Leveraging Sequence-to-Sequence Models for Semantic Annotation of Dutch Pathology Reports

Siepel M[1, 2*], Burger GTN[3,4*], Voorham QJM[5], Cornet R[3], Calixto I[3+], Vagliano I[3+]

1 Amsterdam UMC, University of Amsterdam, Department of Medical Microbiology and infection prevention, Amsterdam Public Health Research Institute, The Netherlands
2 Amsterdam UMC, University of Amsterdam, Department of Internal Medicine, Amsterdam Public Health Research Institute, The Netherlands
3 Amsterdam UMC, University of Amsterdam, Department of Medical Informatics, Amsterdam Public Health Research Institute, The Netherlands
4 Symbiant Pathology Expert Centre, Hoorn, The Netherlands
5 Palga Foundation, Houten, The Netherlands
*Equal contribution
+Equal contribution

Introduction
The Palga Foundation, responsible for indexing pathology data across the Netherlands, plays a critical role in annotating Dutch pathology reports for both patient care and scientific research. However, manual annotation by pathologists is labor-intensive and prone to errors. In this study, we leverage sequence-to-sequence transformer models, particularly T5-based models, to generate these annotations. Additionally, we investigate a constrained decoding approach to enforce domain-specific rules that pathologists must follow when annotating their reports.

Methods
We pre-trained our own T5 model (PaTh5.NL) using Dutch pathology data to ensure optimal alignment with the task. We fine-tuned the PaTh5.NL model using default decoding (DD) and constrained decoding (CD) and compared these two fine-tuned models with a fine-tuned multilingual T5 model (mT5), which was pre-trained on general language data. Additionally, we compared these models with the Pathology Report Annotation Module (PRAM), a rule-based tool currently employed in several pathology laboratories in the Netherlands. Performance was assessed using BLEU scores for quantitative evaluation and manual evaluation for qualitative assessment.

Results
Quantitative evaluations indicated that our two fine-tuned PaTh5.NL models significantly outperformed the fine-tuned mT5 model and the PRAM, particularly for shorter histology and cytology reports. However, performance declined on longer or more complex reports. Despite achieving higher BLEU scores, manual evaluation revealed that the PaTh5.NL models did not

consistently outperform either the mT5 model or the PRAM in generating the most relevant annotations.

Conclusion

This study demonstrates that fine-tuned T5-based models can enhance the annotation process for Dutch pathology reports, though challenges remain. Pre-training on domain-specific data improves the ability to match current annotation quality. The constrained decoding model shows potential for generating adequate annotations but requires further refinement. Future research should focus on improving data quality and developing post-processing algorithms to enhance the generalization of annotations.