SocialLink: Linking DBpedia Entities to Corresponding Twitter Accounts

Yaroslav Nechaev[1,2], Francesco Corcoglioniti[1(✉)], and Claudio Giuliano[1]

[1] Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy
{nechaev,corcoglio,giuliano}@fbk.eu
[2] University of Trento, Via Sommarive 14, 38123 Trento, Italy

Abstract. We present SocialLink, a publicly available Linked Open Data dataset that matches social media accounts on Twitter to the corresponding entities in multiple language chapters of DBpedia. By effectively bridging the Twitter social media world and the Linked Open Data cloud, SocialLink enables knowledge transfer between the two: on the one hand, it supports Semantic Web practitioners in better harvesting the vast amounts of valuable, up-to-date information available in Twitter; on the other hand, it permits Social Media researchers to leverage DBpedia data when processing the noisy, semi-structured data of Twitter. SocialLink is automatically updated with periodic releases and the code along with the gold standard dataset used for its training are made available as an open source project.

Keywords: Social media · Linked open data · Machine learning

Resource Type: Dataset
Persistent URL: http://w3id.org/sociallink/

1 Introduction

Today it is hard to imagine a public person or an organisation that does not have a social media account. Such entities typically have a rich presence in the social media, sharing content, engaging with their audience, and maintaining and expanding their popularity. They typically keep all the information in their profiles and posts as relevant and precise as possible, so that a potential consumer or a fan can be informed about the latest developments in no time. Thus, social media have become a primary source of information providing up-to-date knowledge on a wide variety of topics, from major events to the opening hours of stores or what books or songs a particular celebrity likes.

Coincidentally, such people and organisations often have dedicated Wikipedia pages, and thus corresponding entries in knowledge bases (KB) related to Wikipedia, such as DBpedia, YAGO, or Wikidata. Data in social media and KBs present opposite characteristics. On the one hand, KBs provide high-quality, structured, and easily accessible information, while data from social

© Springer International Publishing AG 2017
C. d'Amato et al. (Eds.): ISWC 2017, Part II, LNCS 10588, pp. 165–174, 2017.
DOI: 10.1007/978-3-319-68204-4_17

media accounts is often noisy, unstructured, and hidden behind restrictive APIs. To extract from social media as much information as typically contained in a KB entry, sophisticated pipelines have to be built implementing tasks like event detection, user profiling, and entity linking. These tasks exploit supervised machine learning requiring large training sets that are scarcely available and expensive to create manually. On the other hand, social media provide up-to-date information, while contents in KBs may lag behind from hours to months, depending on how many people care about those topics; such delays may prevent using these KBs in some areas. In light of these differences, an integration of KBs and social media may be beneficial to address the shortcomings of one leveraging the strengths of the other.

In this paper, we present SocialLink[1], a publicly available Linked Open Data (LOD) dataset that matches social media accounts on Twitter, a popular social network providing up-to-date publicly available contents, to their corresponding entities in DBpedia. This resource creates a bridge between the highly structured LOD cloud and the vibrant and up-to-date social media world. By aligning around 271 K DBpedia persons and organisations to their Twitter profiles, SocialLink serves two 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 [2,6]. 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 is unique in the alignment task it addresses. This task shares some similarities with the *profile matching* task on (different) social media [5], but KBs do not contain attributes that are vital for that task, such as usernames, user-generated content, and social graph, meaning that techniques for profile matching cannot be directly applied in our context.

SocialLink was introduced in mid 2016 using the supervised alignment approach described in [7]. 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, improving its performances 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. SocialLink is repopulated periodically in an automatic way to insure that alignments are always up-to-date.

[1] http://sociallink.futuro.media/—Creative Commons Attribution license (CC BY 4.0).

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

In the remainder of the paper, Sects. 2 and 3 present respectively the SocialLink pipeline and the latest version of the SocialLink dataset. Section 4 discusses some example use cases where SocialLink has been or can be used, while Sect. 5 concludes.

2 SocialLink Pipeline

Figure 1 highlights the three phases of the SocialLink pipeline used to generate the dataset. 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 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 SocialLink dataset. More details on the candidate acquisition and selection phases are available in [7].

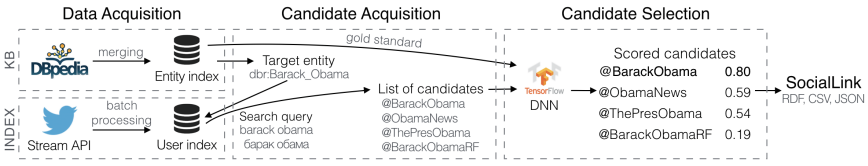

Fig. 1. Principal software components

Data Acquisition. We consider person and organisation entities[3] from all the 128 language chapters of DBpedia (version 2016-04). To speed up processing, we build a local entity index consisting of a Virtuoso triplestore populated with data from multiple DBpedia chapters. We use a *merging* component based on RDFpro [3], which downloads the required DBpedia data, filters out unwanted triples, and merges the remaining ones along owl:sameAs links ('smushing'), so that each entity is assigned a *canonical* URI used in its triples and linked to the entity owl:sameAs aliases. Overall, the merging component downloads 87 GB of compressed RDF corresponding to 7.3 B triples, and populates the index with 1.4 B triples, including 58745 gold standard alignments.[4]

From the social media side, SocialLink requires access to either the Twitter Streaming API or the Twitter Search API in order to populate the list of

[3] These types account for the majority of the DBpedia–Twitter alignments in DBpedia.

[4] Gold alignments derive from selected foaf:isPrimaryTopicOf and wikidata:P2002 triples of entities assumed living based on presence/values of selected properties (full details on website).

candidates. The latter option was used in the original paper and was the main bottleneck of the approach due to strict API rate limits. To avoid this bottleneck, the Twitter Streaming API is now accessed to download a (sampled) stream of tweets from which user accounts and text are extracted and indexed locally according to a continuous process, allowing the system to perform hundreds of queries per second on a single machine and enabling frequent, reliable, and fully automatic population and update of the resource. Additionally, this index provides much more user-related data, thus increasing alignment performances. We currently gathered three years of raw Twitter data, out of which 450 GB of indexed and accessible user data were produced. Processing is implemented using Apache Flink,[5] a framework providing reliability (via automatic checkpoints) and scalability (via automatic horizontal scaling). PostgreSQL is used as a backend for the Twitter user index.

Candidate Acquisition. The introduction of the custom user index allows a great degree of flexibility in acquiring the candidate accounts possibly matching a given DBpedia entity, as different query strategies can be implemented. In this release we employ the strategy that combines all known names of an entity deduplicated and sorted by frequency. Names consisting only of a first or a last name (from foaf:givenName and foaf:surname properties) are filtered out to prevent noisy results. Additionally, the approach now modifies the query (if possible) and performs additional requests in case the original query produces no results or is too broad (i.e., it results in too many candidates). Our index and query strategies are currently based on simple multi-language stemming and tokenisation techniques that perform well on Western languages but weak on Asian and Arabic languages. Despite this, we could further increase the recall of this phase (i.e., the amount of entities with a Twitter account for which some candidate is returned) from 56.5% reported in [7] to 59.2% of the current dataset release.

Candidate Selection. The scoring procedure for each ⟨candidate, entity⟩ pair, used to select the matching candidate given a DBpedia entity (if any), is straightforward. A DNN is trained using the gold standard to perform a binary classification task: to align or not to align. Five kinds of features and all their pairwise combinations are used [7]: (i) name-based features (edit distances); (ii) profile metrics ('is verified' flag, followers/friends/listed/statuses counts); (iii) cosine similarities between profile descriptions, tweets content and DBpedia descriptions; (iv) entity type (person/organisation/other); (v) homepage-related features (e.g., if foaf:homepage property contains a unique reference to the Twitter account). The result is a confidence value representing the probability of a ⟨candidate, entity⟩ pair being a correct alignment. The scoring subsystem has then to decide whether there is a correct alignment for a given entity using 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

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

if they pass the minimum score requirement. We include raw scores for each ⟨candidate, entity⟩ pair in SocialLink to allow tuning thresholds for a desired precision/recall balance. The thresholds for this dataset release are optimized for precision and are set to 0.4 minimum score and 0.4 minimum improvement, leading to 90% precision and 41% recall of generated alignments (vs. 85% precision, 52% recall in [7] where we optimized for F1). Candidate selection alone has 89.4% precision and 69.2% recall, assuming candidate acquisition succeeds in including the correct candidate. Alignments are more reliable for entities having rich Twitter accounts, as amount and relatedness of Twitter content play a key role in our feature set. Up-to-date performance figures are reported on the website.

The SocialLink pipeline is implemented as an open source project that we have been constantly contributing to since the original version described in [7]. The majority of the system is written in Java. Along with the code of the SocialLink pipeline, we are also releasing a complementary web-based test bench called *Social Media Toolkit*. It provides a convenient way to query the SocialLink dataset via the REST API, as well as additional social media-related functionalities that leverage SocialLink, such as a custom Named Entity Linking pipeline that links entities in a free text to social media profiles.

3 SocialLink Dataset

The result of running the SocialLink pipeline is the SocialLink dataset, that we generate periodically to account for updates in DBpedia and Twitter. The SocialLink dataset is distributed in different formats, with RDF being the main one that also includes all the intermediate candidate data. We describe here the modeling choices behind the RDF format of the SocialLink dataset, summarizing the statistics of its latest release and discussing how the dataset is made available online and kept up-to-date.

RDF Format. We encode our alignments in RDF using terms from FOAF, Dublin Core Terms, and our custom SocialLink vocabulary (prefix `sl`), as exemplified in Fig. 2.

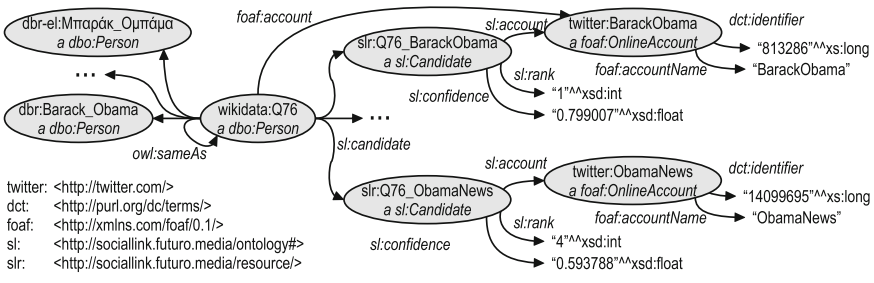

Fig. 2. Representation of alignments in RDF.

DBpedia entities are referenced using canonical URIs possibly taken from Wikidata, like wikidata:Q76 for entity Barack Obama in Fig. 2. 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.

Twitter accounts, like twitter:BarackObama in Fig. 2, 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 Fig. 2) 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
WHERE {?e owl:sameAs <E>; foaf:onlineAccount ?account}
```

Dataset Statistics. Table 1 reports relevant statistics for the latest release of the SocialLink dataset. For the considered dbo:Person and dbo:Organisation DBpedia entity types, and their top five subtypes with the largest number of alignments, we report: (i) the total number of entities of that type in DBpedia (after merging all chapters); (ii) the number of *living* entities that may be

Table 1. Dataset statistics by DBpedia entity type.

Entity type	Entities in DBpedia	Living entities in DBpedia	Entities with candidates	Candidates / entity	Entities aligned
dbo:Person	2 975 645	2 035 590 (68.4%)	737 017 (24.8%)	12.6	234 450 (7.9%)
dbo:Athlete	493 867	412 629 (83.6%)	214 070 (43.3%)	15.1	71 935 (14.6%)
dbo:Artist	269 745	188 095 (69.7%)	104 614 (38.8%)	12.3	41 740 (15.5%)
dbo:Politician	123 460	65 135 (52.8%)	28 554 (23.1%)	11.7	12 400 (10.0%)
dbo:Writer	69 753	37 744 (54.1%)	16 630 (23.8%)	9.6	5 195 (7.4%)
dbo:Model	7 601	7 470 (98.3%)	4 915 (64.7%)	8.4	2 164 (28.5%)
dbo:Organisation	575 644	553 433 (96.1%)	169 332 (29.4%)	13.3	37 374 (6.5%)
dbo:Company	131 056	121 554 (92.7%)	50 778 (38.7%)	12.0	12 972 (9.9%)
dbo:Group	66 868	62 087 (92.9%)	39 472 (59.0%)	19.7	11 198 (16.7%)
dbo:Broadcaster	35 394	35 373 (99.9%)	18 674 (52.8%)	10.9	3 263 (9.2%)
dbo:EducationalInst.	116 139	115 722 (99.6%)	13 515 (11.6%)	5.7	2 366 (2.0%)
dbo:SportsTeam	62 221	60 870 (97.8%)	18 767 (30.2%)	11.5	2 067 (3.3%)
All entities	3 551 289	2 589 023 (72.9%)	906 349 (25.5%)	12.7	271 824 (7.7%)

aligned to Twitter; (iii) the number of living entities for which at least a candidate account was matched in the candidate acquisition phase; (iv) the average number of candidates per living entity (when matches were found); and, (v) the number of entities aligned to Twitter accounts for that type. The percentages in parenthesis refer to the total number of entities of a type. Due to chosen thresholds, the system *abstains* producing an alignment for about 2/3 of all entities having candidates (906 349), corresponding either to ambiguity cases (e.g., cannot distinguish between fake and real Twitter accounts) or to cases where the correct account is not among the produced candidates (e.g., Twitter name not occurring in DBpedia).

Availability and Sustainability. The SocialLink dataset is indexed on DataHub[6] and is available for download on SocialLink website, together with VOID 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 [9] (this release) and Zenodo [8] (all releases) digital repositories. Alignment data is also available and queryable by end users and applications via a publicly accessible SPARQL endpoint[7] 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.

4 Using SocialLink

As stated in Sect. 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 three example use cases where these capabilities can be leveraged.

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

[6] http://datahub.io/dataset/sociallink.
[7] http://sociallink.futuro.media/sparql.

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 [4] 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 [1]. 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.

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. SocialLink was used in this capacity by two teams [2,6] 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 available on SocialLink website.

Twitter to DBpedia: Extracting FOAF Profiles. Up-to-date information about DBpedia persons and organisations can be extracted from Twitter 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.

5 Conclusions and Future Work

In this paper we presented SocialLink, a Linked Open Data dataset that links Twitter profiles to corresponding DBpedia entities in multiple language chapters. By improving our initial approach described in [7], we have made SocialLink 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 enrichment. Our resource is automatically populated using an open source software allowing reproducibility and welcoming contributions from the community.

We will continue to gradually update SocialLink by both improving the approach and expanding the scope to accommodate a larger subset of Linked Open Data entities. A significant goal in our current roadmap consists in the expansion of our approach to other social networks, such as Facebook and Instagram. By introducing more social media to SocialLink we will be able to not only improve coverage but also exploit cross-network information to validate our alignments.

References

1. Besel, C., Schlötterer, J., Granitzer, M.: Inferring semantic interest profiles from Twitter followees: does Twitter know better than your friends?. In: ACM SAC, pp. 1152–1157 (2016)
2. Corcoglioniti, F., Palmero Aprosio, A., Nechaev, Y., Giuliano, C.: MicroNeel: combining NLP tools to perform named entity detection and linking on microposts. In: EVALITA (2016)
3. Corcoglioniti, F., Rospocher, M., Mostarda, M., Amadori, M.: Processing billions of RDF triples on a single machine using streaming and sorting. In: ACM SAC, pp. 368–375 (2015)
4. Farseev, A., Nie, L., Akbari, M., Chua, T.S.: Harvesting multiple sources for user profile learning: a big data study. In: ACM ICMR, pp. 235–242 (2015)

5. Goga, O.: Matching user accounts across online social networks: methods and applications. Ph.D. thesis, LIP6-Laboratoire d'Informatique de Paris 6 (2014)
6. Minard, A., Qwaider, M.R.H., Magnini, B.: FBK-NLP at NEEL-IT: active learning for domain adaptation. In: EVALITA (2016)
7. Nechaev, Y., Corcoglioniti, F., Giuliano, C.: Linking knowledge bases to social media profiles. In: ACM SAC, pp. 145–150 (2017)
8. Nechaev, Y., Corcoglioniti, F., Giuliano, C.: SocialLink dataset. Zenodo (2017). https://doi.org/10.5281/zenodo.820160
9. Nechaev, Y., Giuliano, C., Corcoglioniti, F.: SocialLink: knowledge transfer between social media and linked open data. Figshare (2017). https://doi.org/10.6084/m9.figshare.5235823