Leveraging Article Titles for Cross-lingual Linking of Focal News Events

Goutham Tholpadi and Amogh Param

Computer Science and Automation,
Indian Institute of Science, Bangalore, India
gtholpadi@csa.iisc.ernet.in,amoghparam@gmail.com
http://www.csa.iisc.ernet.in

Abstract. In this paper, we discuss the methods and results from our participation in the Cross-Lingual Indian News Story Search (CL!NSS) track at Forum for Information Retrieval Evaluation (2013). We describe a method that leverages the structure of news articles, especially the title, to achieve good performance on the focal news event linking task. We achieved the best performance among all teams in the NDCG@1 task, and were ranked second and third, respectively, in the NDCG@5 and NDCG@10 tasks. Contrary to popular belief, we find that imposing date constraints did not improve precision.

Keywords: news linking, cross-lingual search

1 Introduction

Linking news stories in different languages on the same news event has several applications. A multilingual reader can compare different reported versions of the same event. News stories in different languages covering the same event contain fragments that are translations or paraphrases. These can be used to learn dictionaries and train machine translation systems [2].

The Cross-Language Indian News Story Search (CL!NSS) track at Forum for Information Retrieval Evaluation (FIRE) 2013 focuses on this task. This track involves news data sets in English and Hindi, and is continuation of last year's track where a similar task was defined (see Section 2 for detailed task definition).

In this paper, we describe a method that achieved the best performance in the NDCG@1 task, and was ranked second and third, respectively, in the NDCG@5 and NDCG@10 tasks. Our method makes explicit use of the structure of news articles, specifically the *title* to achieve high precision. Our method requires machine translation from the target to the source language. We use a small training set for tuning our algorithm, but it is not a required component.

The structure of the paper is as follows. We define the task in Section 2 and describe our approach in Section 3. Section 4 details the experimental details and the main results. We analyse the results in depth in Section 5, and conclude.

2 Definitions

In the following, we define the objects involved in the task, and the task itself.

Article: the basic unit of news reporting that describes an event(s). It consists of three parts:

1. *Content*: a piece of text that actually describes the event(s) of the article.
2. *Title*: a short piece of text that indicates the events described in the content.
3. *Date*: the date of publication of the article.

Focal event: the main event(s) that provide focus for the article.
Background event: event(s) that plays a supporting role in the article, providing context for the focal event, e.g. related events leading up to the focal event, similar events in the past, definitions/explanations/descriptions of things/people/places which play a role in the focal event
News event : a group of related focal events, that is related to the concept of a "real-world event", e.g. presidential elections.

Task.
Given: a source collection S of articles in Hindi, a target collection T of articles in English.
Task objective: For each target article $t \in T$,

- identify articles $s \in S$ that contain the same focal event as t.
- identify articles $s \in S$ such that the focal events of s and t belong to the same news event.

Task definition: For each target article $t \in T$, rank the articles in S and return the top 100 articles. The ranking should be such that the same focal event articles are ranked highest, followed by the news event articles, followed by other articles.

Task evaluation: For each $t \in T$, a human-annotated gold standard set of source articles is available. Each article in the set is assigned a score 2 if it has the same focal event, and 1 if it has the same news event. With this gold standard, the ranked list of articles is scored using NDCG@k, for $k = 1, 5, 10$.

3 Method Description

Our method is motivated by three assumptions:

1. The title indicates the focal event in an article.
2. The content indicates both the focal event and the news event in an article.
3. Source articles containing a target article t's focal event are published at around the same time as t.

The assumptions above are reasonable in the context of news articles, as observed in earlier work [1, 2].

3.1 Method

For each target English article in T, we used an online machine translation service[1] to get its Hindi translation $Q = \{q^T, q^C\}$ where q^T is the set of terms in the title and q^C is the set of terms in the content. Similarly, each Hindi article in S is defined as $D = \{d^T, d^C\}$. We define a similarity scoring function $Sim(Q, D)$ for the target article Q and the source article D as follows.

$$Sim(Q, D) = \alpha^{TT} Sim(q^T, d^T) + \alpha^{TC} Sim(q^T, d^C) + \alpha^{CC} Sim(q^C, d^C)$$

The terms $Sim(q^T, d^T)$ and $Sim(q^T, d^C)$ try to capture the likelihood that the focal event in Q is present in D (using Assumption 1). The term $Sim(q^C, d^C)$ tries to capture the likelihood that the news event in Q is present in D (using Assumption 2).[2] In addition, we considered only those news stories D which were published within a window of α^D days around the date of Q (using Assumption 3). The parameters α are used to tune the algorithm for a particular data set.

In every term above, $Sim(q, d)$ is a variant of the *tf.idf* similarity between documents, weighted by the fraction of query terms $t \in q$ that are present in d. It is defined as follows:

$$\begin{aligned} Sim(q, d) &= \omega(q, d) \sum_{w \in q} (IDF(w))^2 \; TF(w, d) \quad \text{where} \\ \omega(q, d) &= \frac{|q \cap d|}{|q|} \\ TF(w, d) &= \frac{\sqrt{\#\text{occurrences of } w \text{ in } d}}{\sqrt{\#\text{terms in } d}} \\ IDF(w) &= 1 + \log\left(\frac{\#\text{docs in corpus}}{1 + \#\text{docs containing } w}\right) \end{aligned}$$

4 Experiments and Results

Implementation Details.
The CL!NSS data set consists of 50 English articles with relevance judgments, 25 English articles that the teams would be evaluate on, and more than 50,000 Hindi articles. We used the open-source information retrieval library Apache Lucene 4.4[3] in our experiments.

Preprocessing. The text of each article was tokenized according to the Unicode Text Segmentation algorithm[4] (as implemented in Lucene), and dots were removed from acronyms. Tokens consisting of Latin characters were lowercased, and the trailing ''s' (apostrophe followed by 's') was removed if present. Finally, Hindi stop words were removed.

[1] Microsoft Translator API: www.microsoft.com/en-us/translator/developers.aspx

[2] We do not consider $Sim(q^C, d^T)$ since that will boost source articles with other focal events, which happen to be background events in Q.

[3] lucene.apache.org

[4] www.unicode.org/reports/tr29/

Parameters. We did a grid search[5] to arrive at the best values for the parameters α, and the three best-performing parameter configurations on the were used in the runs submitted.

Results

The parameter values used for the three runs, and the results are shown in Table 1. Run 3 had the best NDCG@1, the second-best NDCG@5 and the third-best NDCG@10 among all participating teams.

Table 1. Performance of the method for different runs.

Run	α^{TT}	α^{TC}	α^{CC}	α^{D}	NDCG@1	NDCG@5	NDCG@10
1	0	1	1	7	0.5200	0.4217	0.4084
2	0	3	1	7	0.5400	0.4304	0.4110
3	0	3	1	∞	**0.7800**	0.6783	0.6804
Best result by any team					0.7800	**0.6809**	**0.7268**

5 Analysis

Assumption 1 works, but not always. According to Assumption 1, title-title and title-content similarity are strong indicators of common focal events. The results seem to suggest that the former is false, but the latter is true. We feel the failure of the assumption in the first case is due to severe sparsity—titles have few words, and hence the overlap between titles is very small (usually nil). Hence the $Sim(q^T, d^T)$ term is zero in most cases. Expanding the word set with synonyms might help, but at the cost of precision.

Assumption 3 applies to focal events only. While an infinite date window ($\alpha^D = \infty$) did best in the results, the improvement over the other runs was mainly in terms of recall. Imposing a narrower date window did not hurt focal event identification, but caused the loss of articles on the same news event that were outside the date window. Keeping this in view, it might be fruitful to decouple the two tasks ("same focal event", and "same news event") in the evaluation, in order to accurately measure the impact of different algorithmic decisions on each task.

Special handling of entities is crucial.

[5] The queries and relevance judgments from the 2012 CL!NSS track was used as the development set.

Normalization. The (translated) target article *english-document-00011.txt* contained the word “cell” in the title, while most the documents in the gold standard contained “cell-phone”, which caused them to be ranked lower. Normalization of different surface forms of the same real-world object might be helpful, especially for words that are key entities in the event.

Named entities. It is easy to see that named entities (NEs) are critical for event representation. For the target article *english-document-00005.txt*, the translation API could not translate the NE “Dantewadas” in the title. In later experiments, we found that manually adding this single word translation to the title improved the NDCG@10 from 0.7814 to 0.9366 and the NDCG@5 from 0.8596 to 1.0. This underscores the importance of NEs, and suggests that solutions tailored specifically for identifying and translating them might be worth the effort.

Some comments on the data set.

Articles about long-range events. The gold standard for many of the target articles had source articles from a wide date range. This caused the time-agnostic configuration to perform well. But it is unclear whether this is a characteristic of the articles chosen for the track, or a tendency of the annotators to be lenient when judging “news event” commonality.

Opinion pieces and “celebrities”. Some of the target articles were opinion pieces (e.g. *english-document-00016.txt*) which analyse several events, but have no focal event. Some articles are about personalities who are constantly in the news (e.g. *english-document-00024.txt*) who are a part of numerous events. Our method did not do well on such articles. These illustrate that shallow analysis of article can only take us so far. A richer modeling of news events in terms of actors, actions, locations, and time is needed to achieve a more nuanced distinction between articles.

6 Conclusion

In this paper, we describe a method to link news articles across languages talking about the same focal news event. We leverage the close relationship between the title and the focal event of the article to achieve high precision. We analyze the results and failure cases and identify entity handling as a crucial area for improvement.

References

1. Aker, A., Kanoulas, E., Gaizauskas, R.J.: A light way to collect comparable corpora from the web. In: LREC. pp. 15–20 (2012)
2. Barker, E., Gaizauskas, R.J.: Assessing the comparability of news texts. In: LREC. pp. 3996–4003 (2012)