WiQA: Evaluating Multi-lingual Focused Access to Wikipedia

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

Abstract

We describe our experience with WiQA 2006, a pilot task aimed at studying question answering using Wikipedia. Going beyond traditional factoid questions, the task considered at WiQA 2006 was to identify—given an source article from Wikipedia—snippets from other Wikipedia articles, possibly in languages different from the language of the source article, that add new and important information to the source article, and that do so without repetition.

A total of 7 teams took part, submitting 20 runs. Our main findings are two-fold: (i) while challenging, the tasks considered at WiQA are do-able as participants achieved precision@10 scores in the .5 range and MRR scores upwards of .5; (ii) on the bilingual task, substantially higher scores were achieved than on the monolingual tasks.

1 Introduction

With new types of online content growing rapidly in size and importance, retrieval evaluation platforms are setting up new tasks or tracks around these types of content. E.g., TREC 2006 featured a new blog track, and CLEF 2006 featured a pilot on Question Anwering Using Wikipedia, or WiQA [5], for short.

The idea to organize a pilot track on focused information access using Wikipedia grew from several motivations. First, traditionally, people turn to reference works to get answers to their questions. Wikipedia has become one of the largest reference works ever, making it a natural target for question answering systems. Wikipedia is also an excellent multilingual resource, providing open domain content in 250 languages, including 14 languages with more than 100,000 articles (as of March 2007). Moreover, Wikipedia is a rich mixture of text, link structure, navigational aids, categories, ..., making it extremely appealing for link analysis, text mining, information extraction and information retrieval work. And finally, Wikipedia is simply a great resource. It is something we as researchers want to work with, and contribute to, both by facilitating access to it, and, as the distinction between readers and authors has become blurred, by creating tools to support the authoring process.

One of the aims of the WiQA 2006 pilot was to set up a challenging, but do-able and measurable information access task using Wikipedia. Another was to experiment with different measures for evaluation within this setting. In this overview we first provide a description of the task we selected for the pilot and of the evaluation and assessment procedures (Section 2). After that we describe the runs submitted by the participants (Section 3) and then we detail the results (Section 4). We end with conclusions (Section 5).

2 The Task

Given the properties of Wikipedia outlined above, one can envisage many possible information access tasks including Wikipedia. While defining an information access task suitable for WiQA 2006, we tried to accomodate two possibly conflicting constraints:

- We wanted to define a real-world task such that its effective solution may lead to a useful and useable tool for a real-world problem.
- The task should be clearly defined, the performance of the systems should be measurable and results of manual assessments of systems' output (if manual assessments are needed) should be reusable for automatic evaluation of future systems.

The task we chose for the WiQA 2006 pilot deals with access to Wikipedia's content, where access is considered from the point of view of both *reader* and *author* of articles. We situate our task in the following scenario: a reader or author of a given Wikipedia article (the source article) is interested in collecting information about the topic of the article that is not yet included in the text, but

is relevant and important for the topic. The collected information, for example, can be used to get a broader view on the topic or to update and/or expand the content of the source article. Although the source article is in a specific language (the source language), the reader or author would also be interested in finding information in other languages (the target languages) that she explicitely specifies.

With this user scenario, the task of an automatic system participating in WiQA 2006 is to locate information snippets in Wikipedia which are:

- outside the given source article,
- in one of the specified target languages,
- substantially new w.r.t. the information contained in the source article, and important for the topic of the source article, in other words, worth including in the content of (the future editions of) the article.

One specific application of the task defined in this way can be a system that helps a Wikipedia editor to update or expand an article using the information available elsewhere.

Participants of the WiQA 2006 pilot could take part in two flavors of the task: a monolingual one (where the snippets to be returned are in the language of the source article) and a multilingual one (where the snippets to be returned can be in any of the languages of the Wikipedia corpus used at WiQA).

The input of an automatic system was a topic (i.e., a source Wikipedia article) and a set of allowed target languages. The output of the system is a list of snippets (sentences) from Wikipedia articles in any of the target languages.

2.1 Document Collections

The data collection used at WiQA 2006 consists of XML-ified dumps of Wikipedia in three language: Dutch, English, and Spanish. The three collections differ greatly in size:

- Dutch: 125,004 articles, 857Mb;
- English: 660,762 articles, 5.9Gb; and
- Spanish: 79,237 articles, 677Mb.

The size of the collection is an important factor for the performance of a system addressing the WiQA 2006 task. However, the effect of the corpus size on the difficulty of the task is not obvious:

- on the one hand, a larger collection could potentially provide more additional information snippets about a topic;

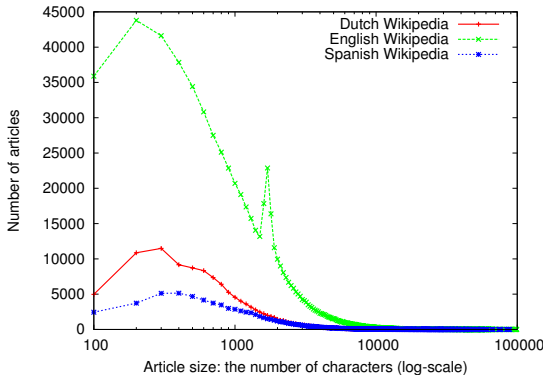

Figure 1. Distribution of the sizes of the articles in the Dutch, English, and Spanish Wikipedia corpora used at WiQA 2006.

- on the other hand, with a larger collection it may be more difficult to filter out noise, i.e., information snippets not important or even not relevant to a topic.

Figure 1 shows the distribution of the article size in Wikipedia of the three languages.

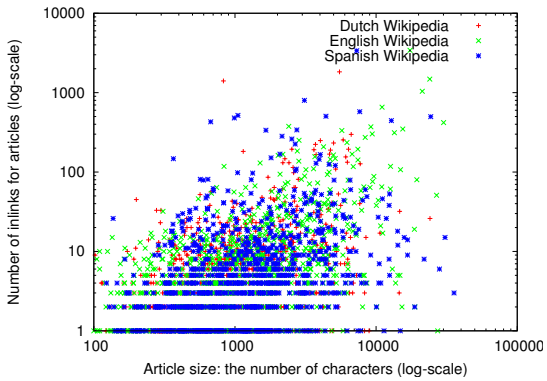

Figure 2. Distribution of the sizes of the articles and the number of incoming links for the Dutch, English, and Spanish Wikipedia collections

Another intrinsic characteristic of a collection important for the WiQA task is the distribution of hyperlinks between articles. Articles linking to a topic are likely to provide some information about the topic. The number of such incoming links (*inlinks*) can be considered as a rough indication of how much relevant information is available in the rest of Wikipedia for a given topic. Figure 2 shows how the number of inlinks correlates with article size. The relation between these parameters is similar for the three languages.

The Wikipedia dumps used at WiQA are based on the XML version of the Wikipedia collections [2] that include the annotation of the structure of the articles, links between articles, categories, cross-lingual links, etc. For the pilot the annotation of articles was automatically extended with XML markup of sentences and classification of articles into named entity classes (person, location, organization). The classification was done in an ad-hoc way using a set of heuristics that employ the category structure of Wikipedie and the uniform structure of "List" articles (e.g., articles entitled *List of living persons*, *List of physicists*, etc.). The table below shows the distribution of the assigned classes in the collection.

Coll.	*person*	*loc*	*org*
English	84,167 (13%)	50,940 (8%)	22,654 (3%)
Spanish	11,009 (14%)	3,980 (5%)	1,292 (2%)
Dutch	10,176 (8%)	7,038 (6%)	1,595 (1%)

We performed a manual assessment of the classes assigned for a random sample of the articles: our heuristic rules resulted in 85% accuracy.

2.2 Topics

For each of the three WiQA 2006 languages (Dutch, English, Spanish) a set of 50 topics correctly tagged as *person*, *loc* or *org* in the XML data collections was released, together with other topics, announced for the participants as optional. These optional topics either did not fall into these three categories, or were not tagged correctly in the XML collections. The optional topics could be ignored by systems without penalty. In fact, all submitted runs provided responses for optional topics as well as for the main topics.

When selecting Wikipedia articles as topics, we included articles marked as stubs,[1] as well as other short and long articles. With all constraints mentioned above, topics for the monolingual tasks were selected by random sampling from the corresponding collections, to ensure that the distrubution of other properties of topic articles (such as size and number of links) follows that of the entire collection (see Figures 1 and 2).

In order to create the topics for the English-Dutch bilingual task, 30 topics were selected from the English monolingual topic set and 30 topics from the Dutch monolingual topic set. The bilingual topics were selected so that the corresponding articles[2] are present in Wikipedia for both languages.

[1] In Wikipedia, a *stub* is an article explicitely marked as requiring update or expansion.

[2] In Wikipedia, versions of articles on the same topic in different languages are indicated explicitly. Articles about the same topic may differ significantly across languages.

The table below shows the number of topics for the four language subtasks.

Task	total	*per*	*loc*	*org*	*other*
English	65	16	18	16	15
Dutch	60	17	16	17	10
Spanish	67	21	22	18	6
English-Dutch	60	18	16	17	9

In addition to the test topics, a set of 80 development topics was released for English.

2.3 Evaluation

Given a source article, an participating system had to return a list of short snippets, defined as sequences of at most two sentences from Wikipedia articles. The ranked list of snippets for the topic were manually assessed using the following binary criteria, inspired by the TREC 2003 Novelty task [4]:

- *support*: the snippet does indeed come from the specified target Wikipedia article, different from the source article;
- *importance*: the information of the snippet is relevant to the topic of the source Wikipedia article, is in one of the target languages as specified in the topic, and is already present on the article (directly or indirectly) or is interesting and important enough to be included in an updated version of the article;
- *novelty*: the information content of the snippet is not subsumed by the information in the source article;
- *non-repetition*: the information content of the snippet is not subsumed by the target snippets higher in the ranked list of snippets produced so far for the given topic.

Note that we distinguish between novelty (subsumption by the source article) and non-repetition (subsumption by the higher ranked snippets) in order for the results of the assessment to be reusable for future automatic system evaluation: novelty only takes the source article and the returned snippet into account, while non-repetition can only be defined for a given ranked list of snippets. Thus, while novelty and importance assessments are reusable between different runs, non-repetition assessments are not.

One of the purposes of the WiQA pilot task was to experiment with different measures for evaluating the performance of systems. WiQA 2006 used the following simple principal measure for accessing the performance of the systems:

- *yield*: the number of supported, novel, non-repetitive, important target snippets, averaged by topic.

We also considered other simple measures:

- *mean reciprocal rank* of the first supported, important, novel, non-repeated snippet, and
- *overall precision*: the percentage of supported, novel, non-repetitive, important snippets among all submitted snippets.

Notice that slightly modified versions of the three measures above can be reused for an automatic system evaluation by ignoring "non-repetition" part (as we will explain below).

2.4 Assessment

To establish the ground truth for the WiQA task, an assessment environment was developed by the task organizers. Individual assessors were not required to provide assessments for all topics,[3] and could choose topics themselves. We made sure that each topic was assessed at least once; several topics received multiple assessments (see Section 4.1 for details). Assessors were given the following instructions. For each system and each source article P the ordered list of the returned snippets was to be manually assessed with respect to importance, novelty and non-repetition following the procedure below:

1. Each snippet was marked as *supported* or not. To reduce the workload on the assessors, this aspect was checked automatically. Hence, unsupported snippets were excluded from subsequent assessments.
2. Each snippet was marked as *important* or not with respect to the topic of the source article. A snippet is important if it contains information that the assessor would like to see in the article P (as a hypothetical author or reader of the article). Snippets were assessed for importance independently of each other and regardless of whether the important information was already present in P (in particular, presence of some information in P did not necessarily imply its importance).
3. Each important snippet was marked as *novel* or not. Assessors were instructed to consider the snippet novel if the important information in the snippet is substantially new with respect to the content of P.
4. Each important and novel snippet had to be marked as repeated or non-repeated with respect to the important snippets higher in the ranked list of snippets.

[3] This was done to reduce load on our assessors.

Following this procedure, snippets were assessed along the four axes (support, importance, novelty, non-repetition). Assessors were not required to judge novelty and non-repetition of snippets that are considered not important for the topic of the source article. The reason for this was to avoid spending time on assessing irrelevant information. Assessors provided assessments for the top 20 snippets for each result list returned. Figure 3 contains a screen shot of the assessment interface.

A total number of 14,203 snippets submitted by the participants had to be assessed. The number of unique snippets assessed was 4,959. Of these, 3,396 were assessed by at least two assessors.

2.5 Submission

For each task (three monolingual and one bilingual), participating teams were allowed to submit up to three runs. For each topic of a run, the top 20 submitted snippets were manually assessed as described above.

3 Submitted Runs

In total, 20 runs were submitted for evaluation:

- 19 run for the monolingual task: 2 for Dutch, 12 for English, and 4 for Spanish;
- 1 run for the bilingual task (English-Dutch).

Most participating systems used a similar three-step architecture: first, identify snippets relevant to the topic, then estimate their importance, and finally, remove duplicate or near-duplicate snippets. However, there was a lot of variation in the wide range of techiques for addressing individual steps:

- For *identifying relevant snippets* outside the source article, systems used traditional IR (with the title of the source articles as a query), string matching, or made use of the in-links of the article;
- For *estimating the importance of a snippet* the systems employed word overlap, as well as Latent Semantic Analysis, Information Gain or they used the category structure of Wikipedia;
- For *removing redundant snippets* the systems used word overlap, cosine similarity, Information Gain as well as Named Entity identification.

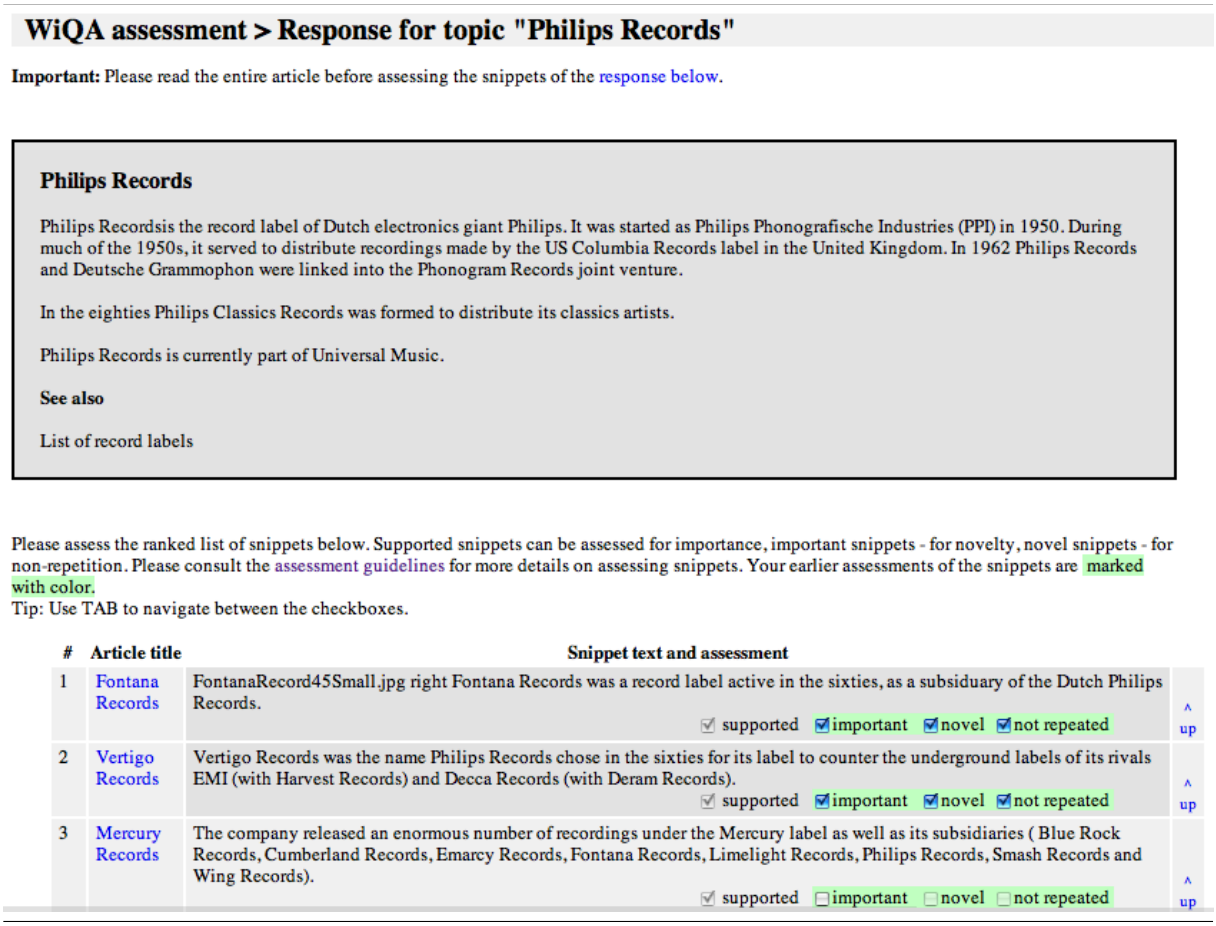

Figure 3. Assessment interface; first snippets of a system's response for topic wiqa06-en-39.

From the text processing perspective, the systems were very diverse. Participants employed techniques ranging from Named Entity tagging to parsing, logic form identification, coreference resolution and machine translation (using Wikipedia as a training resource for translating proper names between languages). For further details of the individual systems, we refer to the official CLEF 2006 proceedings [1].

4 Results

In Table 1 we present the aggregate results of the assessment of the runs submitted to WiQA 2006. Columns 3–7 show the following aggregate numbers: total number of snippets (with at most 20 snippets considered per response for a topic); total number of supported snippets; total number of important supported snippets; total number of novel and important supported snippets; and the total number of novel and important supported non-repeated snippets.

The results indicate that the task of detecting *important* snippets is a hard one: for most submissions, only 50–60% of the found snippets are judged as important. The performance of the systems for detecting *novel* snippets has a substantially higher range: between 50% and 80% of the found important snippets are judged as novel with respect to the topic article.

Table 2 shows the evaluation results for the submitted runs: total yield (for a run, the total number of "perfect" snippets, i.e., supported, important, novel and not repeated), the average yield per topic (only topics with at least one response are considered), the mean reciprocal rank of the first support important novel snippet and the precision of the systems' responses. Due to space limitations, we only show the runs that performed best according to one of the evaluation measures. A detailed analysis of the runs is presented in [3].

Most systems cope well with the pilot task: up to one third of the found snippets are assessed as supported, important, novel and non-repeated for the English and Spanish monolingual tasks, and up to one half for the Dutch monolingual and the English-Dutch bilingual task. Quite expectedly, the relative ranking of the submitted runs

Table 1. Results of the assessments of the submitted runs (only the best runs for each task are shown; highest scores are in bold).

Run	Number of topics with response	Aggregate numbers of snippets (with at most 20 snippets considered per topic).				
		total	supp	supp imp	supp imp novel	supp imp novel not-rep
English monolingual task: 65 topics						
run1	**65**	435	435	226	165	161
run2	**65**	**615**	**614**	**353**	232	**220**
run3	61	526	526	327	**280**	135
Spanish monolingual task: 67 topics						
run4	62	**497**	**497**	**198**	**142**	**113**
run5	**67**	251	251	127	79	69
Dutch monolingual task: 60 topics						
run6	**60**	**455**	**455**	**305**	**236**	**228**
English-Dutch bilingual task: 60 topics						
run7	**60**	**564**	**551**	**456**	**342**	**302**

is different for different evaluation measures: as in many complex tasks, the best yield (a recall-oriented measure) does not necessarily lead to the best precision and vice versa.

Figure 4 shows the performance of all systems for English monolingual topics: minimum, average and maximum yield (i.e., the number of returned important supported non-repeated snippets) for all 65 English topics. As the plot indicates, some topics are clearly harder than others. We did not observe significant correlations between obvious parameters of topics (such as the size of the Wikipedia page and the number of in-links) and topic difficulty.

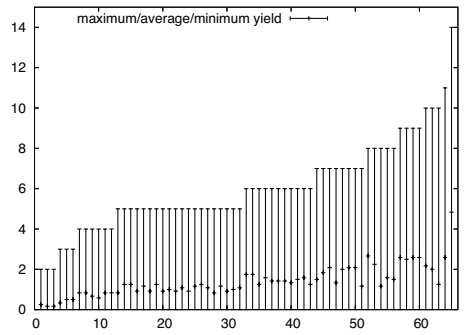

Figure 4. Per-topic performance of all the systems on the English monolingual topics.

An interesting aspect of the results is that the performance of the systems differs substantially for the four tasks. This may be due to the fact that the submissions to WiQA 2006 were assessed by different assessors (native speakers of the corresponding languages), or it may be due to the differences in the sizes and structures of the Wikipedias in these languages. Also, it is worth pointing out that the highest scores were achieved on the English-Dutch bilingual task; this may suggest that different language versions of Wikipedia do indeed present different material on a given topic. We can conclude that, unlike most other retrieval tasks, the bilingual WiQA seems easier that monolingual. This is not surprising if we pause to consider the definition of the task: retrieving *novel* aspects of a topic.

4.1 Inter-annotator agreement

The definition of the WiQA task is quite complicated and the criteria for snippet assessment may be very subjective. To examine this issue, we arranged the assessments so that a portion of the snippets was assessed by two annotators.

Table 3 shows the agreement of pairs of assessors on importance judgments calculated using Cohen's κ. We see that the κ values vary between 0.13 (with an agreement of 56%) to 0.71 (with an agreement of 86%), while and most are above 0.4. This indicates a less than perfect correlation between assessors' judgements, which is also supported by the feedback from the assessors: they often found the *importance* judgements subjective and hard to make.

Table 4 shows overall assessor agreement for all

Table 2. Evaluation results for the submitted runs (calculated for top 10 snippets per topic); only best runs for each task are shown; highest scores are given in boldface.

Run	Number of topics with response	Total yield	Average yield per topic	MRR	Precision
English monolingual task: 65 topics					
run1	65	161	2.46	0.54	**0.37**
run2	65	**220**	**3.38**	0.58	0.36
run3	61	135	2.21	**0.59**	0.26
Spanish monolingual task: 67 topics					
run4	62	**113**	**1.82**	**0.37**	0.23
run5	67	69	1.03	0.30	**0.27**
Dutch monolingual task: 60 topics					
run6	60	**228**	**3.80**	**0.53**	**0.50**
English-Dutch bilingual task: 60 topics					
run7	60	**302**	**5.03**	**0.52**	**0.54**

Pair	**#common**	**Agreement %**	κ
A,B	91	75%	0.49
C,D	242	86%	0.71
C,B	212	77%	0.52
C,A	77	70%	0.38
D,B	573	72%	0.45
D,E	147	56%	0.13
D,A	46	78%	0.57
F,G	643	73%	0.42

Table 3. Agreement for pairs of assessors on *importance* judgements

Judgement	**#snippets**	**Agreement %**
importance	1428	73%
novelty	1428	73%
non-repetition	2806	72%

Table 4. Overall agreement on all doubly-assessed snippets

cases where double assessmets were available: for each such snippet we randomly picked two of the available judgements and checked whether they are the same.

5 Conclusion

We have described the WiQA 2006 pilot: Question Answering Using Wikipedia. Set up as an attempt to take question answering beyond the traditional factoid format and to one of the most interesting knowledge sources currently available, WiQA had 8 participants who submitted a total of 20 runs for 4 tasks. The results of the pilot are very encouraging. While challenging and from being solved, the task turned out to be do-able; several participants managed to achieve precision scores in the 0.3–0.5 range and MRR scores upward of 0.5 (meaning that, on average, they returned the first snippet on rank 1 or 2). Surprisingly, the highest scores were achieved on the bilingual task.

The WiQA 2006 pilot has shown that it is possible to set up tractable yet challenging information access tasks involving the multilingual Wikipedia corpus—but this was only a first step. The next edition of the task will be integrated with the WebCLEF task; the likely scenario to be studied there will be one where an author is writing a an article (or paper, or survey, ...), and needs to collect information to be included in the article; using one or more Wikipedia articles as a starting, the additional information is to be gathered from the web (in a fixed collection of crawled web pages) in addition to Wikipedia articles.

Finally, the collections, topics and the assessed runs of the participants of WiQA are available at `http://ilps.science.uva.nl/WiQA`.

6 Acknowledgments

We are very grateful to the following people and organizations for helping us with the assessments: José Luis Martínez Fernández and César de Pablo from the Daedalus consortium; Silke Scheible and Bonnie Webber at the University of Edinburgh; Udo Kruschwitz and Richard Sutcliffe at the University of Essex; and Bouke Huurnink and Maarten de Rijke at the University of Amsterdam.

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] CLEF 2006 Working Notes, 2006. Working Notes for the CLEF 2006 Workshop URL: http://www.clef-campaign.org/2006/working_notes/CLEF2006WN-Contents.htm%l.

[2] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.

[3] V. Jijkoun and M. de Rijke. Overview of WiQA 2006. In A. Nardi, C. Peters, and J. Vicedo, editors, *Working Notes CLEF 2006*, September 2006.

[4] I. Soboroff and D. Harman. Overview of the TREC 2003 Novelty track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 38–53. NIST, 2003.

[5] WiQA, 2006. Question Answering Using Wikipedia URL: http://ilps.science.uva.nl/WiQA/.