

Protokoll zur Sitzung am 12.06.17

Title: „Analysis of NIL Results in an Entity Linking System“

Student: Mai Linh Pham

Betreuer: Yadollah Yaghoobzadeh

At the begining of the presentatio Pham explain some importants definitions in order to understand later some concepts of this thesis. Entity Linking is the process of linking mentions from text to corresponding entity in knowledge base (KB). NIL Results are entities that are not linked to the KB Fine-Grained tag set has more tags than standard NER tags (LOC, ORG, PER, MISC).

Entity Linking systems can not link all entities, that is why some entities are missing from KB In order to improve EL systems, NIL results should be analysed and for that it should be introduced a new method for clustering. The goal of this thesis is to examine whether fine-grained types are useful for clustering and analyzing NIL mentions.

For this purpose Pham describes her task as follow:

- Combine outputs of
 - (1) an entity annotation tool and
 - (2) an entity linking system
- Extract NIL output
- Cluster NIL output
- Use fine-grained types for clustering task

She uses many tools in this thesis. One of them is FIGER (fine-grained entity annotation system) with 112 tags. This tool allows overlapping types and its output are in Standard BIO-encoded tags. Another tool is WAT (entity linking system). Their components are spotter (scans input text for mentions, retrieves list of candidate entities), disambiguator (ranks candidate entities with different disambiguation algorithms) and pruner (removes useless annotations, aims at increasing the precision). The WAT output is in JASON-format, which has to be in same format as FIGER output.

The student also the Extract NIL mentions, which we mention as follow:

- Unlinked mentions: mentions that are annotated by FIGER but not linked by WAT
 - create list of all entity names in the KB: all Wikipedia titles + all Wikipedia redirect links
 - regard only unlinked mentions that have a corresponding KB entry.

The Cluster NIL Mentions has three clustering approaches:

- (i) coarse-grained type
- (ii) fine-grained type
- (iii) top-level type

The Fine-Grained Clustering divides types into multi-level and single-level types. Also this task clusters types semantically, maps single types to multi-level types and clusters multi-level types.

During the Top-level Clustering many first-level tags in multi-level types are identical. In this case Pham groups these to reduce the number of clusters. But with that action too specific types to be cluster are grouped as *undefined*. That is the reason why she has to create new domains to reduce these undefined types.

She concluded that fine-grained entity types can be used for clustering semantically related NIL mentions. For that it has to be taken into account lexical and contextual properties of entity. She realized that the information anchored in tags can be used for analysis. Fine-grained entity types are more informative than coarse-grained types.