International Journal of Latest Trends in Engineering and Technology
Vol.(10)Issue(1), pp.349-352
DOI: http://dx.doi.org/10.21172/1.101.63
e-ISSN:2278-621X

DBPEDIA BASED INFORMATION EXTRACTION FROM UNSTRUCTURED DATA

Neha Jain[1], Prof. Lalit Sen Sharma[2]

Abstract - In this paper, an Ontology based Named Entity Recognition (NER) technique to annotate text in unstructured data is presented. The proposed technique is used for annotating unstructured text derived from newspaper articles. Rules are defined using a pattern matching grammar called JAPE (Java Annotation Pattern Engine). The annotations obtained from the DBpedia Spotlight are investigated and parsed using the returned annotation property 'types'. JAPE rules are included to create new annotations for specifying particular named entities in the categories of Person names, Places and Organizations. The mentions in these three categories in particular and all other annotations marked using DBpedia are evaluated and the results are presented.

Keywords - Ontology, DBpedia, Ontology based Information Extraction, Semantic Web, JAPE.

1. INTRODUCTION

The Semantic Web is considered to be the next generation of the World Wide Web as marked by the standards specified by the World Wide Web Consortium(W3C). The standard endorses some common formats and protocols for the exchange of data on the WWW. The most common format being the Resource Description Format(RDF). According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across applications, enterprises and community boundaries".[1] The idea of the Semantic Web as visualized by its inventor Tim Berner Lee [2] underlies the viewpoint of the web being a navigable space with a possible mapping from resources to a unique string called URI(Uniform Resource Identifier). The URI serve as a unique identifier for the available resources [3]. Semantic Web Technologies facilitate the creation of data stores on the Web, create vocabularies which are machine understandable and also to create rules for the usage of data. The DBpedia Ontology is a cross-domain ontology based on the concepts derived from the largest online community developed Encyclopedia- 'Wikipedia' [4]. It is a huge information resource covering information about an extremely vast variety of fields and subjects. But the Wikipedia text is readable only by human users. To make the content machine readable and understandable, an initiative towards the Semantic Web is the RDF based counterpart of Wikipedia - DBpedia. It provides a machine readable form of the concepts available in Wikipedia. It provides information about nearly 4.58 million things. DBpedia Ontology is a structural hierarchy of classes and properties in RDF (Resource Description Format) which is a Semantic Web standard for the representation of data. DBpedia consists of almost 3 billion RDF triples, of which 580 million are extracted from the English Edition of Wikipedia and 2.46 billion from other language editions.

The work presents an ontology based technique to annotate text from unstructured data derived from newspaper articles taken from 'The Hindu'. Ontology Based Information Extraction(OBIE) is one of the emerging techniques to be used with the Semantic Web Data. OBIE is different from conventional IE as it locates type of extracted things and entities by linking it to its semantic description available in the linked formal description that is the ontology. The presented technique is related to knowledge representation and has vast scope to support the development of the Semantic Web. An ontology is a formal and explicit specification of a shared conceptualization. Conceptualization refers to an abstract model of some phenomenon occurring in the world by having identified the relevant concepts of that phenomenon. Explicit refers to the type of concepts used and the constraints on their handling are defined clearly and concisely. Formal refers to the notion that the ontology should be machine readable and understandable. Shared points to the constraint that the ontology captures consensual knowledge, that is it should be widely acceptable and procurable. To sum up, the nucleus of an ontology is a data model for expressing and clearly describing the entities in the world. The model essentially contains a set of types, properties and relationship types. DBpedia is one such ontological model derived from Wikipedia. The concepts from Wikipedia are collected using its structured counterpart- the DBpedia ontology for further mining and extracting the mentions of the named entities from the unstructured text data under experimentation. It is a common observation that the text in news articles is often abounded with names of persons, places, organizations, time, spatial information etc. In the experiment, the text is annotated with the structured content derived from DBpedia using the GATE developer environment [5] and rules written in JAPE language which is a JAVA based grammar for annotating text. GATE developer is an integrated development environment for language processing components. JAPE is a finite state transducer and brings in the capability to identify regular expressions in annotations on documents. A JAPE grammar rule comprises of a set of phases, each of which consists

[1] Department of Computer Science and IT, University of Jammu, Jammu, India
[2] Department of Computer Science and IT, University of Jammu, Jammu, India

of a set of pattern/action rules. The phases execute in a sequential manner and represent a cascade of finite state transducers over the annotations. The LHS of the rules represent an annotation pattern depiction. The RHS is composed of the statements to manipulate the annotations.

2. LITERATURE REVIEW:

Daya et al. presented an introduction to ontology-based information extraction systems and also reviewed the details of different available OBIE systems. They investigated different systems and attempted to identify a common architecture among these systems and organized them according to varied parameters, leading to better understandability of the systems. They also discussed the implementation details of these systems, the tools utilized by them and the metrics employed to evaluate their performance. [6]

K. Nebhi [7] in his work described a rule-based methodology to carry out automated semantic annotations of named entities in a newspaper article corpus. He established a connection between the French Named Entities, the DBpedia ontology and the DBpedia databank. In another paper [8] the author presented an ontology based technique to extract useful information from various tweets scrapped from Twitter.

D Maynard et al. [9] in their work investigated various NLP techniques for ontology population, using a combination of rule-based approaches and machine learning. They described a method for term recognition using both linguistic and statistical techniques, making use of contextual information to bootstrap learning. They investigated various means by which of making term recognition techniques useful for the task of information extraction. They also evaluated an ontology based information extraction technique.

Alexiei Dingli et al.[10] presented a method to include machine-understandable meanings in the documents which are available on the World Wide Web by using Natural Language Processing and state of the art web technologies like RDF. They proposed an application that uses Information Extraction techniques to extract patterns from a human readable text and used the derived method to find similar patterns on varied texts extracted from the World Wide Web.

Fernando Gutierrez et al. [11] in their work proposed Ontology Based Components for Information Extraction (OBCIE). The architecture they proposed promote reusability through modularity. They introduced two orthogonal extensions to OBCIE that facilitate the construction of hybrid OBIE systems with a high extraction accuracy and added functionalities. The first extension use OBCIE modularity to integrate different types of implementation into one extractedd system, creating extractions with a higher rate of accuracy. For each concept in the ontology, they either selected the best implementation for extraction, or merged both implementation strategies under a single learning schema. The second extension they proposed is a new ontology-based error detection mechanism. Following a heuristic approach, they proposed a mechanism to recognize sentences that are plausibly conflicting with the domain ontology. Since, the implementation strategy proposed by them for the extraction of concepts is independent of the functionality of the extraction, they presented a design strategy for a hybrid OBIE system with concepts utilizing different implementation strategies for extracting correct or incorrect sentences. Their evaluation results from the conducted experiments mark that their proposed method is more accurate in provisos of correctness and completeness of the extraction process. They also empirically presented that their error detection method identifies the incorrect statements with a higher accuracy.

Ontology Based Information extraction

Information Extraction is the task of extracting the needed content from structured or unstructured text. Ontology based Information Extraction is a sub-field of this vast area of Information Extraction which includes a wide area of tools and techniques. Ontologies provide a formal and explicit specification of conceptualizations[11], hence, they are a potential candidate for automated information extraction process and the development of the Semantic Web. OBIE systems utilize formal ontologies for the task of formally describing the domain knowledge utilized by these systems for the functioning of their processes. This information extraction strategy is deployable at all levels of ontological knowledge ranging from domain entities for the named entity recognition (NER) to the use of conceptual hierarchies for pattern generalization, to the application of conceptual hierarchies for pattern generalization, to the use of properties and relations which do not appear in the hierarchy for pattern acquisition and ultimately to the utilization of the domain model itself for integrating the churned out entities and the instances of relations, in addition to the ascertaining of the implicit facts and the identification of inconsistencies. [13]

In this work, DBpedia ontology is chosen for Information Extraction. Both a general and a domain specific DBpedia ontology based information extraction task have been taken up. DBpedia Spotlight web service has been used for annotating the mentions of DBpedia resources in the investigated text. DBpedia provides capabilities for linking the unstructured information sources to the Linked Open Data Cloud. DBpedia Spotlight executes named entity extraction, including subtasks like entity detection and name resolution or disambiguation. DBpedia Spotlight intends to be adapted for many use cases. This capability of the DBpedia Spotlight has been exploited in the current research work. The annotations provided by the DBpedia Spotlight are worked upon using the JAPE grammar rules to produce customized annotations providing a more interoperable annotation set. The results are evaluated based on their precision, recall and F- measure values.

3. RESULT

3.1 Experimental Settings

A corpus of 40 newspaper articles was prepared, choosing articles from different sections like Sports news, Economics news, International news and local news. The articles were preprocessed using the ANNIE plugin of the GATE(General Architecture for Text Engineering) tool. A pipeline has been created for document preprocessing and annotating text using DBpedia Spotlight and JAPE(Java Annotation Pattern Engine) rules.

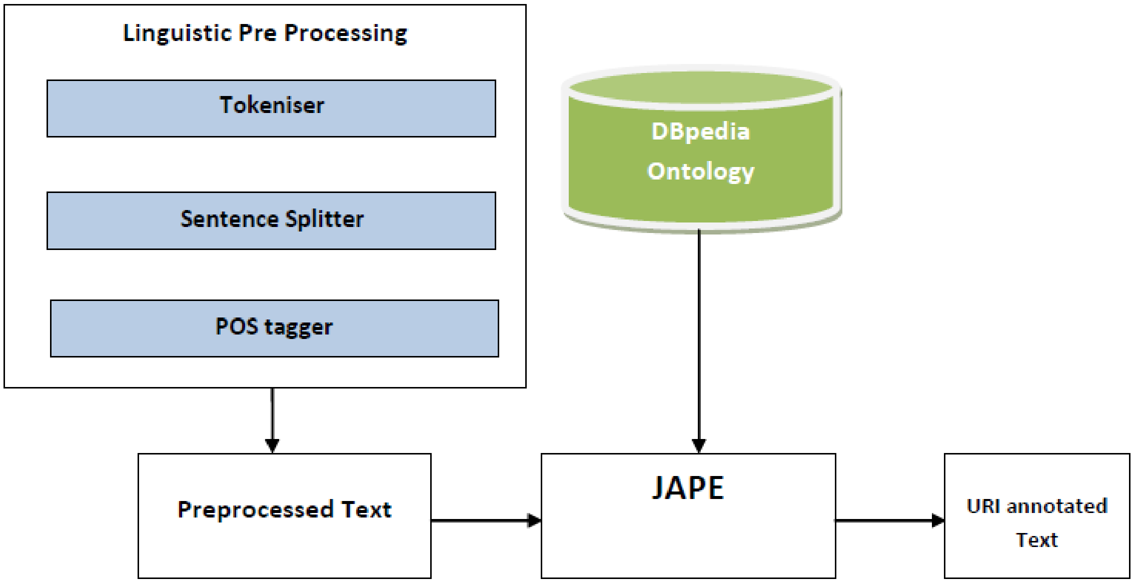

Diagram 1: The experimental Setup

The preprocessing step included ANNIE English Tokeniser, RegEx Sentence Splitter and ANNIE POS (Part-of-speech) Tagger. The Tokenizer split the input text into atomic/ simple tokens. The next process in the pipeline grouped all tokens into sentences using regular expressions and then the resultant text was assigned a correct word class by the Part-of-speech tagger. DBpedia Spotlight annotated the mentions of the resources matched in DBpedia ontology in the investigated text providing capabilities valuable for Named Entity Recognition, Name Resolution etc [12]. The annotations available from the DBpedia Spotlight were parsed using the returned annotation property 'types' using JAPE rules to create new annotations for specifying particular named entities in the categories of Person names, Places and Organizations. The mentions in these three categories in particular and all other annotations marked using DBpedia were evaluated for their precision and recall values against manually annotated values which served as the gold standard for the data set under investigation.

```
if(mention1.getFeatures().get("types").toString().contains("Country"))     {

FeatureMap lookupFeatures = mention1.getFeatures();

gate.FeatureMap features1 = Factory.newFeatureMap();
features1.putAll(loo
outputAS.add(as.firs                                          "DBpedia_Country"
, features1);
```

JAPE rule snippet

3.2 Experimental results:

	Person	Place	Organization	Miscellaneous
True Negatives	NA	NA	NA	NA
True Positives	40	152	34	1299
False Negatives	24	15	15	208
False Positives	15	13	22	814

Table 1: Observations for the chosen mention types.

	Person	Place	Organization	Miscellaneous
Precision	0.72	0.92	0.60	0.61
Recall	0.63	0.91	0.69	0.86

F-Measure	0.64	0.91	0.64	0.71

Table 2: Precision, Recall and F-Measure values for the chosen mention types.

3.3 Observations:

It was observed during experimentation that the URI returned for some common named entities like 'The', 'He', 'This' etc. are highly irrelevant also the feature 'type' is returned none for many entities. This is the prime reason for the low Precision and F-measure values. The values can be significantly improved by fetching the annotation feature 'types' for all the annotations. The type 'Place' has a comparatively high Precision, Recall and F-measure values as most of the geographical locations have their mentions in DBpedia.

4. CONCLUSION AND FUTURE WORK:

DBpedia is a huge collection of data set with an enormous information in the form of RDF about a vast number of entities in the world. The precision and recall values for these entities can be used by adding SWRL rules to the DBpedia ontology for concept extraction. A comprehensive study of the A-box and T-box of the DBpedia ontology and adding of suitable SWRL rules can be undertaken to improvise upon the precision and recall values.

5. REFERENCES

[1] "https://www.w3.org," [Online]. Available: https://www.w3.org/2001/sw/. [Accessed 04 02 2018].

[2] T. B. Lee, Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its inventor, Harper, 1999.

[3] N. Guarino, "The Ontological Level: Revisiting 30 Years of Knowledge Representation," in Lecture Notes in Computer Science, vol 5600, Berlin, Heidelberg, Springer, 2009, pp. 52-67.

[4] "http://wiki.dbpedia.org," [Online]. Available: http://wiki.dbpedia.org/services-resources/ontology. [Accessed 01 02 2018].

[5] H. Cunningham, D. Maynard, K. Bontcheva and V. Tablan, "GATE- A Framework and Graphical Development Environment for Robust NLP Tools and Applications," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 2002.

[6] D. C. Wimalasuriya and D. Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," Journal of Information Science, vol. 36, no. 3, pp. 306 - 323, 2010.

[7] K. Nebhi, "Ontology-Based Information Extraction for French Newspaper Articles," in B. Glimm and A. Krueger (Eds.): KI 2012, LNCS 7526, Springer, 2012, p. 237–240.

[8] K. Nebhi, "Ontology-Based Information Extraction from Twitter," in Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, COLING 2012, Mumbai, December 2012.

[9] D. Maynard, Y. Li and W. Peters, "NLP Techniques for Term Extraction and Ontology Population," Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pp. 107-127, 2008.

[10] A. Dingli and S. Abela, "Using DBPedia to bootstrap new Linked Data," in SEMAPRO 2012 : The Sixth International Conference on Advances in Semantic Process, 2012.

[11] G. Fernando, D. Dejing, F. Stephen, W. Daya and Z. Hui, "A hybrid ontology-based information extraction system," Journal of Information Science, vol. 42 , no. 6, pp. 798-820, 2016.

[12] T. R. Gruber, "A Translation Approach to Portable Ontologies," Knowledge Acquisition, vol. 5, no. 2, p. 199–220, 1993.

[13] K. Vangelis, F. Pavlina, P. Georgios and I. Elias, "Ontology Based Information Extraction from Text," in Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, Vols. Lecture Notes in Computer Science, vol 6050. Springer, Berlin, Heidelberg, 2011, pp. 89-109.

[14] "http://wiki.dbpedia.org," [Online]. Available: http://wiki.dbpedia.org/projects/dbpedia-spotlight. [Accessed 02 02 2018].