Chapter 4

SocialLink Resource

In this chapter, we present a second core contribution of this thesis — the SocialLink dataset, a publicly available Linked Open Data dataset that contains links between the social media accounts on Twitter and the corresponding entities in multiple language chapters of DBpedia. By following the approach presented in Chapter 3 we are able to produce a significant number of alignments to enable the abovementioned knowledge transfer between the two worlds. As a result, on the one hand, we support Semantic Web practitioners in better harvesting the vast amounts of valuable, up-to-date information available in Twitter; on the other hand, the resource presented here permits Social Media researchers to leverage DBpedia data with little effort when processing the noisy, semi-structured data of Twitter.

This resource is part of the SocialLink project and is periodically updated with releases. The code along with the gold standard dataset used to produce it are made available as part of our open source project. This chapter contains details on the design choices, various dataset statistics and provides discussion about some of the general use cases that showcase the above-mentioned knowledge transfer.

4.1 Introduction

The number of existing links to social media for the living people and organizations in the Linked Open Data cloud is very low: the DBpedia 2016-04 that we currently employ in our experiments contains just 56,133 of them. In order to enable the knowledge transfer between the LOD and the social media, the coverage has to be significantly improved. In Chapter 3, we have introduced a scalable approach that can potentially help to fill this gap by providing a significant amount of high-quality links automatically, learning from the existing ones.

In this chapter, we present the SocialLink dataset,[1] a publicly available Linked Open Data dataset based on our state-of-the-art linking approach that matches social media accounts on Twitter to their corresponding entities in DBpedia. Over the last two years we have released three major versions of this dataset. The latest version, *v*3.0, consists of almost 322K high quality (more than 90% precision) alignments, obtained by applying the above linking approach to 2M living people and 500K currently existing organizations in DBpedia. Entities from 128 DBpedia language chapters are considered: while the textual features we extract are mainly designed to work with Western languages, the linking approach employs a wide variety of different feature types allowing us to target entities in other languages as well. Additionally, the dataset contains raw scores for each candidate alignment allowing end users to tune the precision / recall balance as they see fit, potentially obtaining up to 1M of alignments. The dataset is available in many different formats including RDF/OWL, which we then distribute in accordance with LOD best practices (Wilkinson et al., 2016), reusing existing vocabularies and providing a live SPARQL endpoint.

By covering such amount of entities, SocialLink dataset indeed creates a bridge between the highly structured LOD cloud and the vibrant and up-to-date social media world. SocialLink dataset serves two main purposes. On the one hand, it aims at facilitating social media processing by leveraging DBpedia data, e.g., as a source of ground truth properties for training supervised systems for user profiling, or as contextual data in natural language understanding tasks (e.g., Named Entity Linking) operating on social media contents (Corcoglioniti et al., 2016; Minard et al., 2016). On the other hand, SocialLink gives Semantic Web practitioners the ability to populate KBs with up-to-date data from social media accounts of DBpedia entities, such as structured attributes, images, connections, user locations, and descriptions. To the best of our knowledge, SocialLink dataset is unique in the alignment task it addresses providing more than tenfold increase in the number of links to Twitter in the LOD.

SocialLink project focuses on linking living people and organizations. There are two main reasons to that. Firstly, those entities of those types constitute the vast majority of entities that can be reasonably aligned to a social media profile. Indeed, social media were designed to accommodate just living people initially. With such a lucrative commercial opportunity that social media represent, various organizations started to cultivate presence there for themselves too. Both public people and organizations use social media to engage with their audience and potential customers, share relevant information and promote their products and services. Secondly, such entity types constitute the overwhelming majority of existing links between DBpedia and Twitter. While SocialLink approach does support other

[1] `http://w3id.org/sociallink` — Creative Commons Attribution license (CC BY 4.0).

entity types (our preprocessing pipeline and features are entity type-agnostic), as shown in Chapter 3, there are vast differences in linking quality based on entity type: linking of organizations compared to persons exhibit up to 18% lower F_1 using the same approach. The absence of the significant number of training samples will inevitably strengthen this issue. Because of that we decided to limit the scope of the resource to only two entity types.

The first version of SocialLink dataset, $v0.1$-alpha, was released in mid 2016 using the supervised alignment approach described in Nechaev et al. (2017b). Since then, we have significantly expanded its scope and alleviated some of the restrictions of the original system. To name a few, the approach is no longer restricted by the limits of Twitter REST API and is now able to use entity data from 128 DBpedia chapters, allowing us to align DBpedia entities present only in localized DBpedia chapters, and to provide more context to our matching algorithm, helping with disambiguation and increasing the amount of processed entities by a factor of three. The SocialLink pipeline generating the dataset is available open source[2] along with the revised gold standard dataset used to train and evaluate the system. We are able to repopulate the dataset in an automatic way covering the latest data and algorithm improvements to insure that alignments are up-to-date. Relevant statistics and the latest dataset version can be found on our website and Zenodo.[3] At the time of writing, we have released three major versions of the resource, each corresponding to a major milestone in SocialLink development (Nechaev et al., 2017b,d, 2018b). The SPARQL endpoint available on our website always contains the latest public release of our resource.

In the remainder of the chapter, Sections 4.2 and 4.3 present respectively the SocialLink pipeline used to produce the resource and the latest version of the SocialLink dataset. Section 4.4 discusses some of the scenarios where the dataset has been or can be used, while Section 4.5 concludes.

4.2 SocialLink Pipeline

The SocialLink dataset population procedure follows the same three-phase pipeline described in Chapter 3. Briefly, processing starts with the *data acquisition* phase, where the required Twitter and DBpedia data, including preexisting gold standard alignments from DBpedia, are gathered, prepared and indexed locally for further processing. Next, in the *candidate acquisition* phase, for each DBpedia entity, a list of candidate matching Twitter profiles is obtained by querying the indexes. Finally, the *candidate selection* phase uses the gold standard alignments to train a Deep Neural Network (DNN) that scores and selects the

[2] http://github.com/Remper/sociallink

[3] https://doi.org/10.5281/zenodo.1451797

best matching candidate. The system may abstain if there is no suitable candidate. After an entity passes through this pipeline, it is ready to be added to the resource.

In this section, we provide additional details on the linking approach presented in Chapter 3 relevant to the population of the resource. Namely, we discuss coverage issues with some of the feature families, provide details on employing the SocialLink approach for large multi-language KBs and highlight pipeline differences between different versions of the SocialLink dataset. In total, there has been five releases of the resource available on our website[4] including the three major ones covered in the respective publications. The first two versions, $v0.1$-alpha and $v0.5$-beta, were produced during initial experiments and implementation of the base pipeline. The first major version, $v1.0$, uses the approach described in Nechaev et al. (2017b) and is the first release available on Zenodo. The $v2.0$ covers improvements made in Nechaev et al. (2017d), while the most recent release, $v3.0$, uses the updated approach from Nechaev et al. (2018b). In this section, we will refer to $v1.0$ and $v2.0$ as Legacy and $v3.0$ as Current.

4.2.1 Feature Coverage

The Legacy versions of SocialLink employ a simplified BASE feature set, while Current version benefit from the latest BASE_KB_SG_TL system (see Section 3.5). The introduction of graph-based features to Current has significantly improved the performance of our linking approach. However, when applied to entities outside of the gold standard, the coverage aspect of the new features has to be considered. While our Knowledge Base embeddings (Cochez et al., 2017b) cover almost 9M entities of DBpedia, they consider only the English chapter, while the SocialLink dataset covers 128 different language chapters. This yielded $68,9\%$ coverage during population of Current. For the experiments described in this thesis, we consider such coverage sufficient. However, in future a joint embedding can be trained exploiting owl:sameAs links across language chapters to avoid significant drops in coverage. The same can be observed in social media: we approximate the social graph from the incomplete stream of tweets, inevitably losing perfect coverage. The latest version of our approximated social graph contains data for 168M users which covers $65,3\%$ of candidates in Current. Social graph embedding is acquired from the precomputed embeddings of friends, so the process fails if and only if the target user hasn't been observed interacting with any of the known users as sampled from Twitter Streaming API. When we calculate graph-based features, in case where the graph-based features are not available, we default to an average embedding vector for the respective subspaces: the average of all entity embeddings from the KB side and the average of the most popular users from the social media side.

[4] http://w3id.org/sociallink#download

Our BASE (see Table 3.3) feature family can theoretically include features with imperfect coverage as well. For instance, textual features, such as similarities based on names, descriptions and tweets, and profile features default to zeros at training and evaluation time in case the profile object for the candidate was not found in our index. However, for the resource population, we exclude all candidates that did not have at least a profile object in our database always providing at least a minimal textual context and important structured attributes, such as names. The *homepage links* feature family containing 327K candidate-entity pairs has limited coverage as well: it was crawled in mid 2016 based on entities from 2015-10 edition of DBpedia and will become more stale as the time goes by. As this feature family is hard to update and it introduces limited performance improvements, we plan to replace it with a more scalable set of features based on machine-readable properties in next versions of SocialLink. For example, feature families described in Chapter 5 can be employed instead.

Besides the explicit coverage issues, text-based features can discard the data if the input does not match the vocabulary. Firstly, the input tokens are produced based on a simple multi-language stemming and tokenization procedure. It performs well on most Western languages but is weak when applied to Asian and Arabic languages. Secondly, the vocabulary of the produced tokens itself is limited. In Legacy versions, the LSA-based model was employed having $972,001$ most frequent tokens as seen in six-language Wikipedia-based corpus (Aprosio et al., 2013). In Current, we employ three different language models described in Section 3.7, which significantly expands the available vocabulary. However, its performance is still heavily biased towards English and similar languages.

4.2.2 Scoring and Selection Procedures

During evaluation of the SocialLink approach, we employ five-fold cross-validation to provide accurate assessment of the algorithm performance. In order to produce the resource, we acquire pairwise (entity-candidate) predictions from each fold. Then, to ensure stable predictions, we implement a basic ensemble of the models from each fold by averaging the pairwise scores across all folds. After this averaging procedure, for each entity the algorithm has to either select the best suitable candidate (candidate with the largest score) or abstain from selection. To do that, Legacy versions employed two predefined thresholds: *minimum score* required to consider an alignment correct and *minimum improvement* over the second best pick. The latter ensures that the algorithm can abstain if two or more candidates are indistinguishable, even if they pass the minimum score requirement. The thresholds can be selected based on the precision/recall curve produced from the evaluation on a gold standard as shown in Figure 3.4.

For example, for the $v2.0$ release, the thresholds were optimized for the desired precision of 90% and set to 0.4 minimum score and 0.4 minimum improvement, leading to 41% recall of generated alignments (Nechaev et al., 2017d). For the older $v1.0$ release, the F1-optimized setting was used yielding 85% precision and 52% recall (Nechaev et al., 2017b). Since different downstream tasks can impose different precision requirements from alignments, we always provide the raw scores for each entity to enable the user to select appropriate thresholds based on those requirements.

The Current ($v3.0$) release employs an alternative strategy described in Section 3.5 alleviating the need for the *minimum improvement* threshold by rescaling the scores to a proper probability-like distribution considering each candidate plus abstention as possible outcomes. The rescaling approach heavily punishes the scores for entities that are highly ambiguous and has almost no effect in cases where there is a single good choice. The minimum score threshold is still selected based on the evaluation on a gold standard but has one less parameter to tune. As in $v2.0$, we have selected 90% precision target corresponding to 0.28 threshold yielding 55% recall (with 322 307 total alignments) this time due to algorithm and pipeline improvements we have introduced. Have we selected the 0.4 threshold of $v2.0$ for Current release, we would have aligned 248 349 entities corresponding to expected 93% precision and 51% recall. While the new scoring procedure is much more conservative than the previous one leading to more abstentions on ambiguous cases, it simplifies the choice of threshold for the end user and provides more reliable alignments overall.

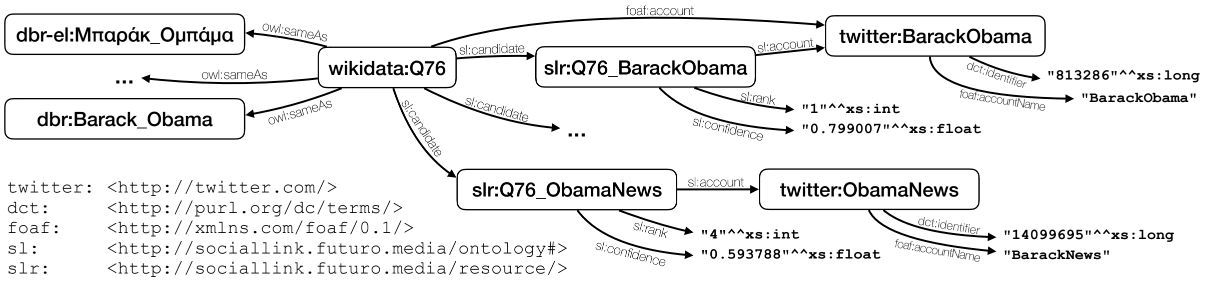
4.2.3 Populating the Resource

The population of the resource follows the general workflow of the SocialLink pipeline. First, the updated data from the social media has to be consolidated. This data (3.3TB at the time of writing) is updated several times during the year and is indexed using a number of Apache Flink[5] pipelines. The resulting index is stored in a PostgreSQL database (747GB). The database schema is available in our repository[6]. Due to Twitter terms of use and privacy concerns we release neither the raw Twitter Streaming API data we used nor the preprocessed index. The user index population typically takes six days to fully complete on our hardware (single Xeon E5-2630v4 with 192GB of RAM). From the KB side, we exploit the same DBpedia index detailed in Section 3.3.1.

Then the resource is produced by first computing the list of candidates for each entity and then scoring each pair independently. The Current version has 14 011 602 of such pairs. Then, based on the selected threshold, the resource is produced in different formats

[5] https://flink.apache.org/

[6] https://github.com/Remper/sociallink/blob/master/alignments/src/main/resources/schema.sql



```
twitter:  <http://twitter.com/>
dct:      <http://purl.org/dc/terms/>
foaf:     <http://xmlns.com/foaf/0.1/>
sl:       <http://sociallink.futuro.media/ontology#>
slr:      <http://sociallink.futuro.media/resource/>
```

Figure 4.1. Representation of alignments in RDF.

detailed in Section 4.3 and on our website. The code for generating the dataset is written in Java and Python as part of our open source project and the documentation for running the pipelines is available online.[7] Additionally, in Chapter 7, we discuss Social Media Toolkit (SMT). Complementary to its entity linking capabilities, SMT acts as a web-based test bench allowing us to debug and validate the resulting SocialLink dataset and to deploy it in production via the API.

4.3 SocialLink Dataset

As mentioned before, the result of running the SocialLink pipeline is the SocialLink dataset, that we are able to generate periodically to account for algorithm updates and new data in DBpedia and Twitter. The dataset is distributed in different formats, with RDF being the main one, including the high quality alignments as well as all the intermediate candidate data. We describe here the modeling choices behind the RDF format of the SocialLink dataset, summarizing the statistics of its three main releases ($v1.0$, $v2.0$, $v3.0$) and discussing how the dataset is made available online and kept up-to-date.

4.3.1 RDF Format

We encode our alignments in RDF using terms from FOAF,[8] Dublin Core Terms,[9] and our custom SocialLink vocabulary (prefix `sl`),[10] as exemplified in Figure 4.1.

DBpedia entities are referenced using canonical URIs possibly taken from Wikidata, like wikidata:Q76 for entity Barack Obama. Each canonical URI has owl:sameAs links to itself and to corresponding URIs in other DBpedia chapters (based on gathered DBpedia data), allowing querying the dataset using localized entity URIs.

[7] https://github.com/Remper/sociallink/wiki/align
[8] http://xmlns.com/foaf/0.1/ (prefix: foaf).
[9] http://purl.org/dc/terms/ (prefix: dct).
[10] http://sociallink.futuro.media/ontology# (prefix: sl).

Table 4.1. Number of entities considered in different versions of SocialLink.

Entity type	Entities in DBpedia		Entities considered (per version)	
	2015-10$_{\text{en only}}$	2016-04$_{\text{multi-lang}}$	v1.0	v2.0 / v3.0
dbo:Person	1 365 651	2 975 645	702 530 (51.4%)	2 035 590 (68.4%)
dbo:Athlete	372 424	493 867	252 268 (67.7%)	412 629 (83.6%)
dbo:Artist	146 759	269 745	84 147 (57.3%)	188 095 (69.7%)
dbo:Politician	67 074	123 460	29 750 (44.4%)	65 135 (52.8%)
dbo:Writer	41 978	69 753	19 055 (45.4%)	37 744 (54.1%)
dbo:Scientist	35 851	64 005	13 634 (38.0%)	28 854 (45.0%)
dbo:SportsManager	17 857	18 281	13 477 (75.7%)	14 860 (81.3%)
dbo:Coach	8 452	8 772	5 142 (60.8%)	5 643 (64.3%)
dbo:Model	2 824	7 601	2 470 (87.5%)	7 470 (98.3%)
dbo:Journalist	2 331	4 019	1 647 (70.7%)	3 285 (81.7%)
dbo:Presenter	821	4 898	643 (78.3%)	4 674 (95.4%)
dbo:Organisation	346 083	575 644	171 187 (49.5%)	531 177 (92.3%)
dbo:Company	75 398	131 056	45 384 (60.2%)	119 365 (91.1%)
dbo:EducationalInst.	62 407	116 139	28 936 (46.4%)	108 353 (93.3%)
dbo:Group	42 056	66 868	31 938 (75.9%)	61 620 (92.2%)
dbo:SportsTeam	41 227	62 221	23 107 (56.1%)	59 257 (95.2%)
dbo:Broadcaster	29 595	35 394	13 151 (44.4%)	35 028 (99.0%)
dbo:MilitaryUnit	19 673	36 151	13 243 (67.3%)	34 574 (95.6%)
dbo:PoliticalParty	12 480	16 611	2 367 (19.0%)	13 767 (82.9%)
dbo:GovernmentAgency	8 461	9 634	1 654 (19.6%)	8 647 (89.8%)
dbo:Non-ProfitOrg.	6 035	8 109	2 354 (39.0%)	7 772 (95.8%)
dbo:TradeUnion	1 773	2 032	153 (8.6%)	2 027 (99.8%)
All entities	1 711 734	3 551 289	873 717 (51.0%)	2 566 767 (72.3%)

Twitter accounts, like twitter:BarackObama, are modeled as foaf:OnlineAccount individuals, using properties foaf:accountName and dct:identifier to respectively encode the account screen name and numeric identifier (useful in applications).

The alignment between a DBpedia entity and the corresponding Twitter account is expressed using property foaf:account. In addition, individuals of type sl:Candidate (e.g., slr:Q76_BarackObama in Figure 4.1) reify the many-to-many relation between DBpedia entities and candidate Twitter accounts, linked via properties sl:candidate and sl:account. This reified relation is enriched with properties sl:confidence and sl:rank encoding the candidate confidence score (i.e., estimated correctness probability) and its rank among the candidates for the entity, to simplify querying for the top candidate.

Based on this modeling, the following SPARQL query retrieves the Twitter account (if any) aligned to an entity identified by any of its localized DBpedia URIs <E>:

```
SELECT ?account {?e owl:sameAs <E>; foaf:onlineAccount ?account}
```

4.3.2 Dataset Statistics

Table 4.1 reports the number of entities we have considered over time. In the first main release of the SocialLink dataset, $v1.0$, only English DBpedia (release 2015-10) was considered, providing 1 771 734 potential entities of types dbo:Person and Organization to align. Given our definition of live person and organization at the time, we have filtered out 49% of those entities ending up with just 873 717 potential targets for alignment. For subsequent releases, including the latest one, we have significantly expanded the scope of the SocialLink dataset taking 128 DBpedia language chapters of the 2016-04 release as the source. The usage of multi-lingual input has increased the number of target entities threefold. It is worth mentioning that we have also improved the filtering conditions employed to discard non-alive entities, which is the main reason behind the percentage-wise increase in considered entities versus the total (from 51% to 72,3%). Another reason is the coverage issue for smaller chapters of DBpedia where for many entities the properties that would indicate its "alive" status are not filled. Additionally, Table 4.1 indicates the number of entities considered belonging to some of the most populous types: for persons we mainly consider athletes, for organizations it is companies.

More importantly, each release of the SocialLink dataset provided different amounts of data. Table 4.2 showcases this difference by detailing (i) the number of entities with candidates, (ii) average number of candidates per entity and (iii) the number of high quality alignments produced for each version of SocialLink dataset. In the first release, $v1.0$, just over 620K entities had at least one candidate, out of which 304K high quality alignments were produced. The redesigned pipeline of $v2.0$ relying on our custom-built index, expanded list of target entities and the refined approach was able to provide candidates for 906K entities. While it is larger than $v1.0$, percentage-wise from the number of considered entities, it is significantly lower (from 71% of all entities to just 40%). This is mainly due to candidate acquisition being more precise in populating the candidates: Twitter API search used in $v1.0$ would try to respond with at least some candidates even if they are unlikely to be a match based on a name, while our user index would require at least a partial name match. Finally, $v2.0$ contains just under 272K high quality alignments, which is an 11% decrease from $v1.0$. This is mainly due to a more strict precision target (90% compared to 85% of $v1.0$ as discussed in Section 4.2) and a more conservative approach overall. Our latest release, version 3.0, improves the overall quality of the pipeline by introducing algorithm improvements at each phase detailed in Chapter 3. This enabled both the increased number of entities with candidates and the increase of 18,5% in high quality alignments (322 124) compared to $v2.0$ without changing the precision requirements.

4.3.3 Availability and Sustainability

The SocialLink dataset is indexed on DataHub[11] and is available for download on SocialLink website, together with VOID (Alexander et al., 2009) statistics, old dataset releases, the gold standard (encoded using the same RDF representation), and non-RDF versions of alignments (JSON, TSV, no intermediate candidate data). Canonical citations (DOIs) for the dataset are available via Springer Nature (Nechaev et al., 2017c) ($v2.0$) and Zenodo (Nechaev et al., 2018a) (all releases) digital repositories. Alignment data is also available and queryable by end users and applications via a publicly accessible SPARQL endpoint[12] using Virtuoso. The SocialLink vocabulary is published according to LOD best practices, and both vocabulary and data URIs are dereferenceable with support of content negotiation.

Extensive documentation is available via the website, covering: (i) dataset scope, format, statistics, and access mechanisms; (ii) instructions for deploying and running the SocialLink pipeline to recreate the resource; (iii) example applications using the dataset; and, (iv) links to external resources like the GitHub repository and issue tracker.

The main requirement for generating the SocialLink dataset is the collection of (at least) some months of raw data from the Twitter Streaming API, e.g., via our data acquisition components. We run a SocialLink pipeline on our premises to continuously collect this data and sustain the periodic update of the dataset. No code modifications are foreseen unless breaking changes occurs in formats and APIs of Twitter and DBpedia.

[11] http://datahub.io/dataset/sociallink

[12] http://sociallink.futuro.media/sparql

Table 4.2. Alignment statistics in different versions of SocialLink. Percentages are calculated from the number of entities considered for each version as reported in Table 4.1.

Entity type	Entities with candidates			Cand. / entity			Alignments produced		
	*v*1.0	*v*2.0	*v*3.0	*v*1.0	*v*2.0	*v*3.0	*v*1.0	*v*2.0	*v*3.0
dbo:Person	524 251 (74,6%)	737 017 (36.2%)	836 490 (41.1%)	7.0	12.6	13.5	246 732 (35.1%)	234 450 (11.5%)	260 628 (12.8%)
dbo:Athlete	187 748 (74.4%)	214 070 (51.9%)	231 036 (56.0%)	7.0	15.1	15.4	80 727 (32.0%)	71 935 (17.4%)	67 266 (16.3%)
dbo:Artist	72 242 (85.9%)	104 614 (55.6%)	116 260 (61.8%)	7.4	12.3	14.8	43 022 (51.1%)	41 740 (22.2%)	48 393 (25.7%)
dbo:Politician	21 223 (71.3%)	28 554 (43.8%)	31 569 (48.5%)	6.4	11.7	10.8	11 049 (37.1%)	12 400 (19.0%)	14 934 (22.9%)
dbo:Writer	14 758 (77.5%)	16 630 (44.1%)	18 904 (50.1%)	6.6	9.6	11.9	7 912 (41.5%)	5 195 (13.8%)	8 020 (21.2%)
dbo:Scientist	8 505 (62.4%)	6 659 (23.1%)	7 982 (27.7%)	6.1	9.4	11.2	3 345 (24.5%)	1 489 (5.2%)	2 374 (8.2%)
dbo:SportsManager	9 509 (70.6%)	7 409 (49.9%)	8 396 (56.5%)	6.9	13.9	15.8	3 492 (25.9%)	1 708 (11.5%)	1 953 (13.1%)
dbo:Coach	4 528 (88.1%)	3 947 (70.0%)	4 315 (76.5%)	7.7	13.2	16.7	2 407 (46.8%)	1 334 (23.6%)	1 695 (30.0%)
dbo:Model	2 239 (90.7%)	4 915 (65.8%)	5 540 (74.2%)	7.3	8.4	11.2	1 470 (59.5%)	2 164 (29.0%)	2 594 (34.7%)
dbo:Journalist	1 505 (91.4%)	2 336 (71.1%)	2 522 (76.8%)	7.3	9.2	12.0	1 075 (65.3%)	1 324 (40.3%)	1 621 (49.3%)
dbo:Presenter	601 (93.5%)	2 608 (55.8%)	2 894 (61.9%)	7.6	5.8	8.3	443 (68.9%)	977 (20.9%)	1 132 (24.2%)
dbo:Organization	96 145 (56.2%)	169 332 (31.9%)	189 923 (35.8%)	6.6	13.3	14.2	58 041 (33.9%)	37 374 (7.0%)	61 496 (11.6%)
dbo:Company	27 691 (61.0%)	50 778 (42.5%)	55 429 (46.4%)	6.4	12.0	12.7	19 288 (42.5%)	12 972 (10.9%)	21 639 (18.1%)
dbo:EducationalInst.	12 659 (43.8%)	13 515 (12.5%)	18 197 (16.8%)	4.6	5.7	5.8	6 375 (22.0%)	2 366 (2.2%)	6 467 (6.0%)
dbo:Group	27 787 (87.0%)	39 472 (64.1%)	42 702 (69.3%)	7.9	19.7	22.5	18 042 (56.5%)	11 198 (18.2%)	11 035 (17.9%)
dbo:SportsTeam	9 888 (42.8%)	18 767 (31.7%)	20 680 (34.9%)	5.5	11.5	11.1	5 218 (22.6%)	2 067 (3.5%)	7 464 (12.6%)
dbo:Broadcaster	9 921 (75.4%)	18 674 (53.3%)	20 528 (58.6%)	8.7	10.9	13.3	5 313 (40.4%)	3 263 (9.3%)	4 938 (14.1%)
dbo:MilitaryUnit	1 934 (14.6%)	1 754 (5.1%)	2 221 (6.4%)	4.7	10.0	9.3	753 (5.7%)	144 (0.4%)	391 (1.1%)
dbo:PoliticalParty	1 019 (43.1%)	2 804 (20.4%)	3 370 (24.5%)	6.2	15.6	14.3	422 (17.8%)	430 (3.1%)	666 (4.8%)
dbo:GovernmentAgency	680 (41.1%)	1 522 (17.6%)	2 134 (24.7%)	5.4	9.8	9.6	225 (13.6%)	301 (3.5%)	664 (7.7%)
dbo:Non-ProfitOrg.	1 315 (55.9%)	2 585 (33.3%)	2 933 (37.7%)	5.9	10.2	10.6	874 (37.1%)	886 (11.4%)	1 610 (20.7%)
dbo:TradeUnion	51 (33.3%)	862 (42.5%)	760 (37.5%)	5.2	18.7	16.2	24 (15.7%)	80 (3.9%)	164 (8.1%)
All entities	620 396 (71.0%)	906 349 (35.3%)	1 026 413 (40.0%)	6.9	12.7	13.7	304 773 (34.9%)	271 824 (10.5%)	322 124 (12.5%)

4.4 Using SocialLink

As stated in Section 4.1, SocialLink establishes a link between DBpedia and Twitter, centered on popular entities occurring in both of them, which enables transferring knowledge from one resource to another and back, as well as comparing and jointly analysing the DBpedia graph and Twitter network. In the following, we describe four example use cases where these capabilities can be leveraged.

4.4.1 DBpedia to Twitter: User Profiling

The task of inferring users attributes based on their digital footprint is typically referred to as *user profiling*. Prediction of various attributes based on a person's social graph, posted content, or other attributes is popular among researchers and companies. However, in most setups, namely supervised machine learning-based ones, user profiling requires significant amounts of manual labour to construct training sets. This both limits the possible attributes that can be inferred and the applicability of approaches operating on large amounts of training data, such as DNNs. Recently, researchers focused on automatic crawling of user profiling datasets from social media. However, even the largest datasets only contain few thousands examples per property (Farseev et al., 2015) and are limited to properties explicitly present in social media.

SocialLink helps tackling user profiling by providing accurate machine-readable descriptions for hundreds of thousands of social media profiles. Any attribute present in DBpedia can now be modeled without relying on expensive manual annotation, and SocialLink can be used both to train and evaluate any proposed attribute classifiers.

Another example is inferring user interests based on social graph. Consider a user following, mentioning, or otherwise interacting with accounts aligned in SocialLink. By using this information, one can try to model interests, location, and language of the user by just looking at the DBpedia properties of these accounts (Besel et al., 2016; Piao and Breslin, 2017). For instance, following dbr:SpaceX and dbr:NASA can point on a dbr:Aerospace_engineering industry fan, while many dbr:Donald_Trump-related tweets can reveal a dbr:GOP supporter. The ability of SocialLink to significantly simplify user profiling pipelines is demonstrated in Chapter 6.

4.4.2 DBpedia to Twitter: Entity Linking

Another use case is the Named Entity Linking (NEL) task, whose goal is to link mentions of named entities in a text to their corresponding entities in a KB such as DBpedia. Challenging on its own, the NEL task presents additional unique challenges when applied to social media posts due to noisiness, lack of sufficient textual context, and informal nature of posts (e.g., use of slang).

Social media posts typically contain explicit mentions of social media accounts in the form of @username snippets. When referring to Twitter, some of these mentions (especially the ones referring to popular accounts) may be aligned in SocialLink, and thus can be directly disambiguated to DBpedia with high precision using our resource. Apart being part of the NEL result, these links provide additional contextual information (injected from DBpedia) that can be leveraged for disambiguating other named entities occurring in the post being processed. Additionally, mentions of the named entities in social media posts are typically done via the specific @username constructs that further hinders the usage of established Natural Language Processing toolsets. SocialLink contain direct links between the social media profile (identified by the unique username) and corresponding DBpedia entities, which is the usual target of the NEL task. Therefore, disambiguating and linking a named entity could be as simple as a simple lookup in our resource. SocialLink was used in this capacity by two teams (Corcoglioniti et al., 2016; Minard et al., 2016) participating to a NEL challenge on Italian tweets (NEEL-IT task) as part of the EVALITA 2016 campaign, allowing both of them to improve their results.

It is worth noting that the two-step approach of the SocialLink pipeline can be adapted to directly disambiguate named entities in texts against the social media. Such functionality is present in the Social Media Toolkit which will be described in Chapter 7.

4.4.3 Twitter to DBpedia: Extracting FOAF Profiles and Type Prediction

Up-to-date information about DBpedia persons and organizations can be extracted from Twitter and brought to DBpedia after an alignment is established through SocialLink. Focusing on persons, different profile properties expressible with FOAF may be extracted from a DBpedia person's Twitter account, including:

- basic properties like foaf:name, foaf:surname, foaf:gender, foaf:birthday, and foaf:depiction linking to user images scarce in DBpedia but available in Twitter profiles;
- acquaintances (foaf:knows), extracted from friends, followers and Twitter accounts a user interacted with that are aligned to DBpedia entities in SocialLink;
- links to homepages (foaf:homepage and similar) and other web resources from a Twitter user description and posts, that can be matched to external links in DBpedia to mine relations with other DBpedia entities (e.g., affiliation, authorship, participation, all expressible in FOAF).

While a basic FOAF profile can be extracted from any Twitter account, the links to DBpedia provided by SocialLink allow grounding the extracted data and disambiguating the values of object properties with respect to a larger KB, this way increasing the usefulness of extracted FOAF profiles.

Going further, as will be shown in Chapter 5, Twitter data imported along the populated links can be used as features to infer missing type information for entities in DBpedia using simple machine learning-based approaches. There we demonstrate that features based on Twitter data can outperform state-of-the-art entity representations built from a knowledge graph on a wide selection of DBpedia types. The performance is further improved by combining those feature families together.

Both of these use cases demonstrate the knowledge transfer from Twitter to DBpedia by showing that social media data can serve to directly enrich a target knowledge graph. Additionally, the results presented in Chapter 5 indicate possibility of using the same approach for ontology population and user profiling.

4.4.4 Twitter to Wikidata: Referencing of Crowdsourced Knowledge

Instead of bringing the knowledge directly, social media can be used as a source of external references for existing claims in datasets in the LOD. Nowadays, public crowdsourced datasets, such as Wikidata, are trying to find ways to corroborate the knowledge contained there with external sources. Currently, only a quarter of all claims contained in Wikidata is supported with external (non-wiki) references. Since many of the profiles in social media are verified and contain first-hand information about the entities they represent, links to them constitute in many cases valid external references. Given sufficient high-quality links between a crowdsourced dataset and social media, which SocialLink could provide, we can significantly improve the coverage of references in such dataset.

The ongoing `soweego`[13] project, recently funded by the Wikimedia Foundation, aims to use SocialLink approach to link external catalogs (including social media) to Wikidata. While SocialLink dataset already provides a significant number of links that could be used to provide references (we mostly use Wikidata URIs in the resource), a custom version explicitly targeting Wikidata can be produced using the same or improved pipeline.

4.5 Conclusions and Future Work

In this chapter, we presented our Linked Open Data dataset that links Twitter profiles to corresponding DBpedia entities in multiple language chapters. Building on the approach described in Chapter 3, we have made the SocialLink dataset a valuable resource for the Semantic Web community and Social Media researchers alike. Use cases of SocialLink include, but are not limited to, user profiling, entity linking, and knowledge base

[13] https://meta.wikimedia.org/wiki/Grants:Project/Hjfocs/soweego

enrichment. Our resource can be automatically repopulated using an open source software allowing reproducibility and welcoming contributions from the community. To this date, we have released three major revisions of the dataset, each marking significant milestones in the SocialLink development.