Patient-friendly Clinical Notes: Towards a new Text Simplification Dataset

Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus, Christin Seifert

TSAR Workshop | EMNLP 2022, Abu Dhabi

Universitätsmedizin Essen
Universitätsklinikum

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Why do we need patient-friendly clinical notes?

Hospitals around the world share clinical notes with patients

Shared decision making

Notes are hard to read for patients

Jargon, detailed analyses, unfamiliar style

Text Simplification can help; But we lack resources

Our contribution: a new parallel dataset

Key facts

851 German pathology reports

Expert simplifications

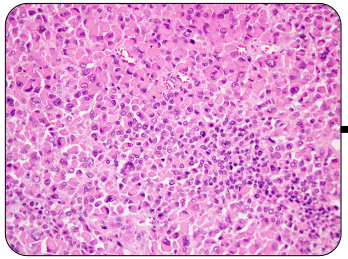
>790,000 tokens

Why is this relevant?

Challenging text characteristics

Document-level simplifications (like Newsela, Xu et al.)

Diversity of resources (clinical German)



Background: Undefined mass proximal thigh left. Suspicion of lymphoma, DDx soft tissue sarcoma. Entity?

Microscopy: The tumor has a relatively monomorphic appearance, with a hemangiopericytoma-like vascular pattern.

Simplify!

Background: The present sample is an undefined tissue growth of the upper left thigh. The question of a diagnosis is raised.

Microscopy: The tumor shows a rather homogeneous picture. The blood vessels grow in a particular pattern.

Talk outline

- Simplification protocol
- Dataset characteristics
- Simplification baselines

Example pathology report

Complex Report

Background:
Undefined mass proximal thigh left. Suspicion of lymphoma, DDx soft tissue sarcoma. Entity?

Macroscopy:
Proximal thigh ventral left: several fragments of beige-brown, partly yellowish tissue of 2 x 2 to 0.3 cm when put together.

Microscopy:
Microscopically, the biopsy shows portions of a spindle-cell shaped tumor. The tumor has a relatively monomorphic appearance, with a hemangiopericytoma-like vascular pattern. The tumor cells have enlarged, slightly vesicular nuclei. Mitotic figures are barely visible (1/10 HPF). The stroma is relatively fine and contains single collagen fibers. Necroses are not detectable. Additionally, immunohistochemical examinations were conducted. The tumor shows strong positivity for CD34 and strong nuclear expression of STAT6. The following antigens are not expressed by the tumor: Actin, caldesmon, pancytokeratin (CKplus), desmin, EMA, MUC4, S100, SOX10, and TLE1.

Conclusion:
BX (proximal thigh ventral left) with a solitary fibrous tumor (SFT), classic type. A supplementary molecular pathological examination (fusion panel) was initiated to validate the findings. There will be a follow-up report on this topic.

Simplification that patients can understand

?

Simplified Report

Background:

Macroscopy:

Microscopy:

Conclusion:

Translated to English (see paper for full example)

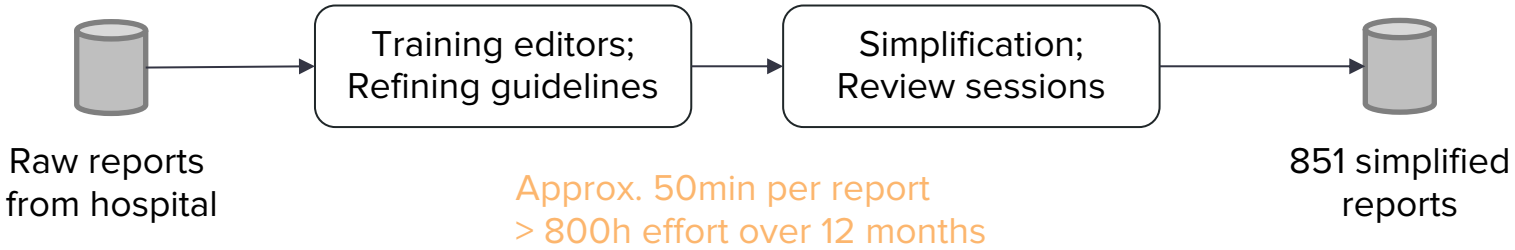
We use a lightweight protocol with experts

Inductive approach: how would a doctor intuitively explain a report to a patient?

Involved 9 medical students (in 4th year of training)

Simulate patient setting with a persona (Cooper, 1999)

Target reader of simplification; Controls variance among editors



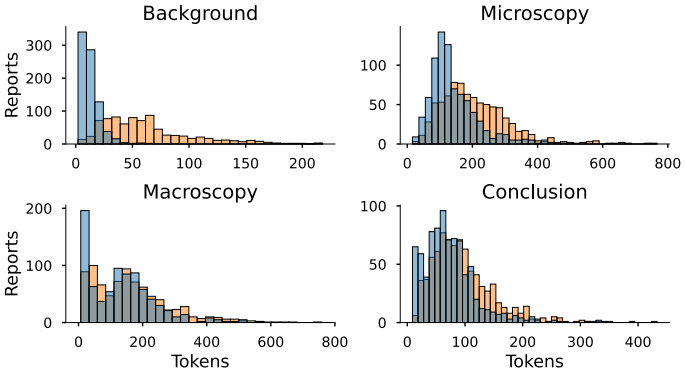
Dataset characteristics and summary statistics

Simplifications are longer
Substantial expansion in background section

Different and more constrained vocabulary
High novelty; lower type-token ratio

Slightly higher readability
Flesch reading ease

	Original	Simplified
Documents	851	851
Tokens	327,466	462,994
Types	10,292	11,229
Novelty	63%	
Avg. TTR	0.47	0.42
Avg. Reading Ease	32.84	40.23



How good are existing methods on this data?

Sequence-to-sequence methods

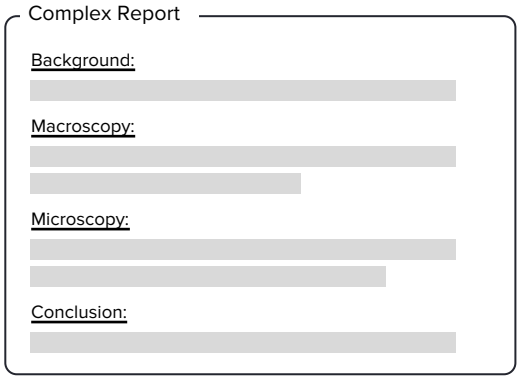
1. Bert2Bert
2. mBART
3. Identity baseline

Paragraph-level simplification (Devaraj et al.)

Each report section is simplified independently

Evaluation

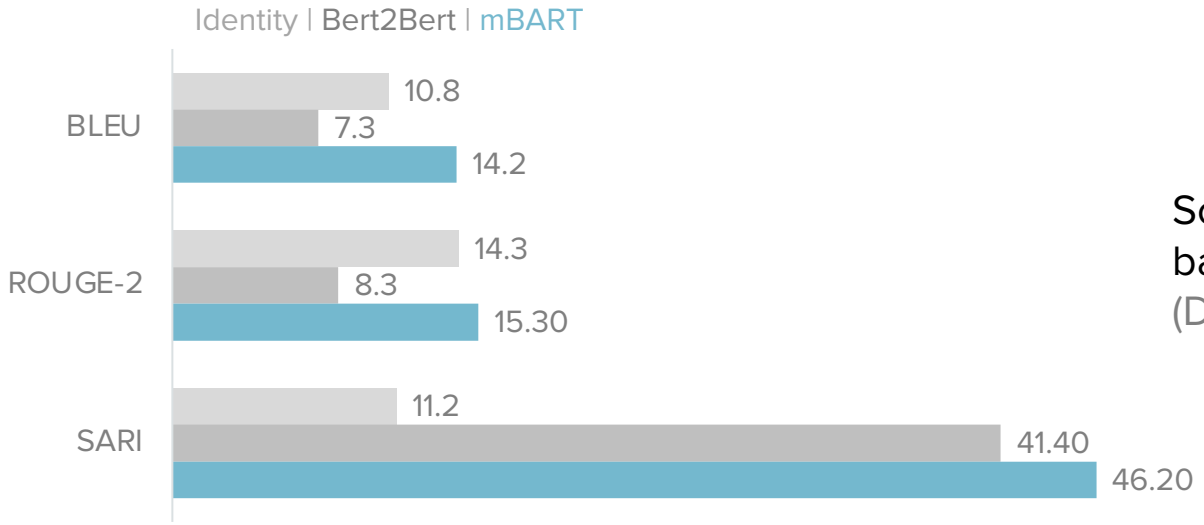
ROUGE (R-1/2/L), SARI, BLEU



Devaraj et al. (NAACL 2021). Paragraph-level Simplification of Medical Texts.

Models improve over identity baseline

Small gain in adequacy (BLEU, ROUGE), Large gain in simplicity (SARI)

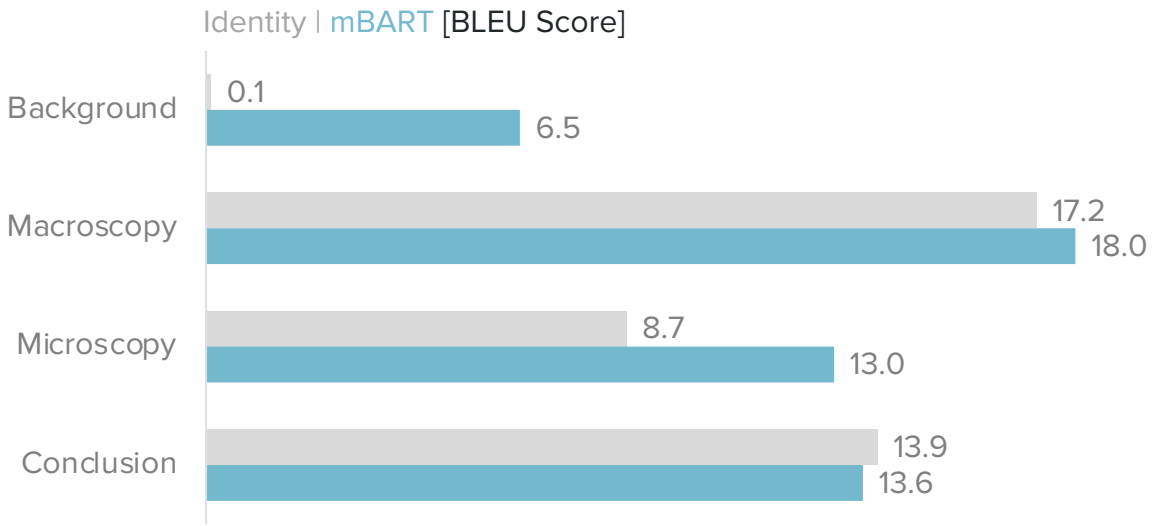


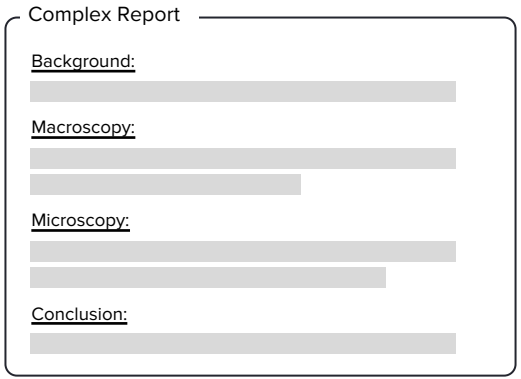
Scores are in a similar ballpark on English datasets (Devaraj et al., 2021)

Devaraj et al. (NAACL 2021). Paragraph-level Simplification of Medical Texts.

Background section is most difficult to simplify

Has explanation/expansion causes low lexical overlap





Reports are fluent but have factual errors

Example mBART output

Input

[...] The tumor shows strong positivity for CD34 and strong nuclear expression of STAT6. [...]

mBART output

[...] Of the tumor markers tested (CD34, STAT6, [...]), CD34 was positive and STAT6 was negative. This combination of tumor markers is suggestive of the presence of a gastrointestinal stromal tumor (GIST). [...]

Negation error

Conceivable, but in the context of this report wrong.

Difficult to capture with automated metrics.

Translated to English (see paper for full example)

Conclusion

Takeaways

- Dataset can help to advance document-level TS in clinical domain
- Challenges in clinical domain (vocabulary, explanations, content selection)
- Factual consistency is a problem

Future Work

- Expand the dataset
- Analyze simplification strategies carried out by experts
- Evaluation with patients/advocacy groups

Thanks!

github.com/jantrienes/simple-patho

jan.trienes@uni-due.de