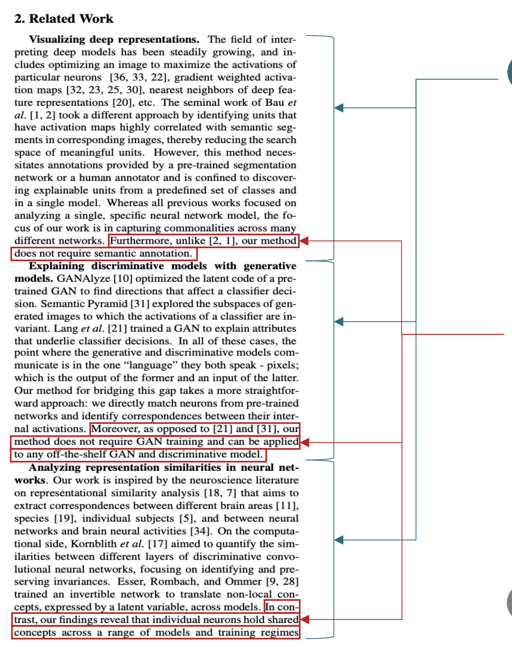
Toward Related Work Generation with Structure and Novelty Statement

Code is here!

Scholarly Document Processing @ACL2024

Kazuya Nishimura[1], Kuniaki Saito[2], Tosho Hirasawa[2], Yoshitaka Ushiku[2]
National Cancer Center Japan, OMRON SINIC X Corp.

What is well-written related work?



Point 1. Structure

Make paragraphs with topics determined from the relations between current work and previous works

Point 2. Novelty statement

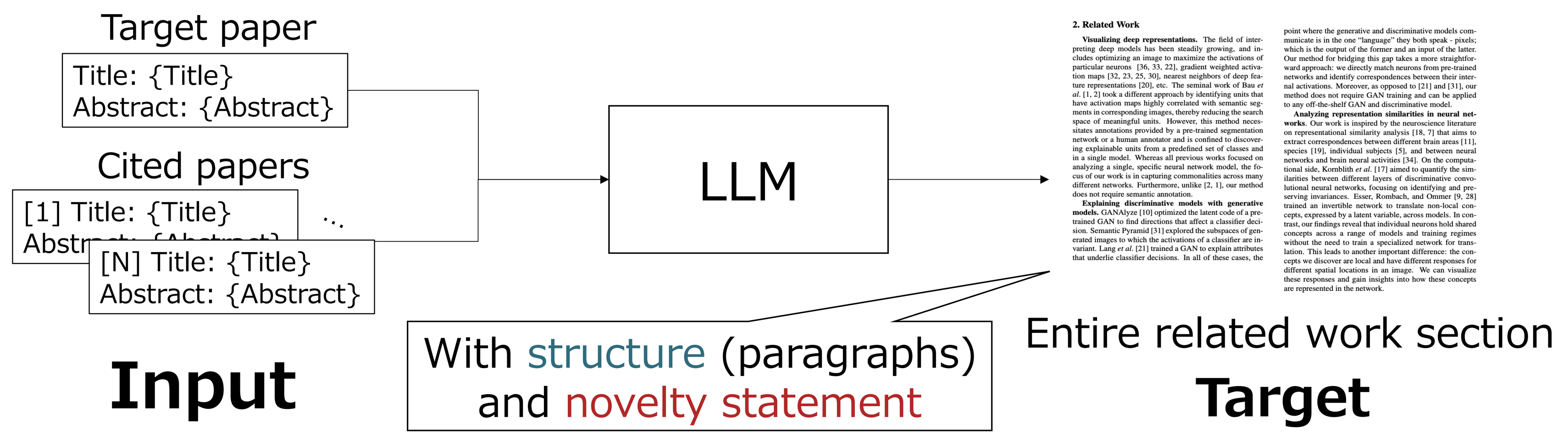
Highlights unique contributions of current work and clearly state difference with previous work

Point 3. Summarization

While structure and novelty statement are important for related work, previous work has treated automated related work generation as **a just summarization**

STructured Related work Generation with Novelty Statemen (STRoGeNS)

Redefine related work generation task as **STRoGeNS**



Problems. Lack of datasets and metrics

Problem 1. no datasets for STRoGeNS

Dataset	Target	Pairs	#Para.
Multi-Xscience [Lu+, 2020]	Only 1 paragraph in related work section	40,528	1
S2ORC [Chen+, 2021]		136,655	1
Delve [Chen+, 2021]		78,927	1
TAS2 [Chen+, 2022]		117,700	1
TAD [Chen+, 2022]		218,255	1
BigSurvey-MDS [Liu+, 2022]		4,478	1
SciReviewGen [Kasanishi+, 2022]		10,130	1

Problem 2. no evaluation metrics for STRoGeNS

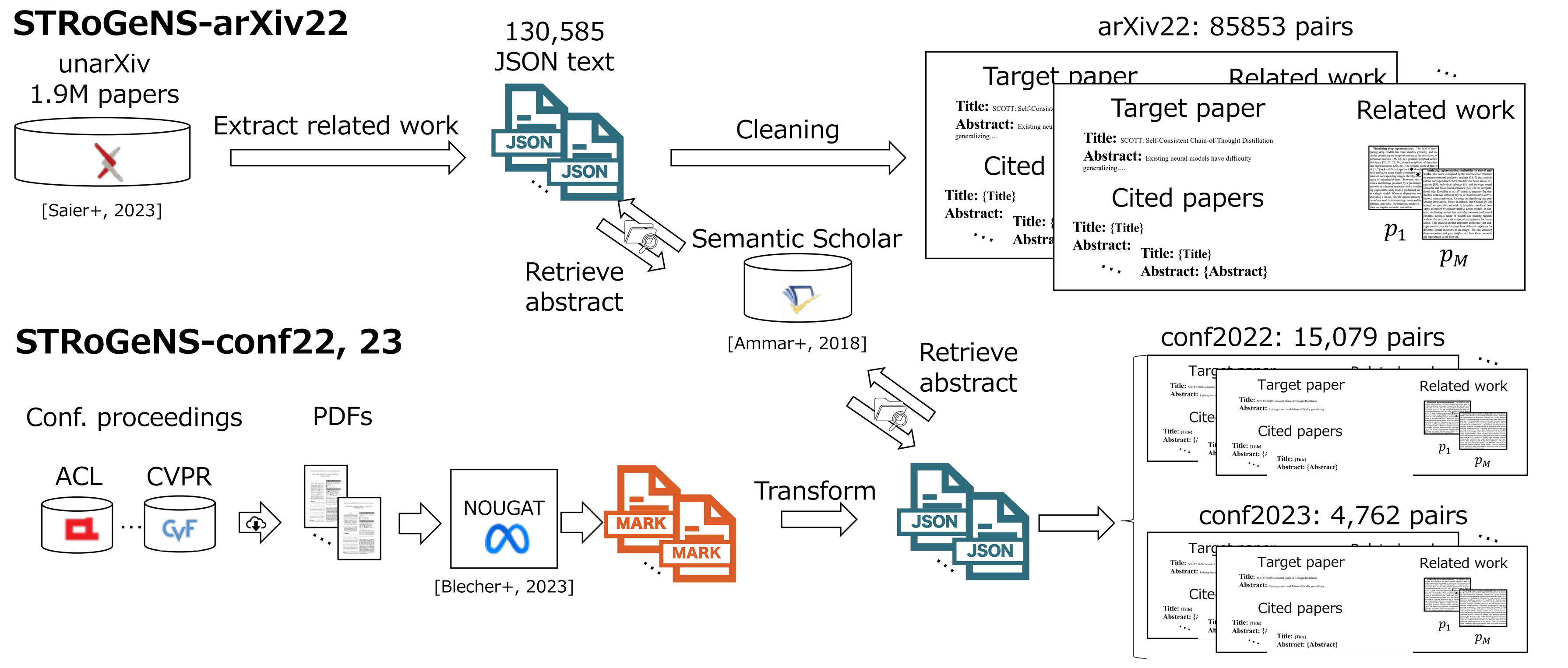
Solutions. Propose datasets and metrics

Solution 1. Propose large-scale datasets for that took into consideration both quality and quantity.

Dataset	Target	Pairs	#Para.
STRoGeNS-arXiv22 (quantity)	Structured related work with novelty statement	85,853	4.22
STRoGeNS-conf22 (quality)		15,079	4.27
STRoGeNS-conf23 (Test)		4,762	4.04

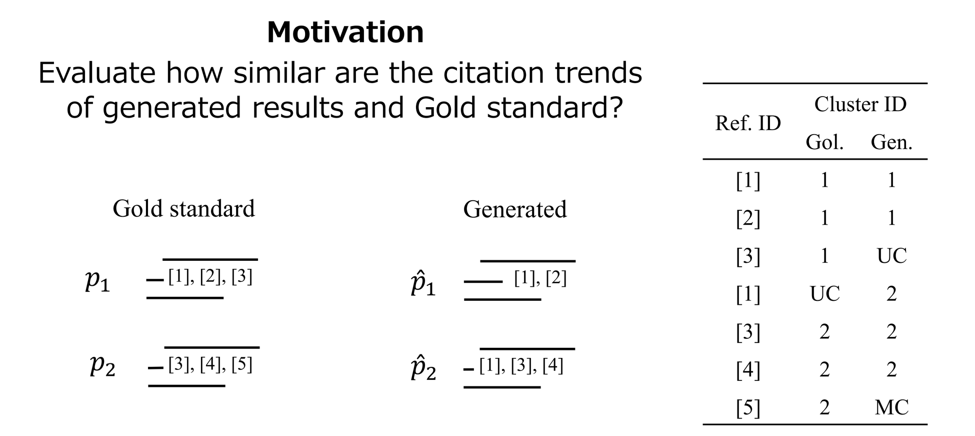
Solution 2. Propose automated evaluation metrics for structure and novelty.

Dataset generation



Evaluation metrics

Structure evaluation based on citation trends



Metrics

F1 measure citation similarity among most similar

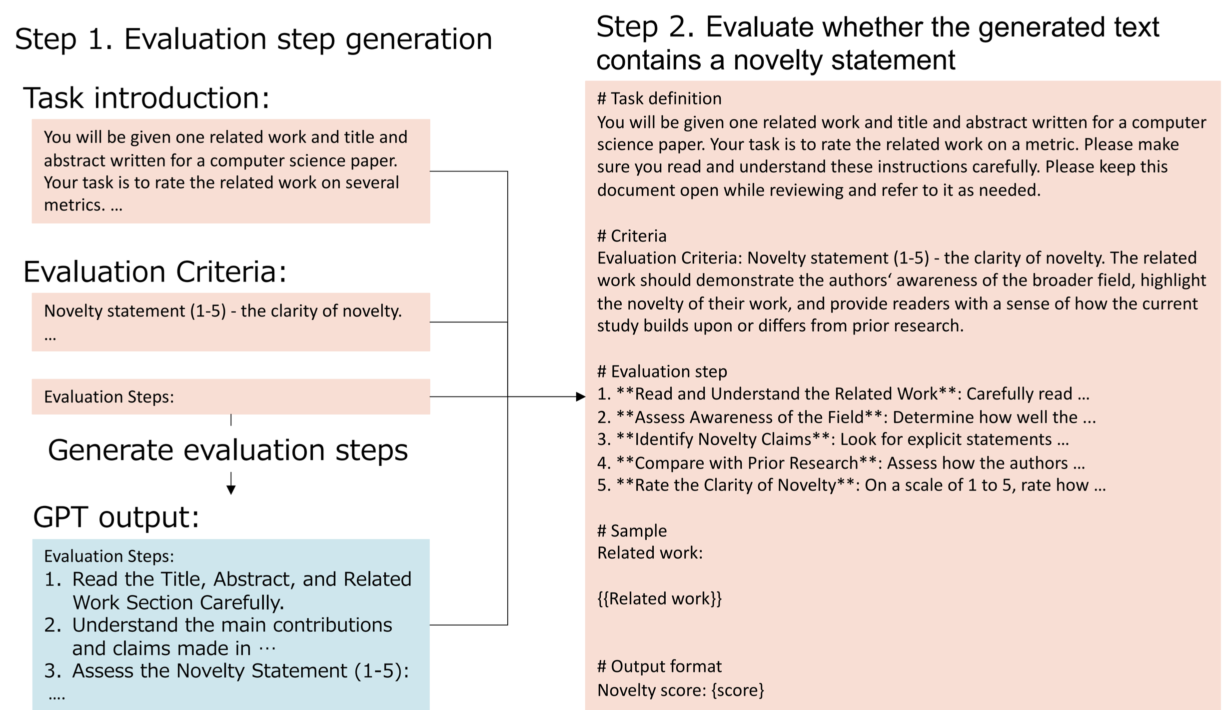
$$\mathrm{F1} = \frac{1}{m}\sum_{i=1}^{m} \max_{j} \mathrm{F1}(f(p_i), f())$$

ARI evaluate citation trends like clustering evaluation

$$\mathrm{ARI}' = \frac{\#\mathrm{TPair}}{\#\mathrm{Pair} - \#\mathrm{UC}}$$

$$\mathrm{ARI}' = \frac{\mathrm{RI}' - \mathrm{ExpectedRI}}{\max(\mathrm{RI}) - \mathrm{ExpectedRI}}$$

Novelty statement evaluation with G-Eval [Liu et al. 2023]



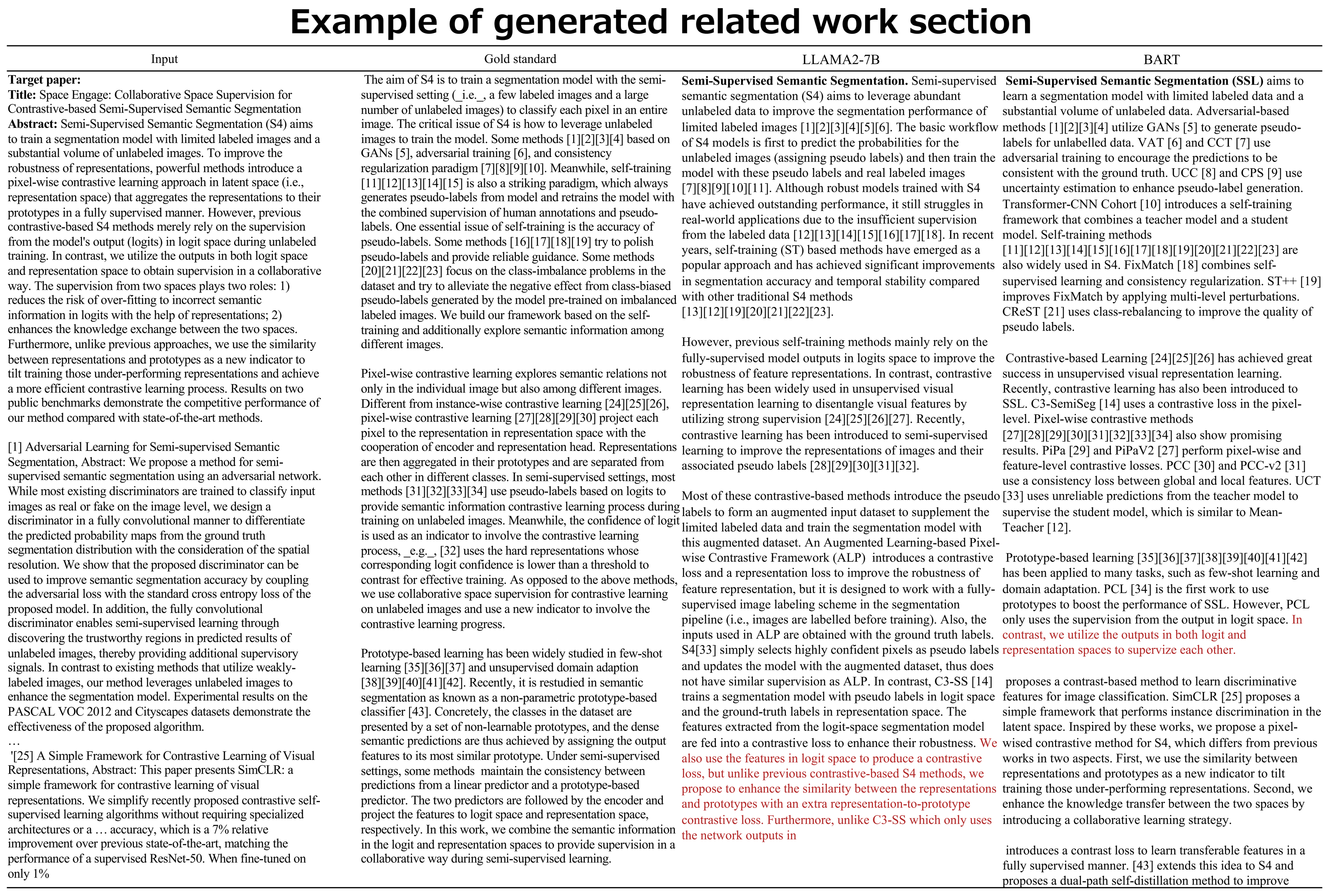
Accuracy of novelty score is 90%.

Experiments

Structure and summarization comparisons

Method	Structure F1↑	Structure ARI'↑	Summarization R1↑	Summarization R2↑	Summarization R-L↑
TextRank [Mihalcea+, 2004]	-	-	31.6	7.5	24.7
LexRank [Erkan+, 2021]	-	-	31.5	8.19	21.6
GPT3.5 [Brown+, 2020]	69.0	33.9	38.2	7.7	28.5
GPT4 [Brown+, 2020]	66.7	29.8	41.4	7.8	29.8
BART [Lewis+, 2019]	69.1	**54.0**	**47.4**	**13.1**	**32.7**
PEGASUS [Zhang+, 2020]	48.8	11.0	25.4	6.2	19.1
LED [Beltagy+, 2020]	59.5	33.7	44.0	11.8	29.2
Llama2-7B [Touvron+, 2023]	**69.8**	47.3	42.7	9.0	29.4

Example of generated related work section



Quality vs quantity

Method	Dataset	R1↑	F1↑
BART [Lewis+, 2019]	conf22	42.5	62.5
	arXiv22	**46.1**	66.4
PEGASUS [Zhang+, 2020]	conf22	18.1	31.2
	arXiv22	17.7	40.2
LED [Beltagy+, 2020]	conf22	41.5	56.1
	arXiv22	40.9	12.3
Llama2-7B [Touvron+, 2023]	conf22	42.6	**70.0**
	arXiv22	37.6	9.0

Novelty statement evaluation

Method	Novelty score
GPT3.5	0.17
BART	**0.91**
LED	0.53
Llamma2	0.56

Human evaluation

Metrics	r	ρ	τ
F1	47.2	56.4	31.9
ARI'	38.3	35.7	24.9