Linking Educational Materials to Encyclopedic Knowledge

Andras CSOMAI and Rada MIHALCEA [1]
University of North Texas

Abstract. This paper describes a system that automatically links study materials to encyclopedic knowledge, and shows how the availability of such knowledge within easy reach of the learner can improve both the quality of the knowledge acquired and the time needed to obtain such knowledge.

1. Introduction

According to studies in cognitive science [1,6], an important aspect of the understanding and learning process is the ability to connect the learning material to the prior knowledge of the learner. Similarly, research in active reading and learning [3,4,5] has shown that active reading skills can improve the effectiveness and efficiency of information comprehension.

The amount of background knowledge necessary for a satisfactory understanding of an educational material depends on the *level of explicitness* of the text. However, it is almost impossible to create pedagogical materials that simultaneously serve the needs of both low- and high-knowledge users. Additionally, although the proactive construction of knowledge through active reading was found to increase the effectiveness of the learning process, such skills are not very advanced in many students [3]. It is therefore desirable to develop methods that can seamlessly integrate additional information in the learning material – such that low-knowledge readers can have immediate access to additional explanatory information that can facilitate a deeper understanding of the study material, and they can consequently develop active reading skills.

In this paper we present how our *Wikify!* system [2] can be used to improve educational materials by automatically selecting keywords, technical terms and other key concepts and linking them to an external knowledge source, in our case an online encyclopedia (Wikipedia). This can be thought of as an artificial extension of the reader's knowledge-base that can be accessed on a per-need basis, and is expected to be particularly useful for the increasingly popular online learning environments, where the lack of a direct communication between the instructor and the student can deprive the learner from the ability to quickly receive answers to questions related to background knowledge.

2. Wikipedia and Text Wikification

Wikipedia (`http://en.wikipedia.org`) is an online encyclopedia that has grown to become one of the largest online repositories of encyclopedic knowledge, with millions of

[1]Correspondence to: Rada Mihalcea, University of North Texas, P.O. Box 311366, Denton, TX. E-mail: rada@cs.unt.edu

articles available for a large number of languages. In fact, Wikipedia editions are available for more than 200 languages, with a number of entries varying from a few pages to more than one million articles per language.

Given a text or hypertext document, we define "text wikification" as the task of automatically extracting the most important words and phrases in the document (keywords), and identifying for each such keyword the appropriate link to a Wikipedia article with detailed explanatory information about the corresponding keyphrase.

Text wikification is performed in two steps. The first step consists of identifying those words and phrases that are considered important for the text at hand. These typically include technical terms, named entities, new terminology, as well as other concepts closely related to the content of the text. This task is similar to the well studied problem of *keyword extraction*, and correspondingly the *Wikify!* system uses state-of-the-art keyword extraction techniques coupled with novel features specific to online encyclopedias.

The second step consists of finding the correct Wikipedia article that should be linked to a candidate keyword. Here, we face the problem of link ambiguity, meaning that a phrase can be usually linked to more than one Wikipedia page, and the correct interpretation of the phrase (and correspondingly the correct link) depends on the context where it occurs. This task is in fact analogous to the problem of *word sense disambiguation* and our system again uses state-of-the-art techniques to address this problem.

The *Wikify!* system creates wikified documents hardly distinguishable from those created by human annotators [2], allowing us to automatically generate high quality hyperlinked texts.

3. Wikification: A History Test

The hypothesis guiding our experiment is that the *Wikify!* system can facilitate the access to information by students, and consequently it can improve the learning process. In order to evaluate this hypothesis, we devised a test that simulates the situation where a student requires additional related information to understand a certain study material.

We created a test by selecting 14 questions from a quiz from an online history course taught at the University of North Texas. The questions were selected so that they inquired about encyclopedic information rather than complex analyses or inferences about a problem. All the questions consisted of multiple choice test items, with a relatively high-degree of difficulty – meaning that in most cases common knowledge was not enough to provide an answer. Out of the 14 questions, half were wikified (including the answer choices), and half were left in their original format (raw text).

60 students were asked to participate in the experiment. Each participant was presented with the set of 14 questions, where randomly either the first seven or the last seven were automatically wikified with the *Wikify!* system. Overall, any single question was presented to an equal number of students in a wikified and non-wikified version. Given the size of the experiment, this setting allows us to compensate for the individual abilities of the participants as well as for the various degree of difficulty of the questions.

All the participants received the same set of instructions, which asked them to answer all the questions in one sitting. The instructions specifically indicated that the students were allowed to use *any* source of information they had access to,[2] including their own knowledge, search engines, or the Wikipedia links provided inside the questions. Students were not required to use the Wikipedia links available in the wikified questions, nor were they

[2]Note that the students taking the original quiz during the history course itself were also allowed to use any additional information they needed in order to answer the quiz questions.

prohibited to use the Wikipedia itself for answering the non-wikified questions. When an answer could not be found, the students were also given the possibility to skip the question.

After the test answers were submitted by all the participants, we post-filtered the submissions by removing all the tests left unfinished (13 tests fell under this category), as well as the tests where all the questions were skipped (3 tests). Both these conditions were interpreted as lack of seriousness, and therefore the corresponding tests were discarded. This left us with a final set of 44 complete tests, leading to a total of 44 answers for each of the 14 questions, accounting for 22 answers collected for the wikified version of the questions, and 22 for the non-wikified version.

We were primarily interested in finding the answer to two main questions: (1) Does the immediate access to relevant encyclopedic information improve the quality of the knowledge acquired for a specific topic? and (2) Does bringing such additional information within easy reach reduce the time required to acquire knowledge?

To answer the first question, we determined the number of questions correctly answered using the wikified version versus the non-wikified one. To answer the second question, we determined the time spent to answer the questions for each of the two versions. Table 1 shows the average calculated over the 22 answers collected for each version for each of the 14 questions.[3]

	Wikified	Non-wikified
CORRECTNESS (%)	78.89	75.97
TIME (SEC.)	62.70	70.33

Table 1. Evaluation results for wikified versus non-wikified test items.

In terms of correctly answered questions, the wikified version of the test items resulted in a relative error rate reduction of 12.2% with respect to the non-wikifed version ($p < 0.1$, paired t-test). Moreover, the time required to collect the required knowledge was also significantly reduced in the case of the wikified questions, with a relative time reduction of 25.5% ($p < 0.05$, paired t-test).

Both questions were answered positively: the immediate access to relevant encyclopedic information seems to improve both the quality of the knowledge acquired, and the time needed to obtain such knowledge. While we do not wish to discuss here the reliability of the information contained in Wikipedia, we believe that these results suggest that providing students with information relevant to the topic of study, and bringing such information within easy reach through hyperlinking, are both successful strategies for increased effectiveness in pedagogical tasks.

References

[1] Walter Kintsch. *Comprehension: A paradigm for cognition*. Cambridge Uniersity Press, New York, United States of America, 1998.

[2] Rada Mihalcea and Andras Csomai. Wikify! linking documents to encyclopedic knowledge. *(Under Review)*, 2007.

[3] T. Murray. Metalinks: Authoring and affordances for conceptual and narrative flow in adaptive hyperbooks. *International Journal of Artificial Intelligence in Education*, 13(1), 2002.

[4] C. Roast, I. Ritchie, and S. Thomas. Re-creating the reader: Supporting active reading in literary research. *Communications of the ACM*, 45(10):109–111, 2002.

[5] B.N. Schilit, M.N. Price, G. Golovshinski, K. Tanaks, and C.C. Marshall. The reading appliance revolution. *IEEE Computer*, 32(1):65–73, 1999.

[6] James F. Voss and Laurie Ney Silfies. Learning from history texts: The interaction of knowledge and comprehension skill with text structure. *Cognition and Instruction*, 14(1):45–68, 1996.

[3] Micro- and macro-average are identical, since the same number of answers is collected for each test item.