The submission of the article „Techniques for High Quality Character Reference Detection on German Historical Novels“ deals with a problem many researchers face: The lack of data in their target domain. In this work, we outline four different methods in order to obtain a suitable automatic component for the recognition of character references in a different domain and in our case even a different label set. We show, that depending on one’s goals and capabilities, Active Learning can be a good choice if one solely wants to minimize the amount of data to be labeled. Using techniques from semi-supervised learning, we were able to integrate prior knowledge in the form of gazetteers and obtained as classifier which was able to reach a classifications F1-score of about 82%. Aside from that, a clever rule based algorithm was able to reach more than 90% F1-score and its global features even showed beneficial when integrated as features into the machine learning algorithms, showing that rule based systems can still perform on par with modern day machine learning algorithms, while not relying on training data.

The detection of character references is a crucial step for the analysis of character networks and other downstream tasks such as coreference resolution, and the results of this work should outline important steps to take in mind when facing a different problem to be automated.