PAN@FIRE 2013: Overview of the Cross-Language !ndian News Story Search (CL!NSS) Track[*]

Parth Gupta[1], Paul Clough[2], Paolo Rosso[1]
Mark Stevenson[2], and Rafael E. Banchs[3]

[1] NLE Lab, Universitat Politècnica de València, Spain
[2] University of Sheffield, UK
[3] Institute for Infocomm Research, Singapore

pgupta@dsic.upv.es p.d.clough@sheffield.ac.uk
prosso@dsic.upv.es M.Stevenson@dcs.shef.ac.uk
rembanchs@i2r.a-star.edu.sg

Abstract. The automatic alignment of documents in a quasi-comparable corpus is an important research problem for a resource poor cross-language technologies. News stories form one of the most prolific and abundant language resource. This edition of PAN@FIRE task, cross-language !ndia news story search (CL!NSS), continues to address the news story linking task across languages English and Hindi. We present the overview of the track with results and analysis.

1 Introduction

Cross-language technologies depend heavily on the natural language resources available for the language pair. Usually these resources are of two types, *(i)* manually generated e.g. linguistic bilingual dictionary and *(ii)* automatically generated e.g. statistical bilingual dictionary. Manually generated resources tend to be accurate but are very costly to produce (requiring significant human effort), can not be generalised across the languages and need to be constantly updated. Automatically generated resources are often more convenient to generate but are dependent on the availability of suitable cross-lingual training data. This training data is normally aligned corpora for many language pairs it is not available or of good enough quality. In the past many approaches have exploited aligned corpora to develop cross-language technologies like cross-language information retrieval (CLIR) [10, 15], machine translation (MT) [2, 12], text mining [21] etc. Bilingual corpora are generally categorised into three categories: *(i)* parallel, *(ii)* comparable and *(iii)* quasi-comparable. In comparable corpora the documents are aligned, i.e. each document has a corresponding one in the other language on the same topic. In quasi-comparable corpora the documents are not topically aligned

[*] This is an updated version of 2012 Overview for Year 2013.

and some off-topic documents are also included. Parallel and comparable corpora are more valuable than quasi-comparable corpora, but are more difficult to obtain.

A convenient source for parallel corpora is parliament proceedings of multilingual regions e.g. the European Union and India. Such corpora include Europarl [4] and JRC-Acquis [5]. However, these corpora are very specific to a single domain and do not reflect the vocabulary required for general applications. One of the major sources of documents is Wikipedia[6] which has been used to create a variety of cross-language technologies [16, 20]. However, Wikipedia's coverage is limited for many languages. Unlike parallel and comparable corpora, quasi-comparable corpora are readily available on the web. One of the major sources is news stories that are published in more than one language.

This edition of the cross-language !ndian news story search (CL!NSS) task continues to focus on journalistic text re-use like last year. News agencies are a prolific source of text on the Web and a valuable source of text in multiple languages. News stories generated by different authors, whether independently or derived, typically exist as separate entities and consequently there is a need to link them.

Linking news stories covering the same events written in different languages offers a number of benefits. For example, in a multilingual environment, such as India, where the same news story is covered in multiple languages, a reader might want to refer to the local language version of a news story. News stories covering the same event(s), published in different languages, may also be rich sources of both parallel and comparable text, for example, parallel fragments in the news story, e.g. direct quotes or translation equivalents; comparable fragments, e.g. paraphrases. Therefore, identification of similar news stories written in multiple languages offers a valuable multilingual resource. In the case of Indian languages there exist limited language resources for natural language processing (NLP) and information retrieval (IR) tasks. For instance, identifying comparable and parallel documents on the web would offer a potential (and abundant) source for deriving bilingual dictionaries and training statistical MT systems [11, 12].

In Section 2 the task description is given followed by the details of corpus in Section 3. Section 4 describes the evaluation and analysis. Final remarks are found in Section 5.

2 Task Description

This year the aim of PAN@FIRE[7] task continues to identify the same news story written in multiple languages (a problem of cross-language news story detection). The results reported in the previous edition of CL!NSS demonstrate the room for further consideration of the topic [7]. The task involves identifying and linking

[4] http://www.statmt.org/europarl/

[5] http://optima.jrc.it/Acquis/index_2.2.html

[6] http://www.wikipedia.org/

[7] http://pan.webis.de

news stories covering the same event, but published in different languages. In the upcoming editions of CL!NSS the aim will be to extract equivalent text fragments (parallel and comparable) and finally to identify cases of potential co-derivation between documents (a common scenario in journalism as content is shared between news agencies and newspapers). The latter task has been extensively studied in monolingual settings, but not as deeply in cross-language ones [3, 4]. We divide the problem of CL!NSS into three distinct tasks:

1. **Story detection**: given a story in one language find the note covering the same story but written in a different language.
2. **Fragment detection**: given a pair of similar (comparable) news reports, extract parallel text fragments (e.g. sentences, phrases etc.).
3. **Story/fragments classification (derived or non-derived)**: in some cases news reports are co-derived, i.e. one of the stories has been based on the other one.

2.1 Definitions

A news story communicates to its readers information about an event or series of events (an event being something that happens at a specific time and location). For example, a news story might follow events in Syria. An article/report refers to the story which is published on a specific date and appears in a newspaper or online. The article will typically report the story (or part of the story) from a particular aspect/viewpoint for a particular audience (e.g. written in a specific language). A news story will often consist of a collection of articles. Some stories will be 'one-off' (those describing events which occur only on one day); others 'running' where events across more than one day will be reported. Another example might be Wimbledon, an event which occurs annually. A particular article might describe the outcome of the final match of the tournament.

As previously stated, locating similar news stories has a number of potential uses. However, a key issue is deciding when two news articles are similar. One would assume that more similar news articles are more comparable and subsequently more useful, e.g. as a source of comparable text. To provide a basis for judging similar news stories, we adopt the scheme devised by [1]. This scheme is well-suited to the CL!NSS task and is applicable for monolingual and cross-language news stories comparison. The scheme is based on identifying the content and structure of news articles consisting of: *focal event*, *background event* and *news event*.

- **Focal event**: The main event or events which provide a focus for the news story. The focus here is considered as a very specific level of information. Very often the most recent event in an unfolding news story, it also provides a particular angle or perspective for the report. For example, "Nagaland Congress seeks NPF backing for Pranab". Here the focal event of the news story is to seek the backing by some entity from another entity in support of Pranab for *Presidential Elections in India.*

- **Background event**: An event that plays a supporting role in the text, providing context for the focal events. It may include: related events leading up to the focal events; examples of similar past events; and definitions, explanations or descriptions of things, people and or places which play a role in the focal events.
- **News event**: A group of related events, broader than and including the focal event, which may be reported over time in different news text installments. This is related to the concept of "real-world event". All the news stories which are related to a particular event taking place in the world share the same news event. For example, all the news articles related to current *Presidential Elections in India* including the early articles on the possible candidates, controversies raised in between to the last stories of the completion of the election and the results of the election fall under the same news event.

2.2 Overview of Current Task and Planned Future Extensions

Fig. 1 summarises the proposed tasks for CL!NSS and highlights different forms of similarity that may exist between news stories. Let A and B be a pair of news reports written in different languages that are loosely related, i.e. on the same theme/topic or category. This is typically the goal of IR systems: to identify documents which are relevant to a given query (where relevance typically reflects topicality). Information about the news story, such as its category (Entertainment, Sports, News etc.), together with the date of publication, is sometimes available in the metadata.

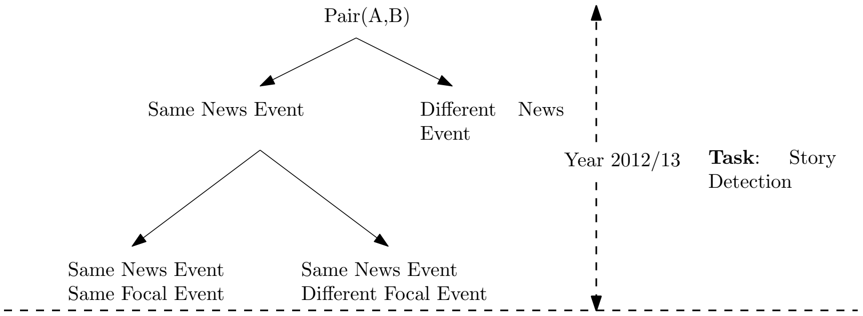

Fig. 1. Summary of tasks in CL!NSS and the relationship between a pair of news articles Pair(A, B)

The second level in Fig. 1 identifies where two news reports are basically describing the same focal events, i.e. they could be the same news report produced in multiple languages. The focus of the news stories are similar and the same events are basically reported in each article.

2.3 Task Statement

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 like last year. The Indian language involved in the source collection is Hindi. The task statement is as below and also depicted in Fig. 2

> *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.*

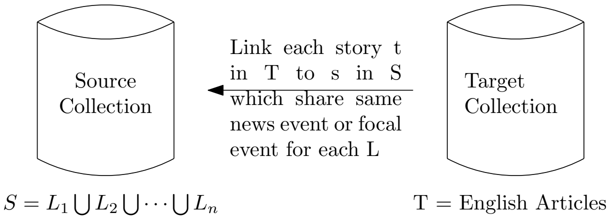
Source Collection
Link each story t in T to s in S which share same news event or focal event for each L
Target Collection
$S = L_1 \bigcup L_2 \bigcup \cdots \bigcup L_n$
T = English Articles

Fig. 2. Framework of the CL!NSS task for 2012/13 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). Table 1 provides an example of relevant and non-relevant focal event for an English-Hindi (en-hi) text pair. Although both source articles share the same news event as the target, the focal event is similar for source article 1 (relevant) but different for source article 2 (non-relevant).

3 Corpus

The corpus contains a set of potential source news stories S written in Hindi and a set of target news stories T written in English. The documents are available with basic meta information like title of the news stories and publication date along with its content and is formatted according to the markup depicted in Fig. 3.

The basic statistics of the corpus and its partitions are presented in Table 2. The source collection is created by crawling and cleaning the online archives from 2010 of the Navbharat Times[8].

[8] http://navbharattimes.indiatimes.com/

Article	**Title**	**Relevance Level**
Target	There's lot more to talk than my 50th Test ton: Tendulkar *cf.* `english-document-00006.txt`	
Source1	मेरी 50वीं सेंचुरी के अलवा भी कई बातें हैं: तेंदुलकर *There are many things except my 50th centurey* *cf.* `hindi-document-24799.txt`	2 (same focal event)
Source2	सचिन ने बनई सेंचुरी की फिफ्टी *Sachin makes fifty in century* *cf.* `hindi-document-08018.txt`	1 (same news event)

Table 1. Example English-Hindi text pairs describing the same news event but different focal events

```
<story>
  <title>xxxxxx</title>
  <date>xx-xx-xxxx</date>
  <content>
    xxxxxx
  </content>
</story>
```

Fig. 3. Text mark-up of the documents in the corpus.

Table 2. CL!NSS 2012 corpus statistics. The statistics are shown for the source partition D_{hi} (Hindi) and a target collection D_{en}. The column headers stand for: $|D|$ number of documents in the corpus (partition), $|D_{tokens}|$ total number of tokens, $|D_{voc}|$ total size of vocabulary (unique terms). k= thousand, M = million.

Partition	$\|D\|$	$\|D_{tokens}\|$	$\|D_{voc}\|$
D_{en}	25	9.3k	2.5k
D_{hi}	50691	15.6M	143k

3.1 Target Collection Generation

In order to prepare the target English collection D_{en}, we first crawled and cleaned the news stories published in the Times of India[9]. After indexing these documents, we retrieved the most relevant document from the index for each query of the FIRE ad-hoc topics[10] according to the BM25 score. Using FIRE topics to sample the collection helped to select a diverse and well distributed set of target news stories.

[9] `http://timesofindia.indiatimes.com/`

[10] `http://www.isical.ac.in/~fire/data/topics/adhoc/en.topics.76-125.2010.txt`

4 Evaluation

All the participants were asked to submit a rank-list of up to 100 source news stories for each target news story in one run. Each team could submit up to three such runs per language pair. To evaluate the performance we prepared a pool of the source news stories for each target news story from the submitted runs. This pool was manually judged to prepare the relevance judgment files (qrels). The annotators were asked to assign each pair of target and source news stories one of the following labels.

- **"0"** different news event
- **"1"** same news event but different focal event
- **"2"** same news event and same focal event

4.1 Evaluation Framework

The participants were asked to submit their runs in the form of a ranked list as shown below.

```
english-document-00001.txt  Q0  hindi-document-00345.txt  1  0.4644
english-document-00001.txt  Q0  hindi-document-42325.txt  2  0.2823
                                 ⋮
english-document-00050.txt  Q0  hindi-document-23443.txt  100  0.1123
```

We use NDCG@k [8] to evaluate the retrieval and the linking of the news stories. As the relevance is not binary and relevance levels are graded categories, NDCG@k is more suitable.

4.2 Participation Overview

In total 8 teams participated who submitted total 23 runs. All the teams opted for very different strategies and a wide variety of settings. A pool or relevant documents is created by taking top 10 source documents for each target document for all the submitted runs. This pool is manually judged to annotate the relevance level.

DCU [14] first retrieved the candidate list of source documents using translated target documents using TF-IDF like algorithms. The translation is obtained from automatic machine translation APIs like Google Translate and Bing. Such candidate source documents are re-ranked using other features like combined translation (from different APIs), transliteration, variable length of summary of target documents instead of complete document, named entities only and publication date based proximity.

APal [13] also used Google Translate to normalise the language of the target stories. Then the set theory based Jaccard Coefficient was used to rank the source documents based on title, unique words and frequent terms of the content of the article. A publication date based similarity booster is also applied.

`IISC` [17] used Microsoft Translator API to translate target documents. The similarity between target and source documents is weighted sum of TF-IDF scores between title-title, title-content and content-content of the target and each source document. The weights are learnt from the training partition using a machine learning method.

`MANIT-2` [5] used a publicly available bilingual dictionary and in-house transliteration engine to compare target and source news stories across the languages. The English terms that are not present in the dictionary are stemmed for look-up.

`MANIT-1` [9] also used the bilingual dictionary and transliteration engine same as [5]. The difference was in the query formulation. They used only title terms with terms from content like proper nouns and/or those with frequency higher than average frequency of terms in the collection as query.

`UHU` [18] uses the contextual n-grams as the unit of comparison. They first translate the target news stories using Google Translate. The source collection is indexed using high accuracy information retrieval system (HAIRS) and during similarity some n-grams like very frequent n-grams are discarded to avoid by-chance matching using 'reference monotony prune strategy' (RMPS). More details on HAIRS and RMPS can be found in [19]. Their system inherently does not address the problem of linking news stories as a ranking problem, instead they return the list of source articles which are quite likely to contain parallel content. It is inspired by the objective of cross-language plagiarism detection.

`IIIT-H` index source collection with their transliteration using a naive rule based phonetic transliteration engine. The query is formulated using the named entities from the target story. The query is further expanded for transliteration equivalents using Editex algorithm [22].

4.3 Results and Analysis

The results obtained by the participants are depicted in Fig. 4. As the ultimate goal is to extract parallel/comparable content from the linked stories, it is important to find stories with same focal event on top ranks. Therefore, the important measure is NDCG@1. The best results according to NDCG@1 are obtained by `IISC` [17] closely followed by `DCU` [14].

In the relevance judgment we found that for some target news stories did not have any relevant source news stories. We found 3 such topics. The frequency of relevant documents for each target news story is plotted in Fig. 5 and evaluation results with those target news stories which have at least one relevant source news story is depicted in Fig. 6.

This year most of the participants exploited the title information of the news stories separately in the similarity estimation. The most popular strategy to normalise language was observed to be automatic machine translation services except `MANIT-1 & MANIT-2` [5, 9] that used bilingual dictionary and `IIIT-H` that

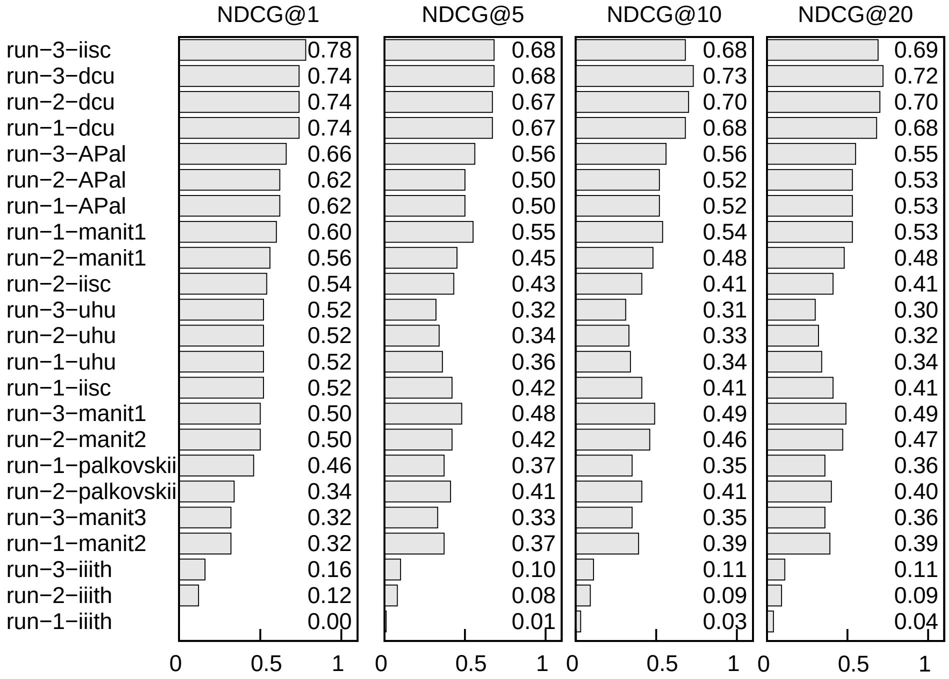

Fig. 4. Overall evaluation results for English-Hindi partition. The left hand side information corresponds to the run. The ranking is upon the NDCG@1 values.

used named entities and their transliteration. It is observed that carefully formulating the query from target document is helpful than simply taking the complete target document as query. Many participants used some smart natural language processing steps like named entity recognition[DCU, MANIT-1, IIITH], transliteration [DCU, MANIT-1, MANIT-2, IIIT-H] and summarization [DCU] for formulating the query.

4.4 Further Challenges

The results achieved in this edition of CL!NSS are significantly higher than previous edition. The approaches in general are motivated to solve the task as a ranking problem to some extent, where stories with same focal event should rank higher than stories with same news event but different focal event. Sometimes, target news stories do not have a focal event at all. In such cases, it is important to model the news story with objective annotation like *who* is doing *what, where* and *when*. The span of news events are very different in nature, from *one-off* news stories which just appear for once to those spanned over a few years *e.g. space missions* to periodic events *e.g. a match between two teams in a tournament.* Some times more than one news events swiftly join one another such that to even

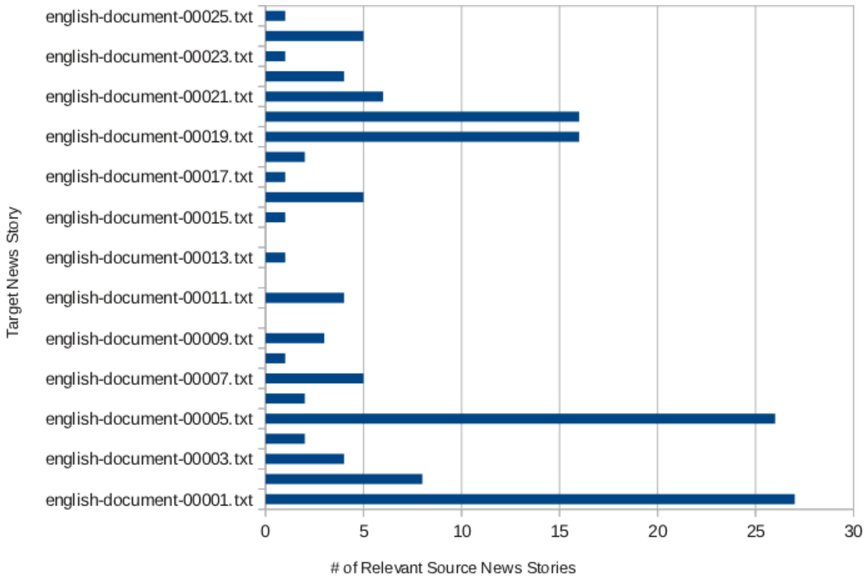

Fig. 5. The frequency of relevant source documents for each target document in the corpus.

manually decide the news event of the target story is difficult. In such cases, if the end goal is to extract parallel or comparable content, one can cascade story linking by a stricter extraction module. Although automatic translation services help to achieve good performance, the quality of translation they provide is also an important factor to analyse especially for resource poor language pair. It is observed that carefully selecting query terms from target documents help to improve the performance and hence a rich analysis of how such terms are handled in the translation service can greatly affect the performance.

5 Remarks and Future Work

In this paper we presented an overview of the second edition of the cross-language !ndian news story search (CL!NSS) track at FIRE where the task was focused on linking news stories across Hindi-English language pair which share same focal event and/or news event. We presented the evaluation of the runs submitted by the participating teams along with analysis. This year 8 teams submitted 23 runs in total. Although, the approaches using automatic machine translation services dominated those which did not rely on them, it is observed that carefully formulating the query help in performance improvement compared to use the complete target document as query. Some further challenges are also identified in this report.

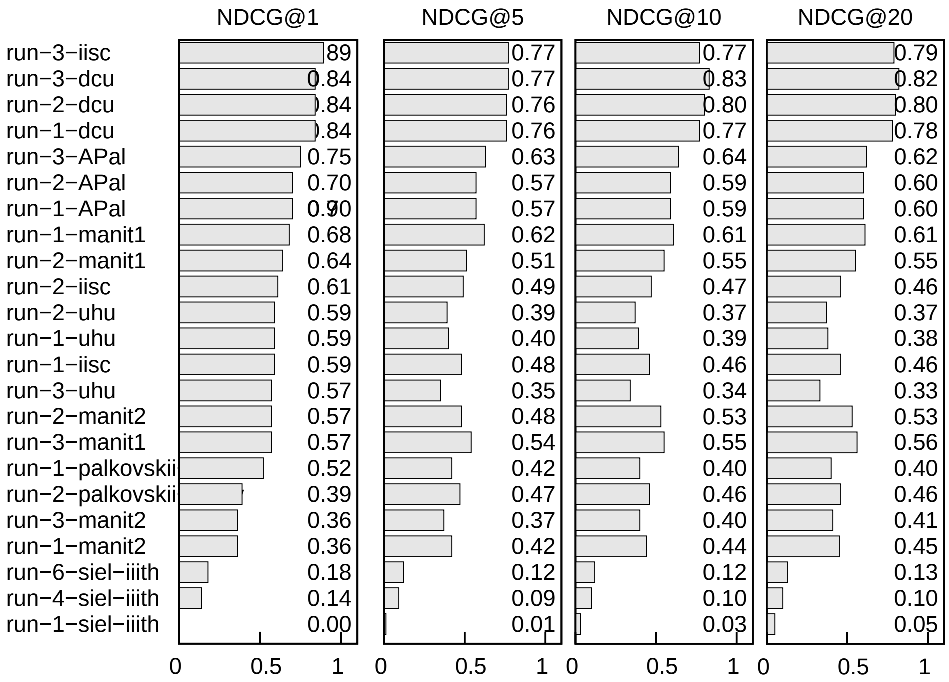

Fig. 6. Evaluation results for target news stories which have at least one relevant source news story. The ranking is upon the NDCG@10 values.

Looking at the high performance scores of the story linking task this year, we believe it will be interesting in the future to identify parallel/comparable fragments from linked articles.

6 Acknowledgment

We thank European Science Foundation (ESF) and ELIAS project for supporting the track. This research work is partially funded by the WIQ-EI (IRSES grant n. 269180) and done in the framework of the VLC/Campus Microcluster on Multimodal Interaction in Intelligent Systems and partially funded by the DIANA APPLICATIONS - Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project.

References

1. Barker, E., Gaizauskas, R.: Assessing the comparability of news texts. In: Chair), N.C.C., Choukri, K., Declerck, T., Doayan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (May 2012)

2. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The Mathematics of Statistical Machine Translation: Parameter Estimation. Comput. Linguist. 19(2), 263–311 (Jun 1993)
3. Clough, P.: Measuring text reuse in a journalistic domain. In: In Proc. of the 4th CLUK Colloquium. pp. 53–63 (2001)
4. Clough, P., Gaizauskas, R., Piao, S.S.L., Wilks, Y.: Meter: Measuring text reuse. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 152–159. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
5. Das, S., Kumar, A.: Performance Evaluation of Dictionary Based CLIR Strategies for Cross Language News Story Search. In: FIRE [6]
6. FIRE (ed.): FIRE 2013 Working Notes. Fifth International Workshop of the Forum for Information Retrieval Evaluation (2013)
7. Gupta, P., Clough, P., Rosso, P., Stevenson, M.: PAN@FIRE: Overview of the Cross-Language !ndian News Story Search (CL!NSS) Track. In: Proceedings of the Fourth Forum for Information Retrieval Evaluation. FIRE '12, India (2012)
8. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20(4), 422–446 (Oct 2002)
9. Kumar, A., Das, S.: Pre-Retrieval based Strategies for Cross Language News Story Search. In: FIRE [6]
10. Littman, M., Dumais, S.T., Landauer, T.K.: Automatic cross-language information retrieval using latent semantic indexing. In: Cross-Language Information Retrieval, chapter 5. pp. 51–62. Kluwer Academic Publishers (1998)
11. Munteanu, D.S., Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora. Comput. Linguist. 31(4), 477–504 (Dec 2005)
12. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. 29(1), 19–51 (Mar 2003)
13. Pal, A., Gillam, L.: Set-based Similarity Measurement and Ranking Model to Identify Cases of Journalistic Text Reuse. In: FIRE [6]
14. Piyush Arora, J.F., Jones, G.J.F.: DCU at FIRE 2013: Cross-Language !ndian News Story Search. In: FIRE [6]
15. Platt, J., Toutanova, K., Yih, W.T.: Translingual Document Representations from Discriminative Projections. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 251–261. EMNLP'10 (2010)
16. Smith, J.R., Quirk, C., Toutanova, K.: Extracting parallel sentences from comparable corpora using document level alignment. In: HLT-NAACL. pp. 403–411 (2010)
17. Tholpadi, G., Param, A.: Leveraging Article Titles for Cross-lingual Linking of Focal News Events. In: FIRE [6]
18. Torrejón, D.A.R., Ramos, J.M.M.: Linking English and Hindi news by IDF, Reference Monotony and Extended Contextual N-grams IR Engine. In: FIRE [6]
19. Torrejón, D.A.R., Ramos, J.M.M.: Text alignment module in coremo 2.1 plagiarism detector - notebook for pan at clef 2013. In: CLEF (Online Working Notes/Labs/Workshop) (2013)
20. Udupa, R., Khapra, M.M.: Improving the multilingual user experience of wikipedia using cross-language name search. In: HLT-NAACL. pp. 492–500 (2010)
21. Udupa, R., Saravanan, K., Kumaran, A., Jagarlamudi, J.: Mint: A method for effective and scalable mining of named entity transliterations from large comparable corpora. In: EACL. pp. 799–807 (2009)

22. Zobel, J., Dart, P.: Phonetic string matching: lessons from information retrieval. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 166–172. SIGIR '96, ACM, New York, NY, USA (1996)