

# Improving Accessibility in an Automated Question-Answering System

*Silvia Quarteroni*

University of Trento

[silvia.quarteroni@disi.unitn.it](mailto:silvia.quarteroni@disi.unitn.it)

## Abstract

We address the problem of accessibility in information retrieval by introducing a Question Answering system able to filter answers based on their reading difficulty. The reading level estimation technique is applicable to any domain and is potentially adjustable to any user category.

## Introduction

Using a computer to answer questions has been a human dream since the beginning of the digital era. A first step towards the achievement of such an ambitious goal is to deal with natural language to enable the computer to understand what its user asks and perform information retrieval.

Question Answering (QA) can be interpreted as a sub-discipline of information retrieval with the added challenge of applying sophisticated techniques to identify the complex syntactic and semantic relationships present in text in order to find concise answers.

However, a common problem in Question Answering and information retrieval is that in most systems results are created independently of the questioner's characteristics, goals and needs. This is a serious limitation: for instance, a primary school child and a History student may need different answers to the question: When did the Middle Ages begin?

So far, "personalized" QA has been advocated in the foremost evaluation campaign of the field, TREC-QA, starting from 2003<sup>1</sup>; however, the issue was solved rather expeditiously by designing a scenario where an "average news reader" (hence one particular user type) was imagined to submit definition questions [12].

In this document, we report on a study where a model of the user's reading abilities and personal interests is used to efficiently improve the quality of the information returned by a Question Answering system.

## A Web-based QA system

Our baseline system is YourQA [10], a QA system able to extract answers to both factoid questions (e.g. about names and dates) and non-factoid (e.g. about definitions) ones from the Web. The QA algorithm follows three phases:

---

<sup>1</sup> "Without any idea of who the questioner is and why he or she is asking the question it is essentially impossible for a system to decide what level of detail in a response is appropriate – presumably an elementary-school-aged child and a nuclear physicist should receive different answers for at least some questions". [12]

1. **Question Processing:** The query's expected answer type (e.g. person, definition) is estimated and the latter is submitted to the underlying search engine (Google, [www.google.com](http://www.google.com));
2. **Document Retrieval:** The top n documents are retrieved from the search engine and split into sentences;
3. **Answer Extraction:**
  - (a) A sentence-level similarity metric combining lexical, syntactic and semantic criteria is applied to the query and to each retrieved document sentence to identify candidate answer sentences;
  - (b) Candidate answers are ordered by relevance to the query; the list of top ranked answers is returned to the user in an HTML page.

The answers returned by YourQA are in the form of sentences with relevant words or phrases highlighted (as visible in Figure 1) and surrounded by their original passage to provide a context to the exact answer (this is especially useful for definitions).

**1. Title:** GradeSaver: ClassicNote: About Pride and Prejudice, **URL:** <http://www.gradesaver.com/classicnotes/titles/pride/about.html>, **Google Rank:** 6, **file:** about.html

About [Pride](#) and [Prejudice](#).

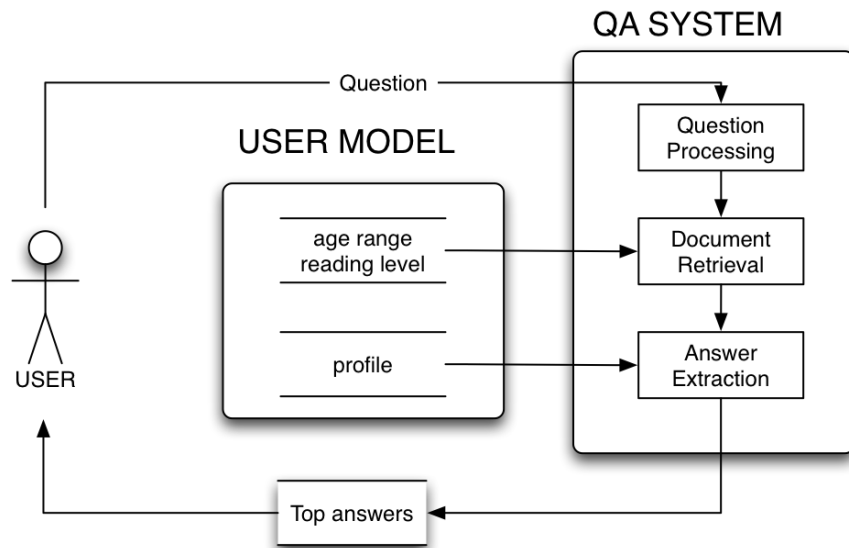
**Pride and Prejudice, published in 1813, is Jane's Austen's earliest work, and in some senses also one of her most mature works.**

Austen began writing the novel in 1796 at the age of twenty-one, under the title First Impressions.

**Fig. 1: Top answer by YourQA to: "When was Pride and Prejudice published?"**

## **A personalized QA system**

The salient feature of the personalized version of YourQA with respect to the standard version described above is the presence of a User Modelling component (as illustrated in Figure 2).



**Fig. 2: Personalized Question Answering Architecture**

As illustrated by the scheme, the interaction of the User Model with the core Question Answering module happens in two phases: first, the User Model provides criteria to filter out unsuitable documents for the user during the document retrieval phase.

Secondly, the User Model provides criteria to re-rank candidate answers based on profile relevance during answer extraction.

### User Model

As a target domain which would be generic enough to be a proof-of-concept of the usefulness of personalized Question Answering and at the same time a concrete, task-oriented application of User Modelling, we chose the education domain: hence, the User Model in YourQA represents students searching for information on the Web for their assignments.

Two basic aspects compose the user representation: on the one hand, the user's interests in terms of answer contents; on the other, the user's preferences in terms of answer presentation. These are modelled using three attributes:

- Age range,  $a \in \{7 - 10, 11 - 16, adult\}$ ; the first two ranges correspond to the primary and secondary school age in Britain, respectively;
- Reading level,  $r \in \{basic, medium, advanced\}$ ;
- Profile,  $p$ , a set of textual documents, bookmarks and Web pages of interest.

Analogous User Model components can be found in the SeAn [1] and SitelF [7] news recommender systems, where information such as age and browsing history, respectively are part of the User Model.

More generally, our approach is similar to that of personalized search systems such as [11], which constructs User Models based on the user's documents and Web pages of interest.

In this paper, we focus on readability as a tool to improve accessibility. For a detailed discussion of the personalization component of YourQA, see e.g. [9].

## Approaching Readability in Question Answering

Among the most widely used approaches to reading level estimation are models based on sentence length, such as "Flesch-Kincaid" [6], Fry [4] or SMOG [8]. The key idea behind these approaches is that the readability of text is inversely proportional to its length, hence readability is assessed using variations of sentence length-based metrics.

However it can be noticed that in Web documents, sentences are generally short and more concise than in printed documents, regardless of the complexity of the text. Hence the discriminative power of the above metrics can be affected by the fact that the difference in length between complex documents and simple ones is often not as wide as the printed text.

As opposed to the previous approaches, the language modelling approach which has been adopted in YourQA and is illustrated below accounts especially for lexical information. The technique has been proved in [3] to be at least as effective as the Flesch-Kincaid approach when modelling the reading level of subjects in primary and secondary school age.

We model reading level estimation as a multi-classification task which consists in assigning a document  $d$  to one of  $k$  different classes, each of which represents one reading level. In order to represent the three different age ranges defined in the corresponding attribute of the User Model, we define the three following classes:

1. *basic*, representing a document suitable for ages 7 – 11;
2. *medium*, representing a document suitable for ages 11 – 16;
3. *advanced*, representing a suitable for adults.

## Reading Level Estimation

We approach reading level estimation as a supervised learning task, where representative documents for each of the three classes are collected as labelled training instances and used to classify previously unseen documents according to their reading levels.

Our training instances consist of about 180 HTML documents, which originate from a collection of Web portals where pages are explicitly annotated by the publishers according to the 7–11, 11–16 and adult age: these contain 33,154, 33,407 and 35,024 words respectively.

Examples of such Web portals include BBC education ([bbc.co.uk/schools](http://bbc.co.uk/schools)), Magic Keys storybooks ([www.magickeys.com/books/](http://www.magickeys.com/books/)), and NASA for kids ([kids.msfc.nasa.gov](http://kids.msfc.nasa.gov)). The readability judgments of the Web portals are our gold standard for learning reading level classification. The fact that our training instances are labelled by an external trusted source contributes to the objectivity and soundness of our approach.

As a learning model, we use the Smoothed Unigram Model, which is a variation of a Multinomial Bayes classifier [3] based on the representation of the data known as unigram language modelling.

Given a set of documents, a unigram language model represents such set of as the vector of all the words appearing in the component documents associated with their corresponding probabilities of occurrence within the set.

In the test phase of the learning process, given an unclassified document  $D$ , a unigram language model is built to represent the single document  $D$  (as done for the training documents). The estimated reading level of  $D$  is the language model  $l_{mi}$  maximizing the likelihood  $L(l_{mi} | D)$  that  $D$  has been generated by  $l_{mi}$ . In our case, three language models  $l_{mi}$

are defined, where  $i \in \{basic, medium, advanced\}$  and the likelihood is estimated using the function:

$$L(lm_i | D) = \sum_{w \in D} C(w, D) \cdot \log[P(w | lm_i)]$$

where  $w$  is a word in the document,  $C(w, d)$  represents the number of occurrences of  $w$  in  $D$  and  $P(w | lm_i)$  is the probability that  $w$  occurs in  $lm_i$  (approximated by its frequency).

Related work. Within computational linguistics, several applications have been designed to address the needs of users with low reading skills. The computational approach to textual adaptation is commonly based on natural language generation: the process “translates” a difficult text into a syntactically and lexically simpler version.

In the case of PSET [2] for instance, a tagger, a morphological analyzer/generator and a parser are used to reformulate newspaper text for users affected by aphasia. Another example of research in this direction is Inui et al.'s lexical and syntactical paraphrasing system for deaf students [5], where the judgment of experts (teachers) is used to learn selection rules for paraphrases acquired using various methods. In the SKILLSUM project [13], used to generate literacy test reports, a set of choices regarding output (cue phrases, ordering and punctuation) are taken by a micro-planner based on a set of rules.

The approach presented in this work is conceptually different from these: exploiting the wealth of information available by using the Web as a source, the QA system can afford to choose among the documents available on a given subject those which best suit the given readability requirements.

## Reading Level Filtering

The first step carried out during personalized document retrieval is the estimation of the reading level of each document returned by Google in response to the query. Such estimation is conducted via language modelling following the technique exposed above. The documents having an incompatible reading level with the user are discarded so that only those having the same estimated reading level as the user are retained for further analysis.

As there can be queries for which the number of retrieved documents matching the requested reading level is less than the number of documents returned by the system (currently five), this condition is relaxed so that part of the documents having other reading levels are accepted in the set of candidate documents for answer extraction.

In particular, if the user's reading level is advanced, medium reading level documents are considered and, in case the threshold number of documents is not met, basic documents complete the set. If the requested reading level is medium, documents having a basic readability are used to complete the set; finally, if the requested reading level is basic, medium documents are accepted in the set. In all cases, due to the absence of other criteria at this stage of the QA algorithm, the choice of which documents to retain for a given reading level is determined by the search engine rank of the former (a higher rank determines preference).

The subsequent QA phase of answer extraction therefore begins with the documents left out of the reading level filtering phase.

## Evaluation

Our evaluation of reading level estimation was conducted according to two criteria: first, an objective assessment of the robustness of the unigram language models created to represent the User Model's reading level; second, an assessment of the agreement of users with the system's estimation.

### Robustness of the Unigram Language Models

The robustness of the unigram language models was computed by running 10-fold cross-validation on the set of documents used to create such models. First, we randomly split all of the documents used to create the language models into ten equally sized folds. Then, estimation accuracy was computed in two ways:

Approach A. Within each fold, the ratio of correctly classified documents with respect to the total number of documents was computed separately for each level. Then, the average between the three reading level estimation accuracies of each fold was used as accuracy of the fold. The final accuracy was thus the average accuracy of the different folds. The results of this experiment gave an average accuracy of 91.49 with a standard deviation of 6.54.

Approach B. The ratio of correctly classified documents with respect to the total number of documents was computed for each fold regardless of the reading level. Such ratio was used as accuracy for the fold and the average accuracy was computed for the ten folds as before. The results of this second experiment gave an average accuracy of 94.23% with a standard deviation of 1.98.

A high level of accuracy is important to ensure the consistency of reading level estimation. These results prove that unigram language models are good predictors of the basic, medium and advanced reading levels. However, this does not prove a direct effect on the user's perception of such levels. The following experiment takes charge of the user-centric aspect of reading level evaluation.

### User Agreement with Reading Level Estimation

The metric used to assess the users' agreement with the system's reading level estimation was called Reading level agreement ( $A_r$ ).

Given the set  $R$  of results returned by the system for a reading level  $r$ , it is the ratio between  $\text{suitable}(R)$ , i.e. the number of documents in  $R$  rated by the users as suitable for  $r$ , and the total number of documents in  $R$ :  $A_r = \text{suitable}(R) / |R|$ .  $A_r$  was computed for each level. The reading level agreement experiment was performed as follows.

**Participants.** The involved participants were 20 subjects aged between 16 and 52. All had a self-assessed good or medium English reading level, and came from various backgrounds (University students/graduates, professionals, high school).

**Materials.** The evaluation was performed by the 20 participants on the results returned by YourQA for 24 questions some of which are reported in Table 1. For each question, the results were returned in three different answer groups, corresponding to the basic, medium and advanced reading levels. As can be seen in Table 1, the answers include factoids ("Who painted the Sistine Chapel?"), lists ("Types of rhyme"), and definitions ("What is chickenpox?").

Query	Aadv	Amed	Abas
Who painted the Sistine Chapel?	0.85	0.72	0.79
Who was the first American in space?	0.94	0.80	0.72
Who was Achilles' best friend?	1.00	0.98	0.79
When did the Romans invade Britain?	0.87	0.74	0.82
Definition of metaphor	0.95	0.81	0.38
What is chickenpox?	1.00	0.97	0.68
Define German measles	1.00	0.87	0.80
Types of rhyme	1.00	1.00	0.79
Who was a famous cubist?	0.90	0.75	0.85
When did the Middle Ages begin?	0.91	0.82	0.68
Was there a Trojan war?	0.97	1.00	0.83
What is Shakespeare's most famous play?	0.90	0.97	0.83
Average	0.94	0.85	0.72

**Table 1. Examples of queries and reading level agreement**

**Procedure.** Each evaluator had to examine the results returned by YourQA to 8 of the 24 questions. For each question, he/she had to assess the three sets of answers corresponding to the reading levels, and specify for each answer passage whether he/she agreed that the given passage was assigned to the correct reading level.

Table 1 reports some sample questions along with their agreement scores. It shows that, altogether, evaluators found our results appropriate for the reading levels to which they were assigned. The accuracy tended to decrease (from 94% to 72%) with the level: this was predictable as it is more constraining to conform to a lower reading level than to a higher one.

## Conclusions and Perspectives

In this article, we address the problem of accessibility in information retrieval by introducing a Question Answering system able to filter answers based on their reading difficulty. The reading level estimation technique based on language modelling has the advantage of being applicable to documents in any domain.

We have demonstrated an application addressing the needs of students of the primary school, secondary school and adult age. However, the method we propose is suitable to model the reading level (and granularity) of any user category (expert/novice, child/adult, foreigner/mother tongue, etc.) provided that training documents are available.

## References

1. L. Ardissono, L. Console, and I. Torre. An adaptive system for the personalized access to news. *AI Communications*, 14(3):129–147, 2001.
2. J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. Simplifying text for language-impaired readers. In *Proceedings of EACL'99*, pages 269–270, 1999.

3. K. Collins-Thompson and J. P. Callan. A language modeling approach to predicting reading difficulty. In Proceedings of HLT/NAACL'04, 2004.
4. E. Fry. A readability formula that saves time. *Journal of Reading*, 11(7):265– 71, 1969.
5. K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura. Text simplification for reading assistance: a project note. In Proceedings of the ACL 2003 Workshop on Paraphrasing: Paraphrase Acquisition and Applications, pages 9–16, 2003.
6. J. Kincaid, R. Fishburne, R. Rodgers, and B. Chissom. Derivation of new readability formulas for navy enlisted personnel. Technical Report 8-75, 1975.
7. B. Magnini and C. Strapparava. Improving user modelling with content-based techniques. In *User Modelling: Proceedings of the 8th International Conference*, volume 2109 of LNCS. Springer, 2001.
8. G. McLaughlin. Smog grading: A new readability formula. *Journal of Reading*, 12(8):693–46, 1969.
9. S. Quarteroni and S. Manandhar. User modelling for personalized question answering. In Proceedings of AI\*IA'07, Rome, Italy, 2007.
10. S. Quarteroni. *Advanced Techniques for Personalized, Interactive Question Answering*. PhD thesis, 2007.
11. J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In Proceedings of SIGIR '05, 2005.
12. E. M. Voorhees. Overview of the TREC 2003 Question Answering Track. In Proceedings of TREC'03, 2003.
13. S. Williams and E. Reiter. Generating readable texts for readers with low basic skills. In Proceedings of ENLG-2005, pages 140–147, 2005.

#### About the author



*Silvia Quarteroni* is a Senior Research Fellow under a Marie Curie Excellence Grant (EXT) on the ADAMACH project at Department of Information and Telecommunication Technology, University of Trento, Italy. She joined the project after receiving her PhD in Computer Science from the University of York, UK in 2007, with the thesis titled "Advanced Techniques for Personalized, Interactive Question Answering."