Enhancing Access To Classic Children's Literature

Claudia Hauff
HMI
University of Twente
Enschede, the Netherlands
c.hauff@ewi.utwente.nl

Dolf Trieschnigg
Database Group
University of Twente
Enschede, the Netherlands
trieschn@ewi.utwente.nl

ABSTRACT

Project Gutenberg is a digital library that contains mostly public domain books, including a large number of works that belong to children's literature. Many of these classic books are offered in a text-only format, which does not make them appealing for children to read. Moreover, stories that were written for children one hundred or more years ago, might not be readily understandable by children today due to diverging vocabularies and experiences. In this poster, we describe ongoing work to enhance the access to this children's literature repository. Firstly, we attempt to automatically illustrate the children's literature. Secondly, we link the text to background information to increase understanding and ease of reading. The overall motivation of this work is to make such publicly available books more easily accessible to children by making them more entertaining and engaging.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval
General Terms: Experimentation
Keywords: story illustration, wikipedia, wiktionary

1. INTRODUCTION

Project Gutenberg[1] is a digital library that contains mostly public domain books, including a large number of works that belong to children's literature. At the time of writing more than 2200 books in the *Juvenile belles lettres*[2] category can be accessed through the project's website. In many cases, these books are offered in a text-only format. One example is the book *Grimms' Fairy Tales*, originally published in the year 1812. The volume available through Project Gutenberg contains 62 separate fairy tales, amounting to 250kb of plain text. Not a single image appears in the digitization, which makes the book, although it is a children's classic, not readily accessible to young readers.

We believe that we can make these works more appealing to children by illustrating them. Naturally, the most accurate method of doing this would be to manually add illustrations to each story. However, such a task would be very time consuming and cost expensive. In this work, we propose an automatic approach to add illustrations to the books instead [4, 5].

A second very important issue arises from the fact that these stories were written one hundred or more years ago[3] in a style that may appear unusual to us today. Furthermore, the authors may have referred to events or occurrences that were common knowledge at the time, while today those references may confuse a reader. Take for instance this excerpt from *The Three Musketeers* by Alexandre Dumas (written in 1844):

> There were nobles, who made war against each other; there was the king, who made war against the cardinal; there was Spain, which made war against the king. Then, in addition to these concealed or public, secret or open wars, there were robbers, mendicants, **Huguenots**, wolves, and scoundrels, who made war upon everybody. The citizens always took up arms readily against thieves, wolves or scoundrels, often against nobles or **Huguenots**, sometimes against the king, but never against cardinal or Spain. It resulted, then, from this habit that on the said first Monday of April, 1625, the citizens, on hearing the clamor, and seeing neither the red-and-yellow standard nor the livery of the **Duc de Richelieu**, rushed toward the hostel of the Jolly Miller. When arrived there, the cause of the hubbub was apparent to all.

The underlined terms are those that we believe are very likely not to be known by children today. The named entities marked in bold on the other hand might be interesting to know more about and helpful for a better understanding of the story. If no aid is given here to make it easier for children to understand these phrases and named entities, they might lose interest in the story very quickly.

Based on this motivation, we propose to investigate mechanisms for the automatic recognition of unusual or outdated phrases, the recognition of "interesting" entities and finally the automatic enhancement of the digitized works with illustrations. So far, we have only taken the first steps in these directions. The initial version of our prototype is described in turn.

[1] `http://www.gutenberg.org/`
[2] *Juvenile belles lettres* is the children's literature category according to the Library of Congress classification.

[3] In the US, depending on the publication date, the copyright of a book usually expires between seventy and ninety-five years after the death of the book's author.

2. PROTOTYPE

The prototype consists of three main components (Figure 1), namely, (i) page illustration, (ii) identification of hard terminology, and, (iii) linking background information.

2.1 Page illustration

The images used for illustration are the freely available works from the OpenClipart[4] library. Out of the $\approx$ 25.000 cliparts that are available for download, we were able to extract meaningful meta-data for $\approx$ 22.000 cliparts. The extracted meta-data includes the number of times a clipart was downloaded, often a short description and a number of tags. For example, a cartoon drawing of a sheep[5] has the description *a cartoon sheep* and the tags

> *media, clip_art, unchecked, public_domain, image, svg, sheep, animal, mammal, colour, cartoon*

The tags thus range from specific (*sheep*) to very general (*animal*). We sorted the terms and phrases extracted from the clipart title, the description and the tags according to their specificity. In the example,*sheep* is the most specific term, followed by *mammal*. To illustrate a page of the book, first, the text is part-of-speech tagged and the nouns are extracted as potential objects of illustration. The most discriminative and concrete nouns (in contrast to nouns describing abstract states such as *warmth*) are considered and a clipart is selected for illustration based on its matching meta-data.

When illustrating pages, at this point in time we do not consider information that can be vital for selecting an accurate image, such as the emotions conveyed in the text, color information provided in the descriptions of a scene or negative polarity among others. Illustrating a paragraph with a crying baby, when the paragraph contains a sentence such as "The baby was happy." is clearly wrong. Likewise, a sentence like "He had a brown hat." at the moment can trigger the illustration of any kind of image that is tagged with *hat* without consideration of color. In future work we aim to address these challenges by considering low-level image features as well natural language processing techniques such as sentiment analysis.

Another potential research direction is to replace the static illustration of a book's pages (which is independent of the reader) by a user-dependent process. For example, if a child prefers a particular color, the illustration selection might be biased towards cliparts containing that color. Additionally, different sets of images could be selected for illustration in each viewing of the book.

2.2 Identification of hard terminology

Children may have difficulty with the terminology they encounter in classic English literature. Their limited vocabulary [2] in combination with nowadays uncommon wordings used in older texts may hamper fluent reading.

To aid the children's ease of understanding we automatically detect these problematic terms and link them to definitions found in Wiktionary[6]. Where possible, we link these terms to the definitions found in the *Wiktionary in Simple English*[7], which is more easy to understand.

Our current prototype is limited to scanning and linking Wiktionary entries without determining the correct sense. Difficult terminology is simply defined as detected entries not found in the Wiktionary category "basic words". The task itself however poses several interesting research challenges. Firstly, there is the challenge of identifying difficult terminology for different age groups, or people with different reading levels. These terms could be identified based on approaches for text simplification [2]. Secondly, disambiguation: determining the sense of the found term and linking the term to the correct definition. Traditional approaches to disambiguation might be used for this purpose [6, 10]. Thirdly, determining whether the found definition is useful for children. Several approaches to predict reading difficulty [1, 3, 9] can be explored for this purpose.

A quantitative evaluation should indicate the accuracy of the identified terms and accuracy of the disambiguated definition. A qualitative evaluation in the form of a user study should indicate whether children appreciate such a form of assisted reading.

2.3 Linking to background information

As a second reading aid, we propose to link the text to publicly available information found in Wikipedia[8] and where possible Simple Wikipedia[9].

We expect Wikipedia to contain useful background information about the entities encountered in the text. Automatically linking text to Wikipedia has been studied by for instance Milne and Witten [8] and Mihalcea and Csomai [7]. A preliminary evaluation with the online version of a Wikification system[10] showed that it performed well in identifying named entities present in Wikipedia. However, the selection process of Wikipedia pages to link to might be inappropriate for this task: for children, only the more interesting terms should be highlighted.

Similar to the identification of hard terminology, our current prototype is limited to scanning and linking to Wikipedia page titles. Again, the task of linking to background information in this setting poses interesting research challenges. A first challenge is of what exactly makes an entity interesting to children. A second challenge is how to filter out information in Wikipedia that is inappropriate for children.

3. CONCLUSION

In this poster, we have described our motivation for enhancing access to publicly available classic children's literature and gave a first outline of our prototype. For each of the three main components, we also outlined the challenges that we expect to face in future work.

References

[1] K. Collins-Thompson and J. Callan. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462, 2005.

[4] http://www.openclipart.org/
[5] http://www.openclipart.org/detail/29313
[6] http://wiktionary.org/
[7] http://simple.wiktionary.org/
[8] http://www.wikipedia.org/
[9] http://simple.wikipedia.org/
[10] http://www.nzdl.org/wikification/

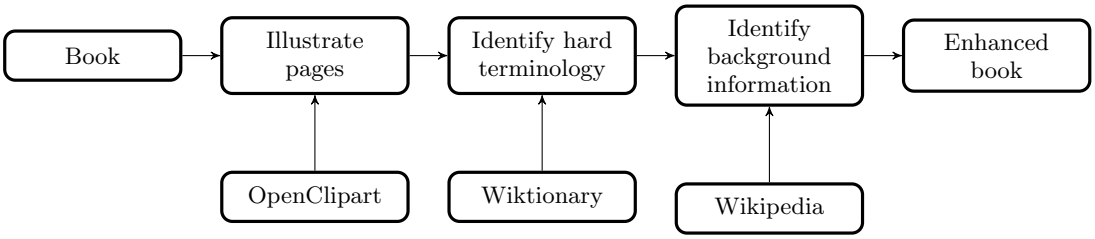

Figure 1: Prototype pipeline.

[2] J. De Belder and M.-F. Moens. Text simplification for children. In *Proceedings of SIGIR 2010 Towards Accessible Search Systems Workshop*, pages 19–26, Geneva, Switzerland, 2010.

[3] L. Feng, N. Elhadad and M. Huenerfauth. Cognitively motivated features for readability assessment. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–237, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[4] Y. Feng and M. Lapata. Topic Models for Image Annotation and Text Illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 831–839, 2010.

[5] D. Joshi, J.Z. Wang and J. Li. The Story Picturing Engine - a system for automatic text illustration. *TOMCCAP*, 1(2):68–89, 2006.

[6] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, USA., 1999.

[7] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA, 2007. ACM.

[8] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518, New York, NY, USA, 2008. ACM.

[9] S. E. Petersen and M. Ostendorf. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23 (1):89–106, 2009. ISSN 0885-2308.

[10] D. Yarowsky. One sense per collocation. In *HLT '93: Proceedings of the workshop on Human Language Technology*, pages 266–271, Morristown, NJ, USA, 1993. Association for Computational Linguistics.