GraWiTas: a Grammar-based Wikipedia Talk Page Parser

Benjamin Cabrera
University of Duisburg-Essen
Lotharstr. 65
47057 Duisburg, Germany

Laura Steinert
University of Duisburg-Essen
Lotharstr. 65
47057 Duisburg, Germany

Björn Ross
University of Duisburg-Essen
Forsthausweg 2
47057 Duisburg, Germany

firstname.lastname@uni-due.de

Abstract

Wikipedia offers researchers unique insights into the collaboration and communication patterns of a large self-regulating community of editors. The main medium of direct communication between editors of an article is the article's talk page. However, a talk page file is unstructured and therefore difficult to analyse automatically. A few parsers exist that enable its transformation into a structured data format. However, they are rarely open source, support only a limited subset of the talk page syntax – resulting in the loss of content – and usually support only one export format. Together with this article we offer a very fast, lightweight, open source parser with support for various output formats. In a preliminary evaluation it achieved a high accuracy. The parser uses a *grammar*-based approach – offering a transparent implementation and easy extensibility.

1 Introduction

Wikipedia is becoming an increasingly important knowledge platform. As the content is created by a self-regulating community of users, the analysis of interactions among users with methods from natural language processing and social network analysis can yield important insights into collaboration patterns inherent to such platforms. For example, Viegas et al. (2007) manually classified talk page posts with regard to the communication type to analyse the coordination among editors. Such insights can be important for research in the area of Computer-Supported Collaborative Learning.

The collaboration patterns among Wikipedia editors are visible on an article's *revision history* and its *talk page*. The revision history lists all changes ever made to an article, but it does not contain explicit information about the collaboration between editors. Most discussions between editors take place on the article's talk page, a dedicated file where they can leave comments and discuss potential improvements and revisions.

The talk page is therefore useful for observing explicit coordination between editors. However, most studies on Wikipedia article quality have focused on easily accessible data (Liu and Ram, 2011), whereas talk pages are not easy to use with automated methods. Essentially, talk pages are very loosely structured text files for which the community has defined certain formatting rules. Editors do not always follow these rules in detail and thus an automated analysis of talk page data requires a parsing of the file into a structured format by breaking it down into the individual comments and the links among them.

In this article, we introduce an open source parser for article talk pages called *GraWiTas* that focuses on a good comment detection rate, good usability and a plethora of different output formats. While a few of such talk page parsers exist, our core parser is based on the Boost.Spirit C++-library[1] which utilises Parsing Expression Grammars (PEG) for specifying the language to parse. This leads to a very fast, easily extensible program for different use cases.

The next section of this paper describes the structure of Wikipedia talk pages. The third section describes the parser we developed. Afterwards, a preliminary evaluation is described. The fifth section gives an overview on related work. Finally, conclusions from our research are given at the end of the paper.

[1] http://boost-spirit.com/, as seen on Feb. 14, 2017

Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, April 3-7 2017, pages 21–24

2 Wikipedia Talk Pages

The talk page of a Wikipedia article is the place where editors can discuss issues with the article content and plan future contributions. There is a talk page for every article on Wikipedia. As any other page in Wikipedia, it can be edited by anybody by manipulating the underlying text file – which uses *Wiki markup*, a dedicated markup syntax. When a user visits the talk page, this syntax file is interpreted by the Wikipedia template engine and turned into the displayed HTML.

Although Wiki markup includes many different formatting (meta-)commands, the commands used on talk pages are fairly basic. Some guidelines as to how to comment on talk pages are defined in the *Wikipedia Talk Page Guidelines*[2]. These rules specify, for example, that new topics should be added to the bottom of the page, that a user may not alter or delete another user's post, that a user should sign a post with their username and a timestamp, and that to indicate to which older post they are referring, users should indent their own post.

The following snippet gives an example of the talk page syntax:

```
== Meaning of second paragraph ==
I don't understand the second paragraph
    ... [[User:U1|U1]] [[User Talk:U1|
    talk]] 07:24, 2 Dec 2016 (UTC)
:I also find it confusing...[[User:U2|U2
    ]] 17:03, 3 Dec 2016 (UTC)
::Me too... [[User:U3|U3]] [[User Talk:
    U3|talk]] 19:58, 3 Dec 2016 (UTC)
:LOL, the unit makes no sense [[User:U4|
    U4]] [[User Talk:U1|talk]] 00:27, 6
    Dec 2016 (UTC)
Is the reference to source 2 correct? [[
    User:U3|U3]] [[User Talk:U3|talk]]
    11:41, 4 Dec 2016 (UTC)
```

The first line marks the beginning of a new discussion topic on the page. The following lines contain comments on this topic. All authors signed their comments with a signature consisting of some variation of a link to their user profile page – in Wiki markup, hyperlinks are wrapped in '[[' and ']]' – and a timestamp. User U2 replies to the first post and thus indents their text by one tab. In Wiki markup this is done with a colon ':'. The third comment, which is a reply to the second one, is indented by two colons, leading to more indentation on the page when it is displayed.

While this structure is easily understood by humans, parsing talk page discussions automatically is far from trivial for a number of reasons. The discussion is not stored in a structured database format, and many editors use slight variations of the agreed-upon syntax when writing their comments. For example, there are multiple types of signatures that mark the end of a comment. Some editors do not indent correctly or use other formatting commands in the Wikipedia arsenal.

In addition, some Wikipedia talk pages contain *transcluded* content. This means that the content of a different file is pasted into the file at the specified position. Also, pages that become "too large" are *archived*[3], meaning that the original page is moved to a different name and a new, empty page is created in its place.

3 GraWiTas

Our parser – GraWiTas – consists of three components, covering the whole process of retrieving and parsing Wikipedia talk pages. The first two components gather and preprocess the needed data. They differ in the used data source: The *crawler component* extracts the talk page content of given Wikipedia URLs while the *dump component* processes full Wikipedia XML dumps. The actual parsing is done in the *core parser component*, where all comments from the Wiki markup of a talk page are extracted and exported into one of several output formats.

3.1 Core Parser Component

The main logic of the core parser lies in a system of rules that essentially defines what to consider as a comment on a talk page. Simply spoken, a comment is a piece of text – possibly with an indentation – followed by a signature and some line ending. As already discussed, signatures and indentations are relatively fuzzy concepts, which means that the rules need to define a lot of cases and exceptions. There are also some template elements in the Wikipedia syntax that can change the interpretation of a comment, e.g. outdents[4].

Grammars are a theoretical concept that matches such a rule system very well. All different cases can be specified in detail and it is easy to build more complex rules out of multiple

[2] https://en.wikipedia.org/wiki/Wikipedia:Talk_page_guidelines, as seen on Feb. 14, 2017

[3] https://en.wikipedia.org/wiki/Help:Archiving_a_talk_page, as seen on Feb. 14, 2017

[4] https://en.wikipedia.org/wiki/Template:Outdent, as seen on Feb. 14, 2017

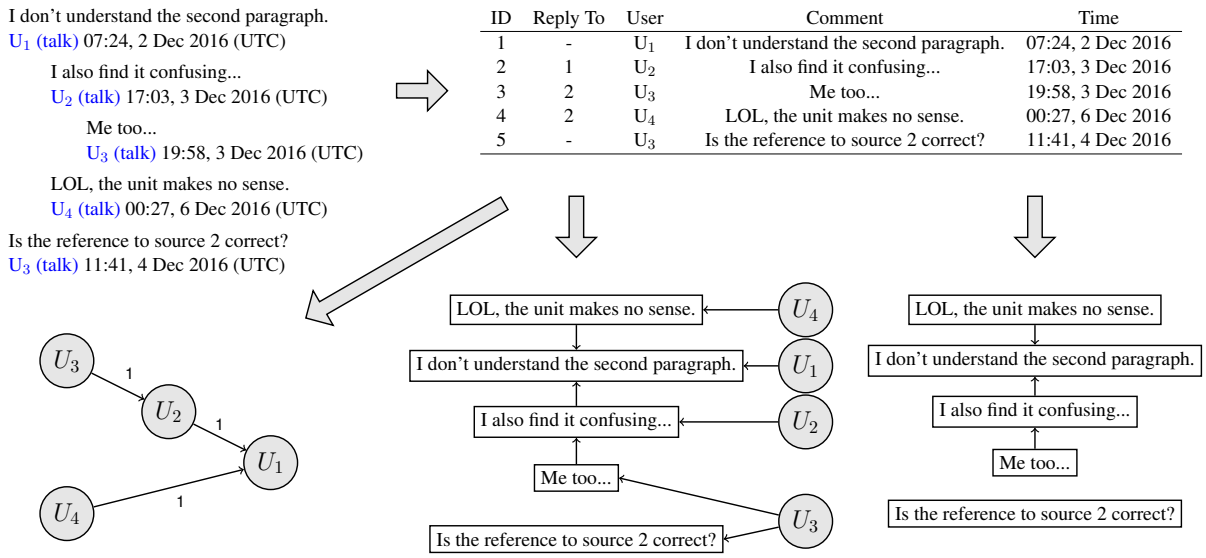

Figure 1: Schematic representation of the core parser component: A table (top right) is extracted from a talk page (top left) using a grammar-based rule set. It can then be transformed into various output formats, e.g. a one-mode (user- / comment-) graph (bottom left and right) or a two-mode graph (bottom center)

smaller ones. In particular we use Parsing Expression Grammars (Ford, 2004) which are very efficiently implemented in the Boost.Spirit library for C++. The library allows developers to essentially write the grammar in C++ syntax, which is then compiled – note the possibility for compile-time optimisation – to a highly optimised parser. Beside its efficient implementation, another advantage of using grammars is extensibility. Whenever we encountered mismatched or incorrectly parsed comments we added another case to our grammar and no real application logic had to be written.

After extracting the comments (with username, date, ...) from the raw talk page the core parser has various output formats including lists of the comments (formats: CSV, JSON, human-readable) and users as well as comment networks (formats: GML, GraphML, Graphviz) (cf. Figure 1).

3.2 Wiki Markup Crawler Component

The crawler component is responsible for obtaining the Wiki markup of a talk page using the Wikipedia website as a source. The program expects a text file containing the URLs to the Wikipedia articles as input. It then connects to the server to retrieve the HTML source code, from which we extract the relevant Wiki markup and which we finally feed to the core parser to get the structured talk page data for each article in the list of URLs.

Our crawler is also able to fetch archived pages and includes their content in the downloaded markup file. Finally, it also fetches transcluded content by searching for the respective Wiki-markup commands. If the transcluded content belongs to the namespace *Talk* – i.e. the name of the transcluded content starts with *Talk:* – it is part of a talk page and is therefore included. All other transcluded content is not included as it is unnecessary for the talk page analysis.

The crawler component should be used for small to medium-scale studies that rely on up-to-date talk page data. Setting up the crawler is very intuitive, but the need to contact the server may make it unfeasible to work with very large Wikipedia data.

3.3 Wiki Markup Dump Component

To be able to work with all data at once – without connecting to their server too often – Wikipedia offers a download of full database dumps that include e.g. all articles, all talk pages and all user pages.[5] For GraWiTas we also provide a program that is able to read such large XML dumps efficiently and feed talk pages to our core parser. Users can select if they want to parse all talk pages

[5] `https://en.wikipedia.org/wiki/Wikipedia:Database_download`, as seen on Feb. 14, 2017

in the dump or only some particular ones characterised by a list of article titles.

The dump component can be used for large-scale studies that look at all talk pages of Wikipedia at once. Compared to the crawler, obtaining the Wiki markdown is faster. However, this comes at the price that users have to download and maintain the large XML dump files.

4 Evaluation

We ran a very small evaluation study to assess the speed and accuracy of our core parser. For analysing the speed, we generated large artificial talk pages from smaller existing ones and measured the time our parser took. Using an Intel(R) Core(TM) i7 M 620 @ 2.67GHz processor, we obtained parsing times under 50ms for file sizes smaller than 1MB (∼2000 Comments). For files up to 10MB (∼20000 Comments) it took around 300ms. This leads to an overall parsing speed of around 30MB/s for average-sized comments. However, real world talk pages very rarely even pass the 1MB mark.

To evaluate accuracy, we picked a random article with a talk page and verified the extracted comments manually. Hereby, an accuracy of 0.97 was achieved. Although such a small evaluation is of course far from comprehensive, it shows that our parser is on the one hand fast enough to parse large numbers of talk pages and on the other hand yields a high enough accuracy for real-world applications.

5 Related Work

There have been previous attempts at parsing Wikipedia talk pages. Massa (2011) focused on user talk pages, where the conventions are different from article talk pages. Ferschke et al. (2012) rely on indentation for extracting relationships between comments, but use the revision history to identify the authors and comment creation dates. However, they do not offer any information on whether archived, transcluded and outdented content is handled correctly, and their implementation has not been made public. Laniado et al. (2011) infer the structure of discussions from indentation similar to our approach. They use user signatures to infer comment metadata (author and date). Their parser is available on Github[6]. However, it does not handle transcluded talk page content, nor the outdent template. Their parser outputs a CSV file with custom fields.

6 Conclusion

The possibility of parsing Wikipedia talk pages into networks provides many new avenues for NLP researchers interested in collaboration patterns. A usable parser is a prerequisite for this type of research. Unlike previous implementations, our parser correctly handles many quirks of Wikipedia talk page syntax, from archived and transcluded talk page contents to outdented comments. It can produce output in a number of standardised formats including GML as well as a custom text file format. Finally, it requires very little initial setup. The GraWiTas source code is publicly available[7], together with a web app demonstrating the core parser component.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media"

References

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the article: Recognizing dialog acts in wikipedia talk pages. In *Proc. of EACL'12*, pages 777–786, Stroudsburg, PA, USA. ACL.

Bryan Ford. 2004. Parsing expression grammars: A recognition-based syntactic foundation. *SIGPLAN Not.*, 39(1):111–122.

David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. 2011. When the wikipedians talk: Network and tree structure of wikipedia discussion pages. In *Proc. of ICWSM, Barcelona, Spain, July 17-21, 2011*.

Jun Liu and Sudha Ram. 2011. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Trans. Manag. Inform. Syst.*, 2(2).

Paolo Massa. 2011. Social networks of wikipedia. In *HT'11, Proc. of ACM Hypertext 2011, Eindhoven, The Netherlands, June 6-9, 2011*, pages 221–230.

Fernanda B. Viegas, Martin Wattenberg, Jesse Kriss, and Frank Van Ham. 2007. Talk before you type: Coordination in wikipedia. In *Proc. of HICSS'07*. IEEE.

[6] `https://github.com/sdivad/WikiTalkParser`, as seen on Feb. 14, 2017

[7] `https://github.com/ace7k3/grawitas`