Corpus2Wiki: A MediaWiki-based Tool for Automatically Generating Wikiditions in Digital Humanities

Alex Hunziker[1] Hasanagha Mammadov[2] Wahed Hemati[3] Alexander Mehler[4]

Abstract: We describe current developments of Corpus2Wiki. Corpus2Wiki is a tool for generating so-called Wikiditions out of text corpora. It provides text analyses, annotations and their visualizations without requiring programming or advanced computer skills. By using TextImager as a back-end, Corpus2Wiki can automatically analyze input documents at different linguistic levels. Currently, it automatically annotates information regarding lemmatization, parts of speech, morphological information, named entities, geolocations and topic labels based on the *Dewey Decimal Classification* (DDC). Any results are stored and displayed by means of a modified and extended MediaWiki which makes it easy to further process texts and their annotations. The aim of this paper is to present the capabilities of Corpus2wiki, to point out the improvements made and to make suggestions for further development.

Keywords: NLP; Annotation; Wikidition; Text Visualization; TextImager; TextAnnotator

1 Introduction

According to Burdick et al. [Bu12], four activities of digital humanities should be technologically supported, namely curation, analysis, editing and modelling. In order to create a valuable and useful tool for this domain, these activities are taken into account in the development of Wikidition [Me16]. In order to further automatize the process of generating Wikiditions, we developed Corpus2Wiki [RHM18]. Corpus2Wiki provides analyses of input texts and annotates information about tokens, sentences and paragraphs of these texts. Texts and annotation results are stored and made accessible using the MediaWiki software. The annotations are easy to edit thanks to the use of an annotation syntax using the functionality provided by MediaWiki. That is, annotations are presented in an intuitive way, keeping the source syntax simple and compact for fast editing and further processing. Corpus2Wiki can be used for determining peculiarities of different texts about the same or different topics, entities, geographical locations etc.

Corpus2Wiki was developed on the basis of established components whose user interfaces were evaluated for their user-friendliness and iteratively improved on the basis of several

[1] Goethe Universität Frankfurt, Text Technology Lab, Frankfurt, Germany s9642857@stud.uni-frankfurt.de
[2] Goethe Universität Frankfurt, Text Technology Lab, Frankfurt, Germany s2436677@stud.uni-frankfurt.de
[3] Goethe Universität Frankfurt, Text Technology Lab, Frankfurt, Germany hemati@em.uni-frankfurt.de
[4] Goethe Universität Frankfurt, Text Technology Lab, Frankfurt, Germany mehler@em.uni-frankfurt.de

doi:10.18420/inf2019_ws14

evaluations. One of these components is MediaWiki, with the help of which users create and edit wiki pages. Vora et al., for example, evaluated the usability of Wikipedia [VKT10]. Based on such findings the usability of MediaWiki was further improved. The use of such properly tested and generally known components contributes to the usability of Corpus2Wiki. Further, Khemakhem et al. found in their study on NLP software that the vast majority of users easily manage to set up software tools with Docker [KHR18]. Docker provides a good basis for easy installation and configuration of NLP tools and is consequently used by Corpus2Wiki. In the version of Corpus2Wiki presented here, the installation, configuration and generation of Wikiditions is automated to the extent described below. Corpus2Wiki remains below the functional scope described in [Me16], but due to its automation depth it facilitates the creation of Wikiditions by far.

Natural Language Processing (NLP) now allows for processing large numbers of texts [HHZ10; Ke98; Ta13]. Corpus2Wiki automatically creates annotations of texts and performs preliminary analyses to support their further processing. For this purpose, Corpus2Wiki uses the services of TextImager which provides various NLP components for several languages [HUM16]. TextImager takes over the resource and time intensive task of automatic annotation. The results received from TextImager are then imported into the MediaWiki database of Corpus2Wiki. In the current version of Corpus2Wiki, we improved its front- and backend by simplifying the installation and configuration, introducing a more powerful and user-friendly import procedure, and adding more text information and visualizations.

2 Related Work

Corpus2Wiki further develops the Corpus2Wiki tool of [RHM18] which is based on Wikidition [Me16; MWG16]. Web-based applications such as WebAnno [Yi13] and WebNLP [Bu14] usually provide easy setup and eliminate the need to maintain private servers. While these tools store the data in the cloud, Corpus2Wiki can store and manage all its files on a local machine. This may be desirable for users who prefer not to store files on a third-party server and may also reduce the costs of external data storage. Furthermore, WebAnno differs from Corpus2Wiki in that the former mainly address manual annotations, while Wikiditions are designed to provide many automatic annotations as a starting point for subsequent wiki-based annotations. On the other hand, advanced tools such as GATE Teamware[5] or FLAT[6] allow for multi-layer annotations and complex structural analyses. However, the installation and use of such sophisticated tools may require in-depth computer knowledge or additional training. Furthermore, some of these tools, such as FLAT, require a special input format (e.g. FoLiA). Corpus2Wiki offers a simple and uncomplicated installation through the use of virtualization software and allows the user to abstract from technical details. As it is based on MediaWiki, many users will be familiar with its interface. Therefore, not much training is required to operate the software.

[5] https://gate.ac.uk/teamware/
[6] https://github.com/proycon/flat

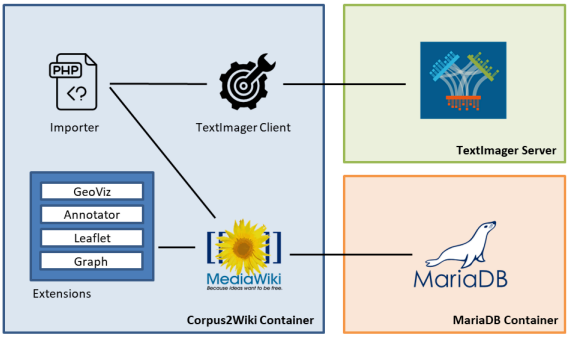

Fig. 1: The architecture of Corpus2Wiki.

3 Implementation

We describe the architecture of Corpus2Wiki and its implementation. In addition, improvements compared to [RHM18] are highlighted. Corpus2Wiki is implemented by means of two docker containers. Docker, an operating system level virtualization software, is used to abstract from the user's operating system and system environment. Using two docker containers as virtual environments allows the automatic configuration and customization of the virtual environment using scripts [Bo15]. Firstly, this concerns a MariaDB container that stores the analyzed corpora. Secondly, this refers to the Corpus2Wiki container, which provides import, analysis and visualization functionality. It is based on a MediaWiki container that is automatically configured and adjusted during installation and extended by a module for the import of corpora. This module contains a web interface and a TextImager client [HUM16], which is required for NLP. To make it easier to import TextImager results into MediaWiki, a module called MediaWiki Writer and post-processing scripts have been added to the TextImager client to convert data into a format that can be read by MediaWiki. In addition, MediaWiki has been extended with standard and custom extensions for annotating and visualizing annotations. Figure 1 gives an overview of the components.

MediaWiki was chosen as the core technology for storing, visualizing and editing annotations generated by TextImager. MediaWiki has an active support and development community and its user interface is familiar to many users [Ko17]. Its editing functionality is versatile and easy to use. This supports the further work with text corpora, which are finally released as Wikiditions.

An advantage of Corpus2Wiki is that it allows users to run their own instance of the software and store results locally or on their server. Installation and configuration do not require detailed technical knowledge for the average user to perform these actions, as Corpus2Wiki abstracts from the system running the software using Docker [Bo15]. This facilitates automatic installation and configuration and makes the software portable for a variety of systems. With this version of Corpus2Wiki, installation is as easy as running a

script; both docker containers are automatically set up and configured. Compared to the previous version, the number of steps required to set up the software was significantly reduced, while the range of functions was considerably expanded.

One focus of Corpus2Wiki is usability. To this end, the import procedure was revised and the command line interface (CLI) was replaced by a graphical web interface. This has a number of advantages: firstly, using a web interface requires much less technical knowledge and is more intuitive than using a CLI. Secondly, the new process is more flexible and configurable and can also be performed from remote computers. Lastly, since the entire process has been moved to Docker containers, it is no longer dependent on the configuration of the host system. This makes it more error-resistant and maintainable. The entire import process is automated and the complexity is abstracted from the user. With the new import interface, the user only has to interact with a simple web form.

Corpus2Wiki can display annotations at different levels: annotations on the word, sentence, paragraph and document level are supported (see Figure 2) (see [Me16] for the underlying text corpus model). A MediaWiki extension called Corpus2Wiki-Annotator was implemented to display graphical tool-tips and provide word highlighting. Since the extension is designed for use with attribute-value pair annotations, all such information can be displayed without changing the visualization component of the software. Currently, the TextImager client and MediaWikiWriter support lemmatization, POS tagging, tagging of morpho-syntactic information, named entity recognition, and topic labeling using the *Dewey Decimal Classification* (DDC) as a reference classification [Us18]. In addition, histograms displaying the distribution of POS frequencies are generated for each input text. Finally, Corpus2Wiki integrates GeoViz [HUM16], a software for displaying geotagging data. It allows the marking of all places mentioned in a text on an interactive map. GeoViz offers a variety of functions, such as displaying relations between events, persons and places. Currently, however, many of its functions are not available in Corpus2Wiki. In order to take full advantage of the functionality of GeoViz, the tool has to be refactored to work as a proper MediaWiki plugin.

A second focus of Corpus2Wiki is on facilitating the editing of processed texts. This is important because a flexible post-processing and revision are required to help ensure the validity of results of NLP in DH. MediaWiki already offers a standard editing interface and also tracks the history of write events. The challenge, however, is that annotations and the use of multiple extensions make the underlying text virtually unreadable for technically inexperienced users. By writing MediaWiki extensions, we were able to better support the visual representation of annotations. These extensions provide a simpler and more intuitive syntax for annotating and visualizing information.

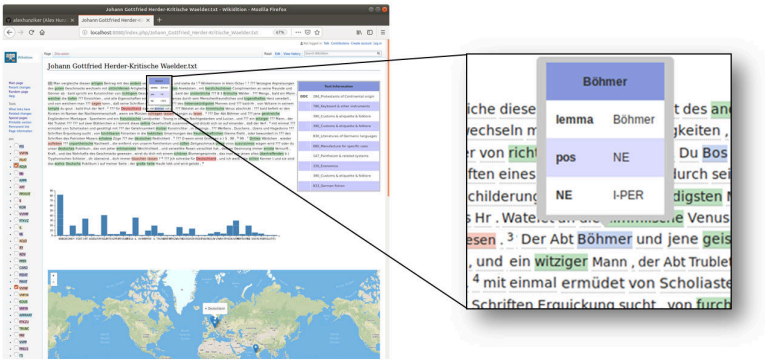

Fig. 2: A screenshot of Corpus2Wiki.

4 Conclusion and Future Work

We presented Corpus2Wiki as a tool for the analysis, storage, visualization and editing of corpora based on the corpus model of Wikidition [Me16]. Since TextImager is used for NLP, a multitude of text information and annotations are generated by means of Corpus2Wiki. The underlying MediaWiki provides simple yet powerful editing functions for processing the resulting Wikiditions. Compared to [RHM18], Corpus2Wiki has an improved software architecture and automates the installation routine. The adaptability of its import component was bettered by software refactoring, as the usability was enhanced by a web interface. More information regarding morphological information, named entities and topic labels can now be further processed. In addition, two custom extensions have been developed for Corpus2Wiki that provide annotation and text highlighting capabilities, as well as the ability to display locations named in a text on an interactive map. Since Corpus2Wiki has a syntax that allows the display of annotations in the form of attribute-value pairs, it is possible to add corresponding annotations without having to change the code. The only change required is to update TextImager's MediaWiki Writer to support these annotations when they are available through TextImager. Since Corpus2Wiki is based on MediaWiki, both standard and custom extensions for this platform can be used to further enrich the visualization of texts. Examples for additional visualizations are parse trees, relation extraction graphs or timelines to display the chronological order of text content as provided by TextAnnotator [Ab19].

As a wiki-based tool, Corpus2Wiki can be extended for many purposes. Using TextImager, it includes powerful NLP functionality, is open source, and can also be run locally on users machines. It is relatively easy to install and use, offers visualizations of textual information and allows quick and easy editing. For these reasons, we believe that its further-development has the potential to become a powerful tool for DH.

5 System Demonstration

The source code of Corpus2Wiki is open and accessible on GitHub:
`https://github.com/texttechnologylab/textimager-corpus2Wiki`

A demo of a running instance can be found on the following website of the Text Technology Lab:
`https://textimager.hucompute.org/corpus2wiki/`

References

[Ab19] Abrami, G.; Mehler, A.; Lücking, A.; Rieb, E.; Helfrich, P.: TextAnnotator: A flexible framework for semantic annotations. In: Proceedings of the Conference: Fifteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation. Gothenburg, Sweden, May 2019.

[Bo15] Boettiger, C.: An introduction to Docker for reproducible research. ACM SIGOPS Operating Systems Review 49/1, pp. 71–79, 2015.

[Bu12] Burdick, A.; Drucker, J.; Lunenfeld, P.; Presner, T.; Schnapp, J.: Digital Humanities. MIT Press, 2012.

[Bu14] Burghardt, M.; Pörsch, J.; Tirlea, B.; Wolff, C.: WebNLP: An Integrated Web-Interface for Python NLTK and Voyant. In (Ruppenhofer, J.; Faaß, G., eds.): Proceedings of the 12th edition of the KONVENS conference : Hildesheim, Germany, October 8-10,2014. Universitätsbibliothek Hildesheim, Hildesheim, pp. 235–240, 2014, URL: `https://epub.uni-regensburg.de/35712/`.

[HHZ10] Hinrichs, E.; Hinrichs, M.; Zastrow, T.: WebLicht: Web-based LRT Services for German. In: Proceedings of the ACL 2010 System Demonstrations. ACLDemos '10, Association for Computational Linguistics, Uppsala, Sweden, pp. 25–29, 2010.

[HUM16] Hemati, W.; Uslu, T.; Mehler, A.: TextImager: a Distributed UIMA-based System for NLP. In: Proceedings of the COLING 2016 System Demonstrations. Federated Conference on Computer Science and Information Systems, Osaka, Japan, 2016.

[Ke98] Kennedy, G.: An introduction to corpus linguistics. Longman, 1998.

[KHR18] Khemakhem, M.; Herold, A.; Romary, L.: Enhancing Usability for Automatically Structuring Digitised Dictionaries. In: GLOBALEX workshop at LREC 2018. Miyazaki, Japan, May 2018.

[Ko17] Koren, Y.: Working with MediaWiki. CreateSpace Independent Publishing Platform, 2017.

[Me16] Mehler, A.; Gleim, R.; vor der Brück, T.; Hemati, W.; Uslu, T.; Eger, S.: Wikidition: Automatic Lexiconization and Linkification of Text Corpora. Information Technology 58/2, pp. 70–79, 2016.

[MWG16] Mehler, A.; Wagner, B.; Gleim, R.: Wikidition: Towards A Multi-layer Network Model of Intertextuality. In: Proceedings of DH 2016, 12-16 July. DH 2016, Kraków, 2016.

[RHM18] Rutherford, E.; Hemati, W.; Mehler, A.: Corpus2Wiki: A MediaWiki based Annotation & Visualisation Tool for the Digital Humanities. In (Burghardt, M.; Müller-Birn, C., eds.): INF-DH-2018. Gesellschaft für Informatik e.V., Bonn, 2018.

[Ta13] Tablan, V.; Roberts, I.; Cunningham, H.; Bontcheva, K.: GATECloud.net: a Platform for Large-Scale, Open-Source Text Processing on the Cloud. Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences 371/1983, ed. by Townend, P.; Xu, J.; Austin, J., p. 20120071, 2013.

[Us18] Uslu, T.; Mehler, A.; Niekler, A.; Baumartz, D.: Towards a DDC-based Topic Network Model of Wikipedia. In: Proceedings of 2nd International Workshop on Modeling, Analysis, and Management of Social Networks and their Applications (SOCNET 2018), February 28, 2018. 2018.

[VKT10] Vora, P.; Komura, N.; Team, S. U.: The N00B Wikipedia Editing Experience. In: Proceedings of the 6th International Symposium on Wikis and Open Collaboration. WikiSym '10, ACM, Gdansk, Poland, 36:1–36:3, 2010.

[Yi13] Yimam, S. M.; Gurevych, I.; Eckart de Castilho, R.; Biemann, C.: WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Sofia, Bulgaria, pp. 1–6, Aug. 2013, URL: https://www.aclweb.org/anthology/P13-4001.