UALR at TREC-ENT 2007

Hemant Joshi[+], Sithu D Sudarsan[#], Subhashish Duttachowdhury[*], Chuanlei Zhang[#], Srini Ramaswamy[*]
+ Acxiom Research, Little Rock, AR, USA
# Applied Science – University of Arkansas at Little Rock, AR, USA
*Computer Science, University of Arkansas at Little Rock, AR, USA
hemant.joshi@acxiom.com {sdsudarsan, sxduttachowd, cxzhang, srini} @ualr.edu

This is the first year we participated in the enterprise track. This year's enterprise track offered completely new enterprise data and two new tasks. The data offered was the CSIRO Enterprise Research Collection corpus[1]. The two new tasks introduced this year are *Expert search* and *Document search*. We participated in both tasks, though Document Search was our primary focus this year. We also believe that the results in our document search task might have a direct impact on the expert search task.

Expert search task was to identify experts or subject matter experts given a particular topic. The goal was to drive queries regarding a certain subject be diverted to a particular set of experts. Identifying experts from the document collection is a challenging problem. We have to assert if the document is informative enough for the given topic and shows the mark of an expert. We have to also find the author of the article or the relevant name or email address mentioned. The results were to be submitted as email addresses with proof of documents that we believe are expert information for the given topic. Fifty new topics were provided by NIST[2] and evaluation for expert search task was conducted with help from real-world CSIRO personnel.

The document search task was to identify documents that are authoritative information about a given topic. Fifty topics were common among the document search and expert search tasks. The challenge was to determine if the document merely contained words associated with the given topic or the document was indeed the authoritative source on that topic. We had to analyze the documents relevant to the given topic and rank them according to how informative those documents are for that particular topic. We experimented with various approaches that can estimate authoritative information content contained within a document. We discuss these approaches and compare them later in this paper.

TREC-ENT Data

CSIRO document collection consists of *370,715* documents with unique IDs provided with the collection. Each individual document is in HTML and is in the TREC format[3].

[1] http://es.csiro.au/cerc/
[2] http://www.nist.gov
[3] http://trec.nist.gov

From the CSIRO website, we gather that the enterprise is organized into flagships and divisions. There are *7* flagships and *18* divisions listed on the website that represent diverse areas of research. We identified a list of *19,073* email addresses that occur at least once in the collection and sort them in descending order of frequency of occurrence. We removed email addresses such as *ento-webmaster@csiro.au* which do not indicate an expert but a generic email address. We also removed email addresses that occur only once in the entire collection. Finally we have *6,754* email addresses that we believe uniquely represent an expert within the organization.

Document Search Task

We submitted *4* runs namely *UALR07Ent1*, *UALR07Ent2*, *UALR07Ent3* and *UALR07Ent4*. UALR07Ent1 was our baseline run. We used the Indri search engine[4] available as a part of Lemur Language Modeling Kit[5]. We used the top 5 document pseudo feedback to boost the accuracy of the search results returned. The *50* topics were modified to follow Indri query syntax[6]. For the base run, we did not use query expansion. Our objective was to experiment with various approaches to detect documents that are authoritative about the given topic and to rank them.

- *UALR07Ent2*: We used MMRSummApp[7] (part of Lemur Applications) which is a complex summarizer that compares passages of the document with respect to the given query and summary length. We passed sample authoritative documents for each query as input to MMRSummApp. Query is important for summarization to know which sentences establish correct context of the query. We limited each summary to 20 words and added those words to the actual query (also known as topic) and re-ranked top *1,500* results of each topic from base run. Re-ranking helps us boost those results which contain most of the words from summary of sample authoritative pages provided.
- *UALR07Ent3*: This run was driven by one question:

  What is so special about documents that will make them authoritative or more informative about a particular topic?

  In order to answer this question, we introduce the '*word difference'* approach. We set out to find what words are in sample authoritative pages that are not in our top *5* documents of the base run. This will tell us not to focus on common words but those special indicative words that indicate authoritative source about the particular topic. So we found words in sample authoritative pages that were not in top results for the topic. Then, we added those words to the query through manual query expansion and re-ranked *1,500* results obtained using Indri. Top *1,000* ranked documents for each query were submitted as run *UALR07Ent3*.

- *UALR07Ent4*: Unlike previous runs, this is a manual query expansion run. We manually selected and modified given topics to yield more informative documents

[4] http://www.lemurproject.org/indri/
[5] http://www.lemurproject.org/lemur/
[6] http://boston.lti.cs.cmu.edu/lti-search/help-qry.html
[7] http://www.cs.cmu.edu/~lemur/doxygen/lemur-2.0/html/MMRSummApp_8cpp.html

as top results for each query. This run was submitted due to encouragement from track organizers to submit manual runs. We were also interested to see if our other two runs *UALR07Ent2* and *UALR07Ent3* can either match or perform better than the manual run.

Expert Search Task

Expert search task was not the primary task we participated in, and so, we decided to utilize our results and runs from the document search task to submit 3 runs for the expert search task.

- *UALR07Exp1*: We used the entire *50,000* results for the *50* topics from *UALR07Ent3* run to identify potential experts for each query and filtered them against our manually created list of *6,754* email addresses. If the email address existed in our master list of expert emails, then we added and used the container document's rank to list up to *100* experts for each topic. The format also required us to submit documents that support the claim that a particular email address is an expert.

- *UALR07Exp2*: This run is different from *UALR07Exp1* in one aspect. Instead of using all *50,000* results from all topics, we focused on each query individually and used a set of *1,000* results from UALR07Ent3 to identify experts and cross-reference them against the master list of *6,754* email addresses. *UALR07Exp1* run is aimed at identifying global experts but *UALR07Exp2* run identifies topic specific experts. We limited the number of experts to *100* though it was suggested that potential number of experts for each run would be at best *2* or *3* for each topic.

- *UALR07Exp3*: This is manual run where we identified expert email addresses for each topic by analyzing top results from *UALR07Ent3*. We used *UALR07Ent3* run as baseline for all *3* runs submitted in expert search task. Run *UALR07Ent3* (discussed earlier) uses word difference approach and at the time we believed that *UALR07Ent3* run would produce best results in identifying authoritative documents. So we decided to exploit that run to enhance our runs submitted for expert search task.

Results and Discussion

For document search task, only *42* topics have been completely judged and so at the time of writing this, we present results obtained in *42* out of the *50* topics. Figure *1* shows the interpolated precision response comparison of *4* runs submitted for document search task. All *3* runs *UALR07Ent2*, *UALR07Ent3* and *UALR07Ent4* perform better than the baseline. Also we observe that runs *UALR07Ent2* and *UALR07Ent3* match or better the performance of manual run *UALR07Ent4*. Figure *2* shows precision response comparison of the 4 runs submitted. Results from figure *2* also reiterate the same observation that word difference approach as well as summarization (MMRSummApp) approach can match human experts in identifying authoritative documents.

We did not perform well in expert search task. We anticipated poor results as we did not get enough time to focus on expert search task. We plan to improve on our expert search task results in near future.

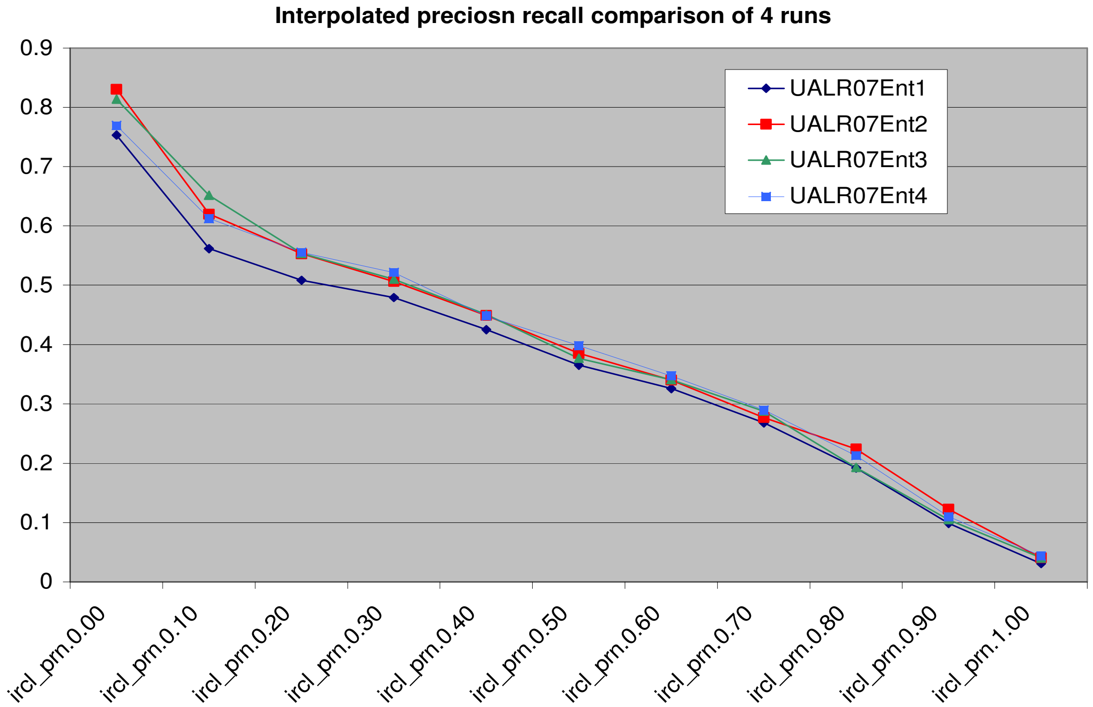

Figure 1: **Document Search**: Comparison of interpolated precision response of *4* runs

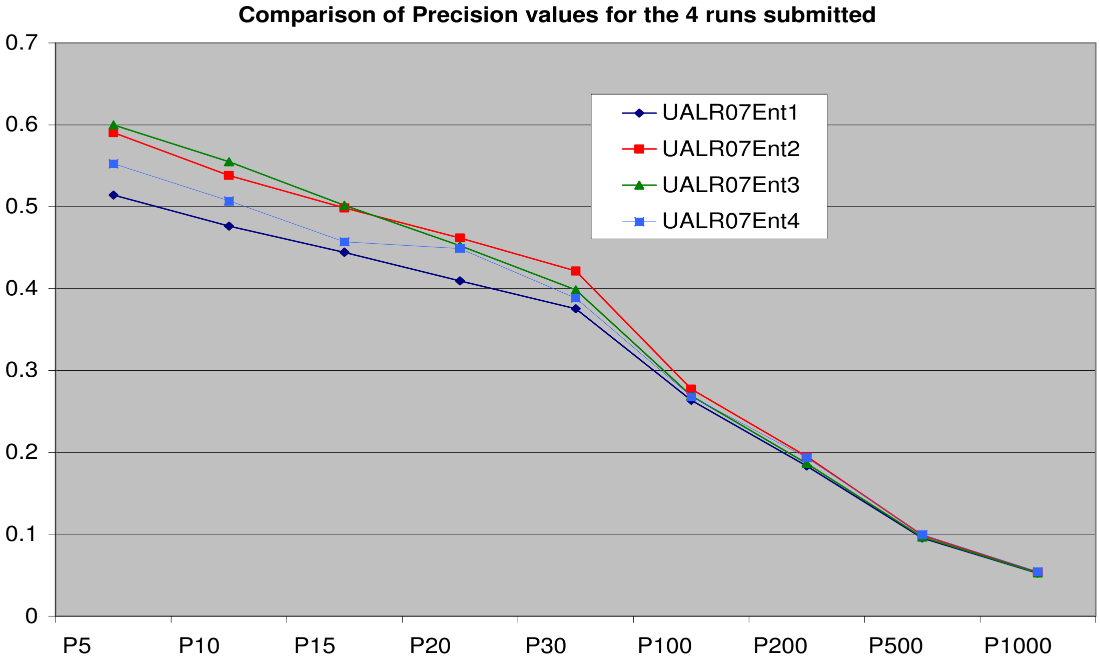

Figure 2: *Document Search*: Comparison of the precision response of *4* runs

Conclusion

We performed extremely well in Document Search task and we were satisfied with our results. We performed better than the baseline run established and matched performance of a manual run and even performed better in early precision values. We plan to continue our research using approaches discussed earlier.