Wiktionary Matcher

Jan Portisch[1,2][0000−0001−5420−0663], Michael Hladik[2][0000−0002−2204−3138], and Heiko Paulheim[1][0000−0003−4386−8195]

[1] Data and Web Science Group, University of Mannheim, Germany
{jan, heiko}@informatik.uni-mannheim.de
[2] SAP SE Product Engineering Financial Services, Walldorf, Germany
{jan.portisch, michael.hladik}@sap.com

Abstract. In this paper, we introduce *Wiktionary Matcher*, an ontology matching tool that exploits *Wiktionary* as external background knowledge source. *Wiktionary* is a large lexical knowledge resource that is collaboratively built online. Multiple current language versions of *Wiktionary* are merged and used for monolingual ontology matching by exploiting synonymy relations and for multilingual matching by exploiting the translations given in the resource.
We show that *Wiktionary* can be used as external background knowledge source for the task of ontology matching with reasonable matching and runtime performance.[3]

Keywords: Ontology Matching · Ontology Alignment · External Resources · Background Knowledge · Wiktionary

1 Presentation of the System

1.1 State, Purpose, General Statement

The *Wiktionary Matcher* is an element-level, label-based matcher which uses an online lexical resource, namely *Wiktionary*. The latter is "[a] collaborative project run by the Wikimedia Foundation to produce a free and complete dictionary in every language"[4]. The dictionary is organized similarly to Wikipedia: Everybody can contribute to the project and the content is reviewed in a community process. Compared to WordNet [4], *Wiktionary* is significantly larger and also available in other languages than English. This matcher uses *DBnary* [15], an RDF version of *Wiktionary* that is publicly available[5]. The *DBnary* data set makes use of an extended *LEMON* model [11] to describe the data. For this matcher, *DBnary* data sets for 8 *Wiktionary* languages[6] have been downloaded

[4] see https://web.archive.org/web/20190806080601/https://en.wiktionary.org/wiki/Wiktionary

[5] see http://kaiko.getalp.org/about-dbnary/download/

[6] Namely: Dutch, English, French, Italian, German, Portugese, Russian, and Spanish.

and merged into one RDF graph. Triples not required for the matching algorithm, such as glosses, were removed in order to increase the performance of the matcher and to lower its memory requirements. As *Wiktionary* contains translations, this matcher can work on monolingual and multilingual matching tasks. The matcher has been implemented and packaged using the MELT framework[7], a Java framework for matcher development, tuning, evaluation, and packaging [7].

1.2 Specific Techniques Used

Monolingual Matching For monolingual ontologies, the matching system first links labels to concepts in *Wiktionary*, and then checks whether the concepts are synonymous in the external data set. This approach is conceptually similar to an upper ontology matching approach. Concerning the usage of a collaboratively built knowledge source, the approach is similar to *WikiMatch* [5] which exploits the *Wikipedia* search engine.

Wiktionary Matcher adds a correspondence to the final alignment purely based on the synonymy relation independently of the actual word sense. This is done in order to avoid word sense disambiguation on the ontology side but also on *Wiktionary* side: Versions for some countries do not annotate synonyms and translations for senses but rather on the level of the lemma. Hence, many synonyms are given independently of the word sense. In such cases, word-sense-disambiguation would have to be performed also on *Wiktionary* [13].

The linking process is similar to the one presented for the *ALOD2Vec* matching system [14]: In a first step, the full label is looked up on the knowledge source. If the label cannot be found, labels consisting of multiple word tokens are truncated from the right and the process is repeated to check for sub-concepts. This allows to detect long sub-concepts even if the full string cannot be found. Label *conference banquet* of concept *http://ekaw#Conference_Banquet* from the *Conference* track, for example, cannot be linked to the background data set using the full label. However, by applying right-to-left truncation, the label can be linked to two concepts, namely *conference* and *banquet*, and in the following also be matched to the correct concept *http://edas#ConferenceDinner* which is linked in the same fashion.

For multi-linked concepts (such as *conference dinner*), a match is only annotated if every linked component of the label is synonymous to a component in the other label. Therefore, *lens (http://mouse.owl#MA_0000275)* is not mapped to *crystalline_lens (http://human.owl#NCI_C12743)* due to a missing synonymous partner for *crystalline* whereas *urinary bladder neck (http://mouse.owl#MA_0002491)* is matched to *bladder neck (http://human.owl#NCI_C12336)* because *urinary bladder* is synonymous to *bladder*.

Multilingual Matching The multilingual capabilities of the matcher presented in this paper are similar to the work of Lin and Krizhanovsky [10] who use

[7] see `https://github.com/dwslab/melt`

data of the English *Wiktionary* (as of 2010) to allow for multilangual matching of the *COMS* matching system [9]. Unfortunately, the matching system never participated in the OAEI *MultiFarm* track. The work presented here is different in that it uses multiple language versions of *Wiktionary*, the corpora are much larger because they are newer, and in terms of the matching strategy that is applied.

The matcher first determines the language distributions in the ontologies. If the ontologies appear to be in different languages, *Wiktionary* translations are exploited: A match is created, if one label can be translated to the other one according to at least one *Wiktionary* language version – such as the Spanish label *ciudad* and the French label *ville* (both meaning *city*). This process is depicted in figure 1: The Spanish label is linked to the entry in the Spanish *Wiktionary* and from the entry the translation is derived.

If there is no *Wiktionary* version for the languages to be matched or the approach described above yields very few results, it is checked whether the two labels appear as a translation for the same word. The Chinese label 决定 (juédìng), for instance, is matched to the Arabic label قرار (qrār) because both appear as a translation of the English word *decision* on *Wiktionary*. This (less precise) approach is particularly important for language pairs for which no *Wiktionary* data set is available to the matcher (such as Chinese and Arabic). The process is depicted in figure 2: The Arabic and Chinese labels cannot be linked to *Wiktionary* entries but, instead, appear as translation for the same concept.

Instance Matching The matcher presented in this paper can be also used for combined schema and instance matching tasks. If instances are available in the given data sets, the matcher applies a two step strategy: After aligning the schemas, instances are matched using a string index. If there are many instances, *Wiktionary* is not used for the instance matching task in order to increase the matching runtime performance. Moreover, the coverage of schema level concepts in Wiktionary is much higher than for instance level concepts: For example, there is a sophisticated representation of the concept *movie*[8], but hardly any individual movies in Wiktionary.

For correspondences where the instances belong to classes that were matched before, a higher confidence is assigned. If one instance matches multiple other instances, the correspondence is preferred where both their classes were matched before.

Explainability Unlike many other ontology matchers, this matcher uses the extension capabilities of the alignment format [2] in order to provide a human readable explanation of why a correspondence was added to the final alignment. To explain the correspondence involving *(http://cmt_de#c-7914897-1988765, http://conference_en#c-0918067-8070827)*, for instance, the matcher gives the explanation *"The label of entity 1 was found in Wiktionary as 'Konferenz' and translated to 'conference' which equals the normalized label of entity 2."* Such

[8] see `https://en.wiktionary.org/wiki/movie`

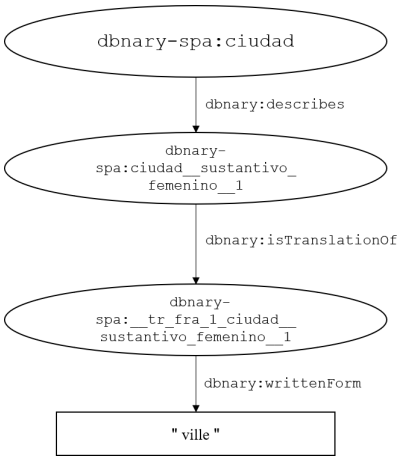

Fig. 1. Translation via the *Wiktionary* headword (using the *DBnary* RDF graph). Here: One (of more) French translations for the Spanish word *ciudad* in the Spanish *Wiktionary*.

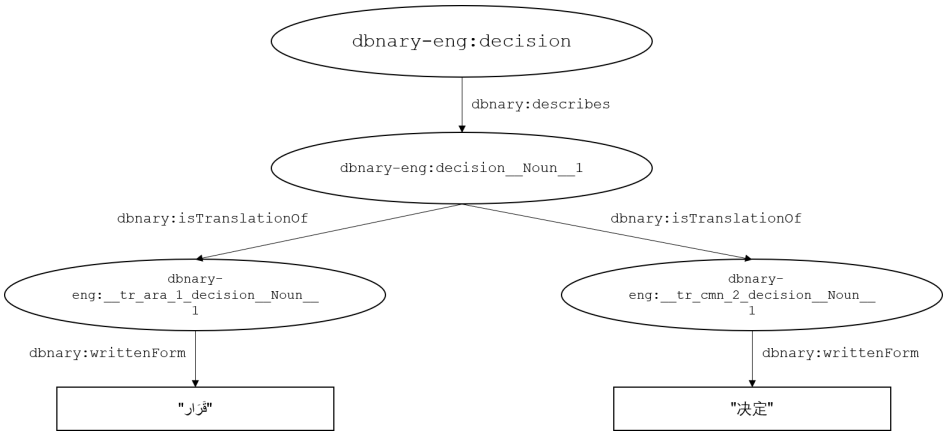

Fig. 2. Translation via the written forms of *Wiktionary* entries (using the *DBnary* RDF graph). Here: An Arabic and a Chinese label appear as translation for the same *Wiktionary* entry (*decision* in the English Wiktionary).

explanations can help to interpret and to trust a matching system's decision. Similarly, explanations also allow to comprehend why a correspondence was falsely added to the final alignment: The explanation for the false positive match *(http://confOf#Contribution, http://iasted#Tax)*, for instance, is given as follows: *"The first concept was mapped to dictionary entry [contribution] and the second concept was mapped to dictionary entry [tax]. According to Wiktionary, those two concepts are synonymous."* Here, it can be seen that the matcher was successful in linking the labels to *Wiktionary* but failed due to the missing word sense disambiguation. In order to explain a correspondence, the `description` property[9] of the *Dublin Core Metadata Initiative* is used.

2 Results

2.1 Anatomy

On the *Anatomy* track [3,1] the matching system achieves a median rank given F_1 scores and significantly outperforms the baseline. The system is capable of finding non-trivial matches such as *temporalis (http://mouse.owl#MA_0002390)* and *temporal muscle (http://human.owl#NCI_C33743)*.

2.2 Conference

The matching system consistently ranks 4th on all reference alignments given F_1 scores in the *Conference* track [16]. Like most matchers, the system achieves better results matching classes compared to matching properties. False positives are in most cases due to string matches and only in some cases due to synonymous relationships such as in (*http://edas#Topic*, *http://iasted#Item*).

2.3 Multifarm

The multilingual approach of the *Wiktionary Matcher* is different from most multilingual ontology matching approaches that use a translation API: Instead of an external function call, multiple multilingual resources are merged and used. Out of the matchers that participated in the *MultiFarm* track [12], *Wiktionary Matcher* performs third with an averaged F_1 score of 0.31 on (i) different ontologies and an averaged F_1 score of 0.12 on (ii) the same but translated ontologies. For the latter task the matching system lacks the ability to recognize that the structure of the ontologies that are to be matched is equal which would be an advantage for this matching problem. As expected, *Wiktionary Matcher* works better for languages for which a data set is available – such as English and French. Compared to other matching systems, the results of this matcher fluctuate more due to missing translation resources for some languages: While the matcher performs competitively for tasks involving the English language, the performance drastically falls when it comes to matching an ontology in the Arabic language.

[9] see `http://purl.org/dc/terms/description`

2.4 Knowledge Graph Track

On the *Knowledge Graph (KG) Track* [8,6], the matcher achieves the second-best result of all submitted matchers on the averaged F_1 scores. Compared to the best matching system, *FCAMap-KG*, the system presented in this paper requires less than a third of the runtime.

The matcher performs better in terms of F_1 on classes and properties compared to instances. This might be due to the fact that the matcher is optimized to match schemas and that the *Wiktionary* background source is only used for the schema matching task.

3 Discussions on the Way to Improve the Proposed System

The current version of *DBnary* does not extract *alternative forms* of words such as *(color, colour)*. This is a limitation by the data set used for this matcher and not by *Wiktionary*. An addition of this relation between lemmas to the data set would likely improve results.

Furthermore, the matching system presented here only uses synonymy and translation relations even though more information is available in the background knowledge source. An extension to other relations that exist between words would help to increase the performance. The false negative match between *intestine secretion* and *intestinal secretion* of classes *http://mouse.owl#MA_0002515* and *http://human.owl#NCI_C32875*, respectively, could be found if the system would exploit the fact that *intestinal* is derived from *intestine* (an information that is available in the data set).

The runtime performance could be improved by loading the background knowledge data (or aggregates) in specialized data structures that allow for a faster data access at runtime, such as key-value stores (rather than querying an RDF graph). This approach could particularly improve the performance on the *MultiFarm* track which has a comparatively slow runtime performance due to complex SPARQL queries.

4 Conclusions

In this paper, we presented the *Wiktionary Matcher*, a matcher utilizing a collaboratively built lexical resource. Given *Wiktionary*'s continuous growth, it can be expected that the matching results will improve over time – for example when additional translations are added. In addition, improvements to the *DBnary* data set, such as the addition of alternative word forms, may also improve the overall matcher performance.

References

1. Bodenreider, O., Hayamizu, T.F., Ringwald, M., de Coronado, S., Zhang, S.: Of mice and men: Aligning mouse and human anatomies.

In: AMIA 2005, American Medical Informatics Association Annual Symposium, Washington, DC, USA, October 22-26, 2005. AMIA (2005), http://knowledge.amia.org/amia-55142-a2005a-1.613296/t-001-1.616182/f-001-1.616183/a-012-1.616655/a-013-1.616652

2. David, J., Euzenat, J., Scharffe, F., dos Santos, C.T.: The alignment API 4.0. Semantic Web **2**(1), 3–10 (2011). https://doi.org/10.3233/SW-2011-0028, https://doi.org/10.3233/SW-2011-0028
3. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., dos Santos, C.T.: Ontology alignment evaluation initiative: Six years of experience. J. Data Semantics **15**, 158–192 (2011). https://doi.org/10.1007/978-3-642-22630-4_6, https://doi.org/10.1007/978-3-642-22630-4_6
4. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Language, Speech, and Communication, MIT Press, Cambridge, Massachusetts (1998)
5. Hertling, S., Paulheim, H.: WikiMatch - Using Wikipedia for Ontology Matching. In: Shvaiko, P., Euzenat, J., Kementsietsidis, A., Mao, M., Noy, N., Stuckenschmidt, H. (eds.) OM-2012: Proceedings of the ISWC Workshop. vol. 946, pp. 37–48 (2012)
6. Hertling, S., Paulheim, H.: Dbkwik: A consolidated knowledge graph from thousands of wikis. In: Wu, X., Ong, Y., Aggarwal, C.C., Chen, H. (eds.) 2018 IEEE International Conference on Big Knowledge, ICBK 2018, Singapore, November 17-18, 2018. pp. 17–24. IEEE Computer Society (2018). https://doi.org/10.1109/ICBK.2018.00011, https://doi.org/10.1109/ICBK.2018.00011
7. Hertling, S., Portisch, J., Paulheim, H.: MELT - Matching EvaLuation Toolkit. In: Semantics 2019 SEM2019 Proceedings. Karlsruhe (2019, to appear)
8. Hofmann, A., Perchani, S., Portisch, J., Hertling, S., Paulheim, H.: Dbkwik: Towards knowledge graph creation from thousands of wikis. In: Nikitina, N., Song, D., Fokoue, A., Haase, P. (eds.) Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 23rd - to - 25th, 2017. CEUR Workshop Proceedings, vol. 1963. CEUR-WS.org (2017), http://ceur-ws.org/Vol-1963/paper540.pdf
9. Lin, F., Butters, J., Sandkuhl, K., Ciravegna, F.: Context-based ontology matching: Concept and application cases. In: 10th IEEE International Conference on Computer and Information Technology, CIT 2010, Bradford, West Yorkshire, UK, June 29-July 1, 2010. pp. 1292–1298. IEEE Computer Society (2010). https://doi.org/10.1109/CIT.2010.233, https://doi.org/10.1109/CIT.2010.233
10. Lin, F., Krizhanovsky, A.: Multilingual ontology matching based on wiktionary data accessible via SPARQL endpoint. CoRR **abs/1109.0732** (2011), http://arxiv.org/abs/1109.0732
11. McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T.: Interchanging Lexical Resources on the Semantic Web. Language Resources and Evaluation **46**(4), 701–719 (Dec 2012). https://doi.org/10.1007/s10579-012-9182-3, http://link.springer.com/10.1007/s10579-012-9182-3
12. Meilicke, C., Garcia-Castro, R., Freitas, F., van Hage, W.R., Montiel-Ponsoda, E., de Azevedo, R.R., Stuckenschmidt, H., Sváb-Zamazal, O., Svátek, V., Tamilin, A., dos Santos, C.T., Wang, S.: Multifarm: A benchmark for multilingual ontology matching. J. Web Semant. **15**, 62–68 (2012). https://doi.org/10.1016/j.websem.2012.04.001, https://doi.org/10.1016/j.websem.2012.04.001

13. Meyer, C.M., Gurevych, I.: Worth its weight in gold or yet another resource - A comparative study of wiktionary, openthesaurus and germanet. In: Gelbukh, A.F. (ed.) Computational Linguistics and Intelligent Text Processing, 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010. Proceedings. Lecture Notes in Computer Science, vol. 6008, pp. 38–49. Springer (2010). https://doi.org/10.1007/978-3-642-12116-6_4, https://doi.org/10.1007/978-3-642-12116-6_4
14. Portisch, J., Paulheim, H.: Alod2vec matcher. In: OM@ISWC. CEUR Workshop Proceedings, vol. 2288, pp. 132–137. CEUR-WS.org (2018)
15. Sérasset, G.: Dbnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. Semantic Web **6**(4), 355–361 (2015). https://doi.org/10.3233/SW-140147, https://doi.org/10.3233/SW-140147
16. Zamazal, O., Svátek, V.: The ten-year ontofarm and its fertilization within the onto-sphere. J. Web Semant. **43**, 46–53 (2017). https://doi.org/10.1016/j.websem.2017.01.001, https://doi.org/10.1016/j.websem.2017.01.001