Coupling Wikipedia Categories with Wikidata Statements for Better Semantics

Houcemeddine Turki[1][0000-0003-3492-2014], Mohamed Ali Hadj Taieb[1][0000-0002-2786-8913] and Mohamed Ben Aouicha[1][0000-0002-2277-5814]

[1] Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia
turkiabdelwaheb@hotmail.fr, {mohamed.benaouicha, mohamedali.hajtaieb}@fss.usf.tn

Abstract. In this position paper, we explain how the combination of Wikipedia Categories already in use for driving semantic applications with Wikidata statements related to the categories and their direct members is possible from a technical perspective thanks to the flexible data models of Wikipedia Categories and Wikidata statements and to the programmatic options provided by the Wikimedia Community. We also outline how such a combination can bring an added value to the development of Wikipedia Projects as well as to the enhancement of knowledge-based systems.

Keywords: Wikipedia Categories, Wikidata, Semantic Applications.

1 Introduction

Nowadays, knowledge-based systems have become interestingly used for automating various tasks [1-3]. By contrast to other types of intelligent systems, knowledge-based systems use information resources as references for verifying and assessing inputs and generating reliable outputs [4]. These resources can be raw texts, semi-structured datasets or fully structured semantic databases [4, 5]. In this context, Wikimedia Projects hosted by Wikimedia Foundation can be valuable to drive knowledge-based systems as they are freely available and retrievable online and provide rich semantic information collaboratively developed by a large community of contributors [2]. In fact, Wikipedia provides a textual outline of multiple entities with a variety of semi-structured and structured information including wikilinks, redirections, infobox data and categories [6]. The sum of these semi-structured data have been useful to construct a series of knowledge graphs such as DBpedia for measuring semantic similarity among other applications [2, 7]. Particularly, Wikipedia Categories has been a very valuable resource to build an “is a” taxonomy that can be used for labelling datasets or supporting semantic applications [2]. Wiktionary provides brief definitions of nouns, adjectives, adverbs and verbs (so-called glosses) for a given language allowing to efficiently find semantic relatedness between concepts [8] and consequently to improve natural language processing algorithms [9]. Again, Wiktionary classifies its entities into categories. However, they rarely represent taxonomic relations between concepts and mostly provide an overview of the grammatical and etymological features of the represented entities [10]. One bright exception that should be generalized is the Russian Wiktionary which has 2 elaborated systems: so-called semantic categories built on the basis of taxonomic relations and semantic relations (synonymy, hypernymy, and so on) in each glossary entry. Wikidata (https://www.wikidata.org) is a large-scale multilingual knowledge graph that represents concepts as items and describes them using statements in the form of triples where Wikidata predicates define the types of the claims [11-12]. Here, Wikidata properties are entities that are described by logical constraints and Wikidata statements specifying the conditions of their practical usage and classes are assigned Shape Expressions (ShEx) defining how their members should be structured [3, 13]. The sum of all this information provides the backbone of a semantically rich information that can be used for various applications [3, 13]. Although every Wikimedia Project has distinctive features that make it useful for driving semantic applications, limited efforts have been made to combine the data provided by two wikis or more to develop semantic computations. Effectively, the only effort in this context was the combination of Wiktionary glosses with Wikipedia Categories for measuring semantic relatedness [14-17].

In this position paper, we introduce the combined use of Wikipedia Categories and Wikidata Statements for various applications related to Semantics and to the development of the two Wikimedia Projects. We begin by explaining the technical approaches that can be used to combine these two types of semantic data (Section 2). Later, we provide a detailed overview of the possible useful applications of such a combination in multiple fields (Section 3). Finally, we draw conclusions for our paper and we outline future directions for our research work (Section 4).

2 Implementation options

Due to the flexible data model of Wikidata and to the easy use of Wikimedia interfaces and tools, there are actually many options that programmatically allow data integration between Wikipedia Categories and Wikidata Statements. As Wikimedia Projects are hosted using MediaWiki software, Wikidata and language editions of Wikipedia have data dumps, especially ones in the XML format, that can be downloaded from https://dumps.wikimedia.org/. Direct download links for the dumps of each project can be easily found in a specific page of its website[1]. As these dumps are in structured formats, they can be easily processed using Python Libraries (e.g. Python XML[2]) to extract and then analyze Wikipedia Categories and Wikidata Statements for the development of interesting knowledge-based systems [18]. As Wikidata dumps can be voluminous and as users can need a unique type of information, WDumper (https://wdumps.toolforge.org/) can be used to create customized Wikidata RDF dumps that only include a needed subset of labels, entities and statements allowing to run the user's bot without having to perform useless tasks [19]. Similarly, Wikipedia dumps can be parsed with the help of several tools such as

[1] https://en.wikipedia.org/wiki/Wikipedia:Database_download (English Wikipedia), https://www.wikidata.org/wiki/Wikidata:Database_download (Wikidata).

[2] e.g., https://pypi.org/project/xml-python/

PetScan and Quarry [20-21]. As well, the MediaWiki API[3] provides a set of actions that can be used to retrieve specific information about Wikipedia Categories and Wikidata items and to adjust them when needed [22]. Such actions can be programmatically performed without any fear of exceeding query limits thanks to several Python Libraries that are designed to parse MediaWiki APIs like Pywikibot[4] and Wikibase Integrator[5]. Given that Wikidata statements are represented in RDF, Wikidata has its own SPARQL endpoint (WDQS, https://query.wikidata.org) for analyzing and extracting features from the Wikidata statements [12, 23]. It can not only allow to process Wikidata statements related to Wikipedia Categories but also to parse the Wikidata statements related to the direct members of Categories thanks to the possibility of a query federation between the Wikidata SPARQL endpoint and the MediaWiki API of a language edition of Wikipedia[6]. WDQS can be a valuable resource for developers as its results can be downloaded in various structured formats (e.g. TSV) and as its queries can be embedded to source codes in various programming languages such as Python. The information returned by the Wikidata Query Service can be augmented through the use of the "Wikidata SPARQL query equivalent" [P3921] statements as well as of the "category combines topics" [P971] and "category contains" [P4224] statements[7]. Further than these two options, Wikipedia and Wikidata pages can be explored using user scripts developed in JavaScript. This is made possible thanks to the Document Object Model of Wikimedia Pages on one hand and to a special set of MediaWiki configurations named mw.config[8]. Furthermore, data about Wikidata items can be automatically retrieved and embedded in Wikipedia Pages using Lua Scripts allowing the creation of pages where the user-generated information about an entity is shown at the same Wikipedia page as the Wikidata statements about it[9].

3 Applications

The combination of Wikidata statements with Wikipedia categories can be useful for various applications. First, the statements involving the Wikidata items corresponding to the members of a given Wikipedia category can be analyzed to infer a structured description of the category to be represented as "category combines topics" [P971] and "category contains" [P4224] statements. Such a description is multilingual and can allow users to understand the content of a category even if they lack consistent proficiency of the language of the assessed Wikipedia edition. As well, this description can provide a snapshot of the needed Wikidata statements to describe, validate or adjust the new inconsistent or incomplete Wikidata items of the category members. Second, the comparison of the outputs of a given Wikipedia Category to the

[3] e.g., https://www.wikidata.org/w/api.php.

[4] https://github.com/wikimedia/pywikibot

[5] https://github.com/LeMyst/WikibaseIntegrator

[6] e.g., https://w.wiki/3nyu, Credit: Houcemeddine Turki, Jan Ainali and Dipsacus fullonum

[7] Examples for the usage of P3921, P971, and P4224 properties can be found at https://www.wikidata.org/wiki/Q7439502.

[8] https://www.mediawiki.org/wiki/Manual:Interface/JavaScript.

[9] https://en.wikipedia.org/wiki/Module:Wikidata.

list of Wikidata items that can be involved in the category can be very efficient to identify the gaps and limitations in the representation of the topic of the category in Wikipedia or Wikidata. The lack of completeness of a given subset of Wikipedia Category Graph can be solved through the machine translation of the labels of the unsupported Wikidata items. Similarly, the lack of assignment of taxonomic relations (e.g., instance of) and non-taxonomic ones (e.g., main subject) to a category member can be fixed through the conversion of the Wikidata statements describing the given Wikipedia category into claims that can be assigned to the Wikidata item. Third, the analysis of Wikidata statements can be efficient to identify whether the subcategorization in the Wikipedia Category Graph is motivated by the fact that one category is a subclass (so-called hyponym) of the other one and consequently to eliminate non-transitive relations from the Wikipedia Category Graph for the generation of an “is a” taxonomy that can be used to drive semantic applications. So far, this process is achieved using complicated and time-consuming statistical and probabilistic approaches [2, 24]. However, this trimming becomes simple thanks to the semantic information provided by Wikidata.

Beyond the Wikimedia Projects, Wikidata statements coupled to Wikipedia categories can be useful to drive knowledge-based systems. The most common examples of such an application are semantic similarity measures driven by Wikipedia Category Graph that can be further enhanced through the use of Wikidata statements [2, 17-18]. Another example can be the use of Wikidata statements to analyze the full texts of category members in Wikipedia editions for modelling the topics of a category as well as for training knowledge graph embeddings and machine learning models to process natural language texts and for enabling knowledge graph enrichment and validation in a multilingual context.

4 Conclusion

In this position paper, we emphasized the usefulness of the combination of Wikipedia Categories with the Wikidata statements involving the categories and their direct members to achieve a better semantic representation for the Wikipedia Category Graph as well as for the Wikidata items. This association can drive multiple semantic applications bringing information retrieval, semantic web and natural language processing to the next stage. As a future direction for this work, we look forward to developing semantic applications driven by the Wikipedia-Wikidata combination.

5 Acknowledgements

This work is developed as a part of the Structured Categories project available at https://www.wikidata.org/wiki/Wikidata:Structured_Categories. We have to thank Anton V. Chupin (Institute of Continuous Media Mechanics of the Ural Branch of the Russian Academy of Sciences, Russian Federation) for providing useful comments and discussion about the topic of the research paper.

References

1. Frikha, M., Turki, H., Ben Ahmed Mhiri, M., Gargouri, F.: Trust Level Computation based on Time-aware Social Interactions for Recommending Medical Tourism Destinations. Journal of Information Assurance and Security 14(3), 86-97 (2019).
2. Ben Aouicha, M., Hadj Taieb, M. A., Ezzeddine, M.: Derivation of "is a" taxonomy from Wikipedia Category Graph. Engineering Applications of Artificial Intelligence 50, 265-286 (2016). doi:10.1016/j.engappai.2016.01.033.
3. Turki, H., Hadj Taieb, M. A., Shafee, T., Lubiana, T., Jemielniak, D., Ben Aouicha, M., Labra Gayo, J. E., Youngstrom, E. A., Banat, M., Das, D., Mietchen, D.: Representing COVID-19 information in collaborative knowledge graphs: the case of Wikidata. Semantic Web Journal (2021).
4. Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H., He, Q.: A survey on knowledge graph-based recommender systems. IEEE Transactions on Knowledge and Data Engineering (2020). doi:10.1109/TKDE.2020.3028705.
5. Hadj Taieb, M. A., Zesch, T., & Ben Aouicha, M.: A survey of semantic relatedness evaluation datasets and procedures. Artificial Intelligence Review 53(6), 4407-4448 (2020). doi:10.1007/s10462-019-09796-3.
6. Ben Aouicha, M., Hadj Taieb, M. A., Ben Hamadou, A.: LWCR: multi-Layered Wikipedia representation for Computing word Relatedness. Neurocomputing, 216(C), 816-843 (2016). doi:10.1016/j.neucom.2016.08.045.
7. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web, 6(2), 167-195 (2015). doi:10.3233/SW-140134.
8. Zesch, T., Müller, C., Gurevych, I.: Using Wiktionary for Computing Semantic Relatedness. In: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, vol. 8, pp. 861-866. AAAI, Chicago (2008).
9. Navarro, E., Sajous, F., Gaume, B., Prévot, L., ShuKai, H., Tzu-Yi, K., Magistry, P., Chu-Ren, H.: Wiktionary and NLP: improving synonymy networks. In: Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, pp. 19-27. Association for Computational Linguistics, Singapore (2009). doi:10.5555/1699765.1699768.
10. Nguyen, K. H., Ock, C. Y.: Using wiktionary to improve lexical disambiguation in multiple languages. In: International Conference on Intelligent Text Processing and Computational Linguistics, pp. 238-248. Springer, Berlin, Heidelberg (2012). doi:10.1007/978-3-642-28604-9_20.
11. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57(10), 78-85 (2014). doi:10.1145/2629489.
12. Turki, H., Shafee, T., Hadj Taieb, M. A., Ben Aouicha, M., Vrandečić, D., Das, D., & Hamdi, H.: Wikidata: A large-scale collaborative ontological medical database. Journal of Biomedical Informatics 99, 103292 (2019). doi:10.1016/j.jbi.2019.103292.
13. Waagmeester, A., Willighagen, E. L., Su, A. I., Kutmon, M., Gayo, J. E. L., Fernández-Álvarez, D., Groom, Q., Schaap, P. J., Verhagen, L. M., Koehorst, J. J.: A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses. BMC biology 19(1), 1-14 (2021). doi:10.1186/s12915-020-00940-y.
14. Zesch, T., Müller, C., Gurevych, I.: (2008, May). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), pp. 1646-1652. Association for Computational Linguistics, Marrakech, Morocco (2008).

15. Müller, C., Gurevych, I.: Using wikipedia and wiktionary in domain-specific information retrieval. In: Workshop of the Cross-Language Evaluation Forum for European Languages, pp. 219-226. Springer, Berlin, Heidelberg (2008).
16. Miller, T., Gurevych, I.: WordNet—Wikipedia—Wiktionary: Construction of a Three-way Alignment. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 2094-2100. Association for Computational Linguistics, Reykjavik, Iceland (2014).
17. Ben Aouicha, M., Hadj Taieb, M. A., Ben Hamadou, A.: Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness. Applied Intelligence 45(2), 475-511 (2016). doi:10.1007/s10489-015-0755-x.
18. Hadj Taieb, M. A., Ben Aouicha, M., Tmar, M., Ben Hamadou, A.: Wikipedia category graph and new intrinsic information content metric for word semantic relatedness measuring. In: International Conference on Data and Knowledge Engineering, pp. 128-140. Springer, Berlin, Heidelberg (2012). doi:10.1007/978-3-642-34679-8_13.
19. Hosseini Beghaeiraveri, S. A., Gray, A. J., McNeill, F.: Experiences of Using WDumper to Create Topical Subsets from Wikidata. In: Second International Workshop On Knowledge Graph Construction @ ESWC 2021, p. 13. CEUR-WS.org, Online (2021).
20. Yoshioka, M.: WC3: Analyzing the style of metadata annotation among Wikipedia articles by using Wikipedia category and the DBpedia metadata database. In International Semantic Web Conference, pp. 119-136. Springer, Cham (2016). doi:10.1007/978-3-319-68723-0_10.
21. Saez-Trumper, D., Redi, M.: Wikimedia Public (Research) Resources. In: WWW '20: Companion Proceedings of the Web Conference 2020, pp. 311-312. ACM, Taipei (2020). doi:10.1145/3366424.3383114.
22. Renzel, D., Schlebusch, P., Klamma, R.: Today's top "RESTful" services and why they are not RESTful. In: International Conference on Web Information Systems Engineering, pp. 354-367. Springer, Berlin, Heidelberg (2012). doi:10.1007/978-3-642-35063-4_26.
23. Malyshev, S., Krötzsch, M., González, L., Gonsior, J., Bielefeldt, A.: Getting the most out of wikidata: Semantic technology usage in wikipedia's knowledge graph. In: International Semantic Web Conference, pp. 376-394. Springer, Cham (2018). doi:10.1007/978-3-030-00668-6_23.
24. Boldi, P., Monti, C.: Cleansing wikipedia categories using centrality. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 969-974. ACM, Montréal (2016). doi:10.1145/2872518.2891111.