**ABSTRACT**

- we present an upgraded version of the Hungarian NYTK-Nerkor named entity corpus, which contains
  - about twice as many annotated spans, and
  - 7 times as many distinct entity types as the original version
- we used an extended version of the OntoNotes 5 annotation scheme
- we also trained and released a transformer-based NER tagger for Hungarian

**RESOURCES**

- many legacy named entity corpora contain an annotation distinguishing four entity types: ORG, PER, LOC, MISC → all existing Hungarian corpora use this schema
- OntoNotes 5: GPE, FAC, PROD, LAW, EVENT, WORK OF ART, dates and times, cardinal and ordinal numbers, quantities, percentages and amounts of money.
- NORP, LANGUAGE
  - biomedical, non-English NER corpora
  - token-level annotation vs nested entities

**ANNOTATION METHOD**

1. created by the DeepPavlov team fine-tuning multilingual BERT
2. based on XLM-RoBERTa

**Error Analysis and Automatic Error Correction** using regular-expression-based patterns

- NameTag 2: Czech model of the NameTag 2 neural named entity tagger trained on the Czech Named Entity Corpus CNEC 2 + fine-grained hierarchy of entity classes having many subclasses within the broader categories

A lemmatized named entity list and automated correction patterns

- all alternative analyses for each entity along with their corpus frequencies → identifies elements frequently misclassified in the OntoNotes model, and entities that should be assigned to distinct classes (e.g. MEDIA, SMEDIA)

**Manual error correction**

- corrected anomalies in the lemmatized named entity list,
- resolved contradictions of the annotations in the original corpus and those generated by the transfer models
- normalized the annotation for references to legislation in the law subcorpus

**FEATURES OF THE CORPUS**

The number of distinguished entity types increased 7-fold while the number of entities marked almost doubled

**MODELS AND PERFORMANCE**

- the zero-shot performance of the transfer-based models
- evaluation with the tagset normalized to the tags present in the original model
- a neural tagger model based on the Hungarian huBERT contextual language model on the training set of the corpus
- performance of the best model trained on Nerkor-Cars on each entity type compared to performance a similar model on the original Nerkor annotation

**THE NERKOR CORPUS**

- size of over one million tokens,
- has a broad coverage of domains and topics
- consists of five 200,000-token corpora: fiction, legal, web, news, wiki
- one-third of the corpus contains coherent text, the rest is a shuffled collection of unrelated sentences or sentence fragments,
- 20% of the corpus includes morphological annotation as well