Dyanmic Entity and Relationship Extraction from News Articles

By

Mazhar Ul Haq

2010-NUST-MS-PhD IT-10

Supervisor

Dr. Ali Mustafa Qamar

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Information Technology (MS IT)

In

School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(June 2013)

Approval

It is certified that the contents and form of the thesis entitled “**Dyanmic Entity and Relationship Extraction from News Articles**” submitted by **Mazhar Ul Haq** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Ali Mustafa Qamar**
Signature: __________________

Date: __________________

Committee Member 1: **Dr Usman Ilyas**

Signature: __________________
Date: __________________

Committee Member 2: **Dr Sarah Shafiq**

Signature: __________________
Date: __________________

Committee Member 3: **Dr Amir Hayat Khan**

Signature: __________________
Date: __________________

Abstract

This thesis summarizes the idea of automatic extraction of the relationship of news articles, and will operate the system line key to reputation management system on the Internet. By an automated system to extract the relationship, we mean a complete system that takes raw news articles as input and returns the entities with their relationships. Echo time in the whole system is very critical because this system will be run from 24 7 365 days a year and will be equipped with millions of news articles per day; Therefore it is very important that the system should respond intelligently against each news item.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Mazhar Ul Haq**
Signature: ________________

Acknowledgments

Firstly I would like to thank Allah Almighty (SWT) for giving me strength and courage to successfully conduct this research. His constant support has enabled me to successfully complete all milestones of this journey.

I offer sincere gratitude to my supervisor Dr. Ali Mustafa Qamar, who provided me a great deal of his knowledge and valuable expertise in this research domain. I have to appreciate his full support, motivation and kind guidance that provided me the possibility to complete this research.

I greatly appreciate the help of Dr Ali Mustafa Qamar in refining major goals of my research. He provided me exemplary guidance, constant encouragement and mentorship to develop a valuable skill set for conducting this research. I am grateful to him for this cooperation during the period of this research. I wish to extend my thanks to my committee members Dr Usman Ilyas and Dr Sarah Shafiq for their kind support and ideas to improve my research goals.

I wish to thank my lovely parents for their complete support who encouraged me to go my own way. It has been a long journey, but their constant support and love helped me to keep on going, even things got difficult. A kind thanks to my brothers, sister and fiance whose motivation and support enabled me to successfully accomplish this goal.

This thesis is dedicated to my lovely parents

For their endless love ,encouragement and support

”We cannot direct the wind but we can adjust the sails.”
Dolly Parton

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In todays fast competitive world every organization or public personality ultimately lean on its repute for success and survival and it could be their biggest credit to make them stand out of the crowed and perceive their good reputation to the public. They should know what is being reported about them and how the community to which they used to deal, have think about them and what their competitors used to do. This is what we called Personal relationship. PR helps the organization or public personality to model the way they work to achieve success. To deal with these types of obstacles, a number of methodologies have been proposed in which a couple of techniques need manual interaction. A system has been proposed in this work to extract entities and finds relationships in them unsupervised way. The system is based on news content monitoring to find the relationship between two entities and proportion of that relationship. So, it can play a role of policy maker by finding out the substantiation of policies.
In Addition to PR analysis this system is capable to help people, for finding out what is total coverage of particular topic, personality, organization in particular area. That helps new entrants to find out how much effort are needed and what are the strength, weaknesses of existing players, what are the current opportunities that they can exploit and what are the major threat to them. This system helps PR analyst to avoid following:

1. Finding Accuracy of the Messages: Mostly for PR activities organizations hire other people to perform analysis for them and the major problem with this activity is the Accuracy of Message. In order to understand the value of message accuracy in a public relations measurement program, it is important to understand the elements that need to be taken into consideration when conducting this type of analysis. Message accuracy is based on an analysis of four basic elements:

the inclusion of basic facts, the inclusion of misstatements about these basic facts, the inclusion of incomplete, deceptive or misleading information that biases the interpretation about basic facts and the omission of basic facts. In Accuracy following are the major areas that generally Impact the PR analyst

   (a) Basic Facts: Basic facts are the fundamental information that is central to any communications program. These facts can be such fundamental information as a definition or description of the product or service. They can also include statements, opinions or points-of-view that can be supported and documented. Examples of opinions or points-of-view that can be considered basic facts are statements about relative value or comparative information that is used to place information in context.

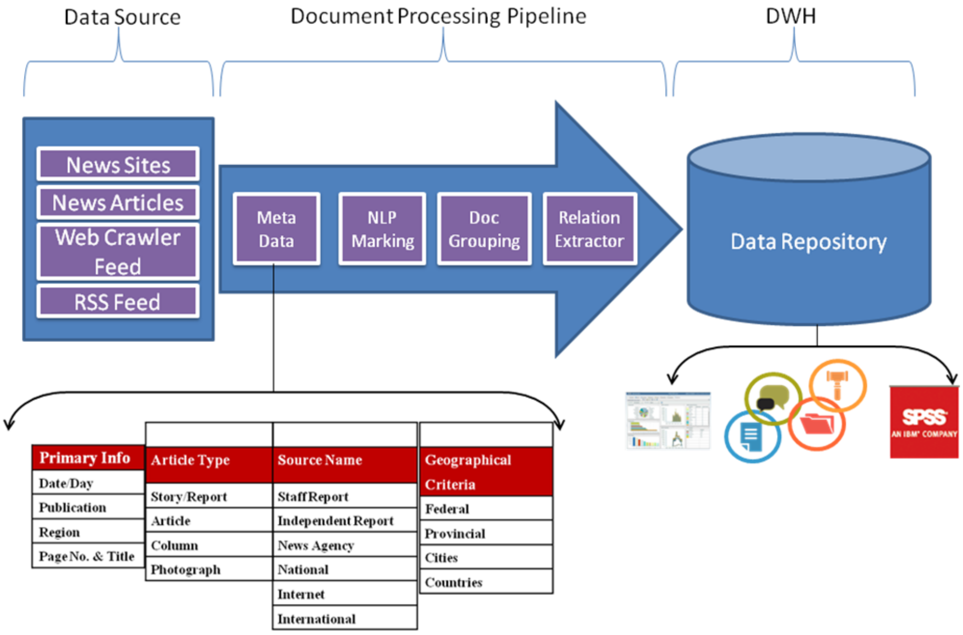
   (b) Misstatements: Misstatements are generally understood as errors or incorrect information included in an article or publication. Misstatements typically result from incorrect data but can also include unsubstantiated opinions or points-of-view from a reporter or interviewee that states a falsehood.

   (c) Incomplete Information: Incomplete information is a statement, opinion or point-of-view that selectively includes some information, but excludes other relevant facts. These apparently accurate statements create a misleading impression or a deception about a product or service and while factually accurate, are in actuality erroneous.

   (d) Omission: Omissions are the absence of key information that should be included in a specific article or publication. Not all basic facts can be considered omissions if they are not included in an article or publication. The key to understanding omissions is in the context of the article. The focus or subject matter of the story has to be relevant to the specific omission and the story or article will be considered incomplete unless that basic fact is included.

2. Not linking Messages to Communication Objective: The other challenge involved in conducting effective content analysis is linking communications objectives with the actual message analysis. Typically communications objectives are directly related to the information needs dictated by a communications life cycle.

In order to model a system that can cater the above needs, we need to have a system which can gather information from different news sites accumulate that information and present it in the form which fulfils these needs as explained in bellow diagram.



Primary Info	Article Type	Source Name	Geographical Criteria
Date/Day	Story/Report	StaffReport	Federal
Publication	Article	Independent Report	Provincial
Region	Column	News Agency	Cities
Page No. & Title	Photograph	National	Countries
		Internet	
		International	

Figure 1.1: End to End System Flow

We have proposed a system which takes document from different WebCrawlers , RSS Feed sites and News sites perform processing on document by marking entities in them, grouping related documents based on the news titles and entities present in them, then model the relationship among those entities which are reported in that particular document group. After extraction of entities the xml format of document is pushed into Data warehouse on which we can use different information modeling tools (like SPSS, Cognos or Hyperion). These reporting tools will in turn help content monitoring people to make their analysis on the information which is Accurate, with no omission, no misstatement and based on complete information. This way of content analysis is also very preferable because it removes all the personal biasness which is common threat when we perform content analysis manually.

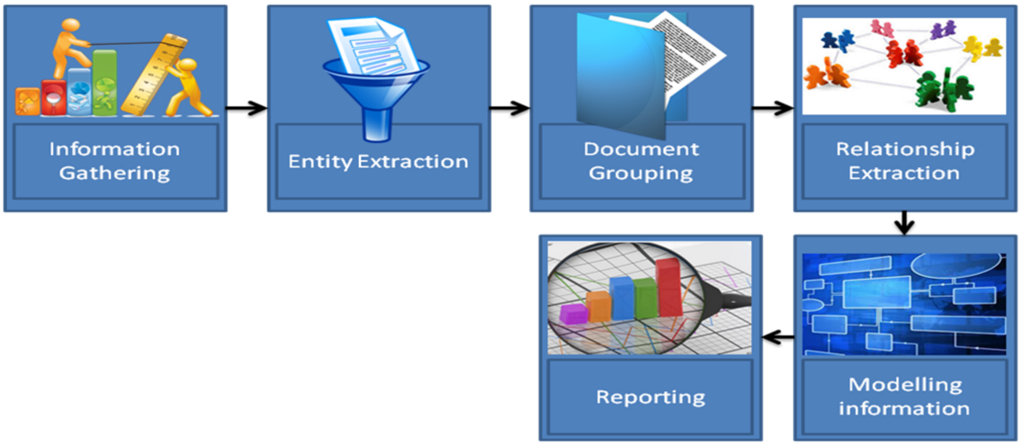


Figure 1.2: Step by Step System Flow

1.1 Information Gathering

In Order to run a system which extracts information (relationship between entities) from news articles, we first need to have news articles. For this research the information gathering part is not in the scope as this research is solely focus on relationship extraction. But for understanding purpose we need to know how news articles will come to the system and what will be their format before entering into document processing pipeline. Documents are gathered by in-house Web Crawlers (in-house because we need transformation on document as per the needs of Document processing pipeline), RSS Feed from different news sites and already gathered information.
Web-Crawlers provide basic source of document for this research. Web-Crawlers are computer programs that surf the World Wide Web for particular documents and downloads them into specified directory. Similarly RSS Feed Reader gets news articles from different news sites and pushes the raw documents into common location. From this common location document transformer reads the raw document gathered from different news site and transforms it into XML formatted document and attaches the formatted meta-data information with each document the overall flow for information gather will be explained in the following diagram

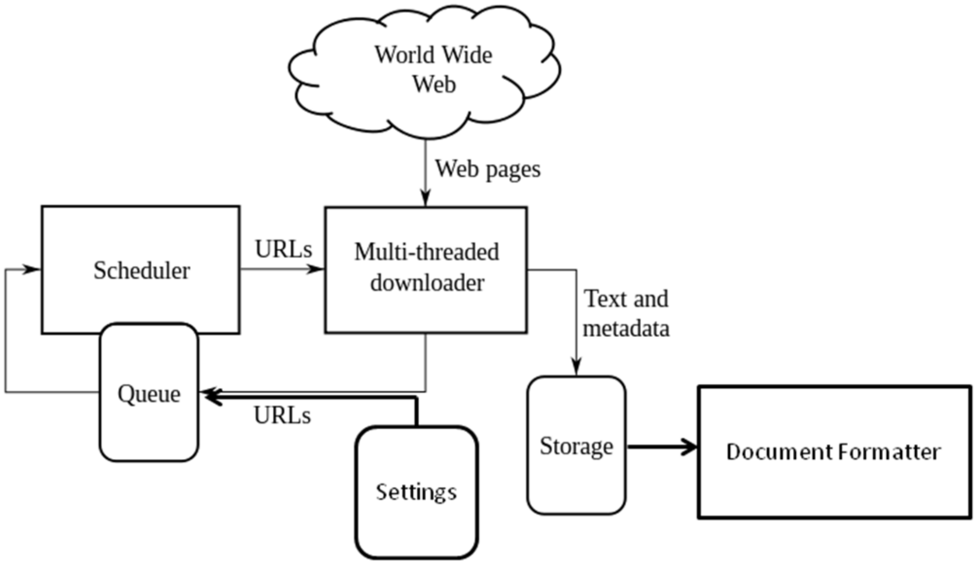


Figure 1.3: Step by Step Information Gathering process

1.2 Entity Extraction

The first step of this automatic analysis includes text pre-processing. First the stream of text is segmented into sentences. Then a named entity tagger is used to specify the named entities. Once Document is made ready from Data Gathering phase it is passed to NLP marking phase which basically marks Part Speech Tagging in each sentence and performs shallow parsing on each sentence to be summarized. After POS marking and shallow parsing we try to identify the entities in the document.

1. First, we used to break up the document into parts of speech and apply tag over each part.

2. After tagging, we pick up the nouns based on the following given criteria:

   (a) Case: we used three features to come up with this for the Case of noun. Significant one is to find whether it starts with a capital letter. The second parameter is whether all letters of the entity name are upper-case or not (usually organization names are all in upper-case) and third is whether it is a mixed case word or not.

   (b) Punctuation: Whether some punctuation is used in the word or not for example hyphen, apostrophe or & sign.

   (c) Digits: If digits are used in a word or not for example 3M or W3C etc.

   (d) Prefixes or suffixes: Whether some prefixes/suffixes are used for example Mr. or Ms.

3. Apply filters to extracted noun and tear off all the nouns which have weights below our threshold.

4. Frequency and percentage of selected noun will also be kept in consideration.

In this step we add some more tags to XML documents like POS tags for Headlines, POS Tags for body part. Meta data and these Tags along with scoring helps us in next step to group related documents.

1.3 Document Grouping

In order to move forward we first have to group the similar event reporting documents together so that we can extract the correct relationship between two Entities. Here documents are grouped based on the candidate Entities list which has been generated by the previous step. Document clustering or grouping has many uses in natural language tools and applications. For example, summarizing sets of documents that all describe the same event requires first identifying and grouping those documents talking about the same event. Document clustering involves dividing a set of texts into non-overlapping clusters, where documents in a cluster are more similar to one another than to documents in other clusters. The term more similar, when applied to clustered documents, usually means closer by some measure of proximity or similarity.
Following are the main steps that are been followed for document grouping:

1. First meta data information is used to identify possible groups for particular document and it select some candidate groups for each document

2. Then we perform Named Entity matching process to find out the similarity between two documents, for matching Named Entity set we basically use two techniques one is TF/IDF matching and second is soundex based matching

3. If Named entity matching process was successful then and document is matched to particular group more than threshold it is assigned to that group
4. The final entity set for whole group is also maintained and it is updated based on the matching results of each document to that group
5. After the grouping is done the groups of documents along with their matrices are passed to next phase which is relationship extraction.

The document grouping step is very important step before entering into the relationship extraction, because it gathers the relevant document related to one news headline and helps system to identify relationship which is non-biased. This is how we remove the biasness of relationship between two entities and then based on the relations extracted we can also perform point-of-view analysis on each news headline for different entities and relationships. This Step helps us to perform following analysis once the information is modelled in the data warehouse:

1. Point-of-view analysis from different news agencies using different modelling techniques on data warehouse
2. Helps us to find-out which parties has reported particular entities and with what frequencies
3. Helps us to find-out which are general head on which particular entities are reported
4. What are the major group of analysis for particular entity (brand, person or company)
5. Different new agencies point-of-views related to particular entity

So therefore we can say document grouping is a crucial component of whole document processing pipeline it helps us in many ways and assist us in relationship extraction. Now the question is why this way of grouping document, the answer to this question is our system is anticipated with heavy loads we cannot deploy some sophisticated techniques for grouping the document, if we do so then the time window to perform analysis and to report to actual entities will be very less and will make overall system less effective.

1.4 Dynamic Relationship Extraction

Relationship between entities that are been reported in some news article is very important it helps us in content analysis process, it helps us to find-out in which way two different entities are related to each other. For example Imran Khan is one entity and he wants to visit to Islamabad for some political reason. If we have a system that has modeled the information related to Imran Khan previous visit to Islamabad will be very helpful. It will help Imran khan by giving information that what are his strong areas related to Islamabad that he can utilize to address people for political reasons. What his other competitor do in Islamabad related to politics. And if by some how we can tag some entities with some topics like problem area for Islamabad we can tell Imran khan that on these problem areas these are the best stances that if he take will help him in his cause. In the similar fashion this system can help brands, companies and celebrities to take public stance related to some issue or related to some particular entities.
Now in order to model the information which is captured in previous steps of document processing pipeline, we need to find out relationship between different entities that are present in the groups identified in the grouping process. So from groups we use redundant information about the same event or personality taken from different sources is mainly used for extract relationship. As the source media report the same news by different context so, this context is really beneficial to extract relationship. Here a list is organized for each document based on entity-connecting phrase-entity i.e. L1= E1 verb E2, L2 = E1 perp E2 etc. For same event there will be more than one list sets. Than these documents will be used to build context set and each one of the set have all the contexts for same two entities.

In this thesis we will cover all these steps and process how we have modeled them what were the previous studies related to these step. We have perform rigorous testing for this thesis that we also cover in testing part. So in general the background study part will contains knowledge of manual way of content analysis, what is the previous research in the areas that are covered in this thesis. Then we come to Problem Statement part which covers what is the problem statement for this thesis what are the assumptions for this thesis what processes are the part of this thesis and what will not be covered in this thesis. Then we have NLP chapter which covers the details of Natural language process and we have use NLP in our thesis to find out relationship between entities dynamically. Then we have a chapter related to methodology which covers detail methodology that is been followed for this thesis which includes entity extraction, document grouping and relationship extraction. Then we have chapter related to testing and finally we will conclude our thesis along with the reference part

Chapter 2

Content Analysis

Content Analysis is described as the scientific study of content of communication. It is the study of the content with reference to the meanings, contexts and intentions contained in messages. The term Content Analysis is 75 years old, and Websters Dictionary of English language listed it since 1961.
In 1952, Bernard Berelson published Content analysis in Communication Research, which heralded recognition for the technique as a versatile tool for social science and media researchers. Some scholars adopted it for historical and political research as well [11]. However, the method achieved greater popularity among social science scholars as well as a method of communication research [12]. The development of content analysis as a full-fledged scientific method took place during World War II when the U.S. government sponsored a project under the directorship of Harold Lasswell to evaluate enemy propaganda. The resources made available for research and the methodological advances made in the context of the problems studied under the project contributed significantly to the emergence of the methodology in content analysis. One of the out comes of the project, the book entitled Language of Politics published in 1940s

2.1 Definition and purpose of Content analysis

Content denotes what is contained and content analysis is the analysis of what is contained in a message. Broadly content analysis may be seen as a method where the content of the message forms the basis for drawing inferences and conclusions about the content. Further, content analysis falls in the interface of observation and document analysis. It is defined as a method of observation in the sense that instead of asking people to respond

to questions, it ”takes the communications that people have produced and asks questions of communications”. Therefore, it is also considered as an unobtrusive or non-reactive method of social research.
A number of definitions of content analysis are available. According to some Authors (like Berelson (1952)) content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication. Some says that it is any technique for making inferences by systematically and objectively identifying specified characteristics of messages. Kerlinger (1986) defined content analysis as a method of studying and analysing communication in a systematic, objective, and quantitative manner for the purpose of measuring variables.

Further, like any other research method, content analysis conforms to three basic principles of scientific method. They are:

1. Objectivity: Which means that the analysis is pursued on the basis of explicit rules, which enable different researchers to obtain the same results from the same documents or messages.

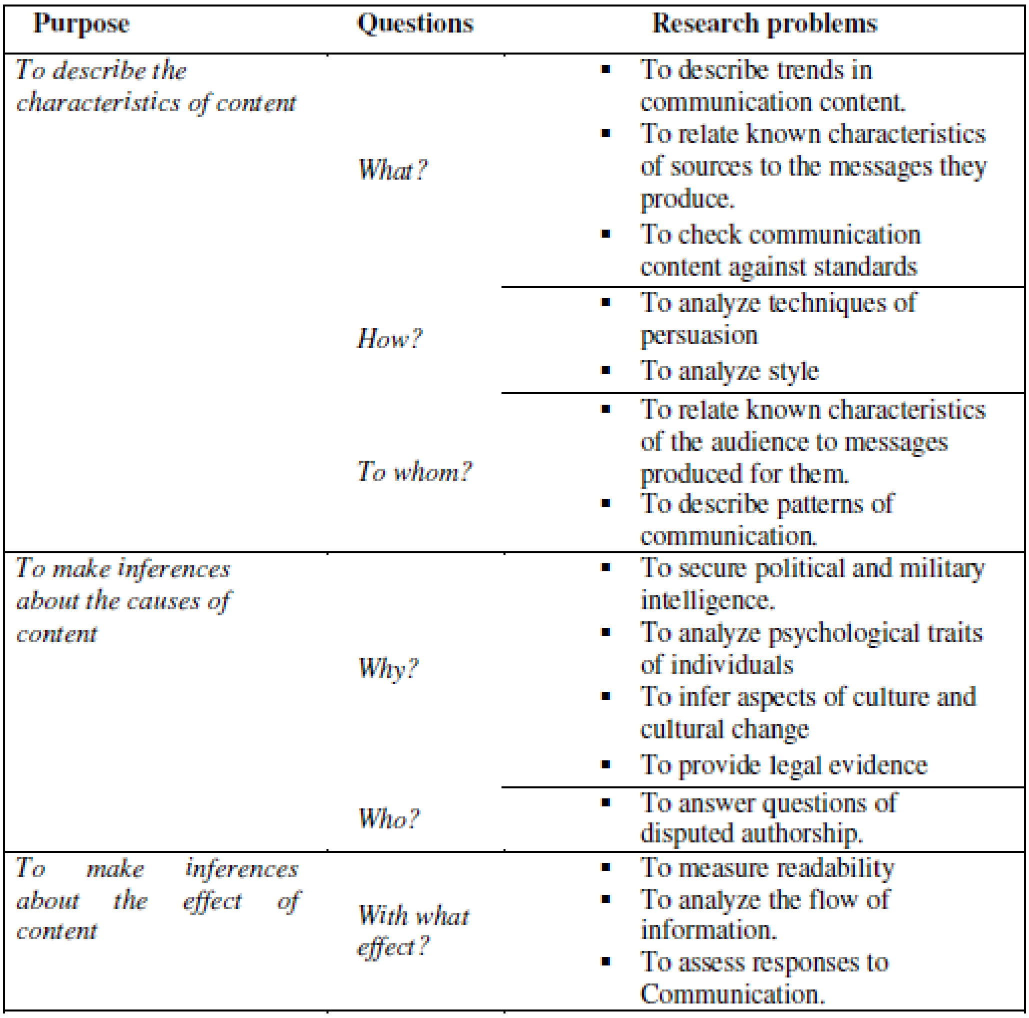
2. Systematic: The inclusion or exclusion of content is done according to some consistently applied rules where by the possibility of including only materials which support the researchers ideas is eliminated.

3. Generalizability: The results obtained by the researcher can be applied to other similar situations.

Now, if content of communication forms the material for content analysis, where does a content analyst find himself/herself in the communication process? Figure shows the communication process and where the analyst figures. [diagram is missing here]

As can be seen, the analyst figures at the point of the message, and as some researchers points out, draws inferences about sender(s) of message, characteristics of message itself, or the effects of the communication on the audience that is the researcher interprets the content so as to reveal something about the nature of the audience or of its effects. Researcher’s incorporated these components in his classical formulation:

WHO says WHAT to WHOM with WHAT EFFECT?



Purpose	Questions	Research problems
To describe the characteristics of content	*What?*	▪ To describe trends in communication content. ▪ To relate known characteristics of sources to the messages they produce. ▪ To check communication content against standards
	How?	▪ To analyze techniques of persuasion ▪ To analyze style
	To whom?	▪ To relate known characteristics of the audience to messages produced for them. ▪ To describe patterns of communication.
To make inferences about the causes of content	*Why?*	▪ To secure political and military intelligence. ▪ To analyze psychological traits of individuals ▪ To infer aspects of culture and cultural change ▪ To provide legal evidence
	Who?	▪ To answer questions of disputed authorship.
To make inferences about the effect of content	*With what effect?*	▪ To measure readability ▪ To analyze the flow of information. ▪ To assess responses to Communication.

Figure 2.1: The Purposes of Content Analysis

2.2 Use of Content Analysis

Now, an attempt is made in this section, using some studies as examples, to explain about the applications of content analysis.

Though scholars from various disciplines such as social sciences, communications, psychology, political science, history, and language studies use content analysis, it is most widely used in social science and mass communication research. It has been used broadly to understand a wide range of themes such as social change, cultural symbols, changing trends in the theoretical content of different disciplines, verification of authorship, changes in the mass media content, nature of news coverage of social issues or social problems such as atrocities against women, dowry harassment, social movements, ascertaining trends in propaganda, election issues as reflected in the mass media content, and so on.

One of its most important applications has been to study social phenomenon such as prejudice, discrimination or changing cultural symbols in the communication content. For example, Berelson and Salter (1948) in their classic content analysis study highlighted the media under-representation of minority groups. They studied prejudice a consistent discrimination against minority groups of Americans - in popular magazine fiction. They content analyzed 198 short stories published in eight of the popular magazines during the period 1937 1943 and discussed their findings under the broad categories such as the distribution of characters, their role, appearance, status and their goals which the authors further classified as head goals and heart goals.

To understand the changing cultural symbols, Taviss (1969) content analyzed popular fiction in the 1900s and the 1950s to test the hypothesis that social alienation had been decreasing in middle class American society, while self-alienation had been increasing. The results indicated, for instance, an overall rise in the appearance of alienation themes, a slight decrease in social alienation and a large increase in self-alienation. Similarly, Lowenthall (1944) in his famous article Biographies in popular magazines examined the changing definitions of heroes in popular magazines in the US, and observed a drift away from working professionals and businessmen to entertainers.

One of the most frequent uses of the content analysis is to study the changing trends in the theoretical content and methodological approaches by content analyzing the journal articles of the discipline (Loy, 1979). Using this approach, Vijayalakshmi et al. (1996) analysed a stratified random sample of 194 research articles published in the Indian Journal of Social Work from 1971 to 1990 to identify characteristics of authors, and document the trends in empirical content, subject areas, and methodological characteristics such as source of data, research design, sampling, and statistical techniques used in the articles. Similarly, public attitude towards important issues such as civic amenities, unemployment and so on were assessed by analysing the content of editorials or letters to the editor in newspapers (Inkeles et.al, 1952, 1953; Devi Prasad et. al. 1992). One significant area of its use has been the analysis of newspaper content of the election coverage and editorial treatment to mould the opinion of voters. For example, Devi Prasad et. al.(1991), analyzed the editorials and letters to the editor published in four dailies in India before the 1991 elections to find out the prominent election related themes which figured in the news and direction of their coverage in the respective newspapers.

As a known unobtrusive research method, content analysis is some times used to study sensitive topics to corroborate the findings arrived at by other methods. Devi Prasad (1994) analyzed the dowry-related news items published in three English and six regional language daily newspapers during the period from 1981 to 1988. The news items were analyzed to understand the background of the dowry victim, persons involved in the conflict, possible causes of conflict, nature of victims abuse and death, and nature of reporting.

Content analysis has also been used to ascertain trends in the communication content of dailies, weeklies, cartoons, and coverage of development news, political news and crime news. Murty (2001) analyzed the news items, letters to the editor, and editorials of four selected dailies in India published during the calendar year of 1995, to make a comparative study of the coverage of development news. Political science researches have used the method to analyze the propaganda devices used by the warring groups (George, 1959; Lasswell et. al., 1965).

Other important applications of the method were systematic analyses of advertisements in newspapers and magazines to draw useful inference on national culture, as well as media preferences of advertisers (Auter and Moore, 1993; Wang, 1996). Similarly, television, radio, and movies offer rich sources of material for content analysis. Many scholars have explored changes in womens roles, sexual behaviour and health, and violence by analysing the content of in television and movie messages (Head, 1952; Lowry, 1989; Olson, 1994).

2.3 How to Perform content Analysis

Content Analysis begins with a specific statement of the objectives or research questions to be studied. The researcher asks the question what do I want to find out from this communication content and frames the objectives for study. The researcher must therefore locate a source of communication relevant to the research question and ask questions that can be solved by content analysis.
The objective of content analysis is to convert recorded raw phenomena into data, which can be treated in essentially a scientific manner so that a body of knowledge may be built up. In fact, the researcher who wishes to undertake a study using content analysis must deal with four methodological issues: selection of units of analysis, developing categories, sampling appropriate content, and checking reliability of coding.

More specifically studies using content analysis usually involve the following six steps:

1. Formulation of the research question or objectives
2. Selection of communication content and sample
3. Developing content categories
4. Finalizing units of analysis
5. Preparing a coding schedule, pilot testing and checking inter coder reliabilities
6. Analyzing the collected data

2.3.1 Formulation of Research Question or Objectives

As mentioned earlier, by making a clear statement of the research question or objective, the researcher can ensure that the analysis focuses on those aspects of content, which are relevant for the research. Content analysis is a method for analysing textual content. Therefore, the selection of topic should be one that can be answered by analysing the appropriate communication content.

2.3.2 Selection of Communication Content And Sample

The next step would be to locate relevant communication content to answer the research question and to determine the time period to be covered. If the body of content is excessive, then a sample needs to be worked out. Though sampling in content analysis is not so much different from sampling in surveys, because of the unique nature of the source material used in this method, there developed some special sampling techniques for content analysis. Thus, depending upon the nature of the communication content whether it is a new item, editorial, short story or a TV serial the sampling techniques differ. For instance, the use of constructed week and consecutive day sampling to control the bias of cyclical trends in news coverage, and the use of basic space unit approach to take a sample from large volumes of newspaper content - are some of the examples.

2.3.3 Developing Content Categories

Content categories can be defined as compartments or "pigeon holes" with explicitly stated boundaries into which the units of content are coded for analysis. They in fact flow from the research question and should be anchored in a review of relevant literature and related studies. Content categories are constructed in response to the query: What classification would most efficiently yield the data needed to answer the research questions raised?
The first step in category construction is preliminary examination of the communications by the researcher on a small-scale or as a pilot study so that such examination will result in the identification of possible content categories into which material can be coded. Usually one experiments with several categories before finalising a set of categories that can be used for the study. Sometimes, category systems already developed by other researchers may also prove useful for your study.
To be useful, every content category must be completely and thoroughly defined, indicating what type of material is and is not to be included. Such definitions in most of the cases should be written down before coding begins. These form the operational definitions of categories. According to Chadwick et al., (1984), categories must be mutually exclusive so that a word, a paragraph or a theme belongs in one and only one category. Also, the categories must be exhaustive so that all units examined fit in an appropriate category. Sometimes, a 'miscellaneous or residual category' is added for units that occur rarely or are un-code-able for other reasons.

2.3.4 Finalizing Units of Analysis

At this stage, that is, once the categories are identified and defined in terms of the research objectives, the content analyst asks two interrelated questions.

They are:

1. What unit of content is to be selected for classification under the categories? and
2. What system of enumeration will be used?

The unit of analysis is the smallest unit of content that is coded into the content category. The units of analysis vary with the nature of data and the purpose of research. Thus, the unit of analysis might be a single word, a letter, a symbol, a theme (a single assertion about one subject), a news story, a short story, a character, an entire article, or an entire film or a piece of programme. There are two kinds of units of analysis. 1) Recording units, 2) Context units. The recording unit is the specific segment of content in which the occurrence of a reference/fact is counted or the unit can be broken down so that reference/facts can be placed in different categories. For example, if it is a single word say 'democracy', the number of times the word appears can be counted. Similarly, a sentence, or a paragraph, a news item or an article containing a symbol or a theme, or a group of facts can also be a unit. Thus, a news item containing a group of facts can be coded in different categories. According to some researchers, five major recording units have been used frequently in content analysis research: words or terms, themes, characters, paragraphs and items . The 'item' may be an entire book, an article, a speech or the like.

The counting or quantification of the units is performed by using three methods of enumeration: a) space/time, b) frequency and c) intensity or direction. A unit of analysis can be measured in terms of space (for example, number of column inches) or time (minutes devoted to a news item on the TV). In the case of frequency, it is the number of times a given unit or theme figured in a body of text - is recorded. Intensity or direction implies the measurement of the direction of the symbolic meaning contained in the message.

2.3.5 Preparing a Coding Schedule, Pilot Testing and Checking Inter Coder Reliabilities

Defining categories and preparing coding schedule for the analysis and coding of content are simultaneous steps. A coding schedule resembles a survey questionnaire and contains different dimensions of the communication content to be coded. Next, piloting the coding schedule is a crucial step before launching the full-scale content analysis. Test coding of a small sample of the material to be analyzed helps reveal inconsistencies and inadequacies in the category construction.

Coding the unit of analysis into a content category is called coding. Individuals who do coding are called coders. The coder may be the investigator himself/herself or employed by the investigator. Careful training of coders, which usually results in a more reliable coding, is an essential task in any content analysis. It is probably desirable to have, even in a small-scale study, more than one coder to independently code the units and to check the inter coder reliabilities.

2.3.6 Analysing the Collected Data

How should the data be analysed? The definition of the research problem gives direction to data analysis, the patterns to be examined, and the relationships to be explored. As in the case of analysis of survey data, the starting point can be the description of the profile of the main categories such as for example characteristics and types of content by period, actors, and so on. Later, the analysis can move to conduct more complex analyses comparing two or more dimensions, periods or data sets.

Tables- univariate and bivariate - can be prepared and cross tabulations can be arrived at. Depending on the nature of data, the statistical principles that apply to other areas of survey research will also apply to content studies with a very few differences. The findings can be presented forcefully even with simple percentages and cross tables. Tables 2 and 3 from the sample article on dowry related violence are shown here. In one case, in Table 3, the death occurred in a place close to the couples residence where the husband, a police constable, murdered his wife and threw her body into the river.

Chapter 3

Related Work

Authors [1] have also proposed a dynamic approach for discovering relations, given a large corpus file. They divided the output into two categories as valid and invalid relations and finally they weighted both. Their system has both the strengths and weakness of traditional relation extraction and open relation extraction. Never-Ending Language Learner (NELL) uses initial ontology and outputs the extraction of facts from the web. In future, their system output will be the input of NELL to gain more accuracy.

Different people have used different techniques to find out entities form the text, [2] they has proposed the conditional random field for identifying entity classes from social-technical system. They have used CRF (Conditional Random Field) and machine learning to extract the relational information from the text corpus. Although this technique is quite useful and has shown good results and has classified the data into semantic model, but these kinds of techniques are not very useful when you have diverse domains and every domain is different from the other. Researchers have also worked on a list of entities which need to be searched from the text document as in [3]. They have proposed the Adhoc entity extraction technique from text collection using inverted Index created on the document and they have shown that their technique is faster than the traditional entity extraction processes. [4] has presented text compression techniques to find out tokens which can be modelled as entities in the text. There approach also requires training document for each domain in news article which might not be practical. All of these methods are based on supervised machine learning techniques in which system analyses the provided corpus and generates the list. By using different techniques it tries to find out the entities from the provided text documents. Similarly, people have also proposed semi supervised and unsupervised way of learning entities from the given text. The most popular technique used in semi supervised machine learning technique is bootstrapping which requires few set of clues to start the learning process for identifying entities, [5] have demonstrated in National Conference the same technique by using pattern generalization techniques. Similarly the unsupervised way of learning entities is to cluster the data based on different heuristics about the Entities e-g entity will be noun, start with capital letter or have the same context.

However there is relatively less research in the area of identifying relationship between entities automatically. Systems generally require that some information about the relationships is specified that need to be learned. For example [6] and [7] systems are based on bootstrapping process i.e. we have to define the relationship name and the entity type to which this particular relation will be applied. [7] has also defined the mutual exclusion can be created between category predicates and that reduce the semantic drift and system achieved the precision of 89%. Although these techniques are quite useful and worked well but they are quite costly when there are thousands of relationships and predicates. [8] has created positive and negative set for extracting relationships ,although in his method, one should not have to provide the seed examples and names of relationship but still we have to provide training documents before it starts working. Unsupervised machine learning techniques have also been used to learn the relationship between entities. In order to remove the short coming of supervised and semi supervised way of learning relationships, [9] has proposed a technique based on creating a vector that contains the features of the Entities and their pair is taken under consideration by creating a similarity matrix. But this similarity matrix does create lots of noise because of writers writing style. [10] has proposed tree based similarity matrix to cluster the similar relationships among the tuples. But the same problem occurs with this technique moreover that writers style of writing may introduce the noise. They are only focused on the single document, but here we are more focused on finding the facts that are reported about entities in different articles. We want to find true relationships between entities so that one can use these facts for defining their strategies.

3.1 Entity Extraction

In this era of electronic text it is a challenge to identify relations within innumerable piles of information. Manually labelling the information is beyond practical work. The paper hence has done a synthesis formalization of the problem and defined an overview of the architecture and the steps that are developed for the solution.

WISE conference T1 Track proposes teams to label entities within plain texts based on a given set of entities. the dataset of wiki-links consist of 40 million

mentions over 3 million entities. the paper shows a straight forward two fold unsupervised strategy to extract and tag entities.
The paper refers data set to the wiki-links that contain over 40 million mentions labelled as referring to 3 million entities. The dataset under consideration is dramatically large and have more variety that makes it more challenging again existing approaches. The paper then describes the formal definition of data then the first step of preprocessing is described. The wiki-link entity file is processed which creates database tables in return. Two separate tables are created for all the entities and their respective mentions that contained reference to the original file. Each entity is then processed to create an auxiliary table that contains tokens of the core words of the entity. After this first step, the preprocessed texts are divided into tokens and they go through an aromatic grammatical tagging to get its appropriate part of speech. In IE tokenization is an important step. After tokenization is completed Sentence Analysis and Entitiy Extraction is done to identify proper noun groups and concrete concepts based on parsing.
Later Entity matching is executed to determine if named entity refers to which wiki-links. The objective of the paper was to achieve high precision level in labelling entities within plain texts. Achieving the objective, architecture of the paper was very simple and scalable.

3.2 Document Grouping

The paper represents an emerging problem of organizing and discovery of information and its solution. In this era of information, retrieval and presentation of information is crucial. For the mentioned purpose the paper discusses two main requirements of information when it comes to presenting the information to a common user. One is finding a specific information and the second one is browsing the topics and structure of a collection of document.
Efficiently organizing documents into hierarchy of topics and subtopics is essential for finding the required information with minimum wastage of time. Document clustering and labelling is one method of effectively organize the documents. This process clusters a document collection into smaller groups with different topics. The process is done repeatedly until the topics are precise enough. The paper uses community mining form social network analysis to generate topic coherent document clusters and give each topic a cluster descriptive label. The paper proposes to use betweenness centrality measure in networks of co-occurring terms to label the clusters. Afterwards the paper includes key phrase extraction and automatic titling in cluster labelling using

KEA method that ending up in generating best labels compared to all other methods. The paper at the end also shows an experiment done using the said technique which expresses that the method have a strong disambiguation ability.

Further the paper describes the methodology in which the basic idea it to reformulate the document clustering problem into a topic community mining problem. As a first phrase of the method, Keywords are extracted. This is a time taking method and can be done off-line along with crawling. Noun phrases are chosen as keywords since they are grammatically consistent. Then to tag each single word with its part of speech, part of speech (POS) tagging is done. After keyword extraction from the document a keyword graph is built. In this graph a node is a keyword whereas an edge is formed when two nodes have co occurred in at least one sentences. Edge weight is the sentence co-occurrences.

Next to detect different topical communities, community mining is done on the keyword graph. Community mining is the grouping of nodes. Fast modularity clustering algorithm (O(nlog2n)) is used. This algorithm is applied recursively in a top down manner until certain conditions are met. Following phase generates document clusters by assigning the documents to the keyword communities. In the last phase the clusters are labelled. Most common method of cluster labelling is to use most frequent or central phrases in a document cluster as labels. Paper uses frequent and Predict Words method that extract terms which are more likely to appears in a cluster than in another clusters as labels as another baseline.

Experiment conducted for the paper constructed data set using Goggle. They experimented with a multitude of queries with ambiguous meanings. Top level and lower level clusters are evaluated separately in the taxonomy. Top level carry the clusters that are under the root directly which shows the disambiguation ability of the method used.

The paper proposes four different labelling methods and found that the labelling method utilizing KEA generates the best over all cluster labels. They included the suggestions for future work as to explore other ways to build the keyword graph so that its sensitive to different subtopics. Another suggestion given is to combine different labelling methods to improve the labelling performance by reflecting the strength of each method.

3.3 Relationship extraction

The paper shows the automatic analysis of television news programs making use of the closed captions that are the steams of plain text mentioned with these news programs. To process this closed caption stream of data the paper uses NLP based pipeline tools to segment, process and annotate captions automatically. Afterwards the linguistic style of the captions accompanying the news is processed. Then an insight on the news providers is made to reveal the biases using automatic methods. Using these new reveals qualitative assessment is also made as to look data from a different perspectives.

Closed captions that comes with every news program is provided by the news network itself. Using these captions the paper studies how to characterize each news network, each news maker and relationship between news maker and the news network. The method mentioned in this paper can be used for any data mining task that uses closed captions.

NLP based pipeline described in the under discussion paper works in the following flow of tasks. It first filters out the non-news programs. Then segmenting the captions into sentences. After words it detects the named entities of news providers and applies a part of speech tagger to find words and qualifiers used together with each entity. Lastly it labels each sentence with an over all sentiment score automatically. The first step of this automatic analysis includes text pre-processing. First the stream of text is segmented into sentences. Then a named entity tagger is used to specify the named entities. Later the captions now processed are match to an aggregator of news stories obtained from online. In matching pre genre classification model trained on thousands of examples labelled by editors is applied.

After text pre-processing, news providers are examined under the query of styles and genres, coverage and timeliness. Factor analysis is applied to obtain a soft bi clustering of news providers according to the words they use more frequently. For measuring coverage, prominence is an important factor, which means by scale provided in the paper the prominence is 1 if the story is covered by all the providers of the same genre. Last but not least timeliness and duration is measured in the means of how different providers cover a story over a period of time. Duration of a story is measured between the first and the last matching of the story.

Next the paper presents the analysis on the news makers. Named entity tagger resolves the names in the captions to obtain more information about the people. After that automatic clustering of news makers is done to complete the analysis of news makers.

Succeeding from here, the paper shows how different news providers frame different people. Positive and negative coverage shows how different people are projected through different networks and there are some outliers who gets different treatment in some providers than the rest. Concluding the work paper says that closed caption data brings the television domain in the stream of data mining research. Datasets demonstrated in this research shows their richness by analysing relationship between genres and styles, coverage and timeliness and sentiments and biases when it comes to the news makers and news providers. The steps provided by the NLP based pipeline can be further utilized in future where caption data is involved.

Chapter 4

Natural Language Processing

In the area of Information extraction from un-structured documents, the role of linguistic syntax is very important. The study of this area helps us to uncover the underlying recursive structure, which governs the human language. It also uncovers the finite ways of words combine together and from phrases, which in turns combine together and form sentences. In addition to this analysis of syntax also reveals us that what are the strings that combines together for valid use of language and therefore gives us the opportunity to define systematic relationship between syntactic structure and it meaning. The Role on NLP (Natural Language Processing) in this area is to define a computer program which takes all these linguistic rules and converts the un-structured document into the form which is useful for modelling the information. So the Question over here is if we have all these linguistic rules and information why NLP is difficult? The short answer to this question is ambiguity. Ambiguity, because each sentence in may be transformed into different meaning based on it context, the speaker and many other factors. The ambiguity of a sentence may take many forms for example, syntactic ambiguity, semantic ambiguity or ambiguity of pronominal reference. Based on these hypotheses the roles of NLP parsers is very important in order to transform unstructured document into structured form. For example we have one sentence Print the file in the buffer and second sentence is print the file on the printer now in these two sentences the prepositional statement can be attached to nearest noun phrase or to the higher verb phrase. Now its parsers job to identify automatically which rules of linguistics should be applied to it.

The research in this area was started in early 90. The studies in that era related to linguistic syntax were based on parsers which were developed on detail representation of linguistics. They had coverage of 70%, which means that rest of 30% sentences had no analysis. Due to detail representation

sometimes very simple sentence comes up with many possible analyses which make them difficult to use. For all those multiple analysis of sentences some of them gives multiple outputs and some of them were changed based some ad-hoc treatment rules. These early parser had also performance implications, for some of them it was not defined and for some of them it was defined based on quantitative analysis, and due to these reason performance comparisons among those parsers was often in-commensurable. In new Era, Statistical parsing was introduced which was very accurate and fast as well. In statistical parsing grammar rules were associated with probability, and these probabilities along with the grammar rules helped new era parser quite a lot in terms of accurate parsing and performance. With these probabilities each sentence is parsed against all the candidate parses and the computation of final parse comes based on the high probability. Most of the work in statistical parsing was based on PTB (Penn Treebank), which is a collection of parsed text corpus that annotates syntactic or semantic structure. This resource acted as valuable resource and test-bed for developing and evaluating the statistical parsing, best performing system achieved F-score of 92%.
In the rest of the chapter we will go through the details of statistical parser, we will cover that why statistical parsers are important what are the areas that they have covered in NLP parsing and how we have used them in our research.

4.1 Usage of Statistical Parsers

Statistical parsers are widely use in these days some of their applications are as follows

1. High Performance Q&A Systems [16]: In this Research they have proposed high performance Question and Answering system which have unparallel performance in TREC-9 Evaluations. And they have used Statistical parsers to find out repeated passages, lexico semantics, and answer caching and taxonomy development, through all these areas in this research they have improved the results of Automated Questioning and answering system. And in results of this paper they have presented their Q&A system as a system that covers wide coverage mechanisms to identify the answers of open domain with multi-feedback retrieval scheme.

2. Improving Biological Named entity Extraction (Jin-Dong KIM, 2004): In this research they have utilised the statistical parsers for classifying

the technical terms into molecular Biology Field. NER is also very important in the field of Bio technology, therefore JNLPBA has shared task for linking Bio-entities, this research has participated in that task by using statistical parsers to identify the field in the area of molecular biology. In order to fulfil their results they have considered features like POS tags, orthography, domain knowledge and fined grilled error analysis for building their statistical parser. And their results have shown that in this area the usage of statistical parser has clearly improved the F-score and accuracy.

3. Syntactically based Sentence Comparison [15]: In this research they have utilize the statistical parser for comparing sentences in the field of biomedical for finding related articles in biomedical domain. They have utilize the syntactic trimming approach which was derived from the newswire domain and this approach has proved great results and second area on which they have worked was extrinsic evaluation and grounding summarization in real-world task. All these areas were based on statistical parser and they have proven significant results.

4. Extracting peoples opinion about Appraisal [14]: In this research they used Natural Language processing using statistical parsers for extracting peoples opinion about appraisal. In this research appraisal is considered as textual unit expressing an evaluative expression towards some target.

5. Improving human interaction in computer games [17]: In this research they have proposed models for intelligent system that can interact with humans and in this system they have used Language processing. All Language processing was done using statistical parsers in this research.

4.2 Statistical Parsing

As we have seen that Statistical parsing in the area of NLP has significant role and many researchers and people have utilize it. The statistical parsing is majorly focused on syntactic analysis, such as Part of Speech Tagging (POS) and grammar induction. So we will go in detail what these areas means and how they are significant in our research.

4.2.1 Part Of Speech Tagging

Part of speech tagging is very hot topic in the area of statistical parsers; in POS we basically identify the parts of sentence into grammatical tags or word category disambiguation. The problem in part-of-speech tagging is to assign to each word in a sentence a part-of-speech label which indicates the linguistic category (e.g. noun, verb, adjective, etc.) to which that word belongs in the context of the sentence. Some part-of-speech tag sets only have a few dozen coarse distinctions, while others include hundreds of categories, distinguishing temporal, mass, and location nouns, as well as indicating the tenses and moods of verbs. Hidden Markov Model (HMM) was the very first technique which was proposed in this area during the lecture by Mercer at MIT and published a paper with this technique on [18]. At BBN [19] has discovered that behavior of HMM based tagging algorithm was not good because it was trained on limited set of data and usage of different model based on HMM tagging revealed many weaknesses.
Decision Tree based Techniques are also proposed in this area like [20].His work was an attempt to extend the usual three-word window made available to a trigram part-of-speech tagger. By allowing a decision tree to select from a larger window those features of the context which are relevant to tagging decisions, he hoped to generate a more accurate model using the same number of parameters as a trigram model. His results, however, were not much better than those of existing taggers. Statistical decision tree (SDT) classification algorithms account for both of these tasks. SDTs can be used to make decisions by asking questions about the situation in order to determine what the best course of action is to take, and with what probability it is the correct decision. For example, in the case of medical diagnosis, a decision tree can ask questions about a patients vital signs and test results, and can propose possible diagnoses based on the answers to those questions. And, using a set of patient records which indicate the correct diagnosis in each case, the SDT can estimate the probability that its diagnosis is correct. For a particular decision-making problem, the SDT growing algorithm identifies the features about the input which help predict the correct decision to make. Based on the answers to the questions which it asks, the decision tree assigns each input to a class indicating the probability distribution over the possible choices. SDTs accomplish a third task which grammarians classically find difficult. By assigning a probability distribution to the possible choices, the SDT provides a ranking system which not only specifies the order of preference for the possible choices but also gives a measure of the relative likelihood that each choice is the one which should be selected. A large problem composed of a sequence of non-independent decisions, like the parsing

problem, can be modeled by a sequence of applications of a statistical decision tree model conditioned on the previous choices. Using Bayes Theorem to combine the probabilities of each decision, the model assigns a distribution to the sequence of choices without making any explicit independence assumptions. Inappropriate independence assumptions, such as those made in P-CFG models, seriously handicap statistical methods.
The decision tree algorithms used in this work were developed over the past 15 years by the IBM Speech Recognition group. The growing algorithm is an adaptation of the CART algorithm in Breiman et.al.[13]. The IBM growing and smoothing algorithm were first published in Lucassens 1983 dissertation [38]. Bahl, et.al., [2] is an excellent discussion of these algorithms applied to the language modeling problem. For this dissertation, I explored variations of these algorithms to improve the performance of the decision trees on the parsing task.

4.2.2 Grammar Induction

Grammar Induction, sometimes also referred as grammatical inference is a process in machine learning, for learning formal grammar to form a set of observations thus forming a model which accounts for the characteristics of the observed object. Most of the work in this area was based on availability of parsed and un-parsed corpus, and due to pre-parsed corpus the major work in this area was based on un-supervised learning or using the Tagged Brown corpus, learning from corpus tagged for part-of-speech. Grammar induction aims to find the hierarchical structure of natural language. Grammar search methods have met with little success, and simple distributional approaches that work for part-of-speech induction do not directly apply. For example, differentiating noun phrases, verb phrases, and prepositional phrases requires discovering the three clusters in the left plot – which seems easy enough. However, deciding which sequences are units at all require telling apart the red and blue clusters on the left – which is much harder.

Up till now we were looking at the components of sentences that is words, moving forward if we want to develop a sophisticated application based on Tagged part of the Sentence we need to have some more understanding that lies in the larger structure of the sentences therefore some properties of language that needed to be analyzed are structure and meaning, Agreement between different tags identified, identification of recursions and long distance dependencies. Therefore the standard way of presenting parsed sentence is in the form of Tree like in the bellow diagram

The Tree like structure allows us to state the dependencies in a consistent and concise manner. For instance English requires number of agreements

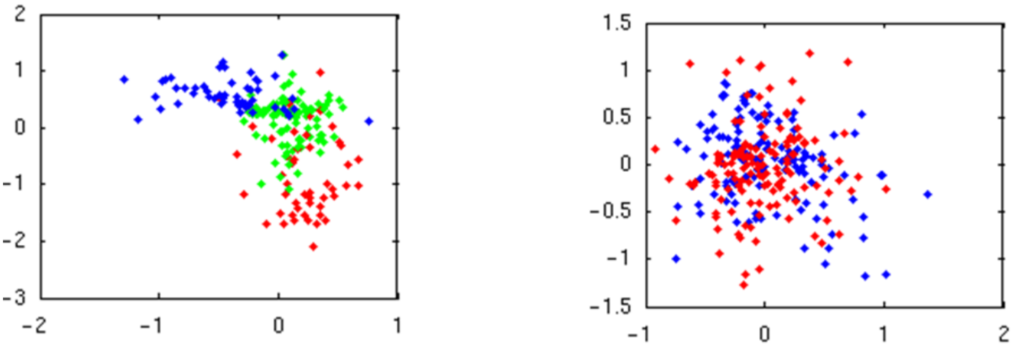


Figure 4.1: Cluster of verb phrases

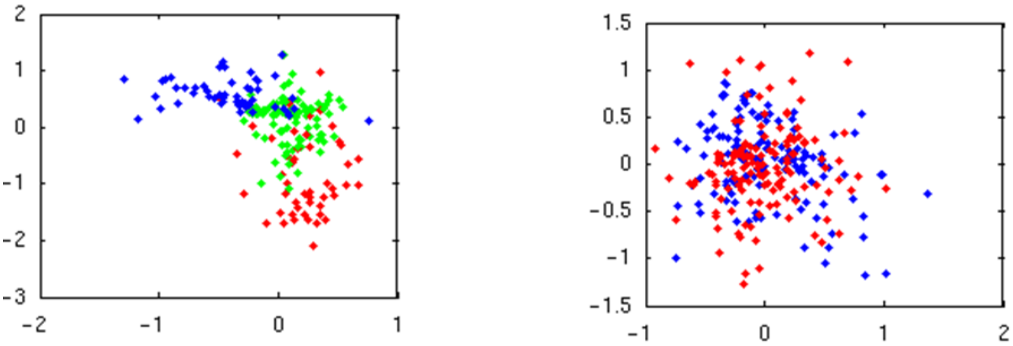


Figure 4.2: Parse Tree for Sentence

between the subject NP (however complex) and main verb. These kinds of constraints can easily be summarized in tree like structure where NP is followed by VP and can produce a reasonable sentence S. Now Question over here is if we have a tree parse for particular sentence how we can verify that this the correct parse for the sentence? The answer is the space for allowable trees are specified by the grammar, and in our case it is context free grammar CFG. A CFG consist of set of the forms category0 -¿ category1 categoryn which states that a labeled with category0 is allowed in tree when if has n children of categories from left to right.

But this tree like structure guided us towards the ambiguity as stated in above diagram. This ambiguity arises from choosing a different lexical category for the word annoying, which then forces different structural analyses. There are many other cases of ambiguity that are purely structural, and the words are interpreted the same way in each case. The classic example of this

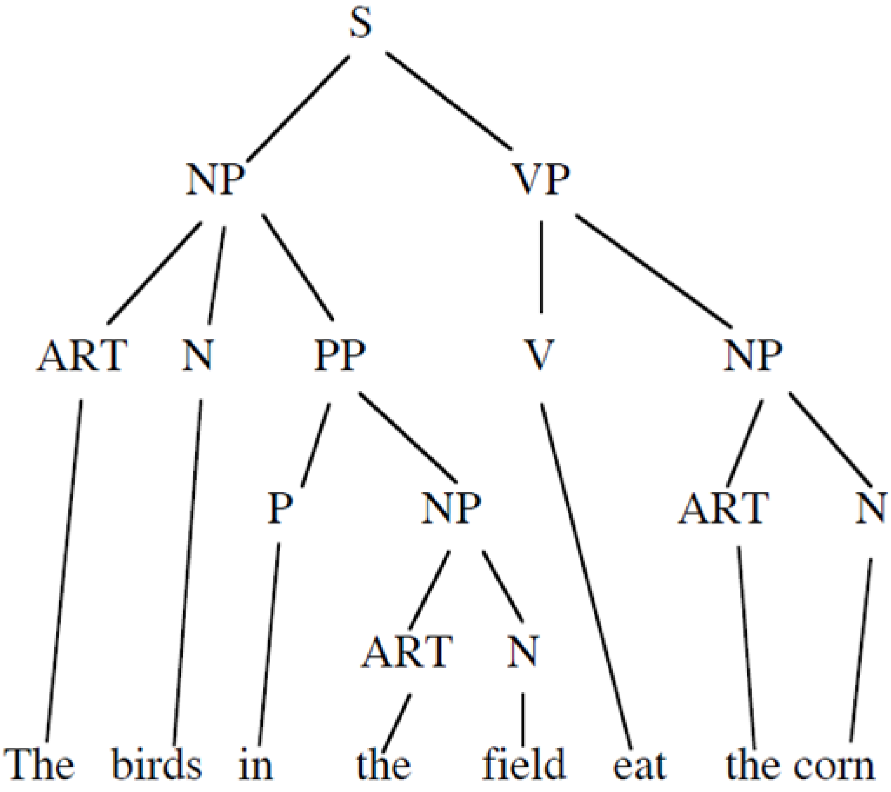


Figure 4.3: 2 Parse Tree for same Sentence

is the sentence I saw the man with a telescope, which is ambiguous between me seeing a man holding a telescope and me using a telescope to see a man. This ambiguity would be reflected in where the prepositional phrase with a telescope attached into the parse tree.
Given a grammar and a sentence, a natural question to ask it 1) whether there is any tree that accounts for the sentence (i.e., is the sentence grammatical), and 2) if there are many trees, which is the right interpretation (i.e., disambiguation). For the first question, we need to be able to build a parse tree for a sentence if one exists. To answer the second, we need to have some way of comparing the likelihood of one tree over another. There have been many attempts to try to find inherent preferences based on the structure of trees. For instance, we might say we prefer tree that are smaller over ones that are larger. But none of these approaches has been able to provide a satisfactory account of parse preferences. Putting this into a probabilistic

framework, given a grammar G, sentence s, and a set of trees Ts that could account for s (i.e, each t has s as its leaf nodes), we want the parse tree that is argmaxt in Ts PG(ts)
To compute this directly, we would need to know the probability distribution PG(T), where T is the set of all possible trees. Clearly we will never be able to estimate this distribution directly. As usual, we must make some independence assumptions. The most radical assumption is that each node in the tree is decomposed independently of the rest of the nodes in the tree. In other words, say we have a tree T consisting of nodes T1, with children T1.1, , T1.n, with Ti.j having children Ti,j,1, , Ti,j,m and so on, as shown in figure

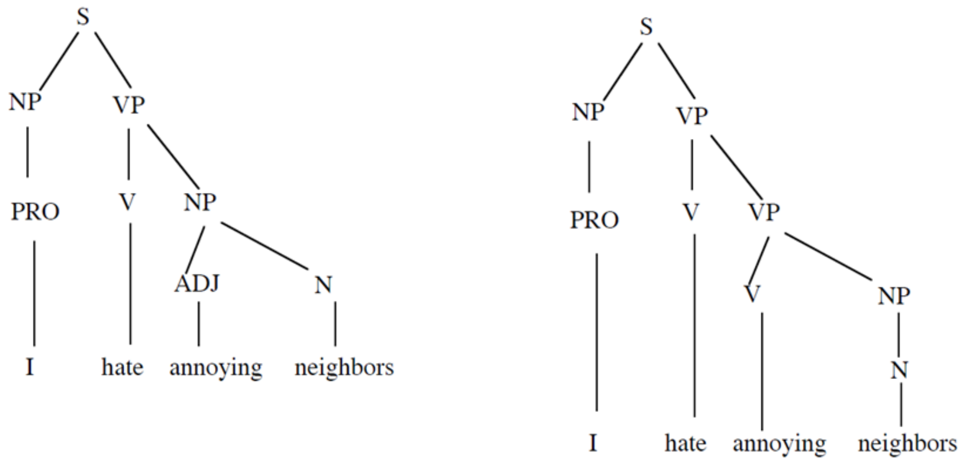


Figure 4.4: Generalize Parse Tree for Sentence

Thus making the independence assumption about, wed say P(T1) = P(T1 T1.1 T1.2) * P(T1.1) * P(T1.2) Using the independence assumptions repeatedly, we can expand P(T1.1) to P(T1.1 T1.1.1 T1.1.2 T1.1.3) * P(T1.1.1) * P(T1.1.2) * P(T1.1.3), and so on until we have expanded out the probabilities of all the subtrees. Having done this, wed end up with the following:

P(T) =S r P(r) where r ranges over all rules used to construct T

The probability of a rule r is the probability that the rule will be used to rewrite the node on the right hand side. Thus, if 3/4 of NPs are built by the rule NP -¿ ART N, then the probability of this rule would be .75. This formalism is called a probabilistic context-free grammar (PCFG).

Number	TAG	Description
1	CC	Coordinating conjunction
2	CD	Cardinal number
3	DT	Determiner
4	EX	Existential there
5	FW	Foreign word
6	IN	Preposition or subordinating conjunction
7	JJ	Adjective
8	JJR	Adjective, comparative
9	JJS	Adjective, superlative
10	LS	List item marker
11	MD	Modal
12	NN	Noun, singular or mass
13	NNS	Noun, plural
14	NNP	Proper noun, singular
15	NNPS	Proper noun, plural
16	PDT	Predeterminer
17	POS	Possessive ending
18	PRP	Personal pronoun
19	PRP$	Possessive pronoun
20	RB	Adverb
21	RBR	Adverb, comparative
22	RBS	Adverb, superlative
23	RP	Particle
24	SYM	Symbol
25	TO	to
26	UH	Interjection
27	VB	Verb, base form
28	VBD	Verb, past tense
29	VBG	Verb, gerund or present participle
30	VBN	Verb, past participle
31	VBP	Verb, non-3rd person singular present
32	VBZ	Verb, 3rd person singular present
33	WDT	Wh-determiner
34	WP	Wh-pronoun
35	WP$	Possessive wh-pronoun
36	WRB	Wh-adverb

Chapter 5

Problem Statement

We are aimed to design a system for news articles, which can detect entities, group documents, and find out the dynamic relationship between entities for that topic

Chapter 6

Methodology

6.1 Entity Extraction

The Named Entity Recognition or simply entity recognition is a process which has been inspired from machine learning techniques. People have worked in all the three areas including supervised learning, semi-supervised learning and unsupervised learning. The area in which most of the work has been done is supervised machine learning which includes rule-based systems and sequence labelling systems. Although the supervised learning methods are quite successful, yet they require a large collection of annotated documents before it can work on actual system. But in our case, we have many diverse areas which may include sports, politics, business, current affairs etc. Therefore, it would be impractical for us to provide documents related to all of these areas.

We have an unsupervised machine learning algorithm in order to model dynamic entity modeling in which case, we are more focused on the process of clustering the groups based on different rules. We first take a document and perform parts of speech (POS) tagging in that document. After tagging, we try to find out the marked nouns in the document before calculating the following features of the gathered nouns:

1. Case: There are three parameters that are defined for this feature. The most important one is whether it starts with capital letter or not, if it starts with capital letter then this might be a possible candidate for entity, second parameter includes all uppercase or not (usually organization names are all in uppercase) and third is whether its a mixed case word or not.

2. Punctuation: Whether some punctuation is used in the word or not for

Table 6.1: candidate entities Extracted From Articles

Candidate For Entities	**Count**	**Percent**
Punjab Assembly	2	5.714286
Bahawalpur	2	5.714286
South Punjab	3	8.571429
Minister Rana Sanaullah Khan	2	5.714286
Speaker Rana Mohammad Iqbal Khan	1	2.857143
Pakistan Peoples Party (PPP)	5	14.28571
Usman Bhatti	1	2.857143
PML-N	4	11.42857

example hyphen, apostrophe or & sign.

3. Digits: If digits are used in word or not for example 3M or W3C etc.
4. Prefixes or suffixes: Whether some prefixes/suffixes are used or not for example Mr. or Ms.

After calculating the features, we filter the words which are not candidates to be entity in the document. We then calculate the occurrence weight of the found word in the document because in News data, people are generally focused on reporting similar kind of entity. They try to present facts related to one or more entities which are present in the reported event. Here we will present our research and perform different experiments to find out what weights will be good to consider something as entity.

We have also calculated the number of times the entities are repeated in the document i.e. the frequency count of the entities along with their percentages.

Let N be the set of Nouns present in document D, such that $N = \{n_1, n_2, n_3, ..., n_j\}$ where j is the total number of nouns marked in the document D. Here each n_i is selected such that it groups the nouns that are adjacent to each other, we perform this step because POS tagging mark each word with some grammatical relation and there is possibility that two NP are marked one after another. But in actual these two NP which came adjacent to each other are one noun for example Pakistan Peoples Party (PPP) all these four words corresponds to one entity but POS is tagging each word thats why they are four NPs adjacent to each others. After creating this set N we will then build an occurrence grid which records the count of the particular noun in the given document. It must be noted that we will count all the occurrences. Such that if a part of that noun is repeating then it will also is considered in the occurrence of that noun. For example Pakistan Peoples Party (PPP) is a noun if rest of document contains PPP 10 times then it will contribute to same noun Pakistan Peoples Party (PPP).

6.2 Document Grouping

In order to move forward we first have to group the similar event reporting documents together so that we can extract the correct relationship between two Entities. Here documents are grouped based on the candidate Entities list which has been generated by the previous step.

Therefore we call this step as document grouping task, in which we try to assign a Boolean value to each pair $(D_j, G_j) \in G$ where $G = \{G_1, G_2, G_3, ..., G_{|G|}\}$ set of dynamically generated groups. A value of T is assigned to (D_j, G_j), shows the decision made by the function F by assigning one value from $\{T, F\}$. As we already know that for each document we have already created the set $N_i = \{n_1, n_2, n_3, ...n_j\}$, we will utilize these set of candidate nouns to categorize the documents into groups.

Initially first document D_1 is assigned to group G_1, and its set $N_1$is assigned to G_1. Then for each document D_i we will take intersection of N_j with group $G_j\,(i.e. N_j \cap G_j))$, that will give us set, we call it U_j. We have two cases here, first one is U_j is null set; if this is the case then clearly the Document D_j does not belongs to G_j. But if U_j is not null then we will calculate the accumulated percentage of D_j in G_j, if accumulated percentage is greater than the defined threshold then it is considered as match, otherwise D_j does not belongs to G_j. If D_j belongs to G_j then we will take union of G_j and N_j, to get new set which is assigned to G_j.

Let $G = \{G_1, G_2, G_3, ...G_{|G|}\}$ it must be noted that each is a noun along with its percentage in the document. When we take intersection of G_j with D_j. It will always give us set U_j. We sum all the percentages associated with the nouns in G_j to get similarity percentage. 6.2 has shown us the same thing, whereas 6.3 is showing the set N for j^{th} document.

Table 6.2: Set G for Group j

No	Candidates For Entities	Percent
n_1	Punjab Assembly	5.714286
n_2	South Punjab	5.714286
n_3	Minister RanaSanaullah Khan	5.714286
n_6	Pakistan Peoples Partys (PPP)	5.714286
n_8	PML-N	5.714286
n_{13}	Federal Government	5.714286
n_{16}	Election Commission	5.714286

If we consider above two sets we will find that there are some entities that are present in both sets like n_1, n_2, n_3, n_6, and n_8. If we sum up their percentages these will add up and makes the total of 52 which is saying that the document D_j is 52% similar to group G_j and we have defined the threshold of 50% therefore this documents could be assigned to this group. But we have also considered the matching of mostly occurred nouns in the both documents i.e. at least one of the mostly occurred nouns must be matched with the atleast one of the noun in Group and similarly for the groups mostly occurred noun. Since we can see it is happening in these two sets therefore we can assign this document to the group and can take union of two sets to make bigger set.

6.3 Relationship Extraction

As we know that the relationship extraction is the most important part of the whole system which must work automatically. Our key idea for extracting relationship between entities is to use the redundant information that has been reported by the different sources about the same event. Because people will report the same relationship using different contexts and we will

Table 6.3: Set N for document j

Candidates for Entities	**Count**	**Percent**
Punjab Assembly	3	7.317073
Pakistan Peoples Party	1	2.439024
PPP	6	14.63415
South Punjab	5	12.19512
Minister Rana Sanaullah	1	2.439024
PML-N	6	14.63415
Bahawalpur	3	7.317073
National Assembly	1	2.439024
Fata	1	2.439024
Hazara	1	2.439024
Pakistan Muslim League Quaid (PML-Q)	1	2.439024
PA	3	7.317073
Deputy Parliamentary Leader	1	2.439024
PA Shaukat-Mehmood Basra	1	2.439024
Article 239	1	2.439024
NA	5	12.19512

use those contexts to extract the information.

In previous steps we have already performed the entity extraction and the classification of document. Now we are aiming to make a List L for each document, which contains entity-connecting phrase-entity as its elements such that these patterns are used more than once in the document and patterns must be like E_1 Verb E_2, E_1 NP Perp E_2, E_1 Verb Perp E_2, and E_1 to Verb E_2 (where E_1 and E_2 are two entities), and all those patterns which are like above will be considered as an part of list L. Now we have list set $L = \{L_1, L_2, L_3, ..., L_k\}$ for k documents belonging to the same event. From these we are aiming to build a context Set C such that each C_i contains all the contexts of the same two entities.

6.4 Algorithm/ Pseudo Code

Input : All context sets $S = \{C_1, C_2, ..., C_m\}$ where m is the total no of different entity pairs, have been found from the articles reporting same event.

Output : Set of entities and their relationships.

Steps :

1. From context sets S build context occurrence weight matrix for each E_i and E_j and normalize the matrix.
2. Rank the relation in each column according to their co-occurrence in the documents
3. Select that relation which got the highest score on the document.

Here for selecting the relationship between entities we have built context occurrence matrix. In this matrix we have put all the possible context into list and calculated their occurrence in the document, for example Punjab Assembly has approved the resolution for Bahawalpur and South Punjab. Some articles have used this as passed the resolution and some has used Adopted as shown in 6.4. We have counted their occurrence and have calculated weight as follows:

$$Matrix\,(i, 1) = \frac{\sum V_i}{\sum_{k=1}^{m} V_k}$$

Here m is the total no of context used for Entity E_i and E_j

Table 6.4: Relation calculation

	Has Approved	**Passed**	**Adoppted**
Weight	0.167	0.667	0.166
Rank	2	1	3

6.5 Extracting Valid Relations Among Entities

There will be some invalid relations that need to be consider here
Incomplete : there will be some areas, where the relation will not be completed with addition of nouns and other entities. For example consider Punjab Assembly has approved resolution for South Punjab and Bahawalpur. In above relation there are three entities, which are involved i-e (Punjab Assembly, South Punjab and Bahawalpur).
Ambiguous clustering : as we know that if we have done wrong clustering then the relations extracted will also be wrong, therefore thresholds for clustering should be set carefully.
We will consider the above issues by adding additional step in our relationship extraction i-e re-validation of relationship. In this step, we will monitor the existence of more than one entity in the same context. In case it exits, we will add them in order to complete our relations between entities.

Chapter 7

Experimental Results

7.1 Data Set

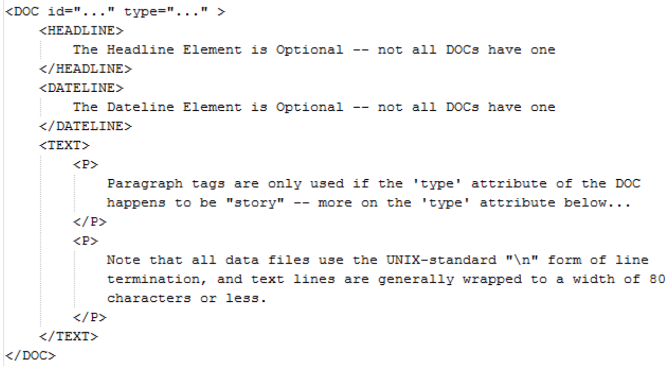
For the validation of our algorithm, we have gathered data from English Gigaword Corpus. The English Gigaword Corpus is a comprehensive archive of newswire text data that has been acquired over several years by the Linguistic Data Consortium (LDC) at the University of Pennsylvania. This is the fifth edition of the English Gigaword Corpus. This edition includes all of the contents in the previous edition (LDC2009T13) as well as new data from the same six sources presented there covering 24-month period of January 2009 through December 2010.

Note that during this period, one of the sources went through a reorganization and name-change: LA Times/Washington Post became Washington Post/Bloomberg. The seven distinct international sources of English newswire included in this edition are the following:

1. Agence France-Presse, English Service
2. Associated Press Worldstream, English Service
3. Central News Agency of Taiwan, English Service
4. New York Times Newswire Service
5. Xinhua News Agency, English Service
6. Los Angeles Times/Washington Post Newswire Service
7. Washington Post/Bloomberg Newswire Service

7.1.1 File Formats and SGML Markup

Each data file name consists of the 7-letter prefix plus another underscore character, followed by a 6-digit date representing the year and month during which the file contents were generated by the respective news source, followed by a ".gz" file extension indicating that the file contents have been compressed using the GNU "gzip" compression utility (RFC 1952). So, each file contains all the usable data received by LDC for the given month from the given news source.
All text data are presented in SGML form, using a very simple, minimal markup structure; all text consists of printable ASCII and whitespace. The file "gigaword.dtd" in the "dtd" directory provides the formal "Document Type Declaration" for parsing the SGML content. The corpus has been fully validated by a standard SGML parser utility (nsgmls), using this DTD file.



```
<DOC id="..." type="..." >
    <HEADLINE>
        The Headline Element is Optional -- not all DOCs have one
    </HEADLINE>
    <DATELINE>
        The Dateline Element is Optional -- not all DOCs have one
    </DATELINE>
    <TEXT>
        <P>
            Paragraph tags are only used if the 'type' attribute of the DOC
            happens to be "story" -- more on the 'type' attribute below...
        </P>
        <P>
            Note that all data files use the UNIX-standard "\n" form of line
            termination, and text lines are generally wrapped to a width of 80
            characters or less.
        </P>
    </TEXT>
</DOC>
```

Figure 7.1: Markup Structure for all data Files

For every "opening" tag (DOC, HEADLINE, DATELINE, TEXT, P), there is a corresponding "closing" tag – always. The attribute values in the DOC tag are always presented within double-quotes; the "id=" attribute of DOC consists of the 7-letter source/language abbreviation (in CAPS), an underscore, an 8-digit date string representing the date of the story (YYYYMMDD), a period, and a 4-digit sequence number starting at "0001" for each date in this way, every DOC in the corpus is uniquely identifiable by the id string.
There are cases where we have assigned a sequence number to a document, and later, we have found out the document is empty or very noisy. In such cases, we have removed the document from the collection, but did not reassign sequence numbers to the rest of the collection for the same day. In addition, there are cases in which data were processed after the bulk of a day's documents; in these cases, additional documents are given sequence

numbers starting at a higher point. As a result there may be some gaps in sequence numbers.

7.1.2 Document Types

The portions of this corpus that were included in the first edition of the English Gigaword corpus have received a uniform treatment in terms of quality control. The new material added in this edition has been initially processed by LDC's daily newswire processing pipeline to create initial mark-up, and then were re-processed follow the design used in the first edition of the Gigaword corpus. The same extent of quality control has been applied to the new material. However, there may be cases where some treatments of data, such as the categorization of DOC units, have changed.
For all of the documents in this corpus, we have applied a rudimentary (and approximate) categorization of DOC units into four distinct "types". The classification is indicated by the " type="string" " attribute that is included in each opening "DOC" tag. The four types are:

1. story : This is by far the most frequent type, and it represents the most typical newswire item: a coherent report on a particular topic or event, consisting of paragraphs and full sentences. As indicated above, the paragraph tag "¡P¿" is found only in DOCs of this type; in the other types described below, the text content is rendered with no additional tags or special characters – just lines of ASCII tokens separated by whitespace.

2. multi : This type of DOC contains a series of unrelated "blurbs", each of which briefly describes a particular topic or event; this is typically applied to DOCs that contain "summaries of todays news", "news briefs in ... (some general area like finance or sports)", and so on. Each paragraph-like blurb by itself is coherent, but it does not bear any necessary relation of topicality or continuity relative to it neighboring sections.

3. advis : (short for "advisory") These are DOCs which the news service addresses to news editors – they are not intended for publication to the "end users" (the populations who read the news); as a result, DOCs of this type tend to contain obscure abbreviations and phrases, which are familiar to news editors, but may be meaningless to the general public. We also find a lot of formulaic, repetitive content in DOCs of this type (contact phone numbers, etc).

4. other : This represents DOCs that clearly do not fall into any of the above types – in general, items of this type are intended for broad circulation (they are not advisories), they may be topically coherent (unlike ”multi” type DOCS), and they typically do not contain paragraphs or sentences (they aren’t really ”stories”); these are things like lists of sports scores, stock prices, temperatures around the world, and so on.

The general strategy for categorizing DOCs into these four classes was, for each source, to discover the most common and frequent clues in the text stream that correlated with the three ”non-story” types, and to apply the appropriate label for the “type=...” attribute whenever the DOC displayed one of these specific clues. When none of the known clues was in evidence, the DOC was classified as a ”story”. // This means that the most frequent classification error will tend to be the use of “ type=”story” ” on DOCs that are actually some other type. But the number of such errors should be fairly small, compared to the number of ”non-story” DOCs that are correctly tagged as such. // Also, since some sources tended to change their delivery methods or format over time, the distribution of non-story types can be seen to vary signficantly by epoch and source. The various ”datastats” tables may be helpful in tracking changes in the nature of the source data (and LDC’s ability to adapt to those changes). // Note that the markup was applied algorithmically, using logic that was based on less-than-complete knowledge of the data. For the most part, the HEADLINE, DATELINE and TEXT tags have their intended content; but due to the inherent variability (and the inevitable source errors) in the data, users may find occasional mishaps where the headline and/or dateline were not successfully identified (hence show up within TEXT), or where an initial sentence or paragraph has been mistakenly tagged as the headline or dateline.

7.1.3 Document Quantities

The ”docs” directory contains a set of plain-text tables (datastats*) that describe the quantities of data by source and month (i.e. by file), broken down according to the four ”type” categories. The overall totals for each source are summarized below. Note that the ”Totl-MB” numbers show the amount of data you get when the files are uncompressed (i.e. approximately 15 gigabytes, total); the ”Gzip-MB” column shows totals for compressed file sizes as stored on the DVD-ROM; the ”K-wrds” numbers are simply the number of whitespace-separated tokens (of all types) after all SGML tags are eliminated.

Table 7.1: Data for Advis data set from different news agencies

	Text in MB's	**K-Words**	**No Of Docs**
Advis			
afp-eng	152	21675	54414
apw-eng	181	27382	39289
cna-eng	0	24	85
ltw-eng	88	14132	28987
nyt-eng	599	95606	157500
wpb-eng	7	1233	2570
xin-eng	12	1920	7522
Total	**1039**	**161972**	**290367**
Multi			
afp-eng	86	13101	37089
apw-eng	244	40000	58570
cna-eng	23	3786	19415
ltw-eng	19	3086	7020
nyt-eng	124	20435	33183
wpb-eng	0	0	0
xin-eng	134	21473	91997
Total	**630**	**101881**	**247274**
Other			
afp-eng	125	18869	133981
apw-eng	337	47208	273377
cna-eng	2	213	1935
ltw-eng	1	228	1063
nyt-eng	116	17681	26601
wpb-eng	1	308	681
xin-eng	130	18448	161724
Total	**712**	**102955**	**599362**

we have utilised this data for testing our algorithm how ever in production environment it is taking data from RSS Feed and Web crawlers.

7.2 Testing

The testing of these technique is not straight forward, because the quality of results are very hard to evaluate there are few techniques that are generally used to validate results.

1. User Inspection
   (a) Study Centroids and Spreads
   (b) Rules for Decision Trees
   (c) For Text documents, one can read some document in cluster
2. confusion matrix can be used to evaluate results and can compute
   (a) Entropy
   (b) Purity
   (c) Precision
   (d) Recall
   (e) F-Score

Testing of above defined technique is done by comparing the data annotated manually by the industry specialist who usually perform these task every day as part of their job.

7.2.1 Manual Marking of the data

In Manual Marking of data the small portion from above corpus has been chosen randomly. Then those document are pushed into the manual marking system, which were marked by the Experts. we have been given 3 person for 10 days, for them we have developed a custom software's for marking entities, grouping of document based on benchmarking system (Benchmarking is system adopted by our company in which entries marked by data entry operator is validated by different techniques), and Relationship extraction. After the manual marking and relationship extraction, the extracted data was transferred to Quality Check process (commonly known as QC process) in which all marking is validated by validation team. After QC process data is then transferred to Reporting Module for validation of algorithm.

Figure 7.2: Lab used for marking data manually

7.2.2 Results for Entity Extraction

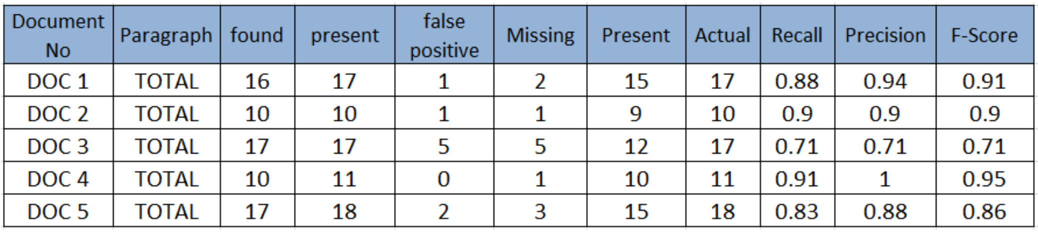
Bellow diagram shows the results of Entity Extraction process of the Algorithm. for reference I have shown results from 5 random documents.

7.2.3 Results for Document Grouping

Bellow diagram shows the results of Document grouping process of the Algorithm. Bellow diagram show the result from the automated system which were compared with manual marking to compute bellow results

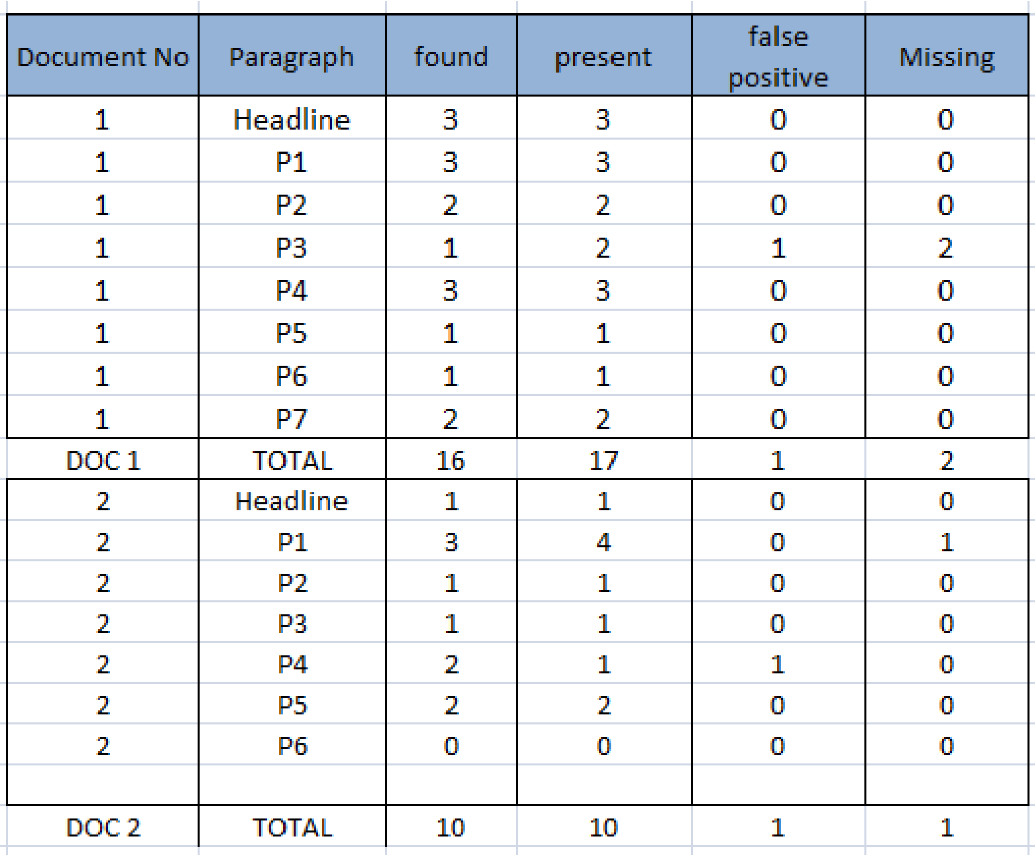
7.2.4 Results for Relationship Extraction

Results of Relation Extraction is shown in bellow diagram



Document No	Paragraph	found	present	false positive	Missing	Present	Actual	Recall	Precision	F-Score
DOC 1	TOTAL	16	17	1	2	15	17	0.88	0.94	0.91
DOC 2	TOTAL	10	10	1	1	9	10	0.9	0.9	0.9
DOC 3	TOTAL	17	17	5	5	12	17	0.71	0.71	0.71
DOC 4	TOTAL	10	11	0	1	10	11	0.91	1	0.95
DOC 5	TOTAL	17	18	2	3	15	18	0.83	0.88	0.86

Figure 7.3: Results of Entity Extraction



Document No	Paragraph	found	present	false positive	Missing
1	Headline	3	3	0	0
1	P1	3	3	0	0
1	P2	2	2	0	0
1	P3	1	2	1	2
1	P4	3	3	0	0
1	P5	1	1	0	0
1	P6	1	1	0	0
1	P7	2	2	0	0
DOC 1	TOTAL	16	17	1	2
2	Headline	1	1	0	0
2	P1	3	4	0	1
2	P2	1	1	0	0
2	P3	1	1	0	0
2	P4	2	1	1	0
2	P5	2	2	0	0
2	P6	0	0	0	0
DOC 2	TOTAL	10	10	1	1

Figure 7.4: Results of Entity Extraction

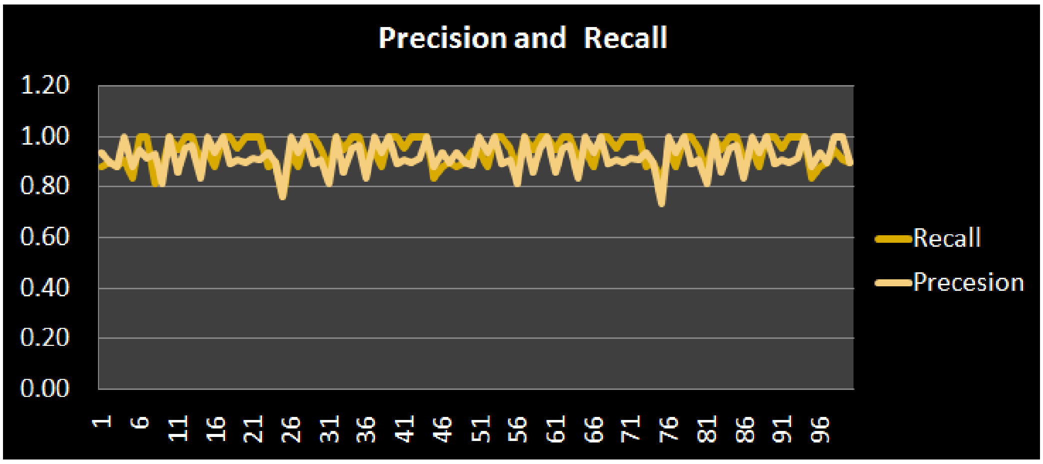
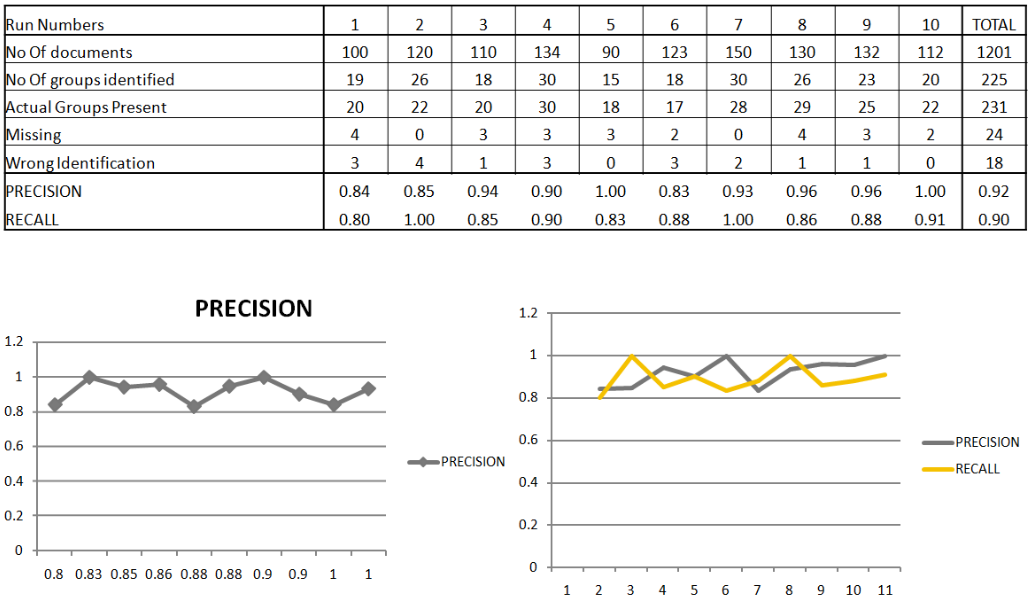
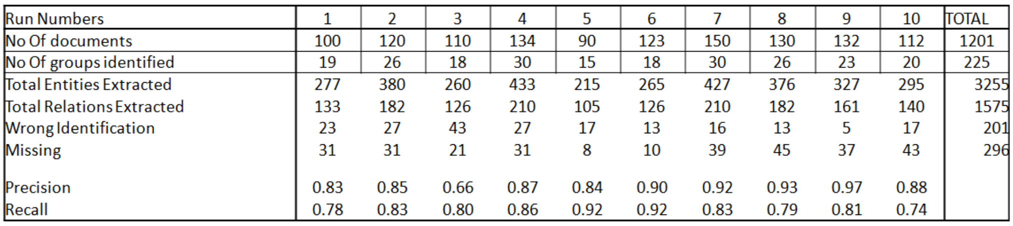


Figure 7.5: Precision and Recall Plot for Entity Extraction



Run Numbers	1	2	3	4	5	6	7	8	9	10	TOTAL
No Of documents	100	120	110	134	90	123	150	130	132	112	1201
No Of groups identified	19	26	18	30	15	18	30	26	23	20	225
Actual Groups Present	20	22	20	30	18	17	28	29	25	22	231
Missing	4	0	3	3	3	2	0	4	3	2	24
Wrong Identification	3	4	1	3	0	3	2	1	1	0	18
PRECISION	0.84	0.85	0.94	0.90	1.00	0.83	0.93	0.96	0.96	1.00	0.92
RECALL	0.80	1.00	0.85	0.90	0.83	0.88	1.00	0.86	0.88	0.91	0.90

Figure 7.6: Results of Document Grouping



Run Numbers	1	2	3	4	5	6	7	8	9	10	TOTAL
No Of documents	100	120	110	134	90	123	150	130	132	112	1201
No Of groups identified	19	26	18	30	15	18	30	26	23	20	225
Total Entities Extracted	277	380	260	433	215	265	427	376	327	295	3255
Total Relations Extracted	133	182	126	210	105	126	210	182	161	140	1575
Wrong Identification	23	27	43	27	17	13	16	13	5	17	201
Missing	31	31	21	31	8	10	39	45	37	43	296
Precision	0.83	0.85	0.66	0.87	0.84	0.90	0.92	0.93	0.97	0.88	
Recall	0.78	0.83	0.80	0.86	0.92	0.92	0.83	0.79	0.81	0.74	

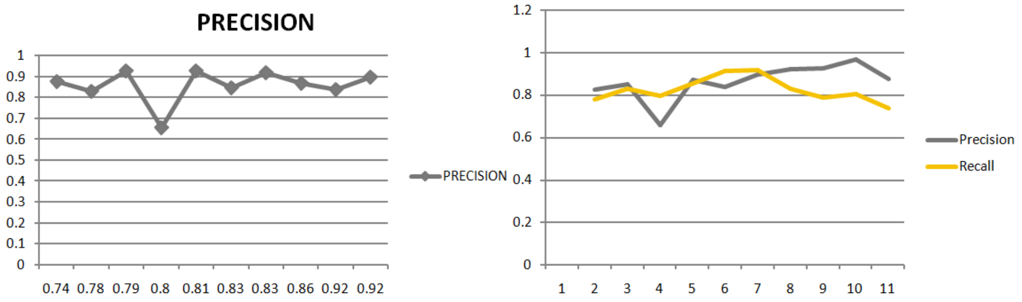


Figure 7.7: Results of Relationship Extraction

Chapter 8

Conclusion and Future Work

Although this technique extracts very less entities and relationships but their validity will always be high and will help to organize the news articles. There is a need to test this technique extensively and find out the possible improvement to increase the Recall of the proposed system. The possible extension of this work will be generation of context matrix for the entities explaining whether the selected entities are used as object or source of the relation which will help further in order to classify the news articles. Bellow are the main area's where this research can be used in future

1. Popular Topic Extraction
2. Images of Parties and personalities portrait
3. Characteristics and Parameters of reported Events
4. Frequency of Appearance
5. Choice Model

8.1 Popular Topic Extraction

Popular Topics are very critical for any PR (Personal Relation) activities, any PR organization for entities requires the popular topics before they define any strategy. In this area one can search the most reported entities for a person and extract what are the popular topics that are been reported. This can be done by searching most relations reported for particular entity and find out popular topics.

8.2 Images for Parties and Personalities Portrait

Major activities that are related to marketing are associated with the images. All Marketing people basically create image of different personalities and brands before they offer them different propositions. Images of Parties and Personalities can be second future enhancement of this research where one can categorize the Relations into different groups and assigned those groups to good and bad image, and offer different reports on it.

8.3 Characteristics and Parameters for Reported Events

When well know entities (Personalities, Parties, or Brands), comes to some crises, they need to analyse all the previous characteristics and parameters which are been portrait to public. These things help them to define new strategies. All these can be done third possible extension of current research where you categorize the relationships to different areas and associate different scores to define the possible available strategies.

8.4 Frequency of Appearance

For some brands and personalities it is very important that they must appear in different news and different areas. The Frequency of appearance is also important when some new entrants are entering into the market and they need to know what should be the minimum Frequency of appearance that need to maintain in order to be in, in the market. This can be done by simple count but further elaboration

Bibliography

[1] Thahir P Mohamed, Estevam R Hruschka Jr and Tom M Mitchell *"Discovering Relations between noun categories",* EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1447-1455, 2011.

[2] Jana Diesner and Kathleen M. Carley, *"Conditional random fields for entity extraction and ontological text coding"*, Springer Science + Business Media LLC, 2008

[3] Sanjay Agrawal, Kaushik Chakrabarti, Surajit Chaudhuri and Venkatesh Ganti, *"Scalable Adhoc Entity Extraction From Text Collections"* Proceedings of the VLDB, 2008.

[4] Ian H.Witten, Zane Bray, Malika Mahoui and W.J. Teahan, *"Using Language Models for generic entity Extraction"* ICML Workshop on Text, 1999.

[5] Marius Pasca , Dekang Lin , Jeffrey Bigham , Andrei Lifchits and Alpa Jain, *"Organizing and Searching theWorldWideWeb of Facts - Step One: The One Million Fact Extraction Challenge"*, in National Conference on Artifical Intelligence, 2006.

[6] Eugene Agichtein and Luis Gravano, *"Snowball: Extracting relations from large plain text collections"*, in Fifth International conference on Digital Libraries, 2000.

[7] Andrew Carlson, Justin Betteridge, Estevam R. Hruschka Jr. and Tom M. Mitchell, *"Coupling Semi-Supervised Learning of Categories and Relationships"*, in NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, 2009.

[8] Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni, *"Open Information extraction from the Web"*, in IJCAI, 2007.

[9] Takaaki Hasegawa, Satoshi Sekine and Ralph Grishman, *"Discovering relation among named entities from large corpora"*, in Proceediing of the 42nd Annual Meeting on Association for ComputaionalLinguitics, 2005.

[10] Min Zhang, Jian Su1, Danmei Wang1, Guodong Zhou, and Chew Lim Tan, *"Discovering relation among named entities from large raw corpus using similarity-bases clustering"*, in IJCNLP'05 Proceedings of the Second international joint conference on Natural Language Processing Pages 378-389.

[11] G.Lindzey, E.Aronson, *"The Handbook of Social Psychology , 2nd Edition Pages 569-692."*, by New Delhi:Amerind Publishing Co.

[12] Wimmer R.D., Dominick J.R. *"Mass media research 4rth Edition"*, California: Wadsworth

[13] Lasswell, H.D., Leites, N., Associates *"Language of politics"*, by Cambridge:MIT Press

[14] Argamon, Kenneth Bloom and Navendu Garg and Shlomo *"Extracting Appraisal Expressions, 308-315"*, Proceedings of NAACL HLT, 2007

[15] Jin-Dong KIM, Tomoko OHTA, Yoshimasa TSURUOKA, Yuka TATEISI *"Introduction to the Bio-Entity Recognition Task at JNLPBA"*, Proceedings of JNLPBA, 2004

[16] Marius A. Pasca, Sanda M. Harabagiu *"High Performance Question/Answering, Pages 366-374"*, Proceedings of SIGIR, 2001

[17] Kai-Yuh Hsiao and Peter Gorniak and Deb Roy *"Systems, NetP: A Network APIforBuilding Heterogeneous Modular Intelligent"*, AAAI 2005 Workshop in Modular Construction of Human-Like Intelligence

[18] Wilbur, Jimmy Lin W. John *"Syntactic sentence compression in the biomedical domain: facilitating access to related articles, Pages 393-424"*, Proceedings of Inf Retrieval Springer, 2007

[19] Rana Forsati, Mehrdad Mahdavi b, Mehrnoush Shamsfarda, Mohammad Reza Meybodi *"Efficient stochastic algorithms for document clustering, Pages 269291"*, Proceedings of Online Fuzzy Machine Learning and Data Mining, 2012

[20] Manjeet Rege , Josan Koruthu and Reynold Bailey *"On Knowledge-Enhanced Document Clustering, Pages 11"*, Journal of Information Retrieval Research, 2013

[21] Tanmay Basu, C.A. Murthy *"A New Hierarchical Approach for Document Clustering"*, Journal of Pattern Recognition Research, 2013

[22] Carolina Abreu, Flvio Costa, Lacio Santos, Lucas Monteiro, Luiz Fernando Peres de Oliveira, Patrcia Lustosa, Li Weigang *"Entity Extraction within Plain-Text Collections WISE 2013 Challenge - T1: Entity Linking Track, Pages 491-496"*, Web Information Systems Engineering WISE 2013

[23] Nathanael Chambers and Dan Jurafsky *"Template-based information extraction without the templates, Pages 976-986"*, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, 2011

[24] Giuseppe Rizzo, Raphal Troncy, Sebastian Hellmann, Martin Bruemmer *"NERD meets NIF: Lifting NLP Extraction Results to the Linked Data Cloud"*, Proceedings of LDOW, 2012