CIRGIRDISCO at TREC 2013 Microblog Track

M. Atif Qureshi[1,2], Colm O'Riordan[1], and Gabriella Pasi[2]

[1] Computational Intelligence Research Group, National University of Ireland Galway, Ireland
[2] Information Retrieval Lab,Informatics, Systems and Communication, University of Milan Bicocca, Milan, Italy
muhammad.qureshi@nuigalway.ie,colm.oriordan@nuigalway.ie, pasi@disco.unimib.it

Abstract. This paper describes our participation in the TREC 2013 Microblog real-time search task. We utilized a query expansion approach and submitted three runs: one without using any form of external evidence and the remaining two runs utilize extended abstracts of Wikipedia articles.

1 Introduction and Task Description

Microblogging platforms have emerged as significant information resources during real-time events with many users turning to the real-time search facility offered by microblogging platforms to learn up-to-date information pertaining to various events and topics [3]. However, it is a significant challenge to perform information retrieval over microblog posts mainly due to the short and informal nature of their contents. The TREC microblog track is an attempt to improve state-of-the-art in microblog retrieval systems. This paper describes our approach during participation in TREC 2013 Microblog track.

The TREC microblog track which has been running since 2011 focuses on Twitter as the microblogging platform of choice; this year is the third time the TREC conference released a task in the microblog track. The task's challenge consists in finding and ranking relevant tweets given a query topic in an attempt to simulate user behaviours wishing to retrieve some information on Twitter using few keywords. The basic goal is to retrieve the most relevant tweet pertinent to the given keyword and discard the irrelevant ones. The corpus for the TREC 2013 Microblog track was distributed using track-as-a-service model where all participants were able to access the tweets through a search API that the organizers offered. The Tweets2013 collection consists of approximately 240 million tweets covering a two-month period from 1st February, 2013 to 31st March, 2013.

2 Methodology

This section describes our approach in dealing with the real-time adhoc search task in microblogs. We developed a baseline system that does not rely on any external sources of evidence. Furthermore, we describe a system that uses extended abstracts of Wikipedia articles.

2.1 Caching and Pre-Processing

Both our approaches utilize the initial retrieval system provided by TREC 2013 microblog track organizers to obtain an initial set of 10,000 tweets across each query. The search API provided by the TREC 2013 microblog organizers was implemented with a query likelihood language model with Dirichlet smoothing. The initially retrieved tweets are cached in order to minimize the load at the remote server provided by TREC 2013 microblog organizers. Pre-processing involved removal of stopwords from the tweets. We then build a query expansion module that retrieves 1,000 tweets[3] from the cached 10,000 tweets.

2.2 Baseline System

The baseline system utilizes a biased version of PageRank [2] applied on terms of the tweet. Across each query topic, the top 20 tweets are retrieved using the search API through a query likelihood language model with Dirichlet smoothing. We build an undirected term graph where each term when occurring in a single tweet is regarded as being adjacent to other terms in the same tweet. We then apply the PageRank algorithm over this undirected term graph, the terms that get the highest PageRank score are used for query expansion. Note that we select the top 15 terms with the highest PageRank score.

2.3 Wikipedia-Based System

After retrieval with the baseline system of section 2.2, we experiment with two versions of Wikipedia-based query expansion. We distinguish between the two versions in that one applies further query expansion for only those queries in which people's names occur[4] and the other applies for further query expansion for all queries[5]. We expand query terms in this step using phrases within the extended abstracts of phrases that match with Wikipedia article titles. The most frequent terms from the extended abstract are used for the query expansion step. Note that the data for Wikipedia articles' titles, and Wikipedia extended abstracts is obtained through a custom Wikipedia API that has pre-indexed Wikipedia data and hence, it is computationally fast[6]. The API is developed using the DBPedia [1] 2012 dumps.

3 Experimental Evaluations

In this section we present the official TREC 2013 evaluation for the microblog real-time search task. As mentioned in previous section, we submitted three runs and Table 1 shows official results released by the TREC organizers.

[3] The task required 1,000 tweets to be submitted for each topic.
[4] This run is referred to as CIRGIRDISCO2.
[5] This run is referred to as CIRGIRDISCO3.
[6] http://www3.it.nuigalway.ie/cirg/prj/WikiMadeEasy.html, we aim to release the API as an open source Wikipedia tool to facilitate other researchers.

Submitted Run	Measures		
	MAP	*RPrec*	*P*@30
Baseline-CIRGIRDISCO4	0.2220	0.2871	0.4078
CIRGIRDISCO2	0.2160	0.2796	0.3817
CIRGIRDISCO3	0.2152	0.2788	0.3828

Table 1: TREC Microblog 2013 Official Results for Runs by *CIRGIRDISCO*

The results show that the baseline system utilizing biased version of PageRank on tweet terms' graph outperforms the runs utilizing Wikipedia as an external source of evidence. As future work, we aim to investigate more sophisticated forms of our Wikipedia-based system.

References

1. C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semant.*, 7(3):154–165, Sept. 2009.
2. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web, 1999.
3. J. Teevan, D. Ramage, and M. R. Morris. #twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 35–44, New York, NY, USA, 2011. ACM.