Chapter 1

Introduction

This thesis is aimed to bring together Semantic Web and Social Media Analysis by enabling the knowledge transfer between the two and, as a result, improving approaches in tasks, such as type prediction, named entity linking and user profiling. In this chapter, first I introduce the relevant context (Section 1.1). Then, I describe the problems that are being addressed in this thesis (Section 1.2). Finally, I present the contributions (Section 1.3), structure (Section 1.4), list of publications that support the contributions (Section 1.5) and the artifacts (Section 1.6) of this thesis.

1.1 Context

Today it is hard to imagine a public person or an organization 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, maintaining and expanding their popularity. They post new content frequently and keep all the information in their profiles 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. An enormous amount of posts, profile updates, comments, images, and videos are being produced each day in response to real-world events ranging from the ones of planet-wide importance like the Olympics all the way to minor local or even personal events. Additionally, structured and semistructured data is voluntarily being filled by users: from the opening hours of stores to what books or songs a particular celebrity likes. Given that Facebook alone has more than 2B monthly active users, the scale and the coverage of such data are hard to overestimate. A desire to benefit from such vast array of knowledge have fueled a more than a decade-long research interest in social media in different scientific communities and companies. While extracting, storing, processing and analyzing the semi-structured data at such a scale is challenging, countless approaches were proposed over the years to tackle various tasks in social media analysis, such as user profiling, profile matching, entity linking, community detection and labeling and many others.

On the other hand, from the very inception of the world wide web, there have been various efforts to gather and systematize human knowledge. One of the most prominent examples, Wikipedia, the largest crowdsourced online encyclopedia, presents such knowledge in a convenient, human-readable fashion. Complementary to this, the Semantic Web community has spawned the Linked Open Data (LOD) project to represent the same knowledge and much more in a machine-readable form. The LOD cloud is a collection of interlinked, standardized and structured datasets, or Knowledge Bases (KB), that have long become an invaluable source of data for many tasks, especially in data mining and knowledge discovery. A Knowledge Base contains a slice of humanity's knowledge in the form of a Knowledge Graph (KG), where entities participate in relationships according to some formal description called ontology and are encoded using RDF as a backbone. The well-understood process of processing the RDF data and the public availability of LOD datasets have made them staple in pipelines that can benefit from the use of large quantities of background knowledge. The community effort has created many billions of RDF triples coming from hundreds of resources over the years. Notoriously, Wikipedia-based knowledge bases, such as DBpedia, YAGO, and Wikidata, created from the labor of millions[1] of Wikipedia editors, are often used in a wide range of Natural Language Processing (NLP) tasks.

Data in social media and KBs present opposite characteristics. On the one hand, KBs provide high-quality structured information (e.g., YAGO has 95% accuracy) that is easily accessible, while data from social media is often noisy, unstructured, and hidden behind restrictive APIs. To extract from the social media as much information as contained in a typical KB entry sophisticated pipelines had to be built. As a result, significant research effort went towards solving tasks on social media, such as event detection, user profiling, and entity linking. These tasks mainly have to exploit supervised learning, which requires training sets that are scarcely available and expensive to create manually. On the other hand, social media provide up-to-date (real-time) information, while contents in KBs may lag behind from hours to months. For Wikipedia-related KBs, such lag comprises both the time for changing the page (hours to months, based on popularity) and, for automatically extracted KBs like DBpedia and YAGO, the time for that change to propagate in the KB (months to years). Such lag may prevent using these KBs in some application scenarios.

Coincidentally, for entities that exist in both the social media and in the KB (mainly people and organizations), the knowledge extracted from the one can complement and confirm the data in other. Data from the social media can update stale entries, fill the

[1] https://en.wikipedia.org/wiki/Wikipedia:Authors_of_Wikipedia

blanks in the KB and act as a reference supporting an existing fact in the KB,[2] while KBs can provide a solid structured background to the analysis performed on the social media. This idea that in essence is a knowledge transfer between the vibrant social media world and the structured Linked Open Data cloud forms the foundation of this thesis. Some studies have already exploited the LOD to augment social media analysis pipelines (Piao and Breslin 2017; Besel et al. 2016) providing the initial proof that such transfer is feasible. The aim of this thesis is to investigate the ways to enable this knowledge transfer and capitalize on it by improving and augmenting existing pipelines in both the Semantic Web and Social Media Analysis fields using the data from the social media and the background knowledge from the KBs respectively.

While the LOD cloud currently contains over 1,000 linked datasets[3] and there is a multitude of social media, in this thesis I specifically explore the knowledge transfer between DBpedia and Twitter. DBpedia is a general purpose KB derived from Wikipedia that forms a foundation of the LOD cloud. It is one of the largest and most popular datasets in the LOD and is the most interlinked one making it a perfect choice for the task at hand. Twitter is one of the major social media having more than 350M monthly active users (statista.com, 2019c). It has the least restrictive open API among big commercial social networks allowing easy access to a significant portion of data and is vastly popular among researchers across many areas of Computer and Social Sciences. Additionally, Twitter functionality is a bare minimum of what the social network typically is (includes the notions of posts, social graph, semi-structured profile, shares, likes and mentions), meaning that many of the approaches designed for Twitter would work on other social media with little or no modification. In sum, bridging DBpedia and Twitter provides the most potential research impact while minimizing the cost of acquiring the needed data and doing experiments.

1.2 Problem definition

In order to enable the knowledge transfer between KBs and social media, a significant amount of links between them has to be present effectively acting as a "bridge" connecting the two worlds. There are more than two million living people and more than half a million of currently existing organizations listed in DBpedia, many of which would have some presence in social media. Given that there are just 56,113 existing links from such entities in Wikidata and DBpedia to Twitter, one could expect a much more significant coverage

[2] Since an account in social media can be designated as official means of communication for an entity in the KB, information extracted from the content that is produced by this account can corroborate facts in the KB.

[3] 1,229 datasets as of November 2018 (see `https://lod-cloud.net`)

to perform the knowledge transfer and to benefit from it successfully. As will be shown in Chapter 5, the amount of available links between KBs and social media significantly affects the performance of the downstream tasks that would like to benefit from such knowledge transfer. In order to overcome this disconnect, in this thesis, I present ways to establish additional links and, as a result, the potential link coverage between DBpedia and Twitter is increased more than tenfold.

With this in mind, I present the task of linking knowledge base entities to social media profiles which is at the core of this work. The goal is to find a profile in a social media for a given entity in a Knowledge Base. The task could be formulated the other way around, i.e., to find an entity in a KB for the chosen social media profile. However, given that there are billions of profiles in a major social media, which can only be partially acquired via expensive crawling, the inverse task will not be touched in this thesis.[4] As mentioned above, in this thesis, I mainly discuss such linking from DBpedia to Twitter, although the task defined here and most of the contributions are general and may apply to other KBs and social media with similar characteristics. To further limit the scope, I will only focus on entities for living people and currently existing organizations. While correct links to social media may be established for other entity types, such as brands, products and events, and some profiles of the deceased people and dissolved organizations are preserved in the social media, their linking will not be covered in this thesis. In DBpedia version 2016-04, this limits the task to a little more than 2.5M entities (see Chapter 4).

While the linking problem presented here is similar to the tasks of entity matching and profile matching that are well-studied by scholars in the fields of Semantic Web (among others) and Social Media Analysis respectively, unique challenges arising from aligning such vastly different resources warrant for a custom solution. Indeed, for entity matching — i.e., the task of finding KB entities referring to the same real-world entity – it is typically assumed to have *structured* information about *all* the entities to be matched in advance, which means that a social network would have to be fully available in some structured form to be amenable to entity matching techniques. As social media are neither openly available nor structured entity matching methods cannot be applied as is. On the other hand, research in profile matching, which is the task of aligning profiles in multiple social networks that correspond to a single person, can not be applied due to its reliance on behavioural patterns that people exhibit when creating multiple social media profiles (e.g., similar social graphs, similarities in a chosen account handle). Such patterns are not typically available in KBs.

Designing a linking approach also requires taking into consideration issues and peculiarities of processing the LOD and social media data. Firstly, different types of information

[4] Nevertheless, some recent works (Besel et al., 2016; Piao and Breslin, 2017) were able to successfully tackle this task for a small subset of Twitter users.

are available. Twitter mainly contains significant amounts of unstructured textual and media data, while DBpedia offers a wide range of structured properties that, unfortunately, can not be aligned to Twitter. Secondly, the quality and the amount of data is varying significantly from entity to entity and among different social profiles. DBpedia suffers from the knowledge lag discussed above, while Twitter data may be noisy, incomplete or deliberately false. In third, Twitter contains an enormous amount of accounts inevitably increasing ambiguity, which requires the approach to be more conservative to take into consideration possible fake and fan accounts, namesakes and even multiple true profiles. All of these issues and challenges will be covered in details in Chapter 3.

After enough links between a KB and a social network are populated, existing approaches for the variety of tasks in both can be modified to benefit from the knowledge transfer. Type prediction, which is the task of predicting missing type information for entities in the KB, is a prominent example. Indeed, as seen in Aprosio et al. (2013), in addition to the target KB's knowledge graph itself, type prediction approaches can exhibit improved performance by ingesting aligned data from other sources. Important tasks in social media can exploit the LOD to improve and simplify approaches as well. User profiling is the task of predicting a target user attribute given all the information that is known about the user. While being immensely popular among researchers in many fields, this task is essential for the commercial success of social media and the surrounding infrastructure. In Besel et al. (2016) and Piao and Breslin (2017), data imported from Wikipedia through Twitter-Wikipedia links is used as an essential piece of the profiling pipeline. In this thesis, I modify the approaches for both tasks to benefit from the data that can be imported along the Twitter-DBpedia links populated by the approach detailed in this thesis.

Moreover, the ingestion of data through the newly found links not only helps with the raw performance of the current systems, but more importantly, it enables new directions to be explored and novel approaches to be proposed that could not be made working by using the existing links. For example, Named Entity Linking (NEL), which is the task of disambiguating the identity of entities mentioned in the text, is typically designed to align entities in the text to entities in some KB. In this thesis, I demonstrate that the populated links and even the proposed linking approach itself can be used to complement typical NEL pipelines and enable linking to social media profiles instead of KB entries.

In sum, the current research efforts to populate and maintain the LOD cloud and the ever-increasing research interest in analyzing social media can both benefit from the knowledge transfer between those two worlds. However, such transfer can only function given that a significant amount of links between the KBs and the social networks can be established and the corresponding approaches are appropriately updated to exploit them.

1.3 Contributions

The core contribution of this thesis is SocialLink — the approach that is designed to align knowledge base entities to social media profiles and the public LOD dataset that contains precomputed alignments between DBpedia and Twitter. SocialLink effectively bridges the two worlds by (i) providing a scalable and robust machine learning-based procedure to find possible alignments for a given knowledge base entity and by (ii) providing a LOD compliant resource that can be used by the Semantic Web practitioners and the social media researchers alike without the need of instantiating the entire linking pipeline. Such bridge enables knowledge transfer in both directions and the approach itself can be used to link social media profiles to entities outside of the LOD cloud. SocialLink approach is a deep neural network-based system that is able to efficiently encode and utilize multiple modalities of data. For example, SocialLink exploits dense embeddings to represent knowledge and social graphs acquired independently in an unsupervised way, and it is then able to learn the similarity function to perform the alignment. Textual and nominal features are combined to assist in the task as well. As of version $v3.0$, released in October 2018, SocialLink performs linking starting from 2.5M entities found in 120 DBpedia language chapters and considers 291M Twitter users for alignment. This results in high quality links for 322K entities and lists candidate alignments with varying levels of confidence for over 600K more entities, providing more than tenfold increase in the number of DBpedia-Twitter links that were available in the LOD before.

SocialLink is an open source project that has been updated with algorithm and pipeline improvements and the dataset releases. The code is publicly available on Github,[5] while all the datasets are released on Figshare (Nechaev et al., 2017c) and Zenodo (Nechaev et al., 2018a). The website[6] contains all of the above along with additional resources, documentation and a public SPARQL endpoint to simplify and facilitate the usage of the dataset. SocialLink slowly but steadily gains adoption. Multiple teams in EVALITA2016 Entity Linking challenge were able to improve their results by exploiting the populated links. Additionally, Wikimedia Foundation has awarded the largest project grant of 2017 to **soweego**:[7] a project with the goal of incorporating SocialLink's linking pipeline as part of Wikidata to automatically align external catalogs and to provide references to claims contained in it.

In order to prove SocialLink's ability to provide added value via the aforementioned knowledge transfer, a number of additional contributions are presented in this thesis. Firstly, a *type prediction* approach was developed using social media data injected along

[5] https://github.com/Remper/sociallink

[6] https://w3id.org/sociallink

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

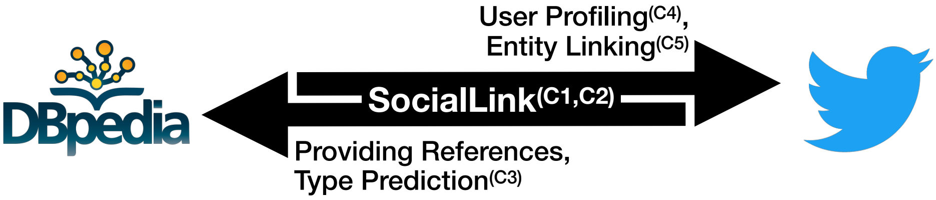

Figure 1.1. A diagram of thesis contributions: SocialLink enables knowledge transfer between the Linked Open Data cloud and social media facilitating tasks in both directions. Automatic referencing of claims in the KB is not a separate contribution of this thesis but it will be mentioned in Chapter 4 as part of the discussion around the **soweego** project.

the populated links. Typical approaches for predicting types in the LOD utilize information already present in the knowledge graph to infer the missing type relations. Alternatively, the information from the related resources, such as Wikipedia, can be imported to facilitate the analysis. The type prediction approach described here is able to improve the state-of-the-art in this task by injecting the social media data as additional features.

Secondly, a concealing approach was presented to hide true user's interests from the user profiling pipelines. The approach exploits the LOD's categorical knowledge, acquired using SocialLink as a bridge, to first predict latent user's interests and then propose actions to perform by the user so that the interests can no longer be reliably predicted. This is done by finding the optimal configuration of user's followed accounts making the resulting distribution over interests, as inferred by the user profiling system, as close as possible to uniform. This approach is a vivid example of a novel task enabled by the knowledge transfer towards social media.

Finally, the Social Media Toolkit (SMT) system was developed as part of the project providing additional functionality over SocialLink. SMT, among other things, uses SocialLink to implement two distinct Named Entity Linking (NEL) scenarios. Firstly, SMT is able to perform direct disambiguation of user mentions in social posts against a knowledge base using the SocialLink dataset. This functionality allows improving the performance of typical NEL pipelines not only by directly solving the task for some of the mentions but by also providing context for the rest of the target text. The MicroNeel system was developed to showcase this scenario reaching the second place in the EVALITA2016 entity linking competition.

In the second scenario, SMT is able to perform NEL on arbitrary text against the social media using SocialLink's linking pipeline. Here the NEL task is modified: instead of linking each mention to a corresponding entity in a KB, a linking to a suitable social media profile is performed. As a clear demonstration of the latter scenario, SMT was incorporated into the Social Media Management platform called Pokedem. Pokedem is designed to recommend specific actions to perform on Twitter in order to increase the popularity of the managed account. The NEL capabilities of SMT are used in Pokedem to augment the proposed tweet with mentions of social media users producing richer content.

To summarize, in this thesis I discuss the knowledge transfer between DBpedia and Twitter, additionally detailing use cases and approaches that are enabled by this transfer in both directions (graphically depicted in Figure 1.1). The major contributions of this thesis, along with their relevant publications and the chapters covering them are as follows:

Contribution C1 – SocialLink Approach An automatic pipeline designed to link LOD-compliant Knowledge Bases to social media profiles (Nechaev et al. 2017b; Nechaev et al. 2018b, Chapter 3).

Contribution C2 – SocialLink Resource The LOD dataset linking 2.5M entities from DBpedia to the corresponding Twitter accounts (Nechaev et al. 2017d, Chapter 4).

Contribution C3 – Type Prediction A type prediction approach employing social media data to consistently outperform state-of-the-art systems (Nechaev et al. 2018c, Chapter 5).

Contribution C4 – Concealing User Interests A system designed to protect interests of social media users from being inferred by the typical user profiling approaches (Nechaev et al. 2017a, Chapter 6)

Contribution C5 – Social Media Toolkit A system providing two novel Named Entity Linking approaches to take advantage of social media data (Corcoglioniti et al., 2016, 2017, 2018, Chapter 7)

1.4 Structure of the Thesis

The remainder of the thesis is structured as follows:

Chapter 2 provides needed background including approaches, techniques, systems and other related work used throughout this thesis. Related work, that is specific for individual contributions, is provided in the end of the respective chapters.

Chapter 3 presents the SocialLink approach (Contribution **C1**). Here I present the linking task in details, identify challenges and go through all the steps of the designed solution. Additionally, I provide the extensive evaluation of the approach along with the thorough error analysis and the discussion on approach limitations.

Chapter 4 describes the SocialLink resource (Contribution **C2**). This chapter presents the resource design aspects, core statistics, formats and provides necessary details for recreating the resource.

The following chapters describe systems and approaches based on SocialLink together covering the rest of the contributions of this thesis:

Chapter 5 describes the state-of-the-art type prediction system exploiting social media data to provide predictions for eight DBpedia types. I provide extensive evaluation showcasing the performance of the social media data extracted exploiting links from DBpedia to Twitter. This chapter covers Contribution **C3**.

Chapter 6 discusses the usage of the LOD data for the user profiling task in social media. Specifically, I tackle the task of interests prediction of passive users. As a main contribution, I present the task of concealing user interests and describe the approaches developed to solve it. This chapter represents Contribution **C4**.

Chapter 7 details the Social Media Toolkit system and it's features. Additionally, this chapter covers the two systems built using the SMT capabilities: MicroNeel and Pokedem. This chapter covers Contribution **C5**.

Finally, I conclude this thesis with the following chapter:

Chapter 8 summarizes the thesis results and provides extensive discussion on the potential uses, limitations, future improvements for the work presented in previous chapters.

1.5 Publications

The core publications supporting the main contributions (**C1**, **C2**) of this thesis are listed below:

- Nechaev, Y., Corcoglioniti, F., and Giuliano, C. (2017b). Linking Knowledge Bases to Social Media Profiles. In *Proceedings of the Symposium on Applied Computing, SAC 2017, Marrakech, Morocco, April 3-7, 2017*, pages 145–150

- Nechaev, Y., Corcoglioniti, F., and Giuliano, C. (2017d). SocialLink: Linking DBpedia Entities to Corresponding Twitter Accounts. In *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II*, pages 165–174

- Nechaev, Y., Corcoglioniti, F., and Giuliano, C. (2018b). SocialLink: Exploiting Graph Embeddings to Link DBpedia Entities to Twitter Profiles. *Progress in AI*, 7(4):251–272

Additional publications supporting the rest of the contributions are as follows:

- Nechaev, Y., Corcoglioniti, F., and Giuliano, C. (2018c). Type Prediction Combining Linked Open Data and Social Media. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 1033–1042

- Nechaev, Y., Corcoglioniti, F., and Giuliano, C. (2017a). Concealing Interests of Passive Users in Social Media. In *Proceedings of the Re-coding Black Mirror 2017 Workshop co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017.*

- Corcoglioniti, F., Giuliano, C., Nechaev, Y., and Zanoli, R. (2017). Pokedem: An Automatic Social Media Management Application. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, pages 358–359, New York, NY, USA. ACM

- Corcoglioniti, F., Nechaev, Y., Giuliano, C., and Zanoli, R. (2018). Twitter User Recommendation for Gaining Followers. In *AI*IA 2018 Advances in Artificial Intelligence - 17th International Conference of the Italian Association for Artificial Intelligence, Trento, Italy, November 20-23, 2018, Proceedings*

- Corcoglioniti, F., Aprosio, A. P., Nechaev, Y., and Giuliano, C. (2016). MicroNeel: Combining NLP Tools to Perform Named Entity Detection and Linking on Microposts. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

1.6 Artifacts

During the development of the approaches and systems detailed here, multiple artifacts were produced ranging from software repositories, such as our main SocialLink project repository, to resources and supplementary materials supporting the publications related to this thesis. They are as follows:

- SocialLink project including SMT – https://github.com/Remper/sociallink
- Type Prediction approach – https://w3id.org/sociallink/type-prediction
- Concealing user interests approach – https://github.com/Remper/re-coding-ws
- Training and model generation code for Pokedem – https://github.com/Remper/pokedem-models
- Twitter knowledge extraction pipelines – https://github.com/Remper/tweetframe

- SocialLink resource – https://w3id.org/sociallink#download
- SocialLink gold standard – https://w3id.org/sociallink#download