# Enhancing multilingual and biomedical named entity recognition using Wikidata semantic relations

Houcemeddine Turki
Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax
Sfax, Tunisia
turkiabdelwaheb@hotmail.fr

Dennis Priskorn
Department of Computer and System Science, Mid Sweden University
Östersund, Sweden
dennis@priskorn.se

Mohamed Ali Hadj Taieb
Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax
Sfax, Tunisia
mohamedali.hajtaieb@fss.usf.tn

Mohamed Ben Aouicha
Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax
Sfax, Tunisia
mohamed.benaouicha@fss.usf.tn

Alejandro Piad-Morffis
School of Math and Computer Science, University of Havana
La Habana 10200, Cuba
apiad@matcom.uh.cu

## Abstract

Named Entity Recognition (NER) is currently a key technique is knowledge engineering, particularly in the context of biomedical informatics. In this context, Wikidata has been used as a Knowledge Graph to drive named entity recognition for various applications such as news tracking and question answering. Current Wikidata NER Systems are mainly based on the labels, aliases and classes of Wikidata items. In this research paper, we propose a new approach to augment and validate the named entity recognition of a type of entities based on Wikidata semantic relations. We evaluate our *semantic relation-based multilingual named entity recognition algorithm* by applying it on a corpus of 8,705 titles of biomedical research publications about drugs extracted from the Wikidata knowledge graph.

***CCS Concepts:*** • **Applied computing** → **Annotation**; • **Information systems** → **Information extraction**; • **Computing methodologies** → **Information extraction**.

***Keywords:*** Named Entity Recognition, Semantic Relations, Wikidata



*Wiki Workshop '22, April 25, 2022, Lyon, France*



ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

**ACM Reference Format:**
Houcemeddine Turki, Dennis Priskorn, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, and Alejandro Piad-Morffis. 2022. Enhancing multilingual and biomedical named entity recognition using Wikidata semantic relations. In *Proceedings of WWW '22 Companion (Wiki Workshop '22).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

Biomedical Named Entity Recognition has evolved for several years as a main technique for knowledge engineering from clinical texts [10]. It allows to identify clinical concepts in sentences enabling the annotation of biomedical publications and electronic health records, the extraction and classification of semantic relations, and the application of data mining techniques on health information for clinical decision support [10]. The accuracy of the algorithms for named entity recognition is consequently important, particularly in a multilingual context, to automate clinical practice and biomedical research [10]. Currently, NER systems are mainly based on machine learning techniques to annotate named entity in biomedical texts although knowledge resources can bring an added value to this task [15].

Particularly, knowledge graphs are structured resources that represent semantic information about entities in the form of triples (*Subject – Predicate – Object*) [12, 13]. Such triples can be taxonomic relations (e.g., *instance of*, *subclass of*, and *part of*), non-taxonomic relations (e.g., *drug used for treatment* and *significant drug interaction*), and non-relational statements (e.g., *external identifier* and *number of cases*) [12, 13]. Given the machine-readable format of knowledge graphs, they can intuitively be reused to support a variety of applications, particularly named entity recognition [12, 13]. This is enabled thanks to programmatic tools such as SPARQL endpoints and APIs [12, 13]. For instance,

Wikidata (https://www.wikidata.org) is an open, collaborative and multilingual knowledge graph that is maintained by the Wikimedia Foundation and that cover various research topics including medicine [13, 14]. Wikidata items cover various aspects of biomedicine, ranging from anatomical structures to drugs and diseases [12, 13]. Each item is assigned a language-independent identifier (Q-number), a set of names (labels and aliases) and glosses (descriptions) in natural languages, and a number of semantic relations characterizing the facets of the Wikidata items [12, 13].

In this research paper, we propose to use the semantic relations describing a type of entities to enrich the named entity recognition of a single class of items (e.g., *drugs*) with the inline annotation of related concepts (e.g., *diseases*, *risk factors*, and *adverse effects*). We will use Wikidata as a reference knowledge graph for our approach. We will begin our work by a brief overview of the related works about biomedical named entity recognition, particularly from a multilingual perspective (Section 2). Then, we will outline our proposed approach for the named entity recognition of drugs based on Wikidata labels and aliases as well as for knowledge-driven named entity recognition augmentation and how we will evaluate it through its application of a brief corpus of the titles of scholarly publications extracted from Wikidata (Section 3). After that, we will provide statistical results of our proposed method, and we will discuss them through comparison with previous research findings and through a comparison with statistical data about the Wikidata knowledge graph (Section 4). Finally, we will conclude our study and provide future directions for our research efforts (Section 5).

## 2 Related Work

Nowadays, three main approaches are applied for biomedical named entity recognition: *knowledge-based*, *rule-based* and *machine learning-based* [15]. Machine learning-based approaches mainly train neural network-based algorithm and classical machine learning models such as Hidden Markov Model and Conditional Random Fields on an annotated corpus where mentions of a particular type of entities are annotated [15, 16]. An example of such a dataset is the NCBI disease corpus [2]. Rule-Based approaches use hard-coded heuristics and conditions to identify various types of entities [3, 15]. The advantage of this type of methods is the possibility of their customization in function of the class of the named entities to be recognized and the input data, particularly clinical texts [11]. However, the drawback of such approaches is the difficulty of the definition of the recognition process through the development of rules, mostly when the recognition complexity is very high [15].

As for knowledge-based approaches, they are also known as dictionary-based methods and are characterized by the use of knowledge resources to extract the list of synonyms and classes of medical concepts so that they can be reused to annotate clinical texts [15]. Although Wikidata as a large-scale knowledge graph includes other interesting semantic information such as semantic relations [12], NER methods such as *NECKAr* and *OpenTapioca* mainly use the names and taxonomic relations (i.e., *instance of*, *subclass of*, and *part of*) of Wikidata items to annotate multilingual texts [1, 4]. This is also applicable even for advanced NER-applications, e.g. those based on advanced machine learning techniques [6] and those developed for specific purposes, particularly question answering [5] and news tracking [8].

## 3 Proposed Approach

To begin, we generate an algorithm for the multilingual named entity recognition of a single class of entities (i.e., drugs) in our returned dataset based on Wikidata labels and aliases (Section 3.1). Then, we propose our method for the augmentation of the output of multilingual named entity recognition through the use of Wikidata relations having a drug as a subject (Section 3.2). Finally, we create a dataset of the titles of scholarly publications about drugs based on Wikidata items about biomedical research papers to evaluate our assumptions (Section 3.3). All the source code and generated outputs are made available on GitHub for reproducibility purposes[1].

### 3.1 Named Entity Recognition of Drugs

We extract the labels and aliases of the Wikidata items about drugs in multiple natural languages using Wikidata Query Service[2] which is a service maintained by Wikimedia Foundation. It is freely accessible to anyone, and the knowledge it generates is licensed under the CC0 License [7, 14]. The query enabling the creation of the multilingual dataset of drug names is shown in Listing 1. As of February 1, 2022, we obtained a dictionary of 100,361 drug names representing 3,398 items in 248 languages. The most represented language is *English* (24,971 drug names) followed by Welsh (6,420), Serbian (5,205), French (4,236), German (3,406), Russian (2,957), Arabic (2,842), Japanese (2,737), and Italian (2,602). The better availability of drug names in English, French, German, Russian, Arabic, Japanese and Italian goes in line with previous studies on the language coverage of medical entities in Wikidata, particularly in the context of the COVID-19 pandemic [12]. However, the coverage of drugs in Welsh and Serbian is surprisingly higher than the average coverage of Wikidata biomedical items in these two languages [12].

**Listing 1.** SPARQL query for extracting the labels and aliases of drugs from Wikidata

```
SELECT ?drug ?label (LANG(?label) AS ?lang) WHERE
{
  ?drug wdt:P31 wd:Q12140.
```

[1] https://github.com/SisonkeBiotik-Africa/Relational-NER
[2] https://www.mediawiki.org/wiki/Wikidata_query_service

```
  {?drug rdfs:label ?label} UNION
  {?drug skos:altLabel ?label}
}
```

After the creation of the dataset of drugs, we filter it to eliminate drug names that are composed of less than four characters, and we screen our corpus to identify the drugs in the titles of scholarly publications. This is done through the detection of language of every title using the *Langdetect* Python Library[3] and then through the search of the related noun phrases in the considered title. The considered titles are not pre-processed, as they are simple statements having a reduced linguistic complexity [9]. That is why there is no need for applying pre-trained models for noun phrase identification and singularization [15] to recognize drug names.

### 3.2 Relation-Based Named Entity Recognition

Our approach augments the named entity recognition of a single class through the annotation of entities from related classes. Our method uses *Langdetect* to identify the language of each considered title. Then, for every title of a biomedical scholarly publication, we find the Wikidata items that are used as objects to Wikidata relations where an annotated drug is the subject, and we extract their names in the language of the analyzed title thanks to a SPARQL query (Listing 2).

**Listing 2.** SPARQL query for finding the labels and aliases of the Wikidata items related to annotated drug

```
SELECT ?relatedentity ?label) WHERE
{
  wd:<DrugID> ?prop ?object.
  {?object rdfs:label ?label} UNION
  {?object skos:altLabel ?label}
  FILTER(LANG(?label)="<Language>")
}
```

Then, we search the extracted related entities in the titles and annotate them if found. For the purpose of our analysis, we extract the type of the relation allowing the named entity recognition (Predicate having a P-number as a language-independent identifier), the Wikidata ID of the annotated drug that is related to the recognized entity, and the Wikidata ID of the recognized item. After that, we add all the retrieved information to the output of our named entity recognition augmentation approach.

### 3.3 Corpus-based evaluation

The dataset of the titles of scholarly publications was created using *ItemSubjector*, a tool implemented in Python that extracts information from Wikidata[4]. The commands used to get the data are highlighted in the Listing 3.

**Listing 3.** Shell commands to create the corpus

```
$ python itemsubjector.py --sparql "SELECT ?item
WHERE {?item wdt:P31 wd:Q35456. MINUS {?item
wdt:P1889 [].}}" -p -c --limit 10000
$ python itemsubjector.py
--export-jobs-to-dataframe
```

The tool was designed to find scholarly articles in Wikidata that have a title that matches a given main subject (e.g. *paracetamol* [Q57055]) but do not yet have it stored as a statement. The extraction was done utilizing the Wikidata Query Service. After extraction, the list was filtered to eliminate duplicated items. Then, we applied our algorithm for the named entity recognition of drugs and subsequently the algorithm for the inline annotation of Wikidata items related to identified drugs. We then analyzed the outputs of the knowledge-driven named entity recognition augmentation using Microsoft Office Excel 2016[5].

## 4 Results and Discussion

As of February 1, 2022, we retrieved a corpus of 8,705 titles of biomedical publications from Wikidata using the *ItemSubjector* tool. The application of our method for the named entity recognition of drugs to the corpus allowed the identification of 37,911 drug mentions of 506 unique medication items. This is absolutely motivated by the existence of several drug mentions in a title of several research papers (an average of 4.4 drug mentions per title). However, this is also due to the drug mentions that are composed of two nouns or more. Such n-grams can correspond to chemical compounds (e.g., *potassium permanganate* [Q190865]) and their split into one-noun terms can result in the named entity recognition of the elements constituting the compound (e.g., *potassium* [Q703] and *permanganate* [Q56809357]). Furthermore, there are several drug items in Wikidata that are either duplicates or are having the same name (e.g., *potassium permanganate* [Q190865, Q28453504]). We found 129 drug items in Wikidata having the same name as another drug in the same database[6]. The output of the drug named entity annotation is limited and lacks precision when compared to other Wikidata-based methods for named entity recognition [1, 4, 6]. However, this limitation of the drug named entity recognition system can be useful to easily identify the effect of considering semantic relations in named entity recognition validation.

When we applied our semantic relation-based multilingual named entity recognition algorithm, we found that the method successfully identified 10,265 mentions of Wikidata items other than drugs. These mentions concern 261 entities. The mostly mentioned items in our corpus ranges from diseases to parent classes of entities and chemical elements as shown in Table 1. The variety of augmented annotations can be confirmed through the analysis of the relation types that

[3] https://pypi.org/project/langdetect/ version 1.0.9

[4] The commit used to get the data was https://github.com/dpriskorn/ItemSubjector/tree/6ed214691c0e6f64b9007ea0ec9fca2528f437c2.

[5] Microsoft Office Professional Plus 2016 Excel Version 2201 Build 14827.20158 was used.

[6] Live query: https://w.wiki/4nbJ

**Table 1.** Top ten mentioned drug-related Wikidata items in our corpus

| Wikidata Item | Wikidata ID | Mentions |
|---|---|---|
| *medication* | Q12140 | 1,833 |
| *nitrate* | Q49916468 | 999 |
| *arthritis* | Q170990 | 983 |
| *Homo sapiens* | Q15978631 | 762 |
| *rheumatoid arthritis* | Q187255 | 740 |
| *leukemia* | Q29496 | 472 |
| *acute myeloid leukemia* | Q264118 | 388 |
| *lymphoma* | Q208414 | 362 |
| *potassium* | Q703 | 350 |
| *rapeseed* | Q177932 | 330 |

**Table 2.** Top Wikidata relation types contributing drug-related item annotation

| Wikidata Item | Wikidata ID | Mentions |
|---|---|---|
| *medical condition treated* | P2175 | 4,419 |
| *instance of* | P31 | 2,070 |
| *found in taxon* | P703 | 1,276 |
| *subclass of* | P279 | 1,083 |
| *has part* | P527 | 585 |
| *route of administration* | P636 | 218 |
| *different from* | P1889 | 144 |
| *subject has role* | P2868 | 128 |

were used for the named entity recognition of items from other semantic classes. In fact, we found that the 10,265 mentions were recognized thanks to 24 distinct relation types led by *medical condition treated* [P2175, 4,419 times] as clearly revealed in Table 2. This proves the ability of using semantic relations to sustain multi-class named entity recognition algorithms through the identification of concepts that can be missed by the machine-learning models. In our corpus, we successfully annotated taxons, parent classes and routes of administration. By contrast, conventional NER does not significantly cover such classes of entities [15]. However, our method highly depends on how well the topic is covered in the reference knowledge graph (i.e., Wikidata). As an open and collaborative knowledge graph, Wikidata can represent several classes in a limited way when an interested community of editors is not available online [12, 13]. Furthermore, our method does not filter the considered relation types and can mistakenly consider unrelated concepts that are linked together in the reference knowledge graph using properties like *different from* [P1889]. Eliminating such insignificant relations can be easily done through a slight adjustment of the SPARQL query in Listing 2 and the use of the FILTER clause [7].

In another context, when we verified the drug annotations that resulted in the recognition of non-drug Wikidata items, we found out that only 8,270 out of 37,911 drug annotations (21.8%) contributed to the annotation of drug-related entities. These drug annotations only cover 258 out of 506 annotated drugs (51.0%). This can be due to the significant lack of representation of semantic knowledge about several drugs in Wikidata [13]. However, this can also be explained by the usefulness of using semantic relations to filter named entity recognition outputs where an item related to the annotated entity cannot be identified in the analyzed text.

## 5 Conclusion

In this research paper, we proposed an approach that augments the named entity annotation of a single class of entities through the annotation of related concepts in the considered text as revealed by a large-scale knowledge graph. We applied our approach to enhance the Wikidata-based named entity annotation of drugs in the titles of 8,705 titles of biomedical scholarly publications. Our approach has been proved as efficient not only for enriching the named entity recognition output through the annotation of non-drug items based on Wikidata semantic relations, but also for validating drug annotations and eliminating the ones that are not related to the context of the analyzed input. As a future direction of this research work, it will be useful to further investigate how knowledge graphs can be used to ameliorate the accuracy of named entity recognition systems and to support other tasks in the context of natural language processing and knowledge engineering.

## Acknowledgments

The contributions of Houcemeddine Turki, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha have been supported by the Federated Research Project PRFCOV19-D1-P1 supervised by the Tunisian Ministry of Higher Education and Scientific Research. The work of Alejandro Piad-Morffis is funded by the University of Alicante, the University of Havana, the Generalitat Valenciana (Conselleria d'Educació, Cultura i Esport), and the Spanish Government through the projects LIVING-LANG (RTI2018-094653-B-C22) and SIIA (PROMETEO/2018/089). This work is encouraged by Sisonkebiotik, an open and inclusive community of African researchers, practitioners and enthusiasts at the intersection of Machine Learning and Healthcare.

## References

[1] Antonin Delpeuch. 2020. OpenTapioca: Lightweight Entity Linking for Wikidata. In *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020 (CEUR Workshop Proceedings, Vol. 2773)*, Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, and Denny Vrandecic (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-2773/paper-02.pdf

[2] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics* 47 (2014), 1–10. https://doi.org/10.1016/j.jbi.2013.12.006

[3] Tome Eftimov, Barbara Koroušić Seljak, and Peter Korošec. 2017. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS One* 12, 6 (2017), e0179488. https://doi.org/10.1371/journal.pone.0179488

[4] Johanna Geiß, Andreas Spitz, and Michael Gertz. 2018. NECKAr: A Named Entity Classifier for Wikidata. In *Language Technologies for the Challenges of the Digital Age*, Georg Rehm and Thierry Declerck (Eds.). Springer International Publishing, Cham, 115–129. https://doi.org/10.1007/978-3-319-73706-5_10

[5] Vladislav Korablinov and Pavel Braslavski. 2020. RuBQ: A Russian Dataset for Question Answering over Wikidata. In *The Semantic Web – ISWC 2020*, Jeff Z. Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal (Eds.). Springer International Publishing, Cham, 97–110.

[6] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. *BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision.* Association for Computing Machinery, New York, NY, USA, 1054–1064. https://doi.org/10.1145/3394486.3403149

[7] Stanislav Malyshev, Markus Krötzsch, Larry González, Julius Gonsior, and Adrian Bielefeldt. 2018. Getting the Most Out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph. In *The Semantic Web – ISWC 2018*, Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl (Eds.). Springer International Publishing, Cham, 376–394. https://doi.org/10.1007/978-3-030-00668-6_23

[8] Charlotte Rudnik, Thibault Ehrhart, Olivier Ferret, Denis Teyssou, Raphael Troncy, and Xavier Tannier. 2019. Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1232–1239. https://doi.org/10.1145/3308560.3316761

[9] Viviana Soler. 2007. Writing titles in Science: An exploratory study. *English for Specific Purposes* 26, 1 (2007), 90–102. https://doi.org/10.1016/j.esp.2006.08.001

[10] Bosheng Song, Fen Li, Yuansheng Liu, and Xiangxiang Zeng. 2021. Deep learning methods for biomedical named entity recognition: A survey and qualitative comparison. *Briefings in Bioinformatics* 22, 6 (2021), bbab282. https://doi.org/10.1093/bib/bbab282

[11] Houcemeddine Turki, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha. 2021. Enhancing filter-based parenthetic abbreviation extraction methods. *Journal of the American Medical Informatics Association* 28, 3 (2021), 668–669. https://doi.org/10.1093/jamia/ocaa314

[12] Houcemeddine Turki, Mohamed Ali Hadj Taieb, Thomas Shafee, Tiago Lubiana, Dariusz Jemielniak, Mohamed Ben Aouicha, Jose Emilio Labra Gayo, Eric A. Youngstrom, Mus'ab Banat, Diptanshu Das, and et al. 2022. Representing COVID-19 information in collaborative knowledge graphs: The case of Wikidata. *Semantic Web* (2022), 1–32. https://doi.org/10.3233/sw-210444

[13] Houcemeddine Turki, Thomas Shafee, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Denny Vrandečić, Diptanshu Das, and Helmi Hamdi. 2019. Wikidata: A large-scale collaborative ontological medical database. *Journal of Biomedical Informatics* 99 (2019), 103292. https://doi.org/10.1016/j.jbi.2019.103292

[14] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. https://doi.org/10.1145/2629489

[15] Xu Wang, Chen Yang, and Renchu Guan. 2018. A comparative study for biomedical named entity recognition. *International Journal of Machine Learning and Cybernetics* 9, 3 (2018), 373–382. https://doi.org/10.1007/s13042-015-0426-6

[16] Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. 2019. CollaboNet: Collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinformatics* 20, S10 (2019), 249. https://doi.org/10.1186/s12859-019-2813-6