Overview of WebCLEF 2008 (Draft)

Valentin Jijkoun and Maarten de Rijke
ISLA, University of Amsterdam
jijkoun,mdr@science.uva.nl

Abstract

We describe the WebCLEF 2008 task. Similarly to the 2007 edition of WebCLEF, the 2008 edition implements a multilingual "information synthesis" task, where, for a given topic, participating systems have to extract important snippets from web pages. We detail the task and the assessment procedure. At the time of writing evaluation results are not available yet.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries

General Terms

Algorithms, Experimentation, Measurement

Keywords

Web retrieval, focused retrieval

The WebCLEF 2008 task is based on its 2007 predecessor [2]: for a given topic (undirected information need of the type *"Tell me all about X"*) automatic systems need to compile a set of snippets, extracting them from web pages found using Google. Thus, WebCLEF 2008 has similarities with (topic-oriented) multi-document summarization.

In the remainder of the paper we describe the task, the submissions and the assessment procedure. At the time of writing (August 2008) evaluation results are not been finalized yet.

1 Task description

The user model for the WebCLEF 2008 is the same as in the 2007 task definition [2]. Specifically, in our task model, the hypothetical user is a knowledgable person writing a survey article on a specific topic with a clear goal and audience (e.g., a Wikipedia article, or a state of the art survey, or an article in a scientific journal). She needs to locate items of information to be included in the article and wants to use an automatic system for this purpose. The user only uses online sources found via a Web search engine.

The user information needs (operationalized as WebCLEF 2008 topics) are specified as follows:

- a short *topic title* (e.g., the title of the survey article),
- a free text *description* of the goals and the intended audience of the article,

- a list of *languages* in which the user is willing to accept the information found,
- an optional list of *known sources*: online resources (URLs of web pages) that the user considers to be relevant to the topic and information from which might already have been included in the article, and
- an optional list of *Google retrieval queries* that can be used to locate the relevant information; each query specifies the expected language of the documents it is supposed to locate.

Below is an example of an information need:

- topic title: *Paul Verhoeven*
- description: I'm looking for information on similarities, differences, connections, influences between Paul Verhoeven's movies of his Dutch period and his American period.
- language: English, Dutch
- known source(s): `http://en.wikipedia.org/wiki/Paul_Verhoeven`, `http://nl.wikipedia.org/wiki/Paul_Verhoeven`
- retrieval queries: "`paul verhoeven (dutch AND american)`", "`paul verhoeven (nederlandse AND amerikaanse OR hollywood OR VS)`"

Each participating team was asked to develop 10 topics and subsequently assess responses of all participating systems for the created topics. In total, 61 multilingual topics were created, of which 48 were bilingual and 13 trilingual; specifically:

- 21 English-Spanish topics
- 21 English-Dutch topics;
- 10 English-Romanian-Spanish topics;
- 6 Russian-English topics;
- 2 English-German-Dutch topics; and
- 1 Russian-English-Dutch topic.

1.1 Data collection

The test collection consists of the web documents found using Google with the queries provided by the topic creators. For each topic the collection includes the following documents along with their URLs:

- all "known" sources specified for the topic;
- the top 100 (or less, depending on the actual availability) hits from Google for each of the retrieval query; in the 2007 edition of the task the test collection included up to 1000 documents per query;
- for each online document included in the collection, its URL, the original content retrieved from the URL and the plain text conversion of the content are provided. The plain text (UTF-8) conversion is only available for HTML, PDF and Postscript documents. For each document, the collection also provides its origin: which query or queries were used to locate it and at which rank(s) in the Google result list it was found.

Participant	Run	Average snippet length	Average snippets per topic	Average response length per topic
	baseline 2007	286	20	5,861
U. Twente	ip2008	450	32	14,580
	ipt2008	464	31	14,678
	ipu2008	439	33	14,607
UNED	Uned RUN1	594	24	14,817
	Uned RUN2	577	25	14,879
	Uned RUN3	596	24	14,861
U .Samalanca	usal 0	851	91	77,668
	usal 1	1,494	86	129,803
	usal 2	1,427	88	126,708

Table 1: Simple statistics for the baseline (one of the systems from WebCLEF 2007) and the 9 submitted runs.

1.2 System response

For each topic description, a response of an automatic system consists of a ranked list of plain text snippets extracted from the test collection. Each snippet should indicate what document in the collection it comes from.

2 Assessment

The assessment procedure was a simplification of the procedure from 2007. The assessment was blind. For a given topic, all responses of all system were pooled into an anonymized randomized sequence of text segments. To limit the amount of assessments required, for each topic only the first 7,000 characters of each response were included (according to the ranking of the snippets in the response); this is also similar to the procedure used at WebCLEF 2007. For the pool created in this way for each topic, the assessors were asked to mark text spans that either (1) repeat the information already present in the known sources, or (2) contain new important information. Unlike in the 2007 tasks, assessors were not asked to group such text snippets into subtopics (by using *nuggets*), as the 2007 assessment results proved inconsistent with respect to nuggets. The assessors used a GUI to mark character spans in the responses.

Similar to INEX [1] and to some tasks at TREC (i.e., the 2006 Expert Finding task [3]) assessment was carried out by the topic developer, i.e., by the participants themselves.

3 Runs

In total, 9 runs were submitted from 3 research groups. For reference and comparison, we also included a run generated by the best system participating in WebCLEF 2007.[1]

Table 1 shows the submitted runs with the basic statistics: the average length (the number of bytes) of the snippets in the run, the average number of snippets in the response for one topic, and the average total length of response per topic.

The evaluation results are not yet available at the time of writing but should be available at the CLEF 2008 workshop.

[1]The source code of the system is publicly available at `http://ilps.science.uva.nl/WebCLEF/WebCLEF2008/Resources`.

4 Conclusions

We detailed the task description and evaluation proceudre for the 2008 edition of WebCLEF, the multilingual web retrieval task at CLEF. At the time of writing, evaluation is still in progress.

Unfortunately, 2008 was the last year in which WebCLEF was run. The track is now being retired, due to a lack interest from the CLEF research community.

5 Acknowledgments

Valentin Jijkoun was supported by the Netherlands Organisation for Scientific Research (NWO) under project number STE-07-012. Maarten de Rijke was supported by NWO under project numbers 220-80-001, 017.001.190, 640.001.501, 640.002.501, 612.066.512, STE-07-012, 612.061.814, 612.061.815 and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

References

[1] N. Fuhr, M. Lalmas, and A. Trotman, editors. *Comparative Evaluation of XML Information Retrieval Systems: 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006*. Springer, 2007.

[2] V. Jijkoun and M. de Rijke. Overview of WebCLEF 2007. In C. Peters, V. Jijkoun, Th. Mandl, H. Müller, D.W. Oard, A. Peñas, V. Petras, and D. Santos, editors, *Proceedings of the 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Revised Selected Papers*, volume 5152 of *Lecture Notes in Computer Science*, pages 725–731, September 2008.

[3] I. Soboroff, A.P. de Vries, and N. Craswell. Overview of the TREC 2006 Enterprise Track. In *The Fifteenth Text REtrieval Conference (TREC 2006)*, 2007.