Zero-Shot Aspect-Based Scientific Document Summarization using Self-Supervised Pre-training

Amir Soleimani, Vassilina Nikoulina, Salah Ait Mokhtar, Benoit Favre

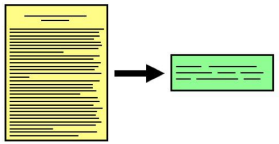
a.soleimani@uva.nl

NAVER LABS Europe

University of Amsterdam

Introduction

- Zero-Shot Aspect-Based Scientific Document Summarization
  - Document Summarization
    - Input: News Articles, Conversations, Scientific Papers
    - Output: Summary
    - Datasets: XSum, CNN, … (400-800 words)



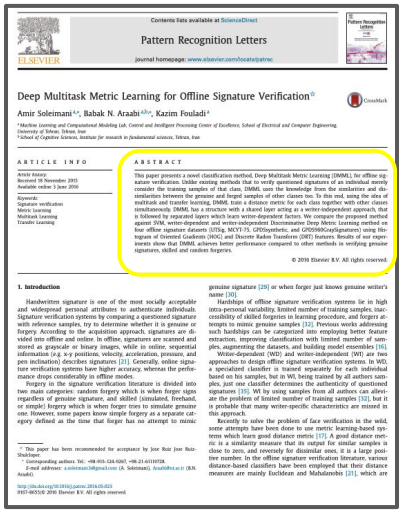
XSum dataset:
BBC News Articles

SUMMARY: *A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.*

DOCUMENT: Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.
The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.
[*6 sentences with 139 words are abbreviated from here.*]
Other reports said the victims had been sunbathing when the plane made its emergency landing.
[*Another 4 sentences with 67 words are abbreviated from here.*]
Video footage from the scene carried by local broadcasters showed a small recreational plane parked on the sand, apparently intact and surrounded by beachgoers and emergency workers.
[*Last 2 sentences with 19 words are abbreviated.*]

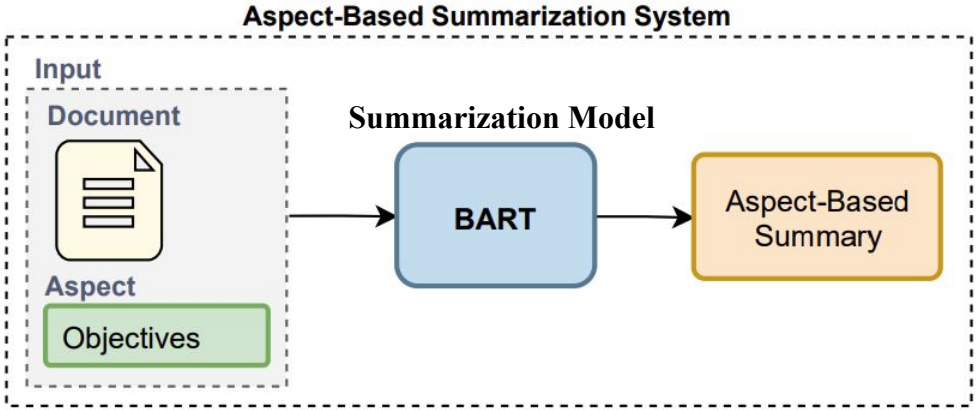
Introduction

- Zero-Shot Aspect-Based Scientific Document Summarization
  - Scientific Document Summarization
    - Papers: longer than news articles (3,000-7,000 words)
    - Datasets: PubMed, arXiv, Emerald
    - Input: Texts from all sections except abstract
    - Output: Abstract



Introduction

- Zero-Shot Aspect-Based Scientific Document Summarization
  - Aspect-Based Scientific Document Summarization
    - To summarize with respect to a query (aspect)
      - News → What's the background?
      - Conversation → Customer question
      - Paper → Objectives, Contribution, …

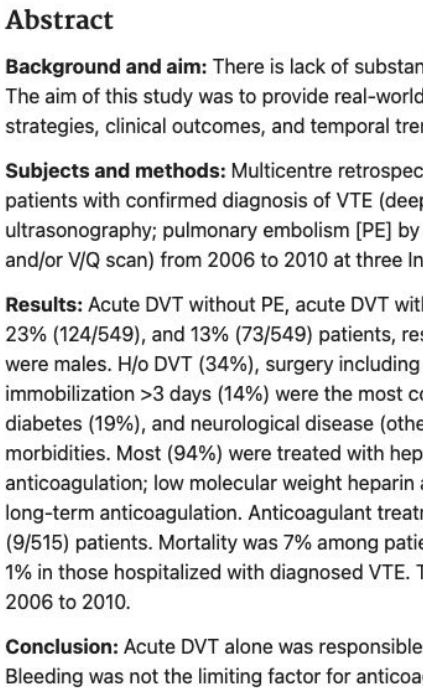


Introduction

- Zero-Shot Aspect-Based Scientific Document Summarization
  - Aspect-Based Scientific Document Summarization
    - Application: Document Assistance Systems
      - Covid-related papers
      - Bibliovid.org

Dataset Challenge!

- Zero-Shot Aspect-Based Scientific Document Summarization
  - Aspect-Based Scientific Document Summarization
    - **Collecting a big dataset is a significant challenge!**
    - Generic Summarization: Numerous sources! No annotation is needed!
    - Aspect-Based Summarization:
      - **Structured Abstracts** (PubMed, Emerald)
        - Introduction, Objectives, Methods, Results, Conclusion
        - No annotation is needed!
      - **Not available in all domains!** (e.g., computer science)
      - **Aspects are extremely limited!** (e.g., strength of evidence)
- Zero-Shot Aspect-Based Scientific Document Summarization
  - **Domain Shift**
  - **Unseen Aspects**



Abstract

Background and aim: There is lack of substan
The aim of this study was to provide real-world
strategies, clinical outcomes, and temporal tre

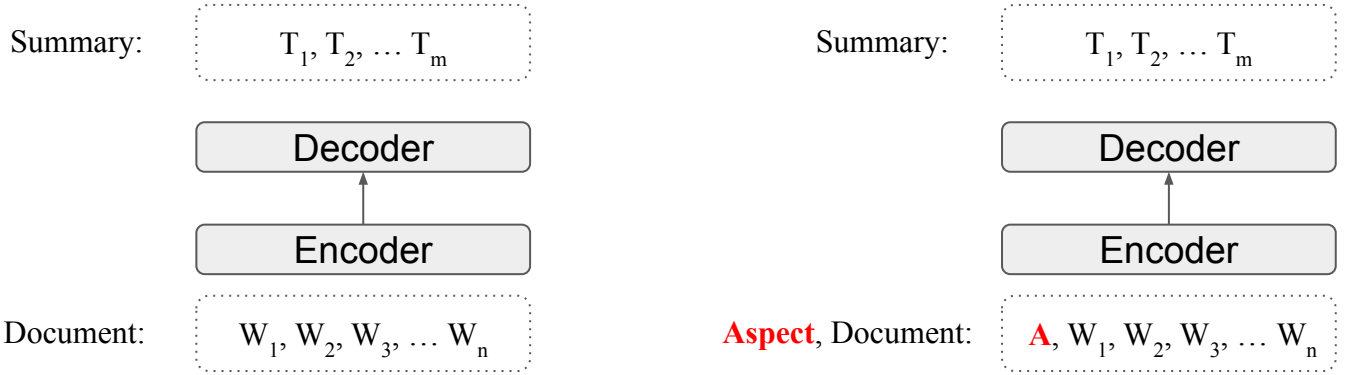
Subjects and methods: Multicentre retrospec
patients with confirmed diagnosis of VTE (deep
ultrasonography; pulmonary embolism [PE] by
and/or V/Q scan) from 2006 to 2010 at three In

Results: Acute DVT without PE, acute DVT with
23% (124/549), and 13% (73/549) patients, res
were males. H/o DVT (34%), surgery including
immobilization >3 days (14%) were the most co
diabetes (19%), and neurological disease (othe
morbidities. Most (94%) were treated with hep
anticoagulation; low molecular weight heparin a
long-term anticoagulation. Anticoagulant treatr
(9/515) patients. Mortality was 7% among patie
1% in those hospitalized with diagnosed VTE. T
2006 to 2010.

Conclusion: Acute DVT alone was responsible
Bleeding was not the limiting factor for anticoa

Summarization Model

- **BART** ("BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension")
  - BART is a pre-trained transformer-based model
  - BERT has only an encoder but BART has both encoder and decoder
  - Input format: X = <s> {A_1, ..., A_K} < /s> {W_1, ..., W_N }



Dataset

- Pubmed
  - Biomedical
- FacetSum (Emerald)
  - Management, Business, Education

PubMed	**# Samples (Aspect, Document)**				
	Train: 139.4K / Validation: 7.9K / Test: 8.1K				
	Average Length (# Words)				
	Documents: 3.5K				
	Summaries:				
	Intro.	Objectives	Methods	Results	Conc.
	53	38	76	94	40
FacetSum	**# Samples (Aspect, Document)**				
	Train: 182.4K/ Validation: 23.7K / Test: 23.7K				
	Average Length (# Words)				
	Documents: 6.6K				
	Summaries:				
		Objectives	Methods	Results	Value
		53	49	66	46

Baselines

Generic:

Aspect-Based:

Generic:

Aspect-Based:

	Model	R-1	R-2	R-L
PubMed Generic	Discourse (Cohan et al., 2018)	38.93	15.37	35.21
	PEGASUS (Zhang et al., 2020)	39.98	15.15	25.23
	BART	45.04	18.45	40.62
PubMed	Greedy Extractive (Oracle)	56.61	39.23	47.58
	BART	39.03	18.47	34.10
	BART-Independent†	38.91	18.21	33.89
	BART Shuffle Aspects	24.21	6.18	19.86
FacetSum Generic	BART (Meng et al., 2021)	45.49	18.10	42.74
	BART-Facet (Meng et al., 2021)	49.29	19.60	45.76
	BART	49.98	19.89	46.68
FacetSum	Greedy Extractive (Oracle)	51.87	32.09	41.55
	BART (Meng et al., 2021)	23.27	10.31	20.29
	BART-Facet (Meng et al., 2021)	37.97	15.17	32.08
	BART	36.97	15.50	31.48
	BART-Independent†	36.77	15.26	31.23
	BART Shuffle Aspects	28.18	6.94	22.71

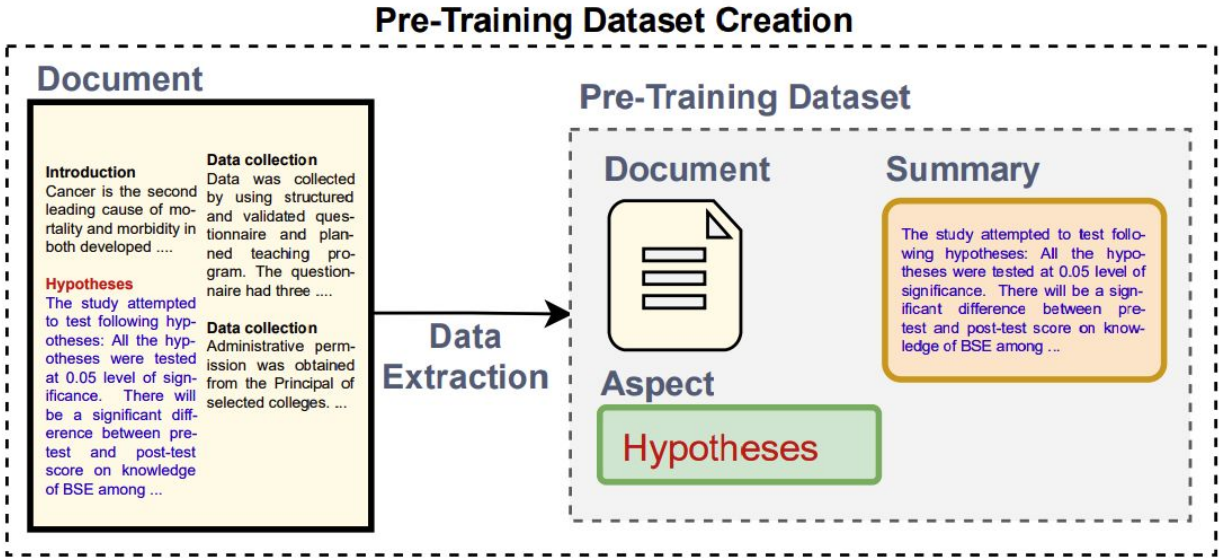
- Evaluation Metric: ROUGE score (calculating Ngram overlapping)
- Greedy Extractive is an oracle (top sentences: highest N-gram overlapping with reference sentences)
- BART-Independent: Trained on each aspect independently
- BART-Shuffle: Train on aspect **A** and evaluate on aspect **B** (same article)
  - Same input article but different aspect
  - To what extent aspects demand difference summaries

How to evaluate Zero-Shot performance!

- Train on dataset A and Evaluate on dataset B
  - PubMed → FacetSum
  - Domain Shift (Biomedical → Business)
  - Unseen Aspects
    - Very Limited: Aspects are almost the same in PubMed and FacetSum.
    - Models cannot learn the concept of **Aspect** with the limited aspects (4-5 aspects)
    - It needs datasets with diverse aspects!
  - We propose an Self-Supervised Pre-training!

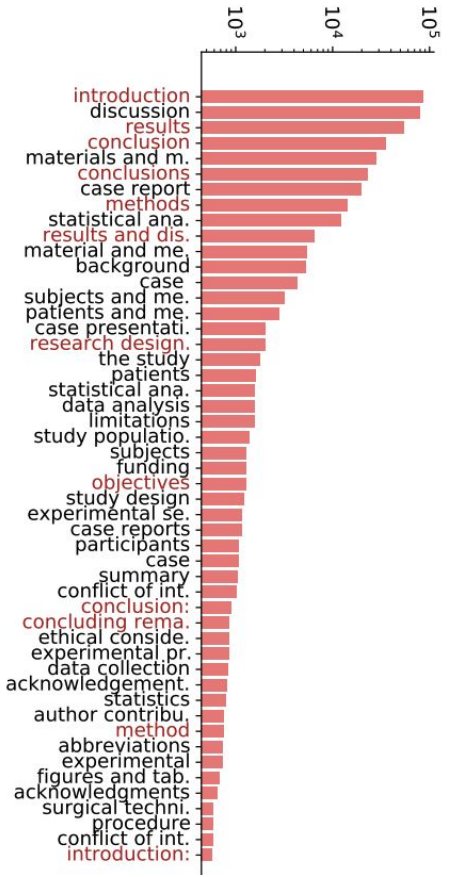
Self-Supervised Pre-training!

- Use section and sub-section headings as **Aspects** and their following sentences as **Summaries**!
- Generate a Summary given a Paper and a Heading!
- Aspect-based text generation!



Self-Supervised Pre-training!

- Histogram of 50 most frequent aspects in PubMed samples
- We have much more diverse aspects!
- PubMed★ dataset (658K)
- FacetSum★ dataset (279K)
- PubMed★ has [150069, 1452, 214, 33] unique aspects with frequency of higher than [1, 10, 100, 1000]
- FacetSum★ → [96525, 841, 120, 21]



Zero-Shot Performance

PubMed					FacetSum				
Pre-Train	**Train**	**R-1**	**R-2**	**R-L**	**Pre-Train**	**Train**	**R-1**	**R-2**	**R-L**
Fully Supervised BART Baseline									
-	PubMed	39.03	18.47	34.10	-	FacetSum	36.97	15.50	31.48
Lower-bound BART Shuffle Aspect Baseline									
-	PubMed	24.21	6.18	19.86	-	FacetSum	28.18	6.94	22.71
Domain Shift: Out-Of-Domain Labelled Data & Unlabelled									
-	FacetSum	28.89	10.20	24.52	-	PubMed	31.03	10.04	25.75
PubMed⋆	FacetSum	**31.31**	**11.53**	**26.79**	FacetSum⋆	PubMed	**31.67**	**10.34**	**26.25**
PubMed⋆ (No Overlap)	FacetSum	30.37	10.68	25.69	FacetSum⋆ (No Overlap)	PubMed	31.17	10.10	25.90
FacetSum⋆	FacetSum	28.92	10.12	24.46	PubMed⋆	PubMed	30.48	9.48	25.29
Only Unlabelled Data									
PubMed⋆	-	**30.76**	**11.64**	**26.16**	FacetSum⋆	-	**28.18**	**7.60**	**23.54**
PubMed⋆ (No Overlap)	-	29.70	10.93	25.20	FacetSum⋆ (No Overlap)	-	26.90	6.67	22.45
FacetSum⋆	-	28.68	9.79	24.30	PubMed⋆	-	27.24	7.01	22.34

Table 5: Performance on PubMed and FacetSum when out-of-domain training data is available (domain shift) or only unlabelled data is available. PubMed⋆ and FacetSum⋆ are the self-supervised datasets for pre-training.

- In-domain pre-training can improve the performance of models that have later an intermediate-training step.
- The best performance: Pre-trained on the same but unlabeled dataset and fine-tuned on the other dataset.
- PubMed⋆ results in more improvement because of its larger size.

Baselines

	Model	R-1	R-2	R-L
PubMed Generic	Discourse (Cohan et al., 2018)	38.93	15.37	35.21
	PEGASUS (Zhang et al., 2020)	39.98	15.15	25.23
	BART	45.04	18.45	40.62
PubMed	Greedy Extractive (Oracle)	56.61	39.23	47.58
	BART	39.03	18.47	34.10
	BART-Independent†	38.91	18.21	33.89
	BART Shuffle Aspects	24.21	6.18	19.86
FacetSum Generic	BART (Meng et al., 2021)	45.49	18.10	42.74
	BART-Facet (Meng et al., 2021)	49.29	19.60	45.76
	BART	49.98	19.89	46.68
FacetSum	Greedy Extractive (Oracle)	51.87	32.09	41.55
	BART (Meng et al., 2021)	23.27	10.31	20.29
	BART-Facet (Meng et al., 2021)	37.97	15.17	32.08
	BART	36.97	15.50	31.48
	BART-Independent†	36.77	15.26	31.23
	BART Shuffle Aspects	28.18	6.94	22.71

Leave-One-Out Performance

			PubMed			FacetSum		
Pre-Train	**Train**	**Test**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
✗	All - Introduction	Introduction	30.88	11.65	25.66	-	-	-
✓	All - Introduction	Introduction	40.07	21.22	35.5	-	-	-
✓✓	All - Introduction	Introduction	38.76	20.29	33.86	-	-	-
✗	All - Objectives	Objectives	28.97	8.97	22.99	29.08	8.33	23.87
✓	All - Objectives	Objectives	34.28	14.26	28.06	36.28	12.92	29.74
✓✓	All - Objectives	Objectives	30.69	10.60	24.84	29.15	8.28	23.77
✗	All - Methods	Methods	25.68	7.03	21.10	27.32	6.59	22.16
✓	All - Methods	Methods	27.28	7.70	22.23	28.13	6.84	22.79
✓✓	All - Methods	Methods	27.41	7.89	22.8	28.07	6.59	22.63
✗	All - Results	Results	21.28	4.68	17.92	23.82	5.25	19.47
✓	All - Results	Results	22.86	5.05	19.51	23.07	4.80	18.90
✓✓	All - Results	Results	21.12	4.67	17.79	24.22	5.28	19.83
✗	All - Conclusion	Conclusion	27.92	7.36	21.86	-	-	-
✓	All - Conclusion	Conclusion	31.23	9.17	24.73	-	-	-
✓✓	All - Conclusion	Conclusion	30.03	8.13	23.49	-	-	-
✗	All - Value	Value	-	-	-	30.41	7.86	24.22
✓	All - Value	Value	-	-	-	31.45	7.92	25.05
✓✓	All - Value	Value	-	-	-	29.25	7.41	23.52

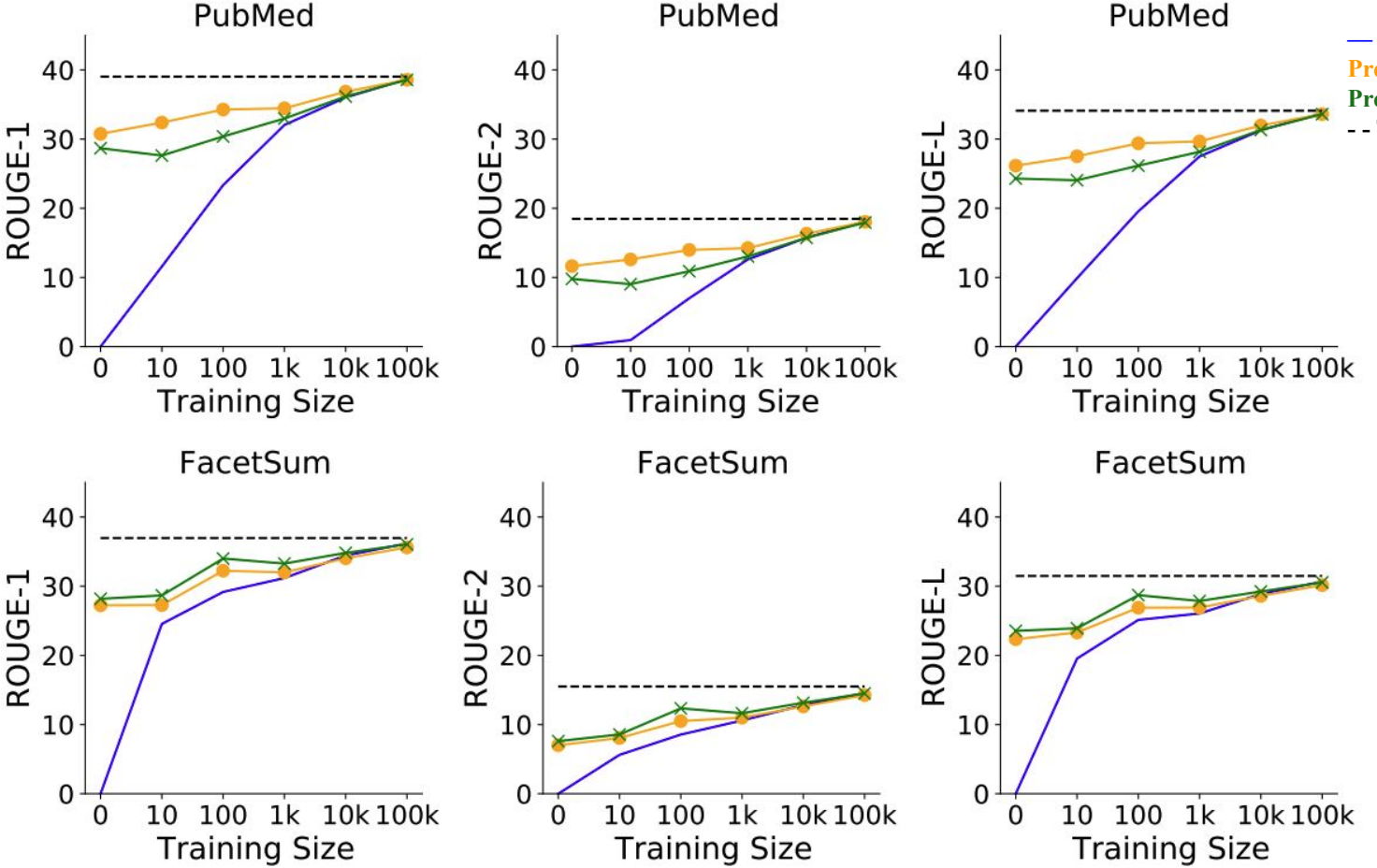
Table 6: Leave-one-out experiment on the PubMed and FacetSum aspect-based summarization datasets. The models are trained on all aspects except the one which the model is tested on. ✗: no pre-training except the BART official pre-training. ✓: model is pre-trained on PubMed* or FacetSum* (in-domain). ✓✓: model is pre-trained on PubMed* (No Overlap) or FacetSum* (No Overlap) (in-domain).

- Pre-training improves the leave-one-out performance!
- No Overlap pre-training is also promising

Paraphrasing

		PubMed			FacetSum		
Pre-Train	**Paraphrased Aspect**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
✗	Introduction	40.66	22.12	36.18	-	-	-
✗	Introduction -> Background	27.98	9.34	23.62	-	-	-
✗	Introduction -> Context	30.37	11.92	25.95	-	-	-
✓	Introduction -> Background	41.47	22.48	36.79	-	-	-
✓	Introduction -> Context	40.28	21.58	35.64	-	-	-
✗	Objectives	51.45	31.79	46.09	48.83	29.10	43.46
✗	Objectives -> Objective	51.37	31.66	46.03	48.91	29.17	43.52
✗	Objectives -> Purpose	36.03	15.93	29.84	46.70	26.11	41.11
✗	Objectives -> Aims	28.89	9.29	23.02	30.95	9.64	25.34
✓	Objectives -> Objective	51.10	31.39	45.60	48.51	28.81	43.14
✓	Objectives -> Purpose	49.77	29.92	44.09	48.28	28.46	42.88
✓	Objectives -> Aims	42.67	22.99	36.72	45.19	24.82	39.55
✗	Methods	40.78	19.08	35.84	32.79	11.71	27.64
✗	Methods -> Method	40.67	18.75	35.753	32.94	11.82	27.73
✗	Methods -> Materials and Methods	40.84	19.16	35.82	32.98	11.75	27.82
✗	Methods -> Research Design	34.82	14.23	29.74	32.68	11.34	27.41
✗	Methods -> Methodology	40.88	19.13	35.90	32.92	11.82	27.81
✓	Methods -> Method	41.13	19.24	36.07	32.85	11.88	27.69
✓	Methods -> Materials and Methods	40.58	19.05	35.58	32.77	11.80	27.69
✓	Methods -> Research Design	38.22	17.18	33.12	32.84	11.81	27.62
✓	Methods -> Methodology	40.82	19.24	35.75	32.77	11.82	27.62
✗	Results	34.73	12.91	30.69	32.67	10.21	27.43
✗	Results -> Result	34.42	12.73	30.30	32.46	10.05	27.21
✗	Results -> Discussion	23.57	7.09	20.09	26.12	5.90	21.25
✗	Results -> Finding	24.85	6.01	21.37	26.63	6.40	21.81
✓	Results -> Result	34.12	12.53	30.00	32.46	9.98	27.22
✓	Results -> Discussion	19.80	4.18	16.65	29.06	7.82	23.93
✓	Results -> Finding	29.11	9.24	25.29	32.46	10.01	27.20
✗	Conclusion	34.03	14.11	28.17	-	-	-
✗	Conclusion -> Conclusions	33.97	14.13	28.16	-	-	-
✓	Conclusion -> Conclusions	33.94	13.92	28.04	-	-	-
✗	Value -> Value	-	-	-	33.58	10.98	27.38
✗	Value -> Values	-	-	-	32.24	10.59	26.98
✓	Value -> Values	-	-	-	33.46	10.99	27.35

Few-Shot Performance



Conclusion

- Self-supervised pre-training improves the zero-shot and few-shot performance.
- It works both for domain-shift and unseen aspects.
- Intermediate training (training on out of domain data) improves the performance.
- What about complex aspects? (introduction and contribution)
- Does the improvement occur only for ROUGE?

Thank You!