NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search

Fredric Gey†, Ray Larson†, Noriko Kando‡, Jorge Machado *, Tetsuya Sakai**
†University of California, Berkeley USA
‡National Institute of Informatics, Tokyo JAPAN
* INESC-ID, National Institute of Electroniques and Computer Systems, Lisbon, PORTUGAL
** Microsoft Research Asia, Beijing, CHINA
gey@berkeley.edu , ray@ischool.berkeley.edu , kando@nii.ac.jp,
jorge.r.machado@ist.utl.pt, tesakai@microsoft.com

ABSTRACT

For the NTCIR Workshop 8 we organized a Geographic and Temporal Information Retrieval Task called "NTCIR GeoTime". The focus of this task is on search with Geographic and Temporal constraints. This overview describes the data collections (Japanese and English news stories), topic development, assessment results and lessons learned from the NTCIR GeoTime task, which combines GIR with time-based search to find specific events in a multilingual collection. Eight teams submitted Japanese runs (including unofficial three teams who provided runs to expand the pools) and six teams submitted English runs. One team participated in both Japanese and English.

Categories and Subject Descriptors

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

1. INTRODUCTION

Cultural Geographic search is quite prevalent in many modern search venues. A great number of documents (web, news, and scientific) have a geographic focus. Geographic search allows for a unique user interface, the interactive map, which can be utilized not only to narrow the user's focus by geography, but also to highlight interesting events. There have been over six workshops [6] on Geographic Information Retrieval (GIR) held in association with SIGIR, CIKM, ECDL or other conferences as well as workshops and conference tracks on location-based search, there has also been 4 years of evaluation of GIR within CLEF (the GeoCLEF track). But, until this task at NTCIR, Asian language geographic search had never been specifically evaluated, even though about half of the NTCIR-6 Cross-Language topics had a geographic component (usually a restriction to a particular country).

Geographic information retrieval is concerned with the retrieval of thematically and geographically relevant information resources in response to a query of the form {<*theme* or *topic*, *spatial relationship*, *location*>}, e.g. ``Temples within 5 km. of Tokyo". [4]. Systems that support GIR, such as geographic digital libraries, and location-aware web search engines, are based on a collection of georeferenced information resources and methods to spatially search these resources with geographic location as a key. Information resources are considered geo-referenced if they are spatially indexed by one or more regions on the surface of the Earth, where the specific locations of these regions are encoded either directly as spatial coordinates, i.e. geometrically, or indirectly by place name [2]. However, in order for place names to support a spatial approach to GIR, they must be associated with a model of geographic space.

The temporal aspects of search have been largely ignored in the IR community, but not in the GIS and information processing communities. There has been a special issue of ACM TALIP on 'Temporal Information Processing' [5], as well as at least two workshops on "Temporal and Spatial Information Processing". The NTCIR-GeoTime organizers wanted to utilize and incorporate past research on this aspect as part of the evaluation.

2. DATA

Two news story collections were used for NTCIR-GeoTime, one Japanese and one English. The Japanese collection was identical to that used for the ACLIA and IR4QA evaluations: Mainichi newspapers for 2002-2005, which had 377,941 documents. The English collection, which was chosen to match with the NTCIR MOAT task on multilingual opinion, consisted of 315,417 New York Times stories also for 2002-2005. Users of the NYT collection had to pay a fee of $50US to the Linguistic Data Consortium to prepare and mail the DVD with this collection. Since we were interested in looking for particular events around which geotemporal topics could be constructed, we ran frequency distributions on both collections by month and discovered gaps in the NYT collection for Jan 2003-July 2004. While the monthly average of documents for 2002 and 2005 was 9,982 and 8,703 respectively, for 2003 and 2004 it was 2,319 and 5,280. Indeed from January 2003 through June 2004 (zero documents), the number of documents per month ranged from 0 to 2209 documents (see the GeoTime collection web page http://metadata.berkeley.edu/NTCIR-GeoTime/databases.php for the complete distributions). A full complement of monthly documents resumed in July 2004. This was described to us by LDC as "a known flaw" because the source document images

used for OCR were too corrupt to produce reliable text, so these documents were omitted. 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.

3. TOPIC DEVELOPMENT

Following the advice of other task organizers, we wanted to create topics which were as realistic as possible. We briefly explored obtaining query logs from actual commercial search engines but quickly abandoned the effort because of institutional barriers and privacy considerations. We downloaded and explored the Excite query log data but found the collection to be too primitive (in the sense of a miniscule number of 'where and when' queries) and too outdated to provide reasonable topics. We then downloaded and explored the TREC million query track (cite needed) by searching for those queries containing keywords 'when' or 'where'. Because the million query track comes from the TREC web evaluation, covering a collection of government web pages, the results of our exploration were miserable, containing items like:

2174: when to clean bird feeder

9375: where's my state refund

In the end we looked at the Wikipedia annual notable events and deaths listing to generated most of the topics, e.g. http://en.wikipedia.org/wiki/2002. From a geographic point of view, this makes our evaluation seem to resemble GikiCLEF [7] the CLEF 2009 track which asked questions against a multilingual subset of Wikipedia.

Prior to posting final topics, five sample topics were posted on the GeoTime task web site and teams were solicited to suggest topics. Organizer Ray Larson indexed both the English and Japanese collections using his Cheshire system, and provided a search engine for testing topics against the collection. This engine (password protected) was made available to participating teams. In addition, a few topics were derived by adding a temporal component to an ACLIA geographic topic such as: Where did Hurricane Katrina make landfall?

Eventually the organizers created 25 topics in English which were translated into Japanese. Each of the 25 topics was vetted to hit at least one relevant document in both languages (the non-Japanese-speaking organizers used Google-Translate to translate the topic and run it against the Mainichi collection and translate and examine the top documents). Unfortunately, topic 17 (When and where was a candidate for president of a democratic South American country kidnapped by a rebel group?) was mistranslated from "South American" to "South African", so the Japanese results omitted this topic and are reported for 24 topics.

Four topics were of the form 'When and where did <person> die?' with one minor variation: GeoTime0007: *How old was Max Schmeling when he died, and where did he die?*

More discussion and evaluation of topic difficulty will follow the presentation of results.

4. PARTICIPATION

While a number of groups signed up to participate in NTCIR-GeoTime, many fewer submitted runs.

Japanese runs were submitted by the following 8 groups

Team Name	Organization
Anonymous	Anonymous submission
BRKLY	University of California, Berkeley
FORST	Yokohama National University, Japan
HU-KB	Hokkaido University, Japan
KOLIS	Keio University, Japan
Anon2	Anonymous submission
M	National Institute of Materials Science, Japan
OKSAT	Osaka Kyoiku University, Japan

English runs were submitted by the following 6 groups:

Team Name	Organization
BRKLY	University of California, Berkeley
DCU	Dublin City University, Ireland
IITH†	International Institute of Technology, Hyderabad
INESC	National Institute of Electroniques and Computer Systems, Lisbon, Portugal
UIOWA	University of Iowa
XLDB	University of Lisbon, Portugal

† Run submitted late, not included in pooling

Each group was allowed to submit up to 5 runs per target language. We encouraged the submission of bilingual runs, and while only BRKLY submitted such runs for JP→EN, three Japanese groups submitted EN→JP runs. The following table summarizes the number of runs submitted by each group:

Team	JA→JA	EN→EN	EN→JA	JA→EN
Anon	3			
BRKLY	3	3	2	2
DCU		5		
FORST	4			
HU-KB	5			
IITH		1		
INESC		5		
KOLIS	`5		4	
Anon2	2		2	
M	3			
OKSAT	1			
UIOWA		5		
XLDB		4		

5. EVALUATION

Relevance judging was done in a traditional manner on a pool of the top 100 documents retrieved from all runs with duplicates removed. Relevance assessment for Japanese was undertaken by a team at NII using the SEPIA system utilized for ACLIA and IR4QA. Because SEPIA was only available for non-English assessment, the third author developed a system at Technical University of Lisbon in Portugal and English assessment was done worldwide with assessors in Portugal, USA and NII. For the first time in NTCIR, participating teams joined in relevance assessment, similar to the participatory assessment done for XML retrieval in the INEX evaluations.[1] The BRKLY, INESC and UIOWA teams assessed topics, in addition to assessors from NII.

For Japanese GeoTime, 15,795 documents were examined and judged. For the English GeoTime, 17,423 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 were aggregated into a single category upon which the following result tables are based. We hope to have a more detailed analysis separating out the categories in the final paper.

Analysis of submitted runs was prepared by Tetsuya Sakai, using the same techniques used for analyzing IR4QA runs. For detail on the methodologies used, please refer to section 3 of the IR4QA overview [8].

6. APPROACHES

A wide variety of approaches were utilized by the different groups. The most conventional was BRKLY's baseline approach of only doing probabilistic ranking coupled with blind relevance feedback. This worked very well for English, but for Japanese it substantially underperformed the approaches by other teams which submitted Japanese runs. Several groups (DCU from Dublin City University, Ireland, IIT-H of Hyderabad, India, and XLDB of University of Lisbon) primarily utilized geographic enhancements (although XLDB did consult DBpedia as an external resource using a timestamp) and did not perform as well as groups which tackled the temporal qualities of the retrieval. The most straightforward of these geotemporal approaches was the KOLIS system of Keio University which merely counted the number of geographic and temporal expressions found in top-ranked documents of an initial search and then re-ranked based upon initial probability coupled with weighting of the counts. The FORST group of Yokahama University used question decomposition to separate out temporal from locational aspects of the topics in order to apply standard factoid question-answering techniques which work well on a single question type (when or where). Both HU-KB of Hokkaido University and the University of Iowa utilized a hybrid approach which combined probabilistic and (weighted) Boolean query formulation. A more elaborate approach was taken by the INESC group from Lisbon, Portugal who utilized a geographic resource (Yahoo PlaceMaker) for extracting geographic expressions and the TIMEXTAG[2] system for locating temporal expressions from within both topic and documents. Document processing was done at both the document and sentence level. Their hybrid approach relied upon the maximum amount of semantic content from the topic, so they utilized both description and narrative components from each topic.

7. RESULTS

7.1 English Results

For search against the English NYT collection, the six groups submitted 25 runs. Table 1a summarizes the results for English sorted by the mean performance over 25 topics showed for three performance measures, Average Precision (AP), Q, and normalized Discounted Cumulative Gain (nDCG). As can be seen from the table, the top performing runs were very close, but performance order differs depending upon metric. The top 10 runs are in identical order for AP and Q; however the order changes substantially when using the nDCG measure. For direct comparison of best results by team, we selected the best team result for description only runs, found in table 1b.

7.2 Japanese Results

For search against the Mainichi Japanese news collection, eight teams submitted runs whose performance is summarized in Table 2a. Table 2b provides best team performance using topic description only and omitting the narrative.

7.3 Topic Difficulty

We can also make an attempt to assess the difficulty of particular topics for both the English and Japanese collections. Figures 1 and 2 average the three performance measures over all submitted runs and plot this average by topic. The data are sorted by average precision in order to more clearly identify which topics presented the most challenge to successful search.

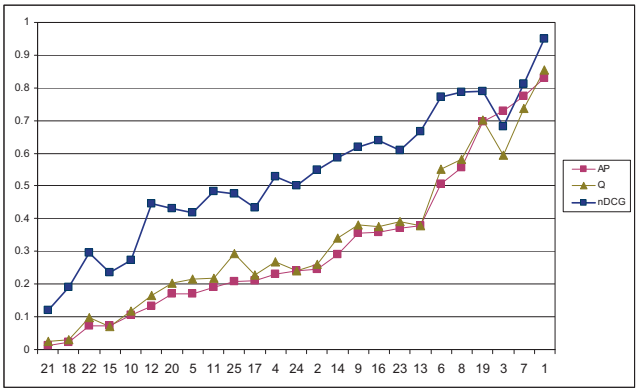

Figure 1: Per-topic AP, Q and nDCG averaged over 25 English runs for 25 topics (pool depth 100), sorted by topic difficulty (AP ascending)

[1] http://inex.is.informatik.uni-duisburg.de/

[2] http://ilps.science.uva.nl/resources/timextag

From the point of view of search of the English NYT collection, the four most difficult topics (less than 0.1 overall average precision) seem to be topic 15 (*What American football team won the Superbowl in 2002, and where was the game played?*), topic 18 (*What date was a country was invaded by the United States in 2002?*), topic 21 (*When and where were the 2010 Winter Olympics host city location announced?*) and topic 22 (When and where did a massive earthquake occur in December 2003?)

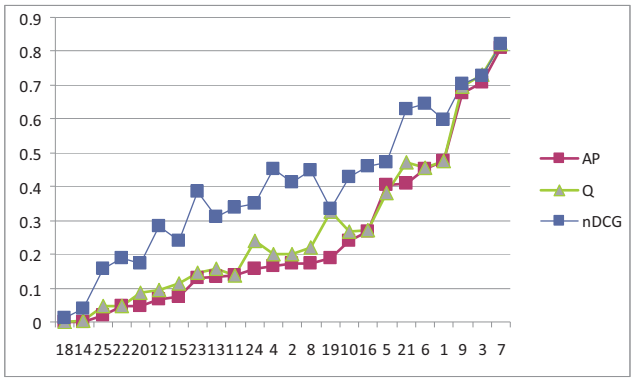

Figure 2: Per-topic AP, Q and nDCG averaged over 34 Japanese runs for 24 topics (pool depth 100), sorted by topic difficulty (AP ascending)

With respect to Japanese search of the Mainichi collection, several other topics (12, 14, and 25) also had average precision below 0.1 while topic 23 searches averaged 0.129.

7.4 Performance Variability across Topics

Another way to assess performance is to examine individual performance variability across topics. Such performance can be displayed by taking individual topic runs and finding the minimum, median and maximum performance for that topic. These are displayed in Figures 3 (English runs) and 4 (Japanese runs). While for nearly all Japanese topics, at least one group had a minimum precision of near zero for that topic, there was still a wide variability of performance from both minimum to median average precision for a topic, as well as from median precision to maximum precision for a topic. Where the median and maximum are very close, we can infer that almost all groups had good performance. An example for English where median and maximum are almost identical is topic 19: *When and where did the funeral of Queen Elizabeth (the Queen Mother) take place?* An example where the best run (UIOWA-EN-03-DN, maximum AP 0.7889) is considerably better than the median (0.177) is for topic 25: *How long after the Sumatra earthquake did the tsunami hit Sri Lanka?*

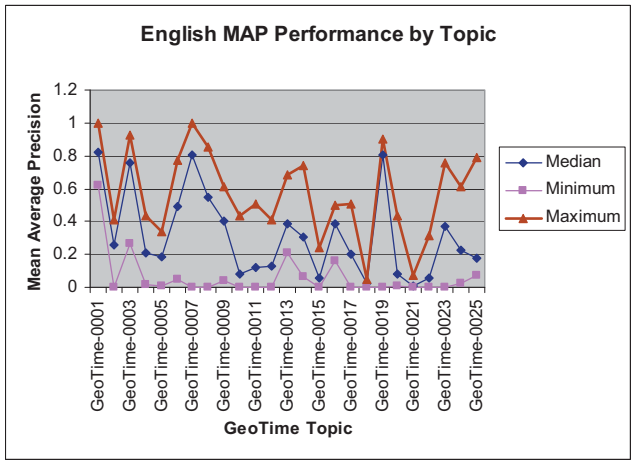

Figure 3: Per-topic AP showing Minimum, Median and Maximum performance for English runs

An example where median and maximum are almost identical (for Japanese) is topic 7: *How old was Max Schmeling when he died and where did he die?* Topic 19: *When and where did the funeral of Queen Elizabeth (the Queen Mother) take place?* which showed almost no variation between median and maximum for English, becomes, for Japanese, an example where the maximum precision (1.000, run FORST-JA-JA-02-D) is more than 7 times better than the median precision (0.1339).

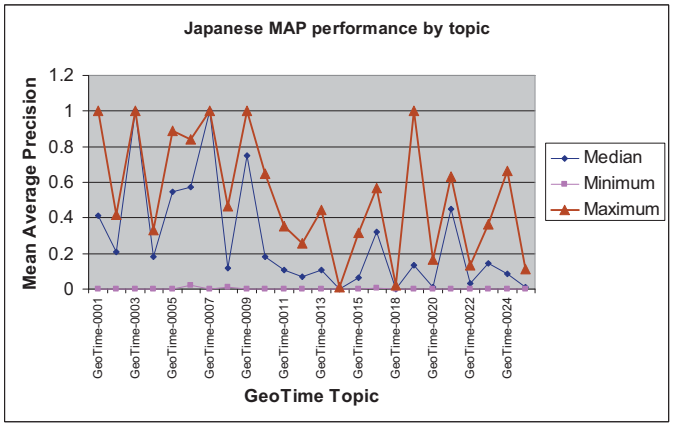

Figure 4: Per-topic AP showing Minimum, Median and Maximum performance for Japanese runs

Table 1a: GeoTime English mean performance for three performance metrics for 25 submitted runs

RUN	mean AP	RUN	mean Q	RUN	nDCG
BRKLY-JA-EN-01-DN	0.4158	BRKLY-JA-EN-01-DN	0.4287	INESC-EN-EN-05-DN	0.6246
BRKLY-EN-EN-02-DN	0.4045	BRKLY-EN-EN-02-DN	0.4197	UIOWA-EN-03-DN	0.6233
UIOWA-EN-01-D	0.3971	UIOWA-EN-01-D	0.4162	UIOWA-EN-01-D	0.6228
INESC-EN-EN-05-DN	0.3879	INESC-EN-EN-05-DN	0.4079	BRKLY-JA-EN-01-DN	0.617
UIOWA-EN-03-DN	0.38	UIOWA-EN-03-DN	0.3933	BRKLY-EN-EN-02-DN	0.6098
BRKLY-JA-EN-02-D	0.3759	BRKLY-JA-EN-02-D	0.3873	UIOWA-EN-04-DN	0.5931
UIOWA-EN-05-DN	0.3659	UIOWA-EN-05-DN	0.3834	UIOWA-EN-05-DN	0.5849
BRKLY-EN-EN-03-D	0.3615	BRKLY-EN-EN-03-D	0.3808	BRKLY-EN-EN-04-DN	0.5769
UIOWA-EN-02-D	0.3605	UIOWA-EN-02-D	0.3765	UIOWA-EN-02-D	0.5758
INESC-EN-EN-03-DN	0.352	UIOWA-EN-04-DN	0.3689	XLDB-EN-EN-02-T	0.5705
UIOWA-EN-04-DN	0.3517	INESC-EN-EN-03-DN	0.3640	XLDB-EN-EN-01-T	0.5701
BRKLY-EN-EN-04-DN	0.3390	XLDB-EN-EN-02-T	0.3584	INESC-EN-EN-03-DN	0.5641
XLDB-EN-EN-02-T	0.3354	BRKLY-EN-EN-04-DN	0.3556	BRKLY-JA-EN-02-D	0.5615
XLDB-EN-EN-01-T	0.3301	XLDB-EN-EN-01-T	0.3543	XLDB-EN-EN-03-T	0.5593
XLDB-EN-EN-03-T	0.3255	XLDB-EN-EN-03-T	0.3482	BRKLY-EN-EN-03-D	0.5566
DCU-EN-EN-02-D	0.3218	DCU-EN-EN-02-D	0.3413	DCU-EN-EN-02-D	0.5513
DCU-EN-EN-01-D	0.3207	DCU-EN-EN-01-D	0.3404	DCU-EN-EN-01-D	0.5506
XLDB-EN-EN-04-T	0.2978	XLDB-EN-EN-04-T	0.3205	XLDB-EN-EN-04-T	0.5325
DCU-EN-EN-03-D	0.2807	DCU-EN-EN-03-D	0.2991	DCU-EN-EN-03-D	0.5129
DCU-EN-EN-04-D	0.2491	DCU-EN-EN-05-D	0.2643	DCU-EN-EN-05-D	0.5042
DCU-EN-EN-05-D	0.241	DCU-EN-EN-04-D	0.2593	DCU-EN-EN-04-D	0.4843
INESC-EN-EN-02-DN	0.2328	INESC-EN-EN-02-DN	0.2338	INESC-EN-EN-04-DN	0.4234
INESC-EN-EN-04-DN	0.2139	INESC-EN-EN-04-DN	0.2223	INESC-EN-EN-02-DN	0.4056
IIIT-H	0.154	INESC-EN-EN-01-DN	0.1536	INESC-EN-EN-01-DN	0.2961
INESC-EN-EN-01-DN	0.137	IIIT-H	0.1447	IIIT-H	0.2224

Table 1b: GeoTime English best team performance for description only runs*

RUN	AP	RUN	Q	RUN	nDCG
UIOWA-EN-01-D	0.3971	UIOWA-EN-01-D	0.4162	UIOWA-EN-01-D	0.6228†
BRKLY-JA-EN-02-D	0.3759	BRKLY-JA-EN-02-D	0.3873	XLDB-EN-EN-02-T	0.5705
XLDB-EN-EN-02-T	0.3354	XLDB-EN-EN-02-T	0.3584	BRKLY-JA-EN-02-D	0.5615
DCU-EN-EN-02-D	0.3218‡	DCU-EN-EN-02-D	0.3413‡	DCU-EN-EN-02-D	0.5513‡
IIIT-H	0.154	IIIT-H	0.1447	IIIT-H	0.2224

***INESC team omitted because no description-only run submitted**

† statistically significant difference (α=0.05) from the value of the run in the next row

‡ statistically significant difference (α=0.01) from the value of the run in the next row

Table 2a: GeoTime Japanese mean performance for three performance metrics for 34 submitted runs

RUN	mean AP	RUN	mean Q	RUN	mean nDCG
HU-KB-JA-JA-02-DN	0.3867	HU-KB-JA-JA-02-DN	0.4268	HU-KB-JA-JA-03-D	0.5881
HU-KB-JA-JA-03-D	0.3719	HU-KB-JA-JA-03-D	0.4162	HU-KB-JA-JA-04-D	0.5717
HU-KB-JA-JA-01-D	0.3697	HU-KB-JA-JA-01-D	0.4117	HU-KB-JA-JA-01-D	0.571
HU-KB-JA-JA-04-D	0.3627	HU-KB-JA-JA-04-D	0.4078	HU-KB-JA-JA-02-DN	0.5685
KOLIS-JA-JA-04-D	0.325	KOLIS-JA-JA-04-D	0.3544	KOLIS-JA-JA-04-D	0.5159
KOLIS-EN-JA-04-D	0.3145	KOLIS-EN-JA-04-D	0.3468	KOLIS-JA-JA-05-DN	0.5095
KOLIS-JA-JA-03-D	0.3139	KOLIS-JA-JA-03-D	0.3459	KOLIS-JA-JA-03-D	0.5063
KOLIS-JA-JA-05-DN	0.3027	KOLIS-JA-JA-05-DN	0.3392	KOLIS-JA-JA-02-D	0.5036
KOLIS-JA-JA-02-D	0.3008	KOLIS-JA-JA-02-D	0.3378	HU-KB-JA-JA-05-D	0.4993
KOLIS-EN-JA-03-D	0.2918	KOLIS-EN-JA-03-D	0.3329	KOLIS-JA-JA-01-D	0.4982
HU-KB-JA-JA-05-D	0.2881	KOLIS-JA-JA-01-D	0.3327	KOLIS-EN-JA-04-D	0.4956
KOLIS-JA-JA-01-D	0.2878	HU-KB-JA-JA-05-D	0.3282	KOLIS-EN-JA-03-D	0.4817
KOLIS-EN-JA-02-D	0.287	KOLIS-EN-JA-02-D	0.3277	KOLIS-EN-JA-02-D	0.4765
FORST-JA-JA-02-D	0.2858	KOLIS-EN-JA-01-D	0.3232	KOLIS-EN-JA-01-D	0.4729
KOLIS-EN-JA-01-D	0.2773	FORST-JA-JA-04-D	0.2865	Anon2-EN-JA-01-T	0.4231
FORST-JA-JA-04-D	0.2762	FORST-JA-JA-02-D	0.2842	Anon2-JA-JA-01-T	0.4045
M-JA-JA-03-D	0.2672	M-JA-JA-03-D	0.2835	BRKLY-JA-JA-01-DN	0.4034
BRKLY-JA-JA-01-DN	0.2472	Anon2-EN-JA-01-T	0.2763	M-JA-JA-03-D	0.3982
M-JA-JA-01-D	0.2472	M-JA-JA-01-D	0.2719	M-JA-JA-02-D	0.3806
Anon2-EN-JA-01-T	0.2379	Anon2-JA-JA-01-T	0.2699	FORST-JA-JA-04-D	0.3772
Anon2-JA-JA-01-T	0.2332	M-JA-JA-02-D	0.2619	M-JA-JA-01-D	0.3766
FORST-JA-JA-01-D	0.233	BRKLY-JA-JA-01-DN	0.2603	FORST-JA-JA-02-D	0.372
M-JA-JA-02-D	0.2305	FORST-JA-JA-01-D	0.2593	BRKLY-JA-JA-03-DN	0.3634
FORST-JA-JA-03-D	0.2056	FORST-JA-JA-03-D	0.2379	FORST-JA-JA-01-D	0.332
BRKLY-JA-JA-03-DN	0.1926	OKSAT-JA-JA-01-D	0.2055	FORST-JA-JA-03-D	0.3244
OKSAT-JA-JA-01-D	0.1835	BRKLY-JA-JA-03-DN	0.2042	BRKLY-EN-JA-01-DN	0.3221
BRKLY-EN-JA-01-DN	0.1788	BRKLY-EN-JA-01-DN	0.1942	OKSAT-JA-JA-01-D	0.3138
BRKLY-JA-JA-02-D	0.1726	BRKLY-JA-JA-02-D	0.1819	BRKLY-JA-JA-02-D	0.3014
Anon-JA-JA-02-UNK	0.1668	Anon-JA-JA-02-UNK	0.1637	BRKLY-EN-JA-02-D	0.2488
Anon-JA-JA-03-UNK	0.1557	BRKLY-EN-JA-02-D	0.1585	Anon2-EN-JA-02-T	0.2343
Anon-JA-JA-01-UNK	0.1474	Anon-JA-JA-03-UNK	0.1559	Anon2-JA-JA-02-T	0.2107
BRKLY-EN-JA-02-D	0.1465	Anon-JA-JA-01-UNK	0.1472	Anon-JA-JA-02-UNK	0.2085
Anon2-EN-JA-02-T	0.0776	Anon2-JA-JA-02-T	0.1033	Anon-JA-JA-01-UNK	0.1983
Anon2-JA-JA-02-T	0.0766	Anon2-EN-JA-02-T	0.1023	Anon-JA-JA-03-UNK	0.1963

Table 2b: GeoTime Japanese best team performance for description only runs

RUN	AP	RUN	Q	RUN	nDCG
HU-KB-JA-JA-03-D	0.3719	HU-KB-JA-JA-03-D	0.4162†	HU-KB-JA-JA-03-D	0.5881†
KOLIS-JA-JA-04-D	0.325	KOLIS-JA-JA-04-D	0.3544	KOLIS-JA-JA-04-D	0.5159†
FORST-JA-JA-02-D	0.2858	FORST-JA-JA-04-D	0.2865	Anon2-EN-JA-01-T	0.4231
M-JA-JA-03-D	0.2672	M-JA-JA-03-D	0.2835	M-JA-JA-03-D	0.3982
Anon2-EN-JA-01-T	0.2379	Anon2-EN-JA-01-T	0.2699	FORST-JA-JA-04-D	0.3772
OKSAT-JA-JA-01-D	0.1835	OKSAT-JA-JA-01-D	0.2055	OKSAT-JA-JA-01-D	0.3138
BRKLY-JA-JA-02-D	0.1726	BRKLY-JA-JA-02-D	0.1819	BRKLY-JA-JA-02-D	0.3014
Anon-JA-JA-02-UNK	0.1668	Anon-JA-JA-02-UNK	0.1637	Anon-JA-JA-02-UNK	0.2085

† statistically significant difference (α=0.05) from the value of the run in the next row

8. DISCUSSION

NTCIR-GeoTime was the first attempt at evaluating geotemporal information retrieval. While Geographic Information Retrieval has had numerous evaluations, the addition of a temporal component has proven very challenging to participants, especially if the topic (question) can be misinterpreted by the automated retrieval process (as in the case of topic 21: *When and where were the 2010 Winter Olympics host city location announced?*) or require a list answer which is time varying (topic 16: *When and where were the last three Winter Olympics held?*). Teams which relied exclusively on geographic enhancements did not perform as well as those which incorporated some temporal expression processing within their methodologies. Questions remain as to why there was so much performance variability across document collection language (Japanese and English) for the same topics.

9. ACKNOWLEDGMENTS

We thank participant assessors Christopher Harris (University of Iowa), Krishna Janakiraman (UC Berkeley), Ricardo Vaz and Flávio Esteves (Technical University of Lisbon).

10. REFERENCES

[1] S Asadi, C.-Y. Chang, X. Zhou, and J. Diederich. Searching the world wide web for local services and facilities: A review on the patterns of location-based queries. In W. Fan, Z. Wu, and J. Yang, editors, WAIM2005, pages 91–101. Springer LNCS 3739, 2005.

[2] L L Hill, *GeoReferencing: The Geographic Associations of Information*, MIT Press, Cambridge, MA 2006.

[3] C. B. Jones, A. I. Abdelmoty, D. Finch, G. Fu, and S. Vaid. The SPIRIT spatial search engine: architecture, ontologies and spatial indexing. In GiScience 2004, Oct. 2004, Adelphi, MD, pages 125–139, 2004. Cunningham,

[4] Larson, R Geographic information retrieval and spatial browsing. In GIS and Libraries: Patrons, Maps and Spatial Information, pages 81–124. UIUC - GSLIS, Urbana-Champaign, IL, 1996.

[5] I. Mani, J. Pustejovsky, and B. Sundheim. Introduction to the special issue on temporal information processing. ACM Transactions on Asian Language Information Processing (TALIP), 3(1):1–10, 2004.

[6] R. Purves, C. Jones, and P. Clough. GIR'10: 6th workshop on geographic information retrieval, 2010. http://www.geo.unizh.ch/ rsp/gir10/index.html.

[7] D Santos, L Cabral, GikiCLEF: Crosscultural issues in an international setting: asking non-English-centered questions to Wikipedia, CLEF 2009 Working Notes, http://www.clef-campaign.org/2009/working_notes/Santos-paperCLEF2009.pdf. September 2009, 21pp.

[8] Sakai, T et al Overview of NTCIR-8 ACLIA IR4QA, In this proceedings.