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

Goutham Tholpadi
Amogh Param

About Me

- BMS College of Engineering, Bangalore - June, 2013
- Major - Computer Science and Engineering
- Interests - NLP – IR, Cross-lingual NLP-IR,
  Machine Learning, Computer Vision.
- Project Assistant at Indian Institute of Science (IISc).
- I've worked for a company that makes virtual farming really fun, Zynga Inc.
- FIRST conference and talk among PhDs.

Introduction

- Method and analysis achieved
  - Best Performance in NDCG@1
  - 2nd and 3rd in NDCG@5 and NDCG@10

- Structure of News Articles
- Requires machine translation from target to source language
- Used a training set from last year's queries to tune our method.

Task

- Given
  - Source Collection S of articles in Hindi
  - Target Collection T of articles in English

- Task Objective
  - For each target article t in target collection T Identify
    - $s \in S$ that contain same focal event as t
    - $s \in S$ such that focal events of s and t belong to same news event.

Task

- Task Definition
  - For each $t \in T$
    - Rank articles in S and return top 100 articles.
    - Ranking:
      - Same focal event articles - ranked highest,
      - Followed by 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.
    - Score 2 if it has the same focal event.
    - Score 1 if it has the same news event.
  - Ranked list of articles is scored using NDCG@k, for k = 1, 5, 10.

Method Description

- Structure of articles
  - Title
  - Date
  - Content

- Motivated by three assumptions
  - Title indicates Focal Event in article
  - Content indicates both Focal Event and News Event in article.
  - Source articles containing target article t's focal event published at around same time as t.

(Based on earlier work, this works reasonably in context of news articles)

Method

- For each English article in query collection T
  - Machine Translation.
    - English query → Hindi query

- Translated Query
  - $Q = \{q^T , q^C \}$
    - q^T - Set of terms in title
    - q^C - Set of terms in content

- Similarly, a Hindi article in source collection S
  - $D = \{d^T , d^C \}$

Similarity Scoring Function

- Sim(Q , D)
  - Q – target article/query
  - D – source article

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

  - Sim(q^T ,d^T) and Sim(q^T ,d^C) capture:
    - Likelihood that focal event in Q is present in D (Assumption 1)
  - Sim(q^C,d^C) captures:
    - Likelihood that news event in Q is present in D (Assumption 2)
  - News stories D published within window of α^D days around the date of Q (Assumption 3)

- Parameters α used to tune algorithm for particular data set

Similarity Scoring Function

- In every term of the equation
  - Sim(q,d) variant of the TF-IDF similarity between documents weighted by fraction of query terms t ∈ q that are present in d.

- Definition

$$Sim(q,d) = \omega(q,d) \sum_{w \in q} (IDF(w))^2 \; TF(w,d) \quad \text{where}$$

$$\omega(q,d) = \frac{|q \cap d|}{|q|}$$

$$TF(w,d) = \frac{\sqrt{\#\text{occurrences of } w \text{ in } d}}{\sqrt{\#\text{terms in } d}}$$

$$IDF(w) = 1 + \log\left(\frac{\#\text{docs in corpus}}{1 + \#\text{docs containing } w}\right)$$

Implementation

- Apache Lucene – Open Source IR library

- Preprocessing
  - Tokenized – Unicode Text Segmentation algorithm (as implemented in Lucene)
  - Dots removed from acronyms
  - Tokens with Latin characters – Lowercased
  - Trailing ‘s (apostrophe followed by s) – removed, if present.
  - Hindi stop words removed

Implementation

- Configurations
  - Grid Search to arrive at parameters (α) resulting in best performance.
  - The queries and relevance judgments from the 2012 CLINSS track was used as the development set.

- Results

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**

Analysis

- Assumption 1 works, but not always.
  - Assumption 1 : Title-title and title-content similarity are strong indicators of common focal events.
  - Title – Title similarity doesn't hold true mostly due to sparsity.
    - Titles have few words, and hence the overlap between titles is very small (usually nil).
  - $Sim(q^T , d^T)$ term = 0 (in most cases).
  - Expanding the word set with synonyms might help, but at the cost of precision.

Analysis

- Assumption 3 applies to focal events only.
  - Infinite date window ($\alpha D = \infty$) - best in the results
  - Imposing a narrower date window
    - Did not hurt focal event identification
    - Cause loss of articles on the same news event that were outside the date window.
  - Therefore, 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.

Analysis

- Special Handling of entities is crucial
  - Normalization.
    - The (translated) target article english-document-00011.txt contained the word “cell” in the title (diagram)
    - 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.

Analysis

- Named entities.
  - Very crucial
  - Translation API could not translate the "Dantewadas" in target article english-document-00005.txt
  - In later experiments, we found that manually adding this single word translation to the title improved the NDCG@10 by 19% and the NDCG@5 16%.
  - This emphasizes the importance of named entities, and suggests that solutions tailored specifically for identifying and translating them might be worth the effort.

Comments

- Articles about long-range events.
  - The gold standard for many of the target articles had source articles from a wide date range.
  - Therefore, time-agnostic configuration performed 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.

Comments

- Opinion pieces and “celebrities”.
  - Some of the target articles were opinion pieces
    - (e.g. english-document-00016.txt) which analyze several events, but have no focal event.
    - Articles about personalities constantly in news (e.g. english-document-00024.txt) who are a part of numerous events.
  - Our method did not do well on such articles.
- Therefore, a richer modeling of news events in terms of actors, actions, locations, and time is needed to achieve a more nuanced distinction between articles.

Conclusion

- Method to link news articles across languages talking about the same focal news event.
- Leverage the close relationship between the title and the focal event of the article to achieve high precision.
- Analyze the results and failure cases and identify entity handling as a crucial area for improvement.

Acknowledgement

- FIRE organizers.
- Parth Gupta
- Goutham Tholpadi, PhD at IISc.
- Everyone here.

THANK YOU