University of Strathclyde at Headline Ranking TREC BLOG 2010

Dmitri Roussinov
Department of Computer and Information Sciences
University of Strathclyde
L13.29 LIVINGSTONE TOWER
16 Richmond Street, Glasgow, UK G1 1XQ
Dmitri.Roussinov@asu.edu

ABSTRACT

The University of Strathclyde participated in TREC BLOG Headline Ranking task only. Our general theme was to explore how the lexical changes in the BLOG corpus can reflect the importance of the articles appearing in the news corpus. Three (*3*) runs were submitted. For automated run "strath1", our algorithm identified the word unigrams, the frequencies of mentioning of which in the blog corpus increased substantially on the day of the query. Up to 100 such words were used as a query to return and rank the headlines using the Terrier platform and its PL2 model. Automated run "strath3" was similar to "strath1" except the weights were estimated based on the amount of the increase in the frequency of use and applied to the query words. "Strath2" was a manual run. Event descriptions were taken from the "Current Event" articles of Wikipedia Portal on the day of each topic (e.g. 22 January 2008 for the first topic). The description of each event was sent to Bing search engine. The words that occurred more frequently within the snippets than on the Web in average, were used to query the headlines corpus. Our participation was in close collaboration with the University of Glasgow group, which provided 1) the index to the news corpus 2) the daily statistics on the lexicons of the blog corpus and 3) the classification of the headlines into the required set of categories.

HEADLINE RANKING TASK

The top stories identification task (headline ranking) was first run as a pilot task in TREC 2009 to address the news dimension of the blogosphere (Macdonald et al., 2009). The task was exploring whether the blogosphere can be used to identify the most important news stories for a given day. In response to a date "query", systems were expected to provide a ranking of 100 headlines (stories) that they think were important on the specified day. For this year, Thomson-Reuters provided the TRC2 newswire corpus covering the same time-span as the corpus used in 2009. Not only the newer TRC2 corpus was much larger, it also included the full content of each story. The corpus was distributed by NIST for the TREC participants free of charge. Different from 2009, this task was treated as an online event detection: only the data created prior to the query date (e.g., an ontology or a Wikipedia article) was allowed to be used by an automatic run.

Two main approaches were used during 2009 TREC to identify top stories (Macdonald et al., 2009; Balog et al., 2006): **(i) News to Blogs,** where mentioning of the headline in the blogs was typically counted as a vote for its importance, and **(ii) Blogs to News**, which generally proceed by the following steps: 1. observe the blog posts from the given date; 2. detect what differentiates these posts from the previous posts; 3. identify the emerging topics; 4. rank the headlines by their similarity to the emerging topics. The overall observation was that "Blogs to News" approaches worked better. This motivated our involvement and specific choices of techniques to explore as we elaborate in the next section. "Related Work" section follows our "Discussion Of Results," followed, in turn, by "Conclusions, Limitations And Future Directions."

SYSTEM DESCRIPTION

AUTOMATIC "BLOGS TO NEWS" RUNS

The primary objective of our involvement in TREC BLOG this year was to explore the simplest possible approaches which would still provide reasonable performance, so can serve as a benchmark for future more elaborate techniques. This year headline (news) corpus, TRC2, consisted of Reuters' articles, which were not as often linked to or referred by their titles (headlines) from the blog posts as the articles in the New York Times corpus used in the preceding year. Thus, the importance of each headline had to be established solely based on the textual content of the article, e.g., by the specific events covered in it.

The simplest possible approach in that scenario would be to treat the entire blog corpus as one giant query which can be used to rank the headlines on the given day. Apart from purely technical challenges that would arise when dealing with a corpus of that size, we decided not to pursue this approach since it disregards the dynamics of the blog corpus. Instead, we looked at the simplest ranking approaches to explore how the lexical changes in the blog corpus can reflect the importance of the articles appearing in the news. Our major underlying hypotheses were that 1) the important events are discussed in blogs and 2) the coverage of the topics related to those events increases on the event day. Thus, an important benefit of a "Blogs to News" approach pursued here over a "News to Blogs" approach is that the dynamics of the blog corpus plays a leading role.

A following simple example illustrates this advantage: a news article about a certain celebrity (e.g. *Lady Gaga*) typically associates with a larger number of blog posts in the entire corpus than a news article about a small town in Italy (e.g. *Aquila*). However, on a certain day, most people would assess an article about *earthquake in Aquila* as more important than the one about *Lady Gaga's concert*, while the letter still may relate to a larger number of blog posts. This dynamics is better captured by "Blogs to News" approach since it would notice the spikes in the use of words related to the earthquake event. Since both approaches have their advantages, we were hoping that future solutions will possibly combine them.

The major challenging in using "Blogs to News" approach is that some words always happen to be mentioned more often than in the background corpus solely due to 1) random chance or 2) "ripple effect" (as related to the events happening days prior). Thus, it was not entirely trivial, before carrying out the experiments reported here, that such a simple approach as we tried, would be at all effective. For example, the terms from the statistical "tail" of the distribution of the frequency of mentioning could have happened to completely dominate any useful terms and lead to almost random ranking of headlines. Thus the mechanism to select the terms is important.

Our selection of terms to query and rank the news corpus was based on their "signal-to-noise" ratios on the given query date, defined for each (unigram) word *t* as the following:

$$s(t) = tf/atf, \quad (1)$$

Where *tf* is the total number of occurrences on the query day and *atf* is the average number of occurrences within twenty (*20)* immediately preceding days. The following thresholds were applied to filter out the terms (words) making only negligible impact on the headline ranking to reduce the processing time:

1. The term has to be mentioned at least once in the news corpus on the day of the query.
2. The term has to be mentioned at least *1000* times in the entire blog corpus.
3. The term has to be mentioned at least *100* times in the blog corpus on the day of the query.
4. The signal-no-noise ratio defined above has to be at least *1.5*.

The top *100* by their signal-no-noise ratios terms satisfying the constraints above, were combined into a single query to order the headlines on the given date. The retrieval and ranking was performed using the Terrier platform (Ounis et al., 2006) and its PL2 model with its default parameters. All the other above mentioned parameters were selected using the 2009 topics (query dates). The "Current Event" articles of Wikipedia Portal, described in the next section, also provided informal benchmarks.

Once the headlines were ranked, the required classification of them into the specified set of categories (*us, world, sports, sci-tech, business*) for all the submitted runs was provided by the University of Glasgow group. When the classifier failed to identify 50 headlines in a certain category on a query date, the remaining spots were filled with the remaining ranked headlines ignoring their category assignments.

Automated run "strath3" was similar to "strath1" except the following "saturating" weights were applied to the terms:

$$w(t) = 1 - e^{-\frac{s(t)-1}{a}},$$

where *s(t)* is defined by formula (1) above and *a* was chosen to be *10* during the parameter tuning stage. Table 2 shows an example of such a weighted query using Terrier syntax.

MANUAL EVENT DESCRIPTION RUN

TREC encourages submitting manual runs (those where human intervention occurs) in order to provide additional estimates of possible performance and to diversify the assessment pools. Our submitted "strath2" was such a manual run. Event descriptions were taken from the "Current Event" articles of Wikipedia Portal on the day of each topic (e.g. 22 January 2008 for the first topic, shown in Figure 1). The descriptions of events are typically only 2-3 sentence long, and do not necessary use the same words as the related news articles. Thus, using them as effective queries warranted application of external expansion techniques (Kwok et al., 2004; Muresan & Roussinov, 2006) as following. First, we sent each description to Bing search engine. Then, we mined the returned snippets for the unigram words that occurred twice as much or more frequently in the snippets than in the background (Web) corpus. The mined terms were combined into a query to rank the headlines on a given day, again using PL2 model from Terrier with default parameters. An example of such a query is shown in Table 1.

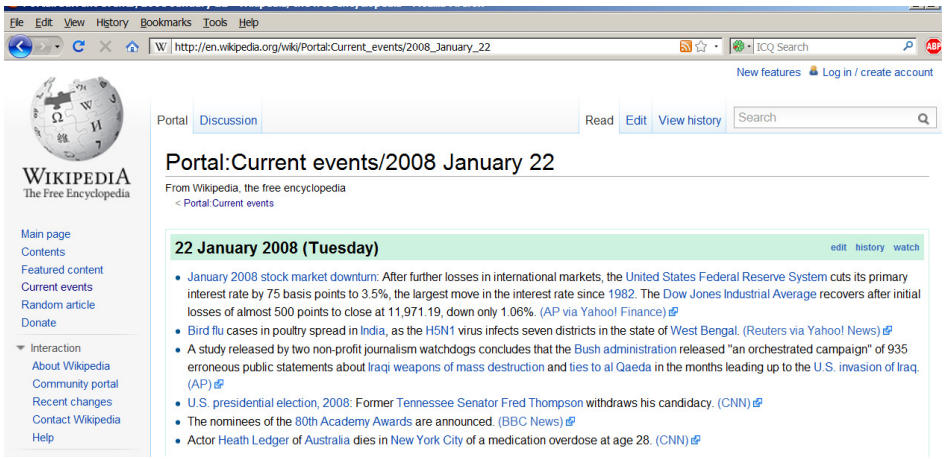

Figure 1. Example of events description on January 22, 2008, in Wikipedia Portal.

```
ledger heath actor overdose actor accidental examiner moog msnw untimely
izabellamikomni683 wcbs gravitated died melvin heroin perth nominee painkillers dies
tumor cnn medication york dead nominated heath medications anxiety ruled manhattan
suspected prescription tragic apartment brady drug brittany australia ruled 22nd
brandon age dies prescribed pointed suicide packets chess pills nyc drugs oscar 1923
city ultram aged
```

Table 1. Example of a query based on Wikipedia event description and external expansion on Bing's snippets.

```
yld^0.920 ieri^0.879 basescu^0.850 sohu^0.806 helplin^0.800 din^0.798 romania^0.794
taxa^0.772 libera^0.761 puls^0.754 bogdan^0.734 cristian^0.715 moldova^0.712
hough^0.708 ledger^0.699 hilit^0.698 raman^0.693 rutherford^0.688 leoni^0.686
tracker^0.677 vesti^0.663 karolina^0.659 walmart^0.638 yam^0.624 hathawai^0.613
baird^0.592 comerci^0.586 inappropri^0.581 rusia^0.572 rezko^0.555 heath^0.543
uae^0.525 ani^0.524 dax^0.507 ticker^0.498 dienstag^0.487 twilight^0.487
enforc^0.477 matthia^0.461 bursa^0.456 vina^0.452 putin^0.450 medvedev^0.446
jelena^0.438 hessen^0.435 australasia^0.433 lor^0.426 loc^0.424 barroso^0.423
brokeback^0.422 intl^0.412 zitat^0.409 mutual^0.408 kreb^0.407 frequent^0.406
shut^0.404 gmbh^0.404 cotillard^0.394 blanchett^0.390 penguin^0.390 tsonga^0.389
tva^0.384 crest^0.384 pool^0.377 liken^0.373 handler^0.368 pune^0.365 mode^0.360
sniff^0.357 blitzer^0.354 confisc^0.353 sanchez^0.350 frankfurt^0.348 nfl^0.348
nato^0.345 ratio^0.345 doubleclick^0.344 bardem^0.340 ion^0.338 ryder^0.338
hudson^0.337 banca^0.337 japon^0.336 sabina^0.336 anymor^0.332 propr^0.329
compliant^0.326 dinar^0.326 deci^0.324 prodi^0.323 cameroon^0.323 dimitri^0.322
gale^0.322 spear^0.322 direkt^0.320 noi^0.318 cisco^0.317 distanc^0.316
schnabel^0.314 crossfir^0.313
```

Table 2. Automatically created weighted query for the first topic (date).

DISCUSSION OF RESULTS

Different from the previous, this year the categories of the deadlines were not given but were expected to be automatically identified. While in general, text categorization is a well researched task and known to be accurate in the 90%+ range, it is still dependent on the definitions of the categories and availability of the training sample. Thus, the impact of category classification on the headline ranking in this year TREC still needs to be determined by additional follow-up analysis, which we left for future research.

Out of all the best runs submitted by each of the six (*6*) participating groups, there are easily visible two clusters by performance: those closely around *.11* and *.21* accordingly. Ours belongs to the lower performing group, which is not surprising since we were looking for the simplest approach that would still be comparable with the other, more sophisticated ones. This comparison happens to be sensitive to the choice of categories to consider: while on the "sport" and "world", ours

would be in the better half (both are #2), "business" and "science" are in the lower one. This may indicate that the relative performance was somewhat determined by the automated classification involved.

The performances of "strath1" and "strath2" are similar, which indicates that the suggested keyword weighting mechanism did not make any tangible impact.

Our best and manual run "strath2" is only marginally better than our automated runs "strath1" and "strath3," which supports the finding that automated analysis of the word dynamics in blogs can be comparable to the keywords obtained from the manually created descriptions of the manually chosen events (Wikipedia "Current Event" articles). The fact, that our manual run is still below 3 other automated runs also indicates that the manual source that we used may not be a very comprehensive one. Indeed, it typically contains no more than 4-6 events on a given day, which, depending on the actual criteria used by the assessors, likely happened to be too few.

RELATED WORK

TREC BLOG

Since the task of headline ranking by its reflection in a large corpus, such as all the blogs combined, is relatively new, most of exploration along this direction has happened within TREC Blog track. A number of different "Blogs to News" and "News to Blogs" approaches were explored by groups participating in 2009 TREC Blog. The University of Glasgow group (uogTr) explored an approach based on their VotingModel for expert search, hypothesizing that the number of blog posts mentioning a headline (aka votes) is a good indicator of the importance of each headline (McCreadie et al., 2009). Their approach happened to be one of the best performing.

The POSTECH KLE group (Pohang University of Science & Technology) estimated the importance of a news headline for a date by linearly combining two probabilities. One is the probability that each news headline generates a given query date, calculated using feed-based or cluster-based approaches. The second is the prior probability that a news headline will be a top story for a given date, estimated using either time-based or term-based evidence. They found out that use of the prior was most important for good performance. They proposed two criteria to estimate the news headline prior that it will be a top story, both working well: 1) The Temporal Profiling approach by Diaz and Jones (2004) applied to the blog posts estimated as relevant to each news headline. 2) The Term Importance criterion that used term information of each news headline estimated by their term frequency of occurrence.

ICTNET (Institute of Computing Technology, Chinese Academy of Sciences) accumulated the BM25 scores for a given headline from the blog posts published that day, and were inspired by topic-focused text summarisation to build diverse blog post rankings (Xu et al., 2009). However, their results were below average.

The University of Amsterdam group used their expert finding model from Balog et al. (2006) as "News to Blogs." For their "Blogs to News", which showed better performance than the former, they identified distinguishing terms from the top 5,000 blog posts from a given date, ordered by their respective number of comments, and the background corpus.

CONCLUSIONS, LIMITATIONS AND FUTURE DIRECTIONS

Based on the informal analysis of the official TREC results, it can be stated that even such a simple implementation of "Blogs to News" approach as described here seems to be effective and promising. It can serve as one of the baselines for the more elaborate future approaches. Our current implementation has a number of technical limitations, mainly due to us only starting to be involved in this task and the short time allocated to preparing our official runs. At present, we consider the following specific improvements:

- Involving phrases in addition to unigram words in order to rank the headlines. Our preliminary analysis of lexical dynamics in the Blogs corpus shows that phrases can serve as very precise indicators of topical dynamics in the blogosphere.
- We used estimated aggregate term statistics (*DF* and *IDF*) based on the news corpus on a given query day only, which may be less reliable than if obtained from a larger corpus, e.g., the blog corpus itself.
- To obtain our signal-to-noise ratio, we used only the simplest metric. Future implementation may involve more sensitive metrics, e.g., involving the variance of the number of occurrences through a *t*-test.
- Using a trainable ranking algorithm for the headlines, rather than applying the one designed for traditional ad-hoc retrieval, e.g. by following Roussinov & Fan (2005).
- Treating the title and the content of a news article differently.

ACKNOWLEDGEMENTS

This work has been partially supported by Dr. Roussinov's start-up grant from the Faculty of Science, Strathclyde University. The i-Lab at the Department of Computer and Information Sciences, lead by professor Ian Ruthven, happened to be supportive and encouraging environment. I am especially grateful to Ian for providing valuable feedback on this report.

The University of Glasgow group has kindly provided the index of the news corpus, the daily lexicons of the Blog corpus and the classification of the headlines into the required set of categories.

I am also grateful to Thomson-Reuters for providing their TRC2 corpus free of charge for all the TREC participants.

REFERENCES

Balog, K., Bron, M., He, J., Hofmann, K., Meij, E., de Rijke, M., Tsagkias, M., Weerkamp, W. (2009). The University of Amsterdam at TREC 2009. Blog, Web, Entity, and Relevance Feedback. *In D. K. Harman, editor, Proceedings of the Twelve Text Retrieval Conference, NIST Special Publication,* 2009.

Diaz, F., Jones, R. (2004) Using temporal profiles of queries for precision prediction. In: SIGIR '04, ACM, pp. 18–24

Kwok, K.L., Grunfeld, L., Sun, H.L., Deng, P. and Dinstl, N. (2004). TREC2004 Robust Track Experiments using PIRCS. *In D. K. Harman, editor, Proceedings of the Twelve Text Retrieval Conference, NIST Special Publication,* 2003.

Macdonald, C., Ounis, I., and Soboroff, I.(2009). Overview of the TREC2009 Blog Track. *In D. K. Harman, editor, Proceedings of the Twelve Text Retrieval Conference, NIST Special Publication,* 2009.

McCreadie, R., Macdonald, C., Ounis, I., Peng, J., Santos, R.L.T. (2009). University of Glasgow at TREC 2009: Experiments with Terrier Blog, Entity, Million Query, Relevance Feedback, and Web tracks. *In D. K. Harman, editor, Proceedings of the Twelve Text Retrieval Conference, NIST Special Publication,* 2009.

Muresan, G. and Roussinov, D. (2006). Where Do Good Query Terms Come From? Proceedings of the Annual Meeting of the American Society for Information Science and Technology. (ASIST 2006), Austin, Texas, November 2006.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., and Lioma, C. (2006). Terrier: a high performance and scalable information retrieval platform. In *Proceedings of OSIR Workshop at SIGIR 2006*.

Roussinov, D., and Fan, W. (2005). Discretization Based Learning Approach to Information Retrieval. In *proceedings of 2005 Conference on Human Language Technologies*.

Xu, X., Liu, Y., Xu, H., Yu, X., Song, L., Guan, F., Peng, Z., Cheng, X. (2009). ICTNET at Blog Track TREC 2009. *In D. K. Harman, editor, Proceedings of the Twelve Text Retrieval Conference, NIST Special Publication,* 2009.