NECKAr: A Named Entity Classifier for Wikidata

Johanna Geiß, Andreas Spitz(✉), and Michael Gertz

Institute of Computer Science, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany
{geiss,spitz,gertz}@informatik.uni-heidelberg.de

Abstract. Many Information Extraction tasks such as Named Entity Recognition or Event Detection require background repositories that provide a classification of entities into the basic, predominantly used classes LOCATION, PERSON, and ORGANIZATION. Several available knowledge bases offer a very detailed and specific ontology of entities that can be used as a repository. However, due to the mechanisms behind their construction, they are relatively static and of limited use to IE approaches that require up-to-date information. In contrast, Wikidata is a community-edited knowledge base that is kept current by its userbase, but has a constantly evolving and less rigid ontology structure that does not correspond to these basic classes. In this paper we present the tool NECKAr, which assigns Wikidata entities to the three main classes of named entities, as well as the resulting Wikidata NE dataset that consists of over 8 million classified entities. Both are available at http://event.ifi.uni-heidelberg.de/?page_id=532.

1 Introduction

The classification of entities is an important task in information extraction (IE) from textual sources that requires the support of a comprehensive knowledge base. In a standard workflow, a Named Entity Recognition (NER) tool is used to discover the surface forms of entity mentions in some input text. These surface forms then have to be disambiguated and linked to a specific entity in a knowledge base (entity linking) to be useful in subsequent IE tasks. For the latter step of entity linking, suitable entity candidates have to be selected from the underlying knowledge base. In the general case of linking arbitrary entities, information about the classes of entity candidates is advantageous for the disambiguation process and for pruning candidates. In more specialized cases, only a subset of entity mentions may be of interest, such as toponyms or person mentions, which requires the classification of entity mentions. As a result, the classification of entities in the underlying knowledge base serves to support the linking procedure and directly translates into a classification of the entities that are mentioned in the text, which is a necessary precondition for many subsequent tasks such as event detection (Kumaran and Allan 2004) or document geolocation (Ding et al. 2000).

G. Rehm and T. Declerck (Eds.): GSCL 2017, LNAI 10713, pp. 115–129, 2018.
https://doi.org/10.1007/978-3-319-73706-5_10

There is a number of knowledge bases that provide such a background repository for entity classification, predominantly DBpedia, YAGO, and Wikidata (Färber et al. 2017). While these knowledge bases provide semantically rich and fine-granular classes and relationship types, the task of entity classification often requires associating coarse-grained classes with discovered surface forms of entities. This problem is best illustrated by an IE tasks that has recently gained significant interest in particular in the context of processing streams of news articles and postings in social media, namely event detection and tracking, e.g., (Aggarwal and Subbian 2012; Sakaki et al. 2010; Brants and Chen 2003). Considering an event as *something that happens at a given place and time between a group of actors* (Allan 2012), the entity classes PERSON, ORGANIZATION, LOCATION, and TIME, are of particular interest. While surface forms of temporal expressions are typically normalized by using a temporal tagger (Strötgen and Gertz 2016), dealing with the classification of the other types of entities often is much more subtle. This is especially true if one recalls that almost all available NER tools tag named entities only at a very coarse-grained level, e.g., Stanford NER (Finkel et al. 2005), which predominately uses the classes LOCATION, PERSON, and ORGANIZATION.

The objective of this paper is to provide the community with a dataset and API for entity classification in Wikidata, which is tailored towards entities of the classes LOCATION, PERSON, and ORGANIZATION. Like knowledge bases with inherently more coarse or hybrid class hierarchies such as YAGO and DBpedia, this version of Wikidata then supports entity linking tasks at state-of-the-art level (Geiß and Gertz 2016; Spitz et al. 2016b), but links entities to the continuously evolving Wikidata instead of traditional KBs. As we outline in the following, extracting respective sets of entities from a KB for each such class is by no means trivial (Spitz et al. 2016a), especially given the complexity of simultaneously dealing with multi-level class and instance structures inherent to existing KBs, an aspect that is also pointed out by Brasileiro et al. (2016). However, there are several reasons to chose Wikidata over other KBs. First, especially when dealing with news articles and social media data streams, it is crucial to have an up-to-date repository of persons and organizations. To the best of our knowledge, at the time of writing this paper, the most recent version of DBPedia was published in April 2016, and the latest evaluated version of YAGO in September 2015, whereas Wikidata provides a weekly data dump. Even though all three KBs (Wikidata, DBpedia, and YAGO3) are based on Wikipedia, Wikidata also contains information about entities and relationships that have not been simply extracted from Wikipedia (YAGO and DBpedia extract data predominantly from infoboxes) but collaboratively added by users (Müller-Birn et al. 2015). Although the latter feature might raise concerns regarding the quality of the data in Wikidata, for example due to vandalism (Heindorf et al. 2015), we find that the currentness of information far outweights these concerns when using Wikidata as basis for a named entity classifying framework and as a knowledge base in particular. While Wikidata provides a SPARQL interface for direct query access in addition to the weekly dumps, this method of accessing the data

has several downsides. First, the interface is not designed for speed and is thus ill suited for entity extraction or linking tasks in large corpora, where many lookups are necessary. Second, and more importantly, the continually evolving content of Wikidata prevents reproducibility of scientific results if the online SPARQL access is used, as the versioning is unclear and it is impossible to recreate experimental conditions. Third, we find that the hierarchy and structure in Wikidata is (necessarily) complicated and does not lend itself easily to creating coarse class hierarchies on the fly without substantial prior investigation into the existing hierarchies. Here, NECKAr provides a stable, easy to use view of classified Wikidata entities that is based on a selected Wikidata dump and allows reproducible results of subsequent IE tasks.

In summary, we make the following contributions: We provide an easy to use tool for assigning Wikidata items to commonly used NE classes by exclusively utilizing Wikidata. Furthermore, we make one such resulting Wikidata NE dataset available as a resource, including basic statistics and a thorough comparison to YAGO3.

The remainder of the paper is structured as follows. After a brief discussion of related work in the following section, we describe our named entity classifier in detail in Sect. 3 and present the resulting Wikidata NE dataset in Sect. 4. Section 5 gives a comparison of our results to YAGO3.

2 Related Work

The DBpedia project[1] extracts structured information from Wikipedia (Auer et al. 2007). The 2016-04 version includes 28.6M entities, of which 28M are classified in the DBpedia Ontology. This DBpedia 2016-04 ontology is a directed-acyclic graph that consists of 754 classes. It was manually created and is based on the most frequently used infoboxes in Wikipedia. For each Wikipedia language version, there are mappings available between the infoboxes and the DBpedia ontology. In the current version, there are 3.2M persons, 3.1M places and 515,480 organizations. To be available in DBpedia, an entity needs to have a Wikipedia page (in at least one language version that is included in the extraction) that contains an infobox for which a valid mapping is available.

YAGO3 (Mahdisoltani et al. 2015), the multilingual extension of YAGO, combines information from 10 different Wikipedia language versions and fuses it with the English WordNet. YAGO[2] concentrates on extracting facts about entities, using Wikipedia categories, infoboxes, and Wikidata. The YAGO taxonomy is constructed by making use of the Wikipedia categories. However, instead of using the category hierarchy that is "barely useful for ontological purposes" (Suchanek et al. 2007), the Wikipedia categories are extracted, filtered and parsed for noun phrases in order to map them to WordNet classes. To include the multilingual categories, Wikidata is used to find corresponding English category names. As

[1] http://wiki.dbpedia.org.
[2] http://www.yago-knowledge.org.

a result, the entities are assigned to more than 350K classes. YAGO3, which is extracted from Wikipedia dumps of 2013–2014, includes about 4.6M entities.

Both KBs solely depend on Wikipedia. Since it takes some time to update or create the KBs, they do not include up-to-date information. In contrast, the current version of Wikidata can be directly queried and a fresh Wikidata dump is available every week. Another advantage is that Wikidata does not rely on the existence of an infobox or Wikipedia page. Entities and information about the entities can be extracted from Wikipedia or manually entered by any user, meaning that less significant entities that do not warrant their own Wikipedia page are also represented. Since Wikipedia infoboxes are partially populated through templates from Wikidata entries, extracting data from infoboxes instead of Wikidata itself adds an additional source of errors. Furthermore, unless all language versions of Wikipedia are used as a source, such an approach would even limit the amount of retrieved information due to Wikidata's inherent multi-lingual design as the knowledge base behind all Wikipedias (Vrandeĉić and Krötzsch 2014).

For completeness, Freebase[3] should be mentioned as a fourth available knowledge base that has historically been used as a popular alternative to YAGO and DBpedia. However, efforts have recently been taken to merge it entirely into Wikidata (Pellissier Tanon et al. 2016). Given the need for current, up-to-date entity information in many event-related applications, the fact that Freebase is no longer actively maintained and updated means that it is increasingly ill-suited for such tasks.

3 The NECKAr Tool

The Named Entity Classifier for Wikidata (NECKAr) assigns Wikidata entites to the NE classes PERSON, LOCATION, and ORGANIZATION. The tool, which is available as open source code (see the URL in the Abstract), is easy to use and only requires a minimum setup of Python3 packages as well as an instance of a MongoDB.

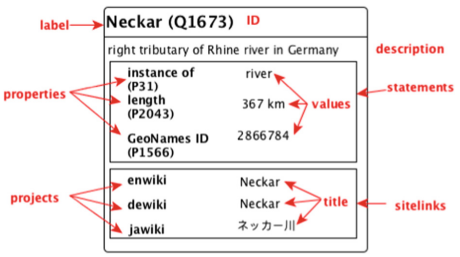

Fig. 1. Wikidata data model

[3] https://developers.google.com/freebase/.

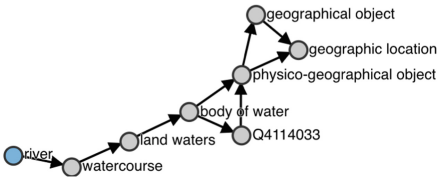

Fig. 2. Class hierarchy for *river*, generated with the Wikidata Graph Builder

3.1 Wikidata Data Model

Wikidata[4] is a free and open knowledge base that is intended to serve as central storage for all structured data of Wikimedia projects. The data model of Wikidata consists primarily of two major components: *items* and *properties*. Items represent all *things* in human knowledge. Each item corresponds to a clearly identifiable concept or object, or to an instance of a concept or object. For example, there is one item for the concept *river* and one item *Neckar*, which is an instance of a river. In this context, a concept is the same as a class. All items are denoted by numerical identifiers prefixed with a *Q*, while properties have numerical identifiers prefixed with *P*. Properties (P) connect items (Q) to values (V). A pair (P, V) is called a statement. A property classifies the value of a statement. Figure 1 shows a simplified entry for the item *Neckar*. Here P2043 describes that the value 367 km has to be interpreted as the length of the river. Both items and properties have a label, a description, and (multilingual) aliases. Property entries in Wikidata do not have statements, but data types that restrict what can be given as a properties value. These data types include items, external identifiers (e.g., ISBN codes), URLs, geographic coordinates, strings, or quantities, to name a few.

When we are interested in the classification of items, we require the knowledge which item is an instance of which class. Class membership of an item is predominately modelled by the property *instance of* (P31). For example, consider the statement Q1673:P31:Q4022 (Neckar is an instance of river), in which *Q4022* can be seen as a class. Classes can be subclasses of other classes, e.g., *river* is a subclass of *watercourse*, which is a subclass of *land water*. Figure 2 shows the subclass graph for *river*.

The property *subclass of* (P279) is transitive, meaning that since *Neckar* is an instance of *river*, which is a subclass of *watercourse*, *Neckar* is implicitly also an instance of *watercourse*. Due to this transitivity rule, in Wikidata there is no need to specify more than the most specific statement[5]. In other words, there is no statement that directly specifies *Neckar* to be a *geographic location*. Thus, we cannot simply extract items that are instances of the general classes. There are, for example, only 1,733 items that are direct instances of *geographic location*. Instead, we need to extract the transitive hull, that is, all items that are an

[4] http://www.wikidata.org.
[5] http://www.wikidata.org/wiki/Help:Basic_membership_properties.

Table 1. Location types with corresponding Wikidata root classes for location types and number of subclasses

Location type	Root classes	# sub
Continent	Q5107	1
Country	Q6256, Q1763527, Q3624078	50
State	Q7275	173
Settlement	Q486972	1224
City	Q515	126
Sea	Q165	12
River	Q4022	34
Mountain	Q8502	81
Mountain range	Q1437459	18
Territorial entity	Q15642541	3,657

instance of any subclass of the general class (henceforward *root classes*). There are several tools available to show and query the class structure of Wikidata. For Fig. 2 we used the *Wikidata Graph Builder* (WGB)[6] to visualize the class tree. For NECKAr, we make use of the SPARQL based *Wikidata Query Service*[7] to extract all subclasses of a root class, e.g., *geographic location*. Once the subclasses of a root class are identified, we can extract all items that are instances of these subclasses.

The task is then to find root classes that, together with their subclasses, best represent the predominately used NE classes LOCATION, ORGANIZATION, and PERSON. In the following we describe how items of these classes are extracted and what kind of information we store for each item. For all items, we store the Wikidata ID, the label (the most common name for an item), the links to the English and German Wikipedia, and the description.

3.2 Location Extraction

To extract all locations from Wikidata, we use the root class *geographic location* (Q222-1906). This class is very large and includes 23,383 subclasses[8]. For each location item, we extract the following statements: coordinate location (P625), population (P1082), country (P17), and continent (P30). Additionally, we assign a location type if an item is an instance of a subclass of the root class for that location type (see Table 1).

In this large set of subclasses of *geographic location*, we encounter several problems. For example, *Food* is a subclass of *geographic location*. *Food* is connected to *geographic location* by a path of length 3 (*food* → *energy*

[6] http://angryloki.github.io/wikidata-graph-builder/.
[7] https://query.wikidata.org.
[8] In the following, all class and subclass sizes are as of February 22, 2017.

Table 2. Example of classified entities

neClass	LOCATION		ORGANIZATION		PERSON
id	Q1796771		Q81230		Q76658
norm_name	Köthen		Siemens		Frank-Walter Steinmeier
description	capital of the district of Anhalt-Bitterfeld Saxony-Anhalt		Engineering and electronics conglomerate		politician
en Wikipedia	Köthen (Anhalt)		Siemens		Frank-Walter Steinmeier
location type	city, settlement	**instance of**	concern,	**occupation**	politician,
population	26,384		bus. enterprise		jurist, lawyer
continent	Europe	**CEO**	Joe Kaeser	**gender**	male
country	Germany		Klaus Kleinfeld	**dob**	1956-01-05
coordinate	51.75	**founder**	Ernst Werner	**dod**	none
	11.916666666667		von Siemens	**alias**	Steinmeier
GeoNames	2885237	**inception**	1847-10-01		
		HQ	Munich		
		country	Germany		
		website	www.siemens.com		

storage → *storage* → *geographic location*). We cannot simply limit the allowed path length since there are other subclasses with a greater path length that we consider a valid location. For example the shortest path for *village of Japan* has a length of 4 (*village of Japan* → *municipality of Japan* → *municipality* → *human settlement* → *geographic location*). In this case we decided to exclude the subtree for *Food*, which reduces the number of subclasses considerably to 13,445. However, there might be other subclasses that are not considered a proper location (e.g., *Arcade Video Game* with the path: *arcade video game* → *arcade game machine* → *computing platform* → *computing infrastructure* → *infrastructure* → *construction* → *geographical object* → *geographic location*). For the time being we only exclude the *Food* subclasses. The identified location items can be filtered for a certain application by using the location type or by only using items for which a coordinate location is given.

3.3 Organization Extraction

The root class *organization* (Q43229) includes 4811 subclasses, such as *nonprofit organization*, *political organization*, *team*, *musical ensemble*, *newspaper*, or *state*. For each item in this category, we extract additional information such as country (sovereign state of this item, P17), founder (P112), CEO (P169), inception (P571), headquarter location (P159), instance of (P31), official website (P856), and official language (P37).

3.4 Person Extraction

To extract all real world persons from Wikidata, we only use the class *human* (Q5) instead of a list of subclasses. In Wikidata, a more specific classification of a person is usually given by the occupation property or by having several *instance of* statements. All items with the statement *is instance of human* are classified as person. Fictional characters, such as *Homer Simpson* or *Harry Potter* and deities that are not also classified as human, are not extracted. For each person, we gather some basic information: date of birth (dob) (P569), date of death (dod) (P570), gender (P21), occupation (P106), and alternative names.

3.5 Extracting Links to Other Knowledge Bases

In addition to the above information, we also record identifiers for the items in other publicly available databases (Wikipedia, DBPedia, Integrated Authority File of the German National Library, Internet Movie Database, MusicBrainz, GeoNames, OpenStreet Map). This information is represented in Wikidata as statements and can be extracted analogously to the examples above.

3.6 Extraction Algorithm

The named entity classes to be extract can be specified in a configuration file. For each chosen class of named entities, the process then works as described in the following. First, the subclasses of the root class are extracted using the Wikidata SPARQL API. The output of this step is a list of subclasses, from which the invalid subclasses are excluded. For locations, we also generate lists of subclasses for the specific location types. The tool then searches the Wikidata dump (stored in a local MongoDB) for all items that are instances of one of the subclasses in the list and extracts the common features (id, label, description, Wikipedia links). Depending on the named entity class, additional information (see above) is extracted, and for locations, the list of location type subclasses is used to assign a location type. This data is then stored in a new, intermediary MongoDB collection. In a subsequent step, we extract for each item the identifiers that link them to the other databases as described in Sect. 3.5 and store them in a separate collection. In the last step, the data is exported to CSV and JSON files for ease of use.

4 Wikidata NE Dataset

The Wikidata NE dataset[9] was extracted using the NECKAr tool. For the version that we discuss in this paper, we extracted entities from the Wikidata dump from December 5, 2016, which includes 24,580,112 items and 2,910 distinct properties.

In total, we extracted and classified 8,842,103 items, of which 51.8% are locations, 37.6% persons, and 10.6% organizations. Table 2 shows examples for each named entity class, including the class specific additional information.

[9] http://event.ifi.uni-heidelberg.de/?page_id=532.

4.1 Location Entities

Of the 4,582,947 identified locations, 51% have geographic coordinates. Location types are extracted for 93% of the location items (see Table 3).

Table 3. Number of entities for location types

Type	Continent	Country	State	City	Territorial entity
Count	10	2,496	4,330	25,470	1,805,213
Type	Settlement	Sea	River	Mountain range	Mountain
Count	1,983,860	183	199,991	7,814	229,853

Most of the classified locations are settlements and territorial entities. We find over 2,400 countries: although there currently are only 206 countries[10], Wikidata also includes former countries like the Roman Empire, Ancient Greece, or Prussia.

4.2 Person Entities

We extracted 3,322,217 persons, of which 78% are male, 15% female, while for 7% of the persons another gender or no gender is specified. Occupations are given for 66% of the person items, where the largest group are *politicians*, followed by *football players* (see Table 4).

Table 4. The five most frequent occupations

#	1	2	3	4	5
Occupation	Politician	Assoc. football player	Actor	Writer	Painter
Count	312,571	206,142	170,291	127,837	99,060

Wikidata covers mostly persons from recent history, so 70% of the persons for whom a birth date is given (over 2,5M persons) were born in the 20th century, while around 20% were born in the 19th century.

4.3 Organization Entities

936,939 items were classified as organizations, of which 11% are business enterprises. Table 5 shows the top 5 organization types. Where possible, we also extracted the country in which the organization is based. Figure 3 shows a heatmap of the number of organizations per country. Most organizations are based in the U.S.A., followed by France and Germany. This is partially due to the fact that *commune of France* and *municipality of Germany* are subclasses of *organization*.

[10] https://en.wikipedia.org/wiki/List_of_sovereign_states.

Table 5. The five most frequent organization types

#	Type	Count
1	Business enterprise	102,129
2	Band	58,996
3	Commune of France	38,387
4	Primary school	36,078
5	Association football club	31,257

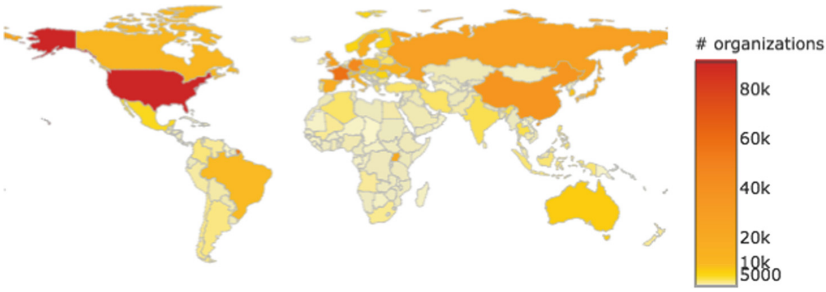

Fig. 3. Heatmap of organization frequency by country

4.4 Assignment to More Than One Class

400,856 Wikidata items are assigned to more than one NE class by NECKAr. The vast majority of this subset (over 99%) are members of the two classes *location* and *organization*. This is mainly caused by a subclass overlap between the root classes *geographic location* and *organization*. In total, they share 1,310 subclasses, e.g., *hospital*, *state* or *library* and their respective subclasses. We do not favour one class over the other, because both interpretations are possible, depending on the context. There are also items that have several *instance of* statements, which in six cases leads to an assignment to all three classes, e.g., *Jean Leon* is described as instance of human and instance of winery, which is a subclass of both organization and geographic location. There are 116 items that are classified as person and location or person and organization, which is again caused by multiple *instance of* statements. In contrast to the subclass overlap between root classes, these cases are caused by incorrect user input into Wikidata.

5 Comparison to YAGO3

In order to get an estimate of the quality of the NECKAr tool, we compare the resulting Wikidata NE dataset to the currently available version of YAGO (Version 3.0.2). When using the YAGO3 hierarchy to classify YAGO3 entities, we find 1,745,219 distinct persons (member of YAGO3 class

`wordnet_person_100007846`), 1,267,402 distinct locations (member of YAGO class `yagoGeoEntity`) and 481,001 distinct organisations (member of YAGO class `wordnet_social_group_107950920`) for a total of 3,493,622 entities in comparison to the 8,8M entities in the Wikidata NE dataset (see Table 6).

YAGO3 entities can be linked to Wikidata entries via their subject id, which corresponds to Wikipedia page names. If a YAGO3 entity is derived from a non-English Wikipedia, the subject id is prefixed with the language code. For 3,430,065 YAGO3 entities we find a corresponding entry in Wikidata (1,715,305 persons, 1,250,409 locations and 464,351 organization). This subset is the basis for our comparison in the following.

To assess the quality of NECKAr, the well-known IR measures F_1-score, precision (P) and recall (R) are used. Precision is a measure for exactness, that is, how many of the classified entities are classified correctly. Recall measures completeness and gives the fraction of correctly classified entities of all given entities. F_1 is the harmonic mean of P and R. The measures are defined as:

$$F_1 = 2 * \frac{P * R}{P + R} \qquad P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \tag{1}$$

Here, TP (true positives) is the number of YAGO3 entites, that NECKAr assigns to the same class, while FP (false positives) is the number of YAGO3 entities that are falsely assigned to that class. $TP+FP$ represents the number of entities assigned to that class by NECKAr. FN (false negatives) is the number of YAGO3 entities in a given class that NECKAr does not assign to that class (these entities might be assigned to a different class or to no class). Thus, $TP + FN$ is the number of YAGO3 entities in a given class. Using these standard metrics, we receive a overall F_1-score of 0.88 with $P = 0.90$ and $R = 0.86$ (see Table 7). The lower recall is due to the fact that NECKAr does not classify all entries that are a *person*, *location*, or *organization* entity in YAGO3. Only about 88% of the YAGO3 entites that correspond to Wikidata entries are classified. For example, NECKAr does not find *Pearson*, a town in Victoria, Australia, because the Wikidata entry does not include any *is instance of* relation. This is true for 290,905 of the 387,259 entities (75.12%) that are not classified by NECKAr. Some entities are missed by NECKAr entirely for a couple of reasons. In some cases, the correct *is instance of* relation is not given in Wikidata. In others, a relevant subclass or property may not have been included. Finally, since YAGO3 was automatically extracted and not every fact was checked for correctness it contains some erroneous claims or classifications. For example, some overview articles in Wikipedia are classified as entities in YAGO, such as *Listed_buildings_in* or *Index of*. The original evaluation of YAGO3 lists the fraction of incorrect facts that it contains as 2% (Mahdisoltani et al. 2015).

5.1 Location Comparison

For LOCATION, NECKAr achieves a F_1-score of 0.88 (P = 0.93 and R = 0.84). 170,869 YAGO3 locations were not classified, of which 81% have no entry in Wikidata for the *instance of* property.

Table 6. Number of entities per class in the Wikidata NE dataset created by NECKAr, YAGO3 and the intersection of YAGO3 and Wikidata

NE	NECKAr	Yago3	Yago3 ∩ WD
LOC	4,582,947	1,267,402	1,250,409
PER	3,322,217	1,745,219	1,715,305
ORG	0936,939	0481,001	0464,351

Table 7. Evaluation results (F_1 score, Precision (P) and Recall (R)) for the Wikidata NE dataset created by NECKAr in comparison to YAGO3

NE	F_1	P	R
LOC	0.88	0.93	0.84
PER	0.97	0.99	0.95
ORG	0.57	0.54	0.60
All	0.88	0.90	0.86

Of the entities that are assigned to a different class, NECKAr classified 97,6% as ORGANIZATION instead of LOCATION. Most of these entities (85%) are radio or television stations for which a classification into either class is a matter of debate. These items are described in Wikidata as *instance of radio station* or *television station* which are subclasses of ORGANIZATION. The majority of the FPs for locations are assigned by NECKAr to two classes (ORGANIZATION and LOCATIONS), whereas in YAGO3 these are only organizations.

5.2 Person Comparison

For entities of class PERSON, we receive the highest F_1-score of all classes with 0.97 (P = 0.99 and R = 0.95). Most of the entities that NECKAr assigned to a different class (90% to ORGANIZATION, 10% to LOCATION) are bands or musical ensembles which are classified as ORGANIZATION.

5.3 Organization Comparison

The class ORGANIZATION shows the lowest F_1-score = 0.57 (P = 0.54 and R = 0.60). The low precision is caused by the high number of false positives. As discussed in the previous section, many entities that are classified as Persons or Locations by YAGO3 are classified as organizations by NECKAr. The low recall is due to the fact that 156,926 YAGO3 organizations were not identified by NECKAr. Again, the majority of these items (82%) has no *is instance of* relation in Wikidata, so NECKAr was not able to classify them. The reason for the missing 18% warrants future investigation in more detail, as it is possible that an important subclass or property was excluded. 29,625 YAGO3 organizations were assigned to another class, 96% to LOCATION and 4% to PERSON.

Many of these items are constituencies or administrative units, which could be seen as organizations and/or locations.

In summary, we find that the application of NECKAr to Wikidata produces a set of classified entities that is comparable in quality to a well known and widely used knowledge base. However, in contrast to existing knowledge bases, which are not updated regularly, NECKAr can be used to extract substantially more entities and up-to-date lists of persons, locations and organizations. Since NECKAr can be applied to weekly dumps of Wikidata, it can be used to extract a reproducible resource for subsequent IE tasks.

6 Conclusion and Future Work

In this paper, we introduced the NECKAr tool for assigning NE classes to Wikidata items. We discussed the data model of Wikidata and its class hierarchy. The resulting NE dataset offers the simple classification of entities into locations, organizations and persons that is often used in IE tasks. The datasets includes basic, class specific information on each item and links them to other linked open data sets. The clear and lightweight structure makes the dataset a valuable and easy to use resource. Much of the original more fine grained classification is preserved and can be used to create application-specific subsets. A comparison to YAGO3 showed that NECKAr is able to create state-of-the-art lists of entities with the added advantage of providing larger and more recent data.

Based on these results, we are further investigating the Wikidata class hierarchy in order to reduce the number of incorrect or multiple assignments. We are also working on an automated process to provide the Wikidata NE dataset on a monthly basis. In future releases of NECKAr, we plan to include the option of choosing between a Wikidata dump and the SPARQL API as source for obtaining the entity data.

References

Aggarwal, C.C., Subbian, K.: Event detection in social streams. In: Proceedings of the Twelfth SIAM International Conference on Data Mining, ICDM, pp. 624–635 (2012). https://doi.org/10.1137/1.9781611972825.54

Allan, J.: Topic Detection and Tracking: Event-based Information Organization, vol. 12. Springer Science & Business Media, New York (2012). https://doi.org/10.1007/978-1-4615-0933-2

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC-2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52

Brants, T., Chen, F.: A system for new event detection. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 330–337 (2003). http://doi.acm.org/10.1145/860435.860495

Brasileiro, F., Almeida, J.P.A., de Carvalho, V.A., Guizzardi, G.: Applying a multi-level modeling theory to assess taxonomic hierarchies in Wikidata. In: Proceedings of the 25th International Conference on World Wide Web, WWW Companion Volume, pp. 975–980 (2016). http://doi.acm.org/10.1145/2872518.2891117

Ding, J., Gravano, L., Shivakumar, N.: Computing geographical scopes of web resources. In: Proceedings of 26th International Conference on Very Large Data Bases, VLDB, pp. 545–556 (2000). http://www.vldb.org/conf/2000/P545.pdf

Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semant. Web **9**(1), 77–129 (2018). https://doi.org/10.3233/SW-170275

Finkel, J.R., Grenager, T., Manning, C.D.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, ACL, pp. 363–370 (2005). http://aclweb.org/anthology/P/P05/P05-1045.pdf

Geiß, J., Gertz, M.: With a little help from my neighbors: person name linking using the Wikipedia social network. In: Proceedings of the 25th International Conference on World Wide Web, WWW Companion Volume, pp. 985–990 (2016). http://doi.acm.org/10.1145/2872518.2891109

Heindorf, S., Potthast, M., Stein, B., Engels, G.: Towards vandalism detection in knowledge bases: corpus construction and analysis. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 831–834 (2015). http://doi.acm.org/10.1145/2766462.2767804

Kumaran, G., Allan, J.: Text classification and named entities for new event detection. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 297–304 (2004). http://doi.acm.org/10.1145/1008992.1009044

Mahdisoltani, F., Biega, J., Suchanek, F.M.: YAGO3: a knowledge base from multilingual Wikipedias. In: Seventh Biennial Conference on Innovative Data Systems Research, CIDR (2015). http://cidrdb.org/cidr2015/Papers/CIDR15_Paper1.pdf

Müller-Birn, C., Karran, B., Lehmann, J., Luczak-Rösch, M.: Peer-production system or collaborative ontology engineering effort: what is Wikidata? In: Proceedings of the 11th International Symposium on Open Collaboration, pp. 20:1–20:10 (2015). http://doi.acm.org/10.1145/2788993.2789836

Pellissier Tanon, P., Vrandeĉić, D., Schaffert, S., Steiner, T., Pintscher, L.: From freebase to Wikidata: the great migration. In: Proceedings of the 25th International Conference on World Wide Web, WWW, pp. 1419–1428 (2016). http://doi.acm.org/10.1145/2872427.2874809

Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, WWW, pp. 851–860 (2010). http://doi.acm.org/10.1145/1772690.1772777

Spitz, A., Dixit, V., Richter, L., Gertz, M., Geiß, J.: State of the union: a data consumer's perspective on Wikidata and its properties for the classification and resolution of entities. In: Wiki, Papers from the 2016 ICWSM Workshop (2016a). http://aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13200

Spitz, A., Geiß, J., Gertz, M.: So far away and yet so close: augmenting toponym disambiguation and similarity with text-based networks. In: Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data, GeoRich@SIGMOD, pp. 2:1–2:6 (2016b). http://doi.acm.org/10.1145/2948649.2948651

Strötgen, J., Gertz, M.: Domain-Sensitive Temporal Tagging. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Rafael (2016). https://doi.org/10.2200/S00721ED1V01Y201606HLT036

Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, WWW, pp. 697–706 (2007). http://doi.acm.org/10.1145/1242572.1242667

Vrandeĉić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014). http://doi.acm.org/10.1145/2629489