Thesis projects organized by research theme

Marijn Schraagen

NLP Group Meeting

April 10th, 2025

Universiteit Utrecht

Research interests

- Feedback on every Assessment & Development interview: choose one research topic and focus on that
- Didn't happen so far
- Some of my research interests (in alphabetical order):
  - digital humanities
  - legal NLP
  - low literacy/language proficiency
  - medical NLP
  - sociolinguistics

Universiteit Utrecht

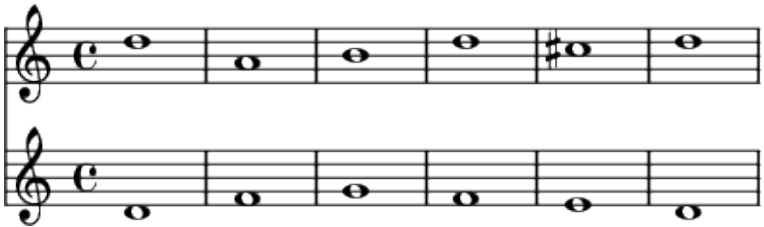
Today's presentation

- A selection of student thesis projects I'm currently involved in
- Grouped by these themes
  - bachelor/master
  - first/second supervisor
  - AI/BI/ADS

Universiteit Utrecht

Counterpoint for composer detection

- Abe Wolthuis, master AI, 2nd supervisor, proposal finished
- Renaissance and Baroque music uses **counterpoint**
- Large set of rules on which notes are allowed to be played together
  1. Begin and end on either the unison, octave, or fifth
  2. Use no unisons except at the beginning or end



Universiteit Utrecht

Composer detection

- Not all composers always followed all rules
- Rule-following patterns as distinctive signature/style of the composer
- Used to detect the composer of unattributed 15th/16th century music pieces
- Thesis project: supervised classifier to identify counterpoint rule usage
- Use as input for composer attribution
- Emphasis on **explainability**: this composer is predicted because of counterpoint rules a, b, c
- Dataset: 800 compositions by a large number of composers
- Various machine learning models

Universiteit Utrecht

Semantic data validation

- 3 ADS master students, 2nd supervisor, start block 4
- One student works at UWV to collect unemployment contributions from employers
- Employers submit forms to UWV with financial details
- These forms are checked for conformance with regulations and laws
- UWV keeps a set of rules to do the checks
- Each year these rules need to be updated based on new legislation and interpretation by the tax authority
- Currently manual process

Rule example

Handbook Salary Premiums (from tax authority)

29.30.1 Disability fund (DF)
Within one category and time period you may provide either *increase DF low* or *increase DF high*. The other rubric should be entered as €0.

UWV ruleset

`rule 2252` IF *increase DF high* not equal to €0 THEN *increase DF low* must be €0
`rule 2254` IF *increase DF low* not equal to €0 THEN *increase DF high* must be €0

Universiteit Utrecht

Project

1. Use information retrieval/text similarity methods to match handbook paragraphs and rules
2. Check if rules are obsolete or missing
3. Check if rules are (still) correct
4. Check against applicable law texts (maybe next year)

- Use existing rules in original or modified form to test
- Expert annotations
- Project starts in block 4
- Project also ends in block 4

Universiteit Utrecht

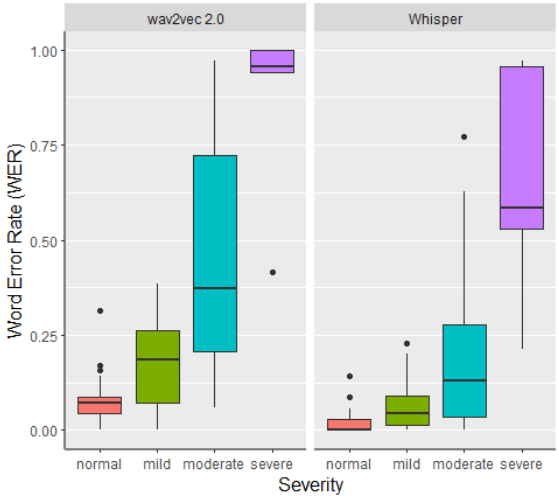
Classifying pathological speech

- Luuk Smiesing, bachelor AI, 1st supervisor
- Classify speech disorders in Dutch COPAS dataset
- 8 pathological speech categories
- Wav2Vec2
- Audio SHAP to see which part of the speech explains the predicted class

Universiteit Utrecht

ASR on the COPAS dataset

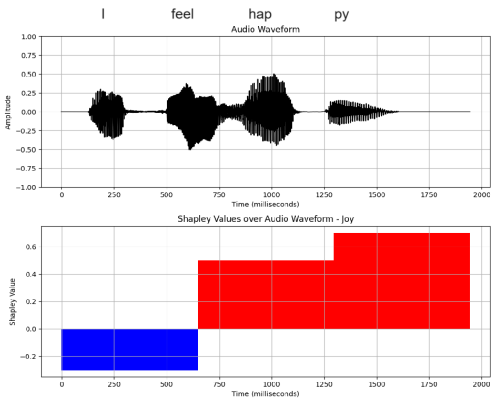
Leo Verspaij 2024 (thesis Radboud Universiteit Nijmegen):



Universiteit Utrecht

Audio SHAP

Kathleen de Boer 2024 (thesis UU):



Universiteit Utrecht

Readability

- Michiel Schouten, bachelor AI, 1st supervisor
- Existing tools for increasing readability of text
- Survey: do these tools align with human choices for improving text readability
  1. Do people agree with the changes proposed by such tools?
  2. What kind of changes would people make themselves?
- Improve tools by not doing types of changes people don't like, implement additional types of changes, apply controversial changes interactively
- Dataset: letters from housing corporations
- Data preparation: run Dutch tool T-Scan and rewrite passages according to T-Scan scores

Universiteit Utrecht

Survey questions

- Text presented, randomly either original or rewritten
  1. How readable is the given text on a 5 point scale?
  2. Do you think the text is patronizing?
  3. What would you change in the text?
  4. (next page) Given the following changes, which would you keep?

  Kindly be invited to an orientation meeting on April 12, 2025 at 12:00 with Alice in our office in Maastricht.

  I would like to invite you for a meeting on April 12, 2025 at 12:00 with Alice. The meeting takes place in our office in Maastricht.
- Change types (not shown to participant): active verbs, remove adjectives without clear added value, short sentences

Preliminary results

- Differences found between acceptance of change types
- Participants (n=20) do not always agree
- Surveys are difficult

Extract medication prescriptions from EHRs

- Maria Dukmak, master AI, 2nd supervisor, proposal phase
- Data and main supervision at UMCU
- Electronic Health Records (EHRs) contain unstructured text with codes and abbreviations
- Medication dosage, duration, usage instructions
  - example usage instructions: 1D1T 3W1WS, 1-2D1T, 1d1t, ???
  - partially standardized
- Goal: transform this into a Common Data Model (CDM) standardized format for use in daily patient care and epidemiological research

Universiteit Utrecht

Constraints

- Domain-specific Dutch data
- Relatively small, 3M EHRs, 1.8M after cleaning (missing or uninformative values)
- Manual annotation for target labels in CDM
- Data cannot go outside trusted UMCU digital research environment (DRE)
  - Limited computing power
  - Limited internet access
  - Limited pipeline/packages support
- BERT-based models and small generative LLMs

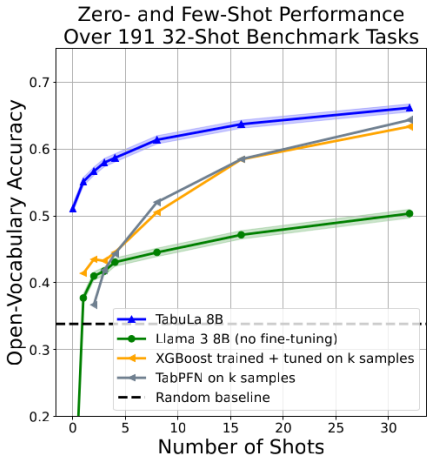
Universiteit Utrecht

Prediction of treatment outcomes and symptom changes

- Omar Taha, bachelor AI, 1st supervisor
- Collaboration with Amsterdam UMC (data, previous work)
- Predict patient responses to psychiatric treatments
- Dataset: tabular sociodemographic, clinical, neuroimaging data from OCD patients at AMC
- Target variables: predict treatment outcome and symptom changes
- Traditional supervised machine learning models
- Recent models: Tabular Prior-data Fitted Network (TabPFN), TabuLa-8B
  - Tabular foundation models: LLMs pre-trained on tabular data with row-based masking

Universiteit Utrecht

Tabular LLMs



Gardner et al. 2024, `https://arxiv.org/abs/2406.12031`

Propaganda detection: entity framing

- Lisa Tanaka, master AI, 2nd supervisor, proposal finished
- **Propaganda**: communication type with the goal of getting people to think or act a certain way
- **Framing**: representing different aspects and perspectives in a certain way in order to convey latent meaning
  - Paying taxes when a person inherits an estate: calling it "death tax" conveys frame of unfair punishment for dying, "estate tax" conveys the frame of justified taxation upon obtaining property
- Data: SemEval 2025 task 10 subtask 1 on Entity Framing
- Approach: feature-based and/or generative LLM based

Universiteit Utrecht

Code Switching

- Nizar Haroun, master AI, 1st supervisor, phase 2
- **Code switching**: a switch is made between two or more languages, either within a conversation or written text
  - *And a scubadiver houdt een schilderij vast*

1. Train supervised classifier on small annotated dataset to detect presence of code switching
2. Apply classifier on large non-annotated dataset
3. Identify interesting linguistic or semantic properties of code switched sentences

Universiteit Utrecht

Approach

- Dataset for training: Kootstra 2010, ∼1100 sentences collected in controlled lab experiment with NL/EN code switching
- Models: T5, mBart
- Dataset for application: collection of Dutch subreddits
- Linguistic/semantic questions:
  - What is the frequency of code-switching, what are the dialogue patterns?
  - Is the sentiment of code-switched sentences different from monolingual sentences?
  - Which topics are more prominent in code-switched data (using topic modelling, clustering)?
  - ...
- Influence of speaker also interesting but this requires different data

Universiteit Utrecht

Thesis projects

- Let me know if you are interested to be first or second supervisor on future thesis projects in these topics
- Or to sit in on a meeting or the defense of the current students
- Or if you want to get in touch with any of the partners
  - UWV, UMCU, AMC, ELSA Lab, smaller companies

- Questions?