

No. DeitY/IPR/1.1/48/2016  
Government of India

Ministry of Electronics and Information Technology  
IPR Division, R&D in Electronics Group


516  
17/MHRD/PM/  
Date: 21.03.2017  
IMPRINT

OFFICE MEMORANDUM

Sub: In-principle concurrence for the proposals received under IMPRINT Program -Reg.

This has reference to your communication vide letter No. 3-18/2015-T.S.I(Vol.III) dated 9<sup>th</sup> January, 2017 requesting this ministry for providing concurrence with regards to the project proposals for 50% funding under IMPRINT India initiative.

2. The matter has been considered and with the approval of Competent Authority, it is conveyed that this ministry approves 50% support to seven shortlisted projects as per details annexed.

  
Dr. A. K. Garg  
Director, IPR Division  
Tel. 011-24364799  
Email: [ajaik@meity.gov.in](mailto:ajaik@meity.gov.in)

Encl: As above

To

Kunden Nath  
Under secretary (IITs)  
430-C, Ministry of HRD  
Shastri Bhawan, New Delhi- 110001

Copy To:

Tripti Gurha,  
Director (IITs),  
427-C, Ministry of HRD  
Shastri Bhawan, New Delhi- 110001







**Title**

A platform for Crosslingual and Multilingual Event Monitoring in Indian Languages

**Name of Principal Investigator**

Sudeshna Sarkar

**Institute**

IIT Kharagpur

**Email**

sudeshna@cse.iitkgp.ernet.in

**Investigator for Correspondence****Co-investigator(s)** (max. 9)

Pawan Goyal (Kgp), Pushpak Bhattacharyya (IITB/P), Asif Ekbal (IITP), Sobha (AuKBC), Malhar Kulkarni (IITB), Ganesh Ramakrishnan (IITB), Prasenjit Majumdar (DAIICT)

**Title** (max. 25 words)

An Open Platform for Multilingual Event Monitoring in Indian Languages

**Domains & Themes** (please select as many as themes as are relevant from any one domain, you can choose multiple themes from a single domain)

**Information & Communication Technology**

Multimodal, Multilingual and Cross-lingual Interfaces

2. Abstract (text only field, max. 150 words)

The project aims to build a cross-lingual, multi-lingual, multi-source and multi-domain platform for information extraction (IE) in Indian languages (ILs) to extract and classify events of interest from news media and to demonstrate the application in a few domains. A pluggable and customizable platform will be developed, enabled for English, Hindi, Bengali, Marathi and Tamil to monitor news events around a few themes such as disaster, conflict events and health. The output will be presented in multiple languages as well as a suitable visualization interface supported by analytics. The system will be extendible to monitor other sources like social media.

3. Objectives (text only field, max. 100 words)



1. To build a multi-lingual event monitoring platform to identify, extract and classify events of interest from news media.
  2. An end-to-end system will be built and demonstrated for English and a few Indian languages such as Hindi, Bengali, Marathi, Tamil and domains such as health, conflict events and disaster.
  3. The necessary NLP tools will be developed for the corresponding Indian languages.
  4. The designed system will be modular and customizable to facilitate addition and adaptation of languages, domains and other sources such as social media.
  5. A user-friendly visualization interface will also be developed, which will support retrieval and analytics driven by user-query.
4. Budget (year-wise) max. total budget Rs. 4 Crores; the heads are: Equipment (less than 30% of the total excluding overhead)  
Manpower, Consumables, Contingency, Travel & Others and Institute Overhead (20% of the total or as per sponsoring Ministry's norms)

#### Year-wise Budget Breakup

	Year 1	Year 2	Year 3	Total
	Rs Lakhs	Rs Lakhs	Rs Lakhs	Rs Lakhs
Equipment	26	7	0	33
Manpower	78	78	87.36	243.36
Travel	9.0	9.0	9.0	27
Consumables	2.35	2.35	2.30	7.00
Contingency	5.35	5.35	6.30	17.00
Project Cost	120.7	101.7	104.96	327.36
Overhead @20%	24.14	20.34	20.992	65.472
Total Incl Overhead	144.84	122.04	125.952	392.832

#### Institutewise Budget (Tentative)

		IITKgp	IITP	AuKBC	IITB	DAIICT	Total
		Rs Lakhs	Rs Lakhs	Rs Lakhs	Rs Lakhs	Rs Lakhs	Rs Lakhs
Manpower	25k to 28k pm	74.88	56.16	65.52	28.08	18.72	243.36
Equipment	Server	6					6
	PC/UPS/Laptop	8	6	7	3	3	27
Travel		8	6	7	3	3	27
Consumables		2.0	1.5	1.75	0.75	1	7.0



Contingency	2.0	1.5	1.75	0.75	1	7.0
Coordination Contingency	10					10
Project Cost	110.88	71.16	83.02	35.58	26.72	327.36
Overhead @20%	22.176	14.232	16.604	7.116	5.344	65.472
TOTAL incl Overhead	133.056	85.392	99.624	42.696	32.064	392.832

#### Tentative Manpower Breakup among Institutes

Manpower	IITKgp	IITP	AuKBC	IITB	DAIICT	
Software Platform	2	2				
IE Platform + English	2		3			
Evaluation					2	
Hindi		3				
Marathi				3		
Tamil			3			
Bengali	3					
Domains	1	1	1			
Total	8	6	7	3	2	26

Out of 3 manpower allocated for each language, 2 will explicitly work on creation of the required NLP tools and modules, while 1 will work on creating the syntactic patterns, thorough testing and evaluation for that language.

Regarding the domains, there is one manpower each at IITKGP (Conflict events), IITP (Health) and AUKBC (Disaster).

#### 5. Current status (max. 600 words or two A4 size pages)

A lot of prior work has focused on event extraction, monitoring, search and recommendation. Most of the event extraction modules are heuristic-based, and use hand-written rules on part-of-speech tags and dependency trees to identify the event patterns from the source sentence (Mausam et al., 2012; Alfonseca et al., 2013). Pighin et al. (2014) developed two data-driven methods for event pattern extraction. The first method uses a sentence compressor to get to the core of the sentence, while the second method relies on a vast collection of human-written headlines and sentences to extract event patterns. Krause et al. (2015) used neural-network based architecture to learn distributed representation of event patterns, which allows them to cluster the related event patterns.

Towards event search and recommendation, EventCube (Tao et al., 2013) is a general online analytical processing framework to effectively organize and search relevant events, and measure event similarity based on multiple dimensions. Recently, meta-paths over heterogeneous information networks have also been utilized. NewsNet (Tao et al., 2014) is a



news information network that extracts types, topical hierarchies and other semantic structures from news data to construct heterogeneous information networks, so that various algorithms based on meta-paths can be used for search and recommendation.

Another line of research focuses on generating event timeline and event chronicle generation, to connect the current event with the relevant past information. Ge et al. (2015) developed a module that given a reference event, generates topically relevant event chronicles from the past events. Yan et al. (2015) developed an evolutionary trans-temporal summarization system, that given the collection of time-stamped web documents, returns event evolution along the timeline.

There have been systems built for multilingual media monitoring (for European languages), a planned collaboration by BBC, and several news event monitoring efforts. Europe Media Monitor (EMM) is one such system, gathering lakhs of online news articles in tens of languages, categorizing news items as per various domains and extracting various information (Atkinson and Van der goot, 2009; Steinberger et al., 2013). Lu et al. (2016) developed a system to search, identify, summarize and organize complex events from cross-media. However, no such platform works directly with ILs, and there is no open platform for the entire system pipeline.

There have been several efforts in building basic resources and capabilities in some of the major Indian languages (ILs) which provide a solid foundation for taking the capabilities of IL processing to the next level. Natural Language Processing (NLP) research has progressed in recent years with new representations and methods being explored. The NLP modules in IL need to be customized or improved and used in building demonstrable and useful applications. For text mining applications, accurate and good quality NLP systems for dependency parsing, semantic role labelling, coreference resolution, relation extraction, event extraction and summarization systems need to be built. Named Entity Recognition (NER) systems for ILs have been developed and the PIs have deep expertise in working on the other tasks.

Thus the technology is in readiness to be taken forward. The proposed development will leverage the expertise of the PIs in building NLP and IR systems, linguistic resources created, and the ongoing research and experience in information extraction and summarization.

#### 6. Motivation and scope (max. 200 words)

A lot of useful information is locked in unstructured textual form, in news, social media and various correspondences. Aggregation and Text Mining of these media enables monitoring, tracking, information management and aid decision making. India is a multilingual country and much of the local information and people's views are reported in Indian language media which may be mined to find important events, identify sentiments and aspirations of the people, and for surveillance.



GDELT (Google) is a platform that monitors global events. EMM (European Media Monitor) monitors events in multiple European languages. BBC and partnering universities have initiated work on a multilingual media monitoring system. However, very little work has been done in Indian language towards event extraction, classification into representative categories and filling in the event roles. We wish to develop an extensible open platform enabled with several Indian language families and a few domains.

## **7. Methodology and research plan (max. 1500 words or five A4 size pages, including figures/charts/tables/images)**

The system will be built to monitor events from given websites, news and RSS feeds and aggregate the content on a regular basis. The full text will be indexed by an Information Retrieval (IR) system. The contents will be processed by an Information Extraction (IE) system which will extract reportable event mentions, along with the event class, and the event arguments. A unique event identification and summarization system will work to link the same event mention from different sources in the same language as well as multiple languages, thus identifying the co-referring events and aggregating event information. A query module will accept queries in multiple languages. The result can be shown in any of the languages enabled in the system and by a dashboard for visual presentation. There will also be a navigation interface for browsing the events based on location and time. The system will be designed so that it is extendible to accept inputs from social media, speech transcripts, etc. and the output can generate speech output, summary feeds, etc.

Next, we discuss various horizontals of the proposed system along with the proposed methodologies:

1. H1:Aggregator/Crawler: This will make use of a list of URLs and feeds and do regular crawling to get content. A CLIR platform will be used to index the full text of the content crawled and make them available for cross-lingual and multi-lingual search. This will make use of Indo-WordNet, bilingual dictionaries and the CLIR platform based on Nutch and Solr developed in the CLIA project. Approaches based on joint learning of cross-lingual word embeddings (Mogadata and Rettinger, 2016) will be utilized for efficiently aggregating the multilingual content.
2. H2:Syntactico-Semantic Processing: This horizontal will ensure that each of the languages considered in the project have the required tool for the Syntactico-Semantic processing of the the raw documents, crawled in H1. The main modules that will be developed and adapted for this particular task include:
  - a. Sentence Splitter
  - b. Morph analyzer
  - c. Part-of-Speech Tagger
  - d. Chunker
  - e. Dependency Parser
  - f. Named Entity Recognizer
  - g. Anaphora Resolution



#### h. Coreference Resolution

3. H3:IE platform with event extraction, open domain IE, event classification, aggregation and timeline summarization. This module will make use of the state-of-the-art methods in cross-lingual embeddings as well as heterogeneous information networks for event aggregation and navigation, to be further used for search and recommendation in H4. In the initial phase of the project, a heuristic-based approach (Alfonseca et al., 2013; Mousam et al., 2012) will be used to extract event patterns from the news and other documents. Syntactico-semantic information from the tools available from H2 will be used to develop the syntactic patterns for this approach. The patterns thus extracted will then be projected in the multi-lingual embedding space to enable alignment of similar events across languages, and provide a richer description based on information fusion from multiple sources. Further, the event patterns in multiple languages will be clustered and ranked to obtain the most important events globally. Projecting the event patterns into a common multi-lingual space will also enable linking events across time (for timeline summarization of events, generating event chronicles) as well as cross-lingual event access.

The crowd-sourced feedback obtained from H4 will help us develop a gold standard for the syntactic patterns identified by the system. Using this data, we plan to experiment using structured prediction algorithm (Li et al., 2014) as well Convolutional Neural Networks (Chen et al., 2015) to enhance the performance of the system.

Using the event embeddings, the event patterns will also be clustered and categorized into various identified domains. For categorization, we plan to build the ground truth in an automated manner, relying on the fact that some of the websites are known to provide events only in a specified domain. Further, the events will also be indexed using the event participants including the persons, locations, time etc. This will allow efficiently navigating through the events using the heterogeneous information networks in H4.

4. H4:Output Interface: A suitable interface will be developed for rich visualization enabled with a mapping interface and timeline. This user-friendly interface will provide the following functionalities to the end-users:
  - a. Event updates in various languages and domains: Once the user selects a language and domain, the interface will retrieve the relevant events from H3 and provide those to the user in a ranked order. This ranking will be determined dynamically using the overall popularity of the event, as well as the previous interactions of the user with the system.
  - b. Event navigation: With each event, a summary will be provided consisting of the important sentence extracts from multiple news articles mentioning this



event. An algorithm based on LexRank (Erkan and Radev, 2004) will be used to generate this summary. Additionally, the user will be shown various related events on demand – a timeline summarization for the event, as well as the events mentioning similar entities involved in the event. The event embeddings as well as information networks, developed in H3 will be utilized for this navigation. Additionally, the user may also demand seeing the corresponding event in other languages, which will also be enabled using the event patterns in the combined embedding space.

- c. Event search: The interface will also support searching the event in various languages. The search mechanism will make effective use of the event embeddings in the combined space. The returned results will also have the functionalities, as in H4(a) and H4(b).
  - d. Gathering user interaction data: The system will make use of the user interaction data such as search queries, selected events while search and recommendation to enhance the performance of the system.
5. H5:Evaluation: A thorough evaluation of each of the modules the proposed system is extremely important, and will be an essential component of the project. The event extraction tasks will be evaluated using standard set-based measures like precision and recall and other clustering metrics. An attempt will be made to make the performance of the system comparable to the state-of-the-art in other languages. While a regular evaluation will be carried out based on user interaction data gathered from the live system, we will also build up a competitive framework for research and development in Event Extraction for Indian languages in a cross-lingual setting. For this annotated data will need to be created for training and reference data for evaluation in a domain(s) in several languages. Shared task on event extraction will be set up in various Indian languages for the proposed domains. The competitions may be held in conjunction with the FIRE workshop.

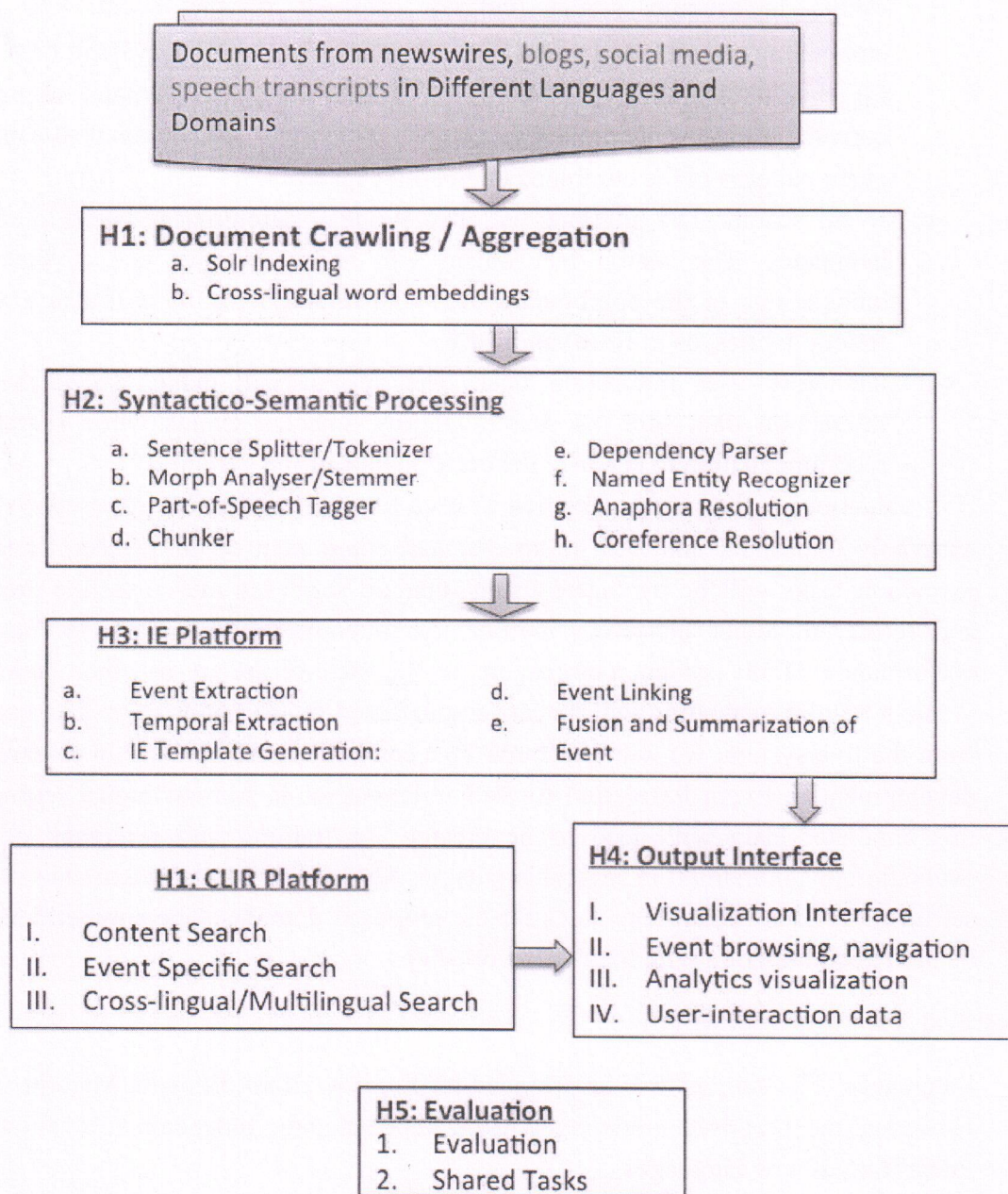
The verticals for the project will be

Languages: The support will be provided for English, Hindi, Bengali, Marathi, Tamil. However, the proposed modeling will be kept language independent to allow easy integration of new languages.

Domain: A few different domains such as Conflict, Disaster and Health events will be monitored initially. The system, however, will be generic enough to allow addition of various other domains, or being made open-domain.

We expect to be able to use and build up on tools and resources already build in Indian languages in the various TDIL projects and contribute the new tools and resources to open source by enabling the use of Indian languages in some standard open source software systems. The proposed system will be hosted and adapted dynamically based on feedback gathered by the users.





**System Diagram: A platform for Cross-lingual and Multilingual Event Monitoring in Indian Languages**



8. Justification and novelty (max. 200 words)

IE plays an important role in developing systems for monitoring, tracking and enabling access to information. At present, no platform works directly with ILs, and there is no open platform for the entire system pipeline. The proposed system addresses this challenge. Setting up an open extendible platform enabled with several IE systems and the creation of annotated resources for training and validation will drive research in this area and enable greater access to Indian language content. We plan to utilize novel techniques based on multi-lingual and cross-lingual embeddings as well as metapaths based navigation over heterogeneous information networks towards the main aim of event extraction, navigation and summarization.

9. Benchmarks milestones and time frames (max. 300 words or one A4 size page, including table/chart)

The time-frames and milestones for the project are mentioned below. A gantt chart has also been shown.

	M1-6	M7-12	M13-18	M19-24	M25-30	M31-36
H1						
H2						
H3						
H4						
H5						
Overall system	System Architecture		Version 1 demo	FIRE Task1	Version 2 demo	Final tuning, demo, FIRE Task 2
Verticals: domains		Conflict Events	Conflict Events	Health	Health, Disaster	All three

**Month 1-6:** Design of the system architecture, specifications of software modules and interfaces. The existing tools in various language verticals will be tested for the



proposed specifications. Work will be immediately started to build up the new tools and adapting the existing tools.

**Month 7-12:** Setting up aggregator, crawler and IR system. Creation of ontologies and annotated resources for the domain of Conflict events. The tools in various language verticals will be tested to be used by the system. Specifications of the event analytics will be prepared.

**Month 13-18:** The event extraction module will be developed for the domain of 'Conflict events' based on the heuristics rules. Visualization Interface with cross-lingual mapping and timeline will also be developed. Specification for the domain of 'Health' will be prepared. Version 1 demo in all languages in Domain1.

**Month 19-24:** Further work on search and summarization for the event navigation part. Feedback gathered from H4 will be used to adapt the system. Version 2 demo in two domains. FIRE shared task on 'Conflict Events' for Hindi, Bengali.

**Month 25-30:** Event categorization, Event Linking. Summarization. Interface to support cross-lingual queries. Thorough evaluation for all the modules will be started. Tools developed under H2 will be adapted based on the testing. Specifications for 'Disaster' domain will be prepared and it will also be supported during this time period.

**Month 31-36:** Last 6 months will be dedicated to tuning and testing of system, so that a robust system is available by the end of 3 years. Final demo will be prepared. FIRE shared task on the 'Health' domain for Marathi, Tamil.

10. Nature and evidence of Industry participation (max. 100 words); *you should also attach a letter from the industry evidencing the support*

11. Equipment, infrastructure and facilities available at the host organization(s) (max. 200 words)

All the Institutes have well developed infrastructure facility such as intra-network, internet, and printers. These would be available for carrying out the project. Initial experiments could be carried out on the high performance computing facilities available at IIT Kharagpur and IIT Patna. However, dedicated servers will be required for the specific applications.



12. Evidence of technology developed (300 words or one A4 size page, evidences such as paper, patent, letter, certificate, photograph, may be included)

Cross Lingual Information Access Project (<http://www.sandhansearch.in>) which caters to 9 Indian languages' search in mono and crosslingual settings (to English and Hindi)-Consortium Lead by Prof. Pushpak Bhattacharyya and participation by IITB, IITKGP, AU-KBC and DAIICT among other Institutes. In this project, the Indian languages included Hindi, Bengali, Marathi and Tamil among others.

Indian Language Wordnet (<http://www.cfilt.iitb.ac.in/indowordnet>) stores concepts and their linkages in unambiguous form with many very useful APIs. Wordnets of 15 Indian languages (scheduled) have been built by IITB, all linked and ready to use in large scale NLP Applications involving Indian languages.

Named Entity Recognition and Classification (NERC): PI(s) have experience in creating resources and tools for NERC in Indian languages. IIT Kharagpur has worked on Hindi and Bengali NER systems, IITB on Hindi and Marathi, IITP PIs on Hindi, Marathi and Bengali and AU-KBC on English and Tamil.

Stemmer: PI(s) have prior experience in developing both linguistic as well as statistical stemmers. YASS (Majumder et al., 2007) which is a statistical stemmer, has been successfully used for several Indian languages for which a linguistic stemmer is not yet available. Stemmers in Hindi, Bengali, Marathi and Tamil were developed for CLIA by the PIs of the current project.

Parser: IIT Kharagpur has developed a high accuracy Hindi dependency parser, and is currently engaged in developing a good quality Bengali dependency parser by using limited Bengali treebank and cross-lingual transfer for Hindi. AU-KBC has experience in developing Tamil dependency parser while IITB had developed a shallow parser for Marathi.

Anaphora and Co-reference Resolution: AU-KBC has worked on various types of anaphora and co-reference resolution in English and Indian languages. IIT Kharagpur and IIT Patna have worked on Bengali anaphora resolution.

AU-KBC has a sanctioned Project from Tamil Nadu Government (Tamil Virtual Academy) on Event-Entity Profiling from Tamil Documents- A Mobile Application which works on domains such as Weather and Sports. The goal of this project is to develop an Event-Entity Profiling application which provides collated information about an event or an entity from



Tamil text documents from various sources such as online Tamil Newswires, web blogs and microblogs such as facebook on smartphone platform.

Information Extraction: The PIs have experience in working on Information Extraction. In particular, the PIs at IIT Kharagpur and AU-KBC have worked extensively on Information Extraction in various domains.

Relevant papers by the PIs are added as attachment.

### 13. Potential users (max. 100 words)

It is expected that in addition to the native speakers, the system may be used by different administrative offices and agencies to track events of interest. Possible end users may be disaster management agencies, health department and home ministry. The system will be customizable to specific requirements.

### 14. Final outcome and deliverables (max. 200 words)

The final outcome of the project will be an open platform enabling Indian language information aggregation, extraction and visualization using open source platform which may be made available for further customizations. A user-friendly interface will enable the users to browse and navigate through the events of their interest. Along with this main outcome, the deliverables will be a). The required NLP tools in Hindi, Bengali, Marathi, Tamil. b). Shared tasks related to Indian language Event Extraction will be run in FIRE 2017 and 2018 by making available resources and encouraging research in IL text mining. c). Algorithms and modules for search, recommendation and (timeline) summarization will be made available for the individual languages, along with the cross-lingual retrieval modules. The platform built will be fully open and designed so that new modules can be plugged in easily and it can be extended to different languages, input media and domains.

### 15. Potential reviewers (up to 5) suggested by PI (please include name, designation, affiliation, verified mobile number and verified email id)

1. Dr. Hema Murthy

Professor

Computer Science Department

IIT Madras

Email: [hema@iitm.ac.in](mailto:hema@iitm.ac.in)



2. Dr. Krithi Ramamritham

Professor

Department of Computer Science

IIT Bombay

Email: [krithi@iitb.ac.in](mailto:krithi@iitb.ac.in)

Phone: +91 22 25767740

3. Dr Mausam

Associate Professor

Department of Computer Science and Engineering

Room 402, School of IT Building

Indian Institute of Technology Delhi

Hauz Khas, New Delhi, 110016, India

Email : [mausam@cse.iitd.ac.in](mailto:mausam@cse.iitd.ac.in), [mausam@cs.washington.edu](mailto:mausam@cs.washington.edu)

Phone : +91-11-2659-6076 (O), +1-206-979-7038 ©

4. Dr. Jayant R. Haritsa

Professor

Department of Computer Science and Automation

IISc Bangalore

Email: [haritsa@dsl.serc.iisc.ernet.in](mailto:haritsa@dsl.serc.iisc.ernet.in)

Phone: +91 80 2293 2793

5. Dr. Mandar Mitra

Associate Professor

Computer Vision and Pattern Recognition Unit

Indian Statistical Institute

203 B.T. Road

Kolkata 700 108.



Email: mandar@isical.ac.in

Ph: +91 33 2575 2858.

16. References (used in proposal, max. 300 words or one A4 size page)

[Alfonseca et al., 2013] Alfonseca, E., Pighin, D., & Garrido, G. (2013). HEADY: News headline abstraction through event pattern clustering. *ACL 2013*, pp. 1243–1253, Sofia, Bulgaria.

[Atkinson and Van der Goot, 2009] Atkinson, M., & Van der Goot, E. (2009, April). Near real time information mining in multilingual news. In *Proceedings of the 18th international conference on World wide web* (pp. 1153-1154). ACM.

[Chen et al., 2015] Chen, Y., Xu, L., Liu, K., Zeng, D., & Zhao, J. (2015). Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Vol. 1, pp. 167-176).

[Erkan and Radev, 2004] Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22, 457-479.

[Ge et al., 2015] Ge, T., Pei, W., Ji, H., Li, S., Chang, B., & Sui, Z. (2015). Bring you to the past: Automatic generation of topically relevant event chronicles. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

[Krause et al., 2015] Krause, S., Alfonseca, E., Filippova, K., & Pighin, D. (2015). Idest: Learning a distributed representation for event patterns. In *Proceedings of NAACL 2015*, pp. 1140-1149, Denver, USA.

[Leetaru and Schrod, 2013] Leetaru, K., & Schrod, P. A. (2013, April). Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention* (Vol. 2, No. 4).

[Li et al., 2014] Li, Qi, Heng Ji, Yu Hong, and Sujian Li. "Constructing Information Networks Using One Single Model." In *EMNLP*, pp. 1846-1851. 2014.

[Lu et al., 2016] Lu, D., Voss, C.R., Tao, F., Ren, X., Guan, R., Korolov, R., Zhang, T., Wang, D., Li, H., Cassidy, T. and Ji, H., Cross-media Event Extraction and Recommendation. In *Proceedings of NAACL-HLT (Demonstrations)*.

[Majumder et al., 2007] Majumder, P., Mitra, M., Parui, S. K., Koley, G., Mitra, P., & Datta, K. (2007). YASS: Yet another suffix stripper. *ACM transactions on information systems (TOIS)*, 25(4), 18.



[Mogadala and Rettinger, 2016] Mogadala, A., & Rettinger, A. Bilingual Word Embeddings from Parallel and Non-parallel Corpora for Cross-Language Text Classification. In Proceedings of NAACL-HLT.

[Mousam et al., 2012] Mousam, Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. (2012, July). Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 523-534). Association for Computational Linguistics.

[Pighin et al., 2014] Pighin, Daniele, Marco Cornolti, Enrique Alfonseca, and Katja Filippova. "Modelling events through memory-based, open-ic patterns for abstractive summarization." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 892-901, Baltimore, USA.

[Steinberger et al., 2013] Steinberger, R., Ehrmann, M., Pajzs, J., Ebrahim, M., Steinberger, J., & Turchi, M. (2013, September). Multilingual media monitoring and text analysis—Challenges for highly inflected languages. In *International Conference on Text, Speech and Dialogue* (pp. 22-33). Springer Berlin Heidelberg.

[Tao et al., 2013] Tao, F., Lei, K. H., Han, J., Zhai, C., Cheng, X., Danilevsky, M., ... & Kanade, R. (2013, August). EventCube: multi-dimensional search and mining of structured and text data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1494-1497). ACM.

[Tao et al., 2014] Tao, F., Brova, G., Han, J., Ji, H., Wang, C., Norick, B., ... & Sun, Y. (2014, June). NewsNetExplorer: automatic construction and exploration of news information networks. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data* (pp. 1091-1094). ACM.

[Yan et al., 2011] Yan, R., Kong, L., Huang, C., Wan, X., Li, X., & Zhang, Y. (2011, July). Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 433-443). Association for Computational Linguistics.

17. You can add up to ten additional attachments, not counting the industry participation letter and the forwarding letter from the institute, with a combined file size less than 2MB

#### Attachments:

1. Relevant papers from PIs
2. System Architecture
18. Forwarding letter for the proposal from institute authority; *this will be an attachment*

Forwarding letter from IITKGP has been attached.