An Evaluation of Geographic and Temporal Search

Fredric Gey†, Noriko Kando‡, Ray Larson†
†University of California, Berkeley USA
‡National Institute of Informatics, Tokyo JAPAN

gey@berkeley.edu , ray@ischool.berkeley.edu , kando@nii.ac.jp,

Abstract

This paper summarizes parts of the NTCIR-Geo-Time Task held in Tokyo June 15-18, 2010. This task was the first evaluation specifically of search with both Geographic and Temporal constraints, i.e. it combines geographic information retrieval (GIR) with time-based search to find specific events in a multilingual collection. We describe the data collections (Japanese and English news stories), topic development, research approaches, assessment results and lessons learned from the evaluation.

1 Introduction

Semantic search queries and factoid questions require semantic processing to deliver results beyond bag-of-words search. Geo-temporal search concerns search which has both geographic and temporal constraints. In particular the search for events or to answer questions about events contains, often, specificity of location (where) and specifity of time (when). A simple example might be "When and where did Rosa Parks die?" in which the user wishes to know a "specific" date (is it year or month-year?) and "specific" location (should it be city?) to answer the question. A more complex question "How long after the Sumatra earthquake did its tsunami hit Sri Lanka?" has geographic constraints and wishes to extract a somewhat specific temporal expression (e.g. "a few hours") from the document collection being searched. The above examples are taken directly from NTCIR-GeoTime, the first evaluation of geo-temporal search recently presented (mid-June 2010) at the eighth NTCIR Workshop in Tokyo. The results clearly demonstrated that semantic markup for geography and time outperformed traditional IR methodologies.

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. 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". [Larson 1996, Jones et al 2004]. It has been estimated that 22 percent of web searches are location based [Asadi et al 2005]. 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 [Hill 2006]. However, in order for place names to support a spatial approach to GIR, they must be associated with a model of geographic space. There have been over six workshops [Purves and Clough 2010] 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 and GikiCLEF tracks [Mandl et al 2008,Santos et al 2010]). But, until this track 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).

The temporal aspects of search have been largely ignored in the IR community, but not in the GIS and computational linguistics communities. There has been a special issue of ACM TALIP on 'Temporal Information Processing' [Mani, Pustejovsky and Sundheim 2004], as well as at least two workshops on "Temporal and Spatial Information Processing". Use of temporal information in web search and results presentation (hit clustering) was explored in [Alonso, Gertz and Baeza-Yates 2007]. The NTCIR-GeoTime task organizers wanted to utilize and incorporate past research on this aspect as part of the evaluation.

2 Data and Test Topics

Two news story collections were used for NTCIR-GeoTime, one Japanese and one English. The Japanese collection was Mainichi newspapers for 2002-2005, which had 377,941 documents. The English collection, 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. Details about these collections and their characteristics may be found in the

GeoTime Overview [Gey et al 2010]. The collections matched those used in other tasks in NTCIR-8 (Advanced Cross-Language Question Answering [Mitamura et al 2010] and Multilingual Opinion Tracking [Seki et al 2010]).

Using Wikipedia as the 'ground truth', the organizers created 25 topics in English , phrased as questions, from the annual notable events summary.[1] 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. The process of topic development is also discussed in the Overview [Gey et al 2010]. 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?* Another question was looking for a fixed list – GeoTime0016: *When and where were the last three Winter Olympics held?* Another, similar question – GeoTime0021: *When and where were the 2010 Winter Olympics host city location announced?* was very difficult because it wanted to know where (Prague, Czech Republic at the 115th session, July 2, 2003) the IOC (International Olympic Committee) announced that Vancouver would host the 2010 Winter games. In the opinion of the organizers, the most difficult topic was expected to be GeoTime0025: *How long after the Sumatra earthquake did the tsunami hit Sri Lanka?* This did prove to be one of the difficult topics, but not necessarily the most difficult. Topics were formatted in XML structures containing a description field and a more extensive narrative field, in both English and Japanese, as in:

```
<TOPIC ID="GeoTime-0001">
<DESCRIPTION LANG="EN">When and where did Astrid Lindgren die?</DESCRIPTION>
<DESCRIPTION LANG="JA">いつ、どこでアストリッド・リンドグレーンは亡くなりましたか？</DESCRIPTION>
<NARRATIVE LANG="EN">The user wants to know when and in what city the children's author Astrid Lindgren died.</NARRATIVE>
 <NARRATIVE LANG="JA">ユーザは、いつ、どの都市で、児童書作家のアストリッド・リンドグレーンが死亡したかを知りたいと思っている。</NARRATIVE>
</TOPIC>
```

The full set of topics may be found at:
http://metadata.berkeley.edu/NTCIR-GeoTime/topics.php

3 Evaluation and Results

An evaluation run consisted of a ranked list of up to 1000 documents for each topic. Relevance judging was done in a traditional manner on a pool of the top 100 documents retrieved from all runs with duplicates removed.

3.1 Teams Submitting Evaluation Runs

Six teams submitted runs for the English collection and five registered teams ran the 25 topics against the Japanese collection (three other groups agreed to submit runs to broaden the pool – two of these groups are labeled 'anonymous' below).

Team Name	Organization submitting English runs
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

Table 1: Groups Submitting English Runs

Team Name	Organization submitting Japanese runs
Anon1	Anonymous submission 1
BRKLY	University of California, Berkeley
FORST	Yokohama National University, Japan
HU-KB	Hokkaido University, Japan
KOLIS	Keio University, Japan
Anon2	Anonymous submission 2
M	National Institute of Materials Science, Japan
OKSAT	Osaka Kyoiku University, Japan

Table 2: Groups Submitting Japanese Runs

The English groups submitted a total of 25 runs (a maximum of 5 different runs per team were allowed) and the Japanese groups submitted 34 distinct runs.

3.2 Results

Results in [Gey et al 2010] are displayed using three relatively well-established evaluation measures: Average Precision (AP), Q Measure, and Normalized Discounted Cumulative Gain (nDCG). Details about these evaluation measures which were also used for the IR4QA (Information Retrieval for Question-Answering) task of NTCIR-8 may be found in [Sakai et al 2010]. For simplicity we only display the nDCG results in the following table to show relative performance. A run is specified by team-name-topic-language-document-language-run_number-D or DN where D means description only which DN means description and narrative were used from the topic (the IIIT submission did not specify which fields were used)

RUN	nDCG
INESC-EN-EN-05-DN	0.6246
UIOWA-EN-EN-01-D	0.6228
BRKLY-JA-EN-01-DN	0.617
XLDB-EN-EN-02-T-DN	0.5705
DCU-EN-EN-02-D	0.5513‡
IIIT-H-EN-EN	0.2224

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

Table 3: Best GeoTime English Run per Team

The most interesting result from this table is that Berkeley had better cross-lingual performance than its monolingual runs. This phenomenon appears occasion-

[1] e.g. http://en.wikipedia.org/wiki/2002

ally in Cross-Language Information Retrieval when blind feedback obtains additional discriminating terms from the top retrieved documents of an initial retrieval (BRKLY used blind feedback as a baseline [Larson 2010] without geotemporal extensions)..

Another way to compare performance is to fix the run type, for example to compare runs which all teams used only the D (description) part of the topic in their runs. The following table compares description only runs against the Japanese collection .

RUN	nDCG
HU-KB-JA-JA-03-D	0.5881†
KOLIS-JA-JA-04-D	0.5159†
Anon2-EN-JA-01-T	0.4231
M-JA-JA-03-D	0.3982
FORST-JA-JA-04-D	0.3772
OKSAT-JA-JA-01-D	0.3138
BRKLY-JA-JA-02-D	0.3014
Anon-JA-JA-02-UNK	0.2085

Table 4: Best Japanese D Run per Team (nDCG)

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

The interesting thing to immediately observe is that BRKLY which did so well in English runs comes in at a relatively low performance using the same blind feedback methodology as for English. Indeed, if we further exclude the anonymous runs (including M) for which we have no methodology , Berkeley's performance is worst among official Japanese runs. The reason for this has yet to emerge, however, all the other Japanese groups except OKSAT utilized sophisticated geotemporal processing in their approaches to retrieval.

4 Technical Approaches to Geo-Temporal Retrieval

In this section we review the technical approaches taken by the best performing teams.

4.1 English 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.

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[1] system from the University of Amsterdam 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. University of Iowa utilized a hybrid approach which combined probabilistic and (weighted) Boolean query formulation.

4.2 Japanese Approaches

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. HU-KB of Hokkaido University , similarly to the University of Iowa for English, also combined probabilistic and Boolean query formulation [Mori 2010]. However, in the case of Hokkaido, the Boolean approach was utilized to filter out unwanted documents from the probabilistic ranking. In order to deal with the Boolean tendency to return the null set, HU-KB expanded the vocabulary using a synonym thesaurus. The FORST group of Yokahama University [Yoshioka 2010] 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). While KOLIS utilized a custom gazetteer of place names and a fixed list of temporal expressions (not including day-of-week), the Hokkaido [HU-KB] approach used the Cabocha system for named entity tagging [Kudo and Matsumoto 2002].

5 Topic Difficulty

There are two methods of assessing topic difficulty: looking at average performance over all runs by topic – the topics with low average precision are assumed to be the most difficult. The other way is to examine differences between median performance and maximum performance – this can demonstrate that particular methods perform better for such topics.

5.1 Topic Difficulty by Average Precision

Figures 1 and 2 average the three performance measures over all submitted runs and plot this average by topic.

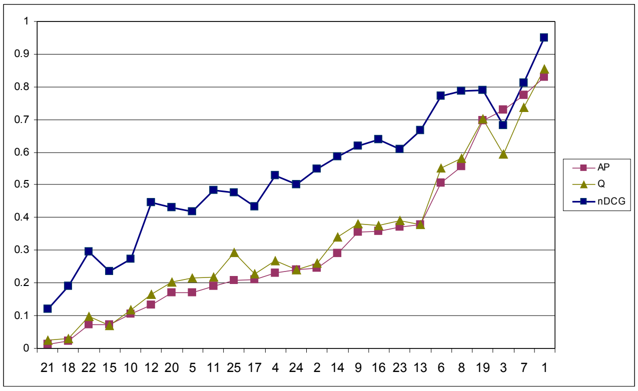

Figure 1: Per-topic AP, Q and nDCG averaged over 25 English runs for 25 topics (pool depth 100), sorted by topic difficulty (AP ascending)

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

The data are sorted by average precision in order to more clearly identify which topics presented the most challenge to successful search.

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?*)

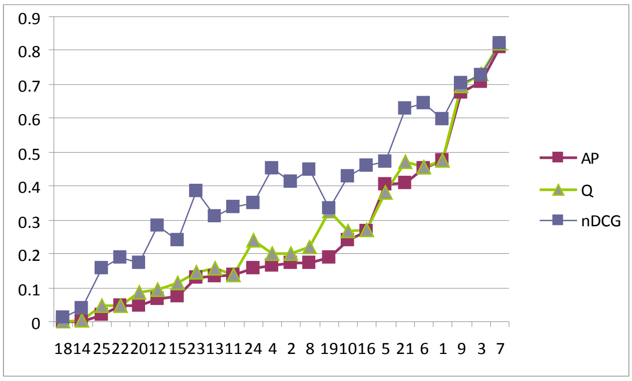

Figure 2: Per-topic AP, Q and nDCG averaged over 34 Japanese runs for 24 topics (pool depth 100), sorted by topic difficulty (AP ascending)

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. Topic 12 is *When and where did Yasser Arafat die?,* Topic 14 is *When and where did a volcano erupt in Africa during 2002?,* Topic 23 is *When did the largest expansion of the European Union take place, and which countries became members?,* and Topic 25, the one predicted by the organizers to be difficult: *How long after the Sumatra earthquake did the tsunami hit Sri Lanka?*

5.2 Median/Maximum Topic Peformance

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.

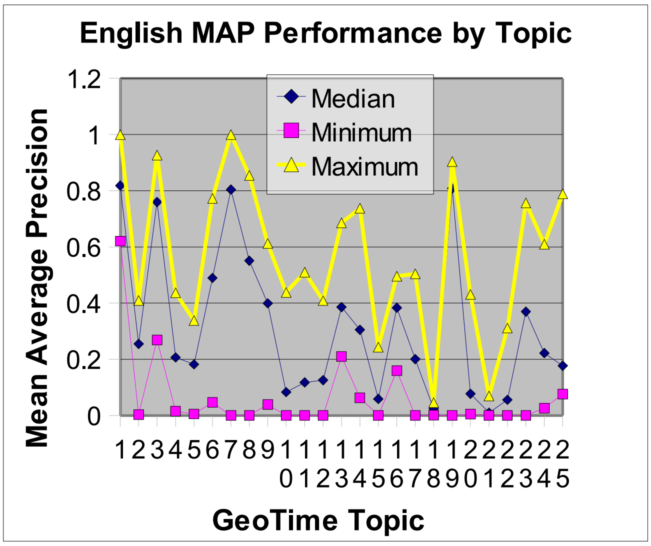

Figure 3: Per-topic AP showing Minimum, Median and Maximum performance for English runs

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 more than four times better than the median (0.177) is for topic 25: *How long after the Sumatra earthquake did the tsunami hit Sri Lanka?*

An example (for Japanese) where median and maximum are almost identical is topic 7: *How old was Max Schmeling when he died and where did he die?* On the other hand, topic 19, 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).

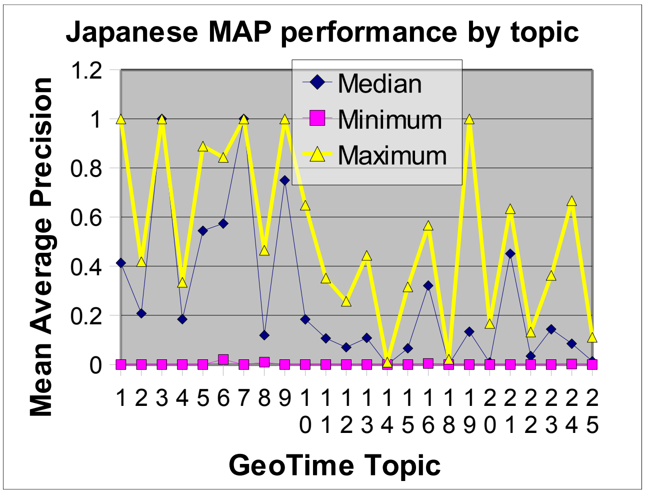

Figure 4: Per-topic AP showing Minimum, Median and Maximum performance for Japanese runs

It is worth noting that the minimum for Japanese was a single run in which the team did very poorly on all topics. It should probably be considered an outlier and removed from future analysis. The median performance is a more reliable statistic from which to draw conclusions.

6 Judgment Approaches for Imprecise Temporal Expressions

One of the difficulties in relevance judgment is how to approach the extreme variability in temporal expressions in text and how to approach judgment, particularly with respect to these expressions. As a point of reference, each document had a specific date upon which it was published. At least for English relevance judgments, imprecise expressions relating to that date were seen as sufficient evidence to judge a document relevant. For example if a document stated "Katherine Hepburn died Wednesday in her home in Connecticut" it was assumed that sophisticated natural language processing could infer the exact date of death from the date of the document. If a document stated (for topic 25) that "*a few hours later* the Sumatra earthquake tsunami hit the coast of Sri Lanka" the document could be judged relevant. Finally, we retrospectively realized that a topic needs to be **date stamped** if it asks a temporally relative question. For example topic 16: *When and where were the last three Winter Olympics held?* was formulated before the 2010 winter Olympics were held in Vancouver. Thus while documents could have known that the 2010 Winter Olympics were to be held in Vancouver, the correct answer (for a topic date-stamped before 2010) would be 1998 (Nagano, Japan), 2002 (Salt Lake City, USA), and 2006 (Turin, Italy).

7 Discussion

7.1 Lessons Learned

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.

7.2 Future Directions

Plans are already being formulated for a second GeoTime evaluation for the NTCIR-9 Workshop in 2011. We are exploring additional languages – Korean and Chinese to the document collection set. For participant groups we will make available a standard set of resources (gazetteers, named entity taggers, TimexTag, etc). In addition, we have a definite desire to evaluate location-based and map-based search simulation, i.e. "What event is happening "here" and "now/tomorrow"" -- where here and now come from the included latitude/longitude coordinates. This should facilitate innovative result visualization using Google/MS Earth/map as well as map-based querying (bounding rectangles).

8 Acknowledgments

We thank participant assessors Christopher Harris (University of Iowa), Krishna Janakiraman (UC Berkeley), Ricardo Vaz and Flávio Esteves (Technical University of Lisbon). Appreciation is also due to Jorge Machado (INESC, Lisbon, Portugal) who programmed the English assessment system, and Tetsuya Sakai (Microsoft Research, Asia) who ran the evaluation statistics. The work of the first author was supported by a Visiting Researcher travel grant from the National Institute of Informatics of Japan during June-July 2010. We thank the LWA/WIR reviewers for suggesting improvements and corrections to the original submission.

9 References

[Alonso, Gertz and Baeza-Yates 2007] On the Value of Temporal Information in Information Retrieval, SIGIR Forum, Vol. 41 No. 2 December 2007, pp 35-41.

[Asadi *et al*., 2005] 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, pp. 91–101. Springer LNCS 3739, 2005.

[Gey et al 2010] F. Gey, R. Larson, N. Kando, J. Machado and T. Sakai, NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search, In Proceedings of the 8th NTCIR Workshop Meeting , Tokyo Japan June 15-18, 2010, ISBN: 978-4-86049-053-9.

[Harris 2010] C. Harris, Geographic Information Retrieval Involving Temporal Components, in Proceedings of the 8th NTCIR Workshop Meeting.

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

[Jones et al 2004] 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

[Kishida 2010] K. Kishida, Vocabulary-based Re-ranking for Geographic and Temporal Searching at NTCIR GeoTime Task, in Proceedings of the 8th NTCIR Workshop Meeting .

[Kudo and Matusomoto 2002] T. Kudo and Y. Matsumoto, Japanese Dependency Analysis using Cascaded Chunking, in CoNLL 2002, Taipei.

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

[Larson 2010] R Larson, Text Retrieval Baseline for NTCIR-GeoTime, in Proceedings of the 8th NTCIR Workshop Meeting .

[Machado, Borbinha and Martins 2010] J. Machado, J. Borbinha and B. Martins, Experiments with Geo-Temporal Expressions Filtering and Query Expansion at Document and Phrase Context Resolution, In Proceedings of the 8th NTCIR Workshop Meeting .

[Mani, Pustejovsky and Sundheim 2004] 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

[Mandl et al 2008] T. Mandl, F. Gey, G. Di Nunzio, N. Ferro, M. Sanderson, D. Santos and C. Womser-Hacker, An Evaluation Resource for Geographic Information Retrieval, In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08) Morocco, May, 2008

[Mitamura et al 2010] T Mitamura, H Shima, T Sakai, N Kando, T Mori, K Takeda, C-Y Lin, R Song, C-J Lin, C-W Lee, Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access, in Proceedings of the NTICIR Workshop 8, Tokyo Japan June 15-18, 2010.

[Mori 2010] T. Mori, A Method for GeoTime Information Retrieval based on Question Decomposition and Question Answering, in Proceedings of the 8th NTCIR Workshop Meeting.

[Purves and Clough 2010] 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.

[Santos et al 2010] D. Santos, L. Cabara, et al, GikiCLEF: Crosscultural Issues in Multilingual Information Access in Proceedings of LREC 2010, Malta, May 2010.

[Sakai et al 2010] T. Sakai, H. Shima, N. Kando, R. Song, C-J. Lin, T. Mitamura, M. Sugimito, C-W. Lee Overview of NTCIR-8 ACLIA IR4QA, in Proceedings of the 8th NTICIR Workshop Meeting, Tokyo Japan June 15-18, 2010.

[Seki et al 2010] Y. Seki, L-W Ku, L. Sun, H-H. Chen and N. Kando,, Overview of Multilingual Opinion Analysis Task at NTCIR-8: A Step Toward Cross Lingual Opinion Analysis, in Proceedings of the 8th NTICIR Workshop Meeting, Tokyo Japan June 15-18, 2010.

[Yoshioka 2010] M. Yoshioka, A Method for GeoTime Information Retrieval based on Question Decomposition and Question Answering, In Proceedings of the 8th NTICIR Workshop Meeting, Tokyo Japan June 15-18, 2010.