SUMMARY

The digital era and the omnipresence of computer systems feature a huge amount of data on identities (profiles) of people, organizations, and other entities, in a digital format. This data largely consists of textual documents, such as news articles, encyclopedias, personal websites, books, and social media, thus transforming identity from a philosophical to a societal issue and motivating the need for robust computational tools that can determine entity identity in text. Determining the identity of the entities described in these documents is non-trivial, given their amount, the contextual dependency of these descriptions, the ambiguity and variance of language, and the interplays described in widely-accepted pragmatic laws of language, such as the Gricean maxims (Grice, 1975).

Today, it is well-understood how to determine identity of entities with low ambiguity and high popularity/frequency in communication (*the head*), as witnessed by the high accuracy scores in the standard Natural Language Processing (NLP) task of Entity Linking. It is unclear, however, how to interpret *long-tail* entities: each different and potentially very ambiguous, with low frequency/popularity, and scarce knowledge.

Expecting that computational systems that establish identity in text struggle with long-tail cases, this thesis investigated how the performance of NLP techniques for establishing the identity of long-tail cases can be improved through the use of background knowledge. It focused on five aspects of this challenge: description/definition, analysis, improvement of evaluation, enabling access to more knowledge, and building knowledge-intensive systems. Concretely, the research questions and corresponding findings of this thesis were:

- *How can the tail entities be distinguished from head entities?* Our experiments showed a positive dependency of system performance on frequency and popularity of entity instances, and a negative one with ambiguity of surface forms. Essentially, this confirms the intuition that system performance is largely based on head cases, and declines strongly on the tail.
- *Are the current evaluation datasets and metrics representative for the long-tail cases?* The commonly used datasets to evaluate disambiguation and reference NLP tasks lack representativeness, as they suffer from low ambiguity, low variance, high dominance, and limited temporal spread.
- *How can we improve the evaluation on the long-tail cases?* On a deliberately created task to evaluate tail instances, we observed very low accuracy of the participating systems. This shows that dealing with high ambiguity and not being able to rely on frequency biases, poses a great challenge for current NLP systems.
- *How can the knowledge on long-tail entities be accessed and enriched beyond DBpedia?* By creating LOTUS, we provided the Linked Open Data community

with the largest centralized text index and access point to the LOD Laundromat data collection. This allows EL systems to use the knowledge found among the billions of statements of the LOD Laundromat collection, thus essentially increasing their recall on the tail instances.

- *What is the added value of background knowledge models when establishing the identity of NIL entities?* Neural background knowledge models ("profiling machines") were built and applied in order to complete the partial profiles extracted from text and establish their identity. The evaluation of these machines on the task of establishing long-tail identity in text showed promising results when applied on top of automatically extracted information from text. We observed no clear patterns between the effectiveness of our profilers and the data ambiguity.

This dissertation thus provided novel insights into an under-explored and difficult NLP challenge: determining identity of long-tail entities in text, demonstrating that better evaluation and more extensive use of knowledge are promising directions forward. The topics covered and the skills employed stemmed from various AI fields: semantic NLP, semantic web, and neural networks, with links to linguistics and philosophy. Besides summarizing a range of learned lessons that are potentially applicable to a number of other disambiguation and reference tasks, this thesis also provoked a long list of future research directions.