Chapter 8

Conclusions

In this thesis, I introduce the task of linking entities in the LOD cloud to social media profiles aiming to enable knowledge transfer between them. Knowledge transfer here is the transfer of data from one medium to another facilitating a wide variety of tasks at the recipient side. The usefulness of such transfer has been repeatedly shown throughout this thesis. However, as the currently available number of links between the LOD and social media is insignificant compared to the potential amount of entities that could be linked to the profiles, a novel approach was required to populate additional links to bridge the two worlds.

To this end, I presented the SocialLink project that aims at the automatic generation of such links as a core contribution of this thesis. SocialLink is designed to target DBpedia, the cornerstone dataset in the LOD, and Twitter, one of the largest social media. However, even though the scope of the approach presented in this thesis is limited to those two particular data sources, particular attention was given to select the features that would most likely be available in other LOD datasets and social networks, thus enabling future extensions. In this chapter, I will briefly summarize the contributions of this thesis, present some of the possible future work directions and discuss privacy concerns that inevitably arise when working with real people's data.

8.1 Summary of Contributions

As presented in Chapter 1, in this thesis, I highlight five main contributions. In this section, I will go over each one of them and briefly summarize the achieved results.

Contribution 1: SocialLink approach. One of the major parts of the SocialLink project is the supervised deep learning-based linking approach that relies on the three-phase procedure described in Chapter 3 to produce links between DBpedia and Twitter. Many challenges had to be addressed in order to develop such an approach. In particular, two of them required a significant amount of novelty to be introduced in order to be

resolved. Firstly, the proprietary nature of Twitter meant that some of the aspects of the approach, such as building the social graph embeddings and searching candidate alignments among hundreds of millions of Twitter users, had to be approximated from the available data. Secondly, Twitter's unstructured and uncurated nature meant that the proposed approach had to take into consideration the presence of fake and fan accounts and correctly work with partially missing features. All these challenges were explored in details in Chapter 3.

Additionally, the introduction of the graph-based features from DBpedia and Twitter required the development of a custom topology for the neural network. Efficiently combining two different vector spaces trained from two independent, unsupervised datasets along with regular numeric and categorical features from our BASE feature set required a much more robust approach. Indeed, the baseline solution based on simple concatenation yielded unreliable convergence and offered less performance.

Finally, the SocialLink approach is extensively evaluated, and error analysis is presented to clearly establish the current performance in this task and highlight strengths and weaknesses. The approach is fully reproducible with the complete source code and documentation available online.[1] The two major revisions of the SocialLink pipeline were presented to the research community as two successive publications in proceedings of ACM SAC conference (Nechaev et al., 2017b) and the Progress in Artificial Intelligence journal (Nechaev et al., 2018b).

Contribution 2: SocialLink resource. The second part of the SocialLink project is the LOD compliant resource that was produced using the presented approach. To this end, in Chapter 4, I discuss the process of building such dataset by taking 2.5M entities of living people and organizations from 120 DBpedia language chapters and linking them to the corresponding Twitter accounts. The final version of the resource, $v3.0$, proposes candidates for over 1M entities and provides high quality (90% precision) links to 322K of them, significantly increasing the number of available links to Twitter for the entities in the Linked Open Data cloud. Chapter 4 contains detailed statistics, design considerations, filtering and preprocessing techniques; it discusses sustainability and presents some of the most important use cases for such a resource both in the field of Semantic Web and Social Media Analysis. The fact that such a resource can be built at scale of millions of links with just conventional hardware proves that the SocialLink approach is efficient and practical. SocialLink dataset is a part of the LOD cloud,[2] available in multiple different formats for download, and a public SPARQL endpoint is maintained to ease access.

[1] http://w3id.org/sociallink
[2] https://lod-cloud.net/dataset/SocialLink

The rapid increase in link coverage between DBpedia and Twitter is a primary prerequisite for the successful knowledge transfer between the two mediums. Having successfully solved this issue, I continued my study of this topic by aiming at proving that such knowledge transfer in both directions is useful in real-world tasks. Specifically, I addressed in details three applications of the SocialLink resource for *Type Prediction*, *User Profiling* and *Named Entity Linking* that form the rest of the contributions of this thesis.

Contribution 3: Type Prediction. To showcase the ability of social media data to benefit Semantic Web community, I set up the pipeline for type prediction on DBpedia. Type prediction, which is the task of predicting missing types for entities in a knowledge base, is typically done by exploiting existing features derived from the knowledge graph of a target KB. Additionally, for such interconnected resources, such as DBpedia, additional resources may be utilized, for example, Wikipedia that has a guaranteed article for each DBpedia entity. In Chapter 5, I defined four feature families (profile, authored text, mentioned text, and social graph) that can be extracted from the Twitter Streaming API and that can be related to the DBpedia entities whose types are predicted through the links of SocialLink. I studied the impact of these additional features enabled by SocialLink, and I have found that in many cases social features allow achieving superior performance compared to the knowledge graph features. Additionally, I studied different combinations of such feature sources, also considering the state-of-the-art Wikipedia-based features as an extra source, finding that the addition of social features provides performance benefits in every case. Such results clearly show that social media data can be used complementary to conventional data sources to improve the type prediction pipelines. Finally, despite the usage of a simple classifier, the combination of all three data sources exhibited great performance levels (e.g., 92% F_1 for Location attribute), which suggests its potential for Ontology Population. To summarize, bringing the social data along the links between the social media and the LOD allows greater performance for the LOD-based task of type prediction.

Contribution 4: Concealing User Interests. Tasks in social media can benefit from the knowledge transfer in the opposite direction: the ingestion of data from the LOD can enable simplified user profiling pipelines. In Chapter 6, I have described a system that is able to identify user interests in an unsupervised manner. Then, I have presented a novel approach that uses the SocialLink resource to propose a set of actions for the user to perform in order to conceal their digital identity. There, the SocialLink resource provides a set of possible profiles with known interests distribution. Then, the novel approach is tasked to find the ideal configuration of profiles to follow in order to confuse profiling pipelines and make them abstain or infer incorrect interest information about the user.

The proposed concealing approach does not degrade user experience as the additional followed profiles can be filtered out on the user side. While the user profiling pipelines relying on the usage of external resources (Piao and Breslin, 2018) were used before, SocialLink uniquely allows the large scale import of knowledge from DBpedia which made the optimization problem proposed in Chapter 6 possible to solve.

Contribution 5: Social Media Toolkit. Named Entity Linking (NEL) is the final task addressed in this thesis exploiting the knowledge transfer. Social Media Toolkit (SMT), presented in Chapter 7, implements two novel NEL scenarios: direct disambiguation of profile mentions found in tweets against DBpedia using the reverse query of SocialLink resource; and the linking of named entities found in a given text to social media profiles employing the custom instance of the SocialLink pipeline. Both of those implementations were embedded into two systems, MicroNeel (Corcoglioniti et al., 2016) and Pokedem (Corcoglioniti et al., 2017, 2018), improving their performances in the respective tasks. Additionally, SMT allows additional customized deployments of the SocialLink pipeline and resource allowing contributors to iterate on the SocialLink project.

8.2 Future Work

This thesis covers multiple research topics spread across its five main contributions. A wide variety of extensions and improvements can be made to the approaches and systems presented in Chapters 3-7. In this section, I propose some of the directions that can be explored in future to provide significant impact on top of the work presented in this thesis.

Firstly, SocialLink pipeline and approach may continue to be gradually updated. A significant improvement would consist in the expansion of the pipeline to other social networks, such as Facebook and Instagram, by generalizing the approach used for Twitter making SocialLink even more of a bridge between the social media world and the LOD cloud. By introducing more social media to SocialLink, the approach can include additional measures to ensure that cross-network information is consistent, improving the precision of the overall system. Another critical direction that could be explored is alleviating the recall drop observed during the *candidate acquisition* phase, for example, by redesigning this phase using machine learning-based techniques. Current *candidate selection* approach can also be improved by learning joint embeddings for both social media profiles and knowledge base entities placing them into the same vector space. Finally, the pairwise candidate selection solution could be reformulated to account for all candidate profiles at once, for example, via learning to rank instead of binary classification.

Additional KBs can also be explored. As part of the `soweego` project, which is a project supported by the Wikimedia Foundation, it is planned to extend SocialLink to

specifically target Wikidata entities instead of DBpedia. The goal of `soweego` is to provide references to claims in Wikipedia across the newly populated links, implementing this way another example of real-world knowledge transfer between social media and knowledge bases.

Secondly, concerning the type prediction system from Chapter 5, Wikidata can be used both as the target knowledge graph and as an extra source of links. The usage of the SocialLink resource can help to bolster coverage even further. These additional links will provide (i) more linked entities whose types (where missing) can be predicted and populated using our approach; and (ii) additional training (where types are known), which in turn may allow targeting a more extensive range of types and performing additional analyses, e.g., of the impact on performances of the amount of available social information. Finally, a joint embedding derived from all user-related data can be learned to address privacy concerns when making the data from social media available to the community. Such embeddings can be released as a LOD dataset allowing researchers to seamlessly use social media data in their type prediction and ontology population pipelines.

In third, the concealing approach, described in Chapter 6, can be extended further by covering other user profiling scenarios that target different attributes (e.g. location) and profiling use cases. The concealing approach itself can be improved making it able to learn from the output of the given interests inference pipeline to provide better results on real-world black box systems, such as Twitter's Who To Follow box.

Finally, the SMT, detailed in Chapter 7, may be improved alongside the SocialLink pipeline and approach to support future releases. The UI component can be improved significantly to cover the entirety of the API functionality and additional downstream tasks can be implemented to assist in testing of the pipeline. Greater variety of downstream tasks can help find and resolve corner cases in SocialLink making it more robust and useful. For example, recent works in emoji prediction (Coman et al., 2018) suggested the usage of user-based features, including the ones coming from DBpedia, to improve classification performance. Pipeline-wise, possible improvement directions include replacing Wikimachine with Tint 2.0 (Aprosio and Moretti, 2018) as a default option for Italian and extend SMT to support other languages.

8.3 Privacy

The SocialLink project is based on processing terabytes of data about hundreds of millions of people on Twitter. When designing, implementing and releasing any of the resources and approaches based on such data, privacy of those individuals have to be taken into account. In this section, I will address some of the privacy aspects that have to be taken into account when working on or with SocialLink and related approaches. Throughout this

thesis, I employ only the publicly available data extracted from Twitter Streaming API. This API provides the same random subset of tweets for all clients connected to it and is routinely used by researchers and companies for many social media analysis, general NLP and other tasks. Subsets of this data have been released as part of the various datasets before.

Even though the release of the raw Twitter data (or at least a portion of it) would benefit the reproducibility and potentially bolster the adoption of both SocialLink and other works described in this thesis, I believe that it wouldn't be appropriate to do so. Once published, there will be no way for the owners of data (i.e., users) to remove, modify or otherwise control the dissemination of it. For this reason, it would have been irresponsible in my opinion to needlessly decrease the ability of users to control their data even if there is a possibility that some other company or researcher may eventually expose the same information publicly anyway. Additionally, such data release may violate Twitter API's terms of use,[3] which also address the user's right to control the dissemination of their data.

Therefore, the SocialLink resource only releases the internal Twitter user identifier and public user handler without releasing any of the profile information or other user data used during linking, such as the estimated social graph or textual content posted by the user. This is a bare minimum of data that is needed to actually establish the link: the public user handler is needed to fill the existing properties in the ontology such as dbo:isPrimaryTopicOf or foaf:account, while the release of the internal Twitter identifier protects the link in case the handler was changed. For the same reason, when releasing data about our type prediction approach,[4] I omit social media features. Such features can be recomputed using the provided code given sufficiently large sample of Twitter Streaming API.

During the discussion around the soweego[5] project, which aims at embedding the SocialLink pipeline into Wikidata, it became apparent that given the particular use case for linking, and the policies and notability requirements of Wikipedia and related initiatives, even stricter privacy requirements have to be put in place. To this end, the linking between Wikidata and Twitter as part of the project was restricted to "verified" profiles to only include people that were willingly publishing their content online in their official (professional or otherwise) capacity.

As mentioned before, SocialLink enables knowledge transfer between the LOD and social media. In Chapter 4, I have argued that the SocialLink resource is able to make user profiling pipelines and other potentially privacy-intrusive tasks in social media analysis much easier to implement. I confirmed this idea by implementing the interests inference

[3] https://developer.twitter.com/en/developer-terms/policy
[4] https://w3id.org/sociallink/type-prediction
[5] https://meta.wikimedia.org/wiki/Grants:Project/Hjfocs/soweego

pipeline targeted at passive users in Chapter 6. However, as shown in the same chapter, SocialLink can also be used as a cornerstone of the novel approaches aimed at protecting users from inferring their private identity. I urge researchers that would like to build systems on top of SocialLink to always consider the privacy impact of their approaches and develop privacy protection mechanisms instead of the privacy breaching ones.