WP 1.3 NOISE-TOLERANT NLP

Veronika Laippala
Jenna Kanerva
Valtteri Skantsi
Anna Salmela
Sampo Pyysalo
Filip Ginter
TurkuNLP, University of Turku
first.last@utu.fi

BACKGROUND AND GOALS

- Natural language processing (NLP) tools can process standard language with extremely high performance
- The performance deteriorates when facing noisy, non-standard input
- This WP develops **datasets** featuring noisy language and **NLP tools** that can survive noise
- Noise can originate from non-standard language use and from the digitization process
  - Historical language varieties, dialects, spoken genres
  - OCR noise, boilerplate remains from Web data, other non textual input

NOISY LANGUAGE USE FROM THE WEB

- Web-crawled datasets include a wide range of noisy language use
- *Register* (Biber 1988) classification provides pathways to extracting potentially noisy text categories
- **Finnish Internet Parsebank** includes nearly 9B words from the Finnish Internet (Luotolahti et al. 2015)
  - Now register-labeled using the FinCORE corpus and XLM-R
  - Register classifier performance 79 % F1-score
- **Oscar** is a a huge multilingual corpus obtained by language classification and filtering the Common Crawl
  - Now register-labeled using a multilingual register model and XLM-R (Rönnqvist et al. 2021)
  - Register classifier performance 77 % F1-score
  - Available as 🤗 dataset

Register labels in the web datasets

Narrative
news report / news blog, sports report, personal blog, historical article, short story, travel blog, community blog, magazine / online article

Opinion
opinion blog, review, religious blog / sermon, advice

Informational description
job description, FAQ, description of a thing, information blog, description of a person, research article, legal terms / conditions, course material, encyclopedia article, report

Interactive discussion
discussion forum, question-answer forum

How-to / instructions
recipe

Informational persuasion
description with intent to sell, news-opinion blog / editorial

Lyrical
poem

Spoken
interview, formal speech

Machine-translated / generated text

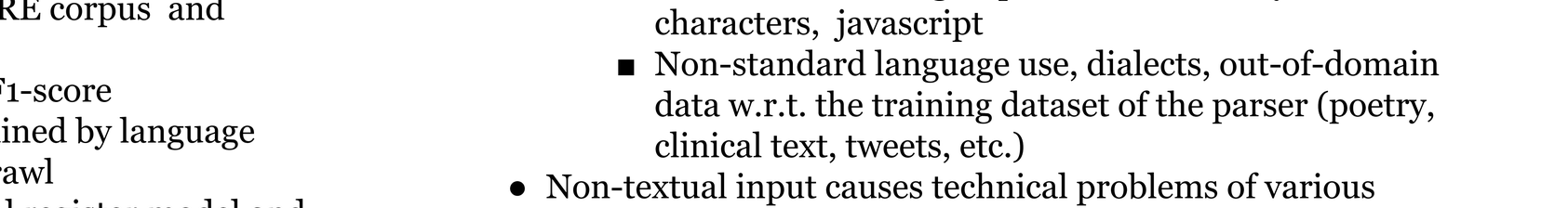
Noisy text identified as discussion forum

Onko sulla haaveena olla joku esim tollanen joka työskentelee laboratoriossa? Niin mä olin joskus viime vuonna kun olin ysillä niin apteekissa ja siellä oli ihan sika kivaa :) Porukka oli mukavaa ja sain järkätä siellä niitä lääkkeitä ja kirjata ylös kun niitä saapu ja kirjoittaa reseptejä. Ja etsiä lääkkeitä niistä huikeen pitkistä…

NOISE-RESISTANT SYNTAX ANALYSIS

- Turku Neural Parser Pipeline is the state-of-the-art Finnish dependency parser with a LAS of 93%
- Different types of noise
  - "Non-textual": long sequences of arbitrary unicode characters, javascript
  - Non-standard language use, dialects, out-of-domain data w.r.t. the training dataset of the parser (poetry, clinical text, tweets, etc.)
- Non-textual input causes technical problems of various kinds because it does not correspond to the assumptions built-in the parser code, especially various maximum lengths in deep learning models
- Out of domain data may go beyond the generalization ability of the parsing model, it is important to understand the limitations
- What we have done:
  - Substantial testing of the parser pipeline code on real Internet data - increased robustness towards malformed textual input. Resulted in new fully-reparsed Suomi24 dataset for Finnish
  - New dataset and manual evaluation on out of domain data: J. Kanerva & F. Ginter (2022) *Out-of-Domain Evaluation of Finnish Dependency Parsing*. Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC'22)

Dependency tree from the clinical domain



References

Biber, D. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge.

Luotolahti, J. Kanerva, J., Laippala, V., Pyysalo, S., Ginter, F. 2015. Towards Universal Web Parsebanks. *Proceedings of the Third International Conference on Dependency Linguistics* (Depling 2015).

Rönnqvist, S., Skantsi, V., Oinonen, M. Laippala, V. 2021. Multilingual and Zero-Shot is Closing in on Monolingual Web Register Classification. *Proceedings of the 23rd Nordic Conference on Computational Linguistics* (NoDaLiDa).