Overview of the INEX 2011 Snippet Retrieval Track

Matthew Trappett[1], Shlomo Geva[1], Andrew Trotman[2], Falk Scholer[3], and Mark Sanderson[3]

[1] Queensland University of Technology, Brisbane, Australia
matthew.trappett@qut.edu.au, s.geva@qut.edu.au
[2] University of Otago, Dunedin, New Zealand
andrew@cs.otago.ac.nz
[3] RMIT University, Melbourne, Australia
falk.scholer@rmit.edu.au, mark.sanderson@rmit.edu.au

Abstract. This paper gives an overview of the INEX 2011 Snippet Retrieval Track. The goal of the Snippet Retrieval Track is to provide a common forum for the evaluation of the effectiveness of snippets, and to investigate how best to generate snippets for search results, which should provide the user with sufficient information to determine whether the underlying document is relevant. We discuss the setup of the track, and the evaluation results.

1 Introduction

Queries performed on search engines typically return far more results than a user could ever hope to look at. While one way of dealing with this problem is to attempt to place the most relevant results first, no system is perfect, and irrelevant results are often still returned. To help with this problem, a short text snippet is commonly provided to help the user decide whether or not the result is relevant.

The goal of snippet generation is to provide sufficient information to allow the user to determine the relevance of each document, without needing to view the document itself, allowing the user to quickly find what they are looking for.

The INEX Snippet Retrieval track was run for the first time in 2011. Its goal is to provide a common forum for the evaluation of the effectiveness of snippets, and to investigate how best to generate informative snippets for search results.

2 Snippet Retrieval Track

In this section, we briefly summarise the snippet retrieval task, the submission format, the assessment method, and the measures used for evaluation.

2.1 Task

The task is to return a ranked list of documents for the requested topic to the user, and with each document, a corresponding text snippet describing the document. This text snippet should attempt to convey the relevance of the underlying document, without the user needing view the document itself.

Each run is allowed to return up to 500 documents per topic, with a maximum of 300 characters per snippet.

2.2 Test Collection

The Snippet Retrieval Track uses the INEX Wikipedia collection introduced in 2009 — an XML version of the English Wikipedia, based on a dump taken on 8 October 2008, and semantically annotated as described in [1]. This corpus contains 2,666,190 documents.

The topics have been reused from the INEX 2009 Ad Hoc Track [2]. Each topic contains a short content only (CO) query, a content and structure (CAS) query, a phrase title, a one line description of the search request, and a narrative with a detailed explanation of the information need, the context and motivation of the information need, and a description of what makes a document relevant or not.

To avoid the 'easiest' topics, the 2009 topics were ranked in order of the number of relevant documents found in the corresponding relevance judgements, and the 50 with the lowest number were chosen.

For those participants who wished to generate snippets only, and not use their own search engine, a reference run was generated using BM25.

2.3 Submission Format

An XML format was chosen for the submission format, due to its human readability, its nesting ability (as information was needed at three hierarchical levels — submission-level, topic-level, and snippet-level), and because the number of existing tools for handling XML made for quick and easy development of assessment and evaluation.

The submission format is defined by the DTD given in Figure 1. The following is a brief description of the DTD fields. Each submission must contain the following:

- participant-id: The participant number of the submitting institution.
- run-id: A run ID, which must be unique across all submissions sent from a single participating organisation.
- description: a brief description of the approach used.

Every run should contain the results for each topic, conforming to the following:

- topic: contains a ranked list of snippets, ordered by decreasing level of relevance of the underlying document.

```
<!ELEMENT inex-snippet-submission (description,topic+)>
<!ATTLIST inex-snippet-submission
  participant-id CDATA #REQUIRED
  run-id CDATA #REQUIRED
>
<!ELEMENT description (#PCDATA)>
<!ELEMENT topic (snippet+)>
<!ATTLIST topic
  topic-id CDATA #REQUIRED
>
<!ELEMENT snippet (#PCDATA)>
<!ATTLIST snippet
  doc-id CDATA #REQUIRED
  rsv CDATA #REQUIRED
>
```

Fig. 1. DTD for Snippet Retrieval Track run submissions

- topic-id: The ID number of the topic.
- snippet: A snippet representing a document.
- doc-id: The ID number of the underlying document.
- rsv: The retrieval status value (RSV) or score that generated the ranking.

2.4 Assessment

To determine the effectiveness of the returned snippets at their goal of allowing a user to determine the relevance of the underlying document, manual assessment has been used. The documents for each topic were manually assessed for relevance based on the snippets alone, as the goal is to determine the snippet's ability to provide sufficient information about the document.

Each topic within a submission was assigned an assessor. The assessor, after reading the details of the topic, read through the top 100 returned snippets, and judged which of the underlying documents seemed relevant based on the snippets.

To avoid bias introduced by assessing the same topic more than once in a short period of time, and to ensure that each submission is assessed by the same assessors, the runs were shuffled in such a way that each assessment package contained one run from each topic, and one topic from each submission.

2.5 Evaluation Measures

Submissions are evaluated by comparing the snippet-based relevance judgements with the existing document-based relevance judgements, which are treated as a ground truth. This section gives a brief summary of the specific metrics used. In all cases, the metrics are averaged over all topics.

We are interested in how effective the snippets were at providing the user with sufficient information to determine the relevance of the underlying document, which means we are interested in how well the user was able to correctly determine the relevance of each document. The simplest metric is the mean precision accuracy (MPA) — the percentage of results that the assessor correctly assessed, averaged over all topics.

$$\text{MPA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{1}$$

Due to the fact that most topics have a much higher percentage of irrelevant documents than relevant, MPA will weight relevant results much higher than irrelevant results — for instance, assessing everything as irrelevant will score much higher than assessing everything as relevant.

MPA can be considered the raw agreement between two assessors — one who assessed the actual documents (i.e. the ground truth relevance judgements), and one who assessed the snippets. Because the relative size of the two groups (relevant documents, and irrelevant documents) can skew this result, it is also useful to look at positive agreement and negative agreement to see the effects of these two groups.

Positive agreement (PA) is the conditional probability that, given one of the assessors judges a document as relevant, the other will also do so. This is also equivalent to the F_1score.

$$\text{PA} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \tag{2}$$

Likewise, negative agreement (NA) is the conditional probability that, given one of the assessors judges a document as relevant, the other will also do so.

$$\text{NA} = \frac{2 \cdot \text{TN}}{2 \cdot \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

Mean normalised prediction accuracy (MNPA) calculates the rates for relevant and irrelevant documents separately, and averages the results, to avoid relevant results being weighted higher than irrelevant results.

$$\text{MNPA} = 0.5\frac{\text{TP}}{\text{TP} + \text{FN}} + 0.5\frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

This can also be thought of as the arithmetic mean of recall and negative recall. These two metrics are interesting themselves, and so are also reported separately. Recall is the percentage of relevant documents that are correctly assessed.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

Negative recall (NR) is the percentage of irrelevant documents that are correctly assessed.

$$\mathrm{NR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{6}$$

The primary evaluation metric, which is used to rank the submissions, is the geometric mean of recall and negative recall (GM). A high value of GM requires a high value in recall and negative recall — i.e. the snippets must help the user to accurately predict both relevant and irrelevant documents. If a submission has high recall but zero negative recall (e.g. in the case that everything is judged relevant), GM will be zero. Likewise, if a submission has high negative recall but zero recall (e.g. in the case that everything is judged irrelevant), GM will be zero.

$$\mathrm{GM} = \sqrt{\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \cdot \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}} \tag{7}$$

3 Participation

Table 1. Participation in the Snippet Retrieval Track

ID	Institute	Runs
14	University of Otago	6
16	Kasetsart University	3
20	Queensland University of Technology	3
23	RMIT University	3
31	Radboud University Nijmegen	6
65	University of Minnesota Duluth	4
72	Jiangxi University of Finance and Economics	8
73	Peking University	8
83	Indian School of Mines, Dhanbad	3

In this section, we discuss the participants and their approaches.

In the 2011 Snippet Retrieval Track, 44 runs were accepted from a total of 56 runs submitted. These runs came from 9 different groups, based in 7 different countries. Table 1 lists the participants, and the number of runs accepted from them.

Participants were allowed to submit as many runs as they wanted, but were required to rank the runs in order of priority, with the understanding that some runs may not be assessed, depending on the total number of runs submitted. To simplify the assessment process, 50 runs were initially accepted, to match the number of topics. This was achieved by capping the number of runs at 8 runs per partipating institute, and discarding any runs ranked below 8.

Six submissions were later rejected. An additional three submissions did not include the full 50 topics, and are thus uncomparable with the remaining 41 submissions. They have been reported alongside the accepted submissions, but have not been assigned a ranking or included in any analysis.

3.1 Participant approaches

The following is a brief description of the approaches used, as reported by the participants.

Queensland University of Technology The run 'QUTFirst300' is simply the first 300 characters of the documents in the reference run.

The run 'QUTFocused', again using the reference run, ignored certain elements, such as tables, images, templates, and the reference list. The tf-idf values were calculated for the key words found in each document. A 300 character window was then moved along the text, counting the total key words found in each window, weighted by their tf-idf scores. The highest scoring window was found, then rolled back to the start of the sentence to ensure the snippet did not start mid-sentence.

The run 'QUTIR2011A' selects snippets, using the reference run to select the appropriate documents. A topological signature is created from the terms of the query. Snippets are determined as 300 character passages starting from the `<p>` tag that is used to delineate paragraphs in the documents. Signatures are created for these snippets and compared against the original query signature. The closest match is used.

RMIT University The snippet generation algorithm was based on selecting highly ranked sentences which were ranked according to the occurrence of query terms. Nevertheless, it was difficult to properly identify sentence boundaries due to having multiple contributors with different writing styles. The main exception was detected when a sentence included abbreviations such as "Dr. Smith". We did not do an analysis of abbreviations to address this issue in detail.

We processed Wikipedia articles before constructing snippets. Specifically, information contained inside the `<title>` and `<bdy>` was used to narrow the document content. We suggest that snippets should include information of the document itself instead of sources pointing to other articles. Therefore, the Reference section was ignored in our summarisation approaches. The title was concatenated to the leading scored sentences.

We used the query terms listed in the title, and we expanded them by addressing a pseudo relevance feedback approach. That is, the top 5 Wikipedia articles were employed for selecting the first 25 and 40 terms.

Radboud University Our previous study found that topical language model improves document ranking of ad-hoc retrieval. In this work, our attention is paid on snippets that are extracted and generated from the provided ranked list of documents.

In our experiments of the Snippet Retrieval Track, we hypothesize that the user recognizes certain combinations of terms in created snippets which are related to their information needs. We automatically extract snippets using terms

as the minimal unit. Each term is weighted according to their relative occurrance in its article and in the entire Wikipedia. The top K scoring terms are chosen for inclusion in the snippet. The term-extraction based snippets are then represented differently to the user. One is a cluster of words that indicate the described topic. Another is a cluster of semi-sentences that contains the topic information while preserving some language structure.

University of Minnesota Duluth The run entitled 'p65-UMD_SNIPPET_RETRIEVAL_RUN_3' was created as follows: Our method of dynamic element retrieval was used to generate a rank-ordered list of all elements associated with each article in the reference run. The elements were focused based on correlation, and the highest correlating element was selected as the basis for the snippet. For this particular run, the corresponding element from the original text (rather than the focused element itself) was selected and further processed by examining each sentence of the element, selecting those containing at least one query term, and ordering the sentences by the number of query terms contained in them. The top 300 characters from this text string were reported.

Jiangxi University of Finance and Economics p72-LDKE-$m_1m_2m_3m_4$, where m_i $(1 \leq i \leq 4)$ equals to 0 or 1, employs four different strategies to generate a snippet. Strategy 1 is dataset selection: using documents listed in reference runs ($m_1 = 0$) or Wikipedia 2009 dataset ($m_1 = 1$). Strategy 2 is snippet selection: using baseline method ($m_2 = 0$) or window method ($m_2 = 1$). According to the baseline method, after the candidate elements/nodes being scored and ranked, only the first 300 characters are extracted as snippet from the element/node has the highest score. Remain part of this snippet are extracted from the successive elements/nodes in case of the precedents are not long enough. While in the window method, every window that contain 15 terms are scored and those with higher scores are extracted as snippets. Strategy 3 is whether using ATG path weight ($m_3 = 1$) or not ($m_3 = 0$) in element retrieval model. The element retrieval model used in our system is based on BM25 and the works about ATG path weight has been published in CIKM 2010. Strategy 4 is whether reordering the XML document according to the reference runs ($m_4 = 0$) or not ($m_4 = 1$) after elements/nodes being retrieved.

Peking University In the INEX 2011 Snippet Retrieval Track, we retrieve XML documents based on both document structure and content, and our retrieval engine is based on the Vector Space Model. We use Pseudo Feedback method to expand the query of the topics. We have learned the weight of elements based on the cast of INEX2010 to enhance the retrieval performance, and we also consider the distribution of the keywords in the documents and elements, the more of the different keywords, the passage will be more relevant, and so is the distance of the keywords. We used method of SLCA to get the smallest sub-tree that satisfies the retrieval. In the snippet generation system, we use

query relevance, significant words, title/section-title relevance and tag weight to evaluate the relevance between sentences and a query. The sentences with higher relevance score will be chosen as the retrieval snippet.

4 Snippet Retrieval Results

In this section, we present and discuss the evaluation results for the Snippet Retrieval Track.

Table 2 gives the ranking for all of the runs. The run ID includes the ID number of the participating organisation; see Table 1 for the name of the organisation. The runs are ranked by geometric mean of recall and negative recall.

The highest ranked run is 'p72-LDKE-1111', submitted by the Jiangxi University of Finance and Economics.

Table 3 lists additional metrics for each run, as discussed in Section 2.5. One statistic worth noting is the fact that no run scored higher than 47% in recall, with an average of 35%. This indicates that poor snippets are causing users to miss more than half of all relevant results. Negative recall is high, with no run scoring higher than 80%, meaning that users are easily able to identify most irrelevant results based on snippets.

Significance tests were performed to determine whether higher ranked systems were significantly better than lower ranked systems. A one-tailed t-test at 95% was used. Table 4 shows, for each submission (shown on the left), whether it is significantly better than each lower ranked run (indicated by "$\star$"). The top run is significantly better than runs 14, 15, 17 and 19–41 – 65% of all lower-ranked runs. However, the average is much lower than this – of the 820 possible pairs of runs, there are only 321 (or 39.2%) significant differences. We should therefore be careful when drawing conclusions based on these results.

5 Conclusion

This paper gave an overview of the INEX 2011 Snippet Retrieval track. The goal of the track is to provide a common forum for the evaluation of the effectiveness of snippets. The paper has discussed the setup of the track, and presented the preliminary results of the track. The results show that, for the submitted runs, users are generally able to identify most irrelevant results, but poor snippets are causing them to miss over half of the relevant results, indicating that there is still substantial work to be done in this area.

References

1. Schenkel, R., Suchanek, F.M., Kasneci, G.: YAWN: A semantically annotated Wikipedia XML corpus. In: 12. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW 2007), pp. 277–291 (2007)

Table 2. Ranking of all runs in the Snippet Retrieval Track, ranked by GM

Rank	Run	Score
1	p72-LDKE-1111	0.5705
2	p23-baseline	0.5505
3	p72-LDKE-0101	0.5472
4	p20-QUTFirst300	0.5416
5	p73-PKU_ICST_REF_11a	0.5341
6	p72-LDKE-1110	0.5317
7	p23-expanded-40	0.5294
8	p72-LDKE-0111	0.5270
9	p65-UMD_SNIPPET_RETRIEVAL_RUN_3	0.5264
10	p20-QUTFocused	0.5242
11	p14-top_tficf_passage	0.5242
12	p23-expanded-25	0.5239
13	p72-LDKE-1121	0.5192
14	p72-LDKE-1101	0.5130
15	p20-QUTIR2011A	0.5122
16	p73-PKU_105	0.5080
17	p73-PKU_102	0.5011
18	p73-PKU_100	0.5001
19	p72-LDKE-1001	0.4919
20	p35-97-ism-snippet-Baseline-Reference-run_01	0.4886
21	p73-PKU_107	0.4805
22	p31-SRT11DocTXT	0.4803
23	p35-98-ism-snippet-Baseline-Reference-run_01	0.4800
24	p72-LDKE-1011	0.4770
25	p73-PKU_106	0.4741
26	p65-UMD_SNIPPET_RETRIEVAL_RUN_4	0.4680
27	p14-top_tf_passsage	0.4648
28	p14-top_tf_p	0.4574
29	p31-SRT11ParsDoc	0.4557
30	p65-UMD_SNIPPET_RETRIEVAL_RUN_1	0.4470
31	p73-PKU_ICST_REF_11b	0.4459
32	p35-ism-snippet-Baseline-Reference-run_02	0.4365
33	p31-SRT11DocParsedTXT	0.4351
34	p14-top_tficf_p	0.4337
35	p65-UMD_SNIPPET_RETRIEVAL_RUN_2	0.4270
36	p31-SRT11ParsStopDoc	0.4157
37	p14-first_p	0.4044
38	p73-PKU_101	0.3956
39	p14-kl	0.3598
40	p31-SRT11ParsStopTerm	0.3458
41	p31-SRT11ParsTerm	0.3392
n/a	p16-kas16-MEXIR-ALL	0.0000
n/a	p16-kas16-MEXIR-ANY	0.0000
n/a	p16-kas16-MEXIR-EXT	0.0000

Table 3. Additional metrics of all runs in the Snippet Retrieval Track (preliminary results only)

Run	MPA	MNPA	Recall	NR	PA	NA
p14-first_p	0.7582	0.6430	0.4641	0.8219	0.3748	0.8292
p14-kl	0.7638	0.6220	0.4115	0.8325	0.3329	0.8313
p14-top_tf_p	0.7684	0.6328	0.4470	0.8187	0.3646	0.8232
p14-top_tf_passsage	0.8022	0.6076	0.3269	0.8884	0.3151	0.8715
p14-top_tficf_p	0.7860	0.6263	0.3888	0.8637	0.3279	0.8587
p14-top_tficf_passage	0.7674	0.6179	0.4058	0.8299	0.3452	0.8364
p20-QUTFirst300	0.7728	0.5919	0.3064	0.8774	0.2727	0.8533
p20-QUTFocused	0.7982	0.5781	0.2446	0.9116	0.2576	0.8715
p20-QUTIR2011A	0.7580	0.6024	0.3716	0.8333	0.2959	0.8247
p23-baseline	0.7702	0.5896	0.3101	0.8692	0.2952	0.8475
p23-expanded-25	0.7988	0.6086	0.3235	0.8938	0.2725	0.8676
p23-expanded-40	0.7690	0.5708	0.2721	0.8695	0.2227	0.8360
p31-SRT11DocParsedTXT	0.8026	0.6047	0.2982	0.9113	0.2900	0.8737
p31-SRT11DocTXT	0.7992	0.6128	0.3231	0.9026	0.3007	0.8721
p31-SRT11ParsDoc	0.7830	0.5704	0.2431	0.8977	0.2171	0.8544
p31-SRT11ParsStopDoc	0.8092	0.6204	0.3513	0.8896	0.3333	0.8672
p31-SRT11ParsStopTerm	0.7830	0.6247	0.3824	0.8670	0.3387	0.8525
p31-SRT11ParsTerm	0.7766	0.6231	0.3866	0.8596	0.3389	0.8514
p35-97-ism-snippet-Baseline-Reference-run_01	0.7958	0.6441	0.4035	0.8848	0.3715	0.8667
p35-98-ism-snippet-Baseline-Reference-run_01	0.7936	0.6233	0.3597	0.8870	0.3319	0.8685
p35-ism-snippet-Baseline-Reference-run_02	0.7652	0.5877	0.3229	0.8525	0.2849	0.8365
p65-UMD_SNIPPET_RETRIEVAL_RUN_1	0.7724	0.5813	0.2904	0.8723	0.2736	0.8500
p65-UMD_SNIPPET_RETRIEVAL_RUN_2	0.7850	0.6207	0.3811	0.8602	0.3498	0.8542
p65-UMD_SNIPPET_RETRIEVAL_RUN_3	0.7976	0.5982	0.3027	0.8937	0.3078	0.8645
p65-UMD_SNIPPET_RETRIEVAL_RUN_4	0.8056	0.6245	0.3534	0.8956	0.3448	0.8706
p72-LDKE-0101	0.8042	0.6165	0.3348	0.8982	0.3161	0.8712
p72-LDKE-0111	0.8090	0.5984	0.2875	0.9093	0.2850	0.8764
p72-LDKE-1001	0.8026	0.6250	0.3706	0.8793	0.3530	0.8697
p72-LDKE-1011	0.7886	0.5732	0.2618	0.8846	0.2398	0.8623
p72-LDKE-1101	0.7580	0.6201	0.4103	0.8300	0.3105	0.8358
p72-LDKE-1110	0.7928	0.6167	0.3731	0.8602	0.3269	0.8612
p72-LDKE-1111	0.7998	0.6076	0.3288	0.8864	0.3050	0.8675
p72-LDKE-1121	0.8054	0.6176	0.3544	0.8809	0.3061	0.8705
p73-PKU_100	0.7808	0.6250	0.3884	0.8617	0.3622	0.8516
p73-PKU_101	0.7892	0.5924	0.2964	0.8883	0.2883	0.8629
p73-PKU_102	0.7656	0.5723	0.2684	0.8762	0.2081	0.8426
p73-PKU_105	0.8144	0.6444	0.3980	0.8909	0.3747	0.8773
p73-PKU_106	0.7772	0.6265	0.3902	0.8627	0.3431	0.8481
p73-PKU_107	0.5908	0.2954	0.0000	0.5908	0.0000	0.6588
p73-PKU_ICST_REF_11a	0.8998	0.4499	0.0000	0.8998	0.0000	0.9386
p73-PKU_ICST_REF_11b	0.8829	0.4414	0.0000	0.8829	0.0000	0.9305
p16-kas16-MEXIR-ALL	0.7726	0.6161	0.3690	0.8633	0.3191	0.8434
p16-kas16-MEXIR-ANY	0.7590	0.6331	0.4347	0.8314	0.3647	0.8301
p16-kas16-MEXIR-EXT	0.7892	0.6279	0.3816	0.8742	0.3522	0.8624

Table 4. Statistical significance of improvement of each run over each lower-ranked run (t-test, one-tailed, 95%)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
1	-	-	-	-	-	-	-	-	-	-	-	-	-	★	★	-	★	-	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★
2		-	-	-	-	-	-	-	-	-	-	-	-	-	★	-	-	-	-	★	★	-	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★
3			-	-	-	-	-	-	-	-	-	-	-	-	★	-	-	-	-	★	-	-	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★
4				-	-	-	-	-	-	-	-	-	-	-	★	-	-	-	-	-	-	-	★	-	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★
5					-	-	-	-	-	-	-	-	-	-	★	-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★
6						-	-	-	-	-	-	-	-	-	★	-	-	-	-	-	-	-	-	-	-	★	★	★	-	★	★	★	★	★	★	★	★	★	★	★	★	★	★
7							-	-	-	-	-	-	-	-	★	-	-	-	-	-	-	-	-	-	-	★	-	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★
8								-	-	-	-	-	-	-	★	-	-	-	-	-	-	-	-	-	-	★	★	★	-	★	★	★	★	★	★	★	★	★	★	★	★	★	★
9									-	-	-	-	-	-	★	-	-	-	-	-	-	-	-	-	-	★	-	★	-	★	★	★	★	★	★	★	★	★	★	★	★	★
10										-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	★	★	-	-	★	★	★	★	★	★	★	★	★	★	★	★	★
11											-	-	-	★	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★	★	★	★	★	★	★
12												-	-	-	-	-	-	-	-	-	-	-	-	-	-	★	-	★	-	★	★	★	★	★	★	★	★	★	★	★	★	★
13													-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★	★	★	★	★	★	★	★
14														-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★	★	★	★	★	★	★	★
15															-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	★	★	★
16																-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	★	-	★	★	★	★	★	★	★	★	★
17																	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★	★	★	★
18																		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	★	-	★	★	★	★	★	★	★	★
19																			-	-	-	-	-	-	-	-	-	-	-	-	-	★	★	-	-	★	★	★	★	★	★	★	★
20																				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★	★	★
21																					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★	★	★
22																						-	-	-	-	-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★	★	★
23																							-	-	-	-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★	★	★
24																								-	-	-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★	★	★
25																									-	-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★	★	★
26																										-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★	★	★
27																											-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★	★
28																												-	-	-	-	-	-	-	-	-	-	★	★	★	★	★	★
29																													-	-	-	-	-	-	-	-	-	-	★	★	★	★	★
30																														-	-	-	-	-	-	-	-	-	-	★	★	★	★
31																															-	-	-	-	-	-	-	-	-	-	★	★	★	★
32																																-	-	-	-	-	-	-	-	-	-	★	★	★
33																																	-	-	-	-	-	-	-	-	-	-	★	★
34																																		-	-	-	-	-	-	-	-	-	★	★
35																																			-	-	-	-	-	-	-	-	★	★
36																																				-	-	-	-	-	-	-	-	-
37																																					-	-	-	-	-	-	-	-
38																																						-	-	-	-	-	-
39																																							-	-	-	-	-
40																																								-	-	-
41																																									-	-

2. Geva, S., Kamps, J., Lehtonen, M., Schenkel, R., Thom, J.A., Trotman, A.: Overview of the INEX 2009 ad hoc track. In: Geva, S., Kamps, J., Trotman, A. (eds.) Focused Retrieval and Evaluation. LNCS, pp. 4–25. Springer, Heidelberg (2010)