Sarah Frank[1,2], Andreas Wagner[1], Christian Gütl[2]

COMBINING TEMPLATES AND LANGUAGE MODELS FOR THE AUTOMATIC CREATION OF SCIENTIFIC OVERVIEWS

[1] CERN, Switzerland
[2] Graz University of Technology, Austria

AISyS 2024 | 3 October 2024

SARAH FRANK

From 2012 to 2021: Graz University of Technology

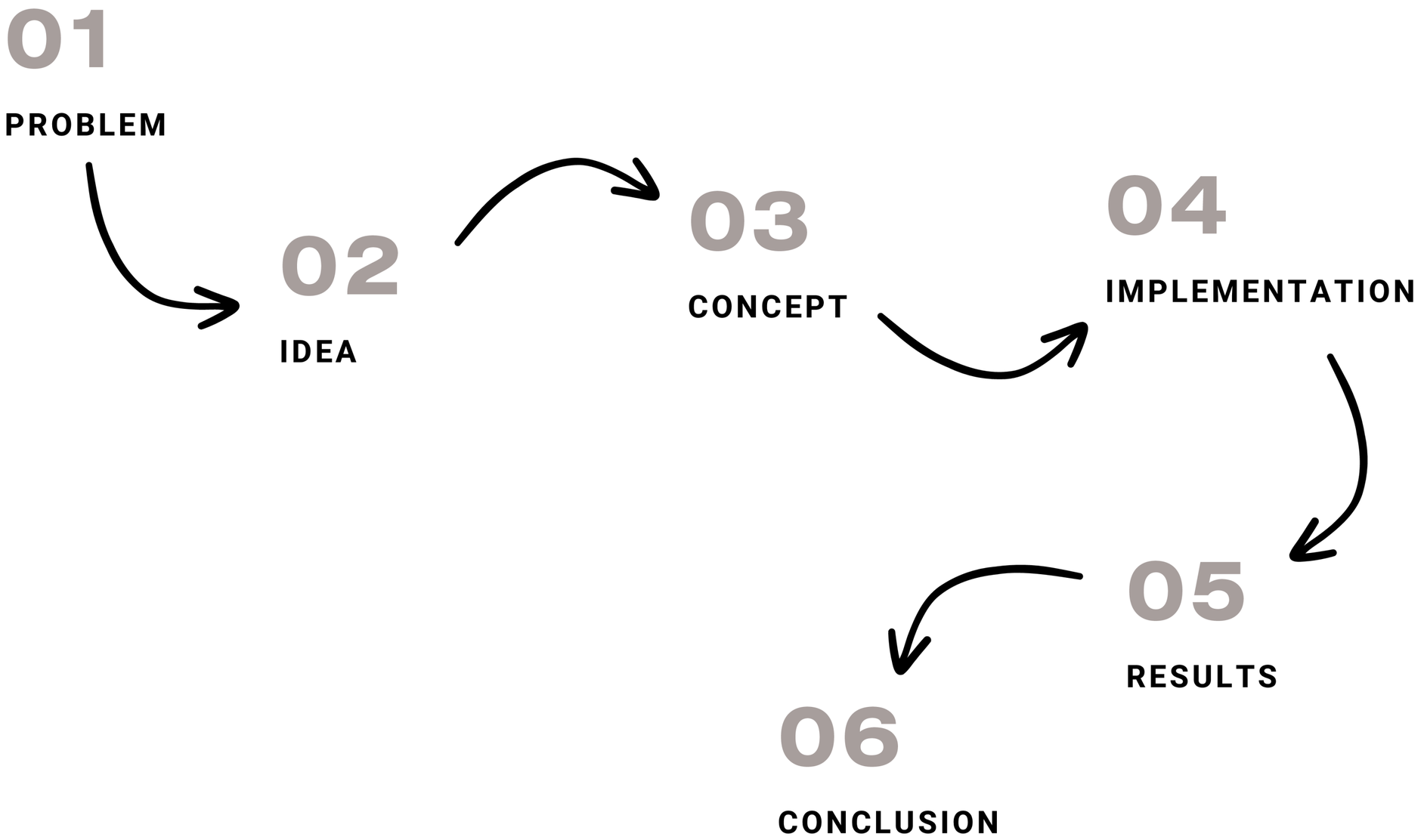
- BSc
- MSc (Robotics and Multimedia Information Systems)

Worked 2-and-something years at Campus 02, Graz

Since 2022: Doctoral student

- Graz University of Technology & CERN
- Automatic summarization of scientific articles

OVERVIEW


01
PROBLEM
02
IDEA
03
CONCEPT
04
IMPLEMENTATION
05
RESULTS
06
CONCLUSION

PROBLEM

The pace of publications is increasingly impossible to keep up with

Scientific environments require trustability

Knowing information sources is important

RQ1

How do existing LMs perform when evaluated for the creation of ultra-short summaries?

RQ2

How can templates tailor results for formulaic texts when used with transformer models?

RQ3

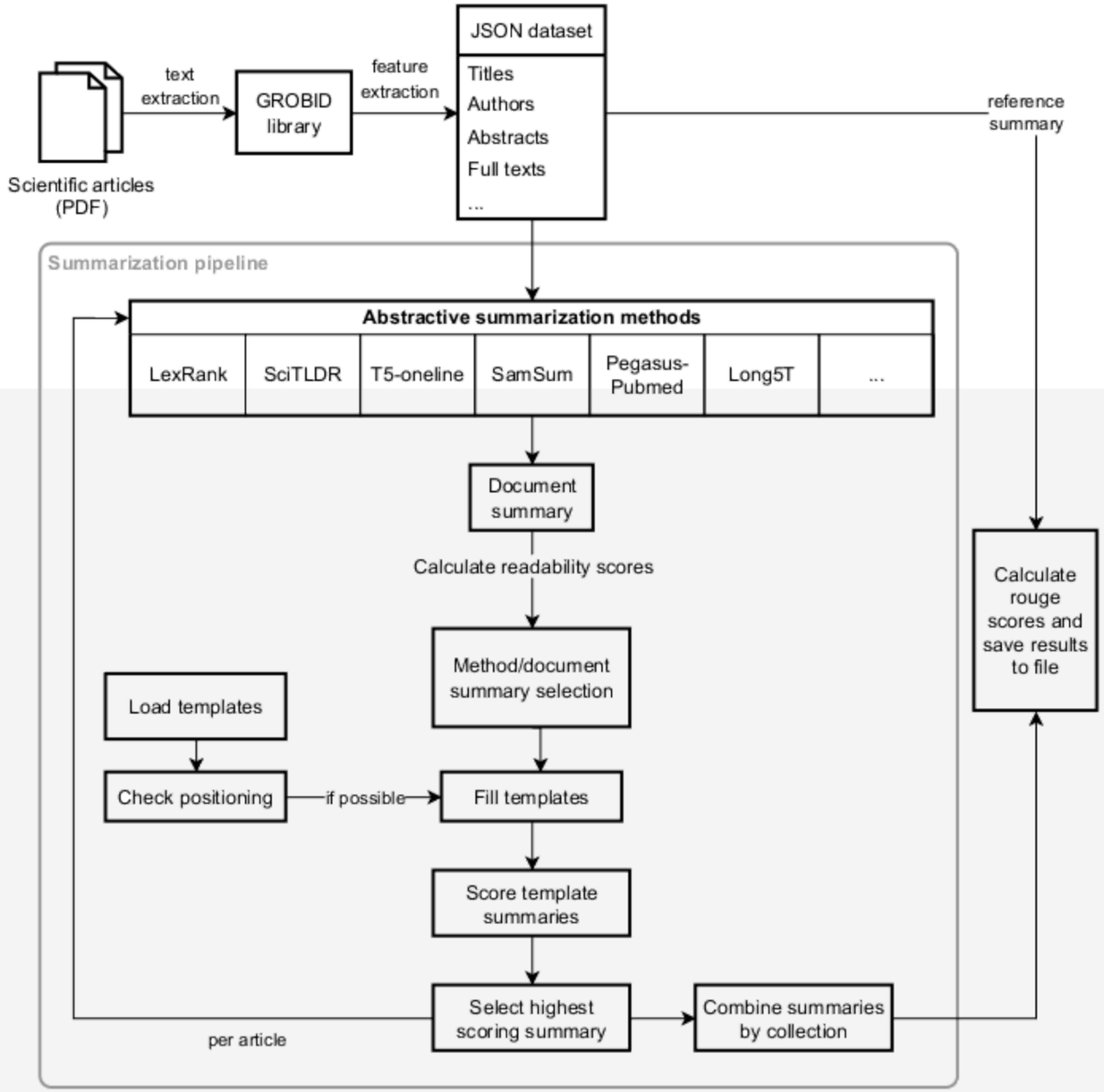
How do the resulting summaries perform when evaluated by automatic and manual means?

IDEA

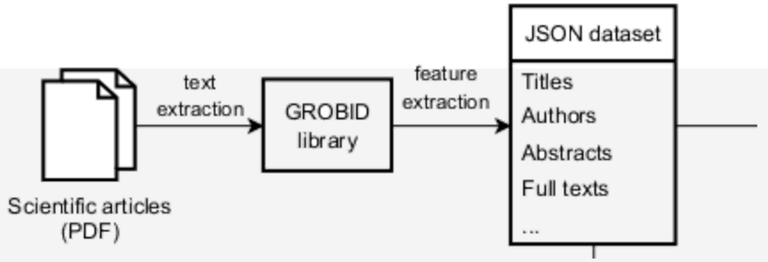
- Ease workload for those who summarize scientific papers
- Create structured summaries according to templates
- Create easily verifiable summaries

CONCEPT

- Ultra-short summaries
- Templates for formulaic texts
- Large language model for content summaries
- Language quality evaluation



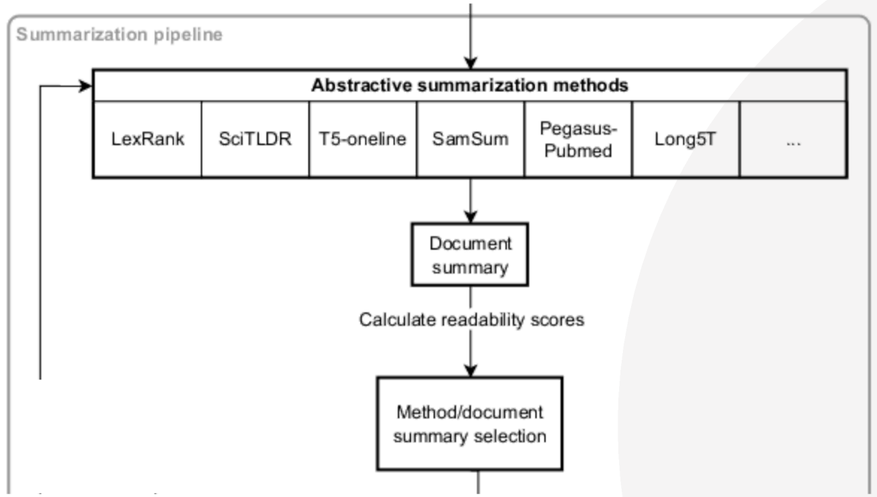
IMPLEMENTATION: DATASET



- 7 issues of “The Journal of Universal Computer science”
- Total of 39 papers

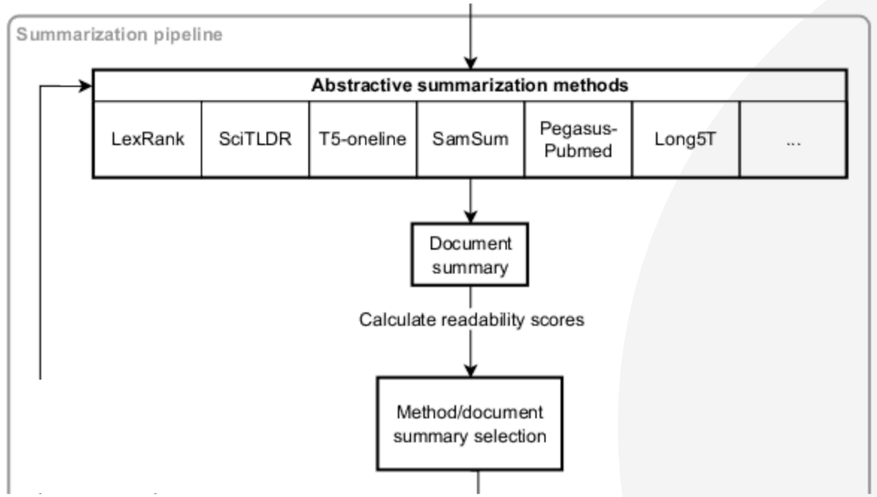
Issue	26/07	26/09	26/10	26/11	27/01	28/03	28/10
# articles	4	9	4	8	3	6	5

- Editorials as reference summaries



IMPLEMENTATION: MODEL SELECTION

- Trained on scientific articles
- Short summary of overall topic
- Full sentences preferred
- Abstractive, single-document

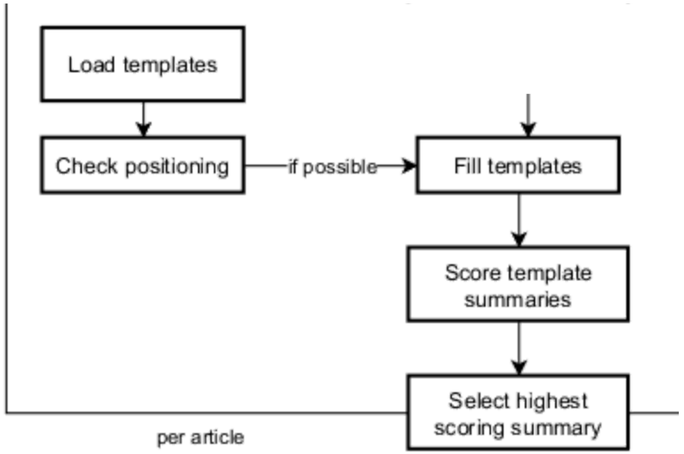


IMPLEMENTATION: MODEL SELECTION

- Trained on scientific articles
- Short summary of overall topic
- Full sentences preferred
- Abstractive, single-document

Method	Flesch	Readability	ROUGE-1	ROUGE-L
Reference	25.20	20.28	-	-
SciTLDR-F	14.98	25.25	0.6402	0.4707
SciTLDR-A	11.79	26.10	0.6994	0.5713
LexRank	10.47	29.37	0.5343	0.3756
...				
T5-oneline	14.78	24.68	0.6794	0.5244

IMPLEMENTATION: TEMPLATES



- Creation by hand
- Placement arrays - [0,1,1]
- Scored with text_standard

‘Finally, in “[TITLE]”, [AUTHORS] [SUMMARY].’

RESULTS

Issue	ROGUE-1	ROGUE-L
26/07	0.91	0.73
26/09	0.68	0.53
26/10	0.60	0.53
26/11	0.77	0.64
27/01	0.64	0.49
28/03	0.63	0.42
28/10	0.70	0.56

RESULTS

Issue	ROGUE-1	ROGUE-L
26/07	0.91	0.73
26/09	0.68	0.53
26/10	0.60	0.53
26/11	0.77	0.64
27/01	0.64	0.49
28/03	0.63	0.42
28/10	0.70	0.56

- Strong variance
- Consistency lacking
- No correlation with number of articles

Issue	26/07	26/09	26/10	26/11	27/01	28/03	28/10
# articles	4	9	4	8	3	6	5

RESULTS

Survey #1

- 11 participants
- 8 completed
- 15 questions

- 11x automated preferred
- 3 equal split
- 1x manual preferred by all

Survey #2

- 14 participants
- 8 completed
- 10 questions

Performance	Accuracy	Coverage	Fluency	Informativeness
Excellent	18	14	23	17
Good	48	35	28	32
Fair	11	13	18	19
Poor	2	16	10	10
Very Poor	1	2	1	2
Average	4	3.54	3.78	3.65
Std. Dev.	0.76	1.07	1.04	1.03

RESULTS - SURVEY #1

“The repetition of full names is entirely irrelevant. It makes the sentences VERY hard to read[...]”

“Nice! Though it is a run-on sentence. May need a period there to separate it [...]”

“#2 gives more information but without any context it’s hard to understand, #1 is more general”

“The second summary is more detailed and fits better to the abstract”

Both summaries are of high quality, but #1 just seems to offer a more rounded and comprehensive snapshot of the abstract [...]

CONCLUSION

APPROACH AND RESULTS

- Combination of LLMs and templates
- Reliability insufficient
- Significant standard deviation
- Grammatical errors
- Consistent quality
- Templates too rigid

FUTURE WORK

- Dynamic creation of templates
- Alternative (newer) LLMs
- Fine-tune own model
- Increase summary lengths

THANK YOU