Converting Wikipedia articles to LaTeX

Dirk Hünniger

Abstract

It is often desirable to have access to Wikipedia articles in LaTeX format. A translation by hand is typically time-consuming and error-prone. Thus it is natural to look for algorithmic solutions to this problem. Our solution is currently available free of charge under an open source license for Windows and Debian GNU/Linux. It is not limited to Wikipedia but supports all servers running the same wiki software (MediaWiki) as Wikipedia. In particular, it is also possible to process local wikis available only on private networks inside institutions.

1 Introduction

A wiki provides a very convenient way of working on a document with many contributors, without needing to learn the details of specialized version control and typesetting software. MediaWiki provides a function to export PDF files. But the possibilities for incorporating individual requirements on the output layout are very limited and usually insufficient for professional publishers. Also the typographic quality of the output is far less elaborate than what is provided by LaTeX. Furthermore, the embedding of formulas as raster graphics is often criticized.

2 User experience

In the default mode, our program takes a url to a web page on a MediaWiki server and writes a PDF version of that page generated with LaTeX to local hard disk. It is also possible to retrieve the corresponding LaTeX source code, including images.

Also in the default mode, the HTML generated by the MediaWiki server is evaluated. There is also an extended mode where the source code of the wiki page written in the wiki markup language is processed. The wiki markup language provides a mechanism similar to the LaTeX `\newcommand` directive, called “templates”. In this mode it is possible to map templates to LaTeX commands and implement them using `\newcommand` or similar methods in the headers. This mechanism provides a fine-grained control over the conversion process and thus gives the user the full flexibility of LaTeX.

3 On the history of the problem

Quite a few attempts have been made to tackle this problem programmatically. We would like to emphasize the successful work of Hans Georg Kluge, who modified MediaWiki’s original parser to produce LaTeX (`http://code.google.com/p/wiki2latex`). Unfortunately it needs to be installed on the server running the wiki in order to run and Wikipedia is currently not attempting to install it. This is partly because the security of the code is currently being discussed, which is particularly a concern since it is written entirely in PHP.

There have also been several attempts approaching the problem with regular expression or Backus-Naur forms. Recently we were able to provide a simple proof, based on the pumping lemma, that improper bracketing of HTML tags, as often found on Wikipedia, causes the grammar to no longer be context free, thus rendering it indescribable by Backus-Naur forms and regular expressions. This, in turn, rules out most standard parsing technology.

In our approach, we run all software on the user machine, thus bypassing any security concerns of Wikipedia. We opted for monadic parser combinators as parsing technology, and were able to handle the non-context-free grammar well with that approach.

4 Technical details of the implementation

The program is entirely written in the purely functional language Haskell. To do the necessary image processing the ImageMagick library is used. We currently use XƎLaTeX as the default compiler, although we recognized that the source (with tiny changes limited to the headers only) does also compile with pdfLaTeX and LuaLaTeX.

Currently there still is no freely available font that covers the whole range of Unicode. A problem in this respect is also that certain code points used for some Asian characters are used for more than one symbol and Wikipedia does not always provide a means to find out which symbol is actually meant by a Unicode character. For now we use FreeSerif as the default font, which omits Asian glyphs entirely. So we also offer a computationally combined font, made of several fonts available under the same open source license that actually covers the full Unicode range. In pdfLaTeX we use just this one font with the CJK package and thus can handle the first 16 bits of the Unicode range. This approach allows the user to still use custom fonts like Utopia, Courier, etc. For XƎLaTeX we provide a set of fonts for bold, italic, typewriter, small caps, and combinations thereof. This approach basically works also with LuaLaTeX, but unfortunately caused huge memory and CPU usage in our tests.

◇ Dirk Hünniger
`http://de.wikibooks.org/wiki/Benutzer:`
`Dirk_Huenniger/wb2pdf`
`dirk dot hunniger (at) googlemail dot com`