Topic Categorization for Relevancy and Opinion Detection

GuangXu Zhou[1], Hemant Joshi[2], and Coskun Bayrak[1]

[1]Computer Science Department
University of Arkansas at Little Rock
Little Rock, AR 72204

[2]Acxiom Research, Acxiom Corporation

{gxzhou@ualr.edu | hemant.joshi@acxiom.com | cxbayrak@ualr.edu}

Introduction

University of Arkansas at Little Rock's Blog Track team participated in only the core task of the blog track this year. The data acquired was identical to that of previous year except some new .retrieval tasks were introduced. The core task was to identify blogs that are opinionated about a certain subject. *Fifty* new topics were provided by National Institute of Standards and Technology (NIST) this year. Apart from the core task, two subtasks were also introduced. Polarity subtask was to detect polarity of the opinionated blog about a given topic. Feed distillation subtask was based on finding feeds rather than individual permalinks. Last year, we participated in the core task [1] and this year we planned to continue on our previous work. Although an attempt was made last year to use Active Learning with Support Vector Machine (SVM) to detect opinionated blog, identifying the opinion expressed about a given topic was unsuccessful. The difference this time around is in the use of search engines to conduct the topic search, categorizations of queries for further training, and a Natural Language based "one-pass-processing" approach.

Data

Total of *6* runs were allowed for each participating team this year. The blog data provided [2] [3] consists of approximately *3.2 million* permalinks from *100,649* feeds. The data consists of spam blogs as well as non-English language blogs. Also the opinions expressed may be in colloquial form or in abbreviations commonly used as chat lingo over the internet. Once opinion is detected, we need to find out if the opinion expressed is about a particular topic. Using paragraphs or passages as the context of opinion detection has shown to produce good results [4]. We intended to limit our focus not through windowing techniques but through combination of Natural Language Processing (NLP) and Machine Learning techniques.

Query Categorization

Apart from blog data, we also analyzed the *50* topics provided this year and divided them into 6 categories namely *thing*, *company*, *food*, *event*, *location* and *person*. Figure 1 shows the distribution of *2007* blog track topics. Categorization of topics helped us get a general idea about the topics. We were able to come up with generic patterns to detect an opinion for each category. Category identification was useful in Categorized Machine Learning approach that we will discuss later in this paper. Majority of the queries were in *person*, *company*, *thing* and *event* categories.

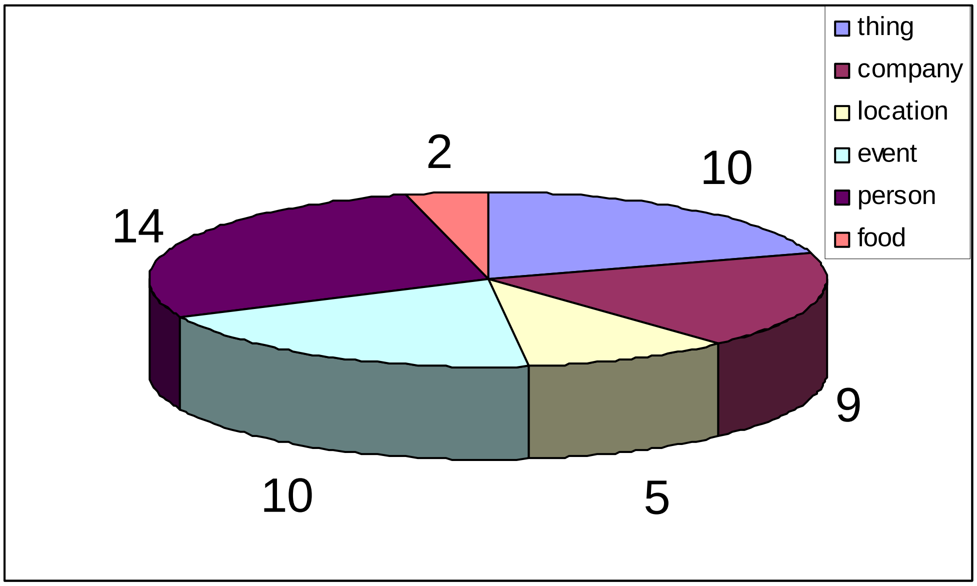

Figure 1: Categorization of *2007* blog track topics into *6* generic categories

We submitted total of *6* runs: *2* baseline and *4* other runs each using a different approach. The baseline run, *UALR07BlogBase,* was obtained with *Indri search engine*, part of *Lemur* language Modeling toolkit [5] with pseudo document feedback. Topics given were simply mapped to Indri syntax. We did not use query expansion for the base run. We used title field only for the *UALR07BlogBase* run. For the second base run, *UALR07TDN* we used Title and description as well as narrative fields. Below we describe the other *4* runs submitted, each with a different approach.

1. *UALR07BlogIU*: IU run was our significant contribution this year to blog track. This run is based on assumption that presence of subjectivity indicators near topic or query words is good indicator of the content opinionated about a given topic. We chose words like “I”, “you”, “we”, “me” as well as words such as “like”, “feel”, “think” etc. and mapped the distance of these words to the topic given a window size of *10* or *20* words. We believe that the presence of subjectivity indicator words near the topic under investigation provides a clue for selecting these blogs with a preference.

   Run *UALR07BlogIU* was a run using modified indri queries. We used unordered window size of 20 words (#uw20 in indri query syntax). Only title field from the given topics was considered for this run. We looked for words such as {*me, we, I, they, you, he, she*} and also words such as {*like, love, hate, suck, nice, good, bad, awesome, awful, never, think and, feel*} within unordered window of 20 words to topics such as *Mozart*. This allowed for weighting those queries that consist of word *Mozart* along with opinion indicators higher than blogs that only contain word Mozart or generic opinion words. Our initial experiments showed an improvement by using relaxed window size of *20* instead of *10*. The unordered window was preferred because we were not certain about the order in which a topic and opinion about that topic will be expressed.

2. *UALR07BlogIU2*: IU2 run was designed to be an improvement over IU run discussed above. Instead of title only, we used narrative and description fields for the given topic and expanded the query to contain those extra words near the opinion indicator words.

The opinion indicator words were the same as those used in *UALR07BlogIU* run. So *UALR07BlogIU* run is title only run whereas *UALR07BlogIU2* is Title-Description and Narrative based run (TDN). The reason for using TDN fields was to increase recall levels, especially in the lower ranks for each topic results. Again the queries were modified to be compatible with Indri format and unordered window of 20 words was used.

3. *UALR07BlogCML*: Categorized Machine Learning (CML) run was designed to take advantage of the 6 general categories of topics. In the previous year we trained SVM on all topics and used active learning approach to judge the most difficult blogs to make a prediction. This approach, although promising, seems to focus on only certain topics rather than all 50 of them. We need to be able to generalize queries and train SVMs to detect opinions in each of those categories. So we trained SVM with linear kernel and C-SVM on each of the 6 categories mentioned earlier. Using active learning we further trained SVM to predict opinionated blogs in that category. Next based on each query we predicted opinionated blogs from the top 1500 results obtained for each query using Indri. We re-ranked these 1500 results giving more importance to opinionated blogs. This approach addresses issues with topic relevance as well as uses machine learning techniques to build category specific models for prediction of opinionated blogs. The other significant difference from that of the last year was limiting the scope of prediction not to all the given categories but only to top 1500 results obtained for each query in that category. This eliminates noise or spam blogs and thus should produce better accuracy than the one reported in the previous year.

4. *UALR07Text*: We used a new natural language based "one-pass-processing[1]" approach for this run. We did not rely on Indri's language model to generate results of the run. Instead we used complete Natural Language Processing approach and regular expressions for retrieval. We parsed permalink documents to extract text from only English blogs. Blogs in languages other than English were ignored. We then segmented each text permalink into passages and used sum of passages matching hand crafted regular expressions of topic to generate final score of each document. Finally, we ranked all documents in descending order of scores. Using passage for establishing a context is a promising idea since it will eliminate possibilities of blogs referring to the words in the topic but not expressing opinion about the topic. Passages are of flexible length and usually referred to <p> and </p> tags in HTML documents. We also wanted to compare performance of NLP approach with that of language modeling where we determine probabilistic values.

Results and Analysis

We compare results from all 6 runs. Table 1 shows the topic relevance results comparison over 50 queries for all runs submitted. The numbers in bold are the highest values for the judging criteria. Mean Average Precision (MAP) values are shown with gray background.

[1] A new dictionary based paragraph scoring method, which is designed to handle the ranking of topic relevance and opinion detection at the same time, since the topic-relevance and opinion-finding have similar effect on scoring and ranking.

Table 1: Topic Relevance Results comparison of 6 runs submitted

	UALR07 Base	*UALR07BlogIU*	*UALR07 CML*	*UALR07 BlogIU2*	*UALR07BlogTDN*	*UALR07 Text*
num_rel	12187	12187	12187	12187	12187	12187
num_rel_ret	7846	**7881**	7846	7256	6817	7612
map	0.3401	**0.3612**	0.3316	0.3292	0.2931	0.3081
P5	0.624	**0.796**	0.58	0.784	0.664	0.556
P10	0.628	**0.734**	0.59	0.74	0.652	0.534
P15	0.6147	**0.708**	0.6027	0.6973	0.6187	0.5253
P20	0.6	**0.675**	0.584	0.67	0.589	0.51
P30	0.578	**0.6353**	0.564	0.6133	0.5533	0.4967
P100	0.471	**0.4864**	0.4714	0.4604	0.419	0.4304
P200	0.3801	**0.3886**	0.3814	0.3628	0.3296	0.3499
P500	0.2468	**0.2534**	0.2496	0.2263	0.2149	0.2248
P1000	0.1569	**0.1576**	0.1569	0.1451	0.1363	0.1522

From Table 1, UALR07BlogIU run reported highest MAP and precision_at_N values. We also notice that MAP is less for *UALR07BlogTDN* run compared to *UALR07BlogBase* run.

Table 2 shows the opinion results comparison of the 6 runs submitted. The highest values are shown in bold numbers for each row. Also for Mean Average Precision comparison, results of *UALR07BlogIU* run reported best values. From tables 1 and 2, it can be concluded that presence of opinion words near a topic is a good approach to identify blogs that are opinionated about a particular topic.

Table 2: Opinion Results comparison for 6 runs submitted

	UALR07 Base	*UALR07BlogIU*	*UALR07 CML*	*UALR07 BlogIU2*	*UALR07Blog TDN*	*UALR07 Text*
num_rel	7000	7000	7000	7000	7000	7000
num_rel_ret	4717	**4736**	4717	4313	3976	4545
map	0.2554	**0.2911**	0.2521	0.265	0.2102	0.2183
P5	0.452	0.632	0.432	**0.636**	0.428	0.456
P10	0.44	**0.58**	0.42	0.558	0.406	0.408
P15	0.4187	**0.5427**	0.4227	0.524	0.3773	0.3787
P20	0.402	**0.51**	0.407	0.492	0.36	0.37
P30	0.374	**0.4647**	0.3753	0.446	0.3333	0.3493
P100	0.283	**0.3274**	0.2872	0.3092	0.2414	0.2716
P200	0.2275	**0.2443**	0.231	0.2261	0.1875	0.2179
P500	0.1484	**0.1544**	0.1506	0.1344	0.1228	0.1361
P1000	0.0943	**0.0947**	0.0943	0.0863	0.0795	0.0909

Figure 2, shows the interpolated precision-recall response comparison between the 6 runs submitted for topic relevance. Similarly, Figure 3 shows the interpolated precision recall response comparison for opinion detection. All *4* runs, *UALR07BlogIU*, *UALR07BlogCML*, *UALR07BlogIU2,* and *UALR07Text* performed better than the *2* baseline runs.

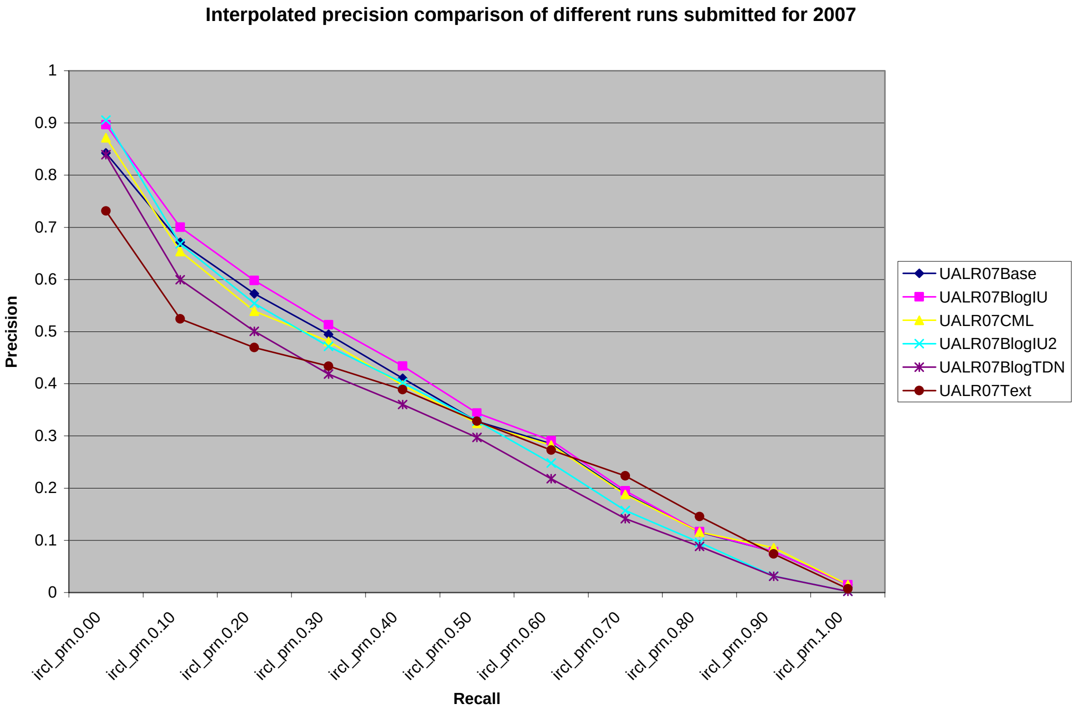

Figure 2: Interpolated precision recall response comparison for topic relevance for the 6 runs

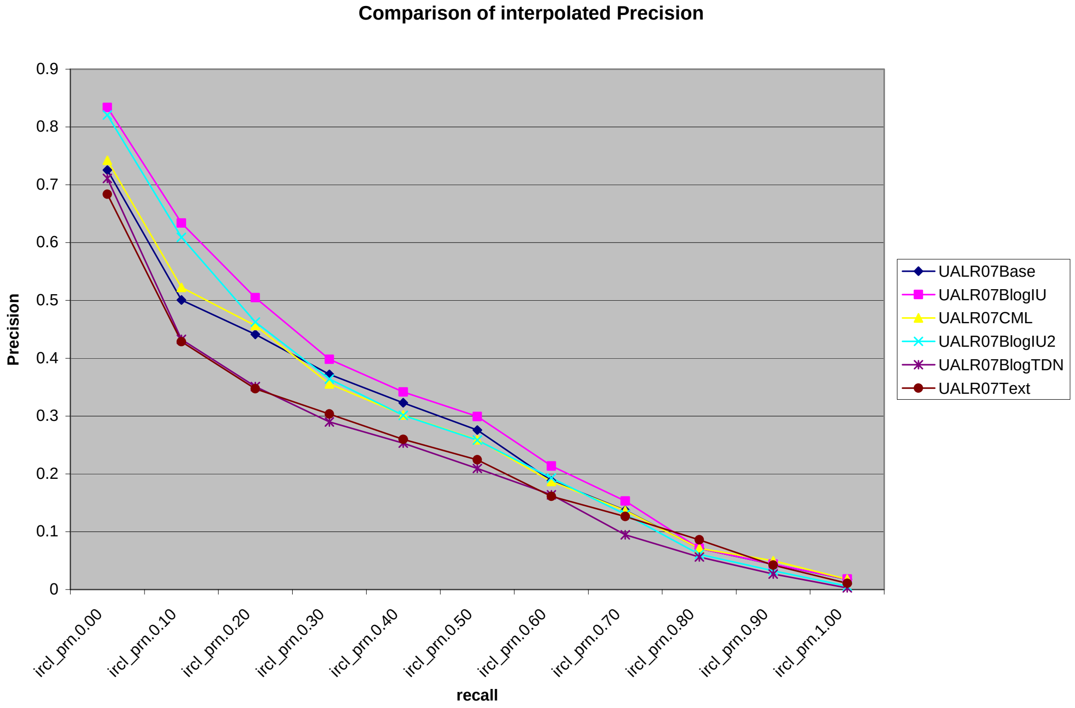

Figure 3: Interpolated precision recall response comparison for opinion detection for the 6 runs

Figure 4 shows MAP response of topics from *901* to *925* for the *4* runs *UALR07BlogIU*, *UALR07BlogCML*, *UALR07BlogIU2* and *UALR07Text* to detect opinions. Figure 5 shows similar response of 4 runs for topics from *926* to *950*.

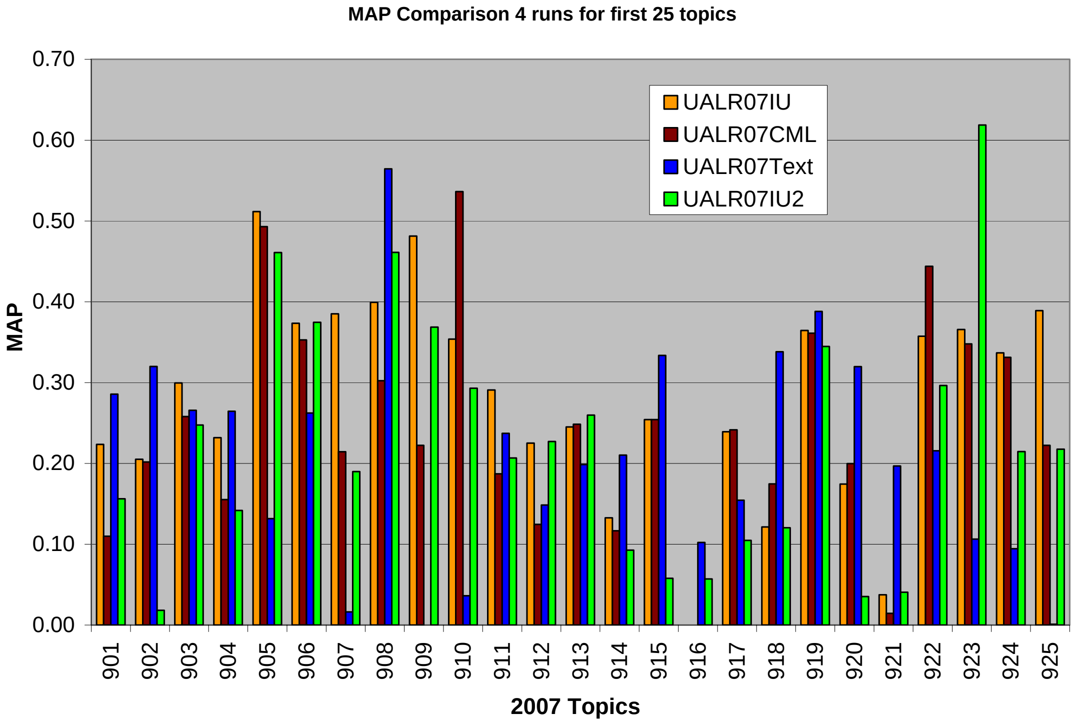

Figure 4: MAP comparison of topics from 901 to 925

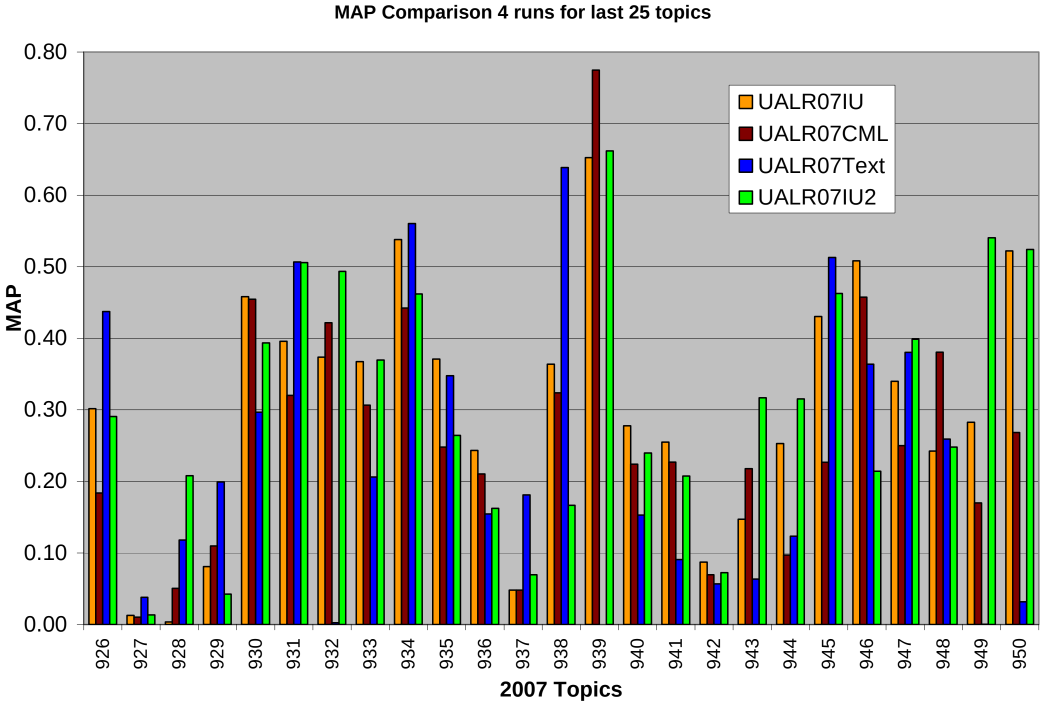

Figure 5: MAP comparison of topics from 926 to 950

Conclusion

We experimented with 3 different approaches and evaluated their performances in detecting blogs that are opinionated about a given topic. UALR07Text run performs as good as indri baseline run and in fact has even better results for certain topics. We intend to refine our technique and make some improvements over the results of UALR07Text run. CML and IU runs also provided new intuitive way of looking at opinion detection problem. Both performed satisfactorily over the baseline.

Acknowledgements

We would like to thank NIST for taking the lead to organize the TREC conference and the people who made it a pleasant experience by participating in it. We also would like to thank Dr. Nasrolah Samadi's constructive comments in revising this manuscript.

References

[1] http://trec.nist.gov/pubs/trec15/papers/uarkansas.blog.final.pdf
[2] http://www.dcs.gla.ac.uk/~craigm/publications/macdonald06creating.pdf
[3] http://ir.dcs.gla.ac.uk/test_collections/blog06info.html
[4] http://trec.nist.gov/pubs/trec15/papers/umd.blog.ent.legal.qa.final.pdf
[5] http://www.lemurproject.org/