DutchMed: a pre-trained clinical language model

First steps: data collection using text classification

Bram van Es, Marijn Schraagen, Forrest Garretsen

UMC Utrecht, Utrecht University

Summary

- **Goal:** Domain-specific Dutch clinical large language model
- **Motivation:** high in-domain performance, safeguard against potentially harmful output, tractable computing requirements
- **Applications:** classification, text embeddings, summarization, medical chatbot
- **Data collection:** locate documents on medical topics in large general-purpose Dutch corpus, collect open Dutch medical texts, translate English clinical corpora
- **Research plan:** Cleaning, deduplication, de-identification, pre-training, fine-tuning

Data collection

- Collect medical documents from many resources
- First resource: SoNaR-500, general-purpose 500 million word Dutch corpus
- Filter for medical documents
- Seed corpus: 2000+ articles from Dutch Journal of Medicine (Nederlands Tijdschrift voor Geneeskunde)
  1. NTvG preprocessing: lower case, remove stopwords, punctuation, numbers, special characters, single-letter terms
  2. tf/idf to extract 5 most distinctive terms in each NTvG article
  3. Clean term list: manually remove non-medical terms
  4. Result: 4864 unique terms
  5. Filter SoNaR-500 on these terms, keep documents with various term matching thresholds

bot	hulpverlener	progressie
braken	laparoscopiegroep	rotatieverschil
cavhd	maagkanker	slaapapnoe
novartis	neuralebuisdefect	schildkliercarcinoom
colostoma	medicamentosa	conversiestoornis
skelet	clindamycine	transplantatie

- Label SoNaR documents with $\geq n$ terms as "medical text", all others as "non-medical text"
- Manual check: 64% correctly labeled, 12% borderline
  - 3.5% of SoNaR corpus remains with cleaned term list and n=4
- Fine-tune RobBERT model
  - Generalize over keyword list
  - Use for other corpora
  - F1 scores up to ∼0.95
  - Need to improve filtering: better cleaning and term thresholds
- Text about health and medical topics vs. actual medical text?

Language model

- Train on open data and **if possible** on EHRs, GP notes, pathology reports etc.
  - Need for thorough de-identification
- Autoregressive decoder training
- Masked Language Modelling encoder training
- 100M–3B parameter Transformer++ models

Medical chatbot

- Use cases: summarization of patient information, diagnostic assistance, medical autocomplete
- Guiding principles:
  - **Truthfulness** over completeness
  - **Specialization** over generality
  - **Parsimony** over exactness
  - **Explainability** or inspectability whenever possible
- **Alignment** after pre-training using health records, (conversational) medical corpora, medical protocols
- Augmented data: concatenate original recording and spoken ASR output
- Ranking/correction by medical specialists

Discussion

- Domain-specific vs general-purpose models
- Feasibility, cost, hardware
  - Choice of model architecture
  - Deployment: local or hosted
  - Potential benefit of data pruning
- Accuracy and grounding with medical protocols
  - Augment with knowledge resources like ICD-10
- Legal issues, personal data leaks in generated text

Acknowledgements

The work received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101057849 (DataTools4Heart project).

UMC Utrecht

Universiteit Utrecht