Spoločný český a slovenský workshop projektu European Language Grid

Počítačové spracovanie jazyka na Slovensku

Radovan Garabík

Jazykovedný ústav Ľ. Štúra SAV

2021-10-18

Akademická sféra

- ťažisko výskumu doteraz
- primárna orientácia na písaný jazyk
- primárna orientácia na slovenčinu
- vieme, ako je financovaná veda na Slovensku

Načo je NLP

- pre (počítačového) lingvistu: strojové učenie pomáha pri výskume, analýze jazyka a anotovaní jazykových zdrojov
- pre dátového vedca: anotované jazykové zdroje pomáhajú pri strojovom učení (chceme použiť aj prirodzený jazyk ako dataset)

NLP

- v súčasnosti prebieha (ďalší) „paradigm shift“ v NLP
- veľa neoznačkovaných dát, málo (prácne) označkovaných dát
- Deep Learning, RNNs, transformers (BERT, GPT-x), unsupervised pretraining, supervised fine-tuning
- hw náročnosť
- 2021: posun od akadémie ku komerčnej sfére
- foundation models
- big players: Google, MS, Facebook, SAS

Korpusy

- Slovenský národný korpus (2002 –) 1.3 G `https://korpus.juls.savba.sk`
- webové korpusy (Aranea...) 4 G `http://aranea.juls.savba.sk`
- manuálne lematizovaný morfologicky anotovaný korpus, 1.2 M
- Universal Dependencies
- špecializované korpusy
  - hovorený `https://korpus.juls.savba.sk/shk.html`
  - historický
  - nárečový
  - paralelné
- speech corpora
  - ÚI SAV
  - KEMT TUKE
- webové korpusy
- Open Data

Základy – lematizácia, MSD

- tokenizácia (slová), segmentácia (vety) – takmer triviálne, ale viď SentencePiece...
- lematizácia, morfologická anotácia (alebo POS)
  `https://morphodita.juls.savba.sk`
  `https://www.juls.savba.sk/bezdiak/`
  - JÚĽŠ: slovník tvarov slov (morfologická databáza), 111k lemmas, 3.6M entries, 1.3M unique wordforms; morphodita
  - tvaroslovník
  - MorfFlex SK
  - ajka
  - ElasticSearch (essential data)
  - Ardevop
- morphodita
- v praxi „vyriešený problém“

Syntax

- (závislostná) syntaktická analýza (parsing)
- JÚĽŠ: manuálne syntakticky anotovaný korpus, 1 M tokenov, 70 k viet
  `https://korpus.sk/synt.html`
- UDPipe `https://lindat.mff.cuni.cz/services/udpipe/`
- dl4dp

Sémantika

- sémantická analýza
- klasika (ontológia): WordNet (10k)
  `https://korpus.juls.savba.sk/WordNet.html`
- word embeddings ($\approx$ 20 jazykov) `https://www.juls.savba.sk/semä.html` (pre ľudí – vyhľadávanie, vizualizácia)
- ale aj *Lexemes in Wikidata* a *Wikifier*

Pomenované entity

- Named Entity Recognition `https://www.juls.savba.sk/nerd/`
- Ardevop
- TUKE FEI (SpaCy)
- ručne značkovaný korpus
- entity linking
- anonymizácia

Speech

- ÚI SAV
- KEMT TUKE
- Newton Dictate

Rôzne

- sentiment analysis
  - Sentigrade (FIIT)
  - JÚĽŠ SAV
- chatbots

Jazykové modely

- sk-bert (Ardevop)
- SlovakBERT (KInIT)

Katalóg zdrojov

- META-SHARE (neaktuálny!) `https://metashare.korpus.sk/`
- `https://github.com/essential-data/nlp-sk-interesting-links`
- `https://github.com/slovak-nlp/resources`
- European Language Grid

Ďakujem za pozornosť