We described entity linking over nested named entities for Russian NEREL dataset

- 933 Wikinews texts
- Nested named entities: longer entities can include shorter entities
- Nesteness of entities enables a more accurate and complete description of relations and links to knowledge bases
- 29 Entity types, more than 56K entities
- 49 relation types
- Wikidata entity links over nested named entities

Entity Linking Annotation

- 16 entity types of 29 are linked to Wikidata
- Numerical and temporal entities are excluded
- 38 thousand entities are linked
- Only one entity mention per document should be annotated (40% reduction)
- Barack Obama, Obama, Obama, ...
- Mentions clusters corresponding to the same entity are created
- Mentions are identified according to
  - The same lemma representations
  - Linking via relations
  - ALTERNATIVE_NAME
  - ABBREVIATION

Results of automatic pre-annotation

Automatic pre-annotation:

- Manual entity links from initial Wikinews texts
- Ranking list of Wikidata entities generated by Elasticsearch retrieval engine
- Page view statistics to exclude noisy candidates
- Matching NEREL named entity types to Wikidata general concept
  - CITY – city/town (Q7930989)
  - AWARD – award (Q618779)
  - Wikidata link should correspond to matched superconcept

If Wikidata link for entity is not found -> special NULL link

Knowledge graph generation from texts

- Named entity recognition
- Relation extraction
- Entity linking

- NEREL – the largest Russian dataset for supervised knowledge graph construction from texts
- Annotation is based on nested named entities
- Relations between nested named entities
- This work is annotation of the third level: annotating Wikidata links

Principles of Linking

- Entity linking annotators rely fully on existing annotations of named entities.
- Some errors can be corrected after agreement with moderator.
- If an entity is absent in Wikidata, then it should be linked to NULL, but its internal entities may still have corresponding links.
- Mayor of Novosibirsk -> NULL link
- Mayor to Q30185, Novosibirsk -> Q683
- Professions are linked to corresponding professions pages
- Nested named entities allow for annotation of so-called iterations of entities
  - {111th [U.S. Congress], ORG}
  - 111th U.S. Congress -> Q170375, U.S. Congress -> Q11258

Linking of Adjectives

- In NEREL adjectives are annotated with entity types of corresponding named entities
  - Moscow -> CITY, Moskovskiy -> CITY.
  - This provides large coverage for establishing relations
- Adjective annotated as named entities are linked to entities according to corresponding nouns
  - Moscow (Q649) -> Moskovskiy (Q649)
  - Such cases are difficult for automatic linking
- Especially difficult for automatic linking nationality-related adjectives (Russian), which can mean in different contexts
  - Nationality, Language, Country

Conclusion

- We described entity linking annotation within the NEREL dataset, the largest Russian dataset for information extraction.
- Entity linking annotation to Wikidata items is provided for 933 documents, 16 entity types, and 38,152 entity mentions.
- The annotation contains a significant share of nested named entities (more than 17%)
- Currently, NEREL is the only Russian dataset with three levels of annotation.

NEREL dataset

- 38,152 entity mentions.
- 38,152 documents, 16 entity types, and 38,152 entity mentions.
- The annotation contains a significant share of nested named entities (more than 17%)
- Currently, NEREL is the only Russian dataset with three levels of annotation.