Overview of WebCLEF 2008

Valentin Jijkoun and Maarten de Rijke

ISLA, University of Amsterdam
Science Park 107, 1098 XG Amsterdam, The Netherlands
`{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, the assessment procedure, the evaluation measures and results.

The WebCLEF 2008 task is based on its 2007 predecessor [4]: 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, the assessment procedure and the results. We also give an analysis of the evaluation measures and differences between the participating systems.

1   Task Description

The user model for WebCLEF 2008 is the same as in the 2007 task definition [4]. Specifically, in our task model, our hypothetical user is a knowledgeable 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 queries; 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.

1.2 System Response

For each topic, 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 systems 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 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 [3] and to some tasks at TREC (i.e., the 2006 Expert Finding task [8]) assessment was carried out by the topic developer, i.e., by the participants themselves.

Out of the total 61 developed topics, 51 topics were actually assessed. For two of these 51 topics assessors did not find any relevant information beside the information from the known sources: topic 30 ("*Thomas Bernhard*") and topic 53 ("*Canned food in Soviet Union*"). Systems were evaluated on the remaining 49 topics.

3   Evaluation Measures

Submissions were evaluated using the following measures:

- *Average character precision (AP)*: the fraction of a system's response that matches at least one of the spans identified by assessors as relevant in the pool of all responses for a given topic; we only used alpha-numerical characters when determining substring matches, but included all characters when computing precision values;
- *Average character recall (AR)*: the sum of the character lengths of the relevant spans that are present in the system's response, divided by the total length of the relevant spans; like for precision, only alpha-numerical characters were used for substring matching, but all characters were used for computing the recall values;
- *ROUGE-1* and *ROUGE-1-2*: the values of the ROUGE evaluation metric [5] computed on word unigrams (ROUGE-1) and word unigrams and bigrams, (ROUGE 1-2); in a nutshell, ROUGE-*n* measures *n*-gram recall: the fraction of *n*-grams of the relevant spans that were found by a system; we excluded stopwords from the ROUGE evaluation.

Similarly to the WebCLEF 2007 task, for a system's response for a given topic, we computed all measures on the first 7,000 bytes of the response.

4   Approaches and Evaluation Results

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

The University of Twente [6] developed three modifications of the baseline, including bugfixes in the baseline's software (namely, in stopword removal). The University of Salamanca [2] implemented three versions of query formulation for estimating query relevance: using only the topic description, using terms extracted from the known sources of the topic, and using only English words from known sources. Finally, UNED [1] extended the baseline with a key term extraction, relevance-based document re-ranking and a method for eliminating cross-lingual redundancy.

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; we also show the four evaluation measures for the runs: average precision, average recall, ROUGE-1 and ROUGE-1-2.

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

Participant	Run	Average snippet length	Average snippets per topic	Average response length	AP	AR	ROUGE 1	ROUGE 1-2
	baseline 2007	286	20	5,861	0.08	0.07	0.14	0.05
U. Twente	ip2008	450	32	14,580	0.23	0.23	**0.20**	0.07
	ipt2008	464	31	14,678	**0.24**	**0.24**	0.19	**0.08**
	ipu2008	439	33	14,607	0.21	0.21	0.17	0.07
UNED	Uned RUN1	594	24	14,817	0.23	0.21	0.18	0.06
	Uned RUN2	577	25	14,879	0.18	0.18	0.18	0.05
	Uned RUN3	596	24	14,861	0.21	0.19	0.18	0.05
U. Samalanca	usal 0	851	91	77,668	0.21	0.23	0.17	0.06
	usal 1	1,494	86	129,803	0.11	0.09	0.16	0.06
	usal 2	1,427	88	126,708	0.09	0.09	0.15	0.05

We see that the best performing runs improve substantially over the baseline run (which was the best performing system in 2007), according to all measures. We looked at the statistical significance of the differences in precision (AP) using the paired two-tailed t-test with $p = 0.05$. There are two groups of statistically indistinguishable runs (when considering AP): {baseline, usal 1, usal 2} and {ip2008, ipt2008, ipu2008, Uned RUN1, Uned RUN2, Uned RUN3, usal 0}. Although all systems improve over the baseline, it is impossible to tell reliably which individual approach gives the best performance.

When we look at the per topic breakdown of the (precision) scores, we see a mixed story. Figure 1 shows the precision scores of the submitted runs for individual topics. On

---

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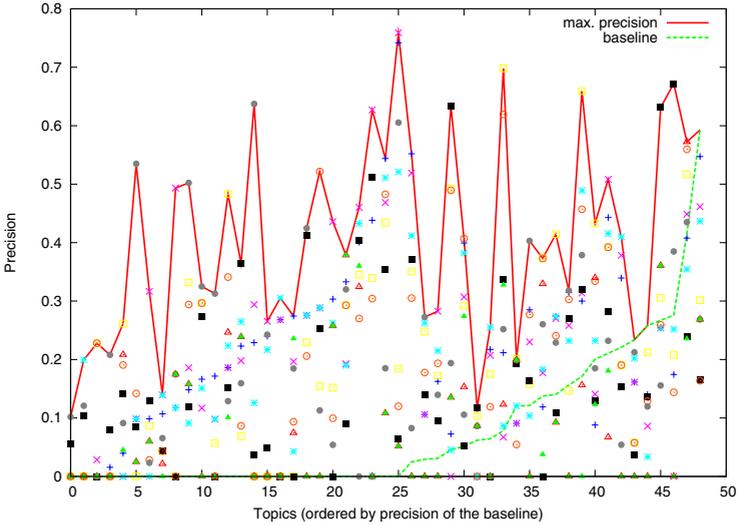http://ilps.science.uva.nl/WebCLEF/WebCLEF2008/Resources.

Fig. 1. Precision for the 49 non-empty topics. Points give precision values for the 9 submitted runs; lines show the maximum precision value for each topic and the precision of the baseline.

many topics, runs that perform poorly on average outperform runs that perform best (on average). Also, there is no run that outperforms all other runs on all (or even on most) topics—this is in line with our observation of a large set of statistically indistinguishable runs.

5   Discussion

In this section we take a brief look at the evaluation measures used at WebCLEF 2008. Figure 2 shows the values of the four evaluation measures for all runs. Clearly, the correlation between different measures is far from perfect. The measures generally agree on the best and worst runs, but the ranking of the runs in the middle is less unanimous.

Table 2 shows Kendall's rank correlation coefficient for the pairs of measures (values close to 1 mean that the two measures rank the runs similarly, values close to 0 indicate no correlation between measures). Note the relatively low correlation between the two ROUGE measures and between precision/recall and ROUGE. Since precision and

Table 2. Kendall's rank correlation coefficient for agreement between evaluation measures

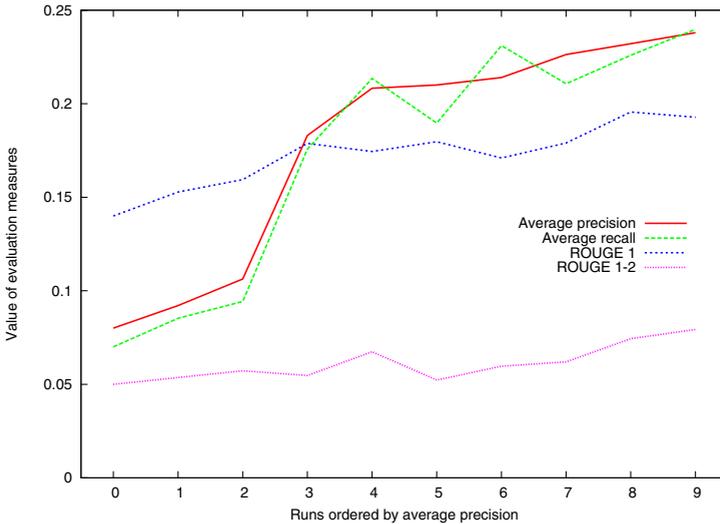|	AP	AR	ROUGE 1	ROUGE 1-2
AP	–	0.82	0.73	0.69
AR	0.82	–	0.56	0.69
ROUGE 1	0.73	0.56	–	0.51
ROUGE 1-2	0.69	0.69	0.51	–

Fig. 2. Values of the evaluation measures for the baseline and the 9 submitted runs (runs ordered by the average precision)

recall are computed straightforwardly from human assessments (in every run, assessors mark up relevant character spans), we conclude that while ROUGE is successfully used in tasks such as summarization or machine translation, it is not fully appropriate for evaluating the WebCLEF task. This is unfortunate, because, as [6] argues, the strict precision/recall-based evaluation of the task does not allow us to reuse the human judgements for evaluating runs that humans have not assessed directly. As a consequence, it is virtually impossible to create a proper test collection for the task.

6   Conclusions

We detailed the task description and evaluation procedure for the 2008 edition of Web-CLEF, the multilingual web retrieval task at CLEF. In 2008, participating systems showed substantial improvements over the best system from 2007 (that was used a baseline). For the best 2008 system, on average 24% of its output is judged relevant by human assessors (compared to 8% for the 2007 baseline). However, all runs with a reasonable performance are statistically indistinguishable from each other. Moreover, we found that the ROUGE measure, often used in machines translation and summarization, is not directly applicable for the evaluating the task.

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.

Acknowledgments

References

1. Amigo, E., Martinez-Romo, J., Araujo, L., Peinado, V.: UNED at WebCLEF 2008: Applying High Restrictive Summarization, Low Restrictive Information Retrieval and Multilingual Techniques. In: Peters, et al [7]
2. Figuerola, C., Berrocal, J., Rodriguez, A., Mateos, M.: Retrieval of snippets of Web pages converted to plain text. More questions than answers. In: Peters, et al. [7]
3. Fuhr, N., Lalmas, M., Trotman, A. (eds.): INEX 2006. LNCS, vol. 4518. Springer, Heidelberg (2007)
4. Jijkoun, V., de Rijke, M.: Overview of WebCLEF 2007. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 725–731. Springer, Heidelberg (2008)
5. Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Proceedings of Workshop on Text Summarization Branches Out, WAS 2004 (2004)
6. Overwijk, A., Nguyen, D., Hauff, C., Trieschnigg, R., Hiemstra, D., de Jong, F.: On the Evaluation of Snippet Selection for WebCLEF. In: Peters, et al. [7]
7. Peters, C., et al.: CLEF 2008. LNCS, vol. 5706. Springer, Heidelberg (2009)
8. Soboroff, I., de Vries, A., Craswell, N.: Overview of the TREC 2006 Enterprise Track. In: The Fifteenth Text REtrieval Conference, TREC 2006 (2007)