Preprocessing Pathology Reports for Vision-Language Model Development

Ruben T. Lucassen [1 2], Tijn van de Luijtgaarden [3], Sander P.J. Moonemans [3], Willeke A.M. Blokx [1], Mitko Veta [2]

1 Dept. of Pathology, University Medical Center Utrecht, the Netherlands
2 Dept. of Biomedical Engineering, Eindhoven University of Technology, the Netherlands
3 Dept. of Mathematics and Computer Science, Eindhoven University of Technology, the Netherlands

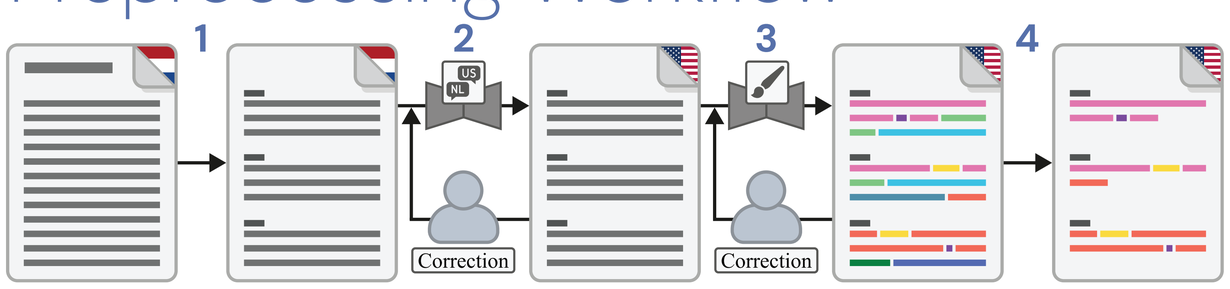
Motivation

Vision-language modeling is gaining attention for histopathology. Whole slide images (WSIs) paired with pathology reports are increasingly being used for developing these models. [1, 2]

Pathology reports often contain information that cannot be derived from WSIs. For tasks like report generation, this can lead to hallucinations (i.e., generated statements that contradict the source content). [3]

As part of a preprocessing workflow for pathology reports of melanocytic lesions, **we developed two language models, one for translation from Dutch to English and one for subsentence segmentation** based on information content. This enables selective use of text data from reports.

Preprocessing Workflow



The preprocessing workflow consisted of the following four steps:

1. Manual de-identification, irrelevant sentence removal, and sectioning.
2. Dutch-to-English translation using iterative model development by expanding the training and validation dataset with manually corrected model predictions for previously unannotated reports, before finetuning the next version of the model (starting from opus-mt-nl-en [4]).
3. Subsentence segmentation using the same iterative model development appoach (starting from FLAN-t5-large [5]).
4. Selection of specific information from the reports.

Segmentation Input & Output

Input `"segment pathology sentence: 'A cutaneous metastasis cannot be excluded for certain, but there is no history of prior melanoma.' Context: 'The Breslow thickness is at least 1.6 mm. <SENTENCE> No ulceration, no regression, no satellites.'"`

Output `"A cutaneous metastasis cannot be excluded for certain, <CON><Nline>but there is no history of prior melanoma.<HIS>"`

Dataset

The dataset consisted of pathology reports for skin biopsy and excision specimens with **melanocytic lesions**, obtained from the UMC Utrecht, the Netherlands.

Set	Reports	Sentences	Words
Training	1,350	20,563	235,194
Validation	150	2,398	27,593
Test	150	3,597	37,465
Unannotated	25,850	179,358	1,980,246
Total	27,500	205,916	2,280,498

Class	Related to
H&E	H&E-stained slides
IHC	IHC-stained slides
IHC+	IHC and H&E
MOL	Molecular testing
CLN	Clinical information
HIS	Medical history
CON	Conclusions and diagnosis
RAD	Radicality assessment
BRS	Breslow thickness
CAL	Calculations and counts
ADV	Treatment advice
SID	Slide identification
UNR	Unrelated information

Annotated Pathology Report

Description
Skin excision specimen reaching into the deep reticular dermis. In the dermis, there are superficial and deep nests of mainly large, epithelioid melanocytes with enlarged nuclei which mostly contain a prominent nucleolus. The nuclei are irregular. In I2, there are up to 5 mitoses per 1 mm2. Towards the base, no maturation is seen. The lesion extends into the basal margin. The distance to the nearest lateral margin is 0.9 mm. Additional stains and molecular diagnostics are requested. Additional immunohistochemistry: Melan-A is positive. The lesion is partly asymmetrical. No clear intraepidermal component. ALK is negative, ROS1 is negative, NTRK is negative, BAP1 is positive, HMB45 is negative, p16 is negative, PRAME is positive.

Discussion
It concerns an atypical, epithelioid melanocytic lesion, somewhat Spitzoid-looking. The lesion has a macroscopic diameter of 5 mm. Based on additional immunohistochemistry, a Spitzoid signature could not be confirmed (ALK, ROS1, and NTRK all negative). NGS shows a mutation in NRAS. The SNP array shows 3 CNV. Often, a total of 3 or more CNVs is regarded as indicative of malignancy. The lesion is, given the morphological and molecular findings, classified as nodular melanoma.

Conclusion
Skin shave excision specimen from the jawline on the right: atypical, epithelioid melanocytic lesion, best to be classified as nodular melanoma. The Breslow thickness is at least 1.6 mm. A cutaneous metastasis cannot be excluded for certain, but there is no history of prior melanoma. No ulceration, no regression, no satellites. pTNM 8th edition 2017: pT2a melanoma. Local re-excision with a margin of 1 cm is advised.

Results

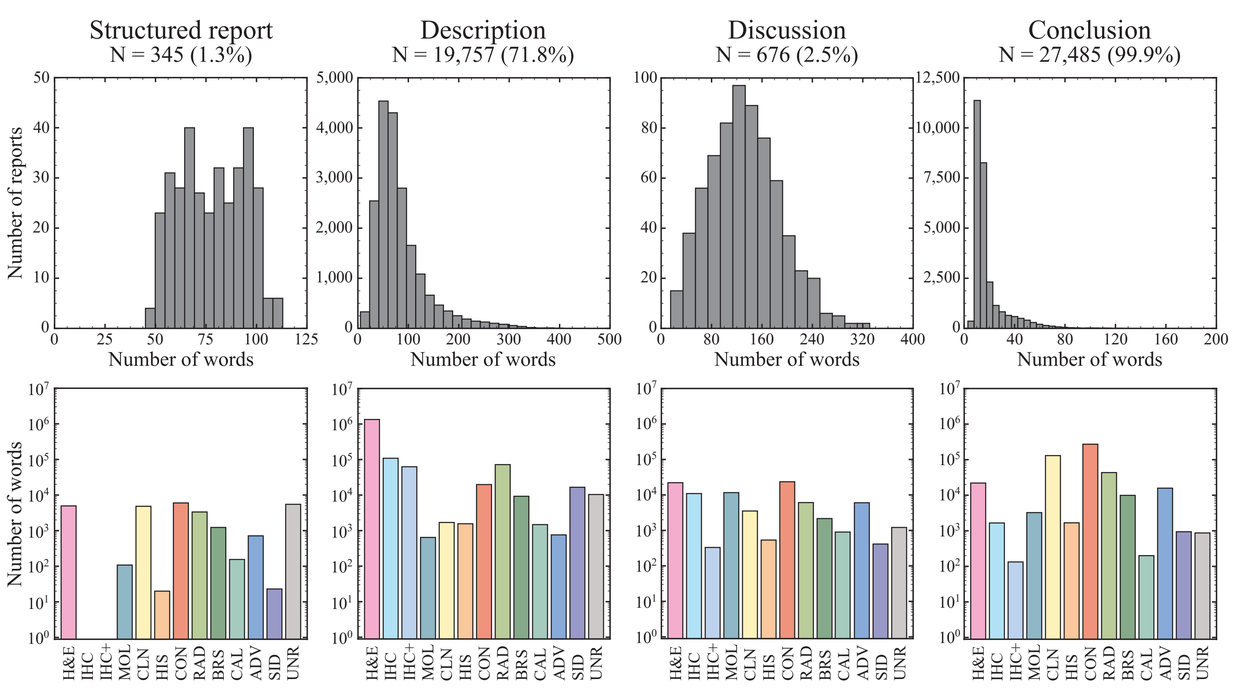
Translation

The translation model was evaluated by counting and categorizing the errors. Translation errors that changed the meaning of the sentence and were undetectable without knowing the input sentence were counted **21** times (0.6%). Spelling and grammar errors were counted **69** times (1.9%). Incorrect translations that were detectable or missing words that did not change the meaning were counted **129** times (3.6%).

Segmentation

The subsentence segmentation model reached a macro average F1-score of **0.921** (95% CI, 0.890-0.940) and a weighted average F1-score of **0.952** (95% CI, 0.944-0.960). The model reached a F1-score of above 0.90 for 11 out of the 13 classes, with only a lower performance for the HIS and IHC+ classes. The model incorrectly reproduced the original subsentence 30 out of 4,660 times (0.6%), which did not affect the predicted class tokens.

Dataset Statistics



[1] Vorontsov et al., "PRISM: a multi-modal generative foundation model for slide-level histopathology". arXiv:2405.10254, 2024.
[2] Lu et al., "A multimodal generative AI copilot for human pathology". Nature, 2024a.
[3] Ji et al., "Survey of hallucination in natural language generation". ACM Computing Surveys, 55(12):1-38, 2023.
[4] Tiedemann and Thottingal, "OPUS-MT-building open translation services for the world". In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, 479-480, 2020.
[5] Chung et al., "Scaling instruction-finetuned language models". Journal of Machine Learning Research, 25(70):1-53, 2024.

HANARTH FONDS
TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY
UMC Utrecht
r.t.lucassen@umcutrecht.nl
linkedin.com/in/ruben-lucassen
github.com/RTLucassen