Multilingual enrichment of disease biomedical ontologies

Léo Bouscarrat, Antoine Bonnefoy, Cécile Capponi, Carlos Ramisch

Multilingual enrichment of disease biomedical ontologies

Léo Bouscarrat[1,2], Antoine Bonnefoy[1], Cécile Capponi[2], Carlos Ramisch[2]
[1]EURA NOVA, Marseille, France
[2]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
{leo.bouscarrat, antoine.bonnefoy}@euranova.eu
{leo.bouscarrat, cecile.capponi, carlos.ramisch}@lis-lab.fr

Abstract

Translating biomedical ontologies is an important challenge, but doing it manually requires much time and money. We study the possibility to use open-source knowledge bases to translate biomedical ontologies. We focus on two aspects: coverage and quality. We look at the coverage of two biomedical ontologies focusing on diseases with respect to Wikidata for 9 European languages (Czech, Dutch, English, French, German, Italian, Polish, Portuguese and Spanish) for both ontologies, plus Arabic, Chinese and Russian for the second one. We first use direct links between Wikidata and the studied ontologies and then use second-order links by going through other intermediate ontologies. We then compare the quality of the translations obtained thanks to Wikidata with a commercial machine translation tool, here Google Cloud Translation.

Keywords: biomedical, ontology, translation, wikidata

1. Introduction

Biomedical ontologies, like Orphanet (INSERM, 1999b), play an important role in many downstream tasks (Andronis et al., 2011; Li et al., 2015; Phan et al., 2017), especially in natural language processing (Maldonado et al., 2017; Nayel and Shashrekha, 2019). Today either the vast majority of these ontologies are only available in English or their restrictive licenses reduce the scope of their usage. There is nowadays a real focus on reducing the prominence of English, thus on working on less-resourced languages. To do so, there is a need for resources in other languages, but the creation of such resources is time and money consuming.

At the same time, the Internet is also a source of incredible projects aiming to gather a maximum of knowledge in a maximum of languages. One of them is the collaborative encyclopedia Wikipedia, opened in 2001, which currently exists in more than 300 languages. As it contains mainly plain text, it is hard to use it as a resource as is. However, several knowledge bases have been built from it: DBpedia (Lehmann et al., 2015) and Wikidata (Vrandečić and Krötzsch, 2014). The main difference between these two knowledge graphs is the update process: while Wikidata is manually updated by users, DBpedia extracts its information directly from Wikipedia. Compared to biomedical ontologies they are structured using less expressive formalisms and they gather information about a larger domain. They are open-source, thus can be used for any downstream tasks. For each entity they have a preferred label, but sometimes also alternative labels that can be used as synonyms. For example, the entity *Q574227* in Wikidata has the preferred label *2q37 monosomy* in English along with the alternative labels in English: *Albright Hereditary Osteodystrophy-Like Syndrome* and *Brachydactyly Mental Retardation Syndrome*. Moreover, entities in these two knowledge bases also have translations in several languages. For example, the entity *Q574227* in Wikidata has the preferred label *2q37 monosomy* in English and the preferred label *Zespół delecji 2q37* in Polish. They also feature some links between their own entities and entities in external biomedical ontologies. For example, the entity *Q574227* in Wikidata has a property *Orphanet ID* (*P1550*) with the value *1001*.

By using both kinds of resources, biomedical ontologies and open-source knowledge bases, we could partially enrich biomedical ontologies in languages other than English. As links between the entities of these resources are already existing, we expect good quality. To further enrich them we could even look at second-order links since many biomedical ontologies also contain some links to other ontologies.

The goal of this work is twofold:

- to study the coverage of such open-source collaborative knowledge graphs compared to biomedical ontologies,
- to study the quality of the translations using first- and second-order links and comparing this quality with the quality obtained by machine translation tools.

This paper is part of a long-term project whose goal is to work on multilingual disease extraction from news with strategies based on dictionary expansion. Consequently, we need a multilingual vocabulary with diseases which are normalized with respect to an ontology. Thus, we focus on one kind of biomedical ontologies, that is, ontologies about diseases.

2. Resources and Related Work

There has already been some work trying to use open-source knowledge bases to translate biomedical ontologies. Bretschneider et al. (2014) obtain a German-English medical dictionary using DBPedia. The goal is to perform information extraction from a German biomedical corpus. They could not directly use the RadLex ontology (Langlotz, 2006) as it is only available in English. So, they first extract term candidates in their German corpus. Then, they try to match the candidates with the pairs in their German-English dictionary. If a candidate is in the dictionary, they

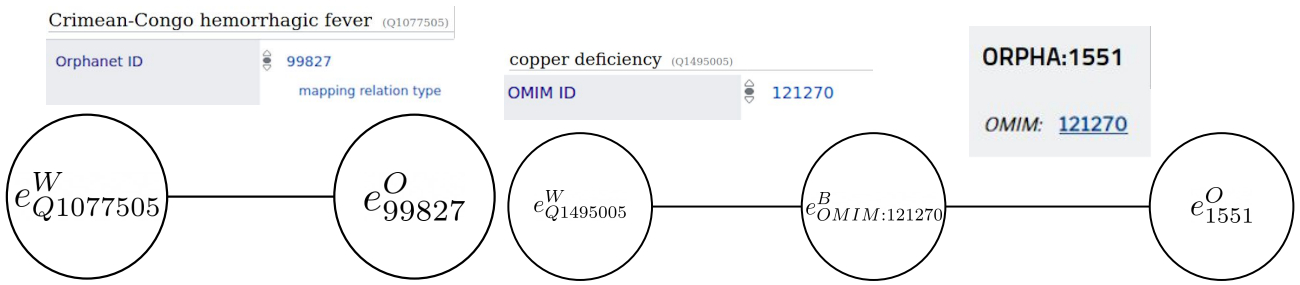

Figure 1: Example of first-order link (left) and second-order link (right)

use the translation to match with the RadLex ontology. Finally, this term candidate alongside with the match in the RadLex ontology is processed by a human to validate the matching.

Alba et al. (2017) create a language-independent method to maintain up-to-date ontologies by extracting new instances from text. This method is based on a human-in-the-loop who helps tuning scores and thresholds for the extraction. Their method requires some "contexts" to start finding new entities to add to the ontology. To bootstrap the contexts, they can either ask a human to annotate some data or use an oracle made by the dictionary extracted from the DBpedia and Wikidata using word matching on the corpus. They then look for good candidates, i.e., a set of words surrounding an item, by looking for elements in similar contexts to the one found using the bootstrapping. Then, a human-in-the-loop validates the newly found entities, adding them to the dictionary if they are correct, or down-voting the context if they are not relevant entities.

Hailu et al. (2014) work on the translation of the Gene Ontology from English to German and compare three different approaches: DBpedia, the Google Translate API without context, and the Google Translate API with context. To find the terms in DBpedia they use keyword-based search. After a human evaluation, they find that translations obtained with DBpedia have the lowest coverage (only 25%) and quality compared to those obtained with Google Translate API. However, to compare the quality of the different methods they only use the translation of 75 terms obtained with DBpedia compared to 1,000 with Google Translate API. They also note that synonyms could be a useful tool for machine translation and that using keyword-based exact match query to match the two sources could explain the low coverage.

Silva et al. (2015) compare three methods to translate SNOMED CT from English to Portuguese: DBpedia, ICD-9 and Google Translate. To verify the quality of the different approaches they use the CPARA ontology which has been hand-mapped to SNOMED CT. It is composed of 191 terms and focused on allergies and adverse reactions. They detect coverage of 10% with the ICD-9, 37% with DBpedia and 100% with Google Translate. To compare the quality of their translations they use the Jaro Similarity (Jaro, 1989).

We elaborate on these ideas by adding some elements. First of all, compared to Hailu et al. (2014) and Silva et al. (2015), we use already existing properties to perform the matching between the biomedical ontology and the knowledge graph, which should improve the quality with regard to the previous works. We also go further than these first-order links and explore the possibility of using second-order links to improve the coverage of the mappings between the sources. Compared to the same works, we also present a more complete study, Hailu et al. (2014) only evaluate on 75 terms and Silva et al. (2015) on 191 terms. We compare the coverage and quality of the entire biomedical ontology containing 10,444 terms. Furthermore, as we want to use the result of this work for biomedical entity recognition, synonyms of entities are really important for recall and also for normalisation, thus we also quantify the difference of quantity of synonyms between the original biomedical ontology and those found with Wikidata.

In this work, as we focus on diseases, we use a free dataset extracted from Orphanet (INSERM, 1999b) to perform the evaluation. Orphanet is a resource built to gather and improve knowledge about rare diseases. Through Orphadata (INSERM, 1999a), free datasets of aggregated data are updated monthly. One of them is about rare diseases, including cross-references to other ontologies. The Orphadata dataset contains the translation of 10,444 entities for English, French, German, Spanish, Dutch, Italian, Portuguese, 10,418 entities in Polish and 9,323 in Czech. All the translations have been validated by experts, thus can be used as a gold standard for multilingual ontology enrichment. One issue of this dataset is that rare diseases are, by definition, not well known. Therefore, one may expect a lower coverage than a less focused dataset; thus we propose to also measure the coverage of another dataset, Disease Ontology (Schriml et al., 2019). However we cannot use it to evaluate the translation task as it does not contain translations.

As an external knowledge base, we use Wikidata. It has many links to external ontologies, especially links to biomedical ontologies such as *wdt:P1550* for Orphanet, *wdt:P699* for Disease Ontology, and *wdt:P492* for the Online Mendelian Inheritance in Man (OMIM). It is also important to note that, over the 9 languages we studied, only the Czech Wikipedia has less than 1,000,000 articles. This information can be used as a proxy for the completeness of the information in each language on Wikidata. We prefer it over DBpedia as we find it easier to use, especially to find the properties.

As a machine translation tool, we use Google Cloud Translation. It is a paying service offered by Google Cloud.

3. Methods and Experiments

In this section, we first define the notations used in this paper, then we describe how we extract the first- and second-order links from our sources. Afterwards, we describe how we perform machine translation. The evaluation metrics are subsequently explained and finally we describe our evaluation protocol.

3.1. Definition and Notations

We define:

- e_i^S as an entity in the source knowledge base S, $S \in [O, W, B]$ where O is Orphanet, W is WikiData and B are all the other external biomedical ontologies used. An entity is either a concept in an ontology or in a knowledge graph.
- $E^S = \{e_i^S\}_{i=1...|E^S|}$ is the set of all the entities in the source S.
- $E = E^O \cup E^W \cup E^B$ is the set of all the entities in all the sources.
- $\mathrm{L}_l(e)$ is the preferred label of the entity e in the language l, or $\emptyset$ if there is no label in this language.
- $\mathcal{L}_l(e)$ represents all the possible labels of the entity e in the language l or $\emptyset$ if there is no label in this language. Furthermore, $\mathrm{L}_l(e) \in \mathcal{L}_l(e)$
- T is a set of links, such that $t \in T$ with $t = (e_i^s, e_j^{s'}), s \neq s'$.
- $G = (E, T)$ is an undirected graph.
- $\mathcal{V}(e_i) = \{e_j \in E | \exists t \in T, t = (e_i, e_j)\}$, defines the set of all the neighbours of the entity e_i.
- $\mathcal{W}(e) = \{v \in \mathcal{V}(e) | v \in W\}$, defines the set of all the neighbours that are in Wikidata of the entity e.
- $MT(\{s_1, ..., s_n\}, l)$ is a function that returns the labels $\{s_1, ..., s_n\}$ translated from English to the language l thanks to Google Cloud Translation.

3.2. Gathering Links between Entities

3.2.1. First-Order Links

The first step of our method consists in gathering all the information about the sources. To obtain the gold translations, we use Orphadata. We collected all the JSON files from their website[1] on January 15, 2020. We extract the OrphaNumber, the Name, the SynonymList and the ExternalReferenceList of each element in the files.

For WikiData we use the SPARQL endpoint[2]. We query all the entities having a property OrphaNumber *wdt:P1550*, and, for these entities, we obtain all their preferred labels (*rdfs:label*) and synonyms (*skos:altLabel*), corresponding to E_i^O in the 9 European languages included in Orphanet. The base aggregator of the synonyms uses a comma to separate them. In our case, this error-prone because the comma can also be part of the label, for example one of the alternative label of the entity *Q55786560* is *49, XXXYY syndrome*. We needed to concatenate the synonyms with another symbol[3]. Thanks to the property which gives the Orphanumber of the related entity in Orphanet we can create links $t = (e^O, e^W)$ between an entity e_i^W in Wikidata and and entity e_i^O in Orphanet.

The mapping is then trivial, as we have the OrphaNumber in the two sources. On the left of Figure 1 we can see that the entity *Q1077505* in Wikidata has a property *Orphanet ID* with the value *99827*, thus we can create $t = (Q1077505^W, 99827^O)$. Nonetheless, the mapping is not always unary, because several Wikidata entities can be linked to the same Orphanet entity.

Formally, the set of Orphanet entities with at least one first-order link is:

$$E^F = \{e \in E^O | \exists w \in W, (e, w) \in T\}$$

3.2.2. Second-Order Links

Orphanet provides some external references to auxiliary ontologies. We add these references to our graph: $t = (e^O, e^B) \in T$. Even if there are already first-order links between Orphanet and Wikidata, we cannot ensure that all the entities are linked. To improve the coverage of translations, we can use second-order links, creating an indirect link when entities from Wikidata and Orphanet are linked to the same entity in a third external source B. For example, on the right of Figure 1, we extract the link between the entity *Q1495005* of Wikidata and the entity *121270* of OMIM. We also extract from Orphanet that the entity *1551* of Orphanet is link to the same entity of OMIM. Therefore, as a second-order relation, the entity *Q1495005* of Wikidata and the entity *1551* of Orphanet are linked.

The objective is to find some links $t' = (e^W, e^B)$ where $\exists v \in \mathcal{V}(e^B)$ and $v \in E^O$. Consequently, we are looking for links between entities from Wikidata and the external biomedical ontologies, whenever the entity in the external biomedical ontology already has a link with an entity in Orphanet.

For that purpose, we extract all the links between Wikidata and the external biomedical ontologies in the same fashion as from Orphanet, using the appropriate Wikidata properties. In the previous example, we create links $(Q1495005^W, OMIM : 121270^B) \in T$ and $(1551^O, OMIM : 121270^B) \in T$.

[1] `http://www.orphadata.org/cgi-bin/rare_free.html`

[2] `https://query.wikidata.org/sparql` can be queried with the interface `https://query.wikidata.org/`

[3] We made a package to extract entities from Wikidata: `https://github.com/euranova/wikidata_property_extraction`

We can now map Wikidata and Orphanet using second-order links. This set of links is denoted as:

$$C = \{e \in E^O | \exists (w,b) \in E^W \times E^B, \\ (e,b) \in T, (w,b) \in T\}$$

We also define the set of all the second-order linked Wikipedia entities of a specific Orphanet entity:

$$\mathcal{C}(e^O) = \{w \in E^W | \exists b \in E^B, (e,b) \in T, (w,b) \in T\}$$

3.3. Machine Translation

We use Google Cloud Translation as a machine translation tool to translate the labels of the ontology from English to a target language. As we want to have the same entities in the test set as for Wikidata, for each language we only translate the Orphanet entities which have at least one first-order link to an entity in Wikidata with a label in the target language. So for an entity e, for the language l the output of Google Cloud Translation is:

$$MT(\mathcal{L}_{en}(e), l)$$

3.4. Definition of Evaluation Metrics

In this section, we define the different evaluation metrics that are used to evaluate the efficiency of the method.

3.4.1. Coverage Metric

To estimate the coverage of Wikipedia on a biomedical ontology we use the following metric:

$$Coverage(E_1, E_2, l) = \frac{|\{e \in E_1 | \, \mathrm{L}_l(e) \neq \emptyset\}|}{|\{e' \in E_2 | \, \mathrm{L}_l(e') \neq \emptyset\}|}$$

where E_1 and E_2 are sets of entities.

3.4.2. Jaro Similarity and n-ary Jaro

In order to evaluate the quality of the translations, we follow Silva et al. (2015) choosing the Jaro similarity, which is a type of edit distance. We made this choice as we are looking at entities. Whereas other measures such as BLEU (Papineni et al., 2002) are widely used for translation tasks, they have been designed for full sentences instead of relatively short ontology labels. The Jaro Similarity is defined as:

$$\mathrm{J}(s,s') = \frac{1}{3}\left(\frac{m}{\mid s \mid} + \frac{m}{\mid s' \mid} + \frac{m-t}{m}\right) \; s, s' \in \{a, ..., z\}^*$$

with s and s' two strings, $\mid s \mid$ the length of s, t is half the number of transpositions, m the number of *matching characters*. Two characters from s and s' are *matching* if they are the same and not further than $\frac{max(|s|,|s'|)}{2} - 1$. The Jaro Similarity ranges between 0 and 1, where the score is 1 when the two strings are the same.

However, since one Orphanet entity may have several neighbour Wikidata entities, we cannot use the Jaro similarity directly. We choose to use the $\max$, for considering the quality of the closest entity:

$$\mathcal{J}_{\max}(s, [s_1, ..., s_n]) = \max_{s' \in [s_1, ..., s_n]} \mathrm{J}(s, s')$$

3.4.3. Quality Metrics

From assessing the quality of the translations, we create 4 different measures with different goals. For each entity in each language, there is a preferred label $\mathrm{L}_l(e)$ and a list of all the possible labels $\mathcal{L}_l(e)$. All of the metrics range between 0 and 1, the higher the better.

$$\mathcal{M}_{\mathrm{pl}}(e, [e_1, ..., e_n], l) = \mathcal{J}_{\max}(\mathrm{L}_l(e), [\mathrm{L}_l(e_1), .., \mathrm{L}_l(e_n)])$$

$$\mathcal{M}_{\mathrm{bl}}(e, [e_1, ..., e_n], l) = \mathcal{J}_{\max}\left(\mathrm{L}_l(e), \bigcup_{i=1}^{n} \mathcal{L}_l(e_i)\right)$$

$$\mathcal{M}_{\mathrm{mbl}}(e, [e_1, ..., e_n], l) = \underset{s \in \mathcal{L}_l(e)}{mean} \mathcal{J}_{\max}\left(s, \bigcup_{i=1}^{n} \mathcal{L}_l(e_i)\right)$$

$$\mathcal{M}_{\mathrm{Mbl}}(e, [e_1, ..., e_n], l) = \max_{s \in \mathcal{L}_l(e)} \mathcal{J}_{\max}\left(s, \bigcup_{i=1}^{n} \mathcal{L}_l(e_i)\right)$$

$\mathcal{M}_{\mathrm{pl}}$, for principal label, compares the preferred labels from Orphanet and Wikidata. This number is expected to be high, but as there is no reason that Wikidata and Orphanet use the same preferred label, we do not expect it to be the highest score. Nonetheless, as Wikidata is a collaborative platform, a score of 1 on a high number of entities in a different language could also indicate that the translations come from Orphanet.

$\mathcal{M}_{\mathrm{bl}}$, for best label, compares the preferred label from Orphanet against all the labels in Wikidata. The goal here is to verify that the preferred label of Orphanet is available in Wikidata.

$\mathcal{M}_{\mathrm{mbl}}$, for mean best label, takes the average of the similarity of one label in Orphanet against all the labels in Wikidata. This score can be seen as a completeness score, it evaluates the ability of finding all the labels of Orphanet in Wikidata.

$\mathcal{M}_{\mathrm{Mbl}}$, for max best label, takes the maximum of the similarity of one label in Orphanet against all the labels in Wikidata. The question behind this metric is: Do we have at least one label in common between Orphanet and Wikidata? A low score here could mean that the relation is erroneous. We expect a score close to 1 here.

We used the same measures for the machine-translated dataset, however, the difference between $\mathcal{M}_{\mathrm{pl}}$ and $\mathcal{M}_{\mathrm{bl}}$ is expected to be smaller, as we are sure that the preferred label from the translated dataset is the translation of the preferred label from Orphanet.

To obtain a score for these measures on the entire dataset, we compute the average of the scores over all Orphanet entities.

3.5. Protocol

The first step of our experiments is the extraction of first-order and second-order links from Wikidata and Orphanet as explained in 3.2.. Once these links are available, we study them, starting with their coverage. To evaluate

	$\mathcal{M}_{\mathrm{p}}\mathrm{l}$			$\mathcal{M}_{\mathrm{b}}\mathrm{l}$			$\mathcal{M}_{\mathrm{m}}\mathrm{bl}$			$\mathcal{M}_{\mathrm{M}}\mathrm{bl}$		
Lang	1st W	1+2nd W	GCT	1st W	1+2nd W	GCT	1st W	1+2nd W	GCT	1st W	1+2nd W	GCT
EN	85.5	**87.5**	N/A	91.5	**92.1**	N/A	**84.1**	80.5	N/A	**97.3**	96.6	N/A
FR	85.3	82.4	**89.8**	87.4	84.2	**90.5**	75.7	69.3	**90.1**	94.1	89.1	**97.7**
DE	77.1	67.8	**80.5**	79.1	70.3	**81.6**	67.5	60.9	**83.4**	88.7	79.0	**95.4**
ES	81.3	70.1	**92.5**	84.4	73.0	**93.0**	68.7	58.4	**90.2**	91.7	89.1	**98.3**
PL	78.0	63.8	**82.0**	82.0	61.3	**83.2**	66.6	55.9	**85.0**	90.7	77.3	**95.7**
IT	79.4	66.7	**88.4**	82.4	68.8	**89.5**	69.1	58.5	**88.1**	90.5	77.4	**97.2**
PT	79.9	64.9	**83.6**	82.1	66.5	**87.6**	**73.7**	60.8	68.4	93.5	83.5	**93.3**
NL	72.9	59.1	**88.0**	75.6	60.9	**88.7**	65.8	55.1	**89.9**	86.5	71.4	**97.2**
CS	76.3	52.8	**81.9**	79.1	54.9	**83.3**	67.5	52.3	**85.4**	88.7	68.8	**95.3**

Table 1: Scores of the different methods with the different metrics in function of the languages. 1st W represents the quality of the first-order links with Wikidata, 1+2nd W the first and second-order links, and GCT the translations obtained by Google Cloud Translation.

the coverage of Wikidata for each language, we compute $Coverage(E^F, E^O, l)$ for the 9 languages. We also compute $Coverage(C, E^O, l)$ for second-order links. As Orphanet is focused on rare diseases, we do not expect a high coverage in Wikidata. To verify this hypothesis, we do the same evaluation on the Disease Ontology, which does not focus on rare diseases.

Then, we study the quality of the different methods. We apply the 4 quality metrics defined in 3.4.3. for each language on each method:

- First-order links: $\underset{e^O \in E^F}{mean}(\mathcal{M}(e^O, \mathcal{W}(e^O), l)$
- Second-order links: $\underset{e^O \in C}{mean}(\mathcal{M}(e^O, \mathcal{C}(e^O), l)$
- Machine translation: $\underset{e^O \in E^F}{mean}(\mathcal{M}(e^O, MT(\mathcal{L}_{e^O}(l), l), l)$

Finally, we look at the number of labels we can obtain for both sources.

- Orphanet: $\underset{e \in E^F}{mean} |\mathcal{L}_l(e)|$
- Wikidata: $\underset{e \in E^F}{mean} \sum_{w \in \mathcal{W}(e)} |\mathcal{L}_l(w)|$
- GCT: $\underset{e \in E^F}{mean} |MT(\mathcal{L}_{en}(e), l)|$

The number of synonyms of an entity e in a language l is: $|\mathcal{L}_l(e)|$, and we also remove the duplicates. We then average this over all the entities which are in a first-order link and in Wikidata and Orphanet.

4. Results

In this part, we first present the results on the coverage of Wikipedia on Orphanet, then we present the quality of the translation. Afterwards, we show results about the number of synonyms in both sources and finally we discuss these results.[4]

4.1. Coverage

4.1.1. Orphanet

First, we evaluate the coverage for each language, i.e., the percentage of entities in Orphanet which have at least one translation in Wikidata.

The Orphadata dataset contains translations of English, French, German, Spanish, Dutch, Italian, Portuguese, Polish and Czech. For Wikidata, the results depend on the language as not all the entities have translations in every language.

Language	Orphanet	Wikidata (%)
English	10,444	8,870 (84.9%)
French	10,444	5,038 (48.2%)
German	10,444	1,946 (18.6%)
Spanish	10,444	1,565 (15.0%)
Polish	10,171	1,329 (13.1%)
Italian	10,444	1,175 (11.3%)
Portuguese	10,444	921 (8.8%)
Dutch	10,444	888 (8.5%)
Czech	9,323	452 (4.8%)

Table 2: Number of translated entities in Orphanet and number of Orphanet entities having at least one translation in Wikidata with first-order links. The percentage of coverage is shown in parentheses.

As we can see in Table 2 that coverage depends on the language. The coverage of English gives us the amount of entities from Orphanet having at least one link with Wikidata. Here, we have 84.9% of the entities which are already linked to at least one entity in Wikidata. It means that the property of the OrphaNumber is widely used. We can also note that the French Wikidata seems to carry more information about rare diseases than the German Wikipedia. Indeed French and German Wikipedias have approximately the same global size[5], but the German Wikidata contains much less information about rare diseases.

[4]The results can be reproduced with this code: `https://github.com/euranova/orphanet_translation`

[5]As of the 6th February 2020: `https://meta.wikimedia.org/wiki/List_of_Wikipedias`

Language	Cov 1st (%)	Cov 1st+2nd (%)
English	8,870 (84.9%)	9,317 (89.2%)
French	5,038 (48.2%)	7,922 (75.9%)
German	1,946 (18.6%)	6,350 (60.8%)
Spanish	1,565 (15.0%)	6,122 (58.6%)
Polish	1,329 (13.1%)	5,797 (57.0%)
Italian	1,175 (11.3%)	5,715 (54.7%)
Portuguese	921 (8.8%)	5,016 (48.0%)
Dutch	888 (8.5%)	5,081 (48.6%)
Czech	452 (4.8%)	3,180 (34.1%)

Table 3: Coverage in terms of number and percentage of entities in Wikidata linked to Orphanet using first-order links (Cov 1st) and first- plus second-order links (Cov 1st+2nd).

The next question is the quantity of new links we can obtain by gathering second-order links.
Table 3 shows that the second-order links improve the coverage. For English, the improvement is small. Thus, for all the other languages, second-order links really help to increase the coverage. It seems to be a good help for average-resourced languages. We have used ICD-10, Medical Subject Heading (MeSH), Online Mendelian Inheritance in Man (OMIM), and, Unified Medical Language System (UMLS) as auxiliary ontologies.

4.1.2. Disease Ontology

Even if the coverage for Orphanet in English is already high, Orphanet is focused on rare diseases, which is really specific. This specificity could have an impact on the coverage as Wikidata is not made by experts. To verify if the specificity of this ontology has an influence on coverage, we have also looked at another biomedical ontology on diseases, Disease Ontology. It is also about diseases but does not focus on rare disease. Thus, this difference in generality is expected to have an impact on the coverage.
The Disease Ontology contains 12,171 concepts. We plan to use it for future works on other languages: Arabic, Russian and Chinese. These three languages also have Wikipedias with more than 1,000,000 articles on which we could rely.
As expected, this less expert ontology seems to have better coverage than Orphanet. Table 4 shows that, even if the coverage for all the languages is better than for Orphanet, the difference is not the same for all the languages. Especially, Spanish has a coverage in Disease Ontology superior to that in Orphadata by more than 11%. We do not have an explanation for these differences.
We do not compute the second-order links for Disease Ontology because 97.2% of the Orphanet entities are already linked using first-order links.

4.2. Quality

The next question concerns the quality of the translations obtained. We can expect high-quality translations from Google Cloud Translation, but to what extent? We also want to compare the quality of translations obtained from Wikidata using first-order and second-order links. The ontology we use is heavily linked directly to Wikidata, but this is not the case for all the ontologies. For ontologies with lower first-order coverage, one could expect higher increase of the second-order coverage as observed in Table 3.

Language	Wikidata (%)
English	11,833 (97.2%)
French	7,156 (58.8%)
Spanish	3,178 (26.1%)
Arabic	2,507 (20.6%)
German	2,500 (20.5%)
Italian	2,098 (17.2%)
Polish	1,869 (15.3%)
Chinese	1,789 (14.7%)
Portuguese	1,748 (14.3%)
Russian	1,706 (14.0%)
Dutch	1,650 (13.6%)
Czech	1,001 (8.2%)

Table 4: Number of entities in Disease Ontology translated, number of Disease Ontology entities having at least one translation in Wikidata with first order links and the percentage of coverage.

The first line of Table 1 shows the matching between the English labels of the entities of Orphanet and Wikidata. $\mathcal{M}_{\mathrm{b}}\mathrm{l}$ and $\mathcal{M}_{\mathrm{M}}\mathrm{bl}$ are interesting here as they can be used as an indicator of a good match. A score of 1 means that one of the labels of Wikidata is the same as the preferred label from Orphanet ($\mathcal{M}_{\mathrm{b}}\mathrm{l}$) or one of the labels from Orphanet ($\mathcal{M}_{\mathrm{M}}\mathrm{bl}$). Considering that the scores are close to 1, the matching seems to be good.
In Table 1 we can see that Google Cloud Translation gives the best translations when evaluated with the Jaro Similarity. Nonetheless, there are still some small dissimilarities depending on the languages, it seems to works well for Spanish and less well for German and Polish. We can also note that for Portuguese, if the preferred label is well translated ($\mathcal{M}_{\mathrm{p}}\mathrm{l}$, $\mathcal{M}_{\mathrm{b}}\mathrm{l}$), it is less the case for the synonyms ($\mathcal{M}_{\mathrm{m}}\mathrm{bl}$).
Then, the first-order links from Wikidata have also some satisfactory results, there are also dissimilarities between the languages. Especially, first-order links seem to work better than the average in French. Compared to second-order links, first-order links are always better and the decrease in quality between both is substantial. Some noise is probably added by the intermediate ontologies.

4.3. Synonyms

Hailu et al. (2014) suggests that synonyms play an important role in translation. Therefore, in addition to high-quality translation, we are also interested in a high number of synonyms. In our case, the synonyms are the different labels available for each language for Orphanet and Wikidata, and the translations of the English labels for Google Cloud Translation. We want to evaluate the richness of each methods in terms of numbers of synonyms. For a fair comparison, for each language we only work on the subset where the entities in Wikidata have at least one label in the evalu-

ated language.

Lang	Orphanet	Wiki 1st	Wiki 1+2nd	GCT
EN	2.3	5.8	166.77	2.3
FR	2.36	1.49	10.59	2.39
DE	2.56	1.84	5.93	2.65
ES	2.26	2.61	9.50	2.39
PL	2.54	2.01	6.88	2.65
IT	2.36	1.85	3.50	2.5
PT	1.62	1.60	2.40	2.41
NL	2.6	1.74	3.74	2.48
CS	2.2	1.74	1.71	2.13

Table 5: Average number of labels in the different sources in function of the language. For Orphanet we only use the subset of entities linked to entities in Wikidata with at least one label in the studied language. For Google Cloud Translation, it is the translation of the English labels of Orphanet.

Table 5 shows that generally Orphanet seems to have more synonyms than Wikidata when using first-order links only. And the fact GCT has more synonyms means that Orphanet has more labels in English than in other languages on the studied subset for majority language, except Dutch and Czech. Thus, this is not the case in English. For this language Wikidata is more diverse.
When using first and second-order links, the number of synonyms is much higher, especially for English. This is related to the fact that second-order links add many new relations. This new relations always have labels in English but not always habe labels in other languages.

5. Discussion

Regarding coverage, in terms of entities only, the coverage of first-order links is already high for Orphanet and Disease Ontology, respectively 84.9% and 97.2% (for English as, in our case, all the entities have English labels). The issue comes from the labels: even if Wikidata is multilingual, in our study we see that the information is mainly in English and French, but for the other studied languages the results are substantially worse. All the entities with a link have labels in English, more than half have labels in French and then for German, only around 20% of the 8,870 linked entities in Wikidata have at least one label in German. The languages we study are among the most used languages in Wikipedia. Thus, it is already an important amount of entities that could have their labels translated from English to another of these languages. As Wikidata is a collaborative project, this number should only increase over time. Second-order links help a lot for languages other than English.
Regarding quality, Google Cloud Translation is the best method. Compared to the results obtained by Silva et al. (2015) on the translation of a subpart of MeSH in Portuguese, the quality of the label translations seems to have greatly improved. Then translations obtained through first-order links are not so distant from Google Cloud Translation. However, the quality of the translations obtained through second-order links has a substantial difference with the translation coming from first-order links. Thus, we can expect Google Cloud Translation to have an advantage as Orphanet is primarily maintained in English and French and then translated by experts to other languages. Even if Google Cloud Translation is not free, translating the entirety of the English labels of Orphanet would only cost around 16$ with the pricing as of February 6, 2020.
For the synonyms, as Orphanet seems to have more labels in English than in the other languages, translating all the labels from English to the different languages allows having more synonyms than Orphanet in other languages. Moreover, Wikidata is poorer in terms of synonyms than Orphanet except for English. This is interesting as Google Cloud Translation seems to perform good translations, and having more synonyms in English also means that if we translate them with Google Cloud Translation we could have also more synonyms in other languages. It is also important to note that Google Cloud Translation only provides one translation by label. Second-order links also bring many more synonyms for all the languages, but especially for those which have a larger Wikidata.

6. Conclusions and Future Work

One of the limitations of this work concerns information that was not used. Especially in Orphanet and Wikidata, when an entity is linked to another ontology, there is additional information about the nature of the link, for example, whether it is an exact match or a more general entity. We did not use at all this information and it could be used to improve the links we create. Wikidata also contains more information about the entities than just the labels, e.g., Jiang et al. (2013) extracts multilingual textual definitions.
We also focus our study on one type of biomedical entities, diseases. The results of this work may not be generalized to all types of entities. Hailu et al. (2014) have found equivalent results for the translation of the Gene Ontology between English and German, but Silva et al. (2015) did not find the same results on their partial translation of MeSH.
Another limitation is our study about synonyms. Having the maximum number of synonyms is useful for entity recognition and normalization. Thus, here we only have quantitatively studied the synonyms, and have not explored their quality and diversity. First- and second-order link extraction from Wikidata seems to be a good method to have more synonyms. A further assessment with an expert that could validate the synonyms could be interesting.
Furthermore, as we are interested in entity recognition, a low coverage on the ontology is not correlated with a low coverage for entities in a corpus. In Bretschneider et al. (2014), by only translating a small sub-part of an ontology they could improve the coverage of the entities in their corpus by a high margin. It will be interesting to verify this on a dataset on disease recognition.
To summarize, as of now, Google Cloud Translate seems to be the best way to translate an ontology about diseases. If the ontology does not have many synonyms, Wikidata could be a way to expand language-wise the ontology. Wikidata also contains other information about its entities which could be interesting, but have not been used in this study such as symptoms and links to Wikipedia pages.

7. Bibliographical References

Alba, A., Coden, A., Gentile, A. L., Gruhl, D., Ristoski, P., and Welch, S. (2017). Multi-lingual Concept Extraction with Linked Data and Human-in-the-Loop. In *Proceedings of the Knowledge Capture Conference on - K-CAP 2017*, pages 1–8, Austin, TX, USA. ACM Press.

Andronis, C., Sharma, A., Virvilis, V., Deftereos, S., and Persidis, A. (2011). Literature mining, ontologies and information visualization for drug repurposing. *Briefings in Bioinformatics*, 12(4):357–368, 06.

Bretschneider, C., Oberkampf, H., Zillner, S., Bauer, B., and Hammon, M. (2014). Corpus-based Translation of Ontologies for Improved Multilingual Semantic Annotation. In *Proceedings of the Third Workshop on Semantic Web and Information Extraction*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Hailu, N. D., Cohen, K. B., and Hunter, L. E. (2014). Ontology translation: A case study on translating the Gene Ontology from English to German. *Natural language processing and information systems : ... International Conference on Applications of Natural Language to Information Systems, NLDB ... revised papers. International Conference on Applications of Natural Language to Info.*, 8455:33–38, June.

INSERM. (1999a). Orphadata: Free access data from orphanet. `http://www.orphadata.org`. Accessed: 2020-02-11.

INSERM. (1999b). Orphanet: an online rare disease and orphan drug data base. `http://www.orpha.net`. Accessed: 2020-02-11.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

Jiang, G. D., Solbrig, H. R., and Chute, C. G. (2013). A semantic web-based approach for harvesting multilingual textual definitions from wikipedia to support icd-11 revision. In *4th International Conference on Biomedical Ontology, ICBO 2013 Workshops on International Workshop on Vaccine and Drug Ontology Studies, VDOS 2013 and International Workshop on Definitions in Ontologies, DO 2013-Part of the Semantic Trilogy 2013*. CEUR-WS.

Langlotz, C. (2006). Radlex: a new method for indexing online educational materials. *Radiographics: a review publication of the Radiological Society of North America, Inc*, 26(6):1595.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., and Lu, Z. (2015). A survey of current trends in computational drug repositioning. *Briefings in Bioinformatics*, 17(1):2–12, 03.

Maldonado, R., Goodwin, T. R., Skinner, M. A., and Harabagiu, S. M. (2017). Deep learning meets biomedical ontologies: knowledge embeddings for epilepsy. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1233. American Medical Informatics Association.

Nayel, H. A. and Shashrekha, H. L. (2019). Integrating Dictionary Feature into A Deep Learning Model for Disease Named Entity Recognition. *arXiv:1911.01600 [cs]*, November. arXiv: 1911.01600.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Phan, N., Dou, D., Wang, H., Kil, D., and Piniewski, B. (2017). Ontology-based deep learning for human behavior prediction with explanations in health social networks. *Information sciences*, 384:298–313.

Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., et al. (2019). Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research*, 47(D1):D955–D962.

Silva, M. J., Chaves, T., and Simoes, B. (2015). An ontology-based approach for SNOMED CT translation. *ICBO 2015*.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.