Overview of the INEX 2012 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 2012 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, details of the assessment and evaluation, and initial participation.

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 goal of the INEX Snippet Retrieval track 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.

This year is the second year in which the INEX Snippet Retrieval track has run. In response to feedback from the first year, search topics have been made more specific, and document-based assessment has been introduced.

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 must return 20 documents per topic, with a maximum of 180 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 by Schenkel et al. [1]. This corpus contains 2,666,190 documents.

This year there are 35 topics in total. The majority of these topics (25 of 35) have been created specifically for this track, with the goal being to create topics requesting more specific information than is likely to be found in the first few paragraphs of a document. The remaining 10 topics have been reused from the INEX 2010 Ad Hoc Track [2].

Each topic contains a short content only (CO) 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.

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

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 will be used. In response to feedback from the previous year, both snippet-based and document-based assessment will be used. The documents will first be 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. The documents will then be assessed for relevance based on the full document text, with evaluation based on comparing these two sets of assessments.

Each topic within a submission will be assigned an assessor. The assessor, after reading the details of the topic, read through the 20 returned snippets, and judge which of the underlying documents seem relevant based on the snippets. The assessor will then be presented the full text of each document, and determine whether or not the document was actually relevant.

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 will be shuffled in such a way that each assessment package contains 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 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 irrelevant, 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 Round 1 of the Snippet Retrieval Track

ID	Institute
20	Queensland University of Technology
46	Jadavpur University
65	University of Minnesota Duluth

Participation in the track has been split into two rounds, the first of which has had a compressed schedule. As of this writing, submissions for round 1 have closed, with submissions received from three participating organisations.

4 Conclusion

This paper gave an overview of the INEX 2012 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, the assessment method and evaluation metrics, as well as initial participation in the track.

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)

2. Arvola, P, Geva, S., Kamps, J., Schenkel, R., Trotman, A., Vainio, J: Overview of the INEX 2010 ad hoc track. In: Geva, S., Kamps, J., Trotman, A. (eds.) Comparative Evaluation of Focused Retrieval. LNCS, pp. 1–32. Springer Berlin / Heidelberg (2011)