Overview of the INEX 2013 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 2013 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. Such snippets 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 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. This allows 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 snippet effectiveness, and to investigate how best to generate informative snippets for search results.

This year is the third year in which the INEX Snippet Retrieval track has run. In response to feedback from the second year, the task has been modified to simplify the assessment process, and to place more emphasis on snippet retrieval rather than document retrieval.

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

A set of topics (or queries) has been provided, each with a corresponding set of search results, taken from the document collection (described below). The task is to automatically generate a text snippet for each of these search results. This text snippet should attempt to convey the relevance of the underlying document, without the user needing to view the document itself.

Each run must give a snippet for each of the 20 documents returned for each topic, with a maximum of 180 characters per snippet.

2.2 Test Collection

The topics for the 2013 track have been reused from the 2012 Snippet Retrieval track. There are 35 topics in total. The majority of these topics (25 of 35) have been created specifically for the Snippet Retrieval 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 [1].

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 irrelevant.

For each topic, there is a corresponding set of twenty documents — the search results for the topics. These XML documents are based on a dump of the English language Wikipedia, from November 2012.

2.3 Submission Format

An XML format was chosen for the submission format. This was due to the human readability, tree structure (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 unique ID identifying the particular run.
- 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.
- 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.

```
<!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

2.4 Assessment

To determine the effectiveness of the returned snippets at the goal of allowing a user to determine the relevance of the underlying document, manual assessment is used. Both snippet-based and document-based assessment are used.

The documents are first 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 is assigned an assessor. The assessor, after reading the details of the topic, reads through the top 100 returned snippets, and judges which of the underlying documents seem relevant based on the snippets alone.

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 are shuffled in such a way that topics from each submission are spread evenly amongst all assessors.

Additionally, each of the 20 documents returned for each of the 35 topics is assessed for relevance based on the full document text. This full set of 700 documents is assessed multiple times, by separate assessors. The consensus formed by all of the document assessments is treated as a ground truth.

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 — this 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.

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

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

In the 2013 Snippet Retrieval track, 4 runs were submitted, from 2 participating groups — 2 runs from Queensland University of Technology, and 2 runs from IRIT.

In addition, a baseline run was generated and evaluated, consisting of the first 180 characters of each document.

4 Snippet Retrieval Results

Table 1. Ranking of all runs in the Snippet Retrieval Track, ranked by GM (preliminary results only)

Rank	Participant	Run	Score
1	IRIT	snippets_2013_knapsack	0.5352
2	QUT	QUT_2013_Focused	0.4774
3	QUT	QUT_2013_Focused_Split	0.4732
4	IRIT	snippets_2013_MW	0.4605
5	-	SR2013-Baseline	0.4025

Table 2. Additional metrics of all runs in the Snippet Retrieval Track (preliminary results only)

Run	MPA	MNPA	Recall	NR	PA	NA
QUT_2013_Focused	0.8171	0.6603	0.3507	0.9700	0.4210	0.8675
QUT_2013_Focused_Split	0.8214	0.6549	0.3684	0.9413	0.4358	0.8624
snippets_2013_knapsack	0.8300	0.6834	0.4190	0.9477	0.4921	0.8673
snippets_2013_MW	0.8300	0.6459	0.3852	0.9067	0.4283	0.8572
SR2013-Baseline	0.8171	0.6414	0.2864	0.9964	0.3622	0.8711

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

At the time of writing, while each of the submissions have had their snippets assessed, the set of full-text documents has been assessed only once. As such, the results presented here are preliminary results only — the final set of results will use the consensus of multiple document assessors as its ground truth relevance judgments. This will be released at a later date.

Table 1 gives the ranking for all of the runs. The runs are ranked by geometric mean of recall and negative recall. The highest ranked run, according to the preliminary results, is 'snippets_2013_knapsack', submitted by IRIT.

Table 2 list additional metrics for each run, as discussed in Section 2.5. It can be seen that no run scored higher than 42% in recall, with an average of 36%. This indicates that poor snippets are causing users to miss over half of all relevant results. Negative recall, on the other hand, is high, with all runs scoring higher than 90%, meaning that users are able to easily identify most irrelevant results based on snippets alone.

5 Conclusion

This paper gave an overview of the INEX 2013 Snippet Retrieval track. The goal of the track is to provide a common forum for the evaluation of snippet effectiveness. The paper has discussed the setup of the track, and presented the preliminary results of the track. The preliminary results show that in all submitted runs, poor snippets are causing users to miss over half of all relevant results, indicating that a lot of work remains to be done in this area. Final results will be released at a later date, once further document assessment has been completed.

References

1. 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)