Report from PAN@FIRE (CL!NSS) Evaluation Lab at FIRE'12

Parth Gupta and Paolo Rosso
Natural Language Engineering Lab - ELiRF
Department of Information Systems and Computation
Universitat Politècnica de València, Spain
http://users.dsic.upv.es/grupos/nle

Alberto Barrón-Cedeño
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya, Barcelona, Spain
http://www.lsi.upc.edu/

Paul Clough and Mark Stevenson
Information Retrieval and Natural Language Processing Group,
University of Sheffield, United Kingdom
http://ir.shef.ac.uk/ & http://nlp.shef.ac.uk/

Sobha Lalitha Devi
Computational Linguistics Research Group
AU-KBC Research Centre, Chennai, India
http://nlp.au-kbc.org/

Webpage: http://www.dsic.upv.es/grupos/nle/clinss.html

February 5, 2013

Abstract

We report on the PAN track at FIRE on Cross-Language !ndian News Story Search (CL!NSS). This report contains the details of the scientific contents and the discussions generated from the workshop. Finally, we discuss the future plans and depict the detailed program of the workshop.

Summary

PAN is a networking initiative that centers around the topics of plagiarism, authorship, and social software misuse. In the PAN@FIRE, this year we shift our focus from plagiarism detection to text reuse and that 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. This was the first edition towards achieving this goal and hence the benchmark data creation was the major milestone. We created a benchmark corpus from the News seeds and generated the relevance judgments for the automatic evaluation. In total we received 8 runs from three teams which employed very different strategies contributing to the diverse pool for the relevance judgment. Out of three teams, two teams submitted working note papers and also presented at the workshop. The workshop was held in one session of 2 hours. The session included the participants talks. The session included a novel component of discussant talk which tried to discuss the highlights of the workshops and its relevance to the current and/past initiatives. There was also a panel discussion focusing on where the cross-language text reuse detection stands and what are the future directions.

Task and Participation Details

The focus of the CL!NSS track this year is to evaluate the identification of news stories with same news event and focal event in a cross-language environment[1]. The Indian languages involved in the source collection are Hindi and 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.*

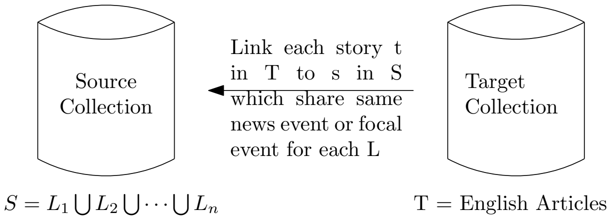
Source Collection
Link each story t in T to s in S which share same news event or focal event for each L
Target Collection
$S = L_1 \bigcup L_2 \bigcup \cdots \bigcup L_n$
T = English Articles

Figure 1: Framework of the CL!NSS task for 2012 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

[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`

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 first edition for this task, majority of time for the task cycle was spent of collection creation and participants were given the time of 45 days to submit the runs. Out of 10 registered teams, 3 teams could submit their runs. Each team was allowed to submit three runs per language pair in order to allow them different strategies or settings of the same system. In total 7 runs were received for English-Hindi pair while only 1 run for English-Gujarati pair. Gujarati is much resource poor compared to Hindi which was the main cause of non-participation.

The track organisers 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 webpage. Interestingly all the three teams tried very different strategies. The strategies without use of machine translation to handle cross-language similarity were also tested among the participant teams and achieved very good results which became a major attraction of discussion. Participants wrote the details about their runs as working note papers included in the FIRE working-notes[2]. At the workshop all the participants, organisers and attendees actively discussed the strategies opted and the results.

Discussant Talk

The lab organisers invited Doug Oard to give a discussant talk in the workshop. The purpose of discussant talk was to take a stand on the task after the participants talks and draw take-away messages. Another interesting and important part of this talk was to compare the CL!NSS task with some other evaluation initiative which in the past or currently running something very near to the lines of CL!NSS and to show how this workshop can benefit from them.

Panel Discussion

The workshop featured the panel discussion on the future directions to the task of CL!NSS with panel members were Mandar Mitra, Doug Oard, Jaap Kamps and Johannes Leveling. There were discussions on the evaluation strategies, less participation and community building. The panel showed the direct interest of Machine Translation community in the task and suggested to actively include them in the further calls of participation. The panel also suggested to continue with the news story linking task for one more year

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

to properly address the issue with time than moving forward to fragment identification in the linked news stories.

Results and Future Plans

The participation results showed that, there is still big scope of improvement and can only be achieved by wide and active participation. This being the first edition of the task, we expect to improve the results of the news story linking task in the upcoming editions. PAN@FIRE had its future plans outlined from the task proposal as shown in Fig. 2. 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.

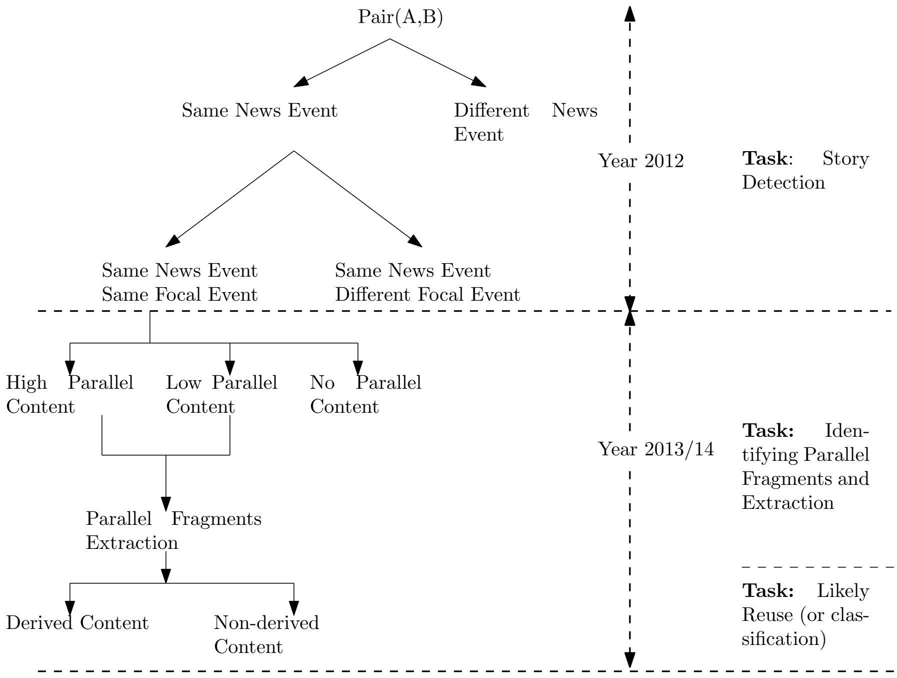
Pair(A,B)
Same News Event
Different News Event
Year 2012
Task: Story Detection
Same News Event Same Focal Event
Same News Event Different Focal Event
High Parallel Content
Low Parallel Content
No Parallel Content
Year 2013/14
Task: Identifying Parallel Fragments and Extraction
Parallel Fragments Extraction
Task: Likely Reuse (or classification)
Derived Content
Non-derived Content

Figure 2: Summary of current and future tasks of the CL!NSS track

Track Program and Speakers

Following table contains the detailed program of CL!NSS which is also available on CL!NSS and FIRE webpage.

Time	**Session: Day 1, 17th Dec**
11:45 – 12:15	CL!NSS Track Overview Parth Gupta (*Universitat Politècnica de València, Spain*)
Time	**Session: Day 2, 18th Dec. Chair:** Paolo Rosso
15:45 – 16:45	CL!NSS track talks Yurii Palkovskii (*Z. S. Uni. & SkyLine LLC, Ukraine*) Nitish Aggarwal (*DERI, Ireland*)
16:45 – 17:15	CL!NSS discussant talk Doug Oard (*University of Maryland, USA*)
17:15 – 17:45	CL!NSS panel discussion Mandar Mitra (*Indian Statistical Institue, India*) Doug Oard (*University of Maryland, USA*) Jaap Kamps (*University of Amsterdam, The Netherlands*) Johannes Leveling (*CNGL, Dublin City University, Ireland*)