Research Networking Programmes

Science Meeting – Scientific Report

Scientific report (one single document in WORD or PDF file) should be submitted online within two months of the event. It should not exceed seven A4 pages.

***Proposal Title**: PAN@ FIRE - Cross Language Indian News Story Search*

Application Reference N°: 5275

1) Summary

PAN is a networking initiative that operates around the topics of plagiarism, authorship, and social software misuse. In the PAN@FIRE, this year we continued our focus on text reuse from a cross- language perspective. The aim of this workshop is to create technologies for extraction of parallel and comparable cross-lingual data from the widely available quasi-comparable data e.g. news stories. As this was the second edition we could make training data available since very beginning to the community. Previous year being the first year, not many trams could participate and hence PAN@FIRE did not see huge participation. Due to less participation, the problem task remained unsolved and hence we decided to continue the task this year as well without making new changes. In total we received 23 runs from eight teams which employed very different strategies contributing to the diverse pool for the relevance judgment. Out of eight teams, six teams submitted the working note papers and four teams presented at the workshop. The session included the participants talks and the overview of the task was given in a separate session as part of the main conference program.

2) Description of the scientific content of and discussions at the event

The focus of the CL!NSS track this year was is to evaluate the identification of news stories with same news event and focal event in a cross-language environment[1]. In order to make it easier for the participating teams and increasing the community interest, we focused only on a single language

1 1 For the definitions of news event and focal event please refer to Task Description page of http://www.dsic.upv.es/grupos/nle/clinss.html

pair English-Hindi unlike two pairs in previous year English-Hindi and English-Gujarati. The task statement is as below and also depicted in Fig. 1

For the given source collection S containing news stories in Indian languages $L_i \in L_s$ and the target collection T, containing news stories in English L_t, the task is to link each news story $t \in T$ to $s \in S$ where (t;s) share shame news event or focal event for each L_i.

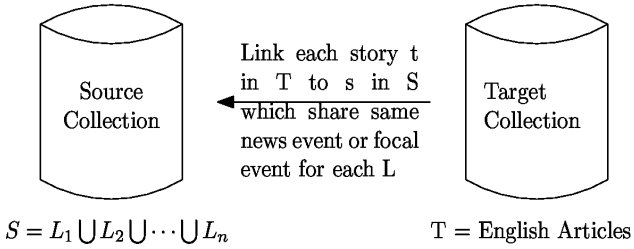

Figure 1: Framework of the CL!NSS task for 2013 edition

The task is similar to a (cross-language) duplicate detection task where the query is an entire document and similar documents must be found from a set of known documents. The task is not trivial because similar stories may exist with varying degrees of overlap (e.g. a story written in English and used as the query text may be a subset of a longer story written in a different language, and vice-versa). This being the second edition for this task, we could provide sufficient training period compared to previous year. Out of 16 registered teams, 8 teams could submit their runs. Each team was allowed to submit three runs in order to allow them different strategies or settings of the same system. In total 23 runs were received. We have written a detailed overview paper of the CL!NSS track which can be accessed from the FIRE working-notes as well as from the CL!NSS web-page. Interestingly all the eight teams tried very different strategies. All the participating teams this year made use of the lessons learnt from the previous year and it reflects from the results. Participants wrote the details about their runs as working note papers included in the FIRE working-notes[2]. At the workshop the participants, organisers and attendees actively discussed the strategies opted and the results.

3) **Assessment of the results and impact of the event on the future directions of the field (up to two pages)**

The previous year results (NDCG@1 = 0.32) showed that, there is still a big scope of improvement and can only be achieved by wide and active participation. This being the second edition of the task, we took measures to increase the participation. The results achieved this year is depicted in Fig. 2.

2 Available at http://www.isical.ac.in/~fire/working-notes.html

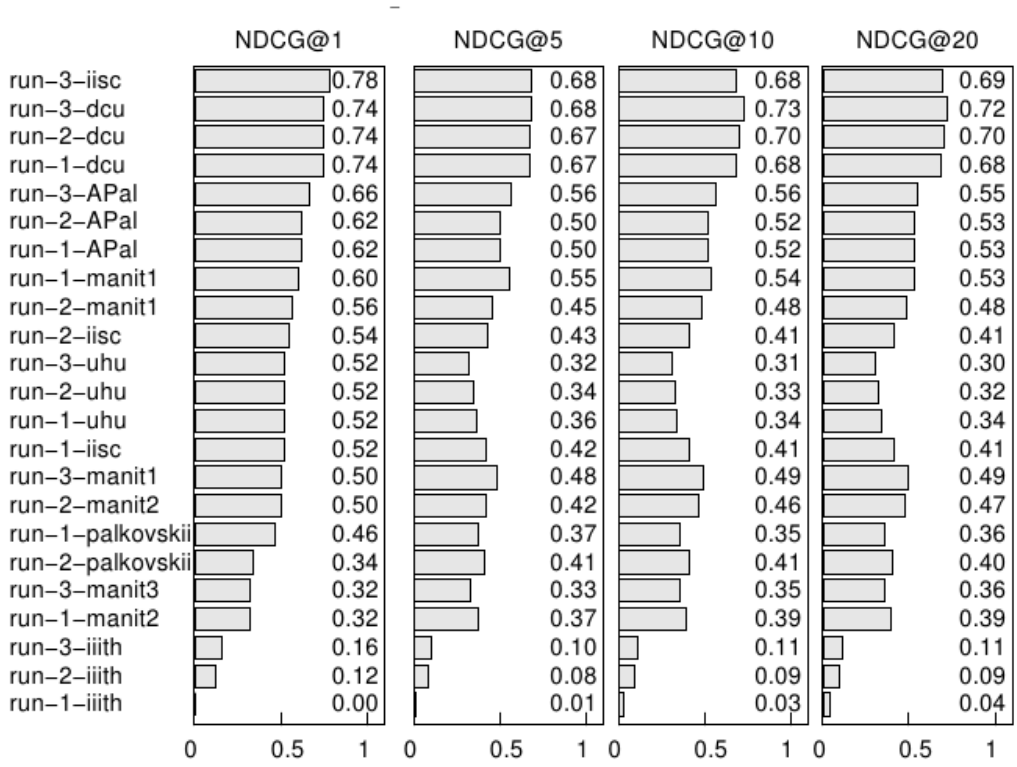

Figure 2: Overall evaluation results for English-Hindi partition. The left hand side information corresponds to the run. The ranking is upon the NDCG@1 values.

Below are some points learnt from this year's participation.

- **The scores achieved this year are quite high NGCD@1 0.78 vs. Last year's best 0.32**
- **Incorporating meta-information explicitly in similarity estimation helps**
- **It is also observed that carefully selecting query terms from target documents help to improve the performance**
- **Although, the approaches are motivated to treat the problem as ranking, more sophisticated modeling of stories would certainly help to determine same focal events**

PAN@FIRE had its future plans outlined from the task proposal as shown in Fig. 3. Based on the discussions and lessons from the workshop meeting we intend to continue with the news story linking task for one more year and then after consolidating the task, we will move forward to fragment extraction.

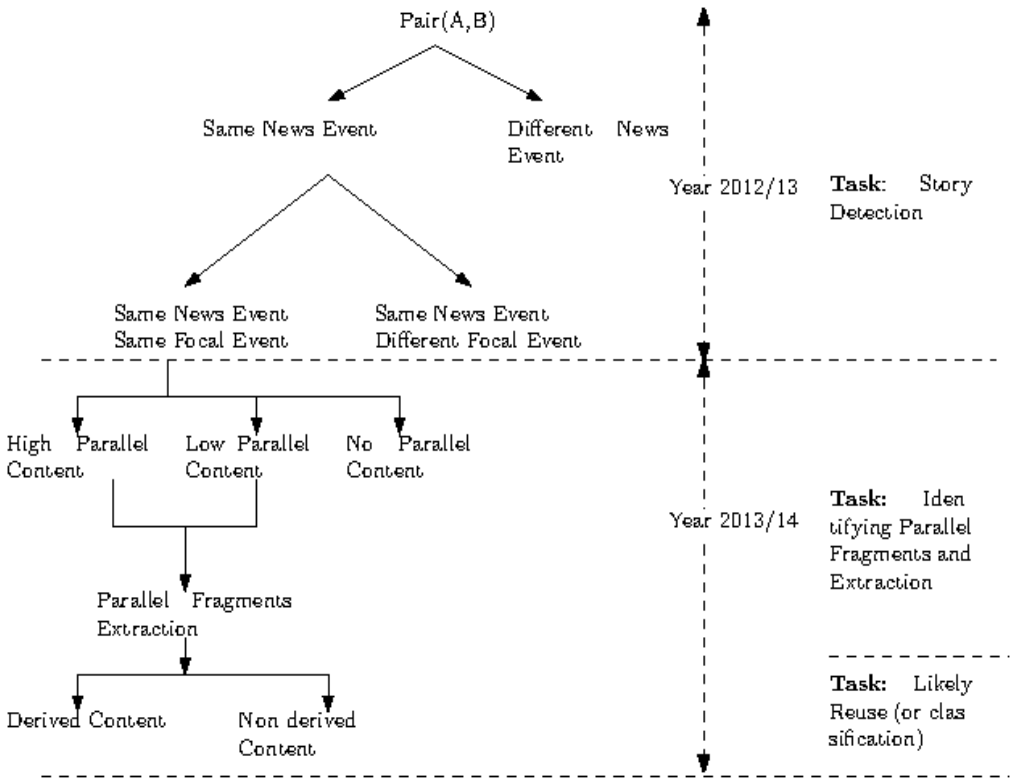

Figure 3: Summary of current and future tasks of the CL!NSS track

Annex 4a: Programme of the meeting and Speakers

Time	Session: Day 1, 4th December
12:00-12:15	CL!NSS Track Overview Parth Gupta (Universitat Politecnica de Valencia, Spain)
	Session: Day 2, 5th December (Chair: Paolo Rosso)
15:30 – 16:45	Participant Talks Amogh Param (IISc Bangalore, India) Piyush Arora (CNGL Dublin City University) Aarti Kumar (MANIT Bhopal, India) Sujoy Das (MANIT Bhopal, India)

Team Members	Institution
Arpan Pal	University of Surrey
Piyush Arora, Jennifer Foster, Gareth Jones	Dublin City University
Goutham Tholpadi, Amogh Param	Indian Institute of Science
Aarti Kumar	Maulana Azad national Institute of Technology, Bhopal
Dr. Sujoy Das	Maulana Azad national Institute of Technology, Bhopal
Yurii Palkovskii, Alexei Belov	Zhytomyr State University Ukraine SkyLine LLC
Diego A. Rodriguez Torrejon, Jose Manuel Martin Ramos	Universidad de Huelva
Ajay Dubey, K Santosh	IIIT-Hyderabad