O uso da Wikipedia como fonte de suporte para pesquisas em idiomas com recursos digitais insuficientes

The Scenario of Wikipedia Usage as a supporting source for under-resource language researches

Lucia Dwi Krisnawati

Duta Wacana Christian University - Indonésia
krisna@staff.ukdw.ac.id

Aditya Wikan Mahastama

Duta Wacana Christian University - Indonésia
mahas@staff.ukdw.ac.id

Resumo

Atualmente, a Wikipedia se tornou a maior e mais importante coleção de dados abertos na Web. Também houve um aumento significativo no uso de artigos da Wikipedia como referência em artigos científicos e relatórios de pesquisa. Neste artigo, revisamos projetos de pesquisa no campo da technologia da informação que mais se beneficiaram com o surgimento da Wikipedia. Em Seguida, descrevemos o cenário de uso da Wikipedia no desenvolvimento do silabificador javanês e sistema de identificação de Idioma. Também descrevemos a participação da comunidade Javanêsa de Wikipedia na anotação dos caracteres Javanêses.

Palavras-chave: Uso da Wikipedia, silabificação, anotação de carateres Javanêses

Abstract

Nowadays, Wikipedia has become the largest and prominent collections of open data on the Web. There has been also a significant increase in the use of Wikipedia articles as a reference in scientific articles and research reports. In this paper, we review research projects within the field of Information Technology which benefited most from the rise of Wikipedia. Then, we described the scenario of Wikipedia usage in developing Javanese syllabifier and language identification system. We also described the participation of Javanese Wikipedia community in annotating the Javanese characters.

Keywords: *Wikipedia Usage, Syllabification, Javanese Character Annotation*

1. Introduction

After almost two decades from its release date, Wikipedia has become the largest free online encyclopedia. The fact that Alexa, the Amazon.com Company, ranked it on the 5th position of the most popular websites proves that it has great number of visit per day. The rate of its visitors is heighten by Google reference which very often puts Wikipedia pages at the top of its query results. There has been also a significant increase in the use of Wikipedia as a reference within all areas of science and scholarship. As an effect of recontextualization efforts of Wikipedia knowledge, Lindgren noted that Wikipedia is used as a complement, repository, or as an unproblematic source of information in areas like Computer Science, Mathematics, Social Sciences, and Arts and Humanities than in Natural Sciences and Medicine (Lindgren, 2014).

The trust to use Wikipedia as a source of reference in scientific articles and reports is inseparable from its article coverage and number. Per 25 January 2019, Wikipedia provided in total 49,446,866 articles displayed in 198,197,107 sites and written in 303 languages[1]. The most contribution has been given by English articles which reached the number of 5,903,325 articles. The German articles are worth being noted as its number reached 2M+ with 48 editors per million speakers which is much higher than the number of English editors (27 per million speakers). Based on this statistics, it is no wonder that there have been abundant Wikipedia-based projects and researches. Among the notable ones are DBpedia which extracts structured data from infoboxes in Wikipedia (Huang, 2015), and Wikidata which extracts structured data from Wikipedia and turns them in form of linked knowledge graphs (Malyshev, Krötzsch, González, Gonsior, & Bielefeldt, 2018).

In contrast to this statistics, the number of Indonesian articles provided by Wikipedia is relatively small with the article count estimating around 400K or 0.9% of the English Articles. The much smaller article count was demonstrated by the ones written in Javanese -- a local language spoken by majority of Indonesians – which estimated around 55K or 0.1% of the English article rate and 12% of the Indonesian article count. This rate could be a good sample of the availability of Javanese documents in digital forms. Based on the number of document availability in digital form, languages are classified into 4 categories: high-resourced languages, under-resourced languages, critical, and endangered languages (Cieri, Maxwell, Strassel, & Tracey, 2016). In Cieri et al. (2016), it is explicitly stated that Indonesian is included in the category of under or low-resourced language. The Javanese would fall into both critical and under-resourced language (Krisnawati & Mahastama, 2018). For this reason, we are interested in digitizing Javanese manuscripts in our long-term research. However, we present some sub-projects dealing with the usage of Wikipedia's articles and resources in this paper.

The paper is organized to comprise 5 sections, in which the first states the introduction. A brief overview on Wikipedia-based projects and researches would be presented on the second section. Following it, the description on usage scenario of Wikipedia's articles as a corpus for Javanese syllabifier is presented on Section 3. Section 4 deals with the discussion on our on-going project of annotating Javanese characters for Wikisource. The last section will present the summary of the paper.

[1] The statistic data were derived from Wikimedia statistics on https://stats.wikimedia.org/EN/

2. A brief Overview on Wikipedia-based Projects

We did a short-term survey on articles describing researches or projects based on Wikipedia and its by-products. The study area of surveyed articles is limited within the scope of Information Technology. For the Wikipedia-based researches in the area of social sciences, readers could refer to the review in (Khoury, 2009). This section would present our review which is grouped according to the fields of Information Technology (IT).

The existence of Wikipedia as a free online encyclopedia has made it to be a data repository for many researches which triggered new innovation in some fields of IT. The most benefitted IT fields from the rise of Wikipedia are Natural Language Processing (NLP), data modeling, and Knowledge Acquisition. In NLP, Wikipedia's category hierarchy was used as classes to improve algorithm of a topic-based Text Classification (Schönhofen, 2006), while Khoury, challenged by the complexity of query processing, developed an automatic query classification system which addressed the problems of query's wide variety and broad range of topics (Khoury, 2011).

Still in the area of NLP, Banerjee and Mitra (2015) built an automatic text summarization system which summarizes existing web content and utilizes the resulted summary to improve the incomplete Wikipedia articles. Unlike Banerjee and Mitra, Milne et al. (2006) were able to automatically build a thesaurus by exploiting Wikipedia's structural similarities. If a thesaurus contains a semantically related words, two documents containing many similar words would be consider as a duplicate or partially duplicate. Measuring text similarity becomes one of the NLP challenges. However, Wee and Hasan (2008) developed a simple but effective way of computing the directional similarity between two texts by using the ratio of the number of Wikipedia articles containing similar words to the total number of articles in which these words occur. Based on topically similar texts, Picardi et al. (2018) built a system which is able to give recommendation to Wikipedia's editors what section to add to the existing or newly created articles. This recommendation system is aimed to solve the difficulties faced by Wiki's writers in structuring new articles and their needs on "significant knowledge about how a well-written article looks for each possible topic" (Piccardi, Zia, Catasta, & West, 2018).

Knowledge acquisition is the process of extracting, structuring and organizing knowledge from either human experts or texts. Researches in knowledge acquisition tend to focus on the mining of semantic information from input sources and represent it into a structured form (Khoury, 2009). Its output, the structured knowledge, is very beneficial for text processing (NLP) and practically contributes to various data models. To begin with, a preprocessing technique for Wikipedia data mining task was proposed in (Boldi & Monti, 2016) by pruning and cleansing Wikipedia category hierarchy with a tunable level of aggregation, while in Gupta et al. (2016), Wikipedia category network was used to automatically construct a unified taxonomy which assembles entities and categories in a *is-a* relation. The automatic extraction of *is-a* and *part-of* relations from Wikipedia articles was also conducted by Arnold and Rahm (2015). Those relations are then used to build up a large and up-to-date thesaurus which provides background knowledge for determining semantic ontology mapping (Arnold & Rahm, 2015). Mining links and text snippets from Wikipedia as a new knowledge base was also conducted by Wira-Alam and Matthiak (2012), and links between topically similar articles of Wikipedia in different languages were automatically created by Wang et al. (2012). The semantic extraction from Wikipedia texts is made possible by semantic annotation (Schindler & Vrandečić, 2011) to add structured data to Wiki pages.

At last, the system performance in extracting the semantic data should be evaluated by reliable metrics which were introduced in (Kruit, Boncz, & Urbani, 2018).

Knowledge extraction from Wikipedia resulted in numerous outstanding works which attracted significant interest in researches. Among these which contribute to the data modeling are DBpedia, Wikidata, and IMGpedia. DBpedia, developed by researchers from Leipzig and Mannheim Universities (Huang, 2015), has grown up to be a crowd-sourced community effort to extract structured content (knowledge) from Wikipedia and make it widely available by semantic web standards and Linked Data (Lehmann, et al., 2012). Started in 2006, DBpedia stores structured information in an open knowledge graph (Huang, 2015) which is a special kind of database in an RDF format. DBpedia becomes the key factor for the success of the Linked Open Data initiative (Lehmann, et al., 2012).

Being a sister project hosted by Wikimedia Germany e.V, Wikidata extracts different kinds of structured information from Wikipedia. It contains various data types as it extracts not only data in form of texts but also images, quantities, coordinates, geographic shapes and dates (Huang, 2015). However, Lehnmann et al. (2012) noted that Wikidata does not explicitly state the truth about the things, but provides statements about them. Thus, given a query on who the wife of Joko Widodo is, Wikidata will return different statements containing Iriana Widodo as the Indonesian first lady and the wife of Mr. Joko Widodo, but not directly state that she is the wife of Joko Widodo. This problem has been tackled by DBpedia. Another large-scale linked dataset is IMGpedia which comprises approximately 15 Million images and 450 million visual-similarity relations between those images (Ferrada, Bustos, & Hogan, 2017). IMGpedia incorporates visual descriptors and visual similarity relations for images in WIKIMEDIA COMMONS linked with relevant knowledge-bases of Wikidata as well as DBpedia datasets (Ferrada, Bravo, Bustos, & Hogan, 2018). The multimedia data in IMGpedia could be browsed and retrieved in a more friendly manner since a new web interface has been created by Ferrada et al. (2018).

3. Wikipedia as a source of corpus building

In this section, the scenario of using Wikipedia articles as a source of corpus for Javanese syllabifier and language identifier system is described.

3.1. Data collection and annotation for an automatic syllabification process

Javanese is considered to be one of world's classical languages (Thompson, 2016) with literary tradition over a thousand years. Javanese has been written in Javanese script which is an Abugida type – a segmental writing system in which consonant-vowel sequences are written as a syllable unit (Krisnawati & Mahastama, 2018). Javanese texts are written from left to right without word boundary (Scriptio Continua). As an effect of the introduction of Latin alphabet by the Dutch in 19th century (colonial era), Javanese is then written in Latin. Nowadays, Javanese scripts occur only on the school books for learning Javanese, the street name posts in some major cities in central Java, on the old manuscripts and historical documents. For this reason, The Javanese syllabifier was constructed with the goal to provide a corpus of Javanese syllables for future-projected word boundary prediction in a process of automatic transliteration of Javanese script into Latin. For this reason, Javanese documents written in Latin were needed as a test set as well as the data set.

The data set was acquired through automatic web scraping by means of a crawler engine which has been assigned to extract Javanese documents written in Latin. Owing to the fact that there has been only a handful of websites written in Javanese, the scrapping activity was directed to specific domains, that is to jv.wikipedia and *Djaka Lodang* which is a website of a Javanese magazine written in Latin. In total, there are 246 documents extracted from Javanese Wikipedia and 20 documents from Djaka Lodang. The small number of scrapped documents from Djaka Lodang is mainly due to the shortage of document availability. Another reason is that the document length in Djaka Lodang has been considered to be inappropriate for the manual annotation process, while the length of Javanese Wikipedia articles has been considered to be ideal. Figure 1 illustrates the process of data collection and its use case scenario for evaluating the syllabifier's performance.

Figure 1 shows that as the crawler engine extracts articles from Javanese Wikipedia and Djaka Lodang, it saves the articles in the plain text format to be a corpus or data set. Some documents from this corpus whose length is tolerable for manual labeling were given to Javanese experts to annotate. The annotation which is a process of segmenting words into syllables was done by inserting a dot as the syllable delimiter. The annotators were informed to segments syllables based on the Javanese orthographic rules. Thus the syllabification process tends to be grapheme-based rather than the phonological one. The annotated documents become the gold labels which are used as a basis of evaluating the syllabifier outputs. The total number of annotated documents reaches 112. However, only 40 documents, in which 20 documents were randomly taken from Wikipedia articles and the rest from Djaka Lodang, were used as test documents. The aim is to have a balance of a test set composition from different sources. The same 40 documents which have not been annotated were fed to the syllabifier, and their outputs were automatically compared to the gold labels to compute syllabification accuracy done by our Javanese syllabifier. Figure 2 displays an example of an input and output documents to our syllabifier.

Fig. 1 the illustration on the process of automatic Web scrapping for data collection and its use case for evaluating our syllabifier performance

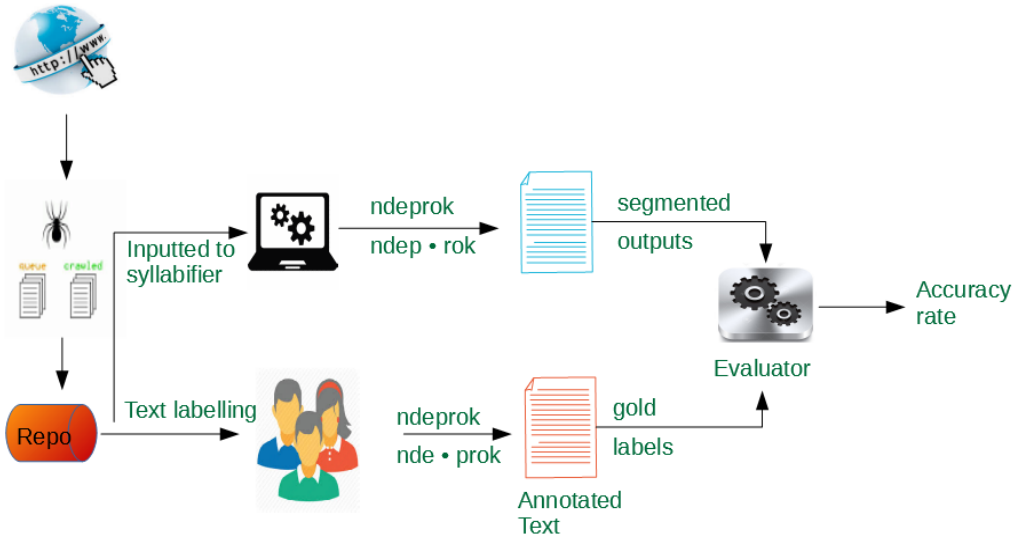
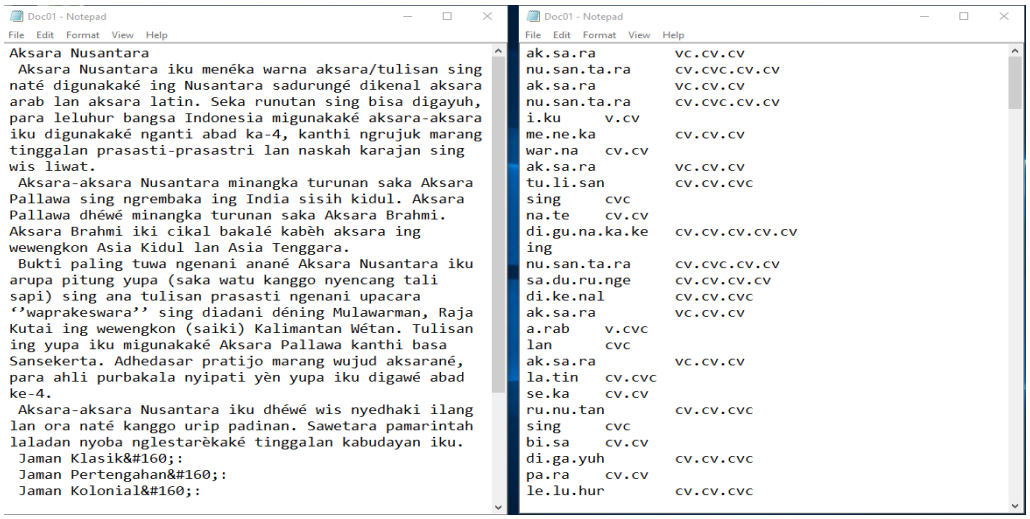


Fig. 2 An example of original document extracted from Wikipedia to be the input of syllabifier (left), and a list of segmented words in syllables and their syllable patterns as the output of our syllabifier (right).



Our syllabifier was built by applying the Finite State Transducer (FST) model which is combined with the hand-written rules for defining the state transition. The states in FST machine were reconstructed to represent each syllable pattern. A state will be visited only if the rules defined in it match the string sequences of a word. Then, the segmentation was executed. The output documents were then evaluated with the accuracy metric. Given 40 test documents, the averaged accuracy rate of documents scraped from Wikimedia reaches 95.56%, while the accuracy rate for Djaka Lodang dokuments achieves 97.92% which is 2.36% higher. Some possible reasons for the higher accuracy rate of the Djaka Lodang's articles are due to the writer homogeneity in using the speech style (*krama* -- polite style). In contrast, the Wikipedia articles were written collaboratively by writers using different dialects of Javanese and speech styles such as *Ngoko* (informal speech), *krama*, a mixture beween *krama* and *ngoko*. Besides, Wiki documents contain more loan and foreign words and a relativley high variety of spelling for the same words. For example, *mahabarata* is also spelled as *mahabharata*. As its consequence, the syllables 'ba' and 'bha' in those spellings would be considered as an unmatched segment by our automatic evaluator.

3.2. Building language profiles for indonesian and javanese

The main goal of building a language identifier system is to enable a smart selection of documents to save as a result of automatic Web crawling and scraping. To do so, a language identifier system should have a knowledge on the profiles of languages being recognized. For this moment, we designed our language identifier to identify three classes of languages, i.e. Javanese, Indonesian, and others. This means that any language of a text which is not recognized as either Javanese or Indonesian will be flagged as 'others'.

To build a language profiles, raw documents in a specific language are needed. For this need, we applied the same method in acquiring digital documents, namely web scrapping. This time, we scrapped from different sites but most of articles were extracted from Wikipedia. The reason is that Wikipedia articles cover a wide range of topics, contain more loan words which are very important in building the language profiles. For increasing the variety of vocabulary, we scraped articles from different genre such as news articles (CNN Indonesia), literary works such as folklore and poems too.

Table 1 presents the statistic of documents used to build the language profiles along with the website names. From this table, it can be seen that 75.26% articles were scraped from Wikipedia.

Table 1 The statistic data on the number of scraped articles and their sources

Website names	# articles in Indonesian	# articles in Javanese	Total number
Wikipedia	41	32	73
Sastra.org (literary works)	-	8	8
Detik.com (weekly magazine)	7	-	7
Cnnindonesia.com (newspaper)	7	-	7
Cerita Rakyat Indonesia (folklore)	1	-	1
Puisi Indonesia (poems)	1	-	1
TOTAL	57	40	97

To get the profiles of Javanese and Indonesian, firstly document normalization was done by changing all characters in lower cases, eliminating all punctuation marks, reducing multiple spaces into a single one, and replacing each space with an underscore. The character n-grams with n between 2-5 were generated from the normalized documents. The occurrence of each n-gram was then summed up and saved along with the n-grams for each language. The n-grams were then sorted based on their frequency of occurrence and only the top 100 n-grams were selected to be the profiles of each language. The similarity of a tested document to each language profile is computed using The Out-Of-Place measure introduced in (Cavnar & Trenkle, 1994). Being tested to 67 documents in Javanese (31), Indonesian (26), Malay (5) and English (5), the accuracy of our language identifier reaches 85.1%.

4. Annotating javanese characters for Wikisource

The Javanese character annotation is a subtask of our Optical Character Recognition (OCR) project. The training data for Javanese characters were acquired from scanned manuscripts. Each page of the scanned manuscript were then automatically segmented into individual characters whose features were also automatically extracted. The annotation task deals with writing metadata for each character which comprises the individual character transliteration in Latin, its features, font, style, the segment condition, and the elements which make a complete character representing a syllable

At first, we considered to use scanned Javanese manuscripts available in Wikimedia Commons. We dropped this idea since the image resolution of scanned manuscript was too low, while our segmentation and feature extraction algorithm requires that the scanned image should have at least 600 DPI in colour (24-bit). We decided to scan the different manuscripts ourselves to meet this requirement. The first manuscript to scan is *Serat-serat Anggitanipun Kangjeng Gusti Pangeran Adipati Arya Mangkunegara IV* (referred as Serat Mangkunegaran). The average number of character in one page reaches 250 characters and we annotated the first 20 pages of this manuscript.

Annotating 5000 characters manually is too laborious for our 3 annotators who are informatics students. The annotation process was considered to be too slow, hence we changed the annotation strategy and tools. In the new strategy, the annotation is crowdsourced. The Wikimedia Indonesia is projected to organize the crowdsourcing process. The annotation resulted from the crowdsourcing would be evaluated by Javanese Wikipedia community whose expertise in reading and writing Javanese characters is very reliable.

Owing to the strategy change, the annotation tool has been redesigned to be more user-friendly for annotators having no background of Informatics. The first annotation tool was a web-based form containing some fields which need to fill in with the metadata which was acquired from running our segmentation and feature extraction software. The annotators should also fill in each feature of a character manually and the annotation result was then saved in an XML format document. This would be cumbersome for both annotators involved in the crowdsourcing process and for the Javanese Wikipedia community.

The new tool which has been developed is also a web-based one but there is only one single field which needs to be filled in manually. This field is the transliteration in Latin for an individual segmented character. The other metadata were served as a drop-down menu to ease the annotators and evaluators. The annotation process starts as a user logs in to the anotation sysem which is hosted in http://trawaca.id/anotasi. The system will display the options of scanned pages in which an annotator is free to choose one of them by clicking the *Select* button (tombol *pilih*). The clicked button will activate the segmentation and feature extraction software which delivers segmented characters. Figure 3 shows the screenshot of the segmented characters which are organized and indexed by the lines and its occurrence order in each line. When a segment is being annotated by a user, other users are unable to annotate it. This is done to prevent data conflicts in sending data to our database. In this use case, there is no different interface between annotators and evaluators. The evaluators are treated as annotators since both of them are able to annotate the unlabeled characters as well as revised the annotation if they think there is a mistake in a character annotation. The annotators are also able to see which characters have been annotated and which have not. Each user's activity over a segment is recorded in a log, so each character segment will have a complete annotation history, which is visible and traceable by the web administrator. Figure 4 displays the screenshot of the page in which an annotator can do the annotation by clicking mostly the provided drop-down menu.

Fig. 3: The character segment selection screen, showing segments from Serat Mangkunegaran page 3, which consists of 8 lines and 154 segmenthe s in red rectangle. The yellow box denotes an already-annotated segmented character, while white ones show unannotated character.

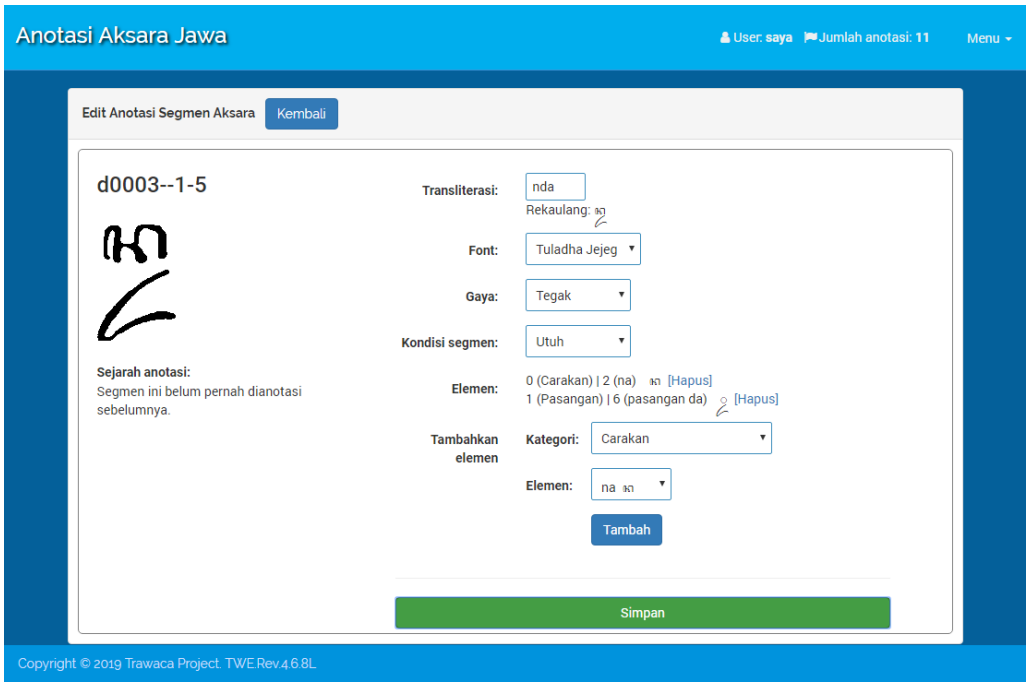
5. Summary

In this paper, we have described our literary review concerning the Wikipedia and its by-product usage in research projects. We have presented also that Wikipedia serves not only as a source of data, but also has stimulated innovation in some fields of Information Technology, for example the DBpedia and Wikidata with their Linked Open Data (LOD). In this article, we have also shared two of our projects which used Wikipedia articles as a source of corpus building and our on-going project concerning the annotation of Javanese characters. From these projects we learnt that Wikipedia becomes a prominent source of digital documents for projects concerning the under-resource languages such as Indonesian and Javanese. Besides, the participation of Wikipedia community also plays an important role in supporting such research projects.

Acknowledgement

We thanks to Wikimedia Indonesia (WMID) which has funded this project. Our special thanks go to LPPM, Duta Wacana Christian University which has partially funded the syllabification project, to Java Wikipedia Community which supports the annotation process, to Samuel Eddijanto, Michell Bernardi S., Nana E. Wulandari, Ofri C. Valent, and Fidelia V. Santoso who have volunteered to the annotation process.

Fig. 4: The annotation screen of a single selected character. This screen shows that the only field to be manually typed is 'Transliterasi' (Latin transliteration), other metadata are selected through a drop-down menu



Referências Bibliográficas

ARNOLD, P., & RAHM, E. (2015). Automatic Extraction of Semantic Relation from Wikipedia. *International Journal on Artificial Intelligence Tool, 24*(2), 24 pages.

BANERJEE, S., & MITRA, P. (2015). WikiKreator: automatic authoring of Wikipedia content. *AI Matters, 2*, 4-6.

BOLDI, P., & MONTI, C. (2016). Cleansing Wikipedia Categories using Centrality. *he International World Wide Web Conference Com-.* Montreal: ACM.

CAVNAR, W., & TRENKLE, J. (1994). N-Gram-Based Text Categorization. *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, (pp. 161-175).

CIERI, C., MAXWELL, M., STRASSEL, S., & TRACEY, J. (2016). Selection Criteria for Low Resource Language Programs. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Portoroz, Slovenia: European Language Resources Association (ELRA).

FERRADA, S., BRAVO, N., BUSTOS, B., & HOGAN, A. (2018). Querying Wikimedia Images using Wikidata Facts. *In WWW'18 Companion: The 2018 Web Conference Companion* (p. 7 pages). Lyon, France: ACM.

FERRADA, S., BUSTOS, B., & HOGAN, A. (2017). IMGpedia: A Linked Dataset with Content-Based Analysis of Wikimedia Image. *International Semantic Web Conference.*

GUPTA, A., PICCINNO, F., KOZHEVNIKOV, M., PASCA, M., & PIGHIN, D. (2016). Revisiting Taxonomy Induction Over Wikipedia. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, (pp. 2300-2309). Osaka, Japan.

HUANG, A. (2015). *A Preliminary Study on Wikipedia, DBpedia, and Wikidata.* Retrieved January 2019, from URL: http://andrea-index.blogspot.tw/2015/06/wikipedia-dbpedia-wikidata.html

KHOURY, R. (2009). The Impact of Wikipedia on Scientific Research. *3rd International Conference on Internet Technologies and Applications*, (pp. 2-11).

KHOURY, R. (2011). Query Classification Using Wikipedia. *International Journal of Intelligent Information and Database Systems, 5*(2), 143-163.

KRISNAWATI, L. D., & MAHASTAMA, A. W. (2018). A Javanese Syllabifier Based on its Orthographic Forms. *International Conference on Asian Language Processing.* Bandung, Indonesia.

KRUIT, B., BONCZ, P., & URBANI, J. (2018). Extracting New Knowledge from Web Tables: *KBCOM'18.* Los Angeles.

LEHMANN, J., ISELE, R., JAKOB, M., JENTZSCH, A., KONTOKOSTAS, D., MENDES, P. N., . . . Bizer, C. (2012). DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web I*, 1-5.

LINDGREN, S. (2014). Crowdsourcing Knowledge Interdiscursive Flows from Wikipedia into Scholarly Research. *Culture Unbound: Journal of current cultural research, 6*, 609-627.

MALYSHEV, S., KRÖTZSCH, M., GONZÁLEZ, L., GONSIOR, J., & BIELEFELDt, A. (2018). Getting the Most out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge. *ISWC'18* (pp. 376-394). Springer.

MILNE, D., MEDELYAN, O., & WITTEN, I. H. (2006). Mining domain-specific thesauri from Wikipedia: a case study. *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, (pp. 442-448).

PICCARDI, T., ZIA, L., CATASTA, M., & WEST, R. (2018). Structuring Wikipedia Articles with Section Recommendations. *SIGIR'18.* Ann Arbor: ACM.

SCHINDLER, M., & VRANDEČIĆ, D. (2011). Introducing New Features to Wikipedia: Case Studies for Web Science. *IEEE Intelligent Systems, 26*, 56-61.

SCHÖNHOFEN, P. (2006). Identifying document topics using the Wikipedia category network. *Proc. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, (pp. 456-462).

THOMPSON, I. (2016, April). *About Wolrd Languages.* Retrieved from Javanese: http://aboutworldlanguages.com/javanese

WANG, Z., LI, J., WANG, Z., & TANG, J. (2012). Cross-lingual Knowledge Linking Across Wiki Knowledge Bases. *International World Wide Web Conference (WWW).* Lyon, France: ACM.

WEE, L. C., & HASSAN, S. (2008). Exploiting Wikipedia for directional inferential text similarity. *Proceedings of the Fifth International Conference on Information Technology: New Generations*, (pp. 686-691).

WIRA-ALAM, A., & MATTHIAK, B. (2012). Mining Wikipedia's Snippets Graph: First Step to Build A New Knowledge Base. *KNOW@LOD*, (pp. 43-48).