Cross-Lingual Linking of News Stories using ESA

Nitish Aggarwal[1], Kartik Asooja[2], Paul Buitelaar[1], Tamara Polajanar[1], and Jorge Gracia[2]

[1]UNLP, DERI, NUIG, Galway, Ireland
[2]OEG, UPM, Madrid, Spain
nitish.aggarwal@deri.org asooja@gmail.com paul.buitelaar@deri.org
tamara.polajanar@deri.org jgracia@fi.upm.es

Abstract. In this paper, we describe our approach for Cross-Lingual linking of Indian news stories, submitted for Cross-Lingual Indian News Story Search (CL!NSS) task at FIRE 2012. Our approach consists of two major steps, the reduction of search space by using different features and ranking of the news stories according to their relatedness scores. Our approach uses Wikipedia-based Cross-Lingual Explicit Semantic Analysis (CLESA) to calculate the semantic similarity and relatedness score between two news stories in different languages. We evaluate our approach on CL!NSS dataset, which consists of 50 news stories in English and 50K news stories in Hindi.

Keywords: Cross-lingual search, Explicit semantic analysis, News linking, Plagiarism detection

1 Introduction

Cross-lingual news story linking aims to identify the same news articles in different languages. It can be important for multilingual applications and information retrieval to connect the available news stories about the same event but in different languages. It may be quite useful for multilingual users also as it allows them to refer the same story from different perspectives and languages.

Cross-language India news story search (CL!NNS) is the task which focusses on the automatic alignment of news stories in a quasi-comparable corpus. We are given a source set containing Hindi stories and a target set of relatively few English stories. The task is to link each English news story from the target set to its corresponding version in the source set. We are asked to provide a list of 100 most similar news stories in Hindi for every English story.

Most of the cross-lingual technologies rely on machine translation to transform the problem into a mono-lingual scenario. In this paper, we describe our approach based on Cross Lingual Explicit Semantic Analysis (CL-ESA) for identifying the

most similar news stories available in a different language. CL-ESA relies on a comparable corpora consisting of concepts with their definitions created explicitly by the humans. In the next sections, we discuss our approaches submitted for the task and the results.

2 Approach

Our approach consists of two major steps; the reduction of search space by using different features and ranking of the news stories based on relatedness scores obtained by CL-ESA.

2.1 Search Space Reduction

We reduced the search space by using two different features, publication dates of the news stories and vocabulary overlap. As all news stories have publication dates, we reduced the search space to those news stories for which publication dates are within K days of the publication date of target news story. The vocabulary overlap were calculated by translating all of the target news stories into the language of the source news stories followed by taking the tf-idf score obtained by Lucene.

2.2 CL-ESA Ranking

Explicit Semantic Analysis (ESA) attempts to represent the semantics of the given term in the high distributional semantic space [2]. These semantics are obtained by use of a high dimensional vector, where each dimension may reflect unique Wikipedia concept. This high dimensional vector is created by taking the tf-idf weight of a given term in the corresponding Wikipedia articles. Semantic relatedness of two given terms can be obtained by calculating the correlation between two high dimensional vectors generated by ESA.

As Wikipedia consists of the articles in different language and their links across the language, it allows a cross-lingual extension of ESA to calculate the semantic relatedness for texts in different languages [5]. Thus, using CL-ESA, we can compare the semantic relatedness score between the news stories written in different languages. We compare each target English story against the source Hindi stories to select the top Hindi stories according to the CL-ESA ranking.

3 Experiment

We submitted three different runs in the CL!NNS task based on the approaches defined in Section 2. The given data includes 50 English news stories and 50000

Hindi news stories. We need to link each English news story to the 100 top Hindi news stories, which has the same news, events or any related topic, with some score defining the relatedness.
Following are the descriptions of the three runs:

- *Run 1:* The data consisted of the dates on which the news articles were published. In this run, we reduced the total number of source Hindi documents to be compared to the target English story. We considered only those which appear in a window of 4 days (2 days before and 2 days after) from the published date of the English story. Then, using CL-ESA ranking as defined in Section 2, the top 100 source Hindi stories were selected for the run.

- *Run 2:* Here, we reduced the search space by selecting only those source Hindi news stories which appear in a window of 14 days (7 days before and 7 days after) from the date of the English story. Also, in this run, we used a modified version of CL-ESA as defined in [1], which searches the whole text together to retrieve the top Wikipedia concepts. It is different from the bag of words approach and adding the vector for each word in the bag to form the vector for the whole text or bag. Searching the whole text together in Lucene also considers the distances between words to rank the documents.

- *Run 3:* All the source Hindi stories were indexed and the target English stories were translated into Hindi using Google translator. Every translated document was searched in the index and top 1000 documents were taken according to the Lucene ranking. Then, from these top 1000 documents, 100 best were selected according to the CL-ESA scores between the original non-translated English document and the 1000 retrieved documents.

4 Evaluation

To evaluate the submissions, the CL!NNS organizers created a manually annotated pool of Hindi news stories against each English story [4]. Each Hindi story in the pool was marked up by one of the following labels:

1. 0 different news event
2. 1 same news event but different focal event
3. 2 same news event and same focal event

All of the submitted runs were against this annotated dataset. NDCG@k [3] was used with different values of k for the evaluation of the linking of the news stories. Table 1 shows the results of various submissions.

5 Conclusion

We presented our approaches for linking news stories published in different languages. Cross lingual explicit semantic analysis was utilized for comparing the

Rank	Run	NDCG@1	NDCG@5	NDCG@10
1	run-1-english-hindi-palkovskii	0.3229	0.3259	0.3380
2	run-2-english-hindi-deriupm	0.2100	0.2136	0.2613
3	run-1-english-hindi-deriupm	0.1900	0.2110	0.2168
4	run-1-english-hindi-iiith	0.1939	0.1994	0.2154
5	run-3-english-hindi-deriupm	0.1500	0.1886	0.2030
6	run-3-english-hindi-iiith	0.1837	0.1557	0.1722
7	run-2-english-hindi-iiith	0.0204	0.0462	0.0512

Fig. 1. Result Table

news stories and selecting the relevant ones. Our second run performed better than the other two runs, probably because a wider window of 14 days was selected and distances between the words were also considered in the CL-ESA model, which made a better ranking of the candidates.

References

1. Aggarwal, N., Asooja, K., Buitelaar, P.: DERI&UPM: Pushing corpus based relatedness to similarity: Shared task system description. In: SemEval-2012, SEM, First Joint Conference on Lexical and Computational Semantics, and co-located with NAACL, Montreal, Canada (6 2012)
2. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: In Proceedings of the 20th International Joint Conference on Artificial Intelligence. pp. 1606–1611 (2007)
3. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20(4), 422–446 (Oct 2002), http://doi.acm.org/10.1145/582415.582418
4. Parth Gupta, Paul Clough, P.R., Stevenson, M.: Pan@fire: Overview of the cross-language !ndian news story search (cl!nss) track. In: Forum for Information Retrieval Evaluation, ISI, Kolkata, India (2012)
5. Sorg, P., Braun, M., Nicolay, D., Cimiano, P.: Cross-lingual information retrieval based on multiple indexes. In: Working Notes for the CLEF 2009 Workshop. Cross-lingual Evaluation Forum, Corfu, Greece (September 2009)

Acknowledgements

This work is supported in part by the European Union under Grant No. 248458 for the Monnet project and by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).