# Using Wikipedia at the TREC QA Track

**David Ahn Valentin Jijkoun Gilad Mishne**
**Karin Müller Maarten de Rijke Stefan Schlobach***

Informatics Institute, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
http://ilps.science.uva.nl/

**Abstract:** We describe our participation in the TREC 2004 Question Answering track. We provide a detailed account of the ideas underlying our approach to the QA task, especially to the so-called "other" questions. This year we made essential use of Wikipedia, the free online encyclopedia, both as a source of answers to factoid questions and as an importance model to help us identify material to be returned in response to "other" questions.

## 1 Introduction

In this paper we describe our participation in the TREC 2004 Question Answering track; our participation in the Web and Terabyte tracks is described elsewhere [8]. This year, we had several aims for the Question Answering track. One was to extend our QA system to handle this year's more complex question presentation, and to see how our existing modules cope with this new setting. Another was to make good use of the English edition of Wikipedia (http://en.wikipedia.org), the open domain encyclopedia, both as an additional stream for answering factoid questions, and as an *importance model* to help us answer "other" questions. Let's explain the latter point before we continue.

When the QA track at TREC was introduced, it focused on so-called "factoid" questions (typically having a short named entity as an answer) such as *How many people live in Tokyo?* or *When is the Tulip Festival in Michigan?*. As the track evolved, it was argued that this type of questions does not accurately model the needs of real users of QA technology. In addition to named entities as answers, users often search for definitions of concepts, or for summaries of important information about them. As a result, in 2003 TREC introduced *definition questions*—questions for which the answer is not a single named entity, but a list of *information nuggets* [14].

We were glad to see that at the TREC 2004 QA track this was taken a step further. The questions were now clustered in small groups, organized around the same topic. For example, the topic *Concorde* included questions such as *How many seats are in the cabin of a Concorde?* and *What airlines have Concordes in their fleets?*. Finally, for every topic, the track guidelines required participants to supply "additional important information found in the corpus about the target, that was not explicitly asked." This last requirement has been dubbed *"other"* questions. In our view, the task presented at the TREC 2004 QA track, and the introduction of the "other" questions makes a big step towards more realistic user scenarios. According to our own analysis of web query logs, users tend to ask much more "knowledge gathering" questions than factoid questions about specific facts.

This new type of "other" questions puts more emphasis on the *user* aspect in the QA process—an issue that has mostly been neglected in the QA community. The TREC criteria for what is a *good answer* to a given question has so far been rather vague, but QA systems dealt with this vagueness fairly effectively for factoid questions. With the "other" questions, where systems are required to re-

*Currently at the Division of Mathematics and Computer Science, Free University of Amsterdam.

turn only *important* information, there is an implicitly assumed user model that can discriminate between important and unimportant facts about a topic. For example, for the topic *Clinton*, his birthday might be considered important, while the day of the week when he left Mexico probably is not. In order to give reasonable responses to "other" questions, a QA system needs to model such preferences.

We present an approach for answering "other" questions using an explicit "importance" model. We describe a method for gathering important facts about an entity from a collection of documents and for ranking the facts with respect to their importance for the user. We show that our ranking improves over plain retrieval of facts from the corpus. The core idea of our method is to estimate the importance of facts found in the target collection by using external "reference" corpora, high-quality sources of information that model a user's ability to distinguish between important and unimportant facts. The proposed method is our first step towards user-oriented QA, and further refinements of the underlying techniques are needed. We identify additional areas where this method is or may be helpful, and discuss its strengths, weaknesses and directions for further research.

The rest of this paper is organized as follows. We give separate accounts of our approaches for answering factoid/list questions and for answering "other" questions. First, though, we address a complication in the presentation of questions in this year's QA task: the grouping of questions by *target*. We then describe our runs, present our results, and conclude.

# 2 Handling Targets

Each target is given explicitly as a phrase, and the questions for the target are presented in sequence. The possibility of anaphoric dependencies of the questions on the target or on preceding questions is thus introduced. We use an anaphoric resolution module to resolve pronouns occurring in the questions. Our module is simple: each pronoun is resolved to the highest ranked compatible antecedent in the antecedent list. The antecedent list consists of the target and all noun chunks occurring in preceding questions (with previously resolved pronouns replaced by their antecedents).

Our heuristic is to rank the target highest and to rank the other noun chunks according to their occurrence order. Compatibility is determined according to a simple type system: pronouns are marked **human** (*he*, *she*, etc.), **non-human** (*it*, *this*, *that*, etc.), or **unknown** (*they*, etc.), while other noun chunks are unmarked until they are resolved to a pronoun. These simple heuristics appear to work well with the question groups, where few entities are introduced and where there is a strong tendency to refer to the target.

# 3 Factoid Questions

Our approach for answering factoid questions is largely based on QUARTZ, our QA system used for experiments in the TREC 2003 QA track [7] and the CLEF 2004 Question Answering track [6]. We use an architecture where several streams run in parallel: each is based on a different approach to QA and is a self contained QA system in itself. A final step of merging the results of the streams is based on both redundancy of answers between streams and a process of learning the strengths and weaknesses of each of them [3].

This year, apart from minor technical modifications of these streams, we employed two new components in QUARTZ: an additional stream exploiting an open-domain encyclopedia and a mechanism for type checking of the answer candidates generated by each stream; see Figure 1. We also implemented a number of simple filtering mechanisms that serve as sanity checks for the answer candidates and performed a number of experiments in our answer justification module. We give an overview of the new components here and refer the reader to [3, 4, 6, 7] for an account of the rest of the system.

## 3.1 Encyclopedia Stream

Many systems participating in the TREC QA track use not only the local (AQUAINT) corpus, but also additional knowledge sources such as the web and various gazetteers [15]. The use of external resources (such as the web) in QUARTZ has proved to be beneficial, and we have therefore decided to employ an additional source of external knowledge into the system: a corpus specifically designed to address the information needs expressed by