A Survey Paper on in Text Citation of Manuscript with Computational Linguistic Tools

[1]Devendra Das, [2]Shika Pandey
Research Scholar, M-Tech (Software Engineering) , Assistant Professor
Computer Science & Engineering
Chhattisgarh India

Abstract— **Due to advancements in the field of computer technology and due to the widespread use of Internet and increasing availability of resources the authenticity of the contents is a major issue. With the availability of the information on WWW and digital libraries, the authenticity of the content became one of the most important issues for universities, schools and researcher's fields. To check the authenticity, various algorithms have been done earlier. The need of citations is extremely helpful to anyone who wants to find out more about your ideas and where they came from not all sources are good or right your own ideas may often be more accurate or interesting than those of your sources. In this paper we are going to survey and list the advantages and disadvantages of the latest and the important effective methods used or developed for reference citations in a manuscript. Mainly methods used are grammar based methods such as dependency parsing, grammar representations.**

KEYWORDS: In_ text-citations, Type dependencies, Parsing, Tokenizer.

I. INTRODUCTION

The aim of this paper is to study and analyze how? To cite the spots in manuscripts from its references .Also to study the description of the grammatical relationships in a sentence that can easily be understood and effectively used by people without linguistic expertise who want to extract textual relations [1]. In text Citation analysis is a new technique which is used to compute quantitatively the value of document through arranging the citations in some kind of category or order· It is also used to study and analyze the authenticity of research work is done on a particular subject or topic. "Citation analysis" refers to references in one text to another text with information on where that text can be found˙ Citation analysis reflects on citation practices˙

More recent studies have shown, however, that this assumption is oversimplified. The reason or motivation for a citation matters

II. NEED OF CITATIONS IN MANUSCRIPT

- To snip and pass off (the ideas or words of another) as one's own.
- To use without accepting the source.
- To constraint literary theft.
- To present as new and original an idea or product derived from an existing source.
- Turning in somebody else's work as your own.
- Copying words or ideas from someone else without giving credit**[3]**.
- Failing to put a citation in quotation marks.
- .Giving incorrect information about the source of a quotation.
- Changing words but copying the sentence structure of a source without giving credit
- .Copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not.

III. DEFINITION OF CITATION, COMPUTATIONAL LINGUISTIC TOOLS

Citation is a way by which we tell the readers that certain materials in your work came from another source [2]. It also gives the essential information to find that source again.

A reference citation is the documentation needed to make your paper acceptable for academic purposes.

Computational linguistics is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective.

Development and need of Citation Analysis.

The development of citation analysis has been marked by the invention of new techniques and measures, the exploitation of new tools, and the study of different units of analysis [3].

1. Citation of a document implies use of that document by the citing author. This assumption actually has two parts: the author refers to all, or at least to the most important, documents used in the preparation of his work; and all documents listed were indeed used, i.e., the author refers to a document only if that document has contrib -butted to his work.
2. Citation of a document (author, journal, etc.) reflects the merit of that document. The underlying assumption in the use of citation counts as quality indicators is that there is a high positive correlation between the number of

citations which a particular document (author, journal, etc.) receives and the quality of that document (author, journal, etc.).

3. A cited document is related in content to the citing document; if two documents are bibliographically coupled, they are related in con- tent; and if two documents are citied, they are related in content.

IV. METHODOLOGY

In previous papers, we have studied that several methods are used:

Grammar-based method

The grammar-based method is one of the important techniques used for citing of contents of manuscripts in systematic way [1]. It focuses on the grammatical structure of documents, and this method uses a string-based matching approach to detect and to measure similarity between the documents. The grammar-based methods is suitable for detecting exact copy without any modification, but it's not suitable for detecting modified copied text by rewriting or switching some words that has the same meaning.

Type Dependency method

The typed dependencies are a way of representation of sentences and paragraphs [1]. It was designed by Stanford and they provided a simple description of the grammatical relationships in sentence. These relationships were easy to understand and effectively used by people without expertise knowledge to extract textual relationships. It displays the phrase representations that have long conquered in the computational linguistic community. It actually represents all sentence relationships consistently as typed dependency relations. It also represents as triples of a relation between pairs of words that may try to fill in next to the graphic .The dependencies map directly to a directed graph representation, in which words available with the sentence are nodes in the graph and grammatical relations are edge labels.

Here is an example sentence:

The dependencies map directly to a directed graph representation, in which words available with the sentence are nodes in the graph and grammatical relations are edge labels [1]. The dependency graph is shown in the figure.

.

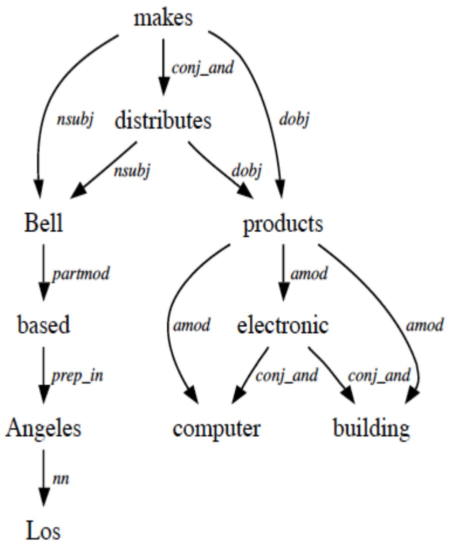

Figure1: A Dependency graph

STYLES OF DEPENDENCY REPRESENTATION

- A tree structure is used as basic typed dependencies which are used in the dependencies for identifying relations. That is, there are no crossing dependencies .They is also known as a projective dependency structure. Each word in the sentence (except the head of the sentence) is the dependent of one other word.
  For the sentence, “Bell, a company which is based in LA, makes and distributes computer products.", the basic typed dependencies will be:
  nsubj(makes-11, Bell-1)
  det(company-4, a-3)

- Another method is the collapsed representation [1]. In this method, dependencies involving are prepositions, conjuncts, as well as information about the referent of relative clauses are collapsed to get direct dependencies between content words.

This “collapsing" is often useful in simplifying patterns in relation extraction applications. For instance, the dependencies involving the preposition “in" in the above example will be collapsed into one single relation:

prep (based-7, in-8)
pub (in-8, LA-9)
Will become
Prep in (based-7, LA-9)

Parsing may be defined as a syntactic analysis [8]. It is the process of analyzing a string of symbols, in natural language, following to the rules of a formal grammar. The term parsing actually originated from Latin pars (orations), meaning part (of speech)[4]. In natural language, a parser is a program that works out the grammatical structure of sentences and it groups the words to go together (as "phrases") and which words are the subject or object of a verb. Probabilistic parsers use knowledge of language gained from hand-parsed sentences to try to produce the most likely analysis of new sentences [5]. These statistical parsers are not perfect and can still make some mistakes, but are commonly used. The development of parser was one of the biggest breakthroughs in natural language processing. In parsing there are two ways to describe a sentence structure in natural language.

- It breaks up the sentence into constituent’s i.e. phrases, which are then broken into smaller constituents as shown in figure.
- Then it draws links connecting individual words. These are called constituency grammar and dependency grammar respectively.

Sentence to be parsed
“Devendra das is an mtech scholar in software engineering”
Tagging
Devendra/NNP das/NNP is/VBZ and/DT mtech/JJ scholar/NN in/IN software/NN engineering/NN. /.
Parse.

```
      (ROOT
        (S
          (NP (NNP Devendra) (NNP das))
          (VP (VBZ is)
            (NP
              (NP (DT an) (JJ mtech) (NN
      scholar))
              (PP (IN in)
                (NP (NN software) (NN
      engineering)))))
   (. .)))
```

Typed dependencies

```
nn(das-2, Devendra-1)
  nsubj(scholar-6, das-2)
  cop(scholar-6, is-3)
  det(scholar-6, an-4)
  amod(scholar-6, mtech-5)
  root(ROOT-0, scholar-6)
  prep(scholar-6, in-7)
  nn(engineering-9, software-8)
  pobj(in-7, engineering-9)
  Typed dependencies, collapsed
  nn(das-2, Devendra-1)
  nsubj(scholar-6, das-2)
  cop(scholar-6, is-3)
  det(scholar-6, an-4)
  amod(scholar-6, mtech-5)
  root(ROOT-0, scholar-6)
  nn(engineering-9, software-8)
  prep_in(scholar-6, engineering-9)
```

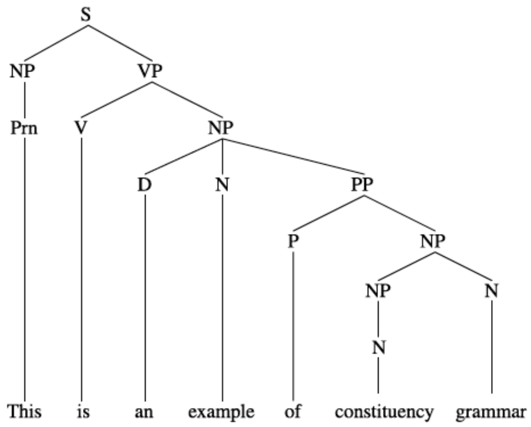

Figure2: A Constituency grammar

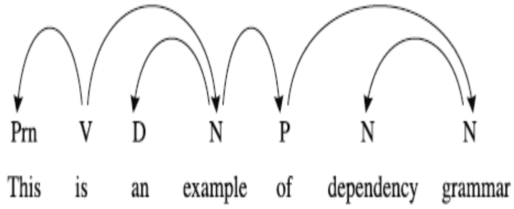

Figure3: A dependency grammar

Dependency trees representations in words.
Whenever two words are connected with the help of a dependency relation, then either it is the head and the otherwise the dependent, and that there is a link connecting them [5]. The dependent is the modifier, object, or complement; the head plays the larger role in determining the behavior of the pair. The dependent assumes the presence of the head; the head may require the presence of the dependent.

Tokenization is the process of replacing thoughtful data with unique identification symbols that retain all the essential information about the data without compromising its security[8]. A simple information extraction system. Tokenization begins by processing a document in the following ways.

The raw text of the document is divided into sentences by using a sentence segmented.
Then each sentence is then further subdivided into words using a tokenizer.
Next, each sentence is tagged with part-of-speech tags as shown in figure[8].

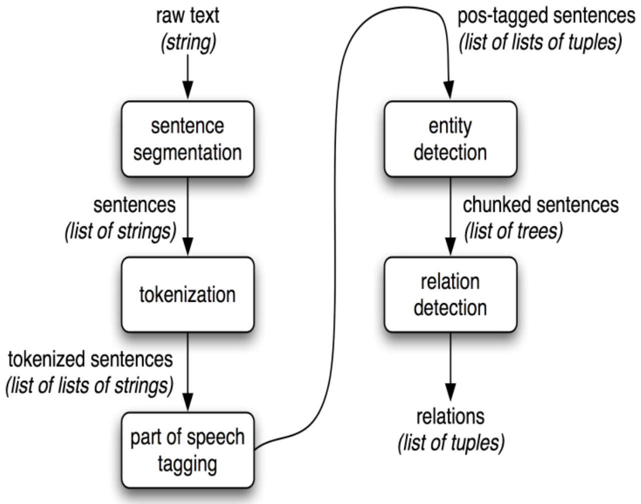

Figure4: An Example of Tokenizer

V. CONCLUSION AND FUTURE WORK

In this study the citation detection was considered as it is one of the most publicized forms of text reuse and check the authencity of the contents around us today. In this paper different methods and techniques that are needed for in text citations. In particular, it has been shown in this study how the citations can be handled using different techniques and tools. However, there are still some weaknesses and shortages in these techniques and tools which will affect the success of citation detection significantly

The proposed work is about performing the valuable in text citation between both the manuscript and the candidate document. Following is the methodology-

- Take the manuscript document as input and extract their keywords and contents.
- Split the documents in paragraph format or in section wise.
- Generate the dependency grammar for each sentence.

VI. REFERENCES

[1] Marie-Catherine de Marne, Christopher D. Manning, “Stanford typed dependencies manual l Software, September 2008, Stanford Parser v. 3.3.

[2] Dr. K.Kumar, T.Raghunadha Reddy,” Citation Analysis of dissertations submitted to the department of library and information science “,venkateswara university,triputi , International Journal of Digital Library Services,vol 1.4,april-june 2012,IssN:2250.1142**.**

[3] Xiaozhong Liu,Jinsong Zhang,Chun Guo” Full-Text Citation Analysis: A New Method to Enhance Scholarly Network”,School of Library and Information Science, Indiana University Bloomington.

[4] Michael A. Covington,” A Fundamental Algorithm for Dependency Parsing”, Artificial Intelligence Center The University of Georgia Athens, GA 30602-7415 U.S.A...

[5] Michael A. Covington,” A Free-Word-Order Dependency Parser in Prolog”, Artificial Intelligence Center The University of Georgia Athens, GA 30602-7415 U.S.A

[6] Michael A. Covington,” ET: an Efficient Tokenizer in ISO Prolog”, Artificial Intelligence Center the University of Georgia Athens, Georgia 30602-7415 U.S.A.

[7] Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, Vaclav Snase,” Survey of Plagiarism Detection Methods”, Department of Computer Science,2011 Fifth Asia Modelling Symposium.

[8] Michael A. Covington,"Important Additional Notes about Dependency Parsing",April 15, 2004,Artificial Intelligence Center The University of Georgia Athens, Georgia 30602-7415