A SUPERVISED APPROACH FOR SEMANTIC ANNOTATION OF ENTITIES IN TEXT

M. Shahzad Akram[1], Imran Sarwar Bajwa[2], M. Asif Naeem[3]
[1] Department of Computer Science& IT
Islamia University of Bahawalpur
[2] School of Computer Science
University of Birmingham, UK
[3] School of Computer and Mathematical Sciences
Auckland University of Technology, New Zealand
shahzad692@gmail.com, i.s.bajwa@cs.bham.ac.uk, mnaeem@aut.ac.nz

ABSTRACT: This paper presents a novel approach to perform the semantic annotation of entities in legal text. We analyze annotations in different domains and develop a new model for semantic annotation of entities in legal text. The presented approach is based on a framework that uses Markov Logic to classify and annotate entities in a piece of text with respect to its context and such effort significantly helps in semantic analysis of entities even in the ambiguous cases. In this paper, we also present of result of the experiments with the approach devised for semantic annotation by analyzing the input legal text by using typical natural language processing techniques, and employ the Markov Logic approach for semantic annotation. The results of the experiments show that the presented approach outperforms the other related approaches used for the similar tasks.

1. INTRODUCTION

Modern business organizations have to face challenges like analytics of big data, information extraction from unstructured data, searching legal text repositories for required information extraction. Since, a piece of legal text is typically builds with a set of entities or entity classes. However, ambiguity in interpretation of legal text is an open problem that makes machine processing and analysis of legal text a challenging task. Various organizations tend to use modern approaches like text mining, graph databases, semantic technology for solving complex data management problems. To find accurate results unstructured data needed to be managed in a better way for example systems like OntoText [1] tool. Various factors are involved in such systems like semantic annotation and semantic curation for the sake of semantic enrichment. Additionally, the semantic interpretation of a metadata can be translated into RDF (Resource Description Framework) format for standardization purposes and information interchange. While in typical semantic curation, various entities phrases and concepts are annotated with the help of tag. However, OntoText tool is particularly designed for life sciences and bio technology. The tool is efficient in identifying named entities and classification of the entities by using a semantic knowledge base to generate semantically enriched bio-medical data. In most of semantic annotation systems, the subject-predicated-object based triple annotation style is used [2].

During the literature review, it is found that a framework is required to annotate entity class in legal text with respect to their background information so that during semantic analysis of entities ambiguity may not affect accuracy of annotation of legal text. There is need of such semantic model for semantic annotation of entities in legal text is motivation of the presented research. In many domains semantic annotation has been already performed like textual data, audio data, and video data and in images. But no previous work has presented in semantic annotation of entities in legal text. An automated approach is required that helps in extracting information from unstructured data through text analysis and annotate the text.

1.1 Named Entity Recognition

Named Entity Recognition (NER) or Named Entity Detection is also called entity chunking and entity identification. It is a process that helps in identifying names in the input text and classifies it into pre-defined categories e.g. name of persons, name of cities, name of organizations, name of different locations, name of quantities, etc. In natural language Entity Name Detection is a used for information extraction. Named entity Detection is used for finding the specific things in the given text. Entity Name Detection is normally classified into different categories these categories are further divided in sub categories. Some Basic categories of Entity Name Detection are

Simple entity types are Person, organization, Location, Time expressions such as time and date, numeric expression such as money, etc. A few sub categories of EntityName Detection are Age, City, Country, State or Province, Weight**.** Entity Name Detection has many benefits some important are

- It is used to identify about the topic of the given.
- Documents are linked with each other on the basics of concepts that are defined within them.
- Entity Name Detection can find name of all the persons in a document.

A legal text is a statement that is different from regular text. Legal texts are used to create, modify, or terminate the different rights of any individuals or any organization. Every organization has its own legal text. In this paper, annotation of entities in a piece legal text is a main objective. However, the limitation of the presented approach is the ability to process only simple sentence that does not involve conjunctions (either...or, both...and, whether...or, etc.) in the input legal texts.

2. Related Work

The annotation is used to attach additional data to other set of data. Annotation is used to add notes, comments, external remarks, to a document without any change and modification to the document [3]. Annotations attachment is also downloaded when the original document is downloaded a user can also update its own annotation to the document. Annotation can be used in many domains [1] .Semantic annotation is a model that is used to find new methods for access new information and enhance the previous information. Lot of work has been done on semantic annotation and indexing [4] some fields are semantic web knowledge [5], knowledge management [6] semantic annotation methods and frameworks [4] .Many efforts have

been done to improve and enhance semantic information systems. These systems can provide automatic annotation in the knowledge base Systems and in the ontology's. Different tools like KIM [7] and many others have been used for the semantic annotation and indexing. A lot of work is also done on semantic web [8] and many applications like QBLS [9] and Trial Solution [10] have already been developed for this purpose. Semantic annotation can be applied on web pages, in databases, and in Text documents.

Many automatic annotation tools are developed that are used on large scale annotation. Lixto [6] is a tool which is used to convert unstructured web contents to structured web. SemTag [11] is another annotation tool which is used large scale annotation of web pages. There are a few tools available for automated semantic annotation first one is for manual annotation and second one is automatic annotation [6]. However, both tools deal with simple text and provide low accuracy for domain specific texts such as legal text, clinical/medical or business text. These tools are based on heuristic approaches and thus are not able to deal with ambiguous text.

3. Used Approach for Semantic Annotation

In this section, the approach used for semantic annotation of entities in Legal text is described in detail. The used approach performs the semantic annotation in three phases: analysis of the input legal text using NLP techniques, recognition of named entities in the Legal text, and finally annotation of the semantic chunks with the appropriate tags. On the basis of these three phases following three modules have been defined that perform distinct functions during the semantic annotation of entities in the legal text:

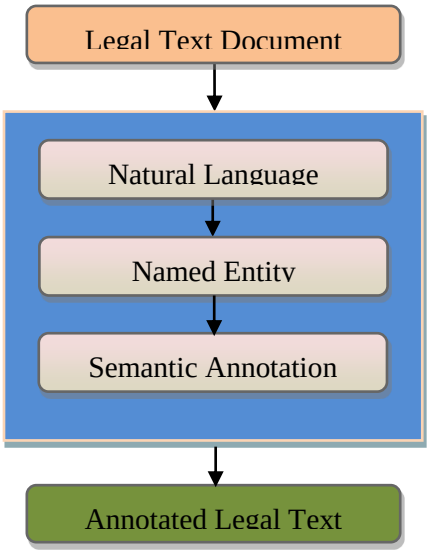

Figure 3.1: Framework of semantic annotation of entities in Legal Text

3.1 Natural Language Processing Module

In this phase, analysis of the input legal text is performed using the typical NLP techniques and libraries. The output of this module is distinct legal words in the form of an array that can be further processed.

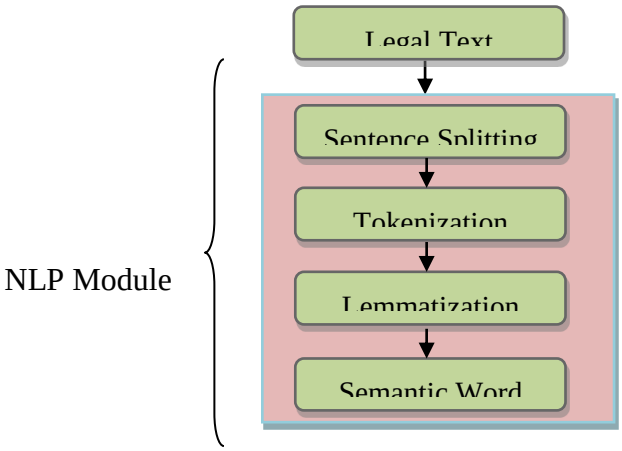

Figure 3.2: NLP module for pre-processing Legal Text

Sentence Splitting: Sentence Splitting is also called boundary disambiguation of sentences. It is an important task in processing of a piece of natural language text that how to decide from where a new sentence starts and where it ends. In the presented approach, sentence splitting is performed using the Stanford Parser [12].

Tokenization: Tokenization is a process in which a sentence is divided into small pieces as words, symbols, keywords or in other elements. Tokens are may be individual phrases, words or some time whole words. Here, for the sake of tokenization, the Stanford Parser [12].is used.

Lemmatization: In Lemmatization phase, a base form of a legal word or token is identified that is called a Lemma. In this phase, the words are converted into their base form in this step.

Semantic Word Grouping: Semantic Word Grouping is a set of different words that are combined into a group by meaning which refers to a specific subject.

3.2 Named Entity Recognition Module

In this phase, entity names are linked and further classified into a set of classes with respect to the domain of output of the previously used NLP module. The output of this module is a list of distinct entities with their respective classes that can be further processed.

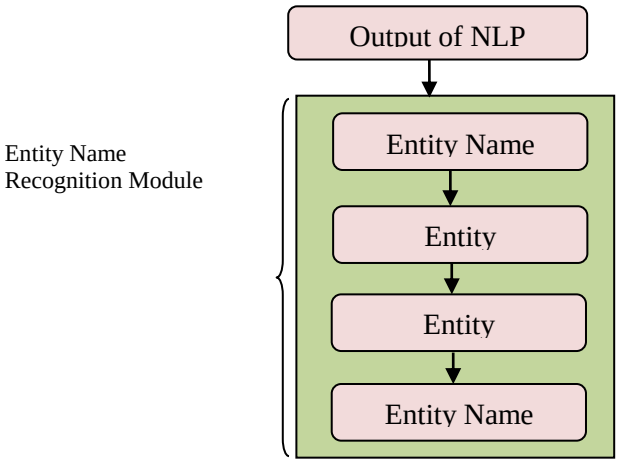

Figure3.3: Entity name recognition module

Entity Name Detection: For detection of named entities, a process is performed to identify proper names in the given text into pre defined categories for example name of persons, name of cities, name of organizations, name of different locations, name of quantities, etc.

Entity Classification: The process in which similar type of entities is grouped together is called entity classification**.** For classification of entities, Markov Logic represents the features of the input data in terms of *n*th joint distribution as shown in equation (1):

$$P(X=x) = \frac{1}{Z}\prod_{k}\phi_k(x_{\{k\}}) \quad (1)$$

Now, the joint distribution [13] of a model is mapped as a set of variables i.e. $X \in (X_1, X_2, \ldots, X_n)$. Typically, in a network of Markov Logic, a set of pairs (F_i, w_i) represent a predicate where a predicate in First Order Logic (FOL) is represented by F_i and a real number depicts w_i that is weight of the predicate/formula. In the used approach, the weights are updated by using equation (2) that is based on statistical relational learning approach and is incorporated by combining probability with the traditional first-order logic. Here, a typical MLN (Markov Logic Network) with a set of weights and formulas can be represented as below:

$$P(X=x) = \frac{1}{Z}\exp\left(\sum_{j} w_j f_j(x)\right) \quad (2)$$

The weights of the formulas are dynamically updated by using diagnolized Newton Method [14]. Here, the weight update formula is shown in equation (3):

$$w = w + D^{-1}g \quad (3)$$

Entity Disambiguation: The ambiguous entities not have clear meaning. Some words may have more than one meaning that depend upon the perspective of sentence. Entity Disambiguation is a process of removing the ambiguity from the ambiguous entities. To disambiguate the entities. Markov Logic approach [15] was used.

Entity Name Linking: Entity Name Linking is also called Named Entity Normalization. It is the process which is used of finding the identity of the entities in the given text. For linking entity names, Wikipedia based support [16] was used.

3.3 Semantic Annotation Module

In this phase, analysis of the input Legal Text is performed using the typical NLP techniques. The output of this module is distinct Business Rule words in the form of an array that can be further processed.

Identify Relationship: The process of looking of relations between the entities is known as relationship identification.

Semantic annotation: The annotation is used to attach additional data to other set of data. Annotation can be used in many domains. Lot of work has been made on semantic annotation some fields are semantic web knowledge, knowledge management.

4. EXPERIMENTS AND RESULTS

To test the performance of Markov Logic based approach for entity classification, a case study of legal text was taken from The Constitution of the Islamic Republic of Pakistan and is done by the framework described in the previous section. Following is one of the examples of legal text taken from the "The Constitution of the Islamic Republic of Pakistan".

Example 4.1.1 A person shall not be appointed a Governor unless he is qualified to be elected as member of the National Assembly and is not less than thirty five years of age.

Following is the output of the designed framework used for processing of Example 4.1.1 of the legal text case study. In this step from legal text 4.1.1 classified Entity Name will get as input and will find the Entity Linking between the classified Entity Name in the legal text

In above step classified Entities are linked with different pages.

In this step from legal text 4.1.1 Identified Relationship will get as input and Semantic Annotation will do on Identified Relationship in the legal text.

The overall results of case study of framework used for semantic annotation of the legal text example 4.1.1 by using the designed framework are shown in Table 4.2.

Besides this case study some other case studies (Table 4.4) were taken from legal documents of banks and universities. All these case studies were unseen. The solved case studies were of different lengths. The largest case study was composed of 209 words and 12 sentences. The smallest case study was composed of 69 words and 5 sentences. Calculated recall, and precision values of the solved case studies are shown in Table 4.3 and the results are visualized in Graph 4.1.

The average overall Recall value (81.93) and Precision value (85.32)is encouraging for initial experiments. We cannot compare our results to any other tool as no other tool is available that can classify and annotate entities in legal text. However, we can note that other language processing technologies, such as information extraction systems, and machine translation systems, have found commercial applications with precision and recall shown in Figure 4.1.

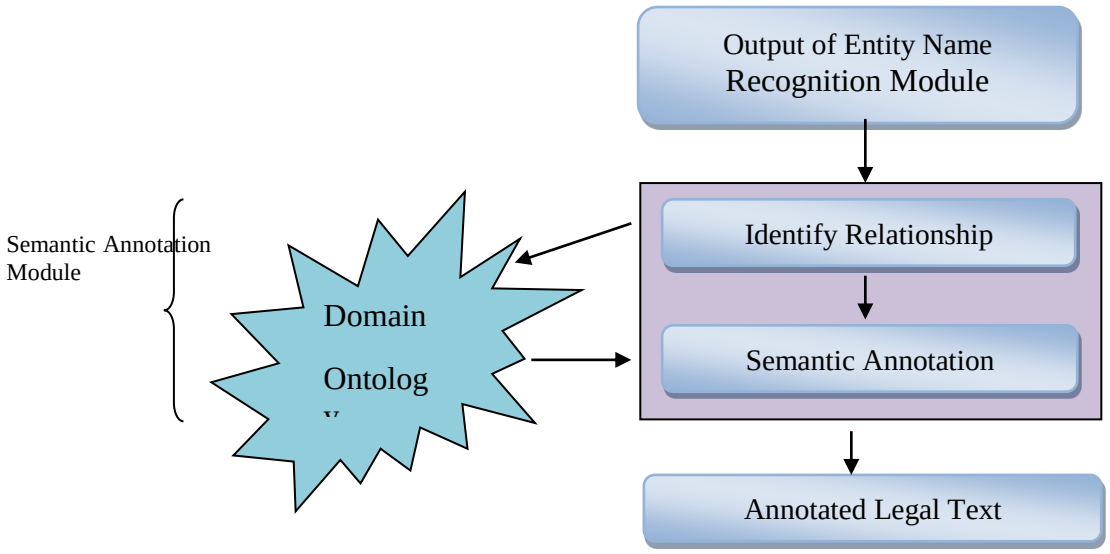

Figure 3.4: Semantic Annotation Module

Table 4.1. Entity Classification and linking of Rule 4.1.1

Entity Classification	A [person] $_{person}$ shall not be appointed a [Governor] $_{Leader}$unless [he]$_{person}$ is qualified to be elected as [member]$_{person}$ of the [National Assembly]$_{Organisation}$ and is not less than [thirty five] $_{Numeric}$ [years] $_{Time}$ of age.
Entity Linking	https://en.wikipedia.org/wiki/Governor_of_Punjab,_Pakistan A [person] $_{person}$ shall not be appointed a [Governor] $_{Leader}$ unless [he] $_{person}$ is qualified to be elected as [member] $_{person}$ of the [National Assembly] $_{Organization}$ and is not less than [thirty five] $_{Numeric}$ [years] $_{Time}$ of age. http://www.na.gov.pk/en/index.php

Table 4.2. Semantic Annotation of Rule 4.1.1

Input	[Person]		[Governor]
	[Member]		[National Assembly]
Semantic Annotation	[Person]	is	[Governor]
	[Member]	of	[National Assembly]

Table 4.3. Overall Results of case study of framework used for semantic annotation.

	Total Entities	Correct Entities	Missed Entities	Incorrect Entities	Recall	Precision
Entity Name Classification	21	18	1	2	85.71%	90.00%
Semantic Annotation	6	5	1	1	83.33%	83.33%

Table 4.4. Evaluation results of experiments

	Total Entities	Correct Entities	Missed Entities	Incorrect Entities	Recall	Precision
C 1	21	18	1	2	**85.71%**	**90.00%**
	6	5	1	1	**83.33%**	**83.33%**
C 2	33	26	3	4	**78.78%**	**86.66%**
	13	11	0	2	**84.61%**	**84.61%**
C 3	19	15	1	3	**78.94%**	**83.33%**
	7	6	0	1	**85.71%**	**85.71%**
C 4	28	24	3	1	**85.71%**	**88.88%**
	11	8	1	2	**72.72%**	**80.00%**
				Average	**81.93**	**85.32**

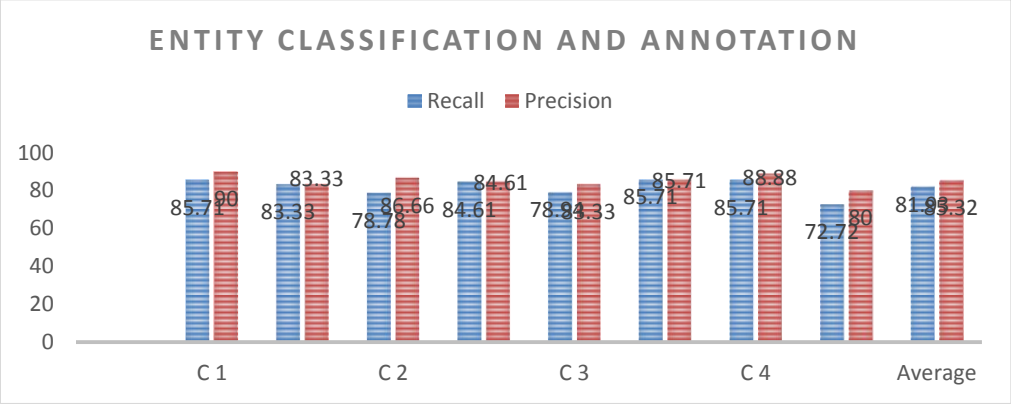

Figure 4.1.The results of the experiments in Recall and precision

5. CONCLUSION AND FUTURE WORK

In this paper,a framework is presented to annotate entity classes in legal texts. The presented approach is based on Markov Logic approach for classification of the entities found from legal text. The present approach was implemented in Java and Java based libraries such as Stanford Core NLP, Stanford NER, etc. Then, to test the performance of the designed approach four case studies were taken from different domain such as legal documents of government, banks, universities, etc. The results of the experiments are also encouraging.

The presented approach for Semantic annotation, takes text document as an input and performs different process on it. The objective of this research is semantic annotation of entities in legal texts. Our future research plans is that it will increase the accuracy of this tool. Our future research will go through Semantic annotation in other legal texts of complex sentences, as well.

6. REFERENCES

[1] Oren, E., Delbru, R., & Decker, S. (2006). Extending faceted navigation for RDF data. In The Semantic Web-ISWC 2006 (pp. 559-572). Springer Berlin Heidelberg.

[2] Sintek, M., & Decker, S. (2001, October). TRIPLE-An RDF Query, Inference, and Transformation Language. In INAP (pp. 47-56).

[3] Schroeter, R., Hunter, J., & Kosovic, D. "Vannotea: A collaborative video indexing, annotation and discussion system for broadband networks", Knowledge Capture, pp. 1-8, 2003.

[4] Popov, B, K, A et.al. "KIM–semantic annotation platform The Semantic Web-ISWC 2003 (pp. 834-849): Springer.

[5] McGlothlin, J. P., Khan, L., & Thuraisingham, B. M. (2011, June). RDFKB: a semantic web knowledge base. In IJCAI Proceedings-International Joint Conference on Artificial Intelligence (Vol. 22, No. 3, p. 2830).

[6] Uren, V. & Cimiano,P. "Semantic annotation for knowledge management: Requirements and a survey of the state of the art", Web Semantics: science, services and agents on the World Wide Web, 4(1), 14-28, 2006.

[7] Kiryakov & Atanas et.al. "Semantic annotation, indexing, and retrieval", Web Semantics: Science, Services and Agents on the World Wide Web, 2(1), 49-79, 2004

[8] Melis, E., Andres, E., Budenbender, J.,et.al. (2001). ActiveMath: A generic and adaptive web-based learning environment. International Journal of Artificial Intelligence in Education (IJAIED), 12, 385-407.

[9] Dehors, S., & Faron-Zucker, C. (2006). Qbls: A semantic web based learning system. In World Conference on Educational Multimedia, Hypermedia and Telecommunications (Vol. 2006, No. 1, pp. 2795-2802).

[10] Xu, Changsheng, Jinjun Wang, et.al. "A novel framework for semantic annotation and personalized retrieval of sports video." Multimedia, IEEE Transactions on 10, no. 3 (2008): 421-436.

[11] Reeve, L., & Han, H. (2005, March). Survey of semantic annotation platforms. In Proceedings of the 2005 ACM symposium on Applied computing (pp. 1634-1638). ACM.

[12] De Marneffe, M. C., & Manning, C. D. (2008, August). The Stanford typed dependencies representation. In Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation (pp. 1-8). Association for Computational Linguistics.

[13] Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible reasoning.

[14] Lawless, A. S., Nichols, N. K., Boess, C., &Bunse-Gerstner, A. "Approximate Gauss–Newton methods for optimal state estimation using reduced- order models",International journal for numerical methods in fluids, 56(8), 1367-1373, 2008.

[15] Richardson, Matthew, and Pedro Domingos. "Markov logic networks."Machine learning 62, no. 1-2 (2006): 107-136.

[16] Hachey, B., Radford, W., Nothman, J., Honnibal, M., & Curran, J. R. (2013). Evaluating entity linking with Wikipedia. Artificial intelligence, 194, 130-150.