Identity of Long-Tail Entities in Text

Filip Ilievski

The digital era has led to a large amount of textual data on identities of people, organizations, and other entities. A digital format is often needed to determine identity, especially in cases with low accuracy and high variability. This is witnessed by the NLP (Natural Language Processing) task of linking, which is more specific to long-tail entities. Each different aspect of this task is challenging, with low-frequency and low-popularity entities. This thesis investigates how the performance of NLP techniques for establishing identity of long-tail cases can be improved through the use of background knowledge. It focuses on five aspects of this challenge: ambiguous data, low-frequency, low-popularity, and semantic knowledge. The techniques and ideas stem from the fields of NLP, Semantic Web, and Machine Learning, with links to linguistics and philosophy.