Wiktionary Matcher Results for OAEI 2021

Jan Portisch[1,2][0000−0001−5420−0663] 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 Business Technology Platform - One Domain Model, Walldorf, Germany
jan.portisch@sap.com

Abstract. This paper presents the results of the *Wiktionary Matcher* in the *Ontology Alignment Evaluation Initiative* (OAEI) 2021. *Wiktionary Matcher* is 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. This is the third OAEI participation of the matching system.[3]

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

CEUR-WS.org/Vol-3063/oaei21_paper14.pdf

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 [3], Wiktionary is significantly larger and also available in other languages than English. This matcher uses *DBnary* [13], an RDF version of Wiktionary that is publicly available[5]. The DBnary dataset makes use of an extended *LEMON* model [6] to describe the data. For this matcher, recent DBnary datasets for 8 Wiktionary languages[6] have been downloaded and merged into one RDF graph. Triples not required for the matching

[3] Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[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.

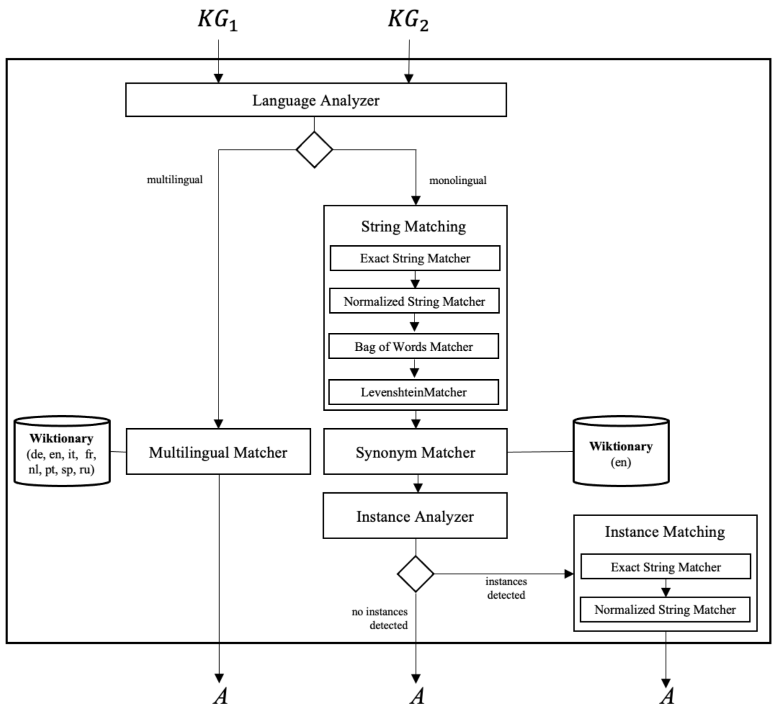

Fig. 1. High-level overview of the *Wiktionary Matcher*. KG_1 and KG_2 represent the input ontologies and optionally instances. The final alignment is referred to as A.

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.

This is the third OAEI participation of this matching system, *Wiktionary Matcher* participated in the OAEI in 2019 [9] and in the OAEI 2020 [11]. The matcher has been implemented and packaged using the *Matching EvaLuation Toolkit (MELT)*[7], a Java framework for matcher development, tuning, evaluation, and packaging [5,8].

1.2 Specific Techniques Used

This matching system system was initially introduced at the OAEI 2019 [9]. An overview of the matching system is provided in Figure 1. The main techniques used for matching are summarized below.

Monolingual Matching For monolingual ontologies, the matching system first applies multiple string matching techniques. Afterwards, the synonym matcher

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

module links labels to concepts in Wiktionary and checks then whether the concepts are synonymous in the external dataset. 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* [4] 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 [7]. Linking labels of entities to Wiktionary is carried out as follows: The full label is looked up in 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 dataset 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 For every matching task, the system first determines the language distributions in the ontologies. If the ontologies appear to be in different languages, the system automatically enables the multilingual matching module: Here, 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 2: 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 dataset is available to the matcher (such as Chinese and Arabic). The process is depicted in Figure 3: The Arabic and Chinese labels cannot be linked to Wiktionary entries but, instead, appear as translation for the same concept.

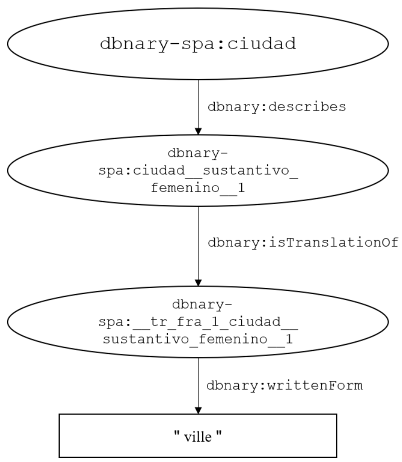

Fig. 2. Translation via the Wiktionary headword (using the DBnary RDF graph). Here: One (of more) French translations for the Spanish word *ciudad* in the Spanish Wiktionary.

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 datasets, the matcher applies a two step strategy: After aligning the schemas, instances are matched using a string index. As there are typically 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 [1] in order to provide a human

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

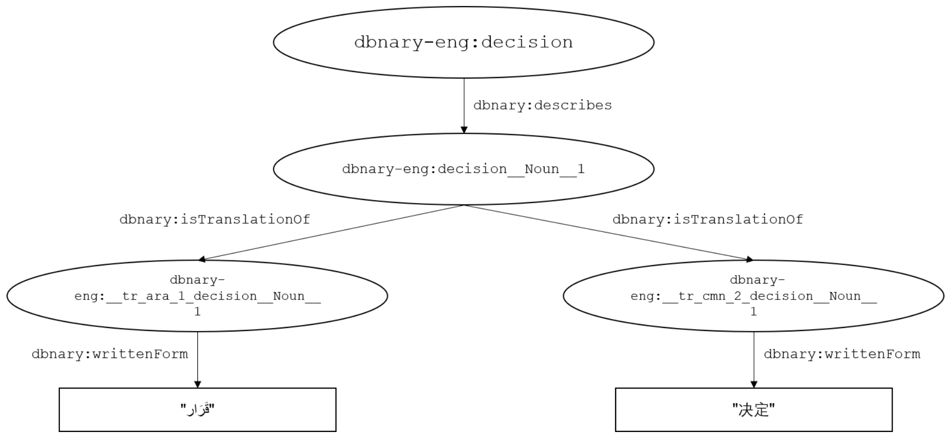

Fig. 3. Translation via the written forms of Wiktionary entries (using the DBnary RDF graph). Here: An Arabic and a Chinese label appear as translation for the same Wiktionary entry (*decision* in the English Wiktionary).

readable explanation of why a correspondence was added to the final alignment. Such 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.

1.3 Extensions to the Matching System for the 2021 Campaign

For the 2021 campaign, the background knowledge has been updated: The system uses DBnary dumps as of late July 2021. The Wiktionary knowledge source grew significantly compared to the version used in the OAEI 2020: In total, 7 million new triples were added; an increase of roughly 9%.

Besides upgrading the background knowledge, the underlying architecture was also improved: Rather than using a custom Wiktionary component, the 2021 version of the matching system was adapted to use the background knowledge modules that were made available with the release of MELT 3.0 [10]. With these changes, the code base is cleaner and better modularized. Improvements to the Wiktionary module will now benefit all MELT users. It is important to emphasize that these architectural improvements do not change the matching algorithm

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

compared to the 2020 version. The system was, furthermore, adapted to be packaged as MELT Web Docker[10] container. The implementation is publicly available on GitHub.[11]

2 Results

2.1 Anatomy Track

On the anatomy track, recall and F_1 could be slightly improved compared to the 2020 version of the matcher. The system performs at the median of all 2021 systems with an F_1 score of 0.843 (precision = 0.956, recall = 0.753).

2.2 Conference Track

The matching system achieves almost the same results as in 2021 on the conference track with a slightly improved recall. With an F_1 score of 0.59 on `rar2-M3`, the system performs above the median in terms of F_1.

2.3 Multifarm Track

The largest overall improvements compared to last year could be observed on the Multifarm track: Here, the F_1 score could be improved through a higher recall (the precision fell slightly). Like in the 2020 campaign, *Wiktionary Matcher* was the system with the overall highest precision and scored third place behind AML and LogMap.

2.4 LargeBio Track

As of writing this article, the results of the LargeBio track were not yet published.

2.5 Knowledge Graph Track

As in 2020, *Wiktionary Matcher* is the best matching system on the knowledge graph track.[12] The performance numbers did not change compared to the 2020 version of the matcher.

2.6 Common Knowledge Graph Track

This year, a new track was added to the OAEI: The *Common Knowledge Graph Track* [2]. Although not optimized for this track, *Wiktionary Matcher* achieved the second best result in terms of F_1 with a score of 0.89.

[10] see `https://dwslab.github.io/melt/matcher-packaging/web`

[11] see `https://github.com/janothan/WiktionaryMatcher`

[12] *2021* [12] achieves the same F_1 score – however, as the performance of the latter matcher on classes and properties is slightly worse, *Wiktionary Matcher* comes in first.

3 General Comments

It is important to note that the matching system currently exploits only a small share of semantic relations available on Wiktionary. The system is restricted by the available relations extracted by the DBnary project. The additional exploitation of the relations *alternative forms* or *derived terms*, for instance, would likely improve the system. However, those are not yet extracted and are consequently not used for the matching task as of today. The improvements observed on Anatomy and Conference are completely due to the updated Wiktionary version since the core matching code was left unchanged.

4 Conclusion

In this paper, we presented the *Wiktionary Matcher*, a matcher utilizing a collaboratively built lexical resource, as well as the results of the system in the 2021 OAEI campaign. Given Wiktionary's continuous growth, it can be expected that the matching results will continue to improve over time – for example when additional synonyms and translations are added. In addition, improvements to the DBnary dataset, such as the addition of alternative word forms, may also improve the overall matcher performance in the future.

References

1. 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`
2. Fallatah, O., Zhang, Z., Hopfgartner, F.: A gold standard dataset for large knowledge graphs matching. In: Shvaiko, P., Euzenat, J., Jiménez-Ruiz, E., Hassanzadeh, O., Trojahn, C. (eds.) Proceedings of the 15th International Workshop on Ontology Matching co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 2, 2020. CEUR Workshop Proceedings, vol. 2788, pp. 24–35. CEUR-WS.org (2020), `http://ceur-ws.org/Vol-2788/om2020_LTpaper3.pdf`
3. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Language, Speech, and Communication, MIT Press, Cambridge, Massachusetts (1998)
4. 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)
5. Hertling, S., Portisch, J., Paulheim, H.: MELT - matching evaluation toolkit. In: Acosta, M., Cudré-Mauroux, P., Maleshkova, M., Pellegrini, T., Sack, H., Sure-Vetter, Y. (eds.) Semantic Systems. The Power of AI and Knowledge Graphs - 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9-12, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11702, pp. 231–245. Springer (2019). https://doi.org/10.1007/978-3-030-33220-4_17, `https://doi.org/10.1007/978-3-030-33220-4_17`

6. 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
7. 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
8. Portisch, J., Hertling, S., Paulheim, H.: Visual analysis of ontology matching results with the MELT dashboard. In: The Semantic Web: ESWC 2020 Satellite Events (2020)
9. Portisch, J., Hladik, M., Paulheim, H.: Wiktionary matcher. In: Shvaiko, P., Euzenat, J., Jiménez-Ruiz, E., Hassanzadeh, O., Trojahn, C. (eds.) Proceedings of the 14th International Workshop on Ontology Matching co-located with the 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 26, 2019. CEUR Workshop Proceedings, vol. 2536, pp. 181–188. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2536/oaei19_paper15.pdf
10. Portisch, J., Hladik, M., Paulheim, H.: Background knowledge in schema matching: Strategy vs. data. In: Proceedings of the International Semantic Web Conference, ISWC 2021 (2021), to appear
11. Portisch, J., Paulheim, H.: Wiktionary matcher results for OAEI 2020. In: Shvaiko, P., Euzenat, J., Jiménez-Ruiz, E., Hassanzadeh, O., Trojahn, C. (eds.) Proceedings of the 15th International Workshop on Ontology Matching co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 2, 2020. CEUR Workshop Proceedings, vol. 2788, pp. 225–232. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2788/oaei20_paper14.pdf
12. Portisch, J., Paulheim, H.: ALOD2vec Matcher results for OAEI 2021. In: OM@ISWC 2021 (2021), to appear
13. 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