Conclusion to Part III

In the final part of the thesis, we addressed the research theme relating topics in different representations. More specifically, we focused on the task of Automatic Link Generation (ALG) with Wikipedia, which aims to identify significant terms or phrases in a piece of text, and for each term or phrase, generate a link to a Wikipedia page that provides background information for the term or phrase. Machine learning approaches using existing Wikipedia links as training data have shown satisfying performance on the related problem of (re)generating links between Wikipedia pages. In this part of the thesis, we evaluated “learning to link with Wikipedia” approaches in two different settings.

First, in Chapter 8 we evaluated the learning approaches in a setting where the task of ALG was formulated as a ranking problem, that is, for a given source text, links are identified and ranked according to their relevance to the source text, and for each anchor text, target pages are ranked according to their relevance to the anchor text. Moreover, the resulting links were evaluated against manual assessments in stead of Wikipedia ground truth, i.e., existing Wikipedia links. Our main findings within this setting are as follows. (i) Linking models trained on a more recent Wikipedia collection (2009) which is of larger size and has more links achieve better performance compared to that achieved by models using an older Wikipedia collection (2008). (ii) Using a ranking based model (i.e., RankingSVM) does not outperform a binary classification based model (i.e., binary SVM), although the goal is to return a ranked list of links. (iii) When evaluating against human assessments, both Wikipedia ground truth and the links generated by models learnt from the Wikipedia ground truth are far from perfect.

Second, in Chapter 9 we turned to a second setting, where we aimed to investigate whether ALG systems that are trained domain independently can effectively link texts from a specific domain to Wikipedia. We conducted a case study in the radiology domain. We found that directly applying the domain independent ALG systems to the radiology data does not yield satisfying results. Further, our proposed ALG approach that considers domain specific properties of the radiology data has shown to effectively improve over the domain independent ALG systems.