Gathering Alternative Surface Forms for DBpedia Entities

Volha Bryl, Christian Bizer, Heiko Paulheim

University of Mannheim, Germany
Research Group Data and Web Science
{volha,chris,heiko}@informatik.uni-mannheim.de

Abstract. Wikipedia is often used a source of surface forms, or alternative reference strings for an entity, required for entity linking, disambiguation or coreference resolution tasks. Surface forms have been extracted in a number of works from Wikipedia labels, redirects, disambiguations and anchor texts of internal Wikipedia links, which we complement with anchor texts of external Wikipedia links from the Common Crawl web corpus. We tackle the problem of quality of Wikipedia-based surface forms, which has not been raised before. We create the gold standard for the dataset quality evaluation, which reveales the surprisingly low precision of the Wikipedia-based surface forms. We propose filtering approaches that allowed boosting the precision from 75% to 85% for a random entity subset, and from 45% to more than 65% for the subset of popular entities. The filtered surface form dataset as well the gold standard are made publicly available.

Keywords: Wikipedia, DBpedia, surface forms, data quality, disambiguation

1 Introduction

In addition to being a large-scale high quality collection of structural knowledge extracted from Wikipedia, DBpedia [1] has proven to be a uselfull source of supporting datasets used in natural language processing (NLP). Specifically, such tasks as entity linking and disambiguation or coreference resolution often rely on knowing *surface forms* of an entity, that is, a collection of strings this entity can be referred as to (synonyms, alternatives names, etc.).

Among examples of such resources based on DBpedia or Wikipedia are the BabelNet [3] multilingual lexicalized semantic network, DBpedia Spotlight datasets [2] used by Spotlight for entity disambiguation, or the surface forms dataset presented in the 2014 edition of the NLP&DBpedia workshop [8].

The problem that is common to all the above resources is the *lack of quality evaluation*: in some cases the indirect evaluation, i.e. using the resource in a concrete NLP task provides evidence of its quality [2, 9]. However, to the best of our knowledge, no direct evaluation and no gold standards have been reported that allow assessing the quality of surface forms extracted from Wikipedia. Just

assuming the high quality of such surface forms is problematic as the meaning Wikipedia editors attribute to redirects or anchor texts of internal Wikipedia links can differ from *same as* or *also known as* towards *related to*, *contains*, etc. In this paper we report the results of quality analysis, which has revealed that the accuracy of the surface forms extracted from Wikipedia labels, redirects, disambiguation pages and anchor texts of the internal Wikipedia links, can be as low as 75%, and drops dramatically (to almost 45%) if we consider popular DBpedia entities that have a large number of extracted surface forms. We provide gold standards we have built for the evaluation.

In most of the Wikipedia-based datasets surface forms come with frequency-related scores (e.g. TF-IDF or PMI), but no or little cleaning or filtering is done. We implement filtering approaches to improve the quality of the extracted surface forms, based on (i) string patterns, (ii) interlanguage links and labels from Wikidata, (iii) TF-IDF scores. With these approaches we were able to improve the precision by 10% for the random evaluation dataset and by more than 20% for the dataset of popular pages.

Finally, we extract surface forms from the Common Crawl, the largest publicly available web corpus, where anchor texts of links to Wikipedia pages are the source of surface form strings.

To summarize, the contributions of the paper are as follows

- quality evaluation of Wikipedia-based surface forms, revealing mistakes most of the similar resources have (but ignore);
- filtering approaches that improve significantly the quality of surface forms;
- extracting, filtering and publishing the surface forms based on Wikipedia and Common Crawl, along with the gold standards for their evaluation.

All the datasets are available at `http://data.dws.informatik.uni-mannheim.de/dbpedia/nlp2014/`.

2 Related work

A number of linguistic resources use Wikipedia or DBpedia as a source of surface form strings for an entity. Labels, redirect and disambiguations are used e.g. in the RDM (Redirect Disambiguation Mapping) dictionary [6] or the BabelNet lexical knowledge base [3].

A number of resources complement this information with that extracted from internal Wikipedia links between pages, e.g. AIDA Means dictionary [7], an extended version of the YAGO means relation, the DBpedia Lexicalizations dataset of DBpedia Spotlight [2], or the recent dataset extending the surface forms with multi-lingual labels and resource co-occurrence information [8]. There exist also resources that make a further step, extracting surface forms from the anchor texts of non-Wikipedia web pages into Wikipedia, such as the Google's Crosswikis dataset [5] or Wikilinks [4].

The quality of the above resources is indirectly evaluated through the task of entity linking and disambiguation to Wikipedia pages [6, 2, 9], where the

Wikipedia labels, redirects and anchors are used as features or/and as training data [9]. To the best of our knowledge, no direct quality evaluation of the Wikipedia-based surface forms has been reported, with the exception of the Wikilinks technical report [4], in which they claim to have inspected manually 100 randomly sampled mentions. Some works admit the lack of quality evaluation attributing it to the absence of gold standards [8].

3 Surface Forms from Labels, Redirects and Disambiguations

In this section we report on the extraction and evaluation of surface forms extracted from DBpedia labels, redirects and disambiguations (LRD). We used English DBpedia 2014 data[1], which correspond to English Wikipedia as of May 2, 2014. The following datasets were used as input:

- Labels dataset (labels_en.nt) contains a label (or a title) for each Wikipedia page, including redirects and disambiguation pages.
- Transitive redirects dataset (redirects_transitive_en.nt) contains redirect links between articles in Wikipedia.
- Disambiguations (disambiguations_en.nt) contains information extracted from Wikipedia disambiguation pages.

For each DBpedia entity E, which is not a redirect and not a disambiguation page, we add as surface forms: (i) label corresponding to E, (ii) labels of all pages that redirect to E, (iii) labels of all pages that disambiguate E. So from the following input:

```
dbpedia:Mars rdfs:label "Mars"
dbpedia:4th_planet rdfs:label "4th planet"
dbpedia:Red_Planet_(novel) rdfs:label "Red Planet (novel)"
dbpedia:Red_planet rdfs:label "Red planet"
dbpedia:4th_planet dbpedia-owl:wikiPageRedirects dbpedia:Mars
dbpedia:Red_Planet dbpedia-owl:wikiPageDisambiguates dbpedia:Red_Planet_(novel)
```

4 surface forms will be extracted:

```
dbpedia:Mars "Mars"
dbpedia:Mars "4th planet"
dbpedia:Red_Planet_(novel) "Red Planet (novel)"
dbpedia:Red_Planet_(novel) "Red planet"
```

While combining the three inputs, we filter away lists (e.g. *List_of_Star_Trek_animals*), numbers (e.g. Wikipedia pages for specific years) and single characters (e.g. pages for letters of the English alphabet), which result in about 3% of the entity-surface form pairs. Table 1 provide numbers describing the resulting dataset.

[1] `http://data.dws.informatik.uni-mannheim.de/dbpedia/2014/en/`

Table 1. Surface forms from DBpedia labels, redirects and disambiguations: statistics.

Labels dataset, triples	4,338,802 (36%)
Redirects dataset, triples	6,363,487 (53%)
Disambiguations dataset, triples	1,368,433 (11%)
Combined (LRD) dataset, lines	12,033,641
Filtered LRD dataset, lines	11,662,936

As we already mentioned, a number of linguistic resources extract surface form strings from the same LRD inputs [3, 8], and use them to support entity linking and disambiguation [7, 6, 9, 2]. However, to the best of our knowledge, no direct (as opposed to task-based) evaluation has been reported in the literature. Such an evaluation is hardly needed for surface forms extracted from labels or disambiguations, as the title of a page itself or title of the corresponding disambiguation page is a straightforward way to refer to an entity. But the situation is different for redirects, which, according to Table 1, account for more than 50% of the surface forms in the LRD dataset.

Wikipedia authors often use redirects to connect an entity with a related one that does not have its own Wikipedia page, for instance a character of a book or movie (*Charlie Babbitt* redirects to Rain Man), company division (*Starbucks Australia* redirects to Starbucks), relative of a famous person (*Suri Holmes* redirects to Tom Cruise), an artifact attributed to a person (*Electric Kiss* song is redirected to Lady Gaga), and so on. Such usage is justified by the fact that redirects indeed help you to find information on Wikipedia about the entity in question. But when assumed to be an alternative name or synonym, it can be very misleading for a human user consulting a resource such as BabelNet, or for a text understanding (e.g. entity disambiguation) software.

Looking at some examples of misleading redirects, we have discovered that in Wikipedia they redirect to page *subsections*, and thus the semantics is clearly different from *alternative name*, as e.g. in the case of *Moses the raven* redirecting to Animal_Farm#Other_animals. This information is, however, not present in DBpedia redirect dumps, where *Moses the raven* redirects to Animal_Farm. This extraction feature is responsible for the decrease in surface forms quality in the resources derived from DBpedia [8, 6] and also BabelNet that presumably processes redirects in the same way.

3.1 Gold standard for DBpedia surface forms

To evaluate the quality of Wikipedia redirects, we annotated two subsets of our LRD dataset. For the first subset, referred to as *popular*, we manually selected 34 popular entities (populated places, famous persons, books, movies, companies, planets) in order to focus on pages for which a lot of editing has been done in Wikipedia (and so, lots of redirects added), and which are often referenced from other Wikipedia pages. For the other subset, referred to as *random*, we randomly selected 81 entities each having at least 5 surface forms. Note that

Table 2. DBpedia surface forms from LRD and WAT: annotated data statistics.

	popular	random
Entities	34	81
Entity - SF pairs	2,776	1,027
SFs from LRD	1,251	551
SFs from WAT	1,924	723
SFs from Redirects	1,160	436
SFs from only WAT	1,525	476
SFs from only LRD	852	304
SFs from both WAT and LRD	399	247
Annotations: correct	1,269	770
Annotations: related, contains	198	88
Annotations: related	1,162	219
Annotations: wrong	345	38
Annotations: correct & LRD	836	477
Annotations: correct & redirects	773	365
Annotations: correct & WAT	741	511
Linked from other Wikipedia pages	813,736	14,760

annotation was done for both surface forms extracted from labels, redirects and disambiguations (LRD), and for the surface forms extracted from anchor texts of internal Wikipedia links (WAT), which we will explain in detail in the next section. Table 2 gives statistics on both annotated sets.

Note the different distribution of surface forms in two subsets: in a *random* set, you have a bit less than 13 surface form strings corresponding to an entity, while in the *popular* set this average is 82. The difference in the number of links from internal Wikipedia page is considerable: an entity in the *random* set is linked 182 times on average from other Wikipedia pages, while for the *popular* set the average is 23,933 times.

The surface forms were annotated not only as correct or wrong, but divided into more fine-grained categories in order to better understand the intended semantics of redirects and anchor texts. Each entity-surface form pair was annotated as one of the following:

- correct (*ok* in the data), with the meaning that a given surface form can indeed be used as an alternative name for a corresponding entity (*the eternal city* for Rome or *red planet* for Mars);
- related, contained (*oi*), when the surface form is part of the entity (*Sao Paulo, Brazil* for Brazil or *Google Japan* for Google), or contains (*og*), when the surface form contains the entity (*Turkey* for Istanbul);

- related, type of (g), when the surface form generalizes the entity (*the city* for Rome or *book* for The Da Vinci Code), or partial (p), when the surface form is an ambiguous partial reference to the entity (*Diego* for Diego Maradona)[2];
- related (r), for the numerous case of the related entities (*Google Blog* for Google, *Martian surface temperatures* for Mars);
- wrong (w), for surface forms that do not refer to the entity (*during World War I* for United States); here we further distinguish wrong because of formatting case (f), e.g. surface forms with residual tags.

Table 2 provides statistics on different annotation categories. Both annotated datasets are available for download[3].

The annotated data allows us to get quite surprising numbers regarding the data quality: if we consider just surface forms extracted from redirects, they are annotated as correct in just 66.6% of the cases for the popular and in 84% for the random set, respectively. For the whole LRD data these are 86.5% and 66.8%, respectively. From examples and discussion above it should be clear what the reasons for such relatively low quality are. But can anything be done to improve it? We address this problem in Section 5, after introducing the second part of the surface forms dataset in the following section.

4 Surface Forms from Internal Wikipedia Links

In this section we describe the extraction of surface forms from the anchor texts of internal Wikipedia links. An example of such link can be found in the source code of the Wikipedia page for Berlin

```
[[Frederick I, Elector of Brandenburg|Frederick I]]
```

where another Wikipedia page, Frederick_I,_Elector_of_Brandenburg is referenced with the string *Frederick I* displayed as anchor text.

We have implemented the new extractor for the DBpedia Extraction Framework[4], based on the existing *PageLinksExtractor*, to collect all the anchor texts of such internal links, which produced a dump containing 137,104,653 triples. We then applied a simple strategy to clean the data: residual HTML tags were removed, anchor texts containing just numbers and *list of* substring, and 1-character strings were removed, as well as surface forms equal to *stop word* strings we identified after the first examination of the aggregated data (e.g. *here*, *more*, *click here*, *details*, *see here*, etc.). This way, around 2.2% of the triples were removed, leaving us with 134,044,488 entity-surface form pairs, many of which were present multiple times in the dataset. After aggregation we received 20,362,516 entity-surface form pairs.

[2] Such entries can be considered correct or related depending on the intended use case: for resolving coreference inside one document they are useful, whereas for entity linking (the use case we are more focused on) they mostly add noise.

[3] `http://data.dws.informatik.uni-mannheim.de/dbpedia/nlp2014/gold/`

[4] `https://github.com/dbpedia/extraction-framework`

Table 3. DBpedia surface forms: statistics for LRD & WAT combined dataset on different TF-IDF (t) thresholds

	LRD	WAT	LRD&WAT	t = 1.8	t = 2.6
Entities	4,515,847	4,323,321	4,515,847		
Entity - SF pairs	11,662,936	12,776,813	18,017,646	16,712,471	13,308,358
Unique SFs	10,551,495	10,115,077	14,564,295	14,445,109	11,557,242
SFs from only WAT			6,354,709	5,133,868	1,729,755
SFs from only LRD			5,240,833	5,180,634	5,180,634
SFs from both WAT&LRD			6,422,104	6,397,969	6,397,969

The next step was to resolve redirects and to filter away entities that do not have a corresponding page in Wikipedia. The latter happens because a Wikipedia editor is free to add a link to a non-existent Wikipedia page, also called "red link"[5]. To filter away red links, only the entities contained in (or redirected to) the LRD dataset, presented in the previous section, were included into the aggregated dataset. The intersection gave us 14,116,929 pairs, and resolving redirects further reduced the dataset to 12,776,813 pairs.

We then calculated the TF-IDF scores for each entity-surface form pair, according to the following formula:

$$TFIDF(E, SF) = log_{10}(count(E, SF) + 1) * log_{10}(\#ofE/count(SF)),$$

where $\#ofE$ is the total number of unique DBpedia entities in the dataset, $count(E, SF)$ is the number of times E is referred to with SF, and $count(SF)$ is the number of times SF is present in the dataset.

The two datasets, the one extracted from labels, redirects and disambiguations, and the one derived from the anchor texts of internal Wikipedia links, are then combined. The statistic is given in Table 3. In total, 18,017,646 entity-surface form pairs for 4,515,847 entities are extracted, with 35.3% coming from anchor texts, 29% from the LRD dataset, and 35.7% extracted from both sources. On average, there are around 4 surface forms per entity.

The two entity subsets for annotation described in the previous section were sampled from this combined dataset, see the summary in Table 2. As for the case of the LRD dataset, we can conclude from Table 2 that for the WAT surface forms, the correct annotation is in place in just 38.5% of the cases for the popular and in 70.7% for the random subset, respectively. In the next section, we address the problem of the (low) surface forms quality.

5 Cleaning and Filtering

For the surface forms extracted from anchor texts of internal Wikipedia links we have frequency counts and TF-IDF scores, which will be the basis for filtering away irrelevant entries. But for the most part of the LRD dataset, these scores

[5] `https://en.wikipedia.org/wiki/Wikipedia:Red_link`

Table 4. Filtering with string patterns and Wikidata: results

LRD	**Popular**			**Random**		
	P	R	F1	P	R	F1
no filtering	0.6683	1.0000	0.8011	0.8657	1.0000	0.9280
just patterns	0.7058	0.9928	0.8250	0.8679	0.9916	0.9256
patterns + wikidata	0.7088	0.9928	0.8271	0.8741	0.9895	0.9282
patterns + wikidata*	**0.7147**	0.9916	0.8307	**0.8820**	0.9874	0.9318
ALL	**Popular**			**Random**		
	P	R	F1	P	R	F1
no filtering	0.4571	1.0000	0.6274	0.7498	1.0000	0.8570
just patterns	0.4704	0.9890	0.6375	0.7515	0.9935	0.8557
patterns + wikidata	0.4729	0.9882	0.6396	0.7542	0.9922	0.8570
patterns + wikidata*	**0.4758**	0.9835	0.6413	**0.7575**	0.9896	0.8581

are not available, and therefore we came up with two filtering strategies we present below.

The first strategy is filtering based on **string patterns**. After examining the data, we found out that surface forms annotated as *related* or *wrong* often follow one of the patterns below:

- URLs: contain .com or .net (*Berlin-china.net* surface form for Berlin);
- of-phrases: contain "of "+entity, with the exceptions for *city of*, *state of*, etc. (*Issues of Toronto* for Toronto);
- in-phrases: contain "in "+entity (*Historical sites in berlin* for Berlin);
- and-phrases: contain "and "+entity or entity + " and" (*Tom Cruise and Katie Holmes* for Tom Cruise);
- list-of: contain *list of* (*List of Toronto MPs and MPPs* for Toronto).

We filtered both popular and random annotated datasets using the above patterns. Results are presented in Table 4, where we report precision (P), recall (R) and F-measure (F1) not only for popular and random sets (ALL), but for subsets containing just surface forms coming from labels, redirects and disambiguations (LRD). Pattern-based filtering removes irrelevant surface forms thus improving the datasets precision: for the popular dataset the increase in precision is 1.33% for the whole dataset and 3.75% for its LRD subset. For the random dataset, the increase is less than 1% due to the much lower number of redirects and thus, less noise in them.

Our second filtering strategy is based on the observation that some surface forms annotated as relevant, e.g. city suburbs, are entities on their own and can have corresponding Wikipedia pages **in other languages**. That is, for instance, the case for *Neckarau* city area of Mannheim, which redirects to Mannheim in English Wikipedia, but has its own page in German Wikipedia.

To implement this strategy, we make use of DBpedia-based RDF dumps for Wikidata released in May 2015[6]. Wikidata is meant to serve as an integrated

[6] `http://wikidata.dbpedia.org/downloads/20150330/`

structured source of multi-lingual Wikipedia data, so for each Wikidata entity, links to all existing Wikipedia pages and multi-lingual labels are available. We use *labels-mappingswiki* and *sameas-all-wikis* dumps, and, for each surface form in our dataset, check whether it corresponds to any of the labels of Wikidata entities that do not have English but have Wikipedia pages in other languages.

We do not apply both pattern- and Wikidata-based filtering to surface forms that have TF-IDF scores greater than a threshold, which we defined to be equal to 5.0. The annotated data is not big enough to cover many of such cases – surface forms that have TF-IDF scores *and* are considered irrelevant by one of the filtering approaches – therefore, we are not able to learn this threshold.

We applied the Wikidata-based filtering, combined with the pattern-based one, on both popular and random annotated datasets. The results are reported in Table 4 (*patterns + wikidata*). As can be seen, the precision does not increase significantly, therefore, we have implemented an extension of the Wikidata-based filtering strategy, in which for each surface form we

- check whether it exactly matches any of the labels of Wikidata entities that do not have English but have other Wikipedia pages;
- if no match found, check whether the surface form is close (Levenshtein distance < 4) to any of the surface forms already discarded for this entity;
- if no match found, check whether the surface form is close (Levenshtein distance < 3) to any of the labels of Wikidata entities that do not have English but have other Wikipedia pages.

Combined with the pattern-based filtering, the results are reported in Table 4 (*patterns + wikidata**). The increase in precision is 0.5% compared to pattern-based filtering. The gain is higher (1.5%) for just LRD data, as the surface forms that can be filtered by applying this strategy most often come from redirects.

5.1 Filtering based on TF-IDF scores

Finally, we use TF-IDF scores to filter surface forms, targeting strings extracted from anchor texts of inter-page Wikipedia links. We worked with the filtered versions of the popular and random annotated datasets, after the application of the *patterns + wikidata** strategy. To determine the threshold at which a surface form is considered irrelevant, we combined the (filtered) popular and random annotated datasets, and randomly split the combined set into training (containing 2/3 of the data) and testing (1/3 of the data). Thresholds for TF-IDF values from 1.0 to 8.0 with a step of 0.2 were evaluated.

Table 5 presents the results of this experiment. Two thresholds were selected based on the training set, that correspond to the two highest values of F1: 1.8 and 2.6. At the first threshold, the recall is still very high, while at the second threshold the trade-off shifts towards the better precision (and still acceptable recall). Threshold equal to 0.0 can be seen as a baseline where no filtering was applied. On the test dataset, the learned thresholds were also corresponding to the two maximum F1 values.

Table 5. Filtering with TF-IDF thresholds: results

	threshold	P	R	F1		threshold	P	R	F1
training	0.0	0.5504	1.0000	0.7100	**test**	0.0	0.5608	1.0000	0.7186
	1.8	0.6123	0.9662	0.7496		1.8	0.6254	0.9676	0.7597
	2.6	0.7234	0.7718	0.7468		2.6	0.7288	0.7729	0.7502
popular	0.0	0.4758	1.0000	0.6448	**random**	0.0	0.7575	1.0000	0.8620
	1.8	0.5468	0.9736	0.7003		1.8	0.7836	0.9554	0.8610
	2.6	**0.6646**	0.7813	0.7182		2.6	**0.8574**	0.7572	0.8042

We applied the obtained thresholds to the popular and random annotated sets – even if they make part of the training data – to compare the precision to its initial values. If your NLP task requires high precision, the second threshold (2.6) should be preferred. As can be seen from Table 5, the second threshold gives 66.5% precision on the popular dataset – compare to 47.6% after pattern- and Wikidata-based filtering, and to 45.7% before any filtering. For the random set, the precision numbers are 85.7% with all three filtering approaches combined, 75.8% with pattern- and Wikidata-based and 75% with no filtering. That is, we were able to increase precision by more than 20% for popular entities, and by more than 10% for the randomly selected entities.

We applied pattern- and Wikidata-based filtering, as well as filtering based on TF-IDF thresholds to the whole WAT&LRD surface forms corpus. The statistics for the resulting datasets are reported in Table 3. All the resulting corpora are available for download[7], includng the non-filtered version.

6 Non-Wikipedia Links Surface Forms Dataset

In this section we present the surface form dataset extracted from the Common Crawl (CC), the largest publicly available web corpus, namely, from the Winter 2014 CC Corpus[8]. The extraction was done in the context of the Web Data Commons project[9], which extracts and provides for public download various types of structured data from the Common Crawl.

Links containing the *wikipedia.* substring were extracted, excluding links from one Wikipedia page to another. The data is available for download[10], and contain 95,271,410 links along with anchor texts (or 14.5GB, unzipped).

Firstly, cleaning (removing HTML tags, replacing spaces in links, etc.) and the following filtering steps were performed:

[7] `http://data.dws.informatik.uni-mannheim.de/dbpedia/nlp2014/lrd-wat/`

[8] `http://blog.commoncrawl.org/2015/01/december-2014-crawl-archive-available/`

[9] `http://webdatacommons.org/`

[10] `http://data.dws.informatik.uni-mannheim.de/structureddata/2014-12/wikianchor/`. We thank Robert Meusel for performing the data extraction.

Table 6. Filtering surface forms extracted from the Common Crawl according to pattern- and Wikidata-based strategies (pt-wd) and different TF-IDF (t) thresholds

	non-filtered	pt-wd	t=1.8	t=2.6	t=3.8
Entities	1,900,205	1,899,203	1,712,581	1,621,070	1,583,139
Entity - SF pairs	5,154,859	5,128,800	4,132,425	2,879,051	2,366,801
Unique SFs	3,963,622	3,940,796	3,920,413	2,735,136	2,260,560
SFs from LRD	2,100,610	2,093,836	2,093,835	2,093,835	2,093,835
SFs from only Common Crawl (CC)	3,054,249	3,034,964	2,038,590	785,216	272,966
SFs from both CC and LRD	2,100,610	2,093,836	2,093,835	2,093,835	2,093,835

- removing links to non-English Wikipedia pages, administrative pages (Help and User pages, User Talks, Wikipedia Files), pages that have numbers as titles, page sections (containing #);
- removing "red links", i.e. leaving only links to entities contained in LRD corpus; resolving redirects;
- removing anchor text strings that are links themselves, as well as strings longer than 100 characters.

As a result, we got 46,263,819 links (51.5% filtered away), with the dump size reduced to 1.5GB. Further cleaning (removing category pages, lists, numeric or 1-character surface forms strings) and aggregation resulted in 5,780,211 entity-surface form pairs for 1,973,873 entities. Out of these 5.8 mln surface forms, 2,100,61 intersect with the LRD corpus, and 3,679,602 come from the Common Crawl corpus. On this corpus, the TF-IDF scores were calculated.

If we exclude surface forms already present in the WAT dataset, we are left with 5,154,859 surface forms, from which 3,054,249 surface forms come from the Common Crawl corpus (see the 1st column of Table 6). That is, we add 3 mln surface forms to the WAT&LRD corpus introduced in the previous sections. In Table 6 we report the size of the corpus after applying the filtering strategies based on string patterns and Wikidata, presented in previous sections, as well as filtering with respect to TF-IDF scores.

We do not report precision and recall values for the CC-based corpus, as we do not have gold standard data for it, leaving it for future work. Quick inspection of several entities shows that the quality of the CC surface forms is lower than the one for LRD&WAT corpus. The reason might lie in the crawling strategy used by Common Crawl, and needs to be further investigated. The data is available for download[11], including the corpus filtered against the higher threshold of 3.8.

7 Conclusion

In this paper we addressed the largely overseen problem of the quality of surface forms extracted from Wikipedia. We found out that the dataset precision is just about 75% for a random data sample, and almost 45% if we consider popular

[11] `http://data.dws.informatik.uni-mannheim.de/dbpedia/nlp2014/lrd-cc/`

DBpedia entities. Three filtering strategies we implemented allowed for significant precision improvements, and are based on string patterns, cross-language information from Wikidata and scores based on anchor texts frequencies. We extract an additional surface form dataset from the links to Wikipedia from the Common Crawl web corpus.

We make all the datasets, including gold standard used for the evaluation, publicly available, and encourage their use in NLP or other tasks. Depending on requirements, different datasets from those we provide might be relevant: in case high precision is crucial, the LRD&WAT surface forms filtered at TF-IDF threshold of 2.6 are an option, while for tasks requiring high recall the combination of LRD&WAT with CC surface forms filtered at a lower threshold might be preferred.

As for future work directions, the evaluation of the quality of the filtered corpus in the task of entity linking is to follow, as well as further work on the gold standard. The resource as well as the filtering approaches can further benefit from using the labels from other DBpedia language editions as surface forms.

References

1. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal*, 2014.
2. Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, 2011.
3. Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.
4. Sameer Singh, Fernando Pereira, and Andrew Mccallum. Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. Technical report, 2012.
5. Valentin I. Spitkovsky and Angel X. Chang. A cross-lingual dictionary for english wikipedia concepts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3168–3175, 2012.
6. Nadine Steinmetz, Magnus Knuth, and Harald Sack. Statistical analyses of named entity disambiguation benchmarks. In *Proceedings of the NLP & DBpedia workshop @ ISWC 2013*, 2013.
7. Mohamed Amir Yosef, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum. AIDA: an online tool for accurate disambiguation of named entities in text and tables. *PVLDB*, 4(12):1450–1453, 2011.
8. Lei Zhang, Achim Rettinger, and Steffen Thoma. Bridging the gap between cross-lingual nlp and dbpedia by exploiting wikipedia. In *Proceedings of the NLP&DBpedia workshop @ ISWC 2014*, 2014.
9. Yiping Zhou, Lan Nie, Omid Rouhani-Kalleh, Flavian Vasile, and Scott Gaffney. Resolving surface forms to wikipedia topics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1335–1343, 2010.