Microblog Retrieval in a Disaster Situation: A New Test Collection for Evaluation

Moumita Basu[1,4]
Anurag Roy[1]
Kripabandhu Ghosh[2]
Somprakash Bandyopadhyay[4]
Saptarshi Ghosh[1,3]

[1]**Department of CST,**
Indian Institute of Engineering Science and Technology Shibpur, India
[2]**Department of CSE,**
Indian Institute of Technology Kanpur, India
[3]**Department of CSE,**
Indian Institute of Technology Kharagpur, India
[4]**Social Informatics Research Group,**
Indian Institute of Management, Calcutta,India

Outline

Role of Microblogs during Disasters

- Lot of useful *situational information* posted on microblogging sites like Twitter during disaster events
- Challenges in extracting the important information
  - Important information obscured amongst lot of sentiment, opinion, ...
  - Microblogs are very short and written informally
  - Large variation in vocabulary of crowdsourced content
- No standard test collection for evaluating strategies for microblog retrieval in disaster scenario

Outline

The Microblog dataset

- Collected tweets posted during Nepal earthquake in April 2015
- Used Twitter Search API with keyword 'nepal'
- About 100K tweets in English collected
- After de-duplication final dataset of 50,068 tweets

Topics for retrieval

- Consulted members of NGOs work in disaster-affected regions – what are typical information requirements during disaster relief operation?
- Identified *five* broad information requirements (*topics*)
  - T1: What resources were available
  - T2: What resources were required
  - T3: What *medical* resources were available
  - T4: What *medical* resources were required
  - T5: What infrastructure damage and restoration were being reported

Developing gold standard for the retrieval

- Two phases, involving human annotation
- Phase 1
  - Each annotator given the microblog collection and topics, asked to identify all tweets relevant to each topic, *independently*
  - Tweets indexed using the Indri IR system
- After Phase 1, the set of tweets identified to be relevant to the same topic by different annotators, was considerably different
- Hence, Phase 2
  - For a topic, all tweets judged relevant by *at least one* annotator considered
  - Relevance finalised through discussion among all the annotators and mutual agreement

Number of tweets in final gold standard

- T1: What resources were available (589 tweets)
- T2: What resources were required (301 tweets)
- T3: What medical resources were available (334 tweets)
- T4: What medical resources were required (112 tweets)
- T5: What infrastructure damage and restoration were being reported (254 tweets)

Examples of relevant tweets

- T1: What resources were available
  - India sends 39 #NDRF team, 2 dogs and 3 tonnes equipment to Nepal Army for rescue operations: Indian Embassy in #Nepal
  - If O+ve Blood is needed around Ilam, I am ready just mention. #NepalQuake
  - Dr. Madhur Basnet leading medical team going to remote villages of Gorkha dist which was epicenter of earthquake. His cell: [number]
- T2: What resources were required
  - Body bags, Tents, water, medicine, pain killers urgently needed in #earthquake stricken #Nepal
  - plz send medicine and food packets to nepal if possible. #NepalEarthquake
  - There is shortage of Blood as well as oxygen cylinders...Nepal is in huge crisis.

Outline

Query Generation from Topic

- Manual Query: Selecting intuitively important terms from topic
- Automatic Query: By pre-processing and POS tagging, nouns, verbs, and adjectives extracted automatically from narrative part of the topics

Pre-processing the Dataset

- Removing standard English stop-words, URLs and punctuation symbols
- Discarding terms having frequency less than 5 in corpus.
- Case-folded and stemmed using Porter stemmer.

Microblog Retieval and Ranking

- Retrieving microblogs using language modeling: Indexed and ranked using Indri
- Retrieving microblogs using word embeddings: Ranked and retrieved using Word2vec

Word2Vec Retrieval Method

- Trained Word2Vec over pre-processed set of tweets
- Continuous bag of words model, Vector size: 2000, Context size: 5, Learning rate: 0.05
- Query-vector: sum of term-vectors of all terms in the query
- Tweet-vector: sum of term-vectors of all terms in the tweet
- Ranked tweets in decreasing order of cosine similarity between corresponding query-vector and each tweet-vector
- Top ranked 1000 tweets are retrieved

Evaluation

Query type	Ranking Model	Prec @20	Recall @1000	MAP @1000	MAP Overall
Manual	Indri	0.3900	0.5635	0.1285	0.1639
Manual	word2vec	**0.6700**	**0.6197**	**0.2343**	**0.2788**
Automatic	Indri	0.3000	0.4357	0.0891	0.1149
Automatic	word2vec	**0.4600**	**0.5591**	**0.1785**	**0.2242**

Table: **Retrieval performance with initial queries (manual and automatic) and two ranking models – language model of Indri and word2vec-based model. The values for the performance measures have been averaged over the five topics. The performances by word2vec are statistically significantly better ($p < 0.05$) than that of Indri for both types of queries.**

Query Expansion

Pseudo (or blind) relevance feed-back: 10 top-ranked tweets retrieved by the original query, and select p = 5 terms from these 10 top-ranked tweets to expand the query

- Rocchio Expansion: tf x idf Rocchio scores for each distinct term is considered
- Expansion using Word2vec:
  - Identifying set of terms (within 10 top-ranked tweets) most related to context of query
  - Checked cosine similarity between term-vector and Query-vector

Evaluation

Query type	Ranking Model	Query Expansion	Prec @20	Recall @1000	MAP @1000	MAP Overall
Manual	Indri	No expansion	0.3900	0.5635	0.1285	0.1639
Manual	Indri	Rocchio	0.3500	0.5532	0.1233	0.1598
Manual	Indri	word2vec	0.3900	0.5518	0.1193	0.1591
Manual	word2vec	No expansion	0.6700	0.6197	0.2343	0.2788
Manual	word2vec	Rocchio	**0.6500**	**0.6281**	**0.2441**	**0.2873**
Manual	word2vec	word2vec	0.6400	0.6080	0.2242	0.2689

Table: **Retrieval performance (averaged over the five topics) for queries expanded using two strategies – Rocchio and Word2Vec-based. For comparison, the performances with the initial queries as was reported in are also shown**

Outline

Future directions

- Lot of improvement in microblog retrieval still necessary
- More standard test collections needed
- Used the collection in FIRE 2016 Microblog Track
- Two more topics included
  - T6: What were the requirements & availabilities at *specific locations*
  - T7: What were the activities of various NGOs / Government organizations
- Improved gold standard by Phase 3 – standard pooling
  - Top 30 results of all the submitted runs pooled and judged by annotators
  - Majority opinion considered for the rest

Acknowledgements

- Anonymous reviewers whose suggestions helped to improve the paper.
- This research was partially supported by a grant from the Information Technology Research Academy (ITRA), DeITY, Government of India(Ref. No.: ITRA/15 (58)/Mobile/DISARM/05).


Thank
You!