Keyphrase Extraction and Enrichment for News Media

A DISSERTATION
submitted towards the partial fulfillment of the
requirements for the award of the degree
of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

Submitted by

NIKITA JAIN

Under guidance of

Dr. DHAVAL PATEL,
ASSISTANT PROFESSOR

Department Of Computer Science And Engineering
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE – 247667 (INDIA)
May, 2016

CANDIDATE'S DECLARATION

I declare that the work presented in this dissertation with title "**Keyphrase Extraction and Enrichment for News Media**" towards the fulfilment of the requirement for the award of the degree of **Master of Technology** in **Computer Science & Engineering** submitted in the **Dept. of Computer Science & Engineering**, **Indian Institute of Technology, Roorkee**, India is an authentic record of my own work carried out during the period **from June 2015 to May 2016** under the supervision of **Dr. Dhaval Patel**, Assistant Professor, Dept. of CSE, IIT Roorkee.
The content of this dissertation has not been submitted by me for the award of any other degree of this or any other institute.

DATE : .... SIGNED: ....

PLACE: .... (NIKITA JAIN)

CERTIFICATE

This is to certify that the statement made by the candidate is correct to the best of my knowledge and belief.

DATE : .... SIGNED: ....

PLACE: .... (DR. DHAVAL PATEL)
ASSISTANT PROFESSOR
DEPT. OF CSE, IIT ROORKEE

PUBLICATION

Part of the work presented in this dissertation has been **accepted to be published** in the following conference:

Nikita Jain, Swati Gupta, Dhaval Patel. **E^3: Keyphrase based News Event Exploration Engine**, In 27^{th} ACM Conference on Hypertext and Social Media (**HT, 2016**).

Acknowledgements

I would like to express my sincere gratitude to my mentor **Dr. Dhaval Patel**, for his invaluable advices, discussions and constructive criticisms which certainly enhanced my knowledge, skills and helped me to bring out the best possible of my potential. I want to thank him for his continuous support and motivation in dealing with difficulties in professional and personal life. Also, thanks for being a great friend.

I thank Smit Parsania, Achuth Kandikunnta, Deepak Jannarapu, Manikanta Nallagatla and Tarun Kumar Tella for their help with keyphrase baselines, preparing dataset and insighful discussions. Also, I want to thank Palash Agrawal, Sachin Aggarwal, Samayak Choudhary, Saurabh Jain, Saurabh Verma, Shivam Dwivedi for helping me in developing a demonstration system.

I also want to thank Tarun Kumar Kulshrestha for being a selfless friend, for providing best suggestions and for always willing to help me in any problem both on and off water. Finally, I thank all my friends and lab mates for being very good friends, for a lot many good discussions and for being along with me in the whole journey and making it so joyful.

Finally, a huge thanks to my **Parents and Siblings** who have taught me the value of a good education. Thanks for their continuous support and encouragement through my life. It would have not been possible to reach here without their driving force.

Abstract

As newswire data is growing continuously at a very fast pace, the need for techniques generating instantly digestible and concise format news information is emerging. My research goal in dissertation thesis is to develop models that can automatically extract summarized and interesting news information. Aiming to solve the problem of low engagement time of news audience and several other news journalism problem.

There has been great progress in automatically extraction and generation of facts, trivias and other interesting information from news media data such as trivia generation, event detection, headlines generation, sentiment analysis, question-answering systems. However, in-spite of these approaches the news audience engagement time is still low. Also, these solutions are often based on different learning models. My goal is to develop general and scalable algorithms that can work over any language, any domain and any media format having textual content. The model (E^3) in this thesis address these shortcomings. They provide effective and efficient keyphrases for multilingual and multi-format news data. They provide a set of features to rank the set of keyphrases. Furthermore, a method is provided to enrich the extracted keyphrases by finding the types and input query related information like role played by person entity. This kind of information is very helpful in cases where many people, multiple organization and multiple location are mentioned. As it is very difficult for a reader to keep track of all the mentioned entities. Henceforth, readers often losses interest in the news concept and the network traffic gets lost. Also, we have specifically chosen the keyphrase based summary as they provide a high-level overview of news data in a short span of time with little effort.

We have evaluated our unsupervised system E^3 on varying input queries, from general topics (E.g. Election) to specific topics (E.g. Bihar Election) to demonstrate the efficiency and effectiveness of our keyphrase extraction and keyphrase enrichment method over existing state-of-the-art. Our experimental results show that E^3 performs significantly better than the defined baselines on seven different parameters. We also investigate the effect of the use of linguistic and syntactical features in keyphrase extraction, with an user case study and found that our system is fairly robust.

Keywords: *Automated News Summarization, Keyphrase Extraction, Keyphrase Enrichment, Keyphrase Ranking, Natural Language Processing.*

Dedication

I lovingly dedicate this thesis and all my achievements
to

my parents, my sister, my brother and my mentors

for their endless love, support and encouragement.

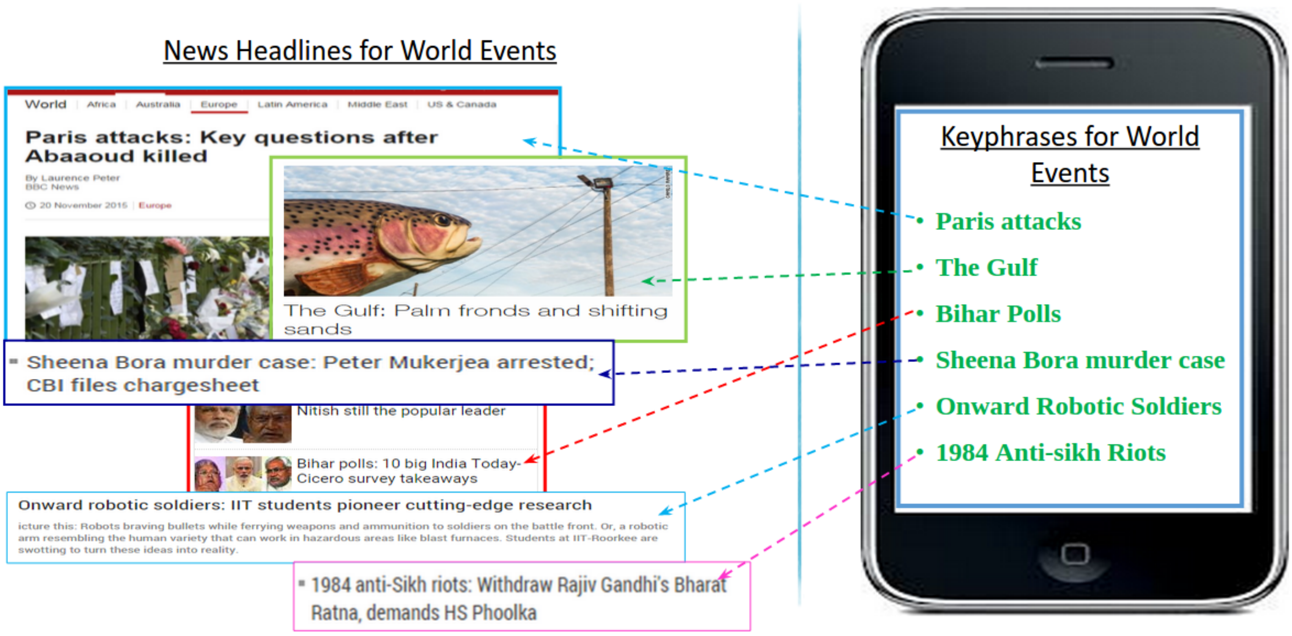
TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

List of Tables

E^3 for Evolving World Events



News Headlines for World Events

World | Africa | Australia | Europe | Latin America | Middle East | US & Canada

Paris attacks: Key questions after Abaaoud killed

By Laurence Peter
BBC News

20 November 2015 | Europe

The Gulf: Palm fronds and shifting sands

Sheena Bora murder case: Peter Mukerjea arrested; CBI files chargesheet

Nitish still the popular leader

Bihar polls: 10 big India Today-Cicero survey takeaways

Onward robotic soldiers: IIT students pioneer cutting-edge research

icture this: Robots braving bullets while ferrying weapons and ammunition to soldiers on the battle front. Or, a robotic arm resembling the human variety that can work in hazardous areas like blast furnaces. Students at IIT-Roorkee are swotting to turn these ideas into reality.

1984 anti-Sikh riots: Withdraw Rajiv Gandhi's Bharat Ratna, demands HS Phoolka

Keyphrases for World Events

- Paris attacks
- The Gulf
- Bihar Polls
- Sheena Bora murder case
- Onward Robotic Soldiers
- 1984 Anti-sikh Riots

CHAPTER 1

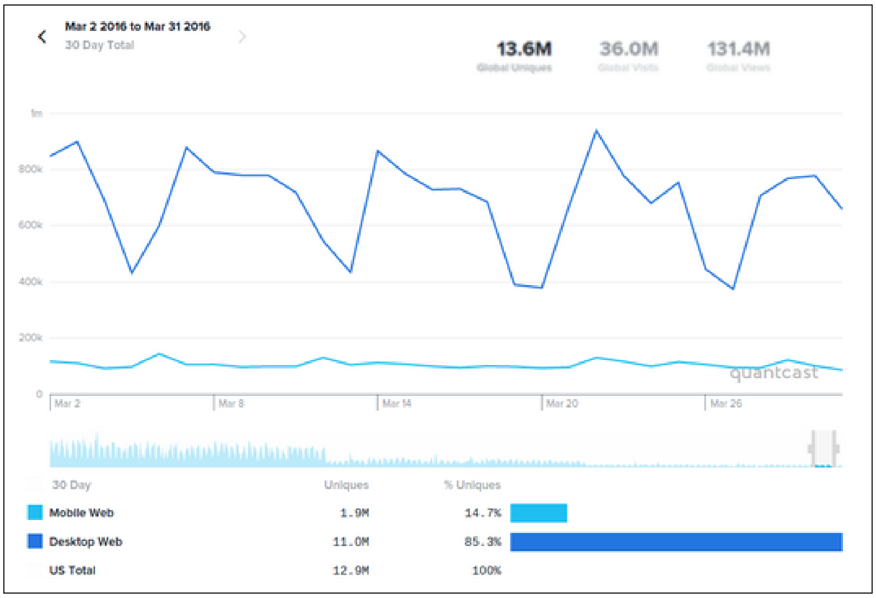
Introduction

News media is a place where ideas, events and opinions are turgid with a wide range of data formats. Such as long text (articles), short text (headlines, tweets, hashtags), videos and others. According to *Reuters*' report also, the content published by news medias is Big Data. The news is continually being published in multiple languages by multiple sources in multiple data formats. Also, the number of news audience is very large and growing significantly for the mobile device users. Figure 1.1 shows that *NBC News* source has itself over 13 million unique visitors each month[1] and around 2 million unique mobile web US viewers. However the amount of time a user gives to a news application is only about 3 minutes per visit, according to the survey conducted by PEW[2]. One of the major reason for news audience's low engagement time is information overload. A news reader itself has to analyze such huge amount of information. Thus, the news audience is compelled to make random *picks* among the search results. Consequently, significant information related to the input query may lose. Currently, there is a high need for instantly digestible, precise and concise information, providing a broad overview of news data.

In view of this, popular news aggregators like Google News, Yahoo News indexes

[1] https://www.quantcast.com/nbcnews.com

[2] http://www.journalism.org



30 Day	Uniques	% Uniques
🟦 Mobile Web	1.9M	14.7%
🟦 Desktop Web	11.0M	85.3%
US Total	12.9M	100%

Figure 1.1: NBC News monthly news traffic (`source: Quancast`)

the information publish by online news media. They provide ergonomic solutions for news exploration, such as news categorization (on topics like Sports, Politics), location specific news, recent and popular news, etc. However, such solutions are adhoc, as they do not provide *quick insights* about the query. News audience is still required to spend more time to understand the happenings and evolution of the news events. For instance, on 26 January 2016 Google Search suggested 227 million results for 'Paris Attack'. Whereas Google News returned just 5% results for the same query. However the 5% of 227 million is not really small, it is about 46 million URLs. And that too the results were multilingual. The reason behind such large number of results is, the search engines do not semantically analyse and summarize the texts. They just return the complete document details using syntactical features like frequency, inverse document frequency, word position and others.

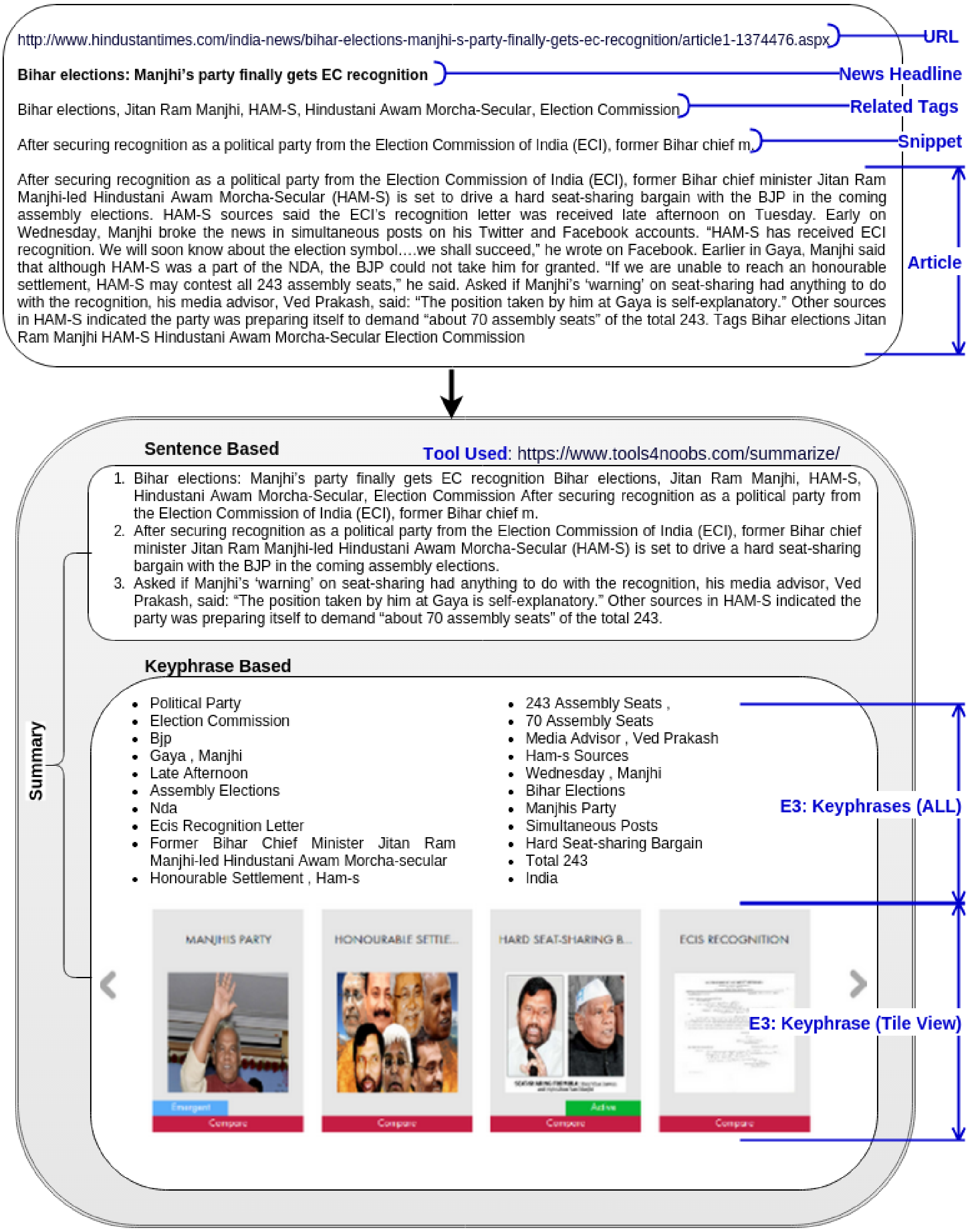
Hence, there is an emerging need for techniques enabling news audience to asses the relevant news information faster and easier. For this problem query relevant summarization system seems a promising solution, as summaries are not cluttered with details. With the help of summary a user can quickly recognize the content and easily filter the irrelevant news stories.

Traditional work done to summarise news content is by generating trending headlines and snippets, having an average length of 10 tokens per headline (for detailed analysis of news headlines, see Appendix A). However, users are seeking instantly digestible news information. They want news in even more compact form, that is sufficient enough to provide all the relevant and important information. Although such general summaries do not work well for all, as some user may be interested in some other news concepts and entities whose details may get missed on standardized summary results. For instance, we passed a news article to an online summarizer tool[3], and its results shown in Figure 1.2. However, if a news reader is interested in reading about 'Kerala Temple', then they may skip the article by reading Figure 1.2(a) sentence based summary. As a result, prominent information may lose. To mitigate the problem we present a keyphrase based news summary, as the information provided by set of keyphrases is usually self contained and complete. Keyphrase mining is the state-of-the-art approach for such scenarios. Here keyphrase is a group of words, describing some prominent information about the input document. However, keyphrase follows no standard structure or grammatical rules like sentences does. Keyphrases are just a permutation of group of words which is contextually meaningful. For example, 'Bihar Election Result', 'Bihar BJP Candidate', 'Grand Alliance'. The result obtained by our keyphrase based system E^3 is shown in Figure 1.2(b). Also, we can observe from the Figure 1.2 that it is comparatively faster and easier to gather knowledge via keyphrase than sentence based news summary.

In this thesis, we aim to convert the voluminous News data into small group tokens which are easy to understand and conveys maximum knowledge to news audiences. Also, our aim is to develop techniques using minimum computational resources, so that applications can be deployed even on low processing devices like mobile phones.

Applications: The Keyphrase extracted from news corpus can be used for news classification, index generation, query expansion, meta-tag suggestion, hashtag generation, disambiguation through related keyphrase sets, and many more. Companies earning on the basis of Ad revenues, can also use appropriate keyphrases to bring related users to the related news stories.

[3] `https://www.tools4noobs.com/summarize/`



http://www.hindustantimes.com/india-news/bihar-elections-manjhi-s-party-finally-gets-ec-recognition/article1-1374476.aspx

Bihar elections: Manjhi's party finally gets EC recognition

Bihar elections, Jitan Ram Manjhi, HAM-S, Hindustani Awam Morcha-Secular, Election Commission

After securing recognition as a political party from the Election Commission of India (ECI), former Bihar chief m

After securing recognition as a political party from the Election Commission of India (ECI), former Bihar chief minister Jitan Ram Manjhi-led Hindustani Awam Morcha-Secular (HAM-S) is set to drive a hard seat-sharing bargain with the BJP in the coming assembly elections. HAM-S sources said the ECI's recognition letter was received late afternoon on Tuesday. Early on Wednesday, Manjhi broke the news in simultaneous posts on his Twitter and Facebook accounts. "HAM-S has received ECI recognition. We will soon know about the election symbol….we shall succeed," he wrote on Facebook. Earlier in Gaya, Manjhi said that although HAM-S was a part of the NDA, the BJP could not take him for granted. "If we are unable to reach an honourable settlement, HAM-S may contest all 243 assembly seats," he said. Asked if Manjhi's 'warning' on seat-sharing had anything to do with the recognition, his media advisor, Ved Prakash, said: "The position taken by him at Gaya is self-explanatory." Other sources in HAM-S indicated the party was preparing itself to demand "about 70 assembly seats" of the total 243. Tags Bihar elections Jitan Ram Manjhi HAM-S Hindustani Awam Morcha-Secular Election Commission

Sentence Based

Tool Used: https://www.tools4noobs.com/summarize/

1. Bihar elections: Manjhi's party finally gets EC recognition Bihar elections, Jitan Ram Manjhi, HAM-S, Hindustani Awam Morcha-Secular, Election Commission After securing recognition as a political party from the Election Commission of India (ECI), former Bihar chief m.
2. After securing recognition as a political party from the Election Commission of India (ECI), former Bihar chief minister Jitan Ram Manjhi-led Hindustani Awam Morcha-Secular (HAM-S) is set to drive a hard seat-sharing bargain with the BJP in the coming assembly elections.
3. Asked if Manjhi's 'warning' on seat-sharing had anything to do with the recognition, his media advisor, Ved Prakash, said: "The position taken by him at Gaya is self-explanatory." Other sources in HAM-S indicated the party was preparing itself to demand "about 70 assembly seats" of the total 243.

Keyphrase Based

- Political Party
- Election Commission
- Bjp
- Gaya , Manjhi
- Late Afternoon
- Assembly Elections
- Nda
- Ecis Recognition Letter
- Former Bihar Chief Minister Jitan Ram Manjhi-led Hindustani Awam Morcha-secular
- Honourable Settlement , Ham-s
- 243 Assembly Seats ,
- 70 Assembly Seats
- Media Advisor , Ved Prakash
- Ham-s Sources
- Wednesday , Manjhi
- Bihar Elections
- Manjhis Party
- Simultaneous Posts
- Hard Seat-sharing Bargain
- Total 243
- India

Figure 1.2: News article summarization (a) Sentence based (b) Keyword based

1.1 Technical Challenges

However, there are some technical challenges that make using keyphrases for news event exploration far from being straightforward. In-spite of keyphrase extraction from multilingual and multi-data format news, some other major challenges are:

- Distinguishing informing news keyphrase from other non-informing keyphrases.
- Identifying duplicates among the keyphrases to minimize redundancy. For instance, the two keyphrases are very similar to each other: "Bihar Election Results", "Bihar Polling Results".
- Ordering the selected keyphrase such that the keyphrase based summary seems interesting to news audiences.

1.2 Thesis Contribution and Outline

The major research contribution in our effort to leverage heterogeneous news data for query based keyphrase mining to improve the news exploration task are:

- We present E^3, a novel keyphrase based news event exploration engine for obtaining a global description of *heterogeneous news*.
- We demonstrate the importance of *linguistic and syntactical features* of text. On comparison with state-of-the-art systems results show that E^3 has significantly improved over keyphrase extraction in terms of informativeness and interestingness.
- We introduce the notion of *novelty, activeness and recency* in time interval as a means for keyphrase ranking along with the frequency of the keyphrase.
- We discovered *temporal roles* played by frequently associated entities in news media and their time period of involvement.
- We provide an effective method to disambiguate entities present in news corpora with no requirement of external knowledge base.

- We propose a new mechanism for evaluating the domain Independence of keyphrase extraction process. For this we use *Plot Keywords from IMDB*, and further use it as testing dataset.

Chapter 2 gives background details for keyphrase extraction and news event exploration approaches. Chapter 3 is an introductory chapter of the E^3 model that introduces the preliminaries of E^3 and keyphrase generation and enrichment process for multilingual, heterogeneous data format news media. The Chapter 4 discusses the linguistic and synthetic Keyphrase Extraction algorithm and Keyphrase Enrichment algorithm in detail. Chapter 5 presents the experiments and results, along with an evaluation of approaches discussed in Chapter 2. Next, in Chapter 6 we demonstrate our system E^3 and gives an overview of our dataset collected for various specific and general input queries. Finally, Chapter 7 concludes the thesis and discusses directions the work can be extended.

CHAPTER 2

Literature Review

This chapter provides the necessary background to explore the subject of this thesis. The review starts with keyphrase extraction which is the main axis of the thesis. It then continues with a discussion of existing news exploratory systems.

2.1 Keyphrase Extraction

Keyphrases helps in analyzing large amount of information quickly, by allowing the readers to skim irrelevant text. To extract a set of keyphrases from textual data various modern techniques, such as Topic Learning [4] and Keyphrase mining [5], have been developed. However, in comparison to topic learning, the output of Keyphrase mining provides a wide range of informative and important phrases. As Keyphrase mining provides good coverage of all the major topics [6].

For efficient Keyphrase extraction, several highly subjective approaches are there. However, still many domains follow manual keyphrase extraction task [7]. For instance, research paper keywords are assigned by researcher, webpage keywords are assigned by SEO specialist, news journalist assigns related tags to news articles. As assigning keyphrases manually is a tedious task requiring complete knowledge of concept. Consequently the great majority of news articles comes without related tags. Hence, automatic extraction of keyphrases will save a lot of time and effort by alleviating the need of manual assignment of keywords.

Our proposed work is also keyphrase centric, as recent literature has shown that Keyphrase mining is the state-of-the-art solution to summarize large documents. Some of the recent automatic keyphrase extraction techniques developed are discussed below.

Keyphrase Extraction Algorithm: KEA [8] employs a supervised Naive Bayes classification model to extract important phrases from a single document. The model is trained using features like term frequency, inverse document frequency, phrase position and probability of a phrase to be a keyphrase in the training dataset. The technique has been improved by Multi-purpose Automatic Topic Indexing (**Maui**) [9] by incorporating the additional set of Wikipedia based features and changing the classifier with bagged decision trees.

Microsoft Web N-gram Language Models: Micro-ngram [10] is also a supervised technique trained on the real world web scale indexed data of Microsoft Bing. The multilingual model is trained to break the input sentence into phrases. Where each phrase has a maximum likelihood of occurrence in the training corpus. Micro-ngram does not directly use static distribution of the corpus. They first smooth the weights (probabilities) of input strings to minimize the prediction errors, so that unseen words do not get assigned with zero probability. They proved that Micro-ngram outperforms the other N-gram models trained on Gigaword corpus and Google Web 1T N-gram corpus on the various natural language processing and web search tasks.

Topical Phrase Mining from Text Corpora: Different from aforementioned supervised methods, ToPMine [11] follows a frequent pattern mining based unsupervised approach. ToPMine discovers phrases that are frequent as well as statistically significant. The approach first converts the bag of words into bag of phrases and then the semantically related bag of phrases are connected to a common topic name, using topic modelling approach. The technique has been improved by **SegPhrase** [12] by incorporating the additional phrase quality estimation and using Random Forest classifier. SegPhrase estimate the quality of phrases using popularity (word frequency), concordance (point-wise mutual information, point-wise KL divergence), informativeness (average IDF) and completeness.

2.1.1 Limitation

The aforementioned approaches performance is good for standard passage text, however their results do not scale well for extracting keyphrases from the news media corpus. They do not provide optimal summary for news data in terms of importance and interestingness. This has happened because the writing style of both passage text and news data, especially headlines differ a lot (see Appendix A). Thus, the systems fail to identify interesting, as well as complete information about the news corpora. As a result why and how the news concept are connected to the input query led to poor results. For instance, given input query 'Election' and a connected concept 'Nitish Kumar', even a supervised system like KEA did not able bring connecting information like 'chief minister candidate'. Also identifying the boundaries of words in news media data where string are not in correct sentential and grammatical form, is another challenge with the existing systems.

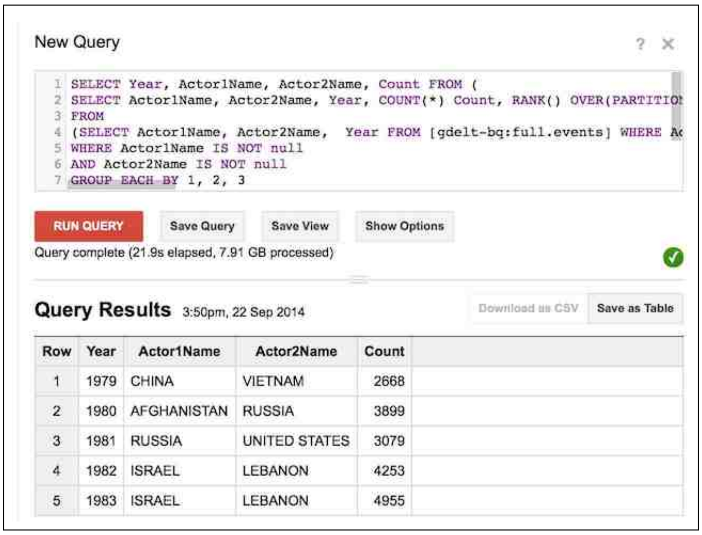
To achieve the above two goals, i.e., news query based keyphrase mining; and using knowledge from multiple data formats in a joint way to improve the news exploration task, we propose a keyphrase mining approach. And bring together the linguistic and syntactical knowledge from different data forms to combine the information in a synergistic way. We also rank the generated list of keyphrases on the basis of phrase novelty, activeness, recency in time period and frequency values according to the input query timeline.

2.2 News Event Exploration

Popular news aggregators like Google news, Yahoo news indexes the information published by online news media and enables users to explore the keyword based news corpus in real time. Since keyword based document retrieval fetches large amount of information, such search engines provide only adhoc solutions for news exploration. They do not provide quick insights and user is still required to spend more time to understand the event happenings and evolution of news. To mitigate the problems with exploration of large amount of news documents, researchers has developed event centric news exploration systems, such has Global Database of Events, Language, and Tone (GDELT) [13], EventRegistry [14], Searching with

Strings, Things, and Cats (STICS) [15] and Europe Media Monitor (EMM) [16].

Global Database of Events, Language, and Tone: GDELT contains more than 200 million news events records with global coverage from 1979 to the present, using news reports from multiple newswire sources. GDELT stores the events records in CAMEO format, capturing two actors and the action performed by $Actor_1$ upon $Actor_2$. Figure 2.1 shows an instance of query made into GDELT database and the results obtained from them.



```
SELECT Year, Actor1Name, Actor2Name, Count FROM (
SELECT Actor1Name, Actor2Name, Year, COUNT(*) Count, RANK() OVER(PARTITIO
FROM
(SELECT Actor1Name, Actor2Name,  Year FROM [gdelt-bq:full.events] WHERE Ac
WHERE Actor1Name IS NOT null
AND Actor2Name IS NOT null
GROUP EACH BY 1, 2, 3
```

Query complete (21.9s elapsed, 7.91 GB processed)

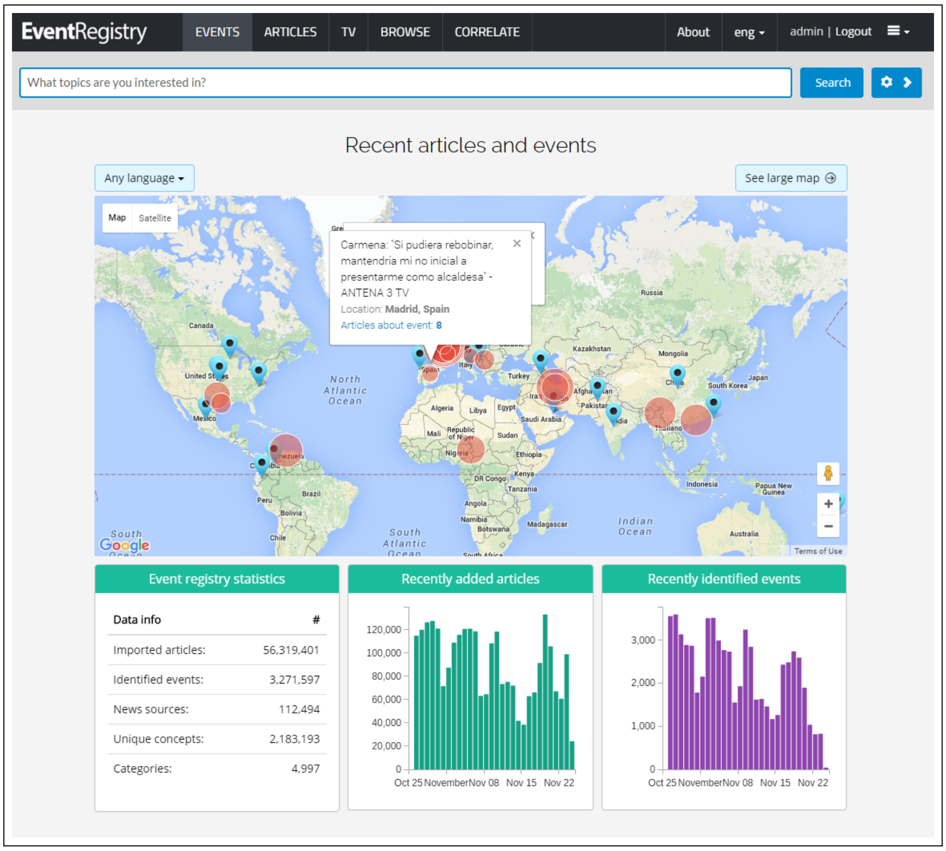
Query Results 3:50pm, 22 Sep 2014

Row	Year	Actor1Name	Actor2Name	Count
1	1979	CHINA	VIETNAM	2668
2	1980	AFGHANISTAN	RUSSIA	3899
3	1981	RUSSIA	UNITED STATES	3079
4	1982	ISRAEL	LEBANON	4253
5	1983	ISRAEL	LEBANON	4955

Figure 2.1: GDelt Database query and result format (`source: [1]`)

EventRegistry: Form groups of cross-lingual articles to describe similar events and represent them as a single event. Using vector space model of article title, body and detected named entities, the clusters are formed. From articles in each cluster, EventRegistry extracts event location, date, entities and what is it about. Homepage of EventRegistry shows real time events, location associated with the event and number of articles comprised by the event. Figure 2.2 is showing the snapshot of EventRegistry homepage. According to the statistics shown in Figure 2.2, on the daily basis EventRegistry captures more than 60 thousand news articles and further summarize them into an average of 2000 events.

Searching with Strings, Things, and Cats: STICS uses keywords, entities, and semantic categories to aggregate documents. Based on named-entity disam-



Carmena: "Si pudiera rebobinar, mantendría mi no inicial a presentarme como alcaldesa" - ANTENA 3 TV
Location: **Madrid, Spain**
Articles about event: **8**

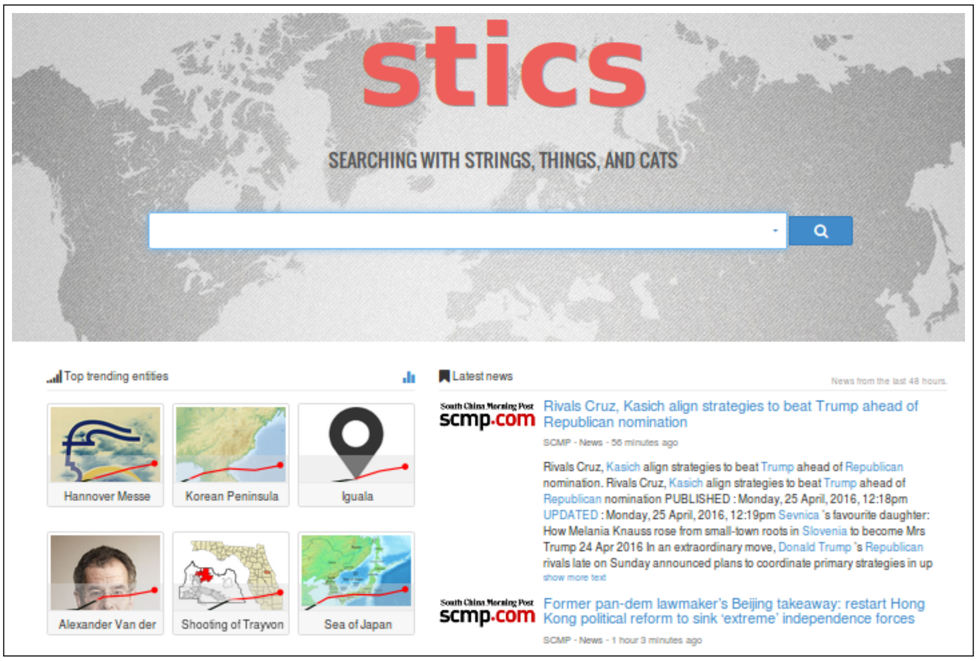
Data info	#
Imported articles:	56,319,401
Identified events:	3,271,597
News sources:	112,494
Unique concepts:	2,183,193
Categories:	4,997

Figure 2.2: EventRegistry Homepage (source: [2])

biguation, the search engine returns documents containing the query's entities. Using newswires RSS feeds news content is gathered in multiple languages and then merge into extracted information according to disambiguated mentions of the persons, organisations and locations. Figure 2.3 shows the homepage of STICS and the top trending entities in news along with latest news articles.

2.2.1 Limitation

Although, both types of system provide up-to-date related news information in real time, but they overload the user with large amounts of results. For instance, given input query '2014 FIFA world cup', event centric EventRegistry suggested 11,504 news event headlines, and the content centric STICS suggested 1,286,369 news articles, however not all of them may be useful. Existing work does not aim to



Rivals Cruz, Kasich align strategies to beat Trump ahead of Republican nomination

SCMP - News - 56 minutes ago

Rivals Cruz, Kasich align strategies to beat Trump ahead of Republican nomination. Rivals Cruz, Kasich align strategies to beat Trump ahead of Republican nomination PUBLISHED : Monday, 25 April, 2016, 12:18pm UPDATED : Monday, 25 April, 2016, 12:19pm Sevnica 's favourite daughter: How Melania Knauss rose from small-town roots in Slovenia to become Mrs Trump 24 Apr 2016 In an extraordinary move, Donald Trump 's Republican rivals late on Sunday announced plans to coordinate primary strategies in up

show more text

Former pan-dem lawmaker's Beijing takeaway: restart Hong Kong political reform to sink 'extreme' independence forces

SCMP - News - 1 hour 3 minutes ago

Figure 2.3: STICS Homepage (source: [3])

provide a comprehensive coverage of entities linked to any event. Like for event based query, the system returns only popular entities linked, not focusing on what connection those entities have to the event. Having no such information, makes it difficult to search relevant information, as we can see in Figure 2.4, the overall distribution of person over an event. We observed that 'General' topic news query (like 'Election', 'ISIS') mostly contains large number of mentions than specific news queries (like 'Bihar Election', 'Paris Attack'). It is difficult to query again and again in order to know the role of each entity is playing in the event. Also, even though news articles still contains majority of the information, the headlines and video data format are moving fast to capture the 'prominent share', and more data forms are about to join them in near future. However, to the best of our knowledge, no attention has been focused on using multiple data formats to extract news content.

Also the aforementioned system does not suggest any important and interesting news concepts, meme, entities that are emerging during the input query. They do not focus on reducing the resulting query specific data, focusing primarily on

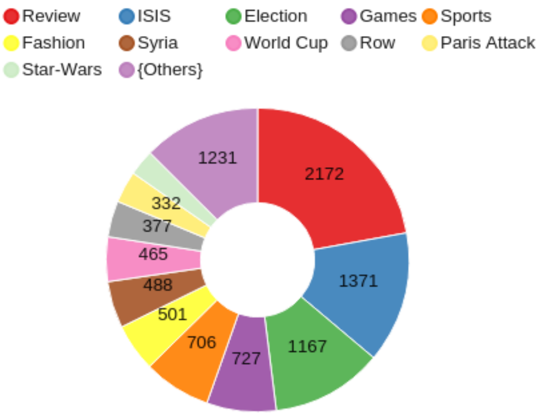


Figure 2.4: Distribution of *person* over event

event coverage. For example, for event 'Bihar Election' more than 4000 articles are retrieved using EMM, thus the information overflow problem is still prevailing. Lacking in some of the important data analysis tasks, like identification of prominent news concepts, additional associated information, and many others. These tasks are left for the reader to analyse. Clearly, there is a need for a system, enabling readers to get a broad overview of news data and thus, reducing the burden of understanding wide information.

Thus, news query based keyphrase mining and knowledge extraction from multiple data formats in a joint way to improve the news exploration task, have not been attained so far. This thesis fills the gap by bringing linguistic and syntactical knowledge together from different data forms, such as headlines and video captions. And then combines the information in a synergistic way to rank the generated list of keyphrases on the basis of phrase novelty and activeness in the input query timeline (an other features mention in Chapter 4). We performed several experiments to evaluate the efficiency of our method on different grounds of quality and quantity. We observe that our system outperforms the state-of-the-art. The user case study also proves that our linguistic and syntactic features improves the keyphrase extraction process over previous methods (in both single data form-news headline and multi-data form), thus showing for the first time the beneficial effects of exploiting multi-data knowledge in a joint fashion.

CHAPTER 3

SYSTEM OVERVIEW OF E^3

In this chapter, we surfaced an introductory overview of the approach to be used to mine important and interesting keyphrases from given entity's news data. The chapter is divided into three sections. In the first section, we have discussed about some preliminary ideas required to formulate the solutions and a formal statement stating the problem we have solved in this dissertation. The next section describes about the type of data we have worked with in proposed engine E^3. Final section of the chapter presents a glimpse of the engine E^3, discussing about various steps required to extract keyphrases from news corpus and for enriching the keyphrases. The detailed procedure of each module has been discussed in later chapters. In particular, this chapter shows how the proposed engine for news event exploration evolves in successive processing stages to gain its final shape.

3.1 Preliminaries

Definition 1. `News Concept` is an abstract idea about the particular goal, action or behavior connected with the news document. Integrating all news concepts provides full congruity of news document.

Definition 2. `Keyphrase` is a short and meaningful group of words describing prominent news concepts and entities mentioned in news document.

Definition 3. `Time Series` T = {v[1], v[2], $\cdots$, v[n]} with length $|T|$ = n is a sequence of real value observations, where v[i] is count of number of times the keyphrase appeared in news headlines for day i. A query q is popular during time interval T[i, j], where $t_i \leq t_j$, such that $\forall_{t \in [t_i, t_j]}$ v[t] ≥ mean(T).

Definition 4. `Novel Keyphrase` is popular only during query time interval, as it appears frequently within the considered time interval, not before.

Definition 5. `Active Keyphrase` is the one appearing frequently within the considered time interval, and have appeared frequently before also.

Problem statement: For given heterogeneous news data, we aim to summarize news results using relevant and prominent keyphrases. Also, our goal is to enrich the extracted keyphrases by assigning tag, rank, class, time-interval, role and related connections with respect to input query in order to make the set of keyphrases interesting to read. The goal is achieved when the set of keyphrases shown to N persons and more than N/2 persons should find it interesting. Where interestingness can be measured on the basis of frequency, collocation and completeness of the topic covered by the output set of keyphrases.

3.2 Glimpse of E^3

Given an input query q, system E^3 extracts keyphrases and stores them in a keyphrase template, as shown in Figure 3.1. Since keyphrases provides a high-level overview of news events in a short span of time. Consequently, using keyphrase we can provide fast information without requiring too much of efforts and time. Over the top, our proposed system E^3 organizes keyphrase into a template consisting three sections: *Type, Ranking* and *InfoBox*. The keyphrase of type $person$, $location$ and $organization$ are stored in Type section. The $novel$, $active$ and $frequent$ keyphrases are stored in Ranking section. Information like role played, time interval, top most associated types and news concept of the selected keyphrase are stored in InfoBox section. For reference, see Figure 3.1 showing details for 'Bihar Election'. This kind of organization of information is very helpful in cases where many people, multiple organization and multiple location are discussed. As it is

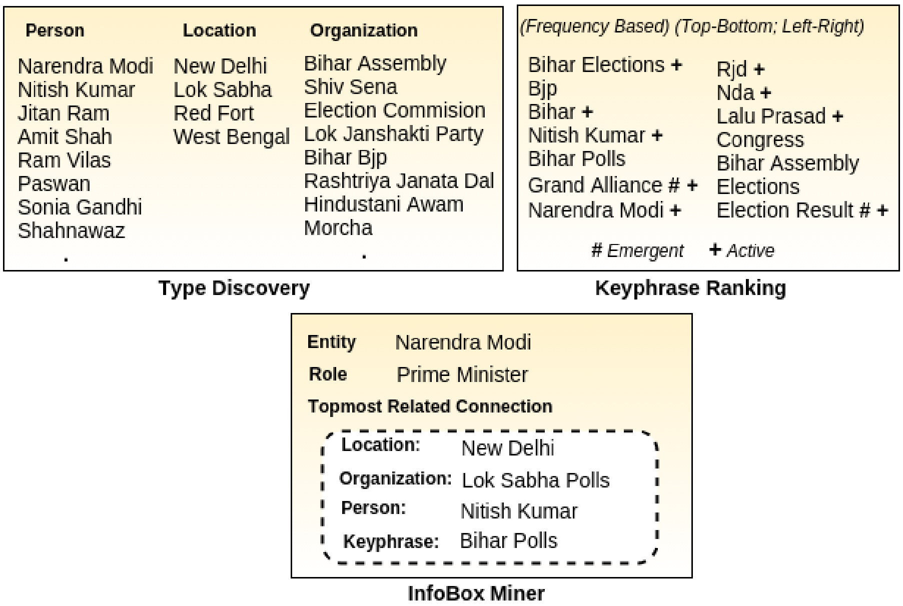


Figure 3.1: E^3 System Working Example for q: 'Bihar election'

very difficult for a reader to keep track of all the mentioned entities, hence they require external engines to know more information about the mention. Henceforth, readers often loses interest in the news concept and the network traffic gets lost. Also the burden on user is increasing as they have to search and analyse the results rigorously in order to learn relation between news concept and entities mentioned.

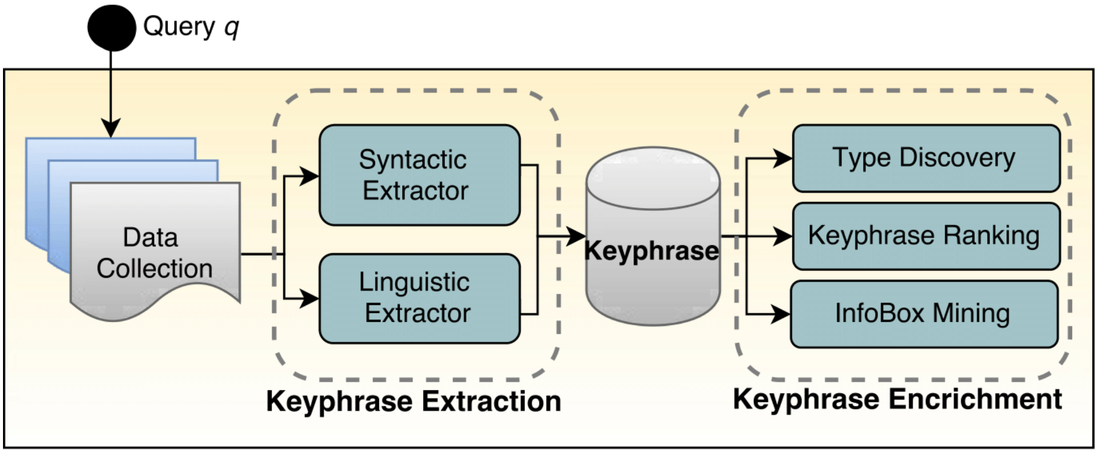


Figure 3.2: E^3 System Architecture

Figure 3.2 depicts the architecture of our system E^3, while Algorithm 1 summarizes the actual procedure performed. The primary contribution of our system E^3 lies in *Keyphrase Extraction* and *Keyphrase Enrichment*. The first component in our pipeline collects news data related to the input query (line 2 to 4). The further steps consist in identifying the candidate phrases, potentially relevant to the news query. For this we pass the collected data to two different extractor, explained in Chapter 4. The series of steps performed to select valid phrases is shown in line 7 to 11. After keyphrase extraction and keyphrase cleaning (line 12), the keyphrases are assigned ranks further (line 13). To enrich the extracted keyphrases, they are passed to following modules (line 15 to 19):

- `Type_Discovery`: Finds the type of keyphrase.
- `Keyphrase_Ranking`: Assigns rank to the keyphrase using global list of ranked keyphrases $kpSet$.
- `Time_Interval`: It gives the time period when the keyphrase was present in our the collected news headlines database.
- `EmergentActive_Classifier`: Tags the novel and active keyphrases.
- `InfoBox_Mining`: It extracts further related information like role played, associated entities and others.

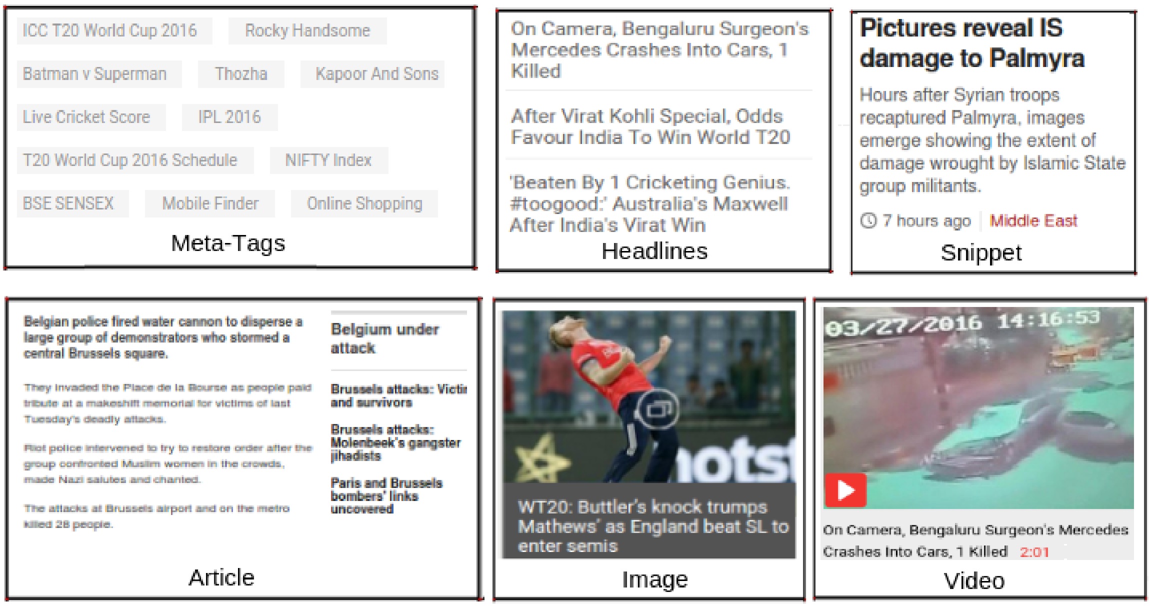
3.3 E^3 Data Collection

As shown in Algorithm 1: Step 3, E^3 collects multi-form data like short text, long text, image and video textual content and stores them as a record. A brief outlook of different data types use by E^3 is shown in Figure 3.3. Where meta-tags, headlines, image and video captions are considered as short text and others as long text.

For input query q, our Data Collection module scrapes related news data published by news media. The module can either use news search engines like Google News, Yahoo News or online news structured repositories like GDELT [13], EMM [16], iMM [17]. In our developed system we are using our in-house iMM system that periodically extracts news headlines (video title) along with their URL,

Algorithm 1 E^3 System

```
Input. q: Query
$USet$ = ObtainURL(iMM, $q$)
for each url $u$ in $USet$ do
    $q_u = \langle$Publication Date, Headline, Keywords, Snippet, Article$\rangle$
    $R_q \cup = (u, q_u)$
end for
$kpSet = \phi$
for each record $r$ in $R_q$ do
    $kpSet \cup = r$.Keywords
    $kpSet \cup =$ Syntactic_Extractor($r$.Headline)
    $kpSet \cup =$ Linguistic_Extractor($r$.Snippet, $r$.Article)
end for
Remove duplicate keyphrases from $kpSet$
$kpSet$.rank = Keyphrase_Ranking($kpSet$, $USet$)
for each keyphrase $k$ in $kpSet$ do
    $k.type$ = Type_Discovery(k)
    $k.rank$ = Keyphrase_Ranking(k, kpSet, USet)
    $k.time$ = Time_Interval(k, q, USet)
    $k.class$ = Emergent-Active_Classifier(k, k.time)
    $k.info$ = InfoBox_Miner(q, k, USet)
end for
```



ICC T20 World Cup 2016 | Rocky Handsome | Batman v Superman | Thozha | Kapoor And Sons | Live Cricket Score | IPL 2016 | T20 World Cup 2016 Schedule | NIFTY Index | BSE SENSEX | Mobile Finder | Online Shopping

Meta-Tags

On Camera, Bengaluru Surgeon's Mercedes Crashes Into Cars, 1 Killed

After Virat Kohli Special, Odds Favour India To Win World T20

'Beaten By 1 Cricketing Genius. #toogood:' Australia's Maxwell After India's Virat Win

Headlines

Pictures reveal IS damage to Palmyra

Hours after Syrian troops recaptured Palmyra, images emerge showing the extent of damage wrought by Islamic State group militants.

7 hours ago | Middle East

Snippet

Belgian police fired water cannon to disperse a large group of demonstrators who stormed a central Brussels square.

They invaded the Place de la Bourse as people paid tribute at a makeshift memorial for victims of last Tuesday's deadly attacks.

Riot police intervened to try to restore order after the group confronted Muslim women in the crowds, made Nazi salutes and chanted.

The attacks at Brussels airport and on the metro killed 28 people.

Belgium under attack

Brussels attacks: Victi and survivors

Brussels attacks: Molenbeek's gangster jihadists

Paris and Brussels bombers' links uncovered

Article

WT20: Buttler's knock trumps Mathews' as England beat SL to enter semis

Image

03/27/2016 14:16:53

On Camera, Bengaluru Surgeon's Mercedes Crashes Into Cars, 1 Killed 2:01

Video

Figure 3.3: E^3 News input data types

publication date, article (or video caption), snippet and meta-keywords. To prepare R related to q, we select the URL from iMM database if q is contained by the URL's headline (video title) or URL's meta-keywords. Then, the URLs are use to extract the other data. After complete data collection, some automatic preprocessing (such as removal of advertisement sentences, de-duplication of news) is performed over the collected data.

In the proposed system, we use our in-house iMM system that periodically extracts news headlines along with their URL, publication date and news keywords. iMM system has been working since January 2014 and has collected more than 30 million news headline till date. Article content and meta-description related to news headline are not extracted by iMM system. These are obtained by our module after performing web-query using the collected URLs. In summary, for given query q, we prepare a set of news records R_q, where each record is described by Quintuple {*Headline, Keywords, Meta-description, Article, Publication Date*}. For 'Bihar Election' query, we prepared 216 records for further processing.

CHAPTER 4

KEYPHRASE EXTRACTION AND ENRICHMENT

To keep track of long-lasting global news stories, keyphrases are one of the current state-of-the-art. They can be leveraged to provide suboptimal summaries in terms of relevance and interestingness. Chapter 3 discussed the heterogeneous data collection stage of E^3. This chapter explains how the prominent and interesting information is extracted from the collected data. Section 4.1 describes how keyphrases are extracted by E^3 in detail, along with supporting examples. Section 4.2 explains how one actually enrich the keyphrases.

4.1 Keyphrase Extraction

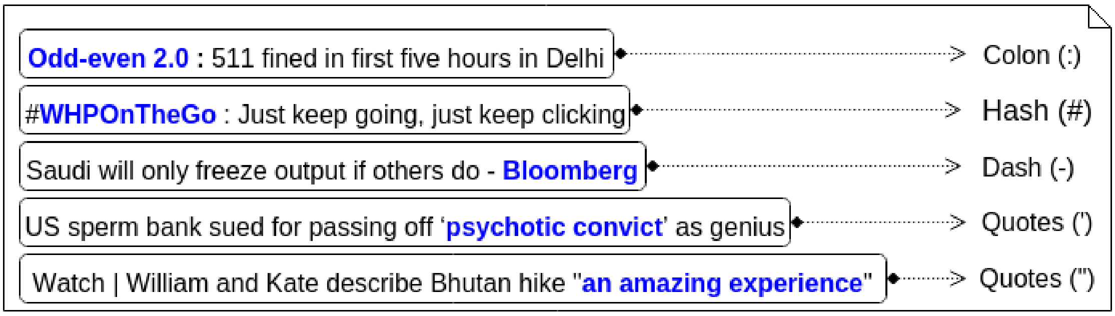
A naive solution is to output the meta-keywords (related tags) associated with news articles as keyphrases. However, the great majority of articles come without meta-keywords. As a result, meta-keywords are not sufficient enough to describe the news data completely. For example, around 20% news URLs, obtained for ‘Bihar Election’, do not have a meta-keyword. Over the top, meta-keywords are very generic (top-level category), assigning them manually is a tedious task, requiring complete knowledge of news concept. Hence we require an efficient selection of relevant phrases. Consider the n-grams ‘he told reporters’, ‘commonly found’, such strings represent valid phrase, but it does not make much sense to mark the n-grams as valid keyphrase, as it does not provide any conceptual knowledge. So we need an efficient way to extract prominent and informative keyphrases from the news data.

On our careful observation, we found that news headlines are short in length and contains special tokens such as colon (:), apostrophe ('), quotes (", '), hash (#), dash (-) to emphasize important information. On the other hand, snippets and news articles are long passage texts and are governed by grammatical rules. Thus, we present two different keyphrase extractors to handle both kinds of writing styles.

Syntactic Extractor utilises special characters as discussed in Algorithm 2 for keyphrase extraction. News headline tokens containing # is directly stored in keyphrase set (line 1). In case of colon (dash), the news headline is splited into two parts using colon (dash) and the shortest part among the two is declared as keyphrases (line 3 to 6). In case of quotes (line 7) the part of text enclosed inside the quotes is declared as keyphrase as shown in line 8 to 10. For reference see Figure 4.1, showing extracted keyphrases in blue for various types of news headlines. From Figure 4.2 and Figure 4.3, we can see that the syntactic approach can also be used effectively to extract keyphrases from multi-lingual news headlines and news tweets.

Algorithm 2 E^3 System: Syntactic_Extractor

```
Input. $text$: Input short text
Output. $kpSet$: Set of keyphrases
 1: $kpSet$ = {words of $text$ containing '#'}
 2: $tokenSet_1$ = { :, ;, - }
 3: for each token $t$ in $tokenSet_1$ do
 4:     $\langle l_1, l_2 \rangle$ = {split $text$ using token $t$ into two parts}
 5:     $kpSet \cup = \{l_1$ if len($l_1$) < len($l_2$) else $l_2\}$
 6: end for
 7: $tokenSet_2$ = { ‘ ’, “ ”, ` ´ }
 8: for each token $t$ in $tokenSet_2$ do
 9:     $kpSet \cup =$ {subpart of $text$ surrounded by $t$}
10: end for
11: return $kpSet$
```



Odd-even 2.0 : 511 fined in first five hours in Delhi

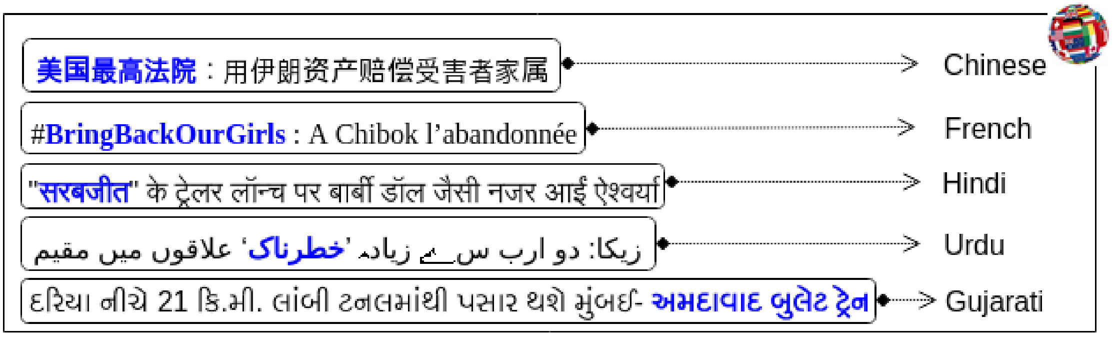
#**WHPOnTheGo** : Just keep going, just keep clicking

Saudi will only freeze output if others do - **Bloomberg**

US sperm bank sued for passing off '**psychotic convict**' as genius

Watch | William and Kate describe Bhutan hike "**an amazing experience**"

Figure 4.1: E^3 Syntactic Extractor on short text (news headlines)



美国最高法院：用伊朗资产赔偿受害者家属

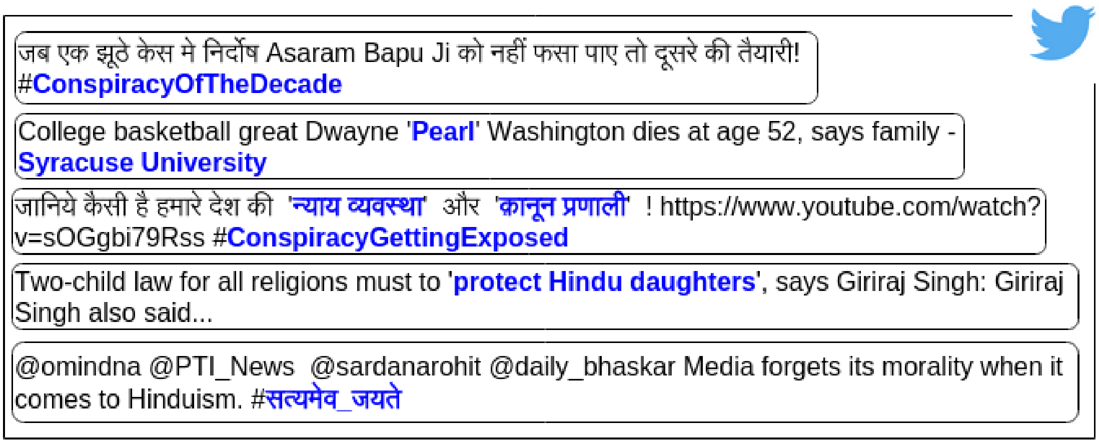
#**BringBackOurGirls** : A Chibok l'abandonnée

"**सरबजीत**" के ट्रेलर लॉन्च पर बार्बी डॉल जैसी नजर आईं ऐश्वर्या

زیکا: دو ارب سے زیادہ '**خطرناک**' علاقوں میں مقیم

દરિયા નીચે 21 કિ.મી. લાંબી ટનલમાંથી પસાર થશે મુંબઈ- **અમદાવાદ બુલેટ ટ્રેન**

Figure 4.2: E^3 Syntactic Extractor on multilingual news



जब एक झूठे केस मे निर्दोष Asaram Bapu Ji को नहीं फसा पाए तो दूसरे की तैयारी! #**ConspiracyOfTheDecade**

College basketball great Dwayne '**Pearl**' Washington dies at age 52, says family - **Syracuse University**

जानिये कैसी है हमारे देश की '**न्याय व्यवस्था**' और '**कानून प्रणाली**' ! https://www.youtube.com/watch?v=sOGgbi79Rss #**ConspiracyGettingExposed**

Two-child law for all religions must to '**protect Hindu daughters**', says Giriraj Singh: Giriraj Singh also said...

@omindna @PTI_News @sardanarohit @daily_bhaskar Media forgets its morality when it comes to Hinduism. #**सत्यमेव_जयते**

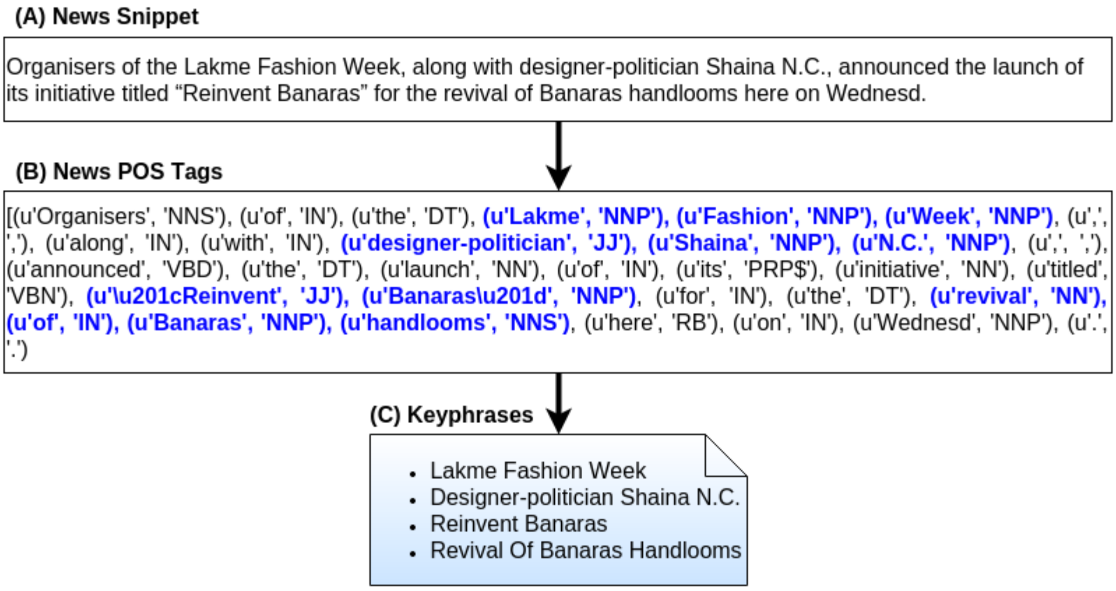
Figure 4.3: E^3 Syntactic Extractor on news tweets

Linguistic Extractor applies language specific part of speech (POS) tagging over snippets and articles. After the POS tagging, we change the part of speech of punctuation marks, adjectives, cardinal numbers, conjunctions to noun, as shown in line 2 to 6. Then we select all the collocated nouns and store them as keyphrases. For reference, a snippet and its POS tagging is shown in Figure 4.4. The adjectives (JJ) and prepositions (IN) collocated with nouns (NN) and its family (NNP) are returned as keyphrase ('Designer-politician Shaina NC', 'Reinvent Banaras' and 'Revival of Banaras Handlooms') along with continually collocated nouns ('Lakme Fashion Week').

Algorithm 3 E^3 System: Linguistic_Extractor

Input. $text_1$: Input Snippet, $text_2$: Input Article
Output. $kpSet$: Set of keyphrases

```
1: for each text $t$ in $text_1$, $text_2$ do
2:     change Punctuation marks (dash (-), apostrophe ('s), and, comma (,) ) POS ⇒ Noun family
3:     change Adjectives and its family (JJ, JJR, JJS ) ⇒ Noun family
4:     change Cardinal number (CD) ⇒ Noun family
5:     change Coordinating conjunction (CC) ⇒ Noun family
6:     $kpSet$ = select all the collocated nouns and its family
7: end for
8: return $kpSet$
```



Organisers of the Lakme Fashion Week, along with designer-politician Shaina N.C., announced the launch of its initiative titled "Reinvent Banaras" for the revival of Banaras handlooms here on Wednesd.

```
[(u'Organisers', 'NNS'), (u'of', 'IN'), (u'the', 'DT'), (u'Lakme', 'NNP'), (u'Fashion', 'NNP'), (u'Week', 'NNP'), (u',', ','), (u'along', 'IN'), (u'with', 'IN'), (u'designer-politician', 'JJ'), (u'Shaina', 'NNP'), (u'N.C.', 'NNP'), (u',', ','), (u'announced', 'VBD'), (u'the', 'DT'), (u'launch', 'NN'), (u'of', 'IN'), (u'its', 'PRP$'), (u'initiative', 'NN'), (u'titled', 'VBN'), (u'\u201cReinvent', 'JJ'), (u'Banaras\u201d', 'NNP'), (u'for', 'IN'), (u'the', 'DT'), (u'revival', 'NN'), (u'of', 'IN'), (u'Banaras', 'NNP'), (u'handlooms', 'NNS'), (u'here', 'RB'), (u'on', 'IN'), (u'Wednesd', 'NNP'), (u'.', '.')
```

- Lakme Fashion Week
- Designer-politician Shaina N.C.
- Reinvent Banaras
- Revival Of Banaras Handlooms

Figure 4.4: E^3 Linguistic Extractor on news snippet and article

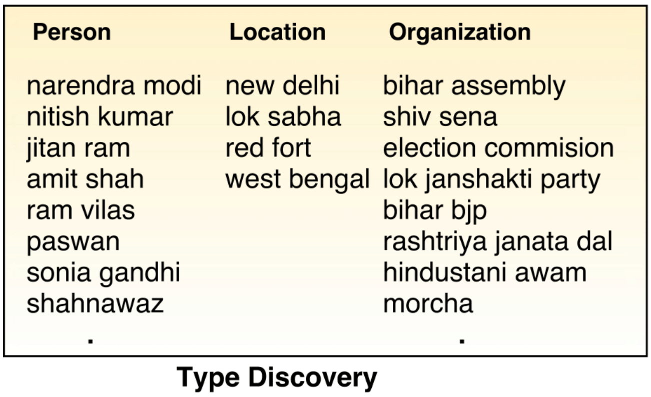
At the end, when all the news records in R are processed, we obtain a set of keyphrases R_k, along with the number of times they are generated.

4.2 Keyphrase Enrichment

The size of generated keyphrases R_k may be large and noisy. To resolve this problem, our Keyphrase Enrichment module helps in extracting valuable and actionable information by filtering and ranking the extracted keyphrases. The keyphrases are filtered using news media specific stopwords such as update, video, photo, pti and others. Next, we apply case normalization and remove the duplicate keyphrases. At this point, noisy keyphrases are removed. The remaining keyphrases are passed through the Type discovery, Novel-Active classifier, Keyphrase ranker, and InfoBox miner modules. The working of each module is explain in the following subsections.

4.2.1 Type Discovery

In *Type Discovery* module, the language specific NER tagger is used to classify keyphrase into three types: $Person$, $Location$ and $Organization$. However, existing NER taggers do not perform well for Indian named entities. Hence, we use a separate list[1] of Indian named entities for the tagging purpose (see Appendix B). A keyphrase without any above NER type, are termed as a *News Concept*. For 'Bihar Election' query, a sample keyphrases for each type is shown in Figure 4.5.



Person	Location	Organization
narendra modi	new delhi	bihar assembly
nitish kumar	lok sabha	shiv sena
jitan ram	red fort	election commision
amit shah	west bengal	lok janshakti party
ram vilas		bihar bjp
paswan		rashtriya janata dal
sonia gandhi		hindustani awam
shahnawaz		morcha
.		.

Type Discovery

Figure 4.5: Type Discovery for 'Bihar Election'

[1] https://github.com/NikkiJain09/Transliteration

4.2.2 Keyphrase Ranking

The module organizes keyphrases according to the value of frequency, novelty, activeness and time period recency. Frequency of keyphrase is already computed during keyphrase extraction process. To compute the value of novelty and activeness, we first extract the time intervals q_t for input query q, during which the q was highly popular in news headlines. Next, a keyphrase is *novel* if its frequency is very high in news headlines only during q_t. Similarly, a keyphrase is *active* if its frequency is high around q_t.

A naive approach for ranking would be treating frequency of keyphrases as prominent feature. However, the most frequent phrase seems more generic and rarely provide interesting insight about the news story. For instance, the top most frequent phrases for various queries is shown in Table 4.1. Like in case of 'Bihar Election' it is obvious for 'Bihar' to come, it is not providing any insightful detail about the query. Since a individual reader processes only a small part of the keyphrase story, such an approach would give an incomplete information view and could therefore easily deviate readers especially if there are data irrelevant to the action of interest.

Table 4.1: Frequency wise topmost keyphrase

Query	Top-most Keyphrase
Election	Bjp, Bihar
Bihar Election	Bjp, Bihar
ISIS	Syria, Islamic State
Paris Attack	Islamic State, France

Therefore, we hypothesize that learning active and novel phrases is prominent for news reader engagement. This is challenging from a frequent phrase model perspective as we have to stamp each keyphrase with time intervals. And annotating the phrase as Novel and/or Active, if satisfies the above mentioned conditions.

For instance, in Figure 4.6 'Grand Alliance' and 'NDA' are discovered as novel and active keyphrases for 'Bihar Election' query, as 'Bihar Election' time interval completely overlaps the 'Grand Alliance' time interval. And 'NDA' occurrence is quite significant during 'Bihar Election' time interval. Also, novel keyphrases interestingness is supported by the fact that most of them have low news article

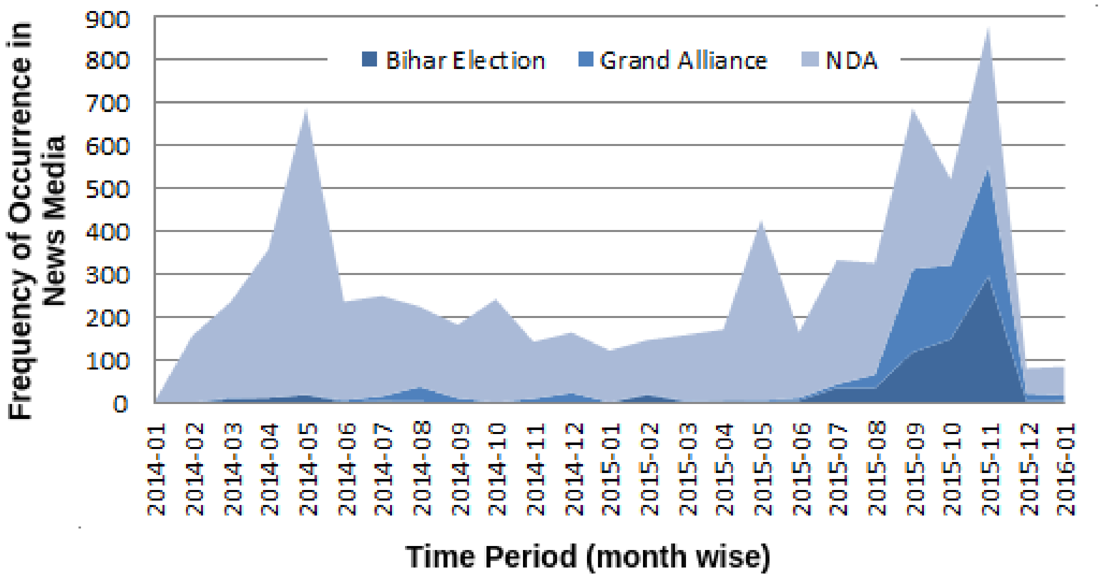
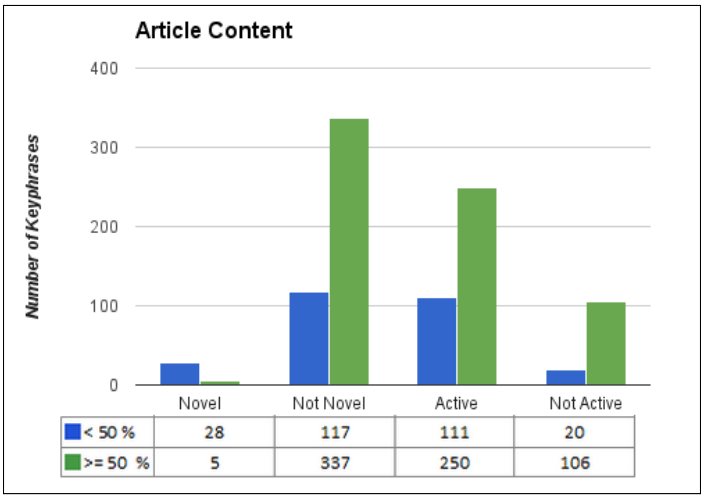


Figure 4.6: ‘Bihar Election’ and Novel Keyphrases

contextual percentage as they are frequently mentioned in meta-keywords (or related tags), snippets, headlines, URLs which users mostly read. For instance, article content present in various keyphrases are shown in Figure 4.7 for ‘Bihar Election’.



	Novel	Not Novel	Active	Not Active
🟦 < 50 %	28	117	111	20
🟩 >= 50 %	5	337	250	106

Figure 4.7: Novel keyphrase distribution

Thus, using following features of keyphrases to rank we can have important and interesting keyphrases on the top. In Keyphrase k `Time Series` T = {v[1], v[2], $\cdots$, v[n]} with length $|T|$ = n, where v[i] is count of number of times the keyphrase appeared in news headlines for month i.

1. **Frequency** (f): It reflects the importance of the keyphrase to news corpus. Larger the normalized value greater the involvement keyphrase have. Note that we have already removed stopwords specific to the news media from the set of keyphrases prepared earlier.

   $$\mathrm{f}_k = \textstyle\sum_{j=1}^{n} v[\mathrm{j}] \, / \, \sum_k \sum_{j=1}^{n} v[\mathrm{j}] \tag{4.1}$$

2. **Novelty** (η): The value reflects whether the keyphrase emerged during the input query time interval (novel).

   Keyphrase k is novel during the input query q time series T_q = {w[1], w[2], $\cdots$, w[m]}

   $$\eta_k = \begin{cases} 1, & \text{if } \forall_{i \in [T_1, T_n]}\ v[\mathrm{i}] = \mathrm{mean(T)} \text{ and } i \ \exists\ [w_1, w_m] \\ 0, & \text{otherwise} \end{cases} \tag{4.2}$$

3. **Activeness** (α): The value reflects whether the keyphrase was actively present during the input query time interval (active).

   Keyphrase k is active during the input query q time series T_q = {w[1], w[2], $\cdots$, w[m]}

   $$\alpha_k = \begin{cases} 1, & \text{if for any } v[\mathrm{i}] = \mathrm{mean(T)} \text{ and } i \ \exists\ [w_1, w_m] \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

4. **Temporal Recency** (τ): The keyphrase having the recent active participation in the news corpus have lower value and vice-versa. During ranking time we take the inverse of this normalized value to bring interesting and important keyphrases on the top.

   Keyphrase k last active month is j such that v[j] ≥ mean(T) and

   $$\tau_k = \mid \text{n-j} \mid \text{ is minimum } \forall_{j \in [T_1, T_n]} \tag{4.4}$$

5. **First Occurrence Count** (ρ): The value indicates the number of times a keyphrase was mentioned in the first half of parent news article. Higher the value of ρ greater the chance of k to occur on top of ranking list.

   $$\rho_k = \begin{cases} 1, & \text{if Quintuple } [Article] \text{ first half section} \\ 0, & \text{otherwise} \end{cases} \tag{4.5}$$

6. **Meta-Tag Presence** (Π): The value indicates the number of times the keyphrase was used by journalist as a related tag to identify the news article. Higher the value of Π greater the chance of k to occur on top of ranking list.

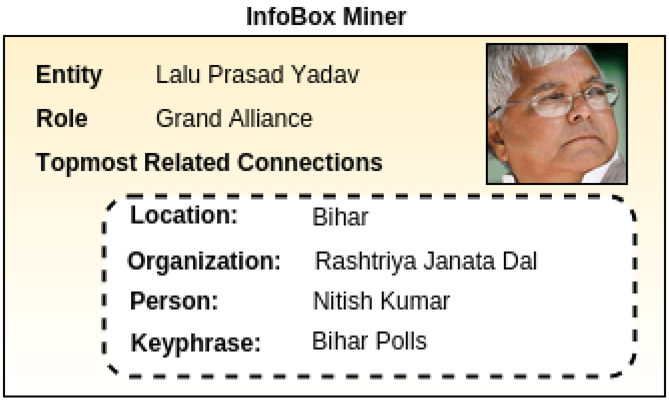
   $$\Pi_k = \begin{cases} 1, & \text{if } k \; \exists \text{ Quintuple } [Keywords] \\ 0, & \text{otherwise} \end{cases} \tag{4.6}$$

7. **Headline Presence** ($\hbar$): The value signifies the number of times the keyphrase was mentioned in the news headline. Higher the value of $\hbar$ greater the chance of k to occur on top of ranking list.

   $$\hbar_k = \begin{cases} 1, & \text{if } k \; \exists \text{ Quintuple } [Headline] \\ 0, & \text{otherwise} \end{cases} \tag{4.7}$$

4.2.3 InfoBox Miner

Using InfoBox Miner we discover *News Related* information for selected keyphrase k with respect to query q. For instance, 'Lalu Prasad Yadav' role present in Google search engine InfoBox is 'Indian Politician' (a general one). Whereas, our system E^3's InfoBox give more specific role. Like as shown in Figure 4.8 'Lalu Prasad Yadav' role returned is 'Grand Alliance', as 'Lalu Prasad Yadav' is one of the member of Grand Alliance group formed. Also E^3's InfoBox displays type-wise top most connections k have. The connections are determined with help of co-occurrence value of the keyphrases in R_k. To find the role played by k we extracts lines from news data were both k and q are present. Then using extracted lines absolute distance is calculated between k and all other the keyphrases present in R_k. The keypharse having minimum mean-distance is selected as k's role. Generally, keyphrases with type person and organization are preferred for InfoBox mining, as they are actors, performing some actions.



InfoBox Miner

Entity Lalu Prasad Yadav

Role Grand Alliance

Topmost Related Connections

Location: Bihar

Organization: Rashtriya Janata Dal

Person: Nitish Kumar

Keyphrase: Bihar Polls

Figure 4.8: 'Lalu Prasad Yadav' InfoBox for 'Bihar Election'

E^3 InfoBox also provides Histogram to compare all associated news concept with k. The horizontal axis in Histogram represents the associated news concepts with k and the vertical axis plots frequency of those keyphrases in news data containing k and q. Figure 4.9 shows the Histogram for 'Lalu Prasad Yadav'.

The timeline of k in the news query is also shown along with the aforementioned information for more insights about the presence of k in q. Figure 4.10 shows one such timeline for 'Lalu Prasad Yadav'. The horizontal axis shows the q monthly

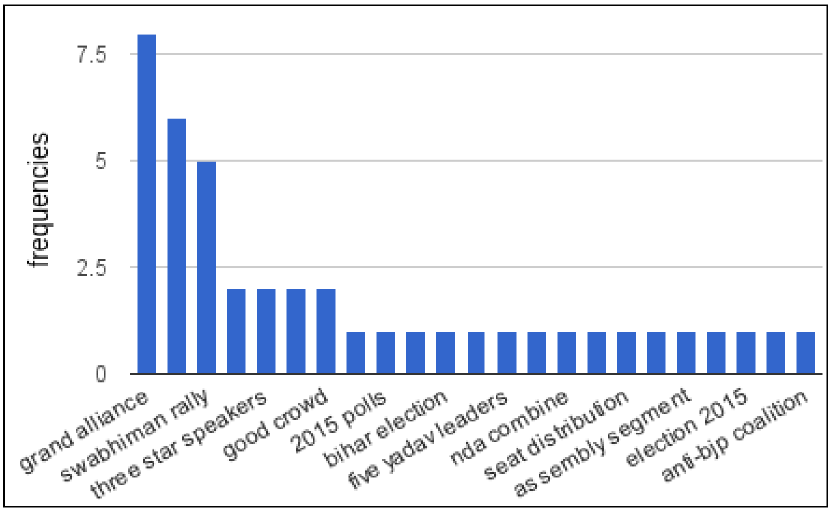


Figure 4.9: 'Lalu Prasad Yadav' associated news concept Histogram

time interval and vertical axis depict the frequency of k in the corresponding month.

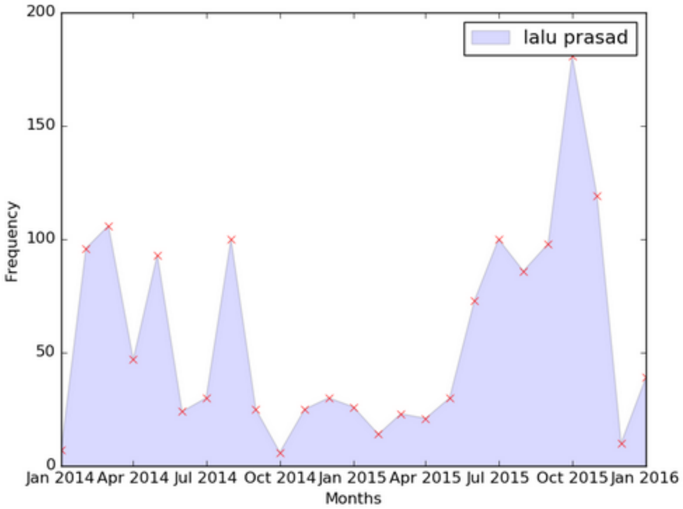


Figure 4.10: 'Lalu Prasad Yadav' associated news concept Histogram

CHAPTER 5

Experimental Work

In this chapter, we present the results of our experimental evaluation of E^3 on both quality and quantity basis. We used both fine and lose-grained news queries to test our system with present state-of-the-art. For evaluation purpose, we used only English news corpora, as many state-of-the-art are not multilingual. However, to the best of our knowledge, there is no such gold standard corpora for such comparison, so we divided our evaluation task into three. First, evaluating the richness of keyphrases discovered by our system. Then, we evaluate the meaningfulness of discovered keyphrases. And finally, we asked human judges to compare the keyphrases obtained from E^3 to find out the overall interestingness of our proposed system.

5.1 Richness of Keyphrases

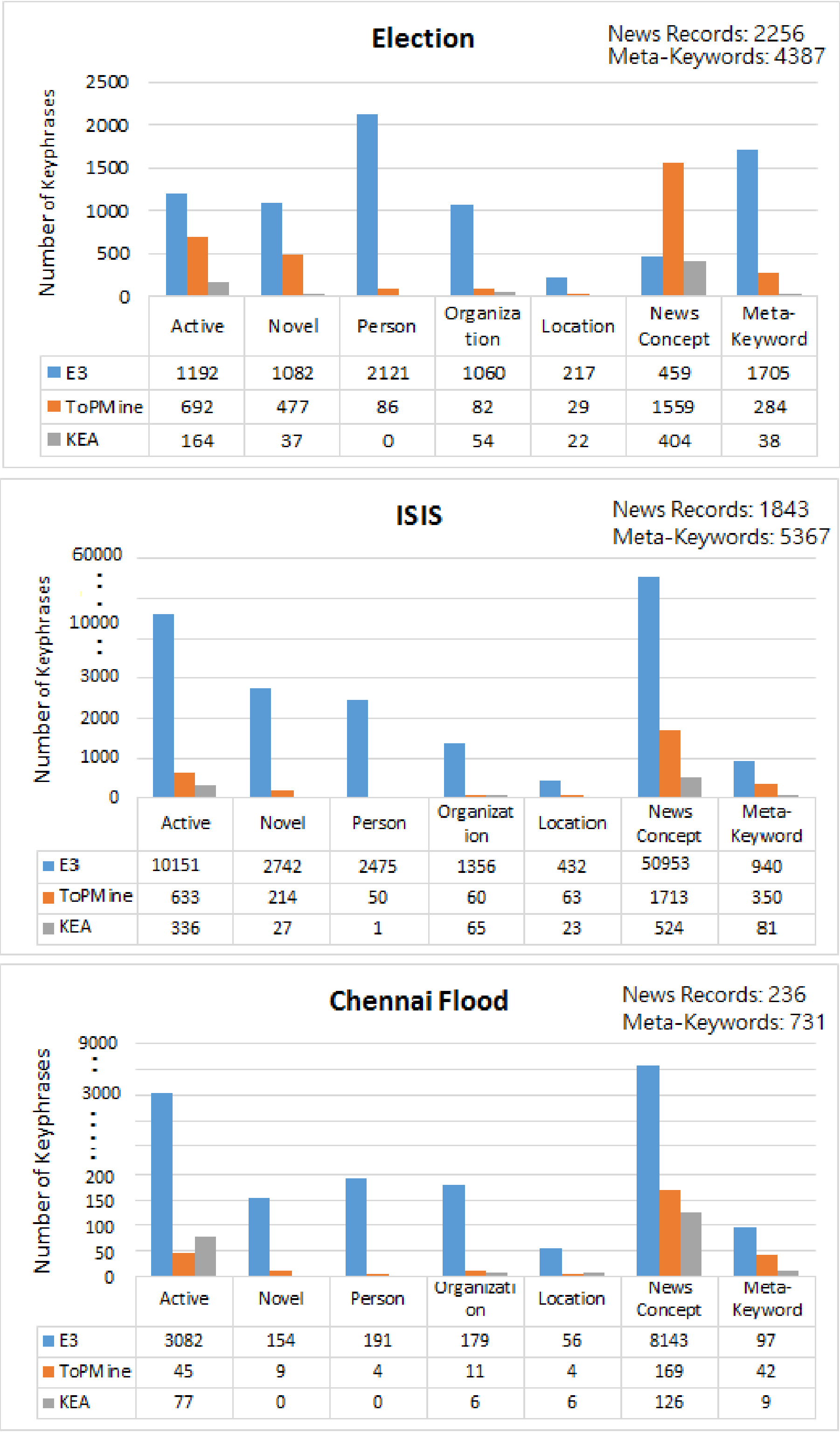
To evaluate the completeness of keyphrase based information extraction by our system E^3, we selected KEA and ToPMine as a baseline. We trained the KEA model over seven different queries' news corpora, collected using our Data Collection module and their respective news records' meta-keywords as the key. The training data have more than 1.08 Million tokens. The training of KEA model is required as available model is trained on standard passage text, which is written in different style from news data.

Figure 5.1 and Figure 5.2 shows the results obtained by KEA, ToPMine and E^3 for both general and specific news queries. The columns describe the Active, Novel, Person, Organization, Location, News Concept and the number of meta-keywords contained in the set of keyphrases obtained from the algorithm. However, the records in each type are not unique, problems like co-resolution exists, hence duplicate named entities are present. The individual row in figures lists the results obtained by using the same news corpora as input to all E^3, ToPMine and KEA for keyphrases discovery.

From Figure 5.1 and Figure 5.2 we can infer that the keyphrases discovered by E^3 have a significant contribution to the task of keyphrase extraction, which show that semantic information, even if minimal, is important for keyphrase based information extraction. Some of the noteworthy observations are:

- While comparing the two systems, we note that our proposed system E^3 discovered more number of associated entities. Hence, making it possible to extract entities information using InfoBox mining, or some other interesting memes, trivia, related to both the query and entity.
- The number of active and novel keyphrases returned by our system is fairly high, hence leading to a better set of news concept, as activeness and novelty are the key for interesting and effective understanding of results.
- Also, on manual inspection of the results, we observed that the news concepts generated by KEA system is more general, not giving much insight about the actions happening inside the news corpus. For instance, given query $paris\ attack$, our system E^3 generated many news concepts giving information about the attack like: Paris Terrorist, Islamic State, Mastermind, Paris Stadium, Bataclan, Paris Theater, 2015, Deaths, Victim, and many others, on the other hand KEA's news concept give any information related to the attack.

Another significant advantage of our linguistic-syntactic approach is that it scales better to unknown languages.



Election

News Records: 2256
Meta-Keywords: 4387

	Active	Novel	Person	Organization	Location	News Concept	Meta-Keyword
🟦 E3	1192	1082	2121	1060	217	459	1705
🟧 ToPMine	692	477	86	82	29	1559	284
▨ KEA	164	37	0	54	22	404	38

ISIS

News Records: 1843
Meta-Keywords: 5367

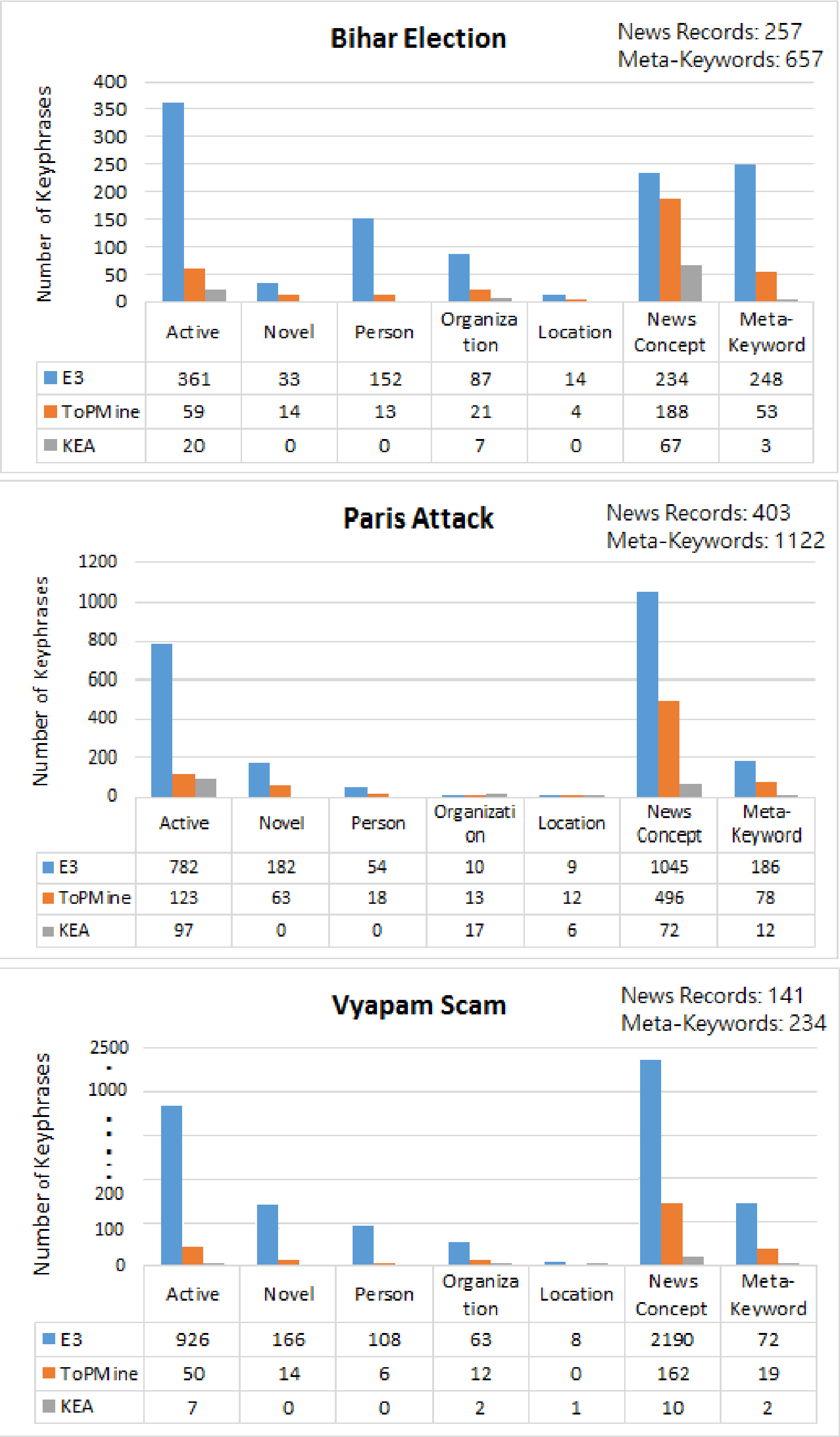
	Active	Novel	Person	Organization	Location	News Concept	Meta-Keyword
🟦 E3	10151	2742	2475	1356	432	50953	940
🟧 ToPMine	633	214	50	60	63	1713	350
▨ KEA	336	27	1	65	23	524	81

Chennai Flood

News Records: 236
Meta-Keywords: 731

	Active	Novel	Person	Organization	Location	News Concept	Meta-Keyword
🟦 E3	3082	154	191	179	56	8143	97
🟧 ToPMine	45	9	4	11	4	169	42
▨ KEA	77	0	0	6	6	126	9

Figure 5.1: Quantity Comparison for General Topics: KEA and E^3's keyphrases



	Active	Novel	Person	Organiza tion	Location	News Concept	Meta-Keyword
■ E3	361	33	152	87	14	234	248
■ ToPMine	59	14	13	21	4	188	53
■ KEA	20	0	0	7	0	67	3

	Active	Novel	Person	Organizati on	Location	News Concept	Meta-Keyword
■ E3	782	182	54	10	9	1045	186
■ ToPMine	123	63	18	13	12	496	78
■ KEA	97	0	0	17	6	72	12

	Active	Novel	Person	Organiza tion	Location	News Concept	Meta-Keyword
■ E3	926	166	108	63	8	2190	72
■ ToPMine	50	14	6	12	0	162	19
■ KEA	7	0	0	2	1	10	2

Figure 5.2: Quantity Comparison for Specific Topics: KEA and E^3's keyphrases

5.2 Meaningfulness of Keyphrases

We were also interested in comparing the keyphrases discovered from E^3 against the searches made by online readers. To evaluate this, we used Google trending searches and observed that the set of keyphrases discovered by our proposed system E^3 contains most of the related searches to the input query, shown by Google Trend. For instance Table 5.1, shows the result. Each row individually compares between Google's related trending search and E^3's keyphrase set for the input query. Cases where system E^3 was not able to match with the Google's trending search are:

1. `Misspelled`. For example, 'Bihar election commison'
2. `Co-reference`. For example, 'Election commission india and Indian Election Commission'; 'Bihar panchayat election and Bihar panchayat polls'
3. `Subset Match`. For example, '2016 election and 2016 election UK'

Table 5.1: Comparison: E^3 keyphrase and Google Trend query

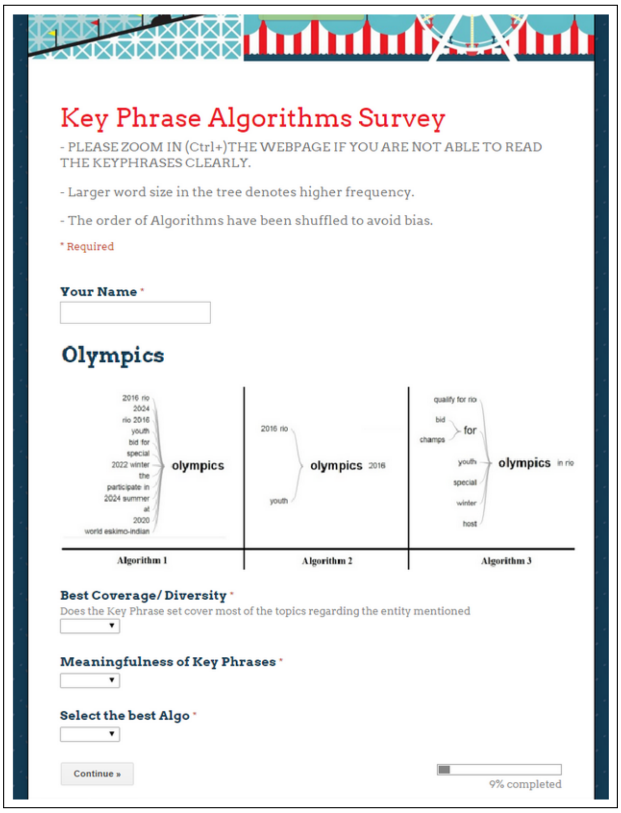
Number of	Bihar Election	Election
Related queries	23	50
Spelling mistakes	3	3
Complete match	7	25
Matches including Subsets	13	35

5.3 Quality of Keyphrases: User Case Study

To ensure a fair comparison of interestingness, importance and collocation of the discovered keyphrases, we asked humans to evaluate the keyphrases generated by current state-of-the-art supervised (KEA), unsupervised (Micro-gram) and our proposed unsupervised linguistic-syntactic (E^3) approaches. We tested all the aforementioned keyphrase extraction methods over news headlines data. To ensure that the results will be totally unbiased, we covered the broad range of all events occurred in past. We prepared a survey form[1] (Figure 5.3) having 10 different

[1] `http://goo.gl/forms/if8lRXEtIJ`

topic keyphrases: Events FIFA, 26/11 attack, Earthquake , Entities Google, Greece, Obama, Salman Khan, Concept Scam, Budget. And asked the evaluators to rank the anonymous results on the basis of *Coverage, Meaningfulness* and *Overall Quality*. Additionally, a question was asked to rate the overall result of the all three algorithms per question. More than 100 computer science students participated in this experiment to evaluate KEA, Micro-gram and E^3. And we found that a careful combination of the features we propose yield up better results. However, in user case study we did not compared E^3 with the results of STICS, EMM, EventRegistry as the results retrieve from them are very large for evaluation in single day.



Key Phrase Algorithms Survey

- PLEASE ZOOM IN (Ctrl+)THE WEBPAGE IF YOU ARE NOT ABLE TO READ THE KEYPHRASES CLEARLY.

- Larger word size in the tree denotes higher frequency.

- The order of Algorithms have been shuffled to avoid bias.

* Required

Your Name *

Olympics

Best Coverage/ Diversity *

Does the Key Phrase set cover most of the topics regarding the entity mentioned

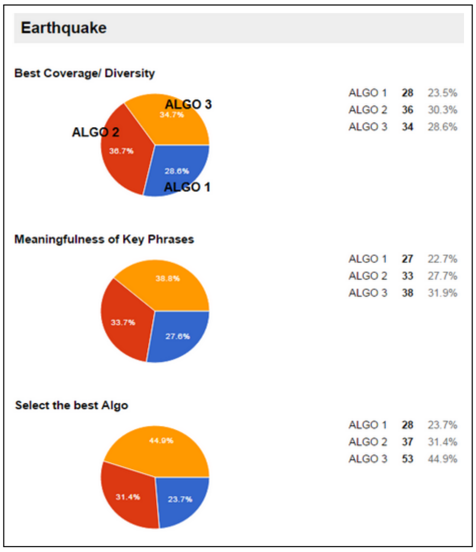
Meaningfulness of Key Phrases *

Select the best Algo *

Figure 5.3: User case study survey form front page

Survey result is available at the link[2] in footnote. Figure 5.4 shows result for keyphrases on the event ‘Earthquake’, given by E^3 (ALGO 3), TopMine (ALGO 1), Micro-ngram (ALGO 2).

[2] `https://goo.gl/fM5yfF`



ALGO 1	28	23.5%
ALGO 2	36	30.3%
ALGO 3	34	28.6%

ALGO 1	27	22.7%
ALGO 2	33	27.7%
ALGO 3	38	31.9%

ALGO 1	28	23.7%
ALGO 2	37	31.4%
ALGO 3	53	44.9%

Figure 5.4: User case study result for 'Earthquake'

The quality based user study showed that our method is fairly robust, as the identification of prominent and interesting news concept does not require any training. The study also show that we are able to improve over previous high performing news exploration, and keyphrase extraction methods (in both single data form-news headline and multi-data form), thus showing for the first time the beneficial effects of exploiting multi-data knowledge in a joint fashion.

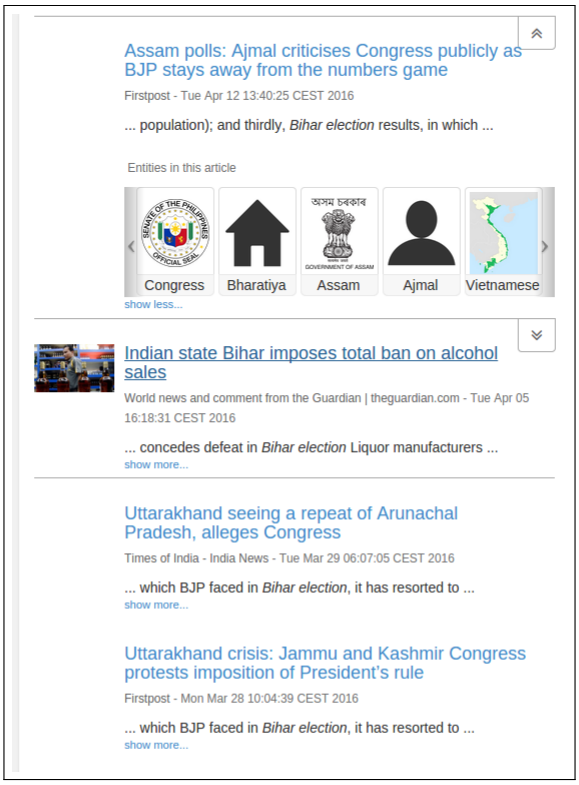
5.4 Quality of System: Example based Comparison

To evaluate the quality of news exploration through our presented system E^3, we compared it with the state-of-the-art event-centric (EventRegistry) and content-centric (STICS) systems. We compared the aforementioned systems with E^3 on the following grounds:

- News Summary generated
- Named Entities discovered
- Information Retrieved with respect to query

5.4.1 News Summary generated

For input query q: 'Bihar Election' results obtained from STICS, EventRegistry and E^3 are shown in Figure 5.5, Figure 5.6 and Figure 5.7 respectively. The total number of documents returned by STICS for q are 355. Number of events returned by EventRegistry are six, where the average number of articles inside each event is 21. And the total number of keyphrases discovered by E^3 for q is 487. E^3 have two views of keyphrase based summary, one is tile based showing keyphrases one by one (Figure 5.7(a)) and another is 3-D tag cloud showing all the keyphrases in one place (Figure 5.7(b)).



Assam polls: Ajmal criticises Congress publicly as BJP stays away from the numbers game

Firstpost - Tue Apr 12 13:40:25 CEST 2016

... population); and thirdly, *Bihar election* results, in which ...

Entities in this article

Indian state Bihar imposes total ban on alcohol sales

World news and comment from the Guardian | theguardian.com - Tue Apr 05 16:18:31 CEST 2016

... concedes defeat in *Bihar election* Liquor manufacturers ...

Uttarakhand seeing a repeat of Arunachal Pradesh, alleges Congress

Times of India - India News - Tue Mar 29 06:07:05 CEST 2016

... which BJP faced in *Bihar election*, it has resorted to ...

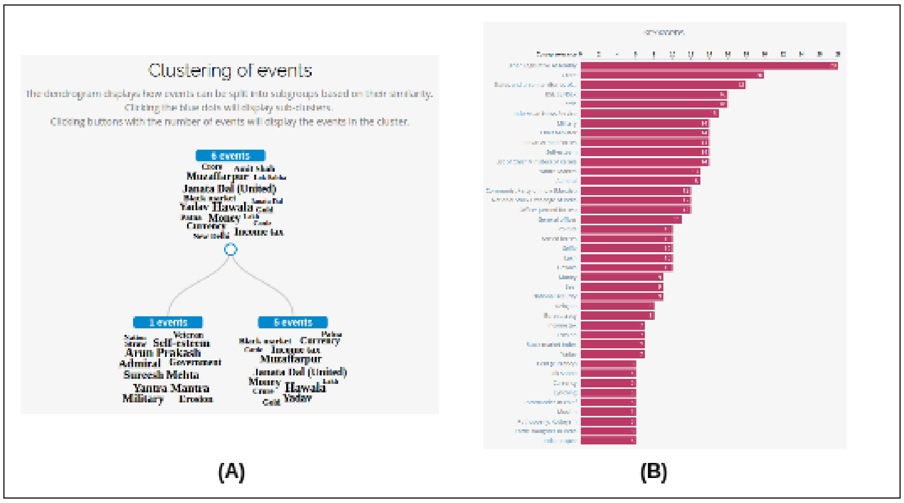
Uttarakhand crisis: Jammu and Kashmir Congress protests imposition of President's rule

Firstpost - Mon Mar 28 10:04:39 CEST 2016

... which BJP faced in *Bihar election*, it has resorted to ...

Figure 5.5: STICS output for 'Bihar Election'

We also compared the aforementioned systems for several other general and specific queries. The number of articles returned by both the systems are still very large to get an overview of an input query and also requires much effort to read large documents. Thus, we can say that here our system E^3 out-performs the



Clustering of events

The dendrogram displays how events can be split into subgroups based on their similarity.
Clicking the blue dots will display sub-clusters.
Clicking buttons with the number of events will display the events in the cluster.

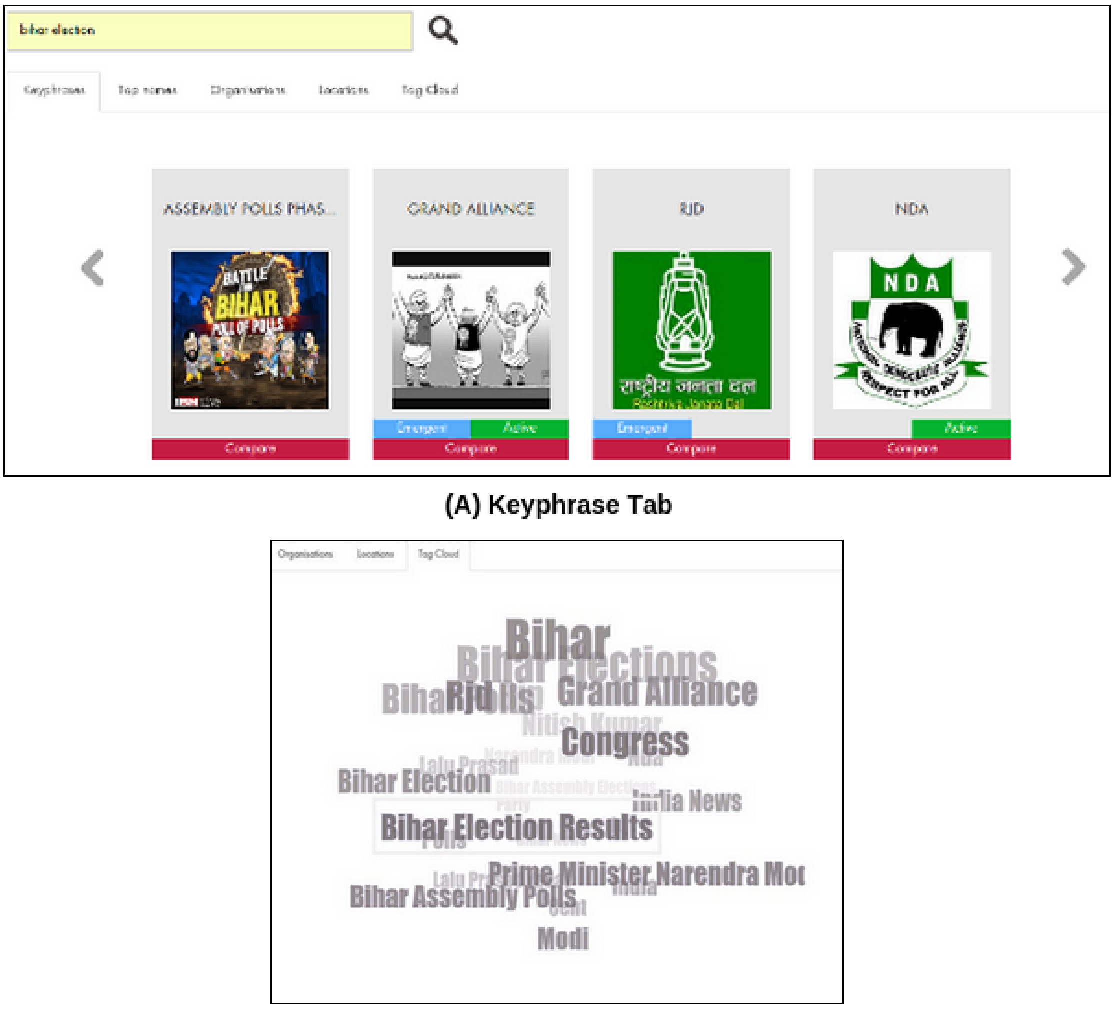
Figure 5.6: EventRegistry output for ‘Bihar Election’

state-of-the-art systems (STICS and EventRegistry).

5.4.2 Named Entities discovered

Since news events are full of actors, having impact on each other. Hence, discovering all named entities present in given news corpora is very important task for further information collection and knowledge enrichment. However, we observed for several news queries ranging from general domain to specific domain very less number of entities are returned by STICS and EventRegistry.

For instance for query ‘Bihar Election’, Figure 5.8 shows that only 20 named entities were found by STICS and Figure 5.9 shows that 40 named entities were discovered by EventRegistry. We observed that STICS at max displays 20 common entities and EventRegistry shows 40. However, for some *General* queries like ‘Olympics’, ‘Elections’, ‘Scam’ have multiple entities are involved which are difficult to summarize in the aforementioned systems. For instance, both the systems did not return some of the popular entities like ‘Jitan Ram Manjhi’, ‘Sonia Gandhi’, ‘Rahul Gandhi’, ‘Congress’ and many others. Whereas our system E^3 returns majority of the named entities involved either having strong impact or low impact. Also E^3 categories the named entities into different sections (Person, Location, Organization and News Concept), for refernce see Figure 5.10.



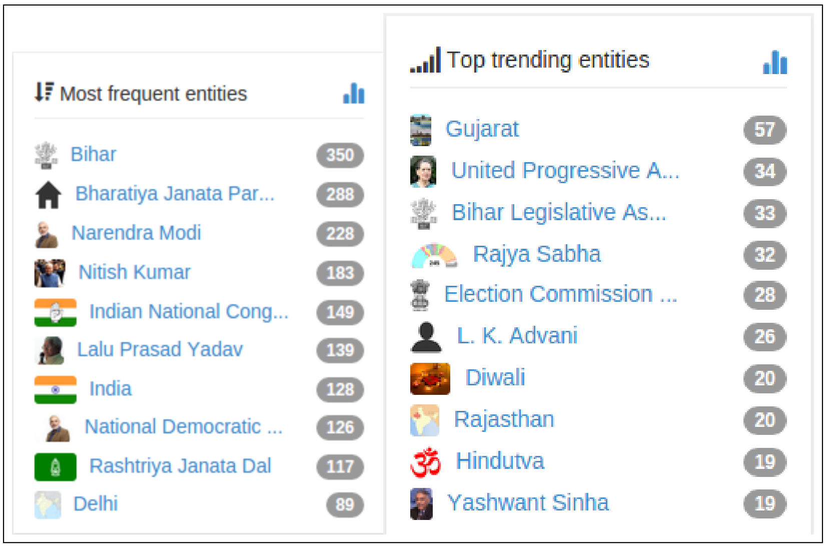
(A) Keyphrase Tab

(B) Tag Cloud

Figure 5.7: E^3 keyphrase outputs for 'Bihar Election' (a) Keyphrase Tab (b) Tag Cloud

5.4.3 Information Retrieved with respect to query

Mostly there are multiple entities associated with a news query. For instance, the number of Person's returned by E^3 for query 'Bihar Election' was more than 127. It's very rare to know about all the entities. Hence, external knowledge base comes to help. However, in order to know details about an entity we need to make a complex query in order to retrieve the correct entity. For instance we queried for 'R. Nagarajan' on Wikipedia, Freebase and BabelNet, present in news concept 'IIT



Most frequent entities

Bihar	350
Bharatiya Janata Par...	288
Narendra Modi	228
Nitish Kumar	183
Indian National Cong...	149
Lalu Prasad Yadav	139
India	128
National Democratic ...	126
Rashtriya Janata Dal	117
Delhi	89

Top trending entities

Gujarat	57
United Progressive A...	34
Bihar Legislative As...	33
Rajya Sabha	32
Election Commission ...	28
L. K. Advani	26
Diwali	20
Rajasthan	20
Hindutva	19
Yashwant Sinha	19

Figure 5.8: Named Entities discovered by STICS for 'Bihar Election'

Fees Hike'. As we can see in the Figure 5.11 'R. Nagarajan' is mislinked to a former Indian cricket umpire, whereas 'R. Nagarajan' is a Professor at IIT Madras. Hence, a reader having poor information about IIT and Indian Cricket may get mislead by the retrieved information. To avoid such misleading and complex query burden, E^3 uses in-built news corpora details to automatically link the entities with correct information.

Also, we found that search engines like Google provides a $General$ role information box for entities, which is mostly not of interest to news audience. Like in Figure 5.12(a) for 'Lalu Prasad Yadav' associated with news query 'Bihar Election', Google search engine returned 'Indian Politician'. Whereas the audience interested in some specific information may not find the information interesting. Hence, to mitigate this problem E^3 displays $Specific$ role played by the entity during the news query time interval. For instance, Figure 5.12(b) shows that for 'Lalu Prasad Yadav' E^3 returned 'Grand Alliance', as he was one of the member of Grand Alliance.

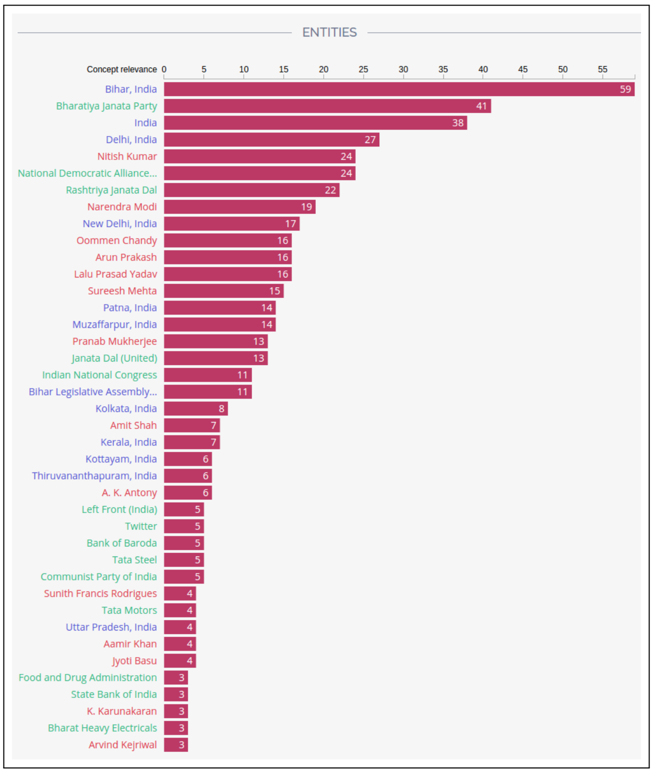


Figure 5.9: Named Entities discovered by EventRegistry for ‘Bihar Election’

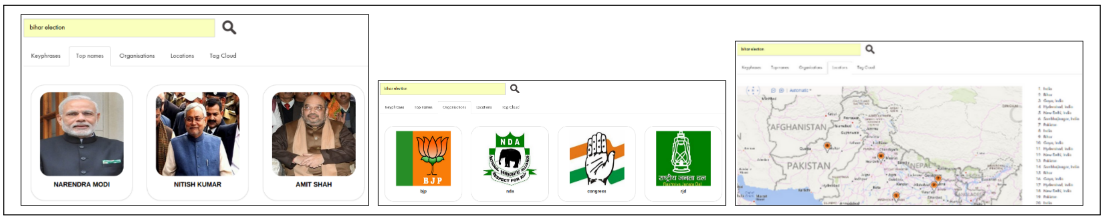
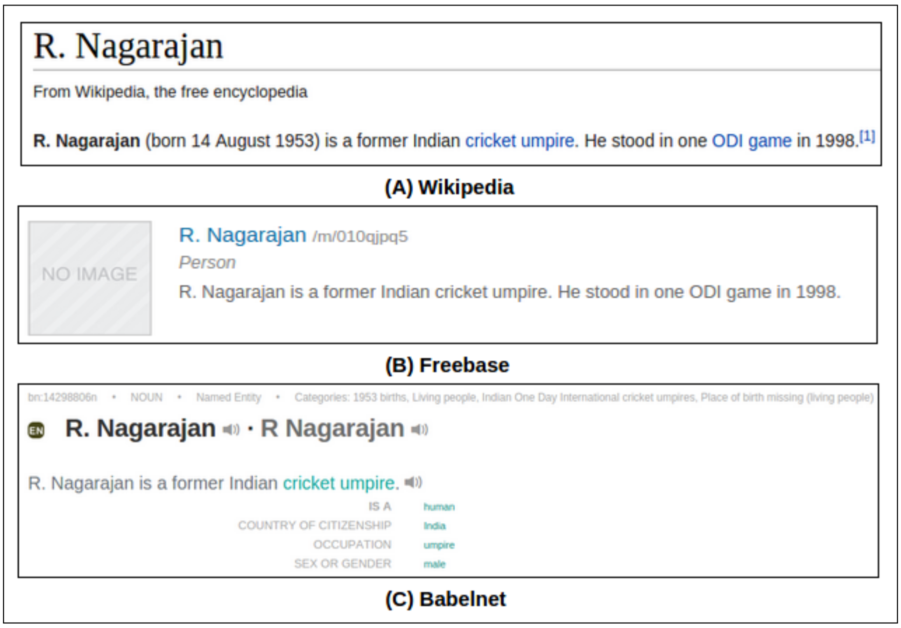


Figure 5.10: Named Entities discovered by E^3 for ‘Bihar Election’ (showing top most row)



R. Nagarajan

From Wikipedia, the free encyclopedia

R. Nagarajan (born 14 August 1953) is a former Indian cricket umpire. He stood in one ODI game in 1998.[1]

(A) Wikipedia

R. Nagarajan /m/010qjpq5

Person

R. Nagarajan is a former Indian cricket umpire. He stood in one ODI game in 1998.

(B) Freebase

bn:14298806n • NOUN • Named Entity • Categories: 1953 births, Living people, Indian One Day International cricket umpires, Place of birth missing (living people)

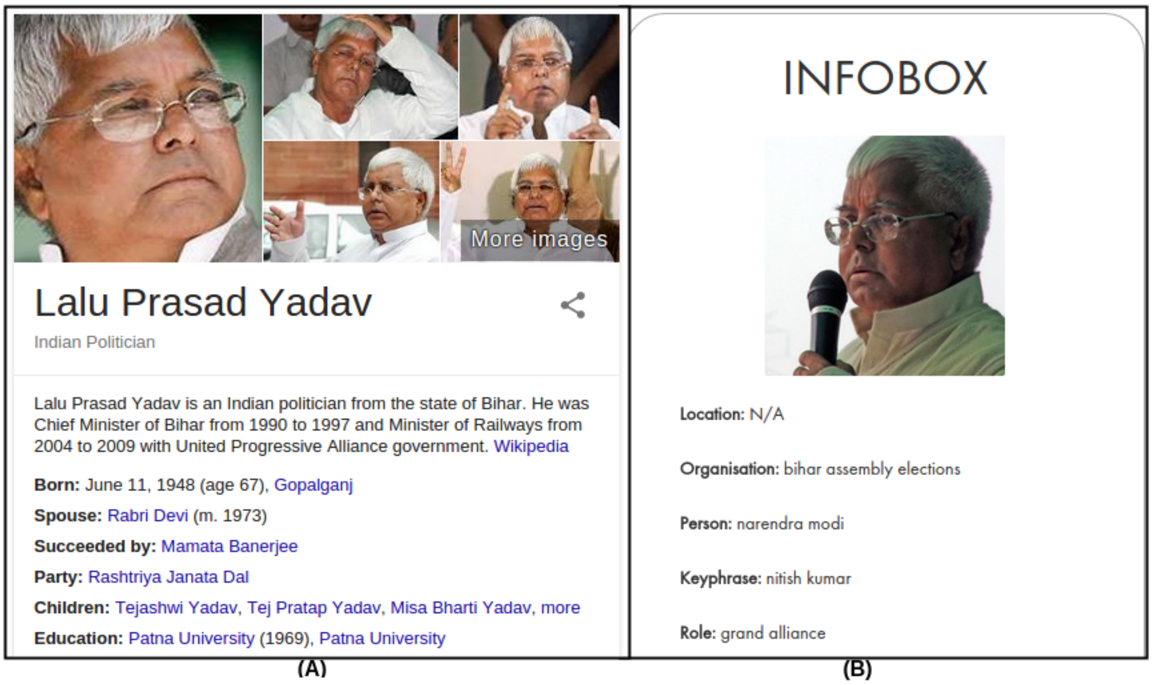
EN **R. Nagarajan** · R Nagarajan

R. Nagarajan is a former Indian cricket umpire.

IS A	human
COUNTRY OF CITIZENSHIP	India
OCCUPATION	umpire
SEX OR GENDER	male

(C) Babelnet

Figure 5.11: Information found in knowledge base for R. Nagarajan



Lalu Prasad Yadav

Indian Politician

Lalu Prasad Yadav is an Indian politician from the state of Bihar. He was Chief Minister of Bihar from 1990 to 1997 and Minister of Railways from 2004 to 2009 with United Progressive Alliance government. Wikipedia

Born: June 11, 1948 (age 67), Gopalganj

Spouse: Rabri Devi (m. 1973)

Succeeded by: Mamata Banerjee

Party: Rashtriya Janata Dal

Children: Tejashwi Yadav, Tej Pratap Yadav, Misa Bharti Yadav, more

Education: Patna University (1969), Patna University

(A)

INFOBOX

Location: N/A

Organisation: bihar assembly elections

Person: narendra modi

Keyphrase: nitish kumar

Role: grand alliance

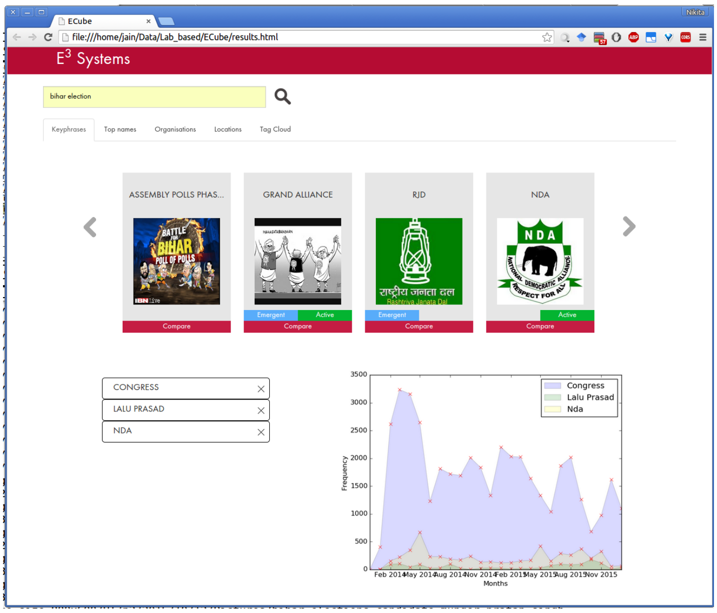
(B)

Figure 5.12: Information Box for 'Lalu Prasad Yadav' associated with 'Bihar Election' (a) Google Search (b) E^3 System

CHAPTER 6

System Demonstration and Results

To mitigate the information overflow problem in news domain, we have contrived an easy to use and robust user interface for efficient keyphrase based news event exploration. Homepage of our system E^3 is shown in Figure 6.15. To enable the audience with broad overview of news data E^3 extracts and enriches keyphrases by tagging, ranking and finding role related to the input query.



file:///home/jain/Data/Lab_based/ECube/results.html

bihar election

Figure 6.1: E^3 Homepage for ‘Bihar Election’

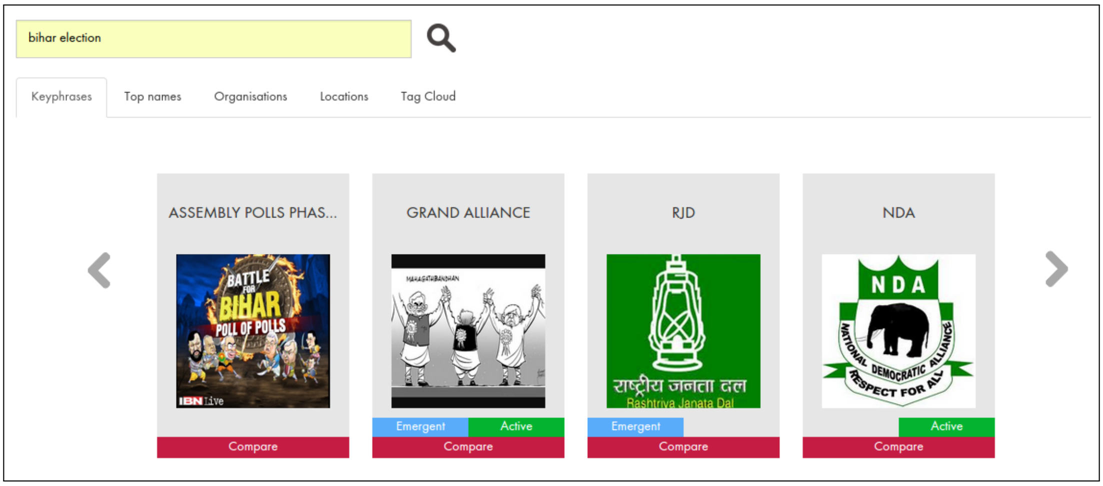
Following are the main functionality offered by the system E^3:

- Keyphrase based Summary
- Timeline based Keyphrase Comparison
- Top associated persons and their InfoBox
- Organisations
- Locations plotted on Bing maps
- Tag Cloud of keyphrases
- Time-interval and histogram of associated keyphrases

The front end of the system is developed in HTML and Bootstrap CSS. The back-end comprises of python scripts and DJango Framework. The system stores the results into a SQLite database so that system does not have to process over and over again on the same data. The detailed information about each of the features provided by the system are:

Keyphrase based Summary: For an input query we display the E^3's extracted keyphrases in a bi-directional lane. The keyphrases are sorted by their importance derived from the number of occurrences in collected news corpus. The keyphrase are also annotated with novel and active tags depending upon their novelty and activeness value. Also, we augment images acquired from DuckDuckGo API to the keyphrases. Figure 6.2 shows a section of keyphrases extracted by E^3 for 'Bihar Election'. Also, in Figure 6.2 novel and active keyphrases annotated with tags at bottom of tile.

Timeline based Keyphrase Comparison: Using this functionality we can compare the keyphrases with help of their relative temporal activities. On clicking 'Plot Timeline', an overlapping time-series plot is generated. The horizontal axis is the time scale, ranging over the time period of input query, while the vertical axis shows the combined frequency of occurrence of keyphrase in the collected news corpus. Figure 6.3 shows the timeline comparison for 'Bihar Election' keyphrases: 'RJD', 'Grand Alliance', 'Bihar Election'. Hence, we can infer that on an average



bihar election

Figure 6.2: E^3 Keyphrase based Summary for 'Bihar Election'

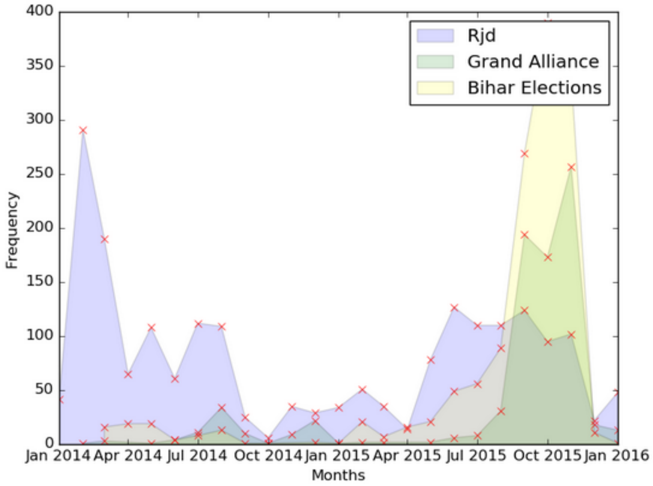


Figure 6.3: E^3 Timeline based Keyphrase Comparison

'RJD' is more popular then 'Grand Alliance', however during September, 2015 to January, 2016 'Grand Alliance' is more on rise in 'Bihar Election'.

Top names and Infobox: Here we provide the top names associated with the input query, sorted by their frequency. We augment each person tile with its image using DuckDuckGo API. Clicking on a person's name generates a InfoBox on right side of the screen. The InfoBox comprises the mostly associated Location, Organisation, People and keyphrases during the input query time interval. The

InfoBox also display information related to the role a person play during the input query time interval.

Figure 6.4(a) Histogram shows the togetherness of a keyphrase with the InfoBox entity using their co-occurrence value. Horizontal axis shows the keyphrases and Vertical axis shows their frequency of occurrence. An option to plot-timeline for the InfoBox entity is also provided. Figure fig:E3-histogram(b) shows the timeline as a pop-up for 'Bihar Election' top associated entity: 'Lalu Prasad Yadav'.

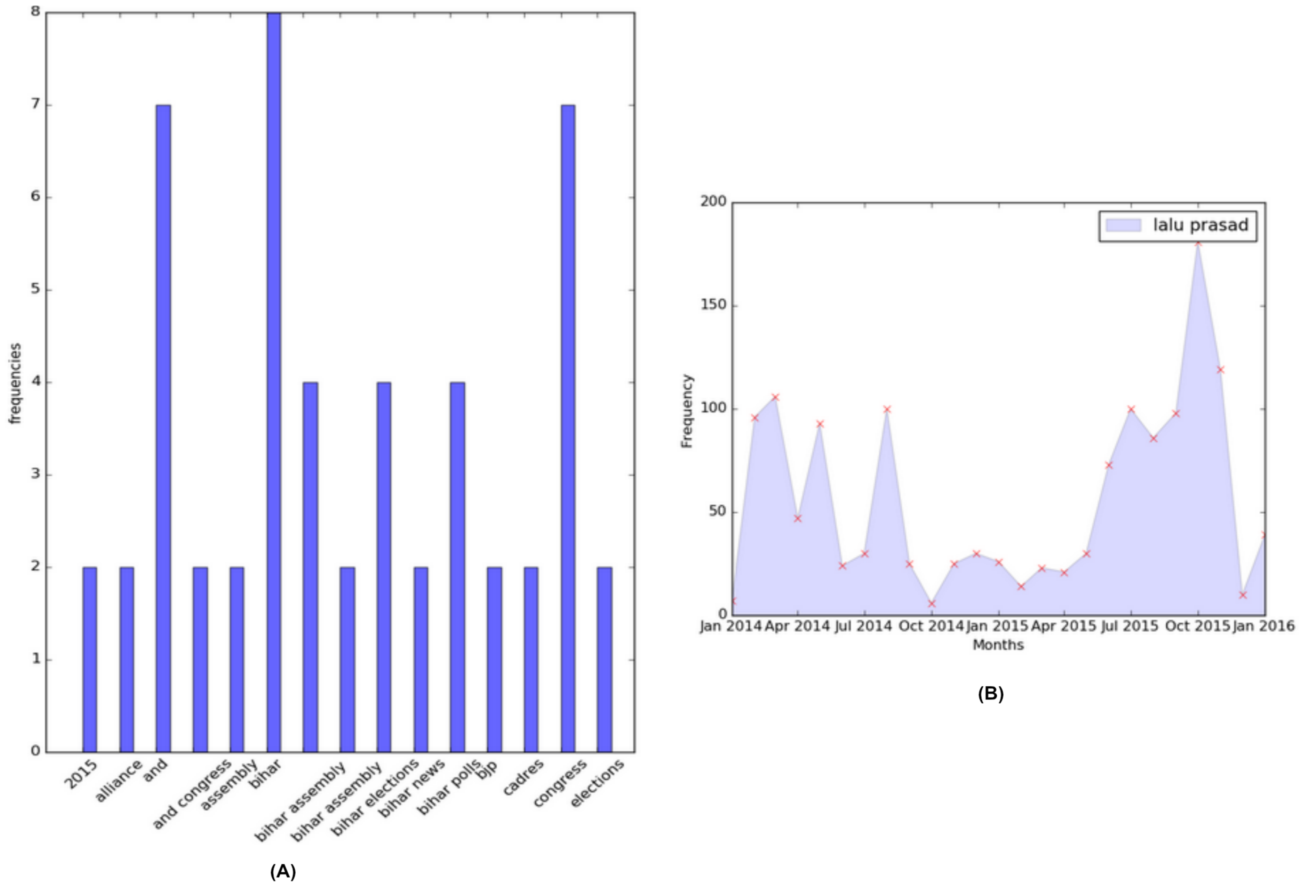
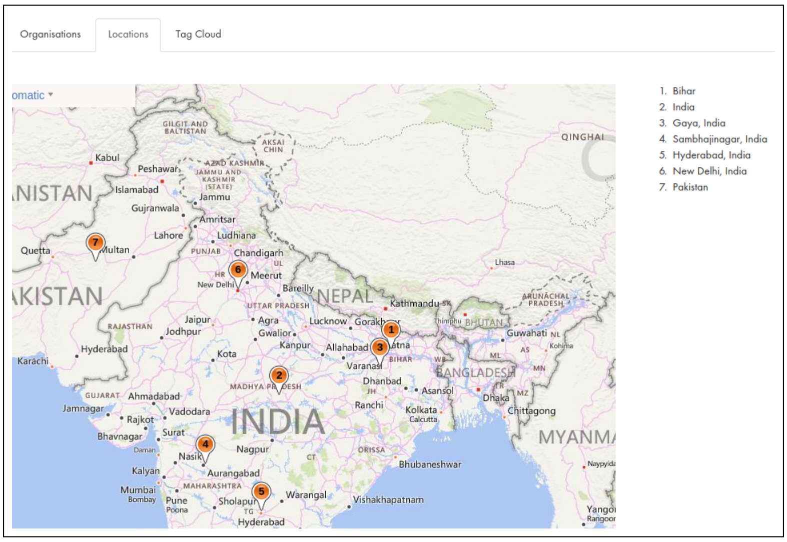


Figure 6.4: E^3 Keyphrase(a) Histogram and (b) Time-plot for 'Lalu Prasad Yadav'

Organisations: In this tab we provide the name of top most associated organisations with the search query.

Locations: Locations related to the input query are shown using Bing Map jotted with push-pins [18]. The legend of the push-pins is provided in a column to the right, see Figure 6.5.



1. Bihar
2. India
3. Gaya, India
4. Sambhajinagar, India
5. Hyderabad, India
6. New Delhi, India
7. Pakistan

Figure 6.5: E^3 Locations for ‘Bihar Election’

Tag Cloud: As shown in Figure 6.6, a 3-D structure cloud is used to summarize all the related keyphrases in an unified place. The distance between the various keyphrases shows their relative degree of relatedness, where the most related ones are near and least related ones are a at distance. The font size of each keyphrase is proportional to its impact over the input query, measured through its number of occurrence in the collected news data. Since the tag cloud is interactive, so a user can focus on his/her interested entity just by moving cursor around.

6.1 News Event Database Collection

We collected keyphrases and enrichment data for more than 50 queries. We collected data on varying input queries, from general topics (Eg. Election) to specific topics (Eg. Bihar Election). And stored in a SQLite database.

We have nearly 8 tables to store all the keyphrases and related data for varying input queries. The tables with data related to search query ‘Bihar Election’ are shown below. Note: In any table, its tupple having value as “N/A” indicates that no information is available for that Table tupple and attribute pair.

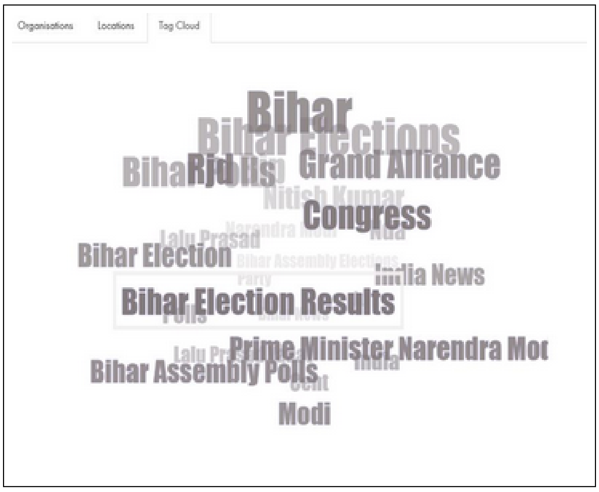
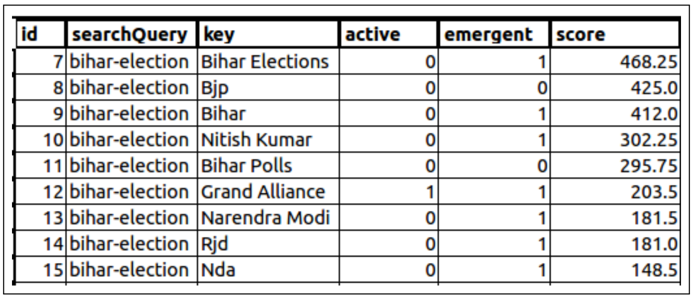


Figure 6.6: E^3 Keyphrase Tag Cloud for ‘Bihar Election’

Figure 6.9 shows the screen-shot for Table storing keyphrases for requested input query. The Primary key for the Table is ⟨`searchQuery, key`⟩. For each ⟨`searchQuery, key`⟩ pair active, novel (emergent) and frequency of occurrence (score) is stored inside the Table, which are further used to display the keyphrases.

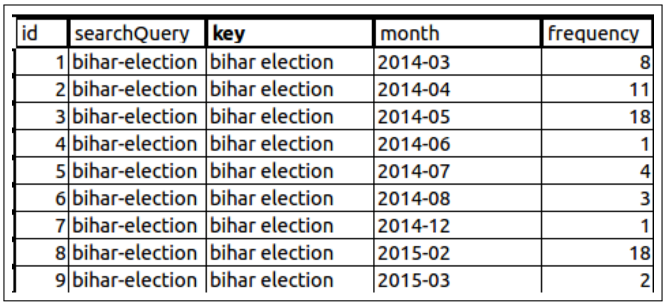


id	searchQuery	key	active	emergent	score
7	bihar-election	Bihar Elections	0	1	468.25
8	bihar-election	Bjp	0	0	425.0
9	bihar-election	Bihar	0	1	412.0
10	bihar-election	Nitish Kumar	0	1	302.25
11	bihar-election	Bihar Polls	0	0	295.75
12	bihar-election	Grand Alliance	1	1	203.5
13	bihar-election	Narendra Modi	0	1	181.5
14	bihar-election	Rjd	0	1	181.0
15	bihar-election	Nda	0	1	148.5

Figure 6.7: E^3 database Table screen-shot for Keywords

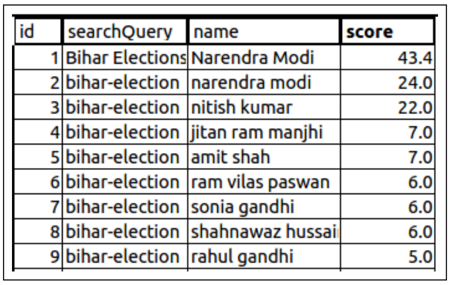
Figure 6.8 shows the screen-shot for Table storing month-wise occurrence of a keyphrase with respect to input query. The Primary key for the Table is ⟨`searchQuery, key, month`⟩. This information is used to plot the timeline for a given entity.

For each searchQuery and the *Person* entities, Table shown in Figure 6.9 stores the data with their frequency of occurrence (score), showing relatedness of entity with the searchQuery. The Primary key for the Table is ⟨`searchQuery, name`⟩.



id	searchQuery	**key**	month	frequency
1	bihar-election	bihar election	2014-03	8
2	bihar-election	bihar election	2014-04	11
3	bihar-election	bihar election	2014-05	18
4	bihar-election	bihar election	2014-06	1
5	bihar-election	bihar election	2014-07	4
6	bihar-election	bihar election	2014-08	3
7	bihar-election	bihar election	2014-12	1
8	bihar-election	bihar election	2015-02	18
9	bihar-election	bihar election	2015-03	2

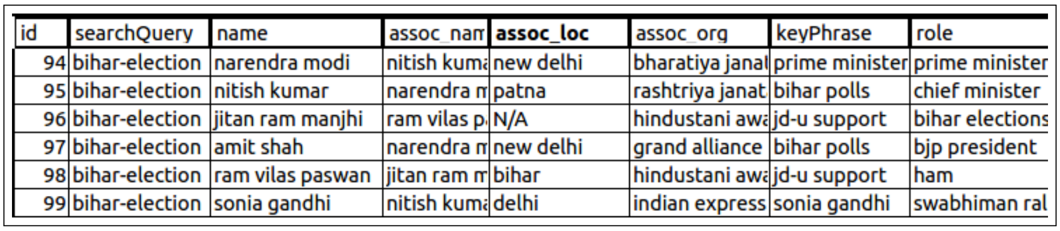
Figure 6.8: E^3 database Table screen-shot for Timeline



id	searchQuery	name	**score**
1	Bihar Elections	Narendra Modi	43.4
2	bihar-election	narendra modi	24.0
3	bihar-election	nitish kumar	22.0
4	bihar-election	jitan ram manjhi	7.0
5	bihar-election	amit shah	7.0
6	bihar-election	ram vilas paswan	6.0
7	bihar-election	sonia gandhi	6.0
8	bihar-election	shahnawaz hussai	6.0
9	bihar-election	rahul gandhi	5.0

Figure 6.9: E^3 database Table screen-shot for Names

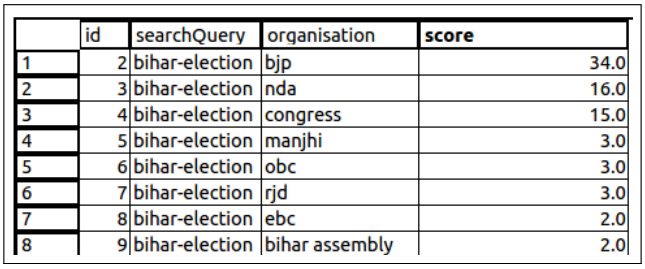
Figure 6.10 shows the screen-shot for Table storing information for each searchQuery and entity most related entities of each category like name (assoc_name), location (assoc_loc), organisation (assoc_org), keyphrase (keyPhrase) and the role of played by entity related to searchQuery. The Primary key for the Table is ⟨`searchQuery, name`⟩. This information is used to design InfoBox for a *Person* entity.



id	searchQuery	name	assoc_nam	**assoc_loc**	assoc_org	keyPhrase	role
94	bihar-election	narendra modi	nitish kuma	new delhi	bharatiya janat	prime minister	prime minister
95	bihar-election	nitish kumar	narendra m	patna	rashtriya janat	bihar polls	chief minister
96	bihar-election	jitan ram manjhi	ram vilas pa	N/A	hindustani awa	jd-u support	bihar elections
97	bihar-election	amit shah	narendra m	new delhi	grand alliance	bihar polls	bjp president
98	bihar-election	ram vilas paswan	jitan ram m	bihar	hindustani awa	jd-u support	ham
99	bihar-election	sonia gandhi	nitish kuma	delhi	indian express	sonia gandhi	swabhiman ral

Figure 6.10: E^3 database Table screen-shot for Person

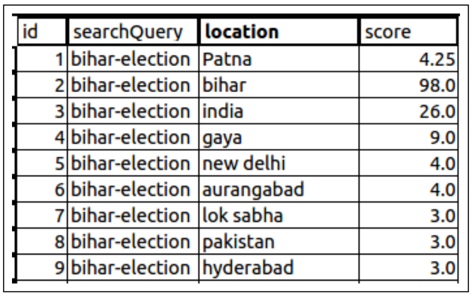
Figure 6.11 shows the screen-shot for Table storing information for each searchQuery and *Organisation* type entity with a score indicating the relative importance of the organisation with respect to the searchQuery. The Primary key for the Table is ⟨`searchQuery, organisation`⟩.



	id	searchQuery	organisation	score
1	2	bihar-election	bjp	34.0
2	3	bihar-election	nda	16.0
3	4	bihar-election	congress	15.0
4	5	bihar-election	manjhi	3.0
5	6	bihar-election	obc	3.0
6	7	bihar-election	rjd	3.0
7	8	bihar-election	ebc	2.0
8	9	bihar-election	bihar assembly	2.0

Figure 6.11: E^3 database Table screen-shot for Organisation

Figure 6.12 shows the screen-shot for Table storing information for each searchQuery and *Location* type entity with a score indicating the relative importance of the location with respect to the searchQuery. The Primary key for the Table is ⟨`searchQuery, location`⟩.

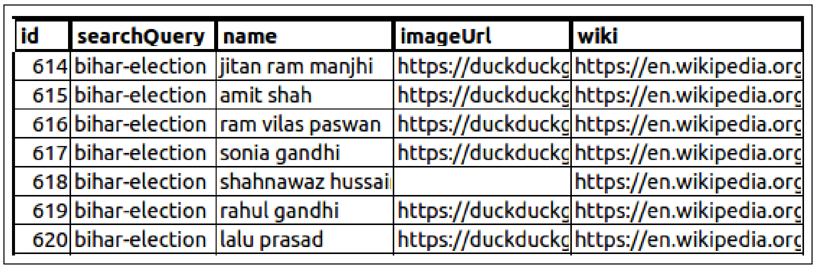


id	searchQuery	location	score
1	bihar-election	Patna	4.25
2	bihar-election	bihar	98.0
3	bihar-election	india	26.0
4	bihar-election	gaya	9.0
5	bihar-election	new delhi	4.0
6	bihar-election	aurangabad	4.0
7	bihar-election	lok sabha	3.0
8	bihar-election	pakistan	3.0
9	bihar-election	hyderabad	3.0

Figure 6.12: E^3 database Table screen-shot for Location

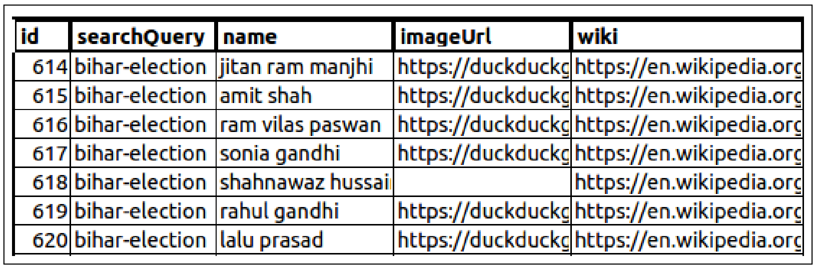
Figure 6.13 shows the screen-shot for Table storing information for each entity found for searchQuery, an URL of image is fetched from DuckDuckGo and stored along with the wikipedia link for the name. This URL and wikipedia link is used load the image and to give a brief description about the entity. The Primary key for the Table is ⟨`searchQuery, name`⟩.

Figure 6.14 shows the screen-shot for Table storing information for each entity (name) found in the searchQuery its associated keyphrase and its frequency is stored. The data stored in Table is further use to plot histogram for entity and its associated keyphrases. The Primary key for the Table is ⟨`searchQuery, name, keyphrase`⟩.



id	searchQuery	name	imageUrl	wiki
614	bihar-election	jitan ram manjhi	https://duckduckg	https://en.wikipedia.org
615	bihar-election	amit shah	https://duckduckg	https://en.wikipedia.org
616	bihar-election	ram vilas paswan	https://duckduckg	https://en.wikipedia.org
617	bihar-election	sonia gandhi	https://duckduckg	https://en.wikipedia.org
618	bihar-election	shahnawaz hussai		https://en.wikipedia.org
619	bihar-election	rahul gandhi	https://duckduckg	https://en.wikipedia.org
620	bihar-election	lalu prasad	https://duckduckg	https://en.wikipedia.org

Figure 6.13: E^3 database Table screen-shot for ImageURL



id	searchQuery	name	imageUrl	wiki
614	bihar-election	jitan ram manjhi	https://duckduckg	https://en.wikipedia.org
615	bihar-election	amit shah	https://duckduckg	https://en.wikipedia.org
616	bihar-election	ram vilas paswan	https://duckduckg	https://en.wikipedia.org
617	bihar-election	sonia gandhi	https://duckduckg	https://en.wikipedia.org
618	bihar-election	shahnawaz hussai		https://en.wikipedia.org
619	bihar-election	rahul gandhi	https://duckduckg	https://en.wikipedia.org
620	bihar-election	lalu prasad	https://duckduckg	https://en.wikipedia.org

Figure 6.14: E^3 database Table screen-shot for Histogram

A series of python scripts are executed whenever a new query is made to E^3 to update the Database. Some of the major tasks done by scripts are: adding all the keyphrases, with the information about their activeness and novelty, histogram and time period of involvement. Finding images and Wikipedia links related to keyphrases, and further inserting them into the database. Also, adds associated names, locations, organizations, top keyphrase and the role of entity in relation to the news query.

6.2 Domain Specific Data Analysis

6.2.1 Entertainment

We collected news data for several movies released in 2016 (see Figure 6.15). Some of them are: 'Baahubali', 'X-men', 'Deadpool', 'Batman v Superman', 'Kung Fu Panda', 'Star Wars'. We extracted all the above movies plot keywords from IMDB site and used them as baseline to compare with our system E^3 generated keyphrases. On comparison we found on an average more than 20% of the IMDB manually created keywords completely match with keyphrases generated from E^3. On careful observation, we found that there are also some partial matches like 'Female Leads' (E^3 Keyword), 'Female Protagonist' (IMDB plot keyword); 'Fathers Ignoble Death' (E^3 Keyword), 'Death of Father' (IMDB plot keyword); and many such having similar meanings to IMDB plot keywords. The number of such partial matches we found on an average is more than 30%. Including both the types of matches more than 53% of matches were in common with the IMDB plot keywords.

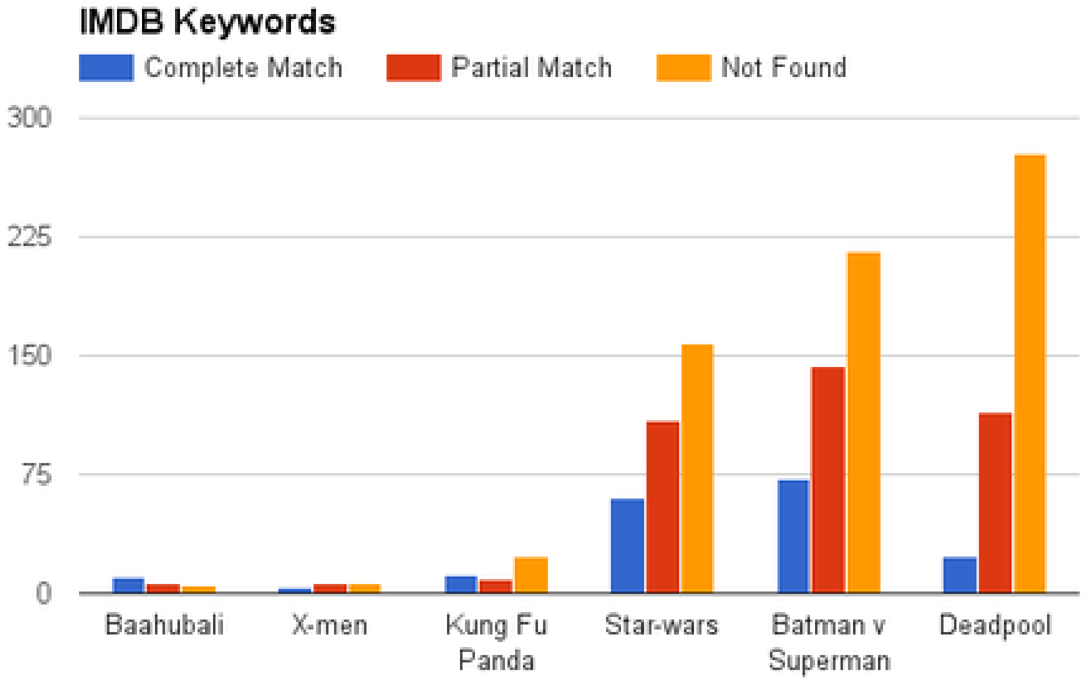
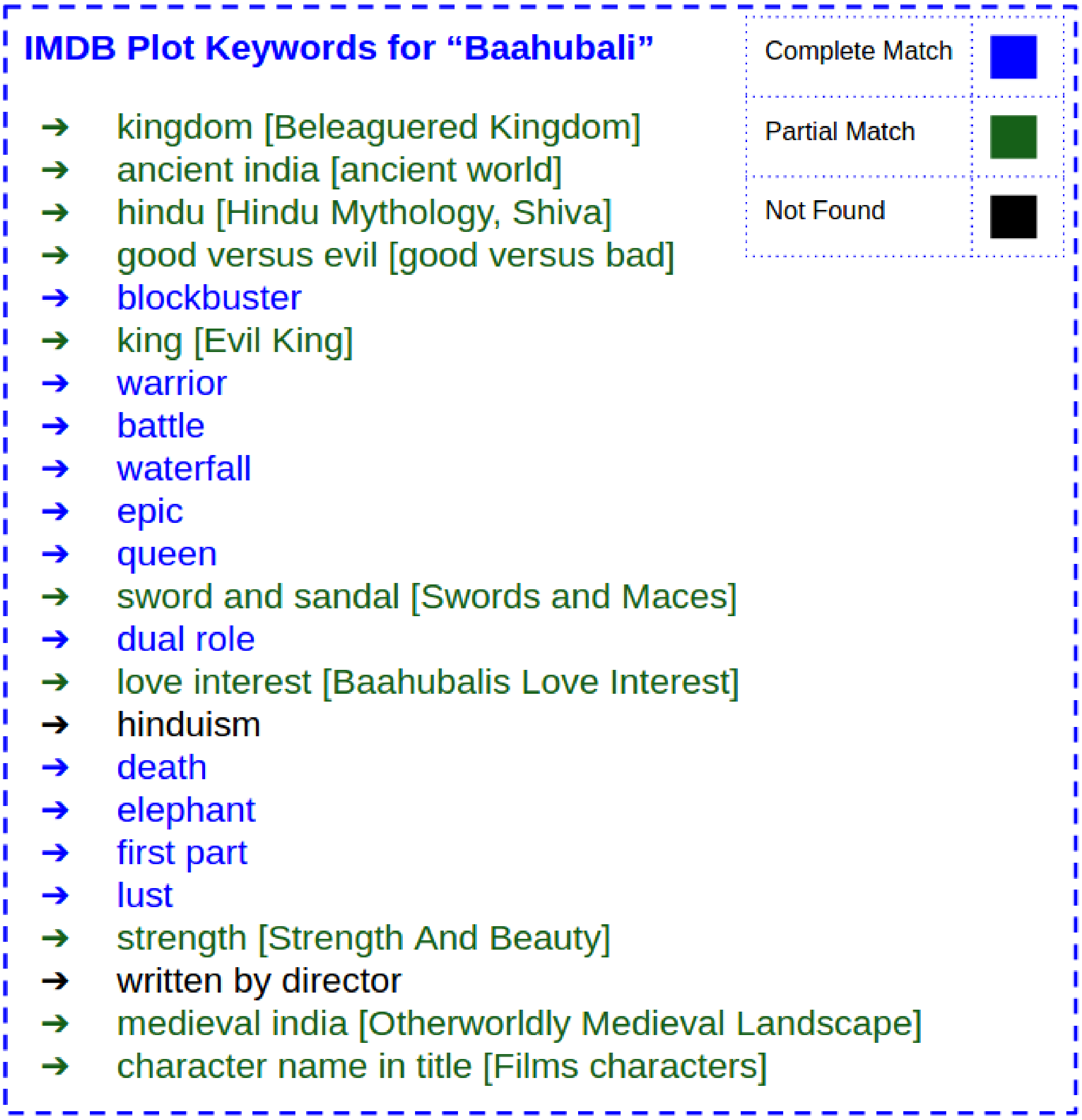


Figure 6.15: IMDB Plot Keywords

On manual inspection, we found that most of the plot keywords like 'world destruction', 'telekinesis' are present in 'Star Wars' IMDB plot keywords, however they are not found in news data related to 'Star Wars'. And as we are extracting keyphrases, not generating them hence its not even possible to retrieve them. In Figure 6.16, we show all IMDB plot keywords for Bollywood Movie *Baahubali*. From the Figure 6.16, we can see that good number of complete and partial matches are found by E^3. Some of the best suitable matches in case of partial match is shown inside the square bracket. Indeed, we observed for other movies also, that in case of partial match E^3 best partial matches are better in explaining the concept.



IMDB Plot Keywords for "Baahubali"

Complete Match	🟦
Partial Match	🟩
Not Found	⬛

- ➔ kingdom [Beleaguered Kingdom]
- ➔ ancient india [ancient world]
- ➔ hindu [Hindu Mythology, Shiva]
- ➔ good versus evil [good versus bad]
- ➔ blockbuster
- ➔ king [Evil King]
- ➔ warrior
- ➔ battle
- ➔ waterfall
- ➔ epic
- ➔ queen
- ➔ sword and sandal [Swords and Maces]
- ➔ dual role
- ➔ love interest [Baahubalis Love Interest]
- ➔ hinduism
- ➔ death
- ➔ elephant
- ➔ first part
- ➔ lust
- ➔ strength [Strength And Beauty]
- ➔ written by director
- ➔ medieval india [Otherworldly Medieval Landscape]
- ➔ character name in title [Films characters]

Figure 6.16: IMDB Plot Keywords v/s E^3 Keyphrases

Hence, we can say that our system E^3 works well even for specific domain data without any formal domain level training.

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this thesis, we presented an approach to summarize the news data using keyphrases. To extract keyphrases from heterogeneous news data we introduced syntactic and linguistic features with multilingual capability. Furthermore, we developed a news event exploration engine E^3 for exploration and enrichment of news event and concepts. The engine classifies the keyphrases into $Active$, $Novel$ and None; tags keyphrases into Person, Location, Organization and News Concept. E^3 also ranks keyphrases using value of activeness, novelty or frequency and identifies the role of associated tags with respect to news query. The quantitative and qualitative results from performed experiments confirms that E^3 outperforms state-of-the-art system in informative and interestingness aspects.

Future Work: The available keyphrase extraction algorithm is not very efficient with respect to running time, because of dependency on Named Entity Recognizer (NER) and Part of Speech (POS) Tagger. Our algorithm is currently for English, but it is possible to convert this algorithm to different languages. The language dependency of this algorithm is mainly caused by NER and POS tagger. Only NER and Part of speech tagger rules are language dependent. In the future, we will study keyphrase extraction from audio scripts for more information retrieval.

APPENDIX A

NEWS HEADLINE AND BNC COMPARISON

The purpose of the comparison is to determine the extent to which the New Headline corpus resembles characteristics of standard text corpus. Our comparison of the News Headlines corpus with the British National Corpus (BNC) is based primarily upon analysis of the part-of-speech content of each corpus. The details of English News headline and BNC corpus are shown in Table A.1 and Table A.2. As there are multiple variations between the news headline and the BNC corpus in terms of POS distribution (see Figure A.1 and Figure A.2). Some notable exceptions are:

- Great proportion of nouns, adverbs and determiners in News Headlines than BNC.
- Wide difference in the proportion of adjectives and verbs.

Hence, the Natural language processing and Text Mining tools developed for standard text does not works well for News Media corpus. Therefore, we need to have News specific tools.

Table A.1: English News Headline Corpus

Number of Words	17 Million
Average Headline Words	10.22
Number of Headlines	10 Lakh

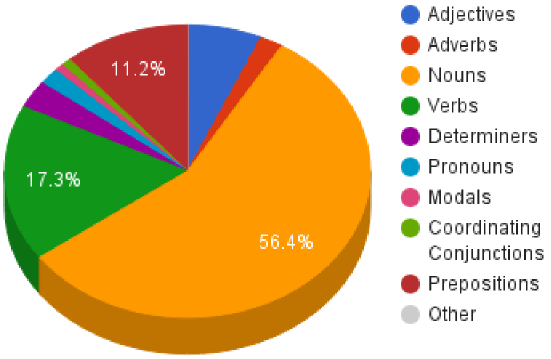


Figure A.1: News Headline Keyword Statistics

Table A.2: English British National Corpus (BNC)

No of Lemmatised Words	6318
Average word occurrence	800
Total number of words	100 Million

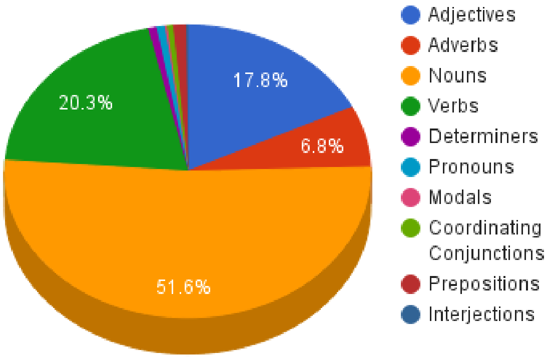


Figure A.2: BNC Keyword Statistics

Appendix B

Indian Named Entity Recognition

SINCE our problem focuses on retrieving the associated named entities, Named Entity Recognition was the natural choice over a list of available Indian names. However, existing NER taggers do not perform well for Indian named entities. Some of the prominent reason behind this anomaly are: `spelling variations` (Example, “Roza” and “Rosa”), `ambiguous names` (Example, “Kamal” is a name of flower and also a name of a person) and several other.

Therefore, we use two different state-of-the-art NER for the tagging purpose. For instance, for the headline “Unstoppable Yuvraj Singh Targets to Take 2019 World Cup By Storm ” Stanford-NER fails to detect “Yuvraj Singh” were as Illinois-NER identifies the entity. Still, there are cases where both fails to detect entities, for instance, “Mahagathbandhan to form govt: Nagmani”, both the state-of-the-art NERs fails to identify “Mahagathbandhan” and “Nagmani” as a named entity. Hence, we use a separate list[1] of Indian named entities generated from the dissertation work of **Avinash Kumar**, Integrated Dual Degree-2015 for the tagging purpose. The list contains more than 55 thousand Hindi, English parallel named entities.

[1] `https://github.com/NikkiJain09/Transliteration`

Bibliography

[1] *GDELT: Global Database of Events, Language, and Tone*.
http://gdeltproject.org/data.html, Last accessed on 27 April, 2016.

[2] *Event Registry*.
http://eventregistry.org, Last accessed on 27 April, 2016.

[3] *STICS: Searching With Strings, Things, And Cats*.
https://stics.mpi-inf.mpg.de, Last accessed on 27 April, 2016.

[4] J. G. Fiscus and G. R. Doddington, *Topic Detection and Tracking*.
In Topic detection and tracking, James Allan (Ed.)., 2002.

[5] P. D. Turney, "Coherent keyphrase extraction via web mining," Computing Research Repository, 2003.

[6] L. Sterckx, T. Demeester, J. Deleu, and C. Develder, "When topic models disagree: Keyphrase extraction with multiple topic models," World Wide Web Companion, 2015.

[7] Y.-f. B. Wu, Q. Li, R. S. Bot, and X. Chen, "Domain-specific keyphrase extraction," Conference on Information and Knowledge Management, 2005.

[8] O. Medelyan and I. H. Witten, "Thesaurus based automatic keyphrase indexing," Joint Conference on Digital Libraries, 2006.

[9] O. Medelyan, *Human-competitive automatic topic indexing*.
The University of Waikato, Ph.D. thesis, July 2009.

[10] K. Wang, C. Thrasher, E. Viegas, X. Li, and B.-j. P. Hsu, "An overview of microsoft web n-gram corpus and applications," Human Language Technologies- ACL, 2010.

[11] C. W. C. R. V. J. H. Ahmed El-Kishky, Yanglei Song, "Scalable topical phrase mining from text corpora," Very Large Data Bases, 2014.

[12] J. Liu, J. Shang, C. Wang, X. Ren, and J. Han, "Mining quality phrases from massive text corpora," Special Interest Group on Management of Data, 2015.

[13] K. Leetaru and P. A. Schrodt, "Gdelt: Global data on events, location, and tone," International Studies Association Annual Convention, 2013.

[14] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik, "Event registry: Learning about world events from news," World Wide Web Companion, 2014.

[15] J. Hoffart, D. Milchevski, and G. Weikum, "Stics: Searching with strings, things, and cats," Special Interest Group on Information Retrieval, 2014.

[16] R. Steinberger, B. Pouliquen, and E. V. der Goot, "An introduction to the europe media monitor family of applications," Special Interest Group on Information Retrieval, 2009.

[17] S. Mazumder, B. Bishnoi, and D. Patel, "News headlines: What they can tell us?," IBM Collaborative Academia Research Exchange, 2014.

[18] *Microsoft Bing Maps*.
https://www.microsoft.com/maps, Last accessed on 27 April, 2016.