A Pilot for Evaluating Exploratory Question Answering

Valentin Jijkoun
ISLA, Informatics Institute
University of Amsterdam
jijkoun@science.uva.nl

Maarten de Rijke
ISLA, Informatics Institute
University of Amsterdam
mdr@science.uva.nl

ABSTRACT

We describe a pilot on evaluating exploratory search in Wikipedia, the free online encyclopedia. The pilot will be held at CLEF 2006, and brings together both search and navigation, and reading and authoring.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation*

General Terms

Experimentation, Measurement

Keywords

Test collection formation, evaluation, question answering, Wikipedia

1. INTRODUCTION

Question Answering (QA) has attracted a great deal of attention, especially since the launch of the QA track at TREC in 1999. While significant progress has been made in technology for answering general factoids (e.g., *How fast does a cheetah run*? or *What is the German population*?), there is a real need to go beyond such factoids [5, 2]. At the TREC QA track this has been recognized through the introduction of definition questions and of so-called "other" questions that are far more exploratory in nature and ask for important information about a topic at hand that the user does not know enough about to ask.

In this paper we describe a pilot evaluation task that takes the "other" questions a step further. The task, called WiQA (Question Answering using Wikipedia [7]), will be organized as part of CLEF 2006. It involves answering undirected informational queries [3] against Wikipedia, the free online encyclopedia [6]. The purpose of the WiQA pilot is to develop novel question answering technologies, ones that go beyond the traditional highly focused factoid questions to include more open and exploratory ones, using the rich structure and reliable content of the Wikipedia.

Below, we first describe our take on different ways of accessing Wikipedia; then we provide details of the WiQA pilot, including a detailed example. After that we briefly describe the assessment criteria and evaluation metrics to be used at WiQA.

2. ACCESSING WIKIPEDIA

We believe that natural ways of accessing the information in Wikipedia mix two types of things:

- search and navigation, and
- reading and authoring.

Given this assumption, there are many natural possible tasks, or aspects of tasks, that are of interest in the WiQA pilot. To start, these include (the usual) highly focused questions. For instance, when using Wikipedia as the source to answer factoid questions (such as *How big is Berlin*? or *Find tennis players born in Berlin*), QA systems can use layout, formatting and wording regularities to pinpoint answers. In addition, they can use explicit semantic annotation: lists (such as *List of male tennis players*), categories (e.g., *Andre Agassi* is categorized into *Las Vegans*, *American tennis players*, etc.), structured tables (so-called *templates* providing standard information about countries, people, etc.).

As to more exploratory types of questions, there are many scenarios that seem very natural in the Wikipedia setting as well as many research questions that such scenarios give rise to. Below we list some of these scenarios and research questions:

Summarizing the content of Wikipedia articles. This corresponds to answering non-factoid questions such as *Tell me important facts about Andre Agassi.* Addressing such information needs raises important research questions. Is the current structure of Wikipedia pages good enough: aren't the "leads" perfect summaries of single pages? Is some level of (user-dependent) summarization needed for very long articles?

Summarizing the structure of Wikipedia. This may allow us to recover relevant information that is not explicitly given on a page, but is rather distributed across entire encyclopedia. E.g., what are important articles that mention Andre Agassi, and what do they say about him?

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...$5.00.

Alice Cooper's musical career begins (from the article *1969 in music*)
Relocating to London in the late '70s, they worked all over the United Kingdom and Europe to establish themselves, ... supporting the top hard-rock acts of the day including Alice Cooper (from the article *AC/DC*)
February 24: Alice Cooper announces that he is going to run for Governor of Arizona. (from the article *1988 in music*)
Throughout the 1960s and 1970s, Ann Arbor was home to many influential rock and roll bands, such as the MC5, Alice Cooper, Iggy Pop... (from the article *History of Ann Arbor, Michigan*)
Arthur Brown, a British rock and roll singer known for his flamboyant, theatrical style and significant influence on shock rockers such as Alice Cooper and Marilyn Manson (from the article *Arthur Brown (musician)*)

Table 1: Snippets to be added to the article on Alice Cooper

Automatic clustering of Wikipedia articles. Systems may use link structure, layout and semantic annotation to find similar or related pages. This is useful to identify potential missing list pages (such as *List of U.S. Open champions*). What techniques are appropriate here? Flat or hierarchical clustering, labelled or unlabelled clusters? Which of the many features of articles are useful for this task?

Handling navigational information needs. This group of tasks includes finding pages similar to a give one, as well as important or popular articles around a given topic. Addressing these needs may also involve generating (ranked) lists of different types of entities related to a given topic (the timeline of events, related locations and organizations), in context, as well as identifying and browsing multiple IS-A hierarchies provided by the catagory structure of Wikipedia.

Multi-lingual aspects. At present (June 2006) there are 10 languages with more than 100,000 articles, and 30 more have over 10,000 articles. Many pages are linked to their counterparts in other languages. Is it possible to automatically detect inconsistencies and missing article alignments? Can we compare existing cross-language alignments of Wikipedia pages, detect missing subgraphs for different languages? Is it possible to use machine translation to generate stubs for pages missing in one language?

As we will see below, the task defined for WiQA 2006 mixes some of these aspects.

3. WIQA 2006

The WiQA 2006 task that we envisage mixes search and navigation, and we are keen on exploring the reader-author inversion, building systems that help provide access to Wikipedia's content and that help author and edit its content. In the WiQA pilot, we will exploit the fact that, in Wikipedia, the distinction between author and reader has become blurred. Specifically, we aim to see how information retrieval and language technology can be effectively used to help readers and authors of articles get access to information spread throughout Wikipedia rather than stored locally on a single page.

Cryptonomicon is a 1999 novel by Neal Stephenson that concurrently follows the exploits of World War II-era cryptographers affiliated with Bletchley Park in their attempts to crack Axis codes... (from the article *Cryptonomicon*; assessed as novel, non-repeated and important)
A rare Abwehr Enigma machine, designated G312, was stolen from the Bletchley Park museum on 1 April 2000... (from the article *Enigma machine*; assessed as novel, non-repeated and important)
Together with the cryptographic efforts centered at Bletchley Park and also at Arlington Hall, the development of radar and computers in the UK and later in the USA, and the jet engine in the UK and Germany, the Manhattan Project represents one of the few massive, secret, and outstandingly successful technological efforts spawned by the conflict of World War II. (from the article *Manhattan Project*; assessed as not novel, non-repeated and important; the important information is actually contained in the original article "Bletchley Park": *The Bletchley Park effort was comparable in influence to other WWII-era technological efforts, such as... Manhattan Project...*)
Olivia's father, Brin Newton-John, originally from Wales, was an MI5 officer attached to the Enigma machine project at Bletchley Park... (from the article *Olivia Newton-John*; assessed as novel, non-repeated and not important)

Table 2: Snippets and their assessments for the Wikipedia article *Bletchley park*

3.1 Task description

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 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.

To illustrate these ideas, let us look at an example. Consider a user wishing to update the article for Alice Cooper. Table 1 lists snippets from other English articles that seem interesting and novel for the topic, thus, worth including in the page.

Participants of the WiQA 2006 pilot will be able to 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 multilingual (where the snippets to be returned can be in any of the languages of the Wikipedia corpus used at WiQA).

3.2 Corpus

The corpus to be used at WiQA 2006 consists of XML-ified dumps of Wikipedia in three language: Dutch, English, and Spanish. The dumps are based on the XML version of the Wikipedia collections [1] that include the annotation of the structure of the articles, links between articles, categories, cross-lingual links, etc. For the WiQA 2006 pilot the collections were enriched with annotations of sentences and classification of pages into named entity classes (person, location, organization).

3.3 Assessment of the systems' results

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 will be manually assessed using the following binary criteria, largely inspired by the TREC 2003 Novelty task [4]:

- *support*: the snippet does indeed come from the specified target Wikipedia article.
- *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
- *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.

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.

To illustrate these ideas, Table 2 provides an example of assessments of snippets found for the target page *Bletchley Park*.

3.4 Evaluation metrics

One of the purposes of the WiQA pilot task is to experiment with different measures for evaluating performance of systems. WiQA will use the following simple principal measure for accessing the performance of the systems:

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

We will also consider other simple measures:

- *success rate*: the number of topics with at least one supported, novel, important target snippet, and
- *overall precision*: the percentage of supported, novel, non-repetitive, important snippets among all submitted snippets.

These choices are considerably "simpler" than the evaluation set-up at today's TREC QA track, where a type of series-based scoring is used that involves requires to identify key information nuggets. For our pilot, we prefer the simpler measures listed above—both to keep the assessment load limited and because we believe they are more transparent.

4. CONCLUSIONS

In this paper we have motivated and described WiQA, a new pilot for evaluating exploratory question answering that will be launched at CLEF 2006. By the time of the EESS workshop we should be able to provide some initial results of the pilot.

5. 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, and 640.002.501.

6. REFERENCES

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

[2] V. Jijkoun and M. de Rijke. Retrieving answers from frequently asked questions pages on the web. In *Proceedings of the Fourteenth ACM conference on Information and knowledge management (CIKM 2005).* ACM Press, 2005.

[3] 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.

[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] R. Soricut and E. Brill. Automatic question answering: Beyond the factoid. In *Proceedings HLT/NAACL*, 2004.

[6] Wikipedia, 2006. Wikipedia. URL: `http://www.wikipedia.org`.

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