A Micro-analysis of Topic Variation for a Geotemporal Query

Fredric Gey†, Ray Larson†, Jorge Machado ‡, Masaharu Yoshioka*
†University of California, Berkeley USA
‡INESC-ID, National Institute of Electroniques and Computer Systems, Lisbon, PORTUGAL
* Hokkaido University, Sapporo, JAPAN
gey@berkeley.edu , ray@ischool.berkeley.edu ,
jorge.r.machado@ist.utl.pt , yoshioka@ist.hokudai.ac.jp

ABSTRACT

Bias introduced in question wording is a well-known problem in political attitude survey polling. For example, the question "The President believes our military mission in Afghanistan is a vital national interest -- agree/disagree?" is quite different from the question: "Do you believe that a military mission in Afghanistan is in the USA’s vital national interest?" Response variation according to different question wording has been studied by researchers in survey methodology. However the influence on search results from variations of topic wording has not been examined for geotemporal information retrieval. For the GeoTime evaluation in NTCIR Workshop 9, the organizers decided to attempt to do an experiment in query variability in order to study variability of performance. We took a single information need and expressed it in three different ways: 1) as a single event question, 2) as a question which would yield an open-ended list (e.g. the classic “which countries did the Pope visit in the last three years”), or 3) a reformulation or the single event question as a location (latitude/longitude) and time inquiry. This paper reports the results of this micro-analysis of variation effects upon a single query expressed in different formats, as well as the degree of success (or failure) which we achieved (or did not achieve) our explicit goal of being able to distinguish performance outcomes for the different formulations.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval—retrieval models, search process. **General Terms:** Experimentation, Performance, Measurement **Keywords:** Geotemporal Search, Geographic Information Retrieval, IR evaluation

1. INTRODUCTION

It is well accepted in information retrieval that query wording and formulation affects retrieval performance. Most evaluation studies have focused upon the process of query expansion using either direct relevance feedback [9] or blind feedback [7]. Ganguly, Leveling, and Jones have recently taken a simulation approach to this problem [4]. An example where manual query reformulation created astounding results was within the TREC-4 Spanish track where the University of Central Florida dedicated approximately 20 hours per query to creating a detailed semantic word structure to express each information request [3]. For example, a query on “news stories about Mexican jewelry” was expanded to include two 'instance' facets – “torquoise” and “silver” as well as “rings”, “bracelets” and “necklaces.” The result of this labor-intensive effort was to produce a convex shaped recall-precision curve instead of the usual concave shape produced by all other participating teams. In the words of David Hull “University of Central Florida did us all a great service by finding a large number of relevant documents which would never have been found by the automatic runs” [Hull 1995 email]. Looking to another discipline, surveys of popular attitudes, we find that question wording in surveys has been well-known to produce profound differences in survey outcomes. As the Pew Center states [8] “The choice of words and phrases in a question is critical in expressing the meaning and intent of the question to the respondent and ensuring that all respondents interpret the question the same way. Even small wording differences can substantially affect the answers people provide.”

To some extent, the effect of variant topic formulations was examined in the TREC Query Track, held in TREC-8 and TREC-9 [1, 2], where the participants in the track generated multiple representations of the base topic and then ran the queries generated by each group. There was not much detailed analysis of the results, though the runs submitted by each participating group in the TREC-9 Query track were made available for further analysis on the TREC web site. The main conclusion was that the forms of queries can have a significant effect on the results when all else is held constant, and that query variations can serve to highlight system issues as well (in [2], for example, the poor performance of one system on a particular query revealed a problem in handling hyphenated words).

In the GeoTime evaluation for NTCIR Workshop 9 we studied three different variations of a single topic:

GeoTime-0035: *When and where did a pipeline explosion occur in Africa killing over 500 people?*

GeoTime-0036: *When and where have there been pipeline explosions in an African country with more than 5 fatalities?*

GeoTime-0037: *What fatal accident occurred near (geographical coordinates 5°52'12"N 5°45'00"E / 5.870°N 5.750°E / 5.870; 5.750), which killed hundreds of people, and when did it occur?*

The contrast would be among a topic with a single answer (35), an 'open-ended' list answer (36), and a topic related to latitude/ longitude, where the answers for all these topics should be either identical or very similar. Our intuitive feel is that topic 36 should be easier than topic 35 because the quantitative restriction of greater than 500 fatalities would require more sophisticated natural language processing. Further, we would intuit that topic 37 would be out of the scope of processing for most traditional IR systems, especially those using 'bag-of-words' approach to similarity matching. It should be noted that this kind of variation in the intended outcome for the topics was different from the form used in the TREC Query track, at least as shown in the examples of [1,2].

2. DATA & TOPIC DEVELOPMENT

For GeoTime in NTCIR-9, two sets of news story collections were used, one Japanese and one English. The Japanese collection consisted of Mainichi newspapers for the periods of 1998-2001 and 2002-2005, while the English collection, consisted, in part, of the NTCIR-8 GeoTime English collection of New York Times stories also for 2002-2005. Because NY Times documents were unavailable for 1998-2001, to cover the same time period in English, the Xinhua Chinese news service English subsection and the English documents from the Korea Times were added and from the English edition of Mainichi. Details about these collections are found in Table 1.

Collection	Language	Time Period	# Documents
Mainichi	J	1998-2001	419,759
Mainichi	J	2002-2005	377,941
Korea Times	E	1998-2001	50,129
Mainichi	E	1998-2001	24,878
NY Times	E	2002-2005	315,417
Xinhua	E	1998-2001	406,792

Table 1: Collections used for NTCIR9-GeoTime

In GeoTime for NTCIR-8 we discovered gaps in the NYT collection for Jan 2003-July 2004. Details about this can be found in [5]. Since in topic development we wished to create topics which had relevant documents in both collections, we had to shy away from events which happened in 2003-June 2004.

For GeoTime in NTCIR-9, we invited participating groups to submit possible topics. The guidelines were to propose questions with time and space aspects. The ground truth evidence was to be provided by Wikipedia articles which specified both time and place. Most topics could be found in the annual notable events in Wikipedia, For example, topic GeoTime-0035 (*In an African country an oil pipeline explosion killed more than 700 people – when and where did this occur*) is found in e.g. http://en.wikipedia.org/wiki/1998, pointing the the article: http://en.wikipedia.org/wiki/1998_Jesse_pipeline_explosion. Each of the 25 topics was vetted to hit at least two relevant documents in both languages.

3. PARTICIPATION

A total of twelve groups participated in GeoTime, with three groups participating in both languages. Only the following groups are identified for this particular evaluation of three topics. For identification of other groups, see the GeoTime Overview [6]. The primary reason for this selection is that these groups performed better than the others on the three topics being analyzed here, and specifically these were the only participants who had Average Precision of greater than 0.0000 for topic 37.

OKSAT	Osaka Kyoiku University, Japan
RMIT	RMIT University, Melbourne Australia
UIOWA	University of Iowa, USA

4. EVALUATION

Relevance judging was done in a traditional manner on a pool of the top 100 documents retrieved from all runs with duplicates removed. For Japanese GeoTime, 15,795 documents were examined and judged. For the English GeoTime, 19,966 were examined and judged. Judgment was graded in that a document could be assessed as "fully relevant" if it contained text which answered both the "when" and "where" aspects of the topic. The document was assessed as 'partially relevant – where' if it answered the geographic aspect of the topic and 'partially relevant – when' if it answered the temporal aspect of the topic. In order to utilize existing evaluation software, the three fully and partially relevant categories (i.e., Both geographically and temporally relevant, geographically relevant only, temporally relevant only) were aggregated into a simple binary relevance categorization upon which the following result tables are based. For the English topics being analyzed here, we have the following:.

English topic	Relevant	Rel-where	Rel-when	Irrele-vant	total
GeoTime-0035	13	3	5	1021	1042
GeoTime-0036	33	4	10	1006	1053
GeoTime-0037	20	5	10	1335	1370

Table 2: Topic Statistics

Topics 35 and 37 were assessed by the first author. The assessment system supplied documents in order of total retrieval status value summed over all systems. Thus for most topics, the relevant documents were usually found in the first hundred or so documents being read, and after that assessment time per document went very quickly. However, the sort order for topic 37 seemed to be entirely random, and every document had to be carefully examined to find the relevant ones.

5. RESULTS

We summarize the results in Table 3 which gives minimum, median, and maximum AP for the three topics, as well as the second-best and best (maximum) team runs for each language:

English topic	**MIN**	**Median**	**MAX**	**UIOWA**	**OKSAT**
GeoTime-0035	0.0000	0.0712	0.5782	0.3974	0.5782*
GeoTime-0036	0.0000	0.1983	0.7491	0.7159	0.7408*
GeoTime-0037	0.0000	0.0000	0.7207	0.1180	0.7207*
Japanese	**MIN**	**Median**	**MAX**	**RMIT**	**OKSAT**
GeoTime-0035	0.0008	0.4563	0.9667	0.2139	0.9667*
GeoTime-0036	0.0010	0.4482	0.8789	0.6952	0.8789*
GeoTime-0037	0.0000	0.0000	0.9514	0.3624	0.9514*

* known manual run

Table 3: Performance Summary for three topics

6. THE TEAMS' METHODS

The team from Osaka Kyoiku University (OKSAT) used external resources such as Wikipedia and Google Maps to construct queries from topics by having team members extract time and place from Wikipedia documents and inserting them into the text of the query [11]. In particular for topic 37, they searched Google Maps for the latitude/longitude location and manually extracted place names found in that neighborhood to add to the final query. In a sense we could say that OKSAT constructed queries which included the essence of the answer to the question posed by the topic. Their results using this approach substantially outperformed other teams' runs. In a sense, their runs provide a goal post to be attained by fully automatic methods. In addition, the OKSAT runs probably retrieved many relevant documents which might not otherwise have been included into the assessment pool. In the results above, all manual runs which included human effort in query construction are marked as such.

The team from RMIT [12] used a novel approach to indexing (*self-indexing)* combined with the well-known Okapi BM-25 ranking algorithm. RMIT also made use of the Japanese Wikipedia for query expansion and feedback in some cases, although they appear to have done this automatically. They did, however, use reverse geocoding on topic 37 using a Yahoo! Geocoding API.

The University of Iowa team [5] used query expansion with geographic terms from the Alexandria Digital Library in conjunction with a language model-based ranking approach. This may have had an impact on the results, although it wasn't clear if the coordinates in topic 37 actually resulted in geographic names in the expanded queries.

Thus, all of the teams examined in this paper used some method of query expansion for the topics, ranging from manual to explicit reverse-geocoding of coordinates.

7. DISCUSSION

Let us draw two (statistically worthless) conclusions from this small case study of query reformulation:

1. A single event topic (35) is more difficult for GeoTemporal retrieval systems than a roughly equivalent topic whose answer results in a multiple list of possible documents (topic 36). This seems to be the evidence from the English runs, but not for the Japanese runs where the median performance of topics 35 and 36 are close. In order to understand the difference, we need to examine whether the Japanese systems used more sophisticated NLP techniques than the English systems.
2. A radically different paradigm in topic expression (topic 37 as a re-phrasing of topic 35 in terms of a latitude/longitude query) results in abysmal performance by systems built according to old paradigm designs. Indeed, *only two groups for each language (identified above in Table 3) had other than zero average precision for this topic.*

8. REFERENCES

[1] C. Buckley and J. Walz. The TREC-8 Query Track, in *Information Technology: The Eighth Text Retrieval Conference (TREC-8)*, Gaithersburg, MD : NIST, 2000.

[2] C. Buckley. The TREC-9 Query Track, in , *in Information Technology: The Ninth Text Retrieval Conference (TREC-9)*, Gaithersburg, MD : NIST, 2001.

[3] J. R. Driscoll, S. Abbott, K. Hu, M. Miller and G. Theis. Multi-lingual Text Filtering Using Semantic Modeling. In *Proceedings of TREC-4,* 1995.

[4] D. Ganguly, J. Leveling, and G. J. F. Jones. Simulation of Within-Session Query Variations using a Text Segmentation Approach. In *Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2011)*, Amsterdam, The Netherlands, September 2011.

[5] C. Harris The Use of Inference Network Models in NTCIR-9 Geotime, in *Proceedings of NTCIR9.,* Tokyo, JAPAN, December, 2011.

[6] F. Gey, R. Larson, N. Kando, J. Machado and T. Sakai. NTCIR-8 GeoTime Overview: Evaluating Geographic and Temporal Search. *Proceedings of NTCIR-8,* Tokyo, JAPAN, June, 2010.

[7] F. Gey, R. Larson, J. Machado and M. Yoshioka. NTCIR9-GeoTime Overview - Evaluating Geographic and Temporal Search: Round 2. In *Proceedings of NTCIR-9,* Tokyo, JAPAN, December, 2011.

[8] P. Ogilvie, E. Voorhees, and J. Callan. On the Number of Terms for Automatic Query Expansion. *Information Retrieval*, 12(6):666-679, 2009.

[9] Pew (2010), Pew Research Center for People & the Press, Question Wording, http://people-press.org/methodology/questionnaire-design/question-wording/

[10] S. Rieh and I. Xie. Analysis of multiple query reformulations of the Web: The interactive information retrieval context. *Information Processing and Managment* 42, 751-768, 2006

[11] T Sato, NTCIR-9 GeoTime at Osaka Kyoiku University-Toward Automatic Extraction of Place/Time Terms, in *Proceedings of NTCIR9.,* Tokyo, JAPAN, December, 2011.

[12] M Yasukawa, J. S Culpepper, F Scholer, Matthias Petri, Language Independent Self Indexing in the NTCIR-9 GeoTime Track. in *Proceedings of NTCIR9.,* Tokyo, JAPAN, December, 2011.

APPENDIX: COMPLETE TOPIC DESCRIPTIONS WITH ANSWER URLS IN XML FORMAT FOR THE THREE TOPICS

```
<TOPIC ID="GeoTime-0035">
  - <DESCRIPTION LANG="EN">
  - <![CDATA[
  When and where did a pipeline explosion occur in
     Africa killing over
  500 people?
    ]]>
    </DESCRIPTION>
  - <DESCRIPTION LANG="JA">
  - <![CDATA[
  500人以上の死者を出したパイプライン事故は、アフリカのどこで
     、いつ起きましたか？
    ]]>
    </DESCRIPTION>
  - <NARRATIVE LANG="EN">
  - <![CDATA[
  An oil pipeline exploded in an African oil-
     producing country and the resulting fire killed
     more than 500 people.  The user wants to know
     where this took place and when was the date of
     the accident.
    ]]>
    </NARRATIVE>
  - <NARRATIVE LANG="JA">
  - <![CDATA[
  アフリカの産油国で起きたパイプラインの爆発で500人以上の死者
     を出す火災が起きた。ユーザはこの爆発が起きた場所と日付を知
     りたい。
    ]]>
    </NARRATIVE>
- <URLs>
   <URL
      LANG="EN">http://en.wikipedia.org/wiki/1998_Je
      sse_pipeline_explosion</URL>
    </URLs>
- <RELs>
   <REL LANG="JA">JA-981020058</REL>
   <REL LANG="JA">JA-981021059</REL>
    </RELs>
   <QUERYDATE YYYYMMDD="20051231" />
      </TOPIC>

   ====================================
<TOPIC ID="GeoTime-0036">
  - <DESCRIPTION LANG="EN">
  - <![CDATA[
  When and where have there been pipeline explosions
     in an African country with more than 5
     fatalities?
     ]]>
     </DESCRIPTION>
  - <DESCRIPTION LANG="JA">
  - <![CDATA[
  5人以上の死者を出したアフリカで起きたパイプライン事故について
     、いつ、どこで、起きましたか？
    ]]>
    </DESCRIPTION>
  - <NARRATIVE LANG="EN">
  - <![CDATA[
  In a single African oil producing country, a number
     of fatal oil pipeline explosions have happened
     between 1999 and 2005.  The user wants to know
    when and where explosions killed more than 5
    persons.
   ]]>
   </NARRATIVE>
 - <NARRATIVE LANG="JA">
 - <![CDATA[
 アフリカの石油会社が1999年から2005年の間に、数件の重大な石
    油パイプラインの事故を起こしている。ユーザはその事故のうち
    、5人以上の死傷者を出した事故について、いつ、どこで起きた
    か知りたい。
   ]]>
   </NARRATIVE>
- <URLs>
                                                        <URL
    LANG="EN">http://en.wikipedia.org/wiki/List_of_p
    ipeline_accidents#Nigeria</URL>
   </URLs>
- <RELs>
  <REL LANG="JA" />
  <REL LANG="EN" />
   </RELs>
  <QUERYDATE YYYYMMDD="20051231" />
   </TOPIC>
 ===================================
<<TOPIC ID="GeoTime-0037">
 - <DESCRIPTION LANG="EN">
 - <![CDATA[
 What fatal accident occurred near (geographical
    coordinates 5°52′12″N
 5°45′00″E / 5.870°N 5.750°E / 5.870; 5.750), which
    killed hundreds of
 people, and when did it occur?
   ]]>
   </DESCRIPTION>
 - <DESCRIPTION LANG="JA">
 - <![CDATA[
 北緯5度52分12秒東経5度45分の近くで起きた数百人の死亡者を出
    した事故は、どのような事故ですか？また、それはいつ起きまし
    たか？
   ]]>
   </DESCRIPTION>
 - <NARRATIVE LANG="EN">
 - <![CDATA[
 This topic requires spatial reasoning, to look up
    places near the geographic coordinates and then
    search for the story about the accident which
    happened there.
   ]]>
   </NARRATIVE>
 - <NARRATIVE LANG="JA">
 - <![CDATA[
 このトピックでは、与えられた地理座標の近くの場所を探すという
    地理空間に関する推論を必要とし、その場所の近くで起きた事故
    について調べる。
   ]]>
   </NARRATIVE>
- <URLs>
  <URL
    LANG="EN">http://en.wikipedia.org/wiki/1998_Je
    sse_pipeline_explosion</URL>
   </URLs>
- <RELs>
  <REL LANG="JA">JA-981020058</REL>
  <REL LANG="JA">JA-981021059</REL>
   </RELs>
  <QUERYDATE YYYYMMDD="20051231" />
   </TOPIC>
 ==============================
```