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

Parth Gupta[1], Paul Clough[2], Paolo Rosso[1], and Mark Stevenson[2]

[1] NLE Lab - ELiRF, Universitat Politècnica de València, Spain
[2] University of Sheffield, UK

pgupta@dsic.upv.es p.d.clough@sheffield.ac.uk
prosso@dsic.upv.es M.Stevenson@dcs.shef.ac.uk

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 India news story search (CL!NSS), focuses to address the news story linking task across languages involving two Indian languages - Hindi and Gujarati. 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) [9, 13], machine translation (MT) [4, 11], text mining [18] 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 and some off-topic documents are also included. Parallel and comparable corpora 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 [3] and JRC-Acquis [4]. 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[5] which has been used to create a variety of cross-language technologies [16, 17]. 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 focuses on journalistic text re-use. 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 [10, 11].

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[6] task is to identify the same news story written in multiple languages (a problem of cross-language news story detection). The task will involve identifying and linking 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 [5, 6]. We divide the problem of CL!NSS into three distinct tasks:

[3] `http://www.statmt.org/europarl/`

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

[5] `http://www.wikipedia.org/`

[6] `http://pan.webis.de`

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 [2]. 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.

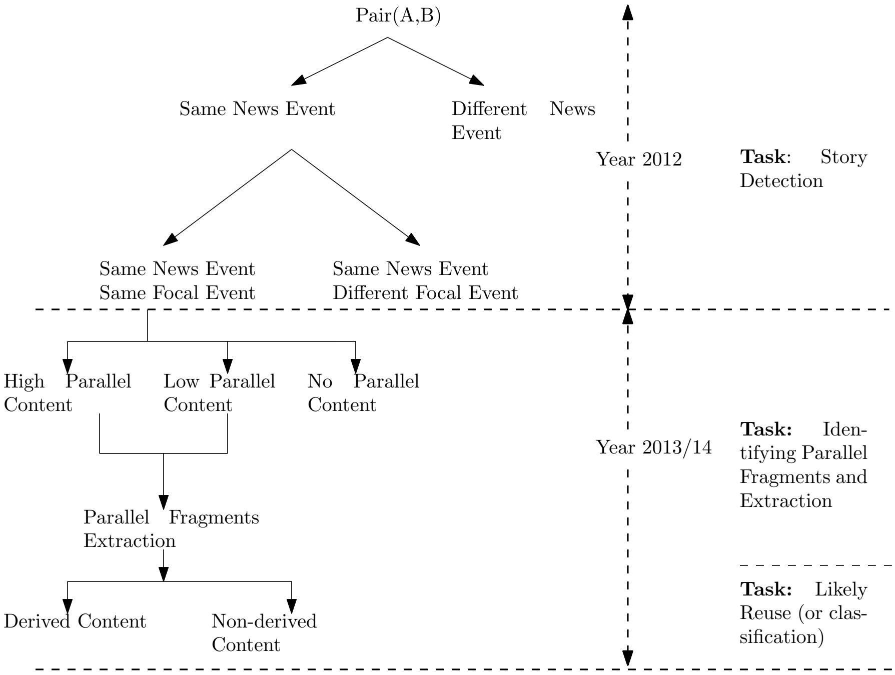
Pair(A,B)
Same News Event
Different News Event
Year 2012
Task: Story Detection
Same News Event Same Focal Event
Same News Event Different Focal Event
High Parallel Content
Low Parallel Content
No Parallel Content
Year 2013/14
Task: Identifying Parallel Fragments and Extraction
Parallel Fragments Extraction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

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. The third level signifies the situation in which fragments between the texts are clearly the same (e.g. they are translation equivalents), although may be subject to forms of paraphrasing or use of colloquial language. This is usually referred as the shared content between them where the granularity of content can be at sentence or sub-sentence level. The final level in Fig. 1 represents the situation in which the similar text fragments are actually derived from each other (e.g. one text fragment is copied from the other or both come from a common source).

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. The Indian languages involved in the source collection are Hindi and Gujarati. 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.*

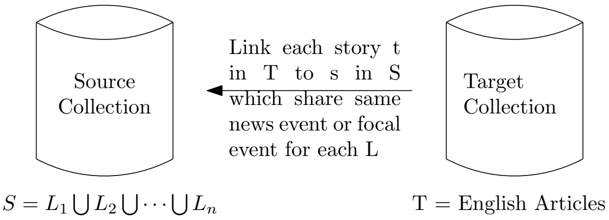
Source Collection
Link each story t in T to s in S which share same news event or focal event for each L
Target Collection
$S = L_1 \bigcup L_2 \bigcup \cdots \bigcup L_n$
T = English Articles

Fig. 2. Framework of the CL!NSS task for 2012 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 Gujarati separately, and a set of target news stories T, written in English.

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

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.

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

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

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 two newspapers: the Navbharat Times[7] for Hindi and Gujarat Samachar[8] for Gujarati.

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

[7] `http://navbharattimes.indiatimes.com/`

[8] `http://www.gujaratsamachar.com/`

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

Table 2. CL!NSS 2012 corpus statistics. The statistics are shown for the two source partitions, D_{hi} (Hindi) and D_{gu} (Gujarati), 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	$\lvert D\rvert$	$\lvert D_{tokens}\rvert$	$\lvert D_{voc}\rvert$
D_{en}	50	21k	4k
D_{hi}	50691	15.6M	143k
D_{gu}	11889	5.8M	282k

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.

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 no longer binary and relevance levels are graded categories, NDCG@k is more suitable.

[10] http://www.isical.ac.in/~fire/data/topics/adhoc/en.topics.176-225.2012.txt

4.2 Participation Overview

In total three teams participated who submitted total 7 runs for en-hi partition and 1 run for en-gu partition. All the teams opted for very different strategies and a wide variety of settings.

Palkovskii [12] primarily translated the english document to the corresponding Indian language, in this case Hindi and Gujarati, using an online machine translation service[11]. Then the following components were used to estimate the similarity between each $t \in T$ and $s \in S$.

- TF-IDF score between title of two documents.
- TF-IDF score between contents of two documents.
- Maximum TF-IDF score between sentences from the contents of the two documents.
- If both documents are published within a window of 10 days, constant 0.5 is added to the final score.

The final score is calculated based on these parameters and the top 100 source news stories returned for each target news story (t) for each language. In addition they harvested a list of the most frequent terms in both languages and ignored them during similarity estimation.

Aggarwal et al. (deriupm) [1] exploited cross-language explicit semantic analysis (CL-ESA) [14] to handle cross-language similarity. In the first run, they reduced the search space of S to those news stories for which publication date is within ± 2 days of that of t. For these pool the similarity is estimated using CL-ESA over a comparable corpus from Wikipedia between the language pair and top 100 documents were returned. In the second run, they increased the window to ± 7 days. While in the third run, they first indexed S using Apache Lucene[12]- an open source search engine and the translated version of $t \in T$ using an online translation service[1]1 was queried to select top 1000 documents. Then top 100 documents were selected by estimating similarity between these retrieved 1000 documents and original t using CL-ESA.

Reddy and Perumal (iiith) [15] first translated the target news story (t) to the language of $s \in S$. They extracted the key-phrases from the target news stories based on n-gram filtration and term weighting techniques. They indexed S collection using Apache Lucene[12] and queried the extracted key-phrases of each t to this index. For each t they obtained several rank-lists depending the number of key-phrases. The similarity score of s was dependent on the number of rank-lists in which it appears (frequency) and its rank in them. Their first run constitute of removing stop-words and taking rank and frequency of s into account. While in second run, they remove stop-words and take only frequency into account. Finally, in their third run they used stop-words and took rank and frequency of s in consideration for similarity estimation.

[11] `http://translate.google.com/`

[12] `http://lucene.apache.org/core/`

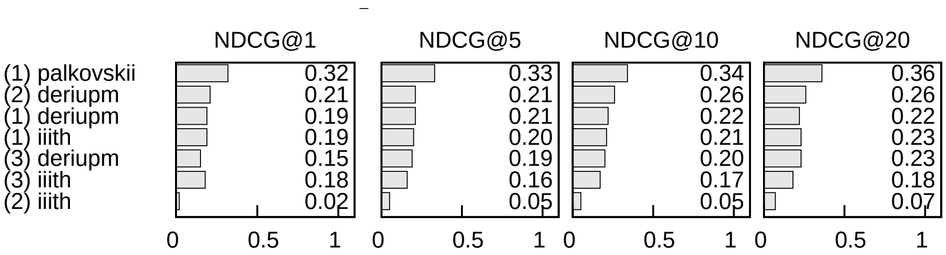

Fig. 4. Overall evaluation results for English-Hindi partition. The left hand side information corresponds to (run number) and team. The ranking is upon the NDCG@10 values.

4.3 Results and Analysis

The results obtained by the participants are depicted in Fig. 4 for English-Hindi (en-hi) and in Table 3 for English-Gujarati (en-gu) partition. The best results for en-hi were obtained by Palkovskii [12] followed by two runs of Aggarwal et al. [1]. Run of Palkovskii [12] benefit by explicitly taking title of the news stories and sentence level similarity into estimation of final similarity. The second run of Aggarwal et al. [1] performs better than the first one where the former run has expanded date window which helps. Surprisingly, the third run of Aggarwal et al. [1] which translates all target news stories as the preprocessing step performs worst compared to both the runs which employ CL-ESA based similarity estimation. Moreover, Aggarwal et al. [1] used Wikipedia reference collection for CL-ESA which in our opinion does not closely cover the vocabulary of the the news stories and may be the reason for underperformance compared to Palkovskii [12]. The keyphrase extraction based approach employed in Reddy and Perumal [15] does not work as well as it did in the previous edition of the CL!TR track [3]. We received only one run for en-gu partition resulting a very small pool for annotation. Compared to Hindi, Gujarati has significantly fewer resources available and consequently machine translation, which is relied on as a pre-processing step, performs badly for en-gu.

Run	NDCG@1	NDCG@5	NDCG@10	NDCG@20
yurii (1)	0.0541	0.0843	0.0955	0.0981

Table 3. Overall evaluation results for English-Gujarati partition.

In the relevance judgment we found that for some target news stories did not have any relevant source news stories. We found 12 such topics. The frequency of relevant documents for each target news story is plotted in Fig. 5 for en-hi partition and evaluation results with those target news stories which have at least one relevant source news story is depicted in Fig. 6.

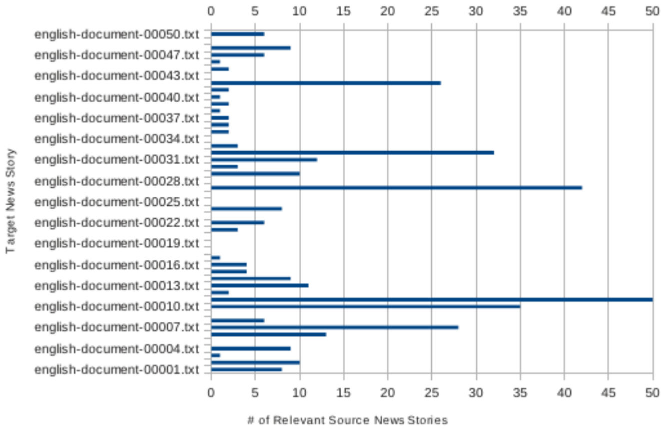

Fig. 5. The frequency of relevant source documents for each target document in the corpus.

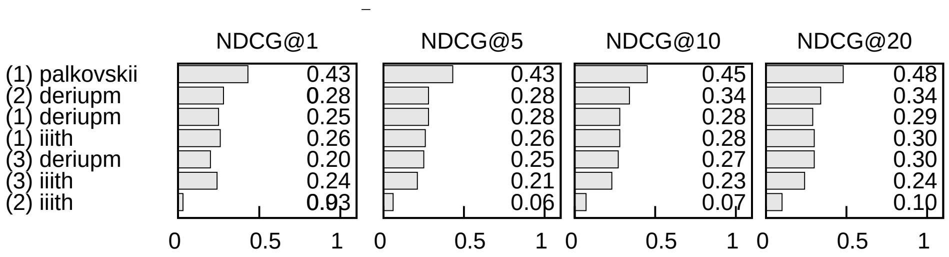

Fig. 6. Evaluation results for target news stories which have atleast one relevant source news story. The left hand side information corresponds to (run number) and team. The ranking is upon the NDCG@10 values.

We analysed the best runs from each team and the target news stories for which they obtained perfect NDCG@10 score. This statistic is shown in Table 4 along with their title and number of relevant documents.

Interestingly each tem performed well on a very different set of news stories. The majority of the target news stories for which Palkovskii [12] achieved the highest performance were "one-off" with a news event which were short lived in press coverage e.g. *Ahmedabad metro declaration, IAG MiG-21 crash, slapgate*⋆ *etc* while those for which Reddy and Perumal [15] performed best were running

⋆ The "slapgate" event drew extreme press attention but it happened in 2008 while this target news story is about a minor reaction in an interview with the cricketer

target	Title	# of rel.
	palkovskii-1	
08	*Madani's role in Bangalore stadium blasts not proved: Police*	6
12	*Ahmedabad to get Metro Rail link*	2
15	*Is Jaswant Singh on his way back into BJP fold?*	5
35	*IAF MiG-21 crashes in West Bengal, pilot safe*	2
36	*Muslim scholars tell Osama he's got it all wrong*	2
37	*No sex in promo Madhur!*	2
39	*'Slapgate' thing of past for Sreesanth, Harbhajan*	2
	deriupm-2	
30	*Lessons to be learnt from Bihar, says Bill Gates*	3
36	*Muslim scholars tell Osama he's got it all wrong*	2
37	*No sex in promo Madhur!*	2
	iiith-1	
04	*9 Indians among 17 dead as Taliban bombers attack Kabul*	9
08	*Madani's role in Bangalore stadium blasts not proved: Police*	6
22	*122 dead in massive quake in Chile; tsunami threatens Pacific*	6
24	*Pak blocks 800 URLs over Facebook cartoon row*	8
44	*HRT team ready to join F1 hall of fame*	2

Table 4. Target news stories for which the NDCG@10 is 1.0 for the best run of each team.

stories with a news event which received wide press coverage e.g. *facebook cartoon row, earthquake in Chile, bomb attack in Kabul etc.*

5 Remarks and Future Work

In this paper we presented an overview of the first edition of the cross-language indian news story search (CL!NSS) track at FIRE where the task was focused on linking news stories across the languages which share same focal event and/or news event. We presented the evaluation of the runs submitted by the participating teams along with analysis. The participating teams opted for highly different strategies and hence their set of best performing target news stories were almost exclusive. Participation for the English-Hindi pair was much higher than for the English-Gujarati pair. Overall participation was lower than the previous year's task and we believe this was in part due to the fact that training data was not provided. We plan to provide this in future versions of this task.

about its mention in 2010 hence there are extremely less relevant source stories with source collection of year 2010.

Our next editions will involve the higher challenges and larger collections for news story linking task and as discussed in Section 2, we will introduce the parallel fragment task.

6 Acknowledgment

We would like to thank Khushboo Singhal, Nitin Ramrakhiyani, Parth Mehta and Shreyansh Bhatt who helped in the annotation. We are highly indebted to European Science Foundation (ESF) 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 MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (plan I+D+i).

References

1. Aggarwal, N., Asooja, K., Buitelaar, P., Polajnar, T., Gracia, J.: Cross-Lingual Linking of News Stories using ESA - Working note for CL!NSS. In: FIRE [7]
2. 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)
3. Barrón-Cedeño, A., Rosso, P., Devi, S.L., Clough, P., Stevenson, M.: PAN@FIRE: Overview of the Cross-Language !ndian Text Re-Use Detection Competition. In: FIRE (ed.) FIRE 2011 Working Notes. Third International Workshop of the Forum for Information Retrieval Evaluation (2011)
4. 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)
5. Clough, P.: Measuring text reuse in a journalistic domain. In: In Proc. of the 4th CLUK Colloquium. pp. 53–63 (2001)
6. 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)
7. FIRE (ed.): FIRE 2012 Working Notes. Fourth International Workshop of the Forum for Information Retrieval Evaluation (2012)
8. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20(4), 422–446 (Oct 2002)
9. 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)
10. Munteanu, D.S., Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora. Comput. Linguist. 31(4), 477–504 (Dec 2005)
11. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. 29(1), 19–51 (Mar 2003)

12. Palkovskii, Y.: Working note for CL!NSS. In: FIRE [7]
13. 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)
14. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-based Multilingual Retrieval Model. In: Proceedings of the IR research, 30th European conference on Advances in information retrieval. pp. 522–530. ECIR'08, Springer-Verlag, Berlin, Heidelberg (2008)
15. Reddy, R., Perumal, K.: Working note for CL!NSS. In: FIRE [7]
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. Udupa, R., Khapra, M.M.: Improving the multilingual user experience of wikipedia using cross-language name search. In: HLT-NAACL. pp. 492–500 (2010)
18. 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)