Automatic Keyphrase Extraction for News Media

"Want information, not documents"

--User

Supervised By :

Dr. Dhaval Patel

Computer Science Department

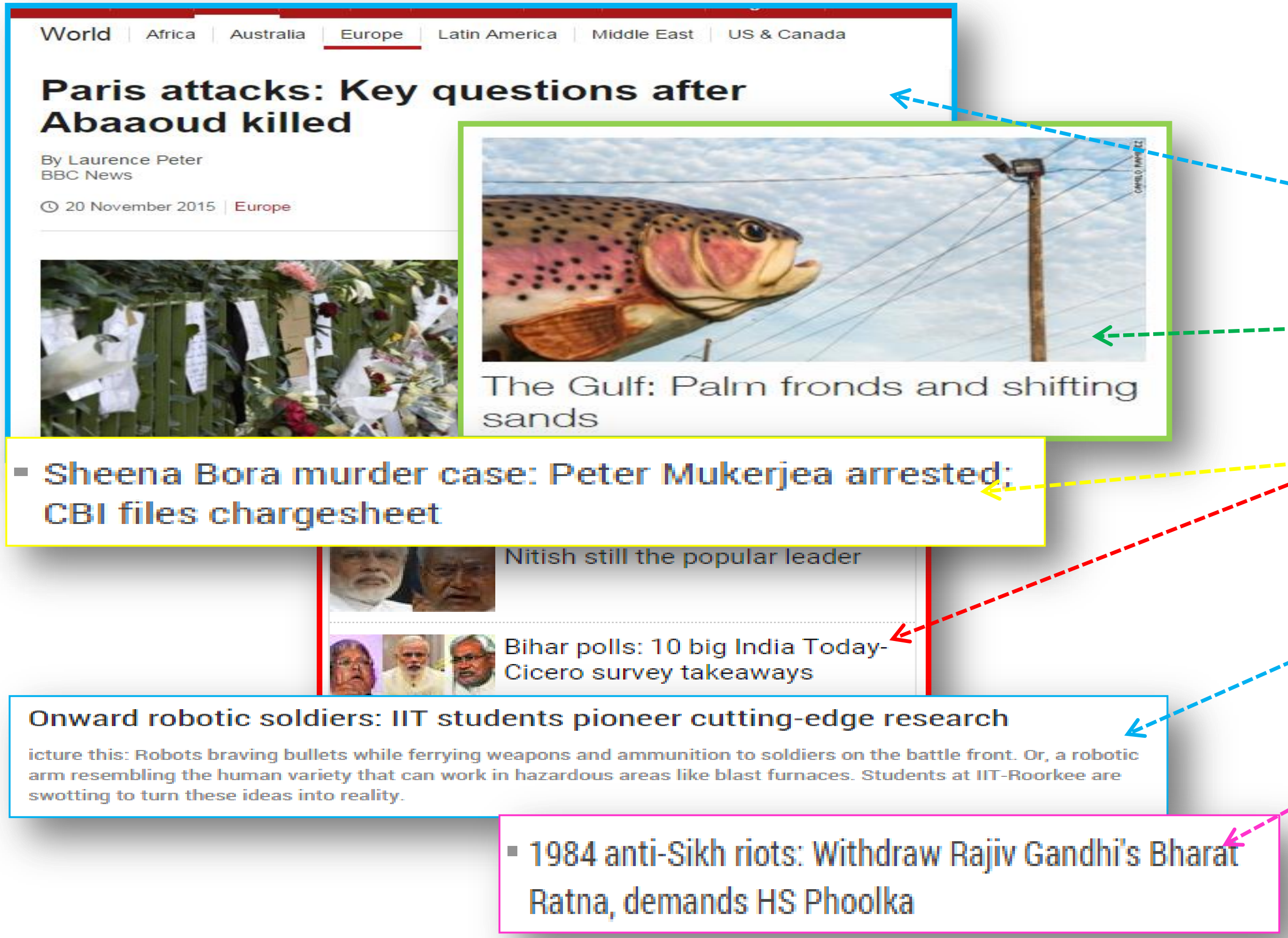
Presented By :

Nikita Jain

14535031

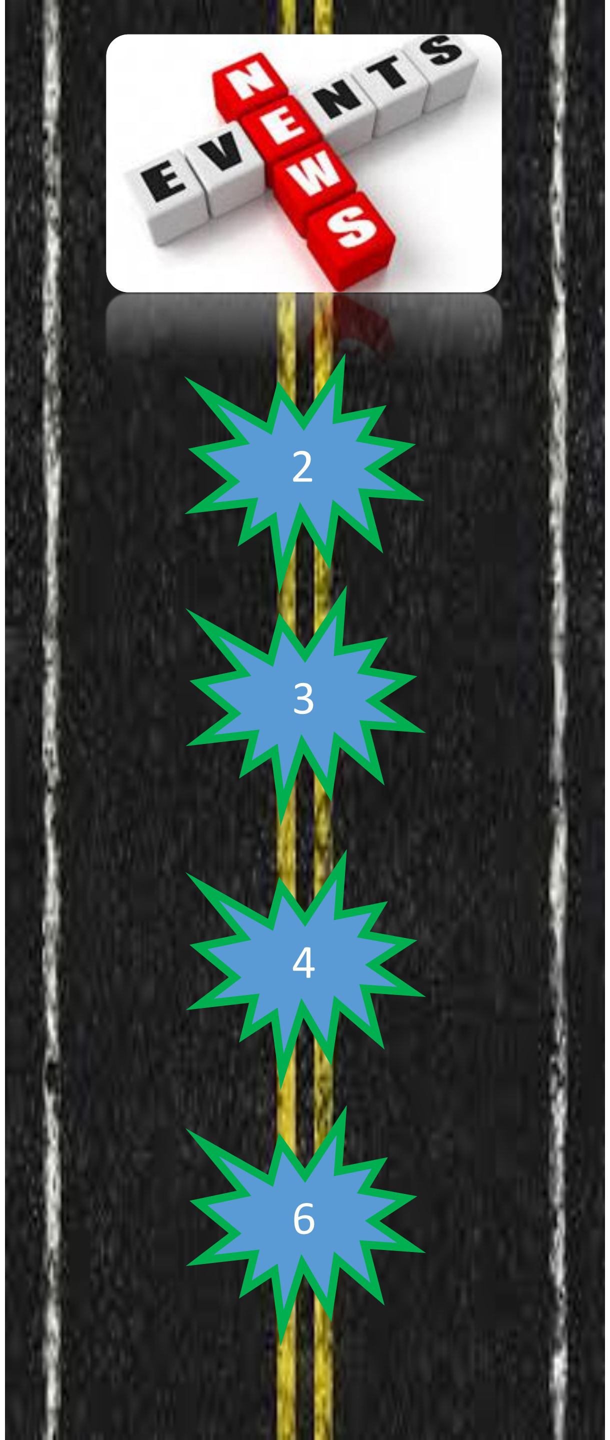
News Observatory System for Evolving World Events

News Headlines for World Events



The Roadmap

1. Introduction
2. Problem Statement
3. Related Work
4. Literature Gaps
5. Keyphrase Quality Checkpoints
6. News-KEA
7. Experimental Work
8. Weakness of our Approach
9. Conclusion & References

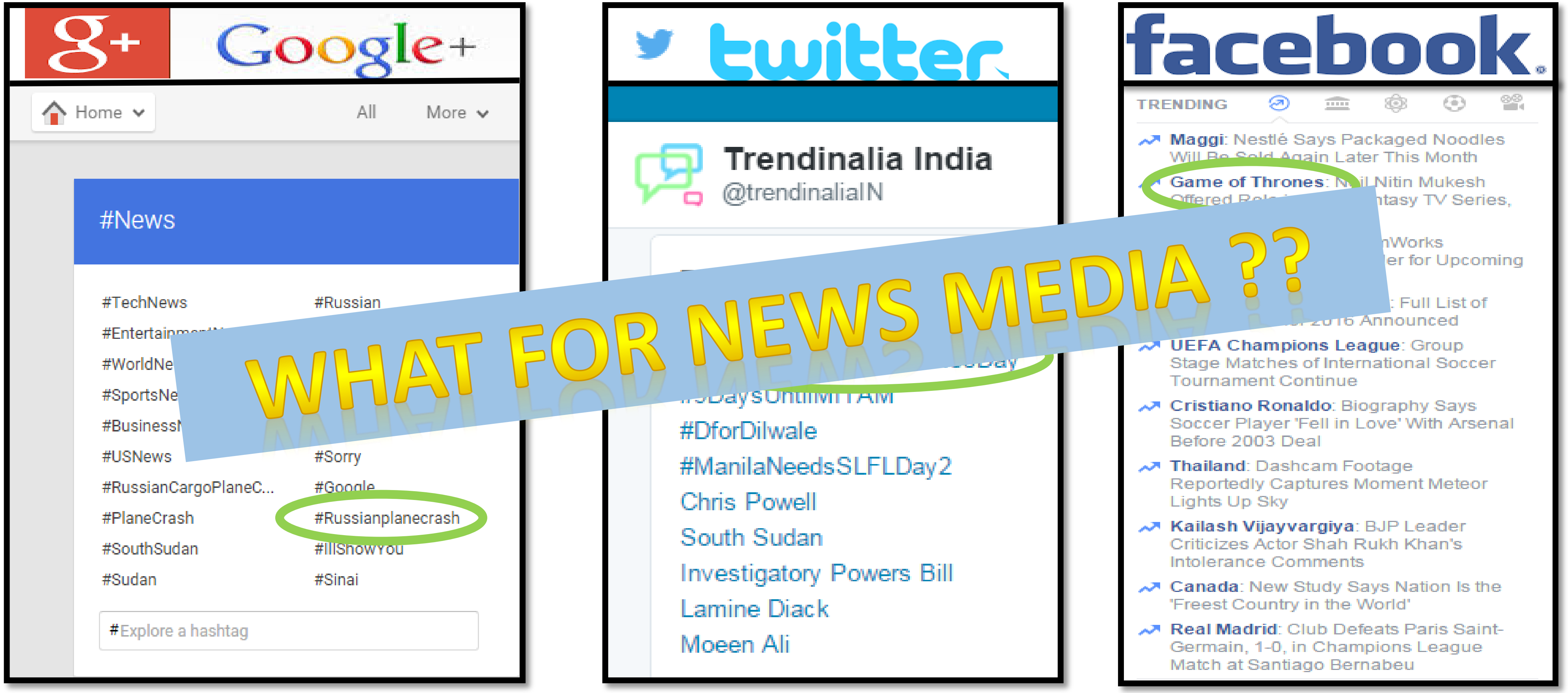


Automatic Keyphrase Extraction for News Media

- *Group of **WORDS**, i.e., Short in length*
- *Expresses an Important **CONCEPT***
- **NO** structure or grammatical **RULES**

Keyphrases in other social media

Keyphrases used by various social media to summarize their **trending content**



An Example of Keyphrases in News

Maruti Suzuki Baleno Receives More Than 21,000 Bookings

By Kritika Sethi | Nov 16, 2015

Maruti Suzuki Baleno, that was launched in India last month, has received over 21,000 bookings. With these numbers, the carmaker expects to garner better numbers this year as compared to that amassed during last year's festive season. In fact, RS Kalsi, Executive Director, Sales and Marketing, MSI, said, "We have witnessed over 56,000 footfalls for Baleno in the last 10 days and have already received 21,000 bookings for the premium hatchback."

While the festive season is now deemed over, Maruti Suzuki hoped to retail around 14,000-15,000 units on Dhanteras alone. Other than the festive spirit and the fact that it is considered an auspicious time to buy new products, Mr Kalsi believes low interest rates and cheaper fuel prices also influenced the market sentiment.

Also Read: **Maruti Suzuki Baleno Prices, Specs, Features, and Photos**

Talking about the company's plans for the Maruti Suzuki Baleno Mr Kalsi had previously said, "We hope to achieve leadership market share in the premium hatchback segment. We have invested about ₹1,060 crore in this project (manufacturing of Baleno). We will be making exports to 100 overseas locations including Europe, Japan, South America in 2016."

Maruti Suzuki Baleno Mileage:

a. 1.2-litre VVT Petrol: 21.4Km/l

b. 1.3-litre DDiS Diesel: 27.39Km/l

(With inputs from PTI)

Related Articles

- 5 Hatchbacks Under ₹5 Lakh That You Can Buy This Diwali
- Maruti Suzuki Plans to Provide Safety Features as Standard on Its Model Range
- Maruti Suzuki Baleno vs Hyundai i20: Battle of the Premium Hatchbacks
- Ask SVP: City vs Verna vs Baleno, Figo Aspire or Nissan Sunny
- Maruti Suzuki Ciaz Photo Gallery

Tags

Maruti Suzuki Baleno | Maruti Suzuki Baleno bookings | Maruti Baleno | Maruti Suzuki

Maruti Suzuki Baleno

₹5.41 lakhs On Road Price (New Delhi)

8.3 NDTV RATING

BOOK TEST DRIVE

Variants

- Sigma Petrol
- Delta Petrol
- Sigma Diesel
- Zeta Petrol
- Delta CVT Petrol
- Delta Diesel
- Alpha Petrol
- Zeta Diesel
- Alpha Diesel

Keyphrase uses

- Better summarisation
- Better indexing
- Better browsing
- Better metadata construction
- Classification and clustering
- Topic/content identification
- Back of book/document

Some more Applications

Keyphrases extracted from News corpus can be used News channels,

- Make it easy to skim news articles by Highlighting keyphrases.
- Search temporal events by using keyphrases as index terms.
- Refine news queries on search engine by using keyphrases as suggestions.
- Find similar news content by using keyphrases as a similarity metric.

Problem Statement

For a given concept*, mine *Interesting Keyphrases* from *News media information*, where a keyphrase is considered interesting if when it is shown to N persons, and more than N/2 persons find it interesting on the basis of *frequency*, *collocation* and *completeness* of the topic covered by the output set of keyphrases.

** Concept can be any name, place, event, organization, etc.*

Related Work

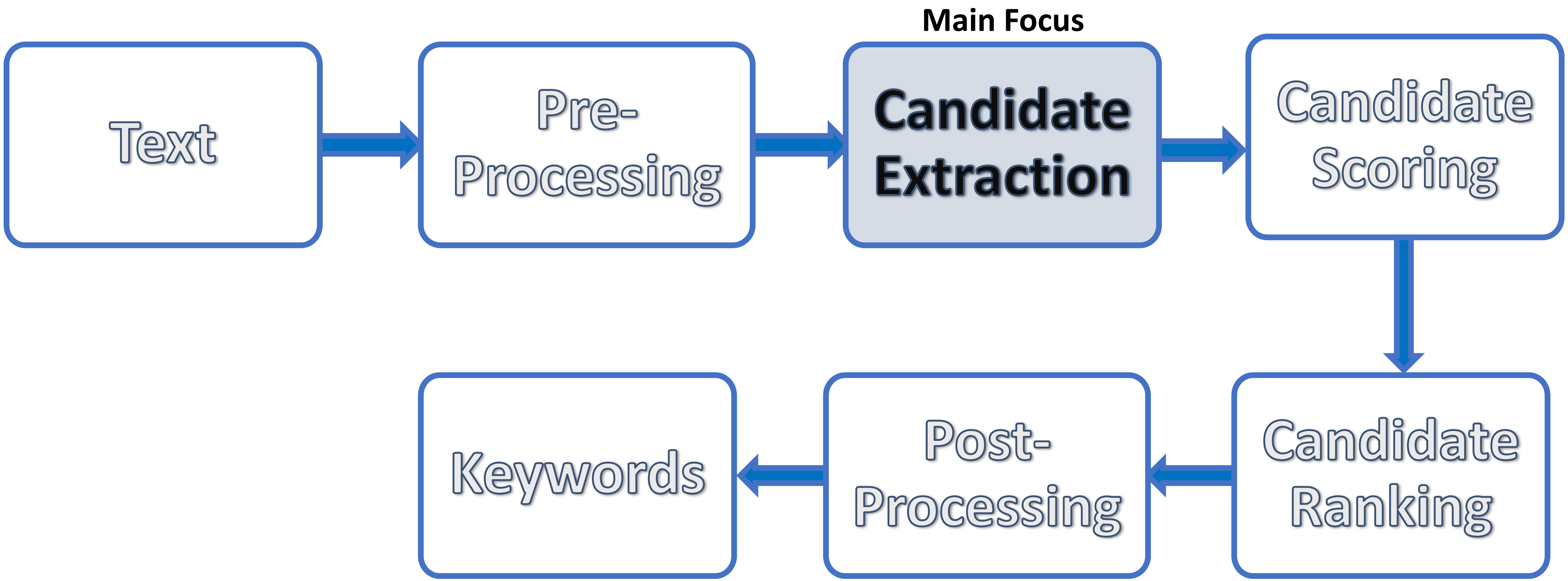
Existing work focuses more on extracting keyphrases from standard **passage text**

- **Keyphrase Extraction (KEA)** (I. H. Witten, 2005)
  - Parameters laid algorithm, Supervised learning task
  - None of the keyphrases describe events properly
  - For special topic, re-work is required
- **Microsoft Web N-Gram Service** (K. Wang, 2010)
  - Gives maximum likelihood of a phrase present
  - Online Algorithm, Limited number of token request
  - Computationally Expensive, as many combinations are tested
  - Large response time
- **ToPMine** (Ahmed El-Kishky, 2014)
  - Topic Modelling approach, converts Bag of (Words → Phrases)
  - Parameters laid algorithm
  - Subset is also returned as a keyphrase
  - Stopwords are present in keyphrase

None of them has leveraged News Media information characteristics

General flow work for Keyphrase Extraction

(Based on Related Work)



Literature Gaps

- *Language Dependent*: No extensive work for Asian languages.
    - Even if we translate the previous work, results will not be good, as reordering of phrases in translation leads to dis-fluencies.
- *Word segmentation*: Identifying the boundaries of word in continuous text, is a fundamental problem of NLP. For News media data its further challenging.
- *Time Complexity:* Baseline systems like Microsoft Web N-gram have very high time complexity and are not good for streaming data like News media.
- *Scalability:* Systems having online services have limited no of token request. Systems like KEA, large size of dataset is required.
- *Pre-trained Model:* Baseline systems depends on a pre-trained datasets, hence do not generalize well on new evolving domains.

Keyphrase Quality Checkpoints

Qualities to validated the candidate keyphrase generated by the algorithms:

1. Frequency: Gives the list of most probable n-grams one will encounter for the concept.

   PMI | Mean and Standard Deviation

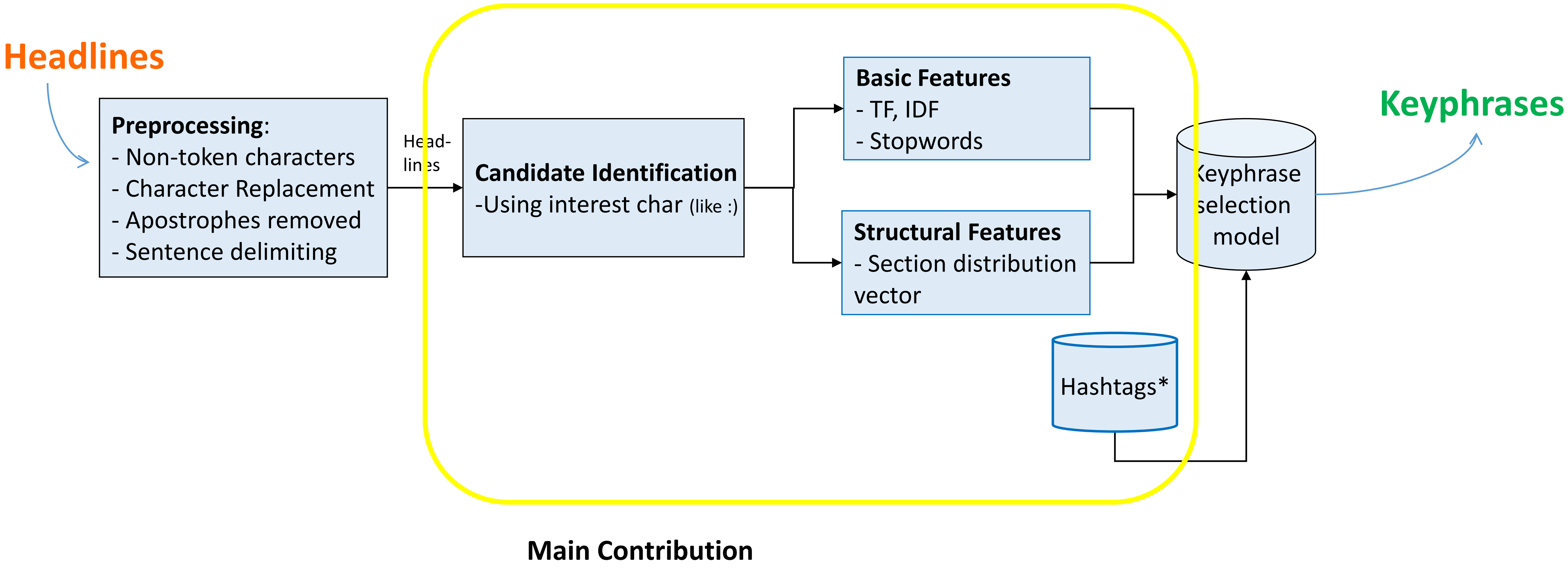
2. Collocation: As juxtaposition of words deviates from what is expected, hence it shows 'interestingness' and 'informative'.

   Term Weights | TF | IDF | X_i measure | Term Length | TF-IDF

3. Completeness: If long frequent phrase satisfy the above criteria, then their subsets also will satisfy the criteria. Phrase-construction algorithm should determine appropriate size.

   Human Interference

News-KEA : System Architecture



* To be done

News-KEA : Preprocessing

Input: English News Headline
Output: Preprocessed English News Headline

String Regular Expression	Replaced String
“(_)+”	“” (empty string)
“(.)+”	“” (empty string)
“@”	“at”
“(!)+”	“” (empty string)
“\’””	“” (empty string)
“(?)+t”	“ not”
“(?)+”	“” (empty string)
“(?)+s”	“ \‘s”
“(?)+r”	“ are”
“(?)+m”	“ am”
“(?)+ve”	“ have”
token starting with “http”	token deleted

Character Replacement

Stemming

Sentence Delimiting

News-KEA : Preprocessing (Example)

Example 1

CADENCE OF HONOR: Floyd Central, Lanesville, New Albany cadets march for veterans

↓ Removal of {Lowercase, Comma}

cadence of honor: floyd central lanesville new albany cadets march for veterans

Example 2

Syrian refugee crisis: It's about compassion and security

↓ Removal of {Lowercase, apostrophe}

syrian refugee crisis: it about compassion and security

Language Independent Step

News-KEA : Candidate Identification

Input: Preprocessed English News Headline

Output: Phrases

Process: Not every character is of interest. We identified a list of interesting characters.

Interested Char: { **:** , ‘ ‘ , “ “ , - }*

Section Distribution Vector:

- Using *Interested Char* we divide the headline into different chunks (vector)
- Smallest length vector is passed to next stage

* To be done

Language Independent Step

News-KEA : Candidate Identification (Example)

- Interest Char (':')

  cadence of honor: floyd central lanesville new albany cadets march for veterans

Keyphrase

- Interest Char ('…')

  zoeller launches 'freeze identity thieves' initiative

News-KEA : Keyphrase Feature

Input: Phrases

Output: Keyphrases

Process: Rank the input set using filters and add Hashtags to the set.

Filters Used:

- Term Frequency (TF)
- Inverse Document Frequency (IDF)
- Stopwords

Hashtags*:

Word having prefix (#), is considered as hashtag

* To be done

News-KEA : Keyphrase Selection Model

Input: Keyphrase Corpus, User request concept

Output: Keyphrases for *user request concept*

Process:

- Select unique keyphrases having similarity with *user request*
- Use External Knowledge base to find similar candidate keyphrases*

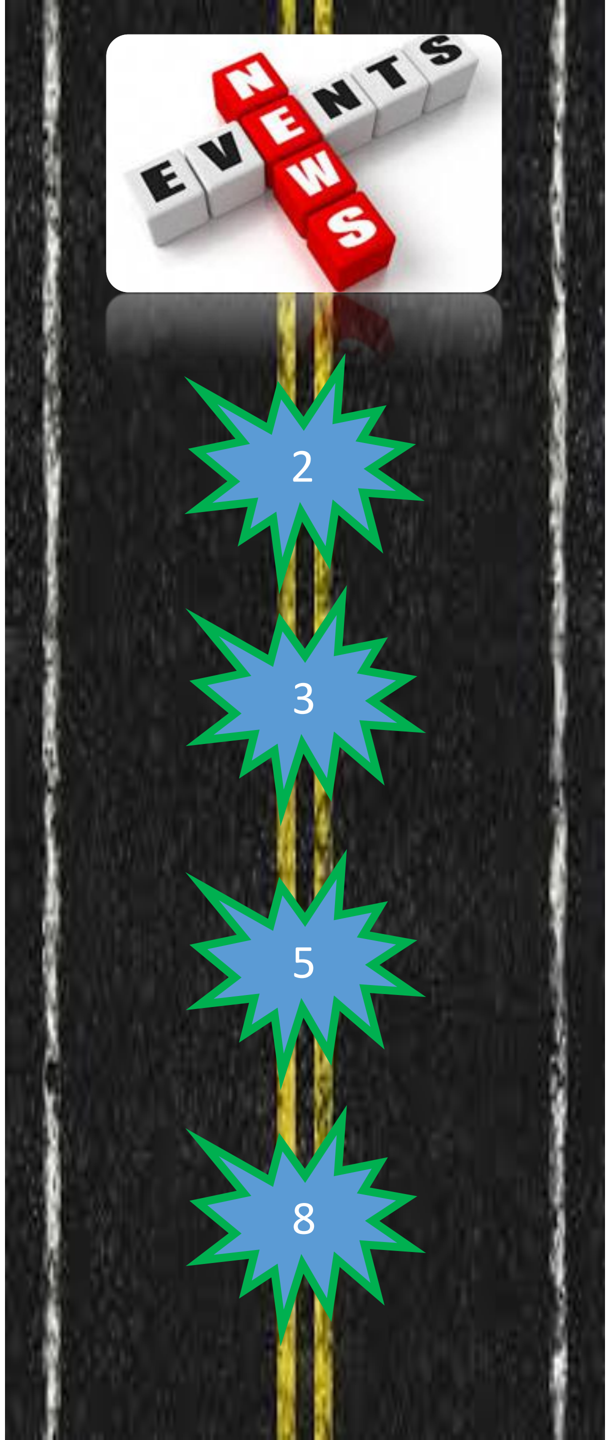
  Eg. {“PM Modi”, “Narendra Modi”, “Modi” }
- Remove less informative paraphrases*

  Eg. {“Sheena Bora Muder Case”, “Sheena Bora Case”}

* To be done

The Roadmap

1. Introduction
2. Problem Statement
3. Related Work
4. Literature Gaps
5. Keyphrase Quality Checkpoints
6. News-KEA
7. **Experimental Work**
8. Weakness of our Approach
9. Conclusion & References



Experimental work (Methods)

To evaluate our algorithm’s keyphrases with other baseline models, we perform comparative and qualitative test using following models:

- Baseline Model:
  - KEA-I
  - KEA-II
  - ToPMine
  - Microsoft Web N-Gram
- Proposed Model
  - News-KEA

Experimental work (Data)

Data Type:

- Training
  - Passage Test (25 Journal Articles, 25 Keys, English)
  - News Headlines (1.1 million, English, 3680 Keys)
  - Bing Search Query (Online, English)
- Testing
  - News Headlines (2.2 million, English) and (3 million, English)

Experimental work (Model Training)

We trained 5 models for evaluating the keyphrases

- Baseline Model:

KEA-I	Passage Text
KEA-II	News Headline
ToPMine	No training dataset
Microsoft Web N-Gram	Bing Search Query

- Proposed Model
  - News-KEA

News-KEA	No training dataset

Experimental work (Model Testing)

We tested 5 models for evaluating the keyphrases

- Baseline Model:

KEA-I	News Headline (3 million)
KEA-II	News Headline (2.2 million)
ToPMine	News Headline (3 million)
Microsoft Web N-Gram	News Headline (3 million)

- Proposed Model
  - News-KEA

News-KEA	News Headline (3 million)

Experimental work (Comparative Evaluation)

Following are the keyphrases returned by respective models for the Concept: *“Scam”.*

M. Web N-Gram [222 Hits]
Saradha scam
Coal scam
scammers
2g scam
A big scam
In adarsh scam
Scam people

ToPMine [48 Hits]
Saradha scam
Coal scam
Saradha scam says CBI
scam case
Coal scam case
scam accused
Coal scam case says

News-KEA [77 Hits]
saradha scam
coal scam
cash-for-vote scam
fodder scam
peb scam
chitfund scam
Ipl scam

KEA-I [0 Hits]

KEA-II [0 Hits]

Testing data contained 2252 news headlines out of 1.1 million* headlines related to scams.

* Preprocessed Headlines

Experimental work (Comparative Evaluation)

Scoring of Models

Model Name	Frequency	Colocation	Completeness	Score (Out of 3)
KEA-I	↓	↑	↓	**1**
KEA-II	↓ +	↑ –	↓ –	**1.25**
Microsoft Web N-Gram	↑	↑	↓	**2**
ToPMine	→	↑	↓	**1.5**
News-KEA	↑	→ *	↑	**2.75**

* As News Headlines have high percentage of Juxtaposition, (892728 headlines out of 1000000 = **89.27%**)

Experiment work (Quality Evaluation)

Conducted User Study

- To assets the keyphrases a Google form was designed.
- No of questions: 10 {Topics} X 3 {Diversity, Meaningfulness, Best Result}.
- To ensure the un-biasedness of results, we randomly renamed each algorithm result.
- We emailed students of Computer Science Department, IIT Roorkee to invite them to participate in the experiment.
- Collected overall 120 responses, selected top 112 for inter-rater reliability test.

10 Topics: Events {FIFA, 26/11 attack, Earthquake} , Entities {Google, Greece, Obama, Salman Khan}, Concept {Scam, Budget}.

Experiment work (Quality Evaluation)

Key Phrase Algorithms Survey

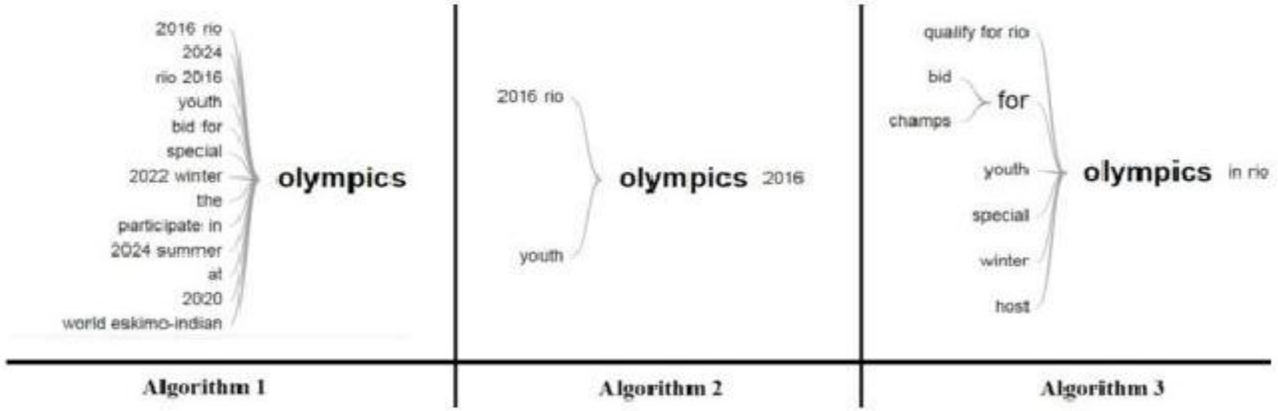
- PLEASE ZOOM IN (Ctrl+)THE WEBPAGE IF YOU ARE NOT ABLE TO READ THE KEYPHRASES CLEARLY.

- Larger word size in the tree denotes higher frequency.

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

* Required

Olympics

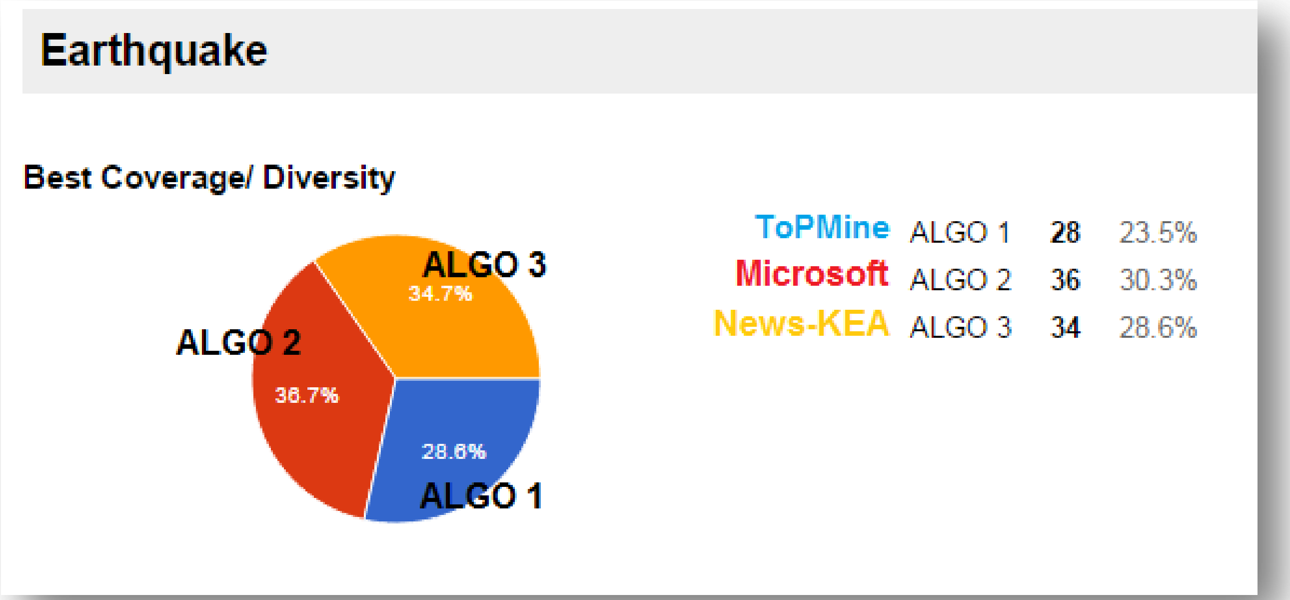


Experiment work (Quality Evaluation)

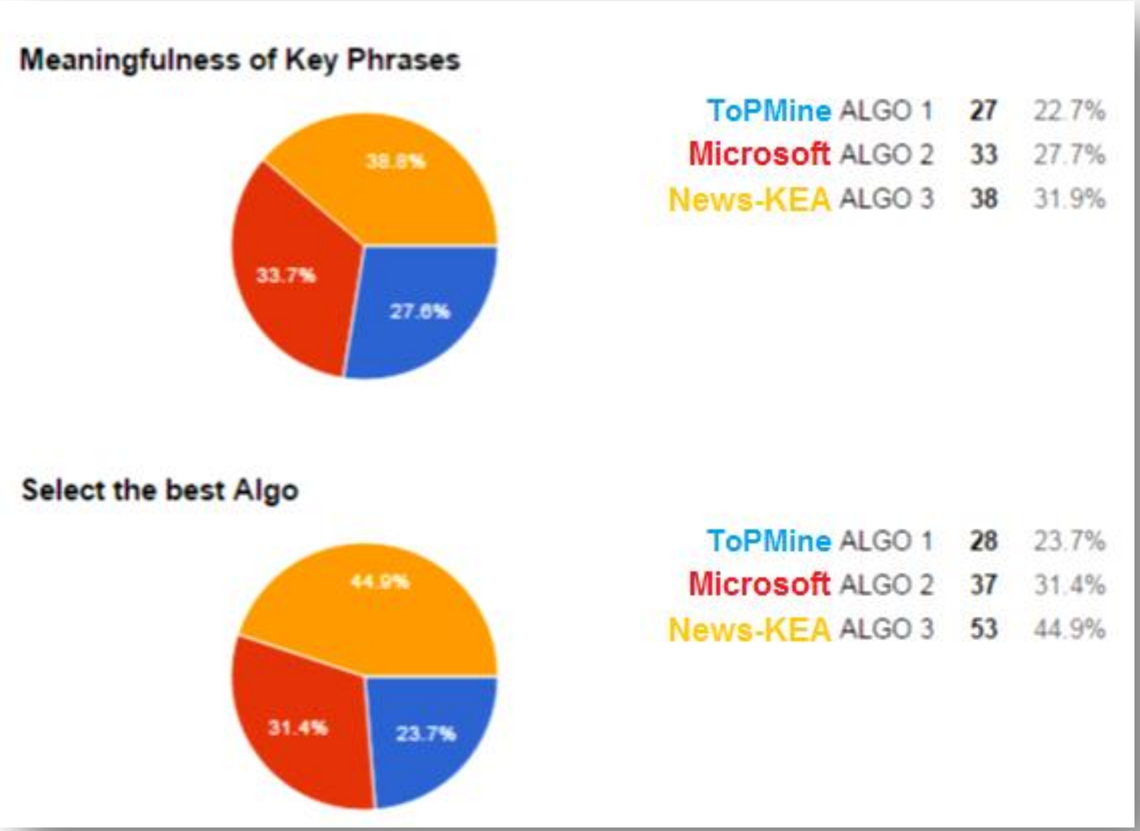
Experiment work (Quality Evaluation)

Result:

For keyphrases on the event "Earthquake", given by News-KEA (Algorithm 3), TopMine (Algorithm 1), Microsoft web N-gram (Algorithm 2).



Experiment work (Quality Evaluation)



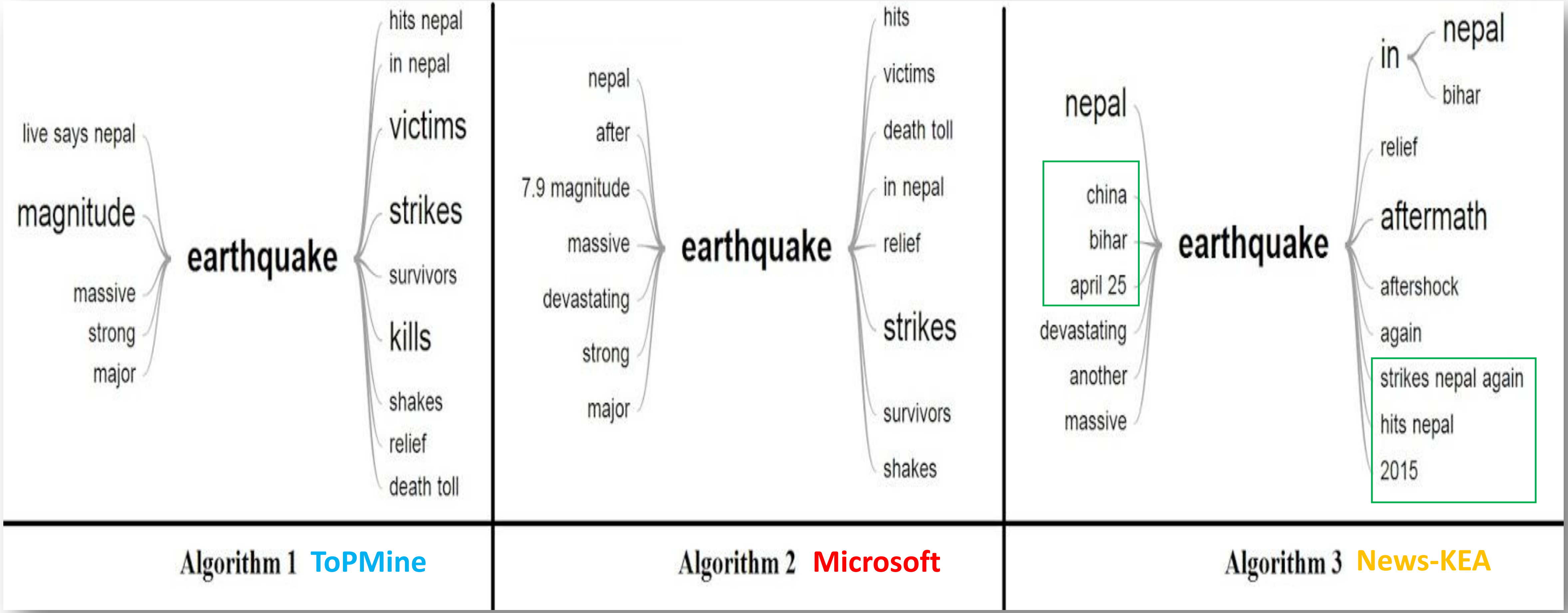
Overall result-

Best Coverage/Diversity: **News-KEA**

Meaningfulness: ToPMine

Best Algorithm: **News-KEA**

Keyphrase Algorithms Survey on *‘Earthquake’*



Inter-Rater Reliability Test

- Measured the degree to which different judges or raters agree in their assessment decisions.
- As human observers will not necessarily interpret answers the same way

Statistical measure of inter-rater reliability applied: Fleiss Kappa

We obtained **slight agreement value** for correlation coefficient, which indicates the stability of the scores.

Weakness of Our Approach

- **Paraphrases** are there.
- **Number** of phrases extracted are less.
- To annotate phrases in News articles, snippets, or other than headlines, our approach depends on **external training** models like KEA.

Conclusions

Contributions: better keyphrase extraction

- Proposed a system automatic keyphrase extraction for News Media
- Compared three algorithms for News related keyphrase extraction
- Shown a way to leveraged publicly available News information
- Created a corpus for keyphrases testing

Current and Future Work

- Enlargement of the keyphrase corpus
- Integrate tags with keyphrases, clustering
- Personalization with respect to user
- To make the algorithm completely language independent

Key References

- Ahmed El-Kishky, Y. S., Scalable Topical Phrase Mining from Text Corpora, (*VLDB Endowment*, 2014).
- K. Wang, C. Thrasher, E. Viegas, Xiaolong, Bo-june (Paul) Hsu., An Overview of Microsoft Web N-gram Corpus and Applications, (NAAACL HLT 2010).
- Gao, J. N., A Comparative Study of Bing Web N-gram Language Models for Web Search and Natural Language Processing, (*ACM SIGIR,* 2010).
- I. H. Witten, G. W.-M., Kea: Practical automatic keyphrase extraction, (*Fourth ACM conference on DL,* 2005).
- Turney, P., Learning to Extract Keyphrases from Text, (*Information Retrieval,* 2000).

Thank You!

End of Presentation