# Overview of the WiQA Task at CLEF 2006

Valentin Jijkoun and Maarten de Rijke

ISLA, University of Amsterdam
{jijkoun,mdr}@science.uva.nl

**Abstract.** We describe 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 return—given an source page from Wikipedia—to identify snippets from other Wikipedia pages, possibly in languages different from the language of the source page, that add new and important information to the source page, 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 impressive scores as measured in terms of yield, mean reciprocal rank, and precision, (ii) on the bilingual task, substantially higher scores were achieved than on the monolingual tasks.

## 1 Introduction

CLEF 2006 featured a pilot on Question Anwering Using Wikipedia, or WiQA, for short. The idea to organize a pilot track on QA using Wikipedia builds on 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. Moreover, Wikipedia is a rich mixture of text, link structure, navigational aids, categories, making it extremely appealing for research on text mining and link analysis. And finally, Wikipedia is simply a great resource. It is something we 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.

In this overview we first provide a description of the tasks considered and of the evaluation and assessment procedures (Section 2). After that we describe the runs submitted by the participants (Section 3 and detail the results (Section 4). We end with some conclusions (Section 5).

## 2 Task Definition

The WiQA 2006 task deals with access to Wikipedia's content, where access is considered both from a point of view and from an author point of view.

As our user model we take the following scenario: a reader or author of a given Wikipedia article (the source page) is interested in collecting information

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 265–274, 2007.

about the topic of the page that is not yet included in the text, but is relevant and important for the topic, so that it can be used to update the content of the source article. Although the source page 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 he explicitely specifies.

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

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

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 page) and a multilingual (where the snippets to be returned can be in any of the languages of the Wikipedia corpus used at WiQA).

### 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 collection and the links structure are important factors for the performance of a system addressing the WiQA 2006 task. Figure 2.1 shows the distribution of the article size and the number of in-links in Wikipedia of the three languages.

The Wikipedia dumps used at WiQA are based on the XML version of the Wikipedia collections [Denoyer and Gallinari(2006)] 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 using a set of heuristics that employ the category structure of Wikipedia 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.

| Collection | *person* | *location* | *organization* |
|---|---|---|---|
| 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.

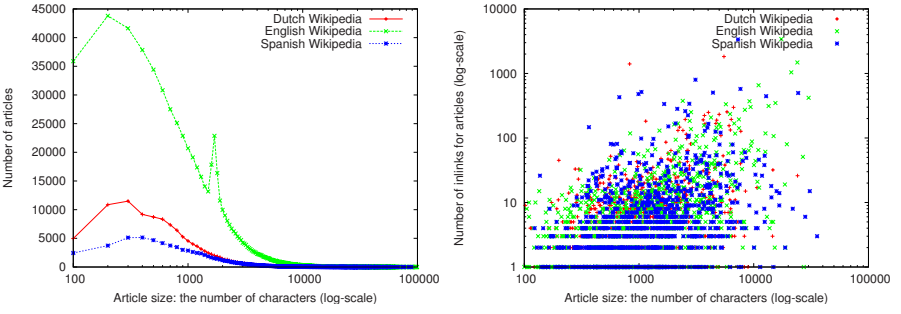


**Fig. 1.** Distribution of the sizes of the articles and the number of in-links in Dutch, English, and Spanish Wikipedia

### 2.2 Topics

For each of the three WiQA 2006 languages (Dutch, English, and Spanish) a set of 50 topics correctly tagged as *person*, *location* or *organization* in the XML data collections was released, together with other topics, announced 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, the submitted runs provided responses for optional topics as well as for the main topics.

When selecting Wikipedia articles as topics, we included articles explicitely marked as stubs, as well as other short and long articles.

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 are present in Wikipedias for both languages.

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

| Task | total | *person* | *location* | *organization* | *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 (English language) development topics was released.

### 2.3 Evaluation

Given a source page, automatic systems return a list of short snippets, defined as sequences of at most two sentences from a Wikipedia page. The ranked list of snippets for the topic were manually assessed using the following binary criteria, largely inspired by the TREC 2003 Novelty task [Soboroff and Harman(2003)]:

- *support*: the snippet does indeed come from the specified target Wikipedia 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 page (directly or indirectly) or is interesting and important enough to be included in an updated version of the page.
- *novelty*: the information content of the snippet is not subsumed by the information on the source page
- *non-repetition*: the information content of the snippet is not subsumed by the target snippets higher in the ranking for the given topic

Note that we distinguish between novelty (subsumption by the source page) and non-repetition (subsumption by the higher ranked snippets) in order for the results of the assessment to be re-usable for automatic system evaluation in future: novelty only takes the source page and the snippet into account, while non-repetition is defined on a ranked list of snippets.

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 main measure for accessing the performance of the systems:

- *yield*: the average (per topic) number of supported, novel, non-repetitive, important target snippets.

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.

### 2.4 Assessment

To establish the ground truth, an assessment environment was developed by the track organizers. Assessors were given the following guidelines. For each system and each source article $P$ the ordered list of the returned snippets was 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 the subsequent assessment.
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 a user of Wikipedia would like to see in $P$ or an author would consider worth to be present in $P$. Snippets were assessed for importance independently of each other and regardless of whether the important information was already

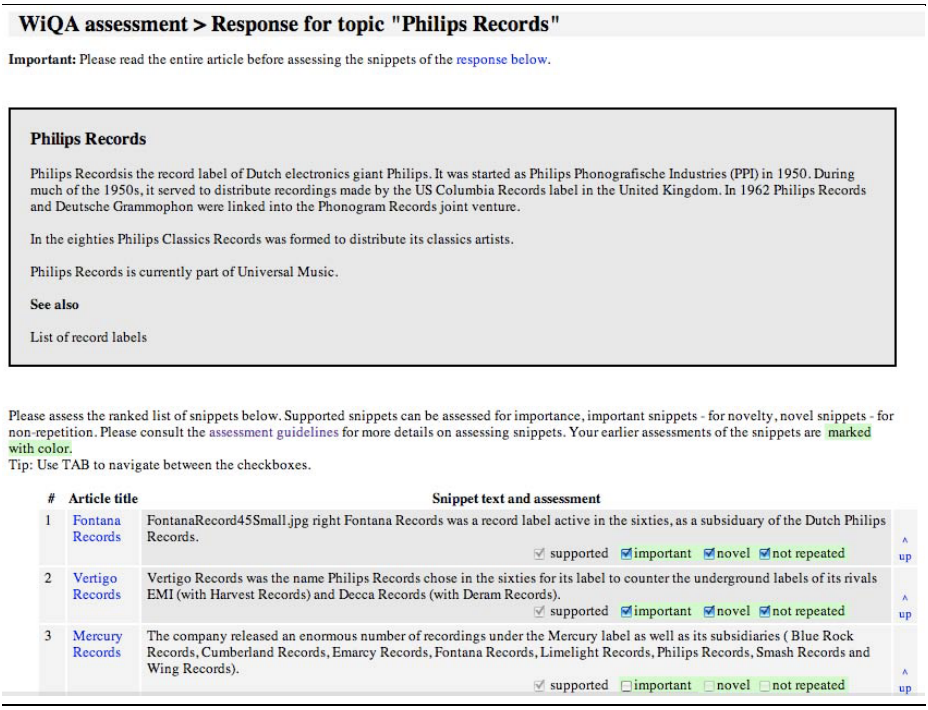


**Fig. 2.** Assessment interface; first three snippets of a system's response for topic wiqa06-en-39

present in $P$ (in particular, presence of some information in $P$ does not necessarily imply its importance).

3. Each important snippet was marked as *novel* or not. It was to be considered novel if the important information in the snippet is substantially new with respect to the content of $P$.
4. Each important and novel snippet was marked as repeated or non-repeated, with respect to the important snippets higher in the ranked list of snippets.

Following this procedure, snippets were assessed along 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 much time on assessing irrelevant information. Assessors provided assessments for the top 20 snippets for each result list returned. Figure 2 contains a screen shot of the assessment interface.

A total number of 14203 snippets had to be assessed; the number unique snippets assessed is 4959. Of these, 3396 were assessed by at least two assessors.

**Table 1.** Summary of runs submitted to WiQA 2006

| Group | Run name | Description |
|---|---|---|
| **English monolingual task** | | |
| LexiClone Inc. | lexiclone | Lexical Cloning method |
| Universidad Politécnica de València | rfia-bow-en | simple "bag of words" submission |
| University of Alicante | UA-DLSI-1 | Near phrase |
| | UA-DLSI-2 | Near phrase temporal |
| University of Essex/Limerick | dltg061 | Limit of ten snippets per topic |
| | dltg062 | Limit of twenty snippets per topic |
| University of Amsterdam | uams-linkret-en | Cross-links and IR for snippet ranking |
| | uams-link-en | Only cross-links for snippet ranking |
| | uams-ret-en | Only IR for snippet ranking |
| University of Wolverhampton | WLV-one-old | No coreference, link analysis |
| | WLV-two | Coreference |
| | WLV-one | no coreference, version 2 |
| **Spanish monolingual task** | | |
| Universidad Politécnica de València | rfia-bow-es | simple "bag of words" submission |
| University of Alicante | UA-DLSI-es | Near phrase |
| Daedalus consortium | mira-IS-CN-N | InLink sentence retr., rank by novelty |
| | mira-IP-CN-CN | InLink passage retrieval, combine cosine and novelty in ranking, no threshold |
| **Dutch monolingual task** | | |
| University of Amsterdam | uams-linkret-nl | Cross-links and IR for snippet ranking |
| | uams-link-nl | Only cross-links for snippet ranking |
| | uams-ret-nl | Only IR for snippet ranking |
| **English-Dutch bilingual task** | | |
| University of Amsterdam | uams-linkret-ennl | Cross-links and IR for snippet ranking |

### 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

Table 1 lists the runs submitted to WiQA 2006: 19 for the monolingual task (3 for Dutch, 12 for English and 4 for Spanish) and 1 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.

From the text processing perspective, the systems were also 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 system descriptions in [CLEF 2006 Working Notes(2006)].

In Table 2 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 with repetition.

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.

## 4 Results

Table 3 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 "perfect" snippet and the precision of the systems' responses.

Clearly, most systems cope well with the pilot task: up to one third of the found snippets are assessed as "perfect" 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 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.

An interesting aspect of the results is that the performance of the systems differs substantially for the four tasks. This can be due to the fact that the submissions for tasks were assessed by different assessors (native speakers of

**Table 2.** Results of the assessment of the submitted runs (at most 20 snippets considered per topic)

| Run name | Topics with response | Aggregate numbers of snippets | | | | |
|---|---|---|---|---|---|---|
| | | total | sup | sup imp | sup imp novel | sup imp novel not-rep |
| English monolingual task: 65 topics | | | | | | |
| lexiclone | 38 | 684 | 676 | 179 | 98 | 79 |
| rfia-bow-en | 65 | 607 | 607 | 255 | 187 | 173 |
| UA-DLSI-1 | 64 | 572 | 571 | 277 | 204 | 191 |
| UA-DLSI-2 | 60 | 489 | 488 | 239 | 173 | 161 |
| dltg061 | 65 | 435 | 435 | 226 | 165 | 161 |
| dltg062 | 65 | 682 | 682 | 310 | 223 | 194 |
| uams-linkret-en | 65 | 570 | 570 | 331 | 202 | 191 |
| uams-link-en | 65 | 615 | 614 | 353 | 232 | 220 |
| uams-ret-en | 65 | 580 | 580 | 325 | 203 | 193 |
| WLV-one-old | 61 | 473 | 473 | 263 | 219 | 142 |
| WLV-two | 61 | 526 | 526 | 327 | 280 | 135 |
| WLV-one | 61 | 473 | 472 | 267 | 221 | 135 |
| Spanish monolingual task: 67 topics | | | | | | |
| rfia-bow-es | 62 | 497 | 497 | 198 | 142 | 113 |
| UA-DLSI-es | 63 | 501 | 501 | 184 | 149 | 111 |
| mira-IS-CN-N | 67 | 251 | 251 | 127 | 79 | 69 |
| mira-IP-CN-CN | 67 | 431 | 431 | 155 | 95 | 71 |
| Dutch monolingual task: 60 topics | | | | | | |
| uams-linkret-nl | 60 | 425 | 425 | 301 | 228 | 210 |
| uams-link-nl | 60 | 455 | 455 | 305 | 236 | 228 |
| uams-ret-nl | 60 | 450 | 450 | 271 | 206 | 192 |
| English-Dutch bilingual task: 60 topics | | | | | | |
| uams-linkret-ennl | 60 | 564 | 551 | 456 | 342 | 302 |

the corresponding languages), as well as due to the differences in the sizes and structures of the Wikipedias in these languages. 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.

### 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.

The table below shows the agreement of pairs of assessors on importance judgments: the percent of matching assessments and Cohen's kappa ($\kappa$).

**Table 3.** Evaluation results for the submitted runs (calculated for top 10 snippets per topic); highest scores per task are given in boldface

| Run name | Number of topics with response | Total yield | Average yield | MRR | Precision |
|---|---|---|---|---|---|
| English monolingual task: 65 topics | | | | | |
| lexiclone | 38 | 58 | 1.53 | 0.31 | 0.21 |
| rfia-bow-en | 65 | 173 | 2.66 | 0.48 | 0.29 |
| UA-DLSI-1 | 64 | 191 | 2.98 | 0.53 | 0.33 |
| UA-DLSI-2 | 60 | 158 | 2.63 | 0.52 | 0.32 |
| dltg061 | 65 | 160 | 2.46 | 0.54 | **0.37** |
| dltg062 | 65 | 152 | 2.34 | 0.50 | 0.33 |
| uams-linkret-en | 65 | 188 | 2.89 | 0.52 | 0.33 |
| uams-link-en | 65 | **220** | **3.38** | 0.58 | 0.36 |
| uams-ret-en | 65 | 191 | 2.94 | 0.52 | 0.33 |
| WLV-one-old | 61 | 142 | 2.33 | 0.58 | 0.30 |
| WLV-two | 61 | 135 | 2.21 | **0.59** | 0.26 |
| WLV-one | 61 | 135 | 2.21 | 0.58 | 0.29 |
| Spanish monolingual task: 67 topics | | | | | |
| rfia-bow-es | 62 | **113** | **1.82** | **0.37** | 0.23 |
| UA-DLSI-es | 63 | 111 | 1.76 | 0.36 | 0.22 |
| mira-IS-CN-N | 67 | 69 | 1.03 | 0.30 | **0.27** |
| mira-IP-CN-CN | 67 | 71 | 1.06 | 0.29 | 0.16 |
| Dutch monolingual task: 60 topics | | | | | |
| uams-linkret-nl | 60 | 210 | 3.50 | **0.53** | 0.49 |
| uams-link-nl | 60 | **228** | **3.80** | **0.53** | **0.50** |
| uams-ret-nl | 60 | 192 | 3.20 | 0.45 | 0.42 |
| English-Dutch bilingual task: 60 topics | | | | | |
| uams-linkret-ennl | 60 | **302** | **5.03** | **0.52** | **0.54** |

| **Assessor pair** | **Common snippets** | **Agreement** | $\kappa$ |
|---|---|---|---|
| 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 |

We see that the $\kappa$ 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.

## 5 Conclusion

We have described the first installment of the WiQA—Question Answering Using Wikipedia—task. 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, the task turned out to be do-able, and in cases several participants managed to achieve impressive yield, MRR, and precision scores. Surprisingly, the highest scores were achieved on the bilingual task.

As to the future of WiQA, as pointed out before we aim to take a close look at our assesssments, perhaps add new assessments, and analyse inter-assessor agreement along various dimensions. 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. In the next edition of the task we would like to put more emphasis on the multilingual aspect of the task and extend the task by allowing systems to locate snippets in a fixed collection of crawled web pages, in addition to Wikipedia articles.

The collections, topics and the assessed runs of the participants of WiQA 2006 are available [WiQA(2006)].

## 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

[CLEF 2006 Working Notes(2006)] CLEF 2006 Working Notes. In: Working Notes for the CLEF 2006 Workshop (2006), `http://www.clef-campaign.org/2006/working_working_notes/CLEF2006WN-Contents.html`

[Denoyer and Gallinari(2006)] Denoyer, L., Gallinari, P.: The Wikipedia XML Corpus. SIGIR Forum (2006)

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

[WiQA(2006)] WiQA, Question Answering Using Wikipedia (2006), `http://ilps science.uva.nl/WiQA/`