Toward Related Work Generation with Structure and Novelty Statement
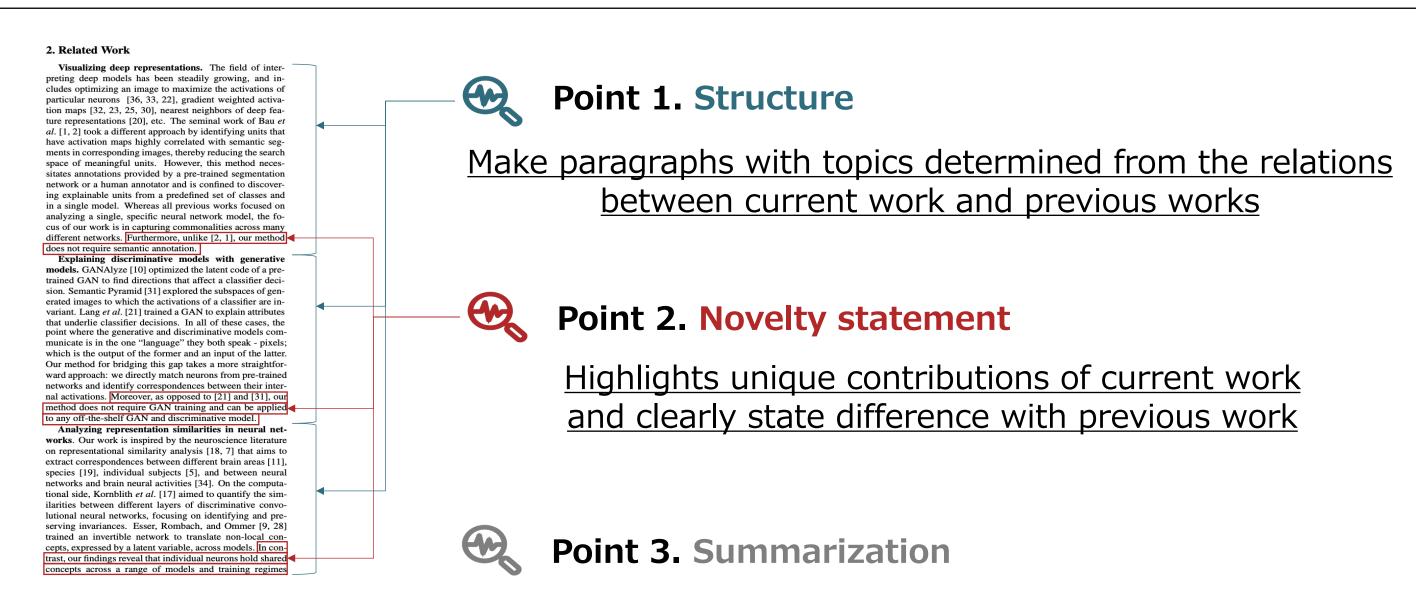
Kazuya Nishimura[1], Kuniaki Saito[2] , Tosho Hirasawa[2] , Yoshitaka Ushiku[2]
National Cancer Center Japan, OMRON SINIC X Corp.

Scholarly
Document
Processing
@ACL2024

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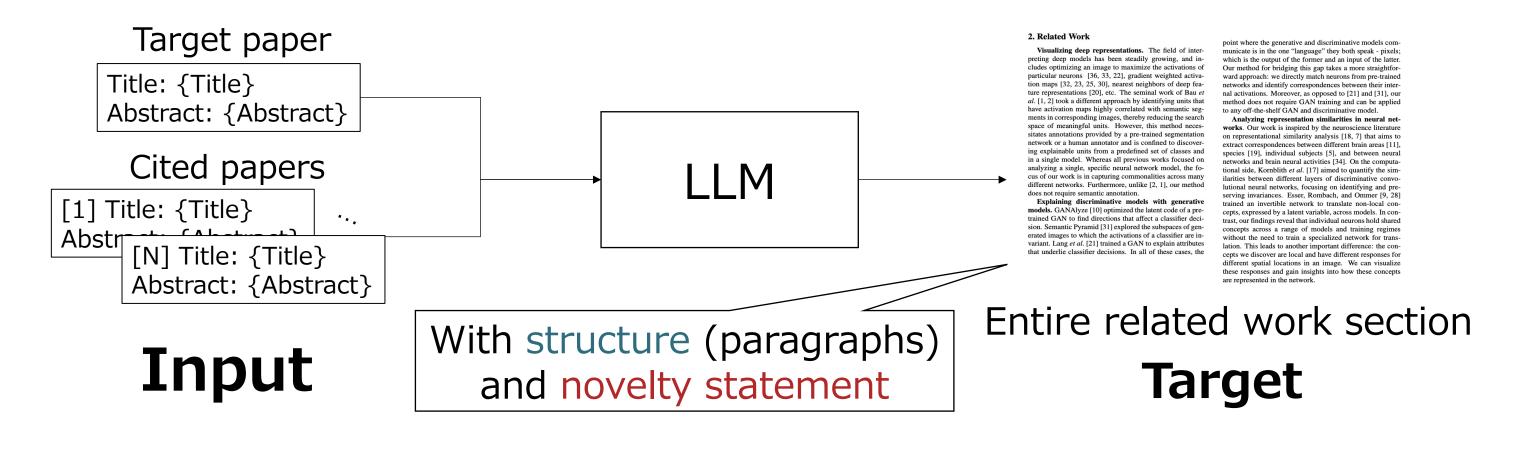
While structure and novelty statement are important for related work, previous work has treated automated related work generation as **a just summarization**

STructured Related work Generation with Novelty Statemen (STRoGeNS)

Redefine related work generation task as **STRoGeNS**

Input

Target paper
Title: {Title}
Abstract: {Abstract}

Cited papers
[1] Title: {Title}
Abstract: {Abstract}
...
[N] Title: {Title}
Abstract: {Abstract}

LLM

With structure (paragraphs) and novelty statement

Target

Entire related work section

Problems. Lack of datasets and metrics

Problem 1. no datasets for STRoGeNS

Dataset	Target	Pairs	#Para.
Multi-Xscience [Lu+, 2020]		40,528	1
S2ORC [Chen+, 2021]		136,655	1
Delve [Chen+, 2021]		78,927	1
TAS2 [Chen+, 2022]	**Only 1 paragraph in related work section**	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)		85,853	4.22
STRoGeNS-conf22 (quality)	**Structured related work with novelty statement**	15,079	4.27
STRoGeNS-conf23 (Test)		4,762	4.04

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

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
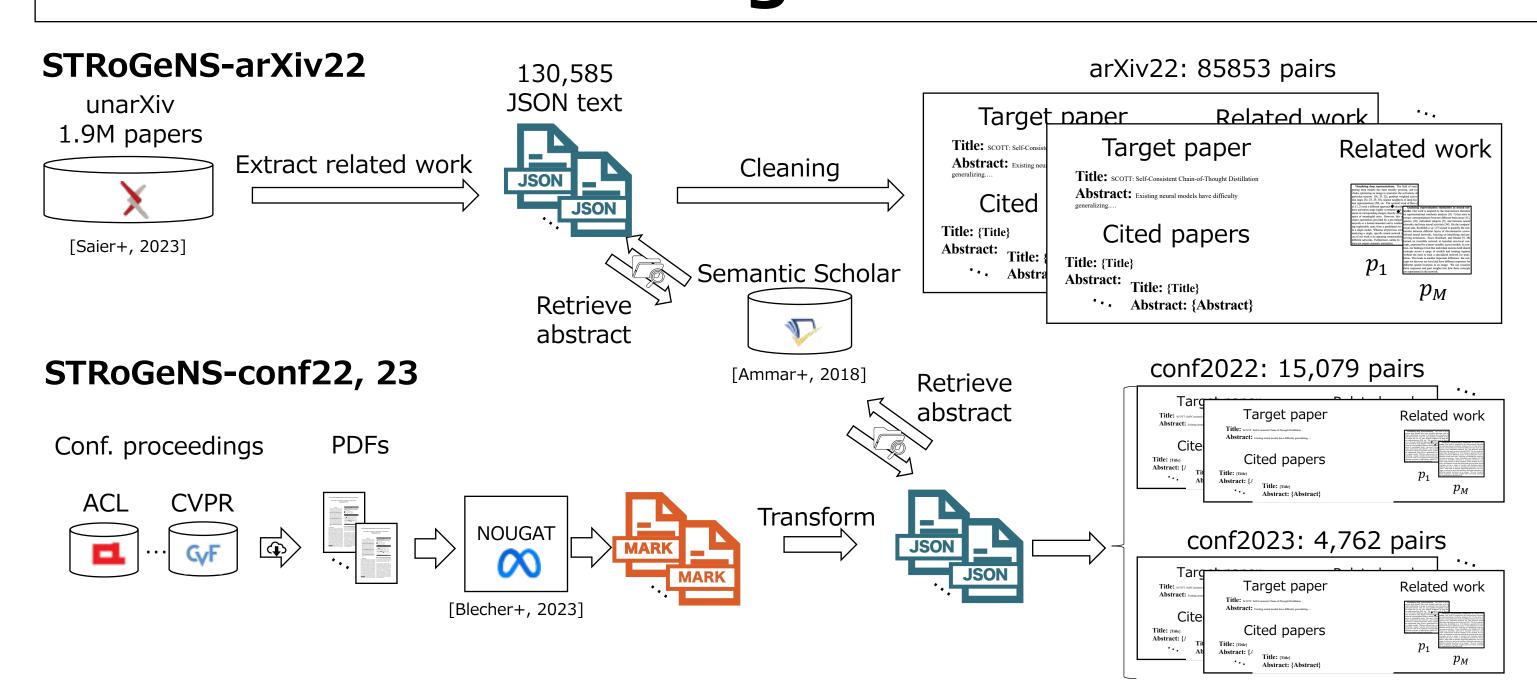
Novelty statement evaluation

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

Dataset generation

STRoGeNS-arXiv22

unarXiv
1.9M papers
[Saier+, 2023]

→ Extract related work →

130,585 JSON text

→ Cleaning →

Retrieve abstract
Semantic Scholar
[Ammar+, 2018]

arXiv22: 85853 pairs

Target paper / Related work
Title / Abstract
Cited papers
$p_1 \quad p_M$

STRoGeNS-conf22, 23

Conf. proceedings
ACL ... CVPR

→ PDFs →

NOUGAT
[Blecher+, 2023]

→ MARK →

Transform →

Retrieve abstract
JSON

conf2022: 15,079 pairs
conf2023: 4,762 pairs

Evaluation metrics

Structure evaluation based on citation trends

Motivation
Evaluate how similar are the citation trends of generated results and Gold standard?

Ref. ID	Cluster ID Gol.	Gen.
[1]	1	1
[2]	1	1
[3]	1	UC
[1]	UC	2
[3]	2	2
[4]	2	2
[5]	2	MC

Gold standard
$p_1 \quad — [1], [2], [3]$
$p_2 \quad — [3], [4], [5]$

Generated
$\hat{p}_1 \quad — [1], [2]$
$\hat{p}_2 \quad — [1], [3], [4]$

Metrics

F1 measure citation similarity among most similar

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

ARI evaluate citation trends like clustering evaluation

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

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

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

Step 1. Evaluation step generation

Task introduction:
You will be given one related work and title and abstract written for a computer science paper. Your task is to rate the related work on several metrics. ...

Evaluation Criteria:
Novelty statement (1-5) - the clarity of novelty. ...

Evaluation Steps:

Generate evaluation steps

GPT output:
Evaluation Steps:
1. Read the Title, Abstract, and Related Work Carefully.
2. Understand the main contributions and claims made in ...
3. Assess the Novelty Statement (1-5): ....

Step 2. Evaluate whether the generated text contains a novelty statement

Task definition
You will be given one related work and title and abstract written for a computer science paper. Your task is to rate the related work on a metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing and refer to it as needed.

Criteria
Evaluation Criteria: Novelty statement (1-5) - the clarity of novelty. The related work should demonstrate the authors' awareness of the broader field, highlight the novelty of their work, and provide readers with a sense of how the current study builds upon or differs from prior research.

Evaluation step
1. **Read and Understand the Related Work**: Carefully read ...
2. **Assess Awareness of the Field**: Determine how well the ...
3. **Identify Novelty Claims**: Look for explicit statements ...
4. **Compare with Prior Research**: Assess how the authors ...
5. **Rate the Clarity of Novelty**: On a scale of 1 to 5, rate how ...

Sample
Related work:
{{Related work}}

Output format
Novelty score: {score}

Accuracy of novelty score is 90%.

Experiments

Structure and summarization comparisons

Method	Structure		Summarization		
	F1 ↑	ARI' ↑	R1 ↑	R2 ↑	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

Input	Gold standard	LLAMA2-7B	BART
Target paper: **Title:** Space Engage: Collaborative Space Supervision for Contrastive-based Semi-Supervised Segmentation. **Abstract:** Semi-Supervised Semantic Segmentation (S4) aims to train a segmentation model with limited labeled images and a substantial volume of unlabeled images. To improve the robustness of representations, powerful methods introduce a pixel-wise contrastive learning approach in latent space (i.e., representation space) that aggregates the representations to their prototypes in a fully supervised manner. However, previous contrastive-based S4 methods merely rely on the supervision from the model's output (logits) in logit space during unlabeled training. In contrast, we utilize the outputs in both logit space and representation space to obtain supervision in a collaborative way. The supervision from two spaces plays two roles: 1) reduces the risk of over-fitting to incorrect semantic information in logits with the help of representations; 2) enhances the knowledge exchange between the two spaces. Furthermore, unlike previous approaches, we use the similarity between representations and prototypes as a new indicator to tilt training those under-performing representations and achieve a more efficient contrastive learning process. ...	The aim of S4 is to train a segmentation model with the semi-supervised setting (_i.e._, a few labeled images and a large number of unlabeled images) to classify each pixel in an entire image. The critical issue of S4 is how to leverage unlabeled images to train the model. Some methods [1][2][3][4] based on GANs [5], adversarial training [6], and consistency regularization paradigm [7][8][9][10]. Meanwhile, self-training [11][12][13][14][15] is also a striking paradigm, which always generates pseudo-labels from model and retrains the model with the combined supervision of human annotations and pseudo-labels. ...	**Semi-Supervised Semantic Segmentation.** Semi-supervised semantic segmentation (S4) aims to leverage abundant unlabeled data to improve the segmentation performance of limited labeled images [1][2][3][4][5][6]. The basic workflow of S4 models is first to predict the probabilities for the unlabeled images (assigning pseudo labels) and then train the model with these pseudo labels and real labeled images [7][8][9][10][11]. ...	**Semi-Supervised Semantic Segmentation (SSL)** aims to learn a segmentation model with limited labeled data and a substantial volume of unlabeled images. Adversarial-based methods [1][2][3][4] utilize GANs [5] to generate pseudo-labels for unlabeled data. VAT [6] and CCT [7] use uncertainty estimation to enhance pseudo-label generation. ...