Output

`Fractivists' Increase Pressure on Hillary Clinton and Bernie Sanders

New York Times April 4, 2016

Below are part of the results from the processing of this article. The NER output is on the left and the Dbpedia type is on the right

('Bernie Sanders', 'PERSON')	OfficeHolder
('Greenpeace', 'ORGANIZATION')	Organisation
('New York', 'LOCATION')	AdministrativeRegion
('Pennsylvania', 'LOCATION')	AdministrativeRegion
('United States', 'LOCATION')	Country

Future Work

- Make use of the wide variety of predicates in DBpedia.
- Use the added information provided by Dbpedia to enhance the quality of the output of clustering and classification algorithms when applied to news articles.
- Use techniques such as locality-sensitive hashing to speed up the string matching process

Web Resources

- Dbpedia: http://wiki.dbpedia.org/
- Malt Parser: http://www.malt.marser.org
- Natural Language Toolkit: http://www.nltk.org/
- Semantic Web: http://semanticweb.org
- Stanford Core NLP Tools: http://stanfordnlp.github.io/CoreNLP/

References

- Daniel Jurafsky and James H. Martin. *Speech and Language Processing*, Second Edition, Prentice Hall. 2008.
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer: DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6(2): 167-195 (2015)
- Jure Leskovec, Anand Rajaram and Jeffrey David Ullman. *Mining of Massive Datasets* (Second Edition). Cambridge University Press. 2014.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language ProcessingToolkit. Pages 55-60 in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014
- Christopher D. Manning, Prabhakar Raghvan and Hinrich Schütze. *Introduction to Information Retrieval*, *Cambridge University Press*. 2008.

Acknowledgments

We thank Teena Gopi for sharing her earlier work on using the Stanford Core NLP Parser and Named Entity Recognizer