Overview of the INEX 2009 Entity Ranking Track

Gianluca Demartini[1], Tereza Iofciu[1], and Arjen P. de Vries[2]

[1] L3S Research Center
Leibniz Universität Hannover
Appelstrasse 9a D-30167 Hannover, Germany
{demartini,iofciu}@L3S.de
[2] CWI & Delft University of Technology
The Netherlands
arjen@acm.org

Abstract. In some situations search engine users would prefer to retrieve entities instead of just documents. Example queries include "Italian Nobel prize winners", "Formula 1 drivers that won the Monaco Grand Prix", or "German spoken Swiss cantons". The XML Entity Ranking (XER) track at INEX creates a discussion forum aimed at standardizing evaluation procedures for entity retrieval. This paper describes the XER tasks and the evaluation procedure used at the XER track in 2009, where a new version of Wikipedia was used as underlying collection; and summarizes the approaches adopted by the participants.

1 Introduction

Many search tasks would benefit from the ability of performing typed search, and retrieving entities instead of 'just' web pages. Since 2007, INEX has organized a yearly XML Entity Ranking track (INEX-XER) to provide a forum where researchers may compare and evaluate techniques for engines that return lists of entities. In entity ranking (ER) and entity list completion (LC), the goal is to evaluate how well systems can rank entities in response to a query; the set of entities to be ranked is assumed to be loosely defined by a generic category, implied in the query itself (for ER), or by some example entities (for LC). This year we adopted the new Wikipedia document collection containing annotations with the general goal of understanding how such semantic annotations can be exploited for improving Entity Ranking.

Entity ranking concerns triples of type <query, category, entity>. The category (i.e. the entity type), specifies the type of 'objects' to be retrieved. The query is a free text description that attempts to capture the information need. The Entity field specifies example instances of the entity type. The usual information retrieval tasks of document and element retrieval can be viewed as special instances of this more general retrieval problem, where the category membership relates to a syntactic (layout) notion of 'text document', or 'XML element'. Expert finding uses the semantic notion of 'people' as its category, where the

query would specify 'expertise on T' for expert finding topic T. While document retrieval and expert finding represent common information needs, and therefore would warrent specific technologies to be developed, the XER track challenges participants to develop generic ranking methods that apply to entities irrespective of their type: e.g., actors, restaurants, musea, countries, etc.

In this paper we describe the INEX-XER 2009 track running both the ER and the LC task, using selected topics from the previous editions over the new INEX Wikipedia collection. For evaluation purpose we adopted a stratified sampling strategy for creating the assessment pools, using xinfAP as the official evaluation metric [5].

The remainder of the paper is organized as follows. In Section 2 we present details about the collection used in the track and the two different search tasks. Next, in Section 3 we briefly summarize the approaches designed by the participants. In Section 4 we summarize the evaluation results computed on the final set of topics for both the ER and LC task. As this year we used a selection of topics from the past editions, in Section 5 we provide an initial comparison of the new test collection with the previous ones. Finally, in Section 6, we conclude the paper.

2 INEX-XER Setup

2.1 Data

The INEX-XER 2009 track uses the new Wikipedia 2009 XML data based on a dump of the Wikipedia taken on 8 October 2008 and annotated with techniques described in [4]. Available annotations could be exploited to find relevant entities. Category information about the pages loosely defines the entity sets. The entities in such a set are assumed to loosely correspond to those Wikipedia pages that are labeled with this category (or perhaps a sub-category of the given category). Obviously, this is not perfect as many Wikipedia articles are assigned to categories in an inconsistent fashion. Retrieval methods should handle the situation that the category assignments to Wikipedia pages are not always consistent, and also far from complete. The intended challenge for participants is therefore to exploit the rich information from text, structure, links and annotations to perform the entity retrieval tasks.

2.2 Tasks

This year's INEX-XER track consists of two tasks, i.e., entity ranking (with categories), and entity list completion (with examples). Entity list completion is a special case of entity ranking where a few examples of relevant entities are provided instead of the category information as relevance feedback information.

Entity Ranking. The motivation for the entity ranking (ER) task is to return entities that satisfy a topic described in natural language text. Given preferred categories, relevant entities are assumed to loosely correspond to those Wikipedia

pages that are labeled with these preferred categories (or perhaps sub-categories of these preferred categories). Retrieval methods need to handle the situation where the category assignments to Wikipedia pages are not always consistent or complete. For example, given a preferred category 'art museums and galleries', an article about a particular museum such as the 'Van Gogh Museum' may not be labeled by 'art museums and galleries' at all, or, be labeled by a sub-category like 'art museums and galleries in the Netherlands'. Therefore, when searching for "art museums in Amsterdam", correct answers may belong to other categories close to this category in the Wikipedia category graph, or may not have been categorized at all by the Wikipedia contributors. The category 'art museums and galleries' is only an indication of what is expected, not a strict constraint (like in the CAS title for the ad-hoc track).

List Completion. List completion (LC) is a sub-task of entity ranking which considers relevance feedback information. Instead of knowing the desired category (entity type), the topic specifies a number of correct entities (instances) together with the free-text context description. Results consist again of a list of entities (Wikipedia pages). If we provide the system with the topic text and a number of entity examples, the task of list completion refers to the problem of completing the partial list of answers. As an example, when ranking 'Countries' with topic text 'European countries where I can pay with Euros', and entity examples such as 'France', 'Germany', 'Spain', then 'Netherlands' would be a correct completion, but 'United Kingdom' would not.

2.3 Topics

Based on the topics from the previous two INEX-XER editions, we have set up a collection of 60 XER topics, with 25 from 2007 and 35 topics form 2008. The <categories> part is supposed to be used exclusively for the Entity Ranking Task. The <entities> part is supposed to be used exclusively for the List Completion Task.

2.4 The INEX-XER 2009 Test Collection

The initial set of topics for the 2009 XER Track consisted of 60 topics which have originally been selected from the last two editions to be run on the new INEX Wikipedia collection. The judging pools have been based on all submitted runs, using a stratified sampling strategy. As we aimed at performing relevance judgments on 60 topics (as compared to 49 in 2008), we adopted a less aggressive sampling strategy that would make the judging effort per topic lower. We used the following strata and sampling rates for the pool construction of INEX-XER 2009:

- $[1, 8]$ 100%
- $[9, 31]$ 70%

– [32, 50] 30%
– [51, 100] 10%

The resulting pools contained on average 312 entities per topic (as compared to 400 in 2008 and 490 in 2007).

All 60 topics have been re-assessed by INEX-XER 2009 participants on the new collection. As in the last edition, from the originally proposed ones, topics with less than 7 relevant entities (that is, 104, and 90) and topics with more than 74 relevant entities (that is, 78, 112, and 85) have been excluded (see [3]). The final set consists of 55 genuine XER topics with assessments.

Out of the 55 XER topics, 3 topics have been excluded for the LC task (i.e., 143, 126, and 132). The reason is that example entities for these topics were not relevant as the underlying Wikipedia collection has changed. After this selection, 52 List Completion topics are part of the final set and are considered in the evaluation.

2.5 Not-an-entity annotations

An additional difference from the relevance judgments performed during the past editions of INEX-XER is the possibility to mark a retrieved result as not being an entity. This choice is intended for those Wikipedia pages that do not represent an entity and, thus, would be irrelevant to any XER query. Examples include "list-of" or "disambiguation" pages.

Differentiating between a non-relevant and not-an-entity result does not influence the evaluation of INEX-XER systems as both judgments are considered a wrong result for XER tasks. These judgments may however be useful as training data, for example, to train classifiers for entity/non-entity pages.

3 Participants

At INEX-XER 2009 five groups submitted runs for both the ER and LC tasks. We received a total of 16 ER runs and 16 LC runs. In the following we report a short description of the approaches used, as reported by the participants.

Waterloo. Our two runs for each task is based on Clarke et al.'s question answering technique that uses redundancy [1]. Specifically, we obtained top scoring passages from each article in the corpus using topic titles (for ER task) and topic titles+examples (for LC task). For LC task, we estimated the categories of entities to return as the union of categories in the examples. Within each top scoring passage, we located candidate terms that have a Wikipedia page that fall under the desired categories. We ranked the candidate terms by the number of distinct passages that contain the term.

AU-CEG (Anna University,Chennai). In our approach, we have extracted the Entity Determining Terms (EDTs), Qualifiers and prominent n-grams from the query. As a second step, we strategically exploit the relation between the extracted terms and the structure and connectedness of the corpus to retrieve links which are highly probable of being entities and then use a recursive mechanism for retrieving relevant documents through the Lucene Search. Our ranking mechanism combines various approaches that make use of category information, links, titles and WordNet information, initial description and the text of the document.

PITT team (School of Information Sciences, University of Pittsburgh). As recent studies indicate that named entities exist in queries and can be useful for retrieval, we also notice the ubiquitous existence of entities in entity ranking queries. Thus, we try to consider entity ranking as the task of finding entities related to existing entities in a query. We implement two generative models, i.e. MODEL1EDR and MODEL1EDS, both of which try to capture entity relations. These two models are compared with two baseline generative models: MODEL1D, which estimates models for each entity using Wikipedia entity documents; MODEL1E, which interpolates entity models in MODEL1D with entity category models.

UAms (Turfdraagsterpad). We rank entities by combining a document score, based on a language model of the document contents, with a category score, based on the distance of the document categories to the target categories. We extend our approach from last year by using Wordnet categories and by refining the categories we use as target categories.

UAms (ISLA). We propose a novel probabilistic framework for entity retrieval that explicitly models category information in a theoretically transparent manner. Queries and entities are both represented as a tuple: a term-based plus a category-based model, both characterized by probability distributions. Ranking of entities is then based on similarity to the query, measured in terms of similarities between probability distributions.

Discussion. It is possible to notice that a common behavior of participants this year was to identify entity mentions in the text of Wikipedia articles, passages, or queries. They then applied different techniques (e.g., detect entity relations, exploit category information) to produce a ranked list of Wikipedia articles that represents the retrieved entities. The best performing approach exploited a probabilistic framework ranking entities using similarity between probability distributions.

4 Results

The five groups submitted 32 runs to the track. The evaluation results for the ER task are presented in Table 1, those for the LC task in Table 2, both reporting xinfAP [5].

Table 1. Evaluation results for ER runs at INEX XER 2009.

Run	xinfAP
2_UAmsISLA_ER_TC_ERreltop:	0.517
4_UAmsISLA_ER_TC_ERfeedbackSP:	0.505
1_AU_ER_TC_mandatoryRun.txt:	0.270
3_UAmsISLA_ER_TC_ERfeedbackS:	0.209
2_UAmsISLA_ER_TC_ERfeedback:	0.209
1_TurfdraagsterpadUvA_ER_TC_base+asscats:	0.201
3_TurfdraagsterpadUvA_ER_TC_base+asscats+prfcats:	0.199
2_TurfdraagsterpadUvA_ER_TC_base+prfcats:	0.190
1_UAmsISLA_ER_TC_ERbaseline:	0.189
4_TurfdraagsterpadUvA_ER_TC_base:	0.171
1_PITT_ER_T_MODEL1EDS:	0.153
1_PITT_ER_T_MODEL1EDR:	0.146
1_PITT_ER_T_MODEL1ED:	0.130
1_PITT_ER_T_MODEL1D:	0.129
1_Waterloo_ER_TC_qap:	0.095
5_TurfdraagsterpadUvA_ER_TC_asscats:	0.082

Table 2. Evaluation results for LC runs at INEX XER 2009.

Run	xinfAP
5_UAmsISLA_LC_TE_LCexpTCP:	0.520
3_UAmsISLA_LC_TE_LCreltop:	0.504
6_UAmsISLA_LC_TE_LCexpTCSP:	0.503
1_UAmsISLA_LC_TE_LCexpTC:	0.402
1_UAmsISLA_LC_TE_LCtermexp:	0.358
2_UAmsISLA_LC_TEC_LCexpTCS:	0.351
3_UAmsISLA_LC_TE_LCexpT:	0.320
1_AU_LC_TE_mandatoryRun.txt:	0.308
2_UAmsISLA_LC_TE_LCbaseline:	0.254
4_UAmsISLA_LC_TE_LCexpC:	0.205
4_TurfdraagsterpadUvA_LC_TE_base+wn20cats:	0.173
3_TurfdraagsterpadUvA_LC_TE_base+wiki20cats+wn20cats:	0.165
2_TurfdraagsterpadUvA_LC_TE_base+wiki20cats+prfcats:	0.160
5_TurfdraagsterpadUvA_LC_TE_base+wiki20cats:	0.157
1_TurfdraagsterpadUvA_LC_TE_base+wiki20cats:	0.156
1_Waterloo_LC_TE:	0.100

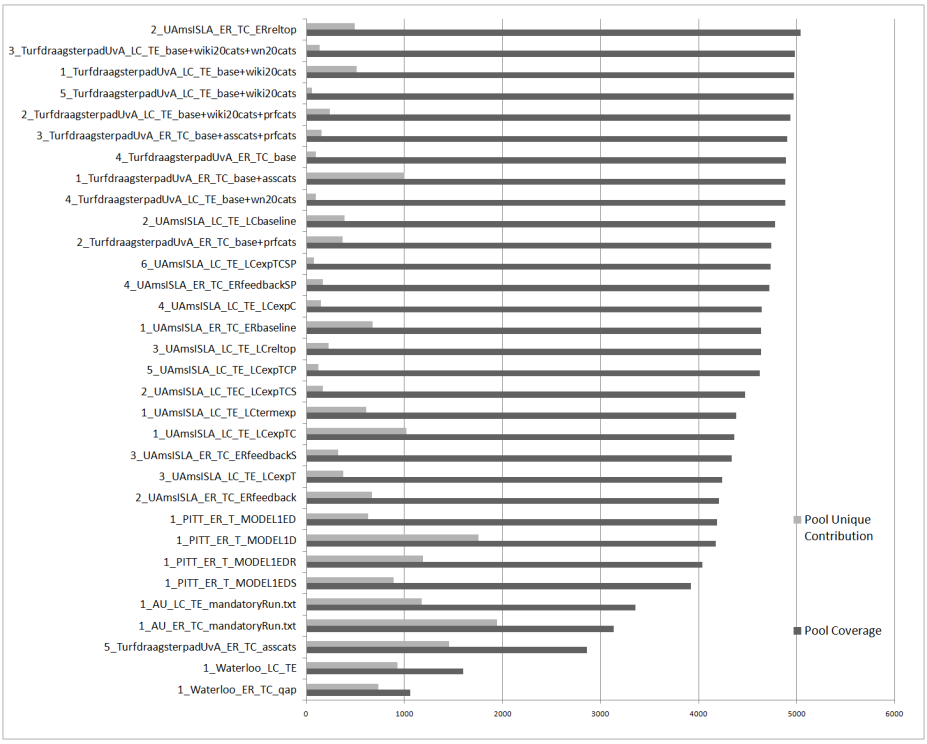

Fig. 1. Pool coverage: number of entities retrieved by the runs and present in the pool. Pool Unique Contribution: number of entities sampled only in this run.

As we considered all the runs during the pooling phase and as some groups submitted more runs than others, we performed an analysis of possible bias in the pool. Figure 1 shows both the *pool coverage* (i.e., the number of entities retrieved by the run which are present in the pool and, therefore, have been judged) and *pool unique contribution* (the number of entities in the pool which were sampled only in this run) for each run submitted to INEX-XER 2009. We can see that the runs having worse coverage from the pool are also those that contribute most unique entities. This means that such runs are "different" from others in the sense that they retrieve different entities. We can see in Figure 2 that a relatively high proportion of retrieved entities belong to strata 1 and 2, which guarantees a fair evaluation. However, as some runs did not retrieve up to 8 results for a topic and as some systems did not run all the topics, not all runs have an equal number of entities covered in stratum 1 (which considers a complete sampling). For the Waterloo runs for example, only few entities have been sampled due to the low number of entities retrieved per topic.

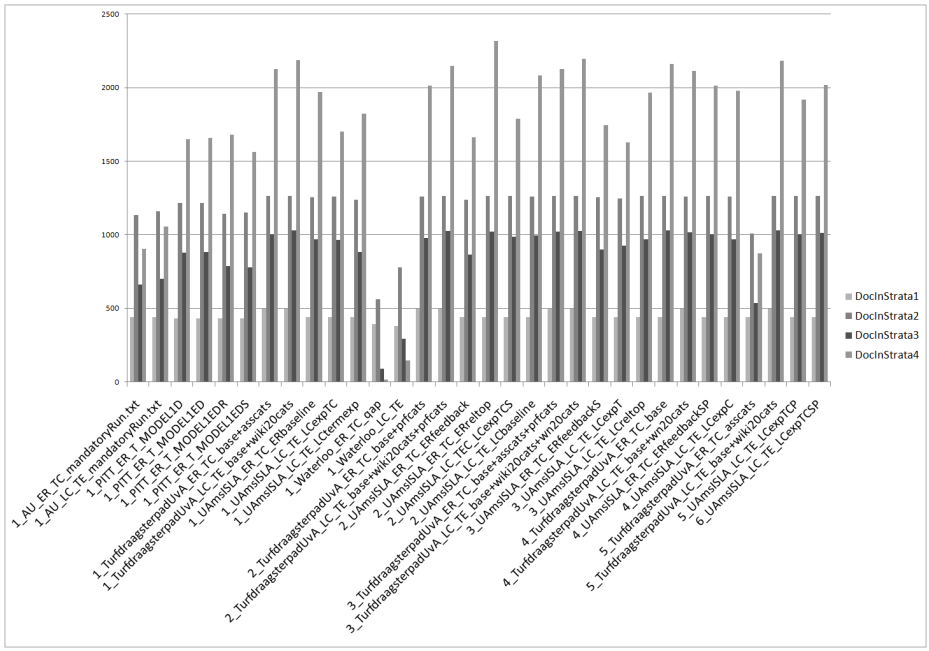

Fig. 2. Per-stratum pool coverage: number of entities retrieved by runs in different strata and present in the pool.

5 Comparison with Previous INEX-XER Collections

At INEX-XER 2009 we used a selected set of topics from the previous editions while using the newer and annotated Wikipedia collection. This allows us to perform some comparisons with previous collections.

Comparison on the number of relevant entities. Figure 3 shows the number of entities judged relevant for each topic at INEX-XER 2009 as well as in the previous editions. While we would expect to find the same number of relevant entities while re-judging the same topic, we must take into account that the new Wikipedia is bigger and contains more up-to-date information. Thus, we expect the number of relevant entities to be greater or equal to that in the past edition. This is not the case for 12 topics. The highest difference can be seen for topic 106 "Noble English person from the Hundred Years' War".

Preliminary Comparison on Samples and Judgments. Based on the titles of the sampled pages we compared the pools and assessments against the previous years. As the 2007 and 2008 topics have been assessed on the old corpus (with different IDs), we had to make the comparison based on the entity title performing a simple textual comparison. Thus minor changes in the title of an entity in the two collections would lead to the entity not being identified as the same in

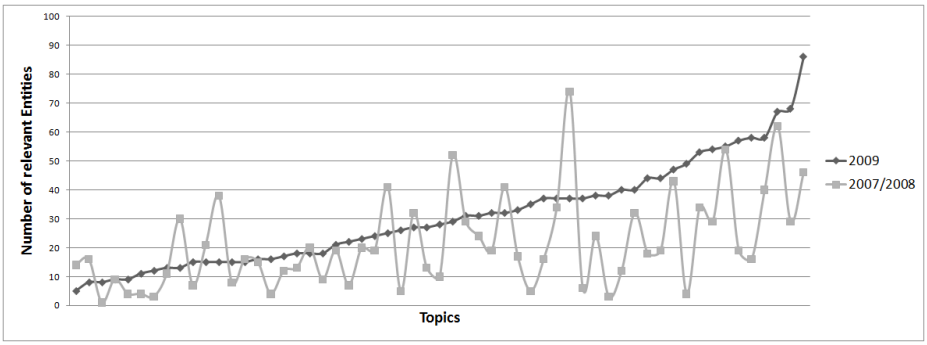

Fig. 3. Number of relevant entities per topic compared to previous editions.

the two collections. Table 3 shows the comparison results for 55 topics assessed in 2007/2008 and 2009. We show the following indicators in the table:

- S-co: the number of entities that have been sampled in both years
- S-past: the total number of entities that have been sampled in the past edition
- S-2009: the total number of entities that have been sampled at INEX-XER 2009
- R-past: the total number of relevant entities in the past edition
- R-2009: the total number of relevant entities at INEX-XER 2009
- R-co: the number of entities assessed as relevant in both years
- I-co: the number of entities assessed as not-relevant in both years
- Ryear1-Iyear2: the number of entities assessed as relevant in year1 and as not-relevant in year2
- UniRelYear: the number of entities that were both sampled and assessed as relevant only in the respective year

Table 3. Comparison of samples and judgments between INEX-XER 2009 and previous editions.

						agreement		disagreement			
Year	S-co	S-past	S-2009	R-past	R-2009	R-co	I-co	Rpast-I09	R09-Ipast	UniRelPast	UniRel09
2007	57.86	490	295.27	16.36	26.18	9.59	41.86	3.86	2.55	2.91	14.05
2008	79.24	400.03	314.55	26.09	31.64	16.91	54	4.94	3.24	4.24	11.48

For the set of topics assessed both in 2007 and 2009, from the entities sampled in both years (S-co), 17% were relevant in both years, and 72% were not relevant (the agreement between the assessments being of 89%). On the other hand, 6.7% entities were relevant in 2007 and assessed as not relevant in 2009, and 4.4% the other way around, thus amounting to a disagreement of 11%. Additionally, on

average 2.9 entities relevant in 2007 have not been sampled in 2009 (UniRelPast), and 14 entities not sampled in 2007 have been sampled and are relevant in 2009 (UniRel09).

For the set of topics assessed both in 2008 and 2009, from the entities sampled in both years, 21% were relevant in both years, and 68% were not relevant (the agreement between the assessments being of 89%). On the other hand, 6.2% entities were relevant in 2008 and assessed as not relevant in 2009, and 4.1% the other way around, thus amounting to a disagreement of 10%. Additionally, on average 4.2 entities relevant in 2008 have not been sampled in 2009 (UniRelPast), and 11 entities not sampled in 2008 have been sampled and are relevant in 2009 (UniRel09).

In conclusion, we observe that for the both sets of topics, the agreement between assessments (R-co + I-co) is much larger than the disagreement (Rpast-I09 + R09-Ipast).

6 Conclusions and Further Work

After the first two editions of the XER Track at INEX 2007 and INEX 2008 [2, 3], INEX-XER 2009 created additional evaluation material for IR systems that retrieve entities instead of documents. The new aspect in INEX XER 2009 is the use of the new annotated Wikipedia collection, re-using topics developed for the two previous editions of this track. The track created a set of 55 XER topics with relevance assessments for the ER task and 52 for the LC task.

References

1. Charles L. A. Clarke, Gordon V. Cormack, and Thomas R. Lynam. Exploiting redundancy in question answering. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 358–365, New York, NY, USA, 2001. ACM.
2. Arjen P. de Vries, Anne-Marie Vercoustre, James A. Thom, Nick Craswell, and Mounia Lalmas. Overview of the INEX 2007 Entity Ranking Track. In Norbert Fuhr, Jaap Kamps, Mounia Lalmas, and Andrew Trotman, editors, *INEX*, volume 4862 of *Lecture Notes in Computer Science*, pages 245–251. Springer, 2007.
3. Gianluca Demartini, Arjen P. de Vries, Tereza Iofciu, and Jianhan Zhu. Overview of the INEX 2008 Entity Ranking Track. In Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors, *INEX*, volume 5631 of *Lecture Notes in Computer Science*, pages 243–252. Springer, 2008.
4. R. Schenkel, FM Suchanek, and G. Kasneci. YAWN: A semantically annotated Wikipedia XML corpus. In *Symposium on Database Systems for Business, Technology and the Web of the German Socienty for Computer science (BTW 2007)*, 2007.
5. Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, *SIGIR*, pages 603–610. ACM, 2008.