The NLP Task at INEX 2005

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

Alan Woodley
Queensland University of Technology
Brisbane, Australia
ap.woodley@student.qut.edu.au

Abstract

With XML information retrieval, like in traditional IR, the user's information need is loosely defined, linguistic variations are frequent, and answers are a ranked list of relevant elements. Like in database querying, structure is of importance and a simple list of keywords may not be sufficient to define an XML query. Structured query languages for XML have been developed, but appear to be difficult to use even by system-level users, let alone end-users. Therefore developing natural language interfaces for XML-IR requires innovative solutions.

INEX provides a framework (documents, topics and relevance assessments) for independent evaluation of XML-IR systems and approaches. In 2002 and 2003 systems in the Ad-hoc task accepted formal language queries (i.e. <*title*> elements) and produced results lists of relevant XML elements (usually well below the document root). In 2004 INEX introduced a Natural Language Processing task that operated in parallel with the Ad-hoc task (using the same topics and assessments) except that systems operated on XML natural language queries (i.e. <*description*> elements).

In this report we describe the motivation for using NLP in XML IR, the general approaches that were tested at INEX 2005, and comment on the results.

1 Introduction

Although queries that are based on natural language description are not new (e.g. TREC), The information need in XML information retrieval is more complex than in traditional IR. Users must express both content and structural requirements and therefore, XML-IR users need a more sophisticated interface than just keywords or simple phrases. Historically, this information need had been captured by formal query languages; however, these are almost invariably too complex – hence impractical - for end-users, for two main reasons: Firstly, formal languages are too difficult for ad-hoc use for both casual and expert users. Experiences within INEX has shown the difficulty that both expert users (O'Keefe and Trotman, 2004) and casual users (van Zowl et. al, 2005) have in writing correct formal queries in the first or even second attempt. In contrast, a natural language is an easy and intuitive way for users to express both content and structural requirements.

```
<inex topic topic id="256" query type="CAS" >
  <InitialTopicStatement>Find Information regarding data embedding using watermarking.
  </InitialTopicStatement>
  <castitle>
  //article[about(.//p,data embedding] //p[about(.,watermarking)]
  </castitle>
  <description>
  We are looking for paragraphs describing watermarking in articles which describe data embedding.
  </description>
  <narrative>
  In today's world the issue of data security is highly significant. One such technique to ensure data security
  is steganography where data is embedded in various media files…
  </narrative>
</inex topic>
```

Figure 1. An example of an INEX topic

Secondly, formal languages are bound to the structure of the documents. For instance, in order to retrieve information from abstracts, sections or bibliographic items, users need to know the corresponding markup tags (for example <*abs*>, <*sec*> and <*bb*> respectively). It would be difficult for users to remember every tag name in collection, and near impossible in a rich heterogeneous collection. A natural language interface could help resolve this problem, since users could express their information need conceptually and let the system construct the specific query.

3. INEX Topics

Figure 1 is an example of an INEX topic. The format of INEX topics is similar to that of TREC topics. The <*castitle*> and <*description*> elements are used as input for the Ad-hoc and NLP tasks respectively. The <*castitle*> *(content and structure title)* represents users' information need as a formal XPath-like language called NEXI (O'Keefe and Trotman, 2004) while the <*description*> expresses the users' information need in a natural language (e.g. English). NEXI's syntax encapsulates both an element's structural path and content requirements. In 2005 two types of topics were used in the INEX NLP task: COS and CAS. COS queries that represented a relatively simple information need, such as "*curricula vitae about information retrieval students*") and CAS queries that represented somewhat more complex requirements (such as "*retrieve abstracts about information retrieval in articles about natural language processing*").

2 Methodology

The INEX NLP task has run in 2004 and in 2005. In 2004, participants produced complete NLQ XML-IR systems that accepted natural language queries as input and produced ranked lists of XML elements as output. In 2005, an additional ***NLQ2NEXI*** task was added. Participants in the NLQ2NEXI task accepted natural language queries as input and produced NEXI queries as output. The NEXI queries were then executed on a third-party Ad-hoc XML-IR system (GPX, Geva, 2005) that produced the ranked lists of XML elements. The result lists were evaluated as if they were conventional Ad-hoc submissions against the same topics and same relevance assessments. This

approach was adopted since it: provided a more meaningful comparison between systems by using a standard baseline search engine across all NLQ2NEXI submissions; It provided a lower cost of entry for participation, with participants not having to implement an XML search engine.

3 INEX 2005 Approaches

Following the INEX 2005 CFP 11 groups registered for participation. By the end of the process only three groups made submissions: Ecoles des Mines de Saint-Etienne (EMSE), University of Klagenfurt (KLU), and Queensland University of Technology (QUT). While the final number of participating systems was disappointing, the result were certainly the complete opposite. In fact, all 3 submissions demonstrated that XML natural language queries are a viable approach. Each of the participating systems used a different approach, however, all were based on the following four steps.

Step 1: Detection of structural and content constraints in the natural language query. Both KLU and QUT set up a template matching based on words and parts-of-speech; however, QUT also preformed syntactic parsing as a precursor to template matching. EMSE used a deep syntactic analysis, complemented by specific semantic rules concerning query structure

Step 2. Mapping of structural constraints to corresponding XML markup tags. QUT used a dictionary look-up based up the properties of the collection's DTD. EMSE added grammatical knowledge to identify frequent linguistic constructions that implicitly refer to structure.

Step 3. Derivation of content requirements. Most of the users' content requirements occurred in noun phrases that were identified either by explicit delimitations (e.g. quotation marks) or as specific sequences of parts-of-speech. EMSE was able to use content terms to set up a contextual search along the entire structure of the collection.

Step 4. NEXI query formation. The final stage of translation is the formulation of NEXI queries that were based on the derived structural and content constraints.

4 INEX 2005 NLP Results

The full results for the 2005 NLP task are available in the INEX proceedings (Fuhr. et. al, 2006). The proceedings contain results for both the COS and CAS tasks for each of the NLP participating system. A fourth "baseline" system which used the original (manually constructed) NEXI expressions as input (<*castitle*> elements). This allowed us to compare the performance of the NLP systems with the underlying search engine using the Ad-hoc track topics. The results are interesting since the approaches are often comparable, and some times outperform the baseline system. This is a significant improvement over the results in INEX 2004 and demonstrates the potential for natural language queries as an alternative to formal language interfaces in XML-IR. Furthermore, the baseline system itself proved to be rather successful at INEX 2005 and consequently the performance of the NLQ2NEXI systems that utilised it were also very good in comparison with the performance of standard Ad-hoc submissions. The MAP values of the NLP submissions were usually at about 80% of the MAP values that were obtained by the best INEX submission.

5 Conclusion and Future Directions

Here, we have presented a summary of the INEX 2005 NLP task. While the numbers of participants in 2005 was small, the systems were rather successful, sometimes outperforming an equivalent Ad-hoc system. In 2006 the task will again run in parallel with the ad-hoc task and the collection that will be used is a snapshot of the entire Wikipedia in XML format. This collection introduces new

complexities. Many documents, and document sets, uses their own unique extensions to the dtd (the template facility of the Wikipedia). There are Xlinks throughout the collection, and the semantics of the XML is much richer than it was with the IEEE Computer Society articles collection that was previously used. All this makes for a very challenging task not only for the NLP participants, but for the Ad-hoc task participants as well.

6 Acknowledgements

The authors would like to acknowledge the participants in the 2005 NLP task especially Xavier Tannier and Marcus Hassler. We would also like to the acknowledge and thank the organisers of INEX Norbert Fuhr and Mounia Lamlas.

7 References

Fuhr, N., Lalmas, M, Malik, S. and Kazai, G. (eds). Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Dagstuhl 28-30 November 2005, Lecture Notes in Computer Science, Vol 3977, Springer-Verlag, 2006.

Geva, S. 2005. GPX - Gardens Point XML Information Retrieval at INEX 2005. in Fuhr et. al (2006)

O'Keefe, R. and Trotman, A. The simplest query language that could possibly work. In *INEX 2003 Workshop Proceedings*, (Dagstuhl, Germany, December 15-17 2003), 2004, 167-174.

Tannier, X. From Natural Language to NEXI, an interface for INEX 2005 queries. In Fuhr et. al (2006).

van Zowl, R, Bass, J., van Oostendorp, H., Wiering, F. Query Formulation for XML Retrieval with Brick. In Fuhr, N., Lamas, M., Trotman, A. (eds.). *In Proceedings of INEX 2005 Workshop on Element Retrieval Methodology vol 2*, (Glasgow, Scotland 2005) 2005, 75-83.

Woodley A. and Geva, S. NLPX at INEX 2005. In Fuhr et. al (2006) LNCS Proceeding of INEX 2005 *(to appear)*.