The NLP Task at INEX 2004

Shlomo Geva
Faculty of Information Technology
Queensland University of Technology
Brisbane, Australia
s.geva@qut.edu.au

Tony Sahama
Faculty of Information Technology
Queensland University of Technology
Brisbane, Australia
t.sahama@qut.edu.au

Abstract

The INEX workshop is concerned with Evaluating the effectiveness of XML retrieval systems. In 2004 a natural language query task was added to the INEX Ad hoc track. Standard INEX Ad hoc topic titles are specified in NEXI – a simplified and restricted subset of XPath, with a similar feel, and yet with a distinct IR flavour and interpretation. The syntax of NEXI is rigid and it imposes some limitations on the kind of information need that it can faithfully capture. At INEX 2004 the NLP question to be answered was simple – is it practical to use a natural language query that is the equivalent of the formal NEXI title? The results of this experiment are reported and some information on the future direction of the NLP task is presented.

1 Introduction

In 2004 a natural language query task was added to the INEX Ad hoc track. Standard INEX Ad hoc topic titles are specified in NEXI – a simplified and restricted subset of XPath with a distinct IR flavour and interpretation. The syntax of NEXI is rigid and it imposes limitations on the kind of information need that it can faithfully capture. However, it is expressive enough to allow the formulation of fairly complex content and structure constraints that are intended to capture an information need.

NEXI is not intended as an end-user query language. The Natural Language processing task at INEX 2004 focused on the question of whether it is possible to use an alternative natural language expression of the title – a natural language query (NLQ.) This task is simpler than the more ambitious and perhaps obvious task of directly interpreting the information need expression in the narrative. This is indeed a longer term goal for the NLP track. At INEX 2004 the question to be answered was simpler – is it practical to search on the basis of a natural language expression that is the equivalent of the formal NEXI title? Of course the advantage of this approach is in lifting the strict syntactic constraints that NEXI imposes, and the need for exact knowledge of the XML document structure, while simplifying the task of interpreting natural language expressions by significantly narrowing the context and the form of natural language expressions.

For example, consider the NEXI title element:

```
//article[about(.//abs, XML information retrieval)]//fm[about(.//au, A B Smith)]
```

Equivalent (more or less) English expressions might be:

"Return the front matter of articles about XML information retrieval where the author is A B Smith"

Or

"Find the general details of articles on XML information retrieval that were written by A B Smith"

Of course there are many other alternative English expressions that might correspond acceptably well to the NEXI expression. Importantly, it is much easier to form a natural language expression than it is to form the NEXI expression. This was evidenced by the large number of initial errors in topic formulation, and by the many rounds of corrections in 2003 and in 2004. It was left to the authors of the INEX topics to formulate the description element of a topic as a natural language expression of the formal NEXI title (or vice versa). There was no restriction on the use of natural language. No specific names of XML element were specified (end users need not and may not know that <fm> stands for front matter, or that <au> stands for author details.) However, end users might have some reasonable general knowledge about the structure of documents in a collection, even if not the specific DTD or XML Schema, or knowledge of the NEXI syntax and semantics, with which to specify NEXI queries. In preparing topics for the 2004 INEX NLQ task some mild limitations on the formulation of the description were imposed. The topics were reviewed and corrected where required - by the original authors - so that the content and structure constraints of the title appeared in some form in the description. The challenge to the NLP systems was to interpret the natural language description and to formulate a suitable NEXI query to produce the desired results.

Assessment of retrieval runs for topics was, as usual, based on the information need. Since both the Ad hoc track and the NLQ track used the same topics, it was possible to use the Ad hoc track assessments in evaluating the NLQ retrieval runs. This allowed us to compare the performance of systems that used the standard NEXI title, to find relevant elements, and the systems that used an equivalent natural language description. This ability to compare with the Ad hoc results was very important in providing a baseline for the assessment of the NLQ approach. The Ad hoc track at INEX is relatively mature. The performance of the NLQ systems in isolation might have been meaningless in the absence of a robust baseline – particularly given the small number of submissions.

2 INEX 2004 Approaches and Results

Following the INEX 2004 CFP there were 9 groups who registered for participation. By the end of the process only 4 run submissions were received from participating groups: Ecoles des Mines de Saint-Etienne, IBM Haifa Research, The Robert Gordon University, and Queensland University of Technology. The groups whose systems achieved top ranking in the Ad hoc track with content only task (CO), and the content and structure task (CAS), participated in the NLQ track. This allowed us to obtain a comparison of performance with the best baseline available – that of the highest ranked systems at the INEX 2004 Ad hoc track.

The CO task is somewhat similar to traditional IR in text collections, but with a difference in the kind of results returned. A set of keywords is specified and used to identify results. However, rather than identify relevant documents, the objective of the CO track is to identify suitable units of retrieval. Elements are assessed for relevance along two dimensions – exhaustivity and specificity. A relevant element may be highly exhaustive if it satisfies the information need comprehensively, or it may be moderately or mildly exhaustive if it satisfies the information need partially. A relevant element may be

highly specific if it is focused on the information need, or it may be moderately or mildly specific if it also contains other irrelevant material. The objective is to return elements of optimal size.

The approach of the participants to the NLQ CO task was rather simple – the query was specified in the description element of a topic. Systems only needed to extract relevant search keywords from the description. After that the topic could be handled in the same manner as Ad hoc track CO titles which specified a set of keywords, but without the verbosity of natural language. So it was just a matter of filtering the "noise" from the natural language expressions. The IBM Haifa submission topped the ranking in the Ad hoc CO task and it also topped the NLQ CO task. The average precision of this submission dropped from 0.144 when using the NEXI title to 0.129 when using the natural language description. Interestingly though, had this NLQ submission been entered in the Ad hoc track, it would have achieved 3rd position overall, from 70 Ad hoc submissions.

The CAS task was trickier. The title specified not only keywords, but also structural information (where to look, what to return). All such information was considered as a set of hints to the search engine, much in the same way that keywords are taken as hints in text retrieval. Here the NLQ search engine had to not only filter "noise" from verbose queries, but to also differentiate between search terms (keywords) and structural cues (tags). Furthermore, the structural cues were in natural language and had to be somehow mapped into XML tags so that they could be used by the XML search engine.

The approach taken by QUT was to translate the description into a NEXI expression by distilling the structural and content constraints from the natural language description and forming a NEXI expression. Then the search was performed by the same search engine that was designed for the Ad hoc track, and operated on NEXI expressions. The translation was based on part of speech analysis, on question templates, and on a dictionary to map from possible English expressions to tag names. Care was taken not to bias the NLQ analysis towards the 2004 topics. The 2004 topics had to remain previously unseen when retrieval runs were performed. The process was based on information derived from 2002 and 2003 INEX topics. The QUT submission in the Ad hoc track was ranked 1st from 52 submissions, with an average precision of 0.126. The NLQ submission was also ranked 1st in the NLQ track, with an average precision of 0.064. There was a marked drop in precision between the NLQ based search and the NEXI based search. The QUT NLQ submission would have been ranked 9th from 52 submissions had it been entered in the Ad hoc track. Clearly, it is a much harder task, but nevertheless the system performed reasonably well by comparison with systems that used a much more precise definition of the constraints in the form of a NEXI expression.

Given the limited participation and the limited scope of the task, the results of the NLQ track at INEX 2004 can only be regarded as preliminary. The results seem to suggest that a natural language alternative to an XPath like language such as NEXI is viable and that this alternative deserves further attention.

3 Planned Natural Language Query task at INEX 2005

In 2005 the NLQ task will be refined. NEXI itself is still evolving and will probably become more expressive and less restrictive over time to allow for more complex information needs to be expressed.

We are considering direct use of the narrative as input in the natural language query task. This will allow NLQ systems access to more accurate information regarding the information need. In 2004 the NLQ system was restricted to a description that was meant to be a faithful representation of the NEXI

expression. This meant that any limitations of NEXI carried over to the NLQ expression. It makes much more sense to use the narrative without the intervening translation from the narrative to NEXI and to NLQ.

There is nevertheless a significant advantage to using NEXI to express queries. In order to facilitate easier access to INEX, and to lower the effort that is involved with participation at INEX, we intend to run the NLQ task in two distinct modes. Full NLQ mode, and NLQ2NEXI mode.

3.1 NLQ mode

In this mode participants submit retrieval runs that are based on the natural language specification of the information need (probably using the narrative). This is an open ended task and participants may use any NLP search technology or any approach in retrieval runs. The ultimate goal is of course to do as well as, or better than, NEXI based system. There are no restrictions other than the use of a specification of the information need in natural language.

3.2 NLQ2NEXI mode

This mode is meant to encourage researchers, with NLP expertise, but without a search engine or database engine at their disposal, to enter the INEX evaluation program. It is designed to eliminate the need for the high cost incurred in the development of a database and search engine. In order to allow new participants entry without a search engine, or even without the document collection, the track is focused on the NLQ task and not on the retrieval task:

Given the set of INEX topics, the DTD, the NEXI language specification, and the NEXI syntax checker program, participants need to

- Provide a new set of topics, having replaced the original NEXI titles with ones derived by a program from natural language expressions of the information need.
- These derived NEXI topics form the "submission". Participants need not perform any retrieval runs. This will be done by the organizers on a standard search engine/s that operate on NEXI queries.
- The submissions will be evaluated against the assessments of the Ad hoc topics (sharing the same information need).
- The outcome of evaluation is a comparison between NLQ2NEXI systems without requiring participants to index or search the collection.

4 References

Details about INEX and about the specific Ad hoc and NLP approaches can be found in the INEX 2004 Workshop pre-proceedings (soon to appear as a volume in Lecture Notes in Computer Science, Springer-Verlag.) Editors N Fuhr, M Lalmas, S Malik, and Z Szlavik *http://inex.is.informatik.uni-duisburg.de:2004/pdf/INEX2004PreProceedings.pdf*

The INEX 2004 URL provides general information on the workshop: *http://inex.is.informatik.uni-duisburg.de:2004*

NEXI is described by Trotman, A., & Sigurbjson, B. (2004). Narrowed Extended XPath I (NEXI). In INEX 2004 Workshop Pre-Proceedings, (pp. 219-236). The online NEXI parser can be found here: *http://metis.otago.ac.nz/abin/nexi.cgi*