Overview of WebCLEF 2007

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

Abstract

This paper describes the WebCLEF 2007 task. The task definition—which goes beyond traditional navigational queries and is concerned with undirected information search goals—combines insights gained at previous editions of WebCLEF and of the WiQA pilot that was run at CLEF 2006. We detail the task, the assessment procedure and the results achieved by the participants.

The WebCLEF 2007 task combines insights gained from previous editions of WebCLEF 2005–2006 [6, 1] and the WiQA 2006 pilot [4, 3], and goes beyond the navigational queries considered at WebCLEF 2005 and 2006. At WebCLEF 2007 we consider so-called undirected informational search goals [5] in a web setting: "I want to learn anything/everything about my topic." A query for topic X might be interpreted as "Tell me about X."

In the remainder of the paper we detail the task, the assessment procedure and the results achieved by the participants.

1 Task description

As key starting points for defining the WebCLEF task we took several issues into account. First, the task should correspond as close as possible to some real-world information need with a clear definition of a user. Second, multi- and cross-linguality should be natural (or even essential) for the task in the CLEF setting. Next, the collection(s) used in the task should be a natural source of choice for the user's information need. Then, collections, topics and assessors' judgements, resulting from the task should be re-usable in future. Finally, the task should be challenging for the state-of-the-art IR and NLP technology.

1.1 Task model

Our hypothetical user is a knowledgable person, perhaps even an an expert, writing a survey article on a specific topic with a clear goal and audience, for example, 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 to help with this. The user does not have immediate access to offline libraries and only uses online sources.

The user formulates her information need (the topic) by specifying:

- 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 found information,
- 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; the queries may use site restrictions (see examples below) to express the user's preferences.

Here's an example of an information need:

- topic title: *Significance testing*
- description: I want to write a survey (about 10 screen pages) for undergraduate students on statistical significance testing, with an overview of the ideas, common methods and critiques. I will assume some basic knowledge of statistics.
- language(s): English
- known source(s): `http://en.wikipedia.org/wiki/Statistical_hypothesis_testing` ; `http://en.wikipedia.org/wiki/Statistical_significance`
- retrieval queries: significance testing; site:mathworld.wolfram.com significance testing; significance testing pdf; significance testing site:en.wikipedia.org

Defined in this way, the task model corresponds to addressing undirected informational search goals, that are reported to account for over 23% of web queries [5].

Each participating team was asked to develop several topics and subsequently assess responses of all participating systems for the created topics and

1.2 Data collection

In order to keep the idealized task as close as possible to the real-world scenario (i.e., there are many relevant documents) but still tractable (i.e., the size of the collection is manageable), our collection is defined per topic. Specifically, for each topic, the subcollection for the topic contains the following set of documents along with their URLs:

- all "known" sources specified for the topic;
- the top 1000 (or less, depending at the actual availability) hits from Google for each of the retrieval queries specified in the topic, or for the topic title if the queries are not specified;
- 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 conversion is only available for HTML, PDF and Postscript documents. For each document, the subcollection 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.

1.3 System response

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

2 Assessment

In order to comply with the task model, the manual assessment of the responses of the systems was done by the topic creators. The assessment procedure was somewhat similar to assessing answers to OTHER questions at TREC 2006 Question Answering task [8].

The assessment was to be blind. For a given topic, all responses of all system were pooled into anonymized sequence of text segments. To limit the amount of required assessments, for each topic only first 7,000 characters of each response were included (according to the ranking of the snippets in the response). The cut-off point 7,000 was chosen so that for at least half of the submitted runs

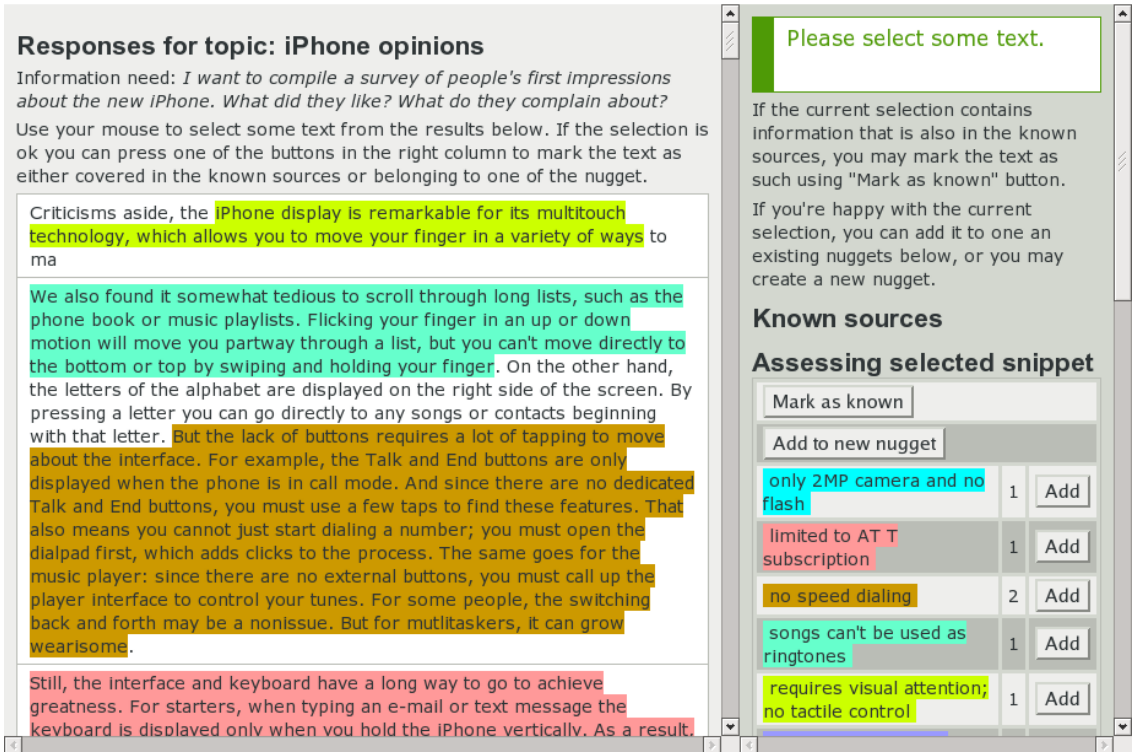

Figure 1: A screenshot of the WebCLEF 2007 assessment interface: the topic definition (top), responses of the systems with annotated character spans (left), list of nuggets created by an assessor (right).

the length of the responses was at least 7,000 for all topics. For the pool created in this way for each topic, the assessor was asked to make a list of nuggets, atomic facts, that, according to the assessor, should be included in the article for the topic. A nugget may be to character spans in the responses, so that all spans linked to one nugget express this atomic fact. Different character spans in one snippet in the response may be linked to more than one nugget. The assessors used a GUI to mark character spans in the responses and link each span to the nugget it expresses (if any). Assessors could also mark character spans as "known" if they expressed fact relevant for the topic but alredy present in one of the known sources.

Figure 1 shows the assessment interface for the topic *"iPhone opinions"*. Snippets (left bottom of the figure) are separated by grey horisontal lines. Note that only part of the first snippet is marked as relevant and the second snippet contain two marked spans that are linked to distinct nuggets. This example illustrates the many-to-many relation between nuggets and character spans in the response.

Similar to INEX [2] and to some tasks at TREC (i.e., the 2006 Expert Finding task [7]) assessment was carried out by the topic developer, i.e., by the participants themselves.

Table 1 gives the statistics for the 30 test topics and for the assessments of the topics.[1]

3 Evaluation measures

The evaluation measures for the task are based on standard precision and recall. For a given response R (a ranked list of text snippets) of a system S for a topic T we define:

[1]Full definition of the test topics is available from `http://ilps.science.uva.nl/WebCLEF/WebCLEF2007/Topics`.

Id	Topic title	Assessor	Languages	# known sources	total # snippets	# known spans	# marked spans	# chars in marked spans
1	Big Bang Theory	A	EN	3	258	2	164	36591
2	Symptoms Avian Influenza or bird flu	A	EN	2	384	0	46	12595
3	The architecture of Borobudur Temple	A	EN	2	249	0	29	12198
4	Sistemas de calefacción por Biomasa	B	ES,EN,PT,IT	6	324	2	5	3739
5	Sistemas biométricos de autentificación	B	ES,EN,PT,IT	6	241	7	17	4714
6	revistas científicas open access	B	ES,EN,PT,IT	5	341	4	13	1586
7	Magnum Opus	C	EN	1	308	3	3	765
8	Bloomsday (Band)	C	EN	1	261	6	4	596
9	Belayneh Densamo	C	EN	1	412	16	1	197
10	The Empire in Africa	C	EN	2	235	3	25	6402
11	Gaelic Games	C	EN	1	261	17	11	2706
12	schatten van voorwaardelijke kansen	D	NL	1	291	14	0	0
13	sentiment analysis for European languages other than English	D	NL,EN	1	254	4	2	450
14	European blog search engines	D	NL,EN	1	273	0	6	497
15	verhuistips	D	NL	2	238	26	4	948
16	Yasujiro Ozu	D	NL,EN	3	268	10	10	4570
17	Visa regulations regarding HIV status of applicants	E	EN,NL	0	269	0	31	6237
18	Holidays on Maldives	E	EN	0	281	0	21	1798
19	Comparison of retrieval models in Information Retrieval	E	EN	2	238	0	29	5915
20	ATM (automated teller machine) fraud	E	EN,NL	1	264	0	72	13158
21	iPhone opinions	E	EN,NL	0	290	0	35	4388
22	school education in The Netherlands	E	EN,NL	1	251	9	16	4476
23	Details on obtaining Russian turist visa for foreigners	E	EN,NL	0	285	0	39	7553
24	Albrecht Drer's "Passions" engravings and woodcuts	E	EN	1	387	0	6	807
25	Human-induced climate change: pro and cons	E	EN,NL	0	260	0	27	7345
26	Plastic tableware and the environment	E	EN,NL	0	275	0	26	3887
27	Details on interpretation of Maurice Ravel's "Gaspard de la Nuit"	E	EN	1	250	4	21	8024
28	Nabokov's "Invitation to a Beheading"	E	EN	1	258	11	20	2702
29	Durability of digital storage media	E	EN	0	279	0	9	1605
30	Madonna's books for children	E	EN	1	253	0	45	9019

Table 1: Statistics for WebCLEF 2007 topics and assessments. For each topic we show: topic id and title, author/assessor (anonymized), accepted languages, the number of "known" sources (web pages), the total number of snippets assessed from all submissions, the total number of spans marked as "known", the total number of spans attached to one of the nuggets, and the total length (the number of characters) in these spans.

Participant	Run	Average snippet length	Average snippets per topic	Average response length per topic
Baseline	Google snippets	145	898	131041
School of Computing, Dublin City University	DCU run1 simple	118	30	3552
	DCU run2 parsed	137	27	3770
	DCU run2 topfilter	112	29	3346
Faculty of Computer Science, University of Indonesia	UIWC07odwgstr	151	10	1522
	UIWC07uw10	155	10	1559
	UIWC07wstr	152	10	1530
REINA Research Group, University of Salamanca	USAL reina0.25	833	50	41680
	USAL reina0	832	50	41658
	USAL reina1	833	50	41708
ISLA, University of Amsterdam	UVA par vs	254	29	7420
	UVA par wo	277	25	7158
	UVA sent wo	214	33	7225

Table 2: Simple statistics for the baseline (Google snippets) and the 12 submitted runs.

- *recall* as the sum of character lenghts of all spans in R linked to nuggets, divided by the total sum of span lengths in the responses for T in all submitted runs.
- *precision* as the number of characters that belong to at least one span linked to a nugget, divided by the total character length of the system's response.

Note that the evaluation measures described above differ slightly from the measures originally proposed in the task description. [2] The original measures were based on the fact that spans are linked to nuggets by assessors: as described in section 2, different spans linked to one nugget are assumed to bear approximately the same factual content. Then, in addition to character-based measures above, a *nugget-based recall* can be defined based on the number of nuggets (rather than lengths of character spans) found by a system. However, an analysis of the assessments showed that some assessors used nuggets in a way not intended by the assessment guidelines: namely, to group related rather than synonymous character spans. We believe that this misinterpretation of the assessment guidelines indicates that the guidelines are overly complicated and need to be simplified in future edition of the task. As a consequence, we did not use nugget-based measures for evaluation.

4 Runs

In total, 12 runs were submitted from 4 research groups. To provide a baseline for the task, we created an artificial run: for each topic, a response of the baseline was created as the ranked list of at most 1000 snippets provided by Google in response to retrieval queries from the topic definition. Note that the Google web search engine was designed for a task very different from WebCLEF 2007 (namely, for the task of web page finding), and therefore the evaluation results of our baseline can in no way be interpreted as an indication of Google's performance.

Table 2 shows the submitted runs with the basic statistics: the average length (the number of bytes) of the snippets in the run, the averate number of snippets in the response for one topic, and the average total length of response per topic.

[2] See `http://ilps.science.uva.nl/WebCLEF/WebCLEF2007/Tasks/\#Evaluation_measures`.

Run	@ 1,500 bytes		@ 3,500 bytes		@ 7,000 bytes	
	P	R	P	R	P	R
Google snippets	0.13	0.3	0.11	0.07	0.08	0.11
DCU run1 simple	0.07	0.02	0.08	0.05	–	–
DCU run2 parsed	0.10	0.03	0.10	0.06	–	–
DCU run2 topfilter	0.08	0.02	0.08	0.04	–	–
UIWC07odwgstr	0.11	0.03	–	–	–	–
UIWC07uw10	0.09	0.02	–	–	–	–
UIWC07wstr	0.11	0.03	–	–	–	–
USAL reina0.25	0.11	0.03	0.14	0.09	0.16	0.20
USAL reina0	0.11	0.03	0.13	0.08	0.14	0.18
USAL reina1	0.11	0.03	0.14	0.09	0.16	0.21
UVA par vs	0.19	0.05	0.20	0.13	0.20	0.26
UVA par wo	0.15	0.04	0.20	0.13	0.20	0.25
UVA sent wo	0.10	0.03	0.09	0.06	0.09	0.11

Table 3: Evaluation results for the baseline (Google snippets) and the 12 submitted runs calculated at cut-off points 1,500, 3,500 and 7,000 bytes for a response for one topic.

5 Results

Table 3 shows the evaluation results for the baseline and the submitted runs: precision and recall at three different cut-off points. Since the sizes of the submitted runs varied substantially (Table 2), the cut-off points were chosen to enable comparison across runs.

Table 3 indicates that most runs outperform the baseline, or show a similar performance. Two of the runs (*UVA par vs* and *UVA par wo*) show the best performance for all cut-off points. Two other runs (*USAL reina0.25* and *USAL reina1*) show a comparable performance. We will provide a more detailed per-topic analysis and significance testing results by the time of the workshop.

One unexpected phenomenon is that for all runs (except the baseline) the precision grows as the cut-off point increases. This might indicate that although systems manage to find relevant information snippets in the collection, the ranking of the found snippets is far from optimal. This issue, however, also needs a detailed analysis.

6 Conclusions

We described WebCLEF 2007. This was the first year in which a new task was being assessed, one aimed at undirected information search goals. While the number of participants was limited, we believe the track was a success, as most submitted runs outperformed the Google-based baseline. For the best runs, in top 7,000 bytes per topic about 1/5 of the text was found relevant and important by the assessors.

The WebCLEF 2007 evaluation also raised several important issues. The task definition did not specify the exact size of a system's response for a topic, which has make a comparison across systems problematic. Furthermore, assessor's guidelines appeared to be overly complicated: not all assessors used nuggets as was intented by the organizers.

Our plans for future work include a detailed per-topic analysis of runs and a study of re-usability of the results of the assessments for improving the systems.

7 Acknowledgments

Valentin Jijkoun was supported by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 600.065.120 and 612.000.106. Maarten de Rijke was supported

by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.065.120, 612-13-001, 612.000.106, 612.066.302, 612.069.006, 640.001.501, 640.002.501, and and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

References

[1] K. Balog, L. Azzopardi, J. Kamps, and M. de Rijke. Overview of WebCLEF 2006. In *CLEF 2006*, 2007.

[2] 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.

[3] V. Jijkoun and M. de Rijke. Overview of the WiQA task at CLEF 2006. In *CLEF 2006*, 2007.

[4] V. Jijkoun and M. de Rijke. WiQA: Evaluating Multi-lingual Focused Access to Wikipedia. In T. Sakai, M. Sanderson, and D.K. Evans, editors, *Proceedings EVIA 2007*, pages 54–61, 2007.

[5] D.E. Rose and D. Levinson. Understanding user goals in web search. In *WWW '04: Proceedings of the 13th intern. conf. on World Wide Web*, pages 13–19, New York, NY, USA, 2004. ACM Press.

[6] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. Overview of WebCLEF 2005. In C. Peters, F.C. Gey, J. Gonzalo, H. Müller, G.J.F. Jones, M. Kluck, B. Magnini, and M. De Rijke, editors, *Accessing Multilingual Information Repositories*, volume 4022 of *Lecture Notes in Computer Science*, pages 810–824. Springer, September 2006.

[7] 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.

[8] E.M. Voorhees and H.T. Dang. Overview of the TREC 2005 question answering track. In *The Fourteenth Text REtrieval Conference (TREC 2005)*, 2006.