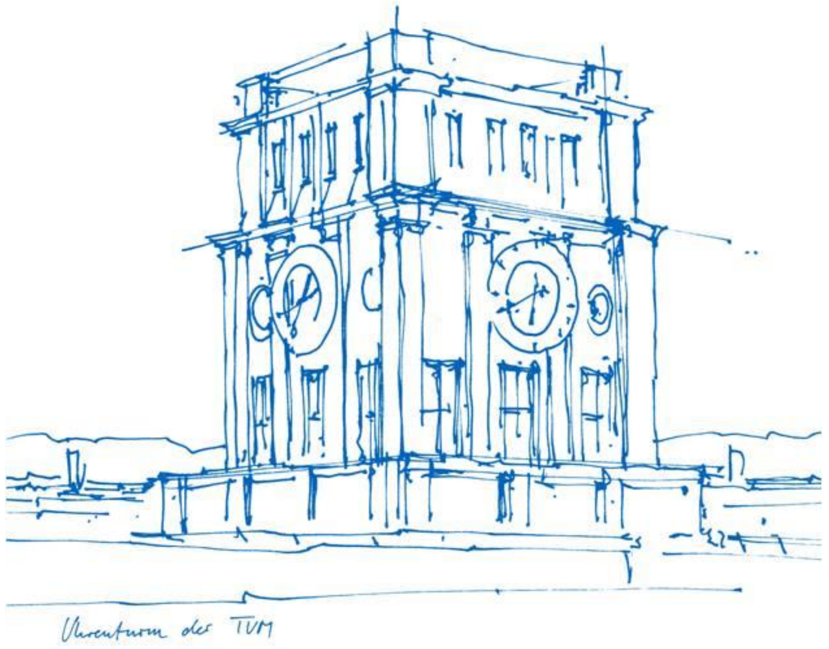
Scientific Document Representation Learning

Divya Bansal

M.Sc. Data Engineering and Analytics



Scientific Documents

- Store vast amounts of knowledge, amassed through many decades of research
- Primary means of disseminating new ideas, knowledge, and research findings

Information Overload!!

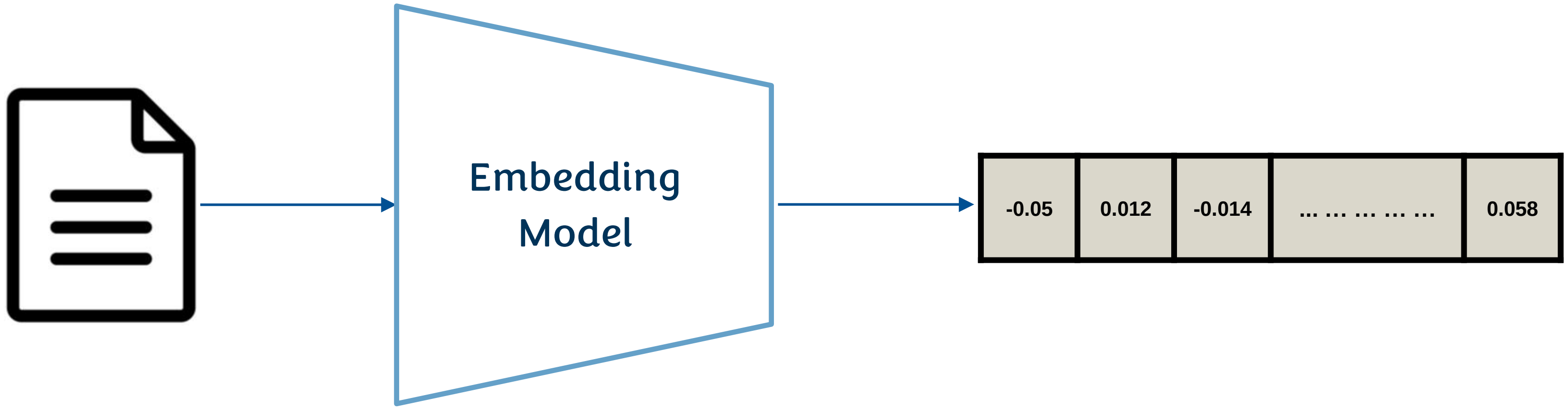
Solution: NLP

Applications:

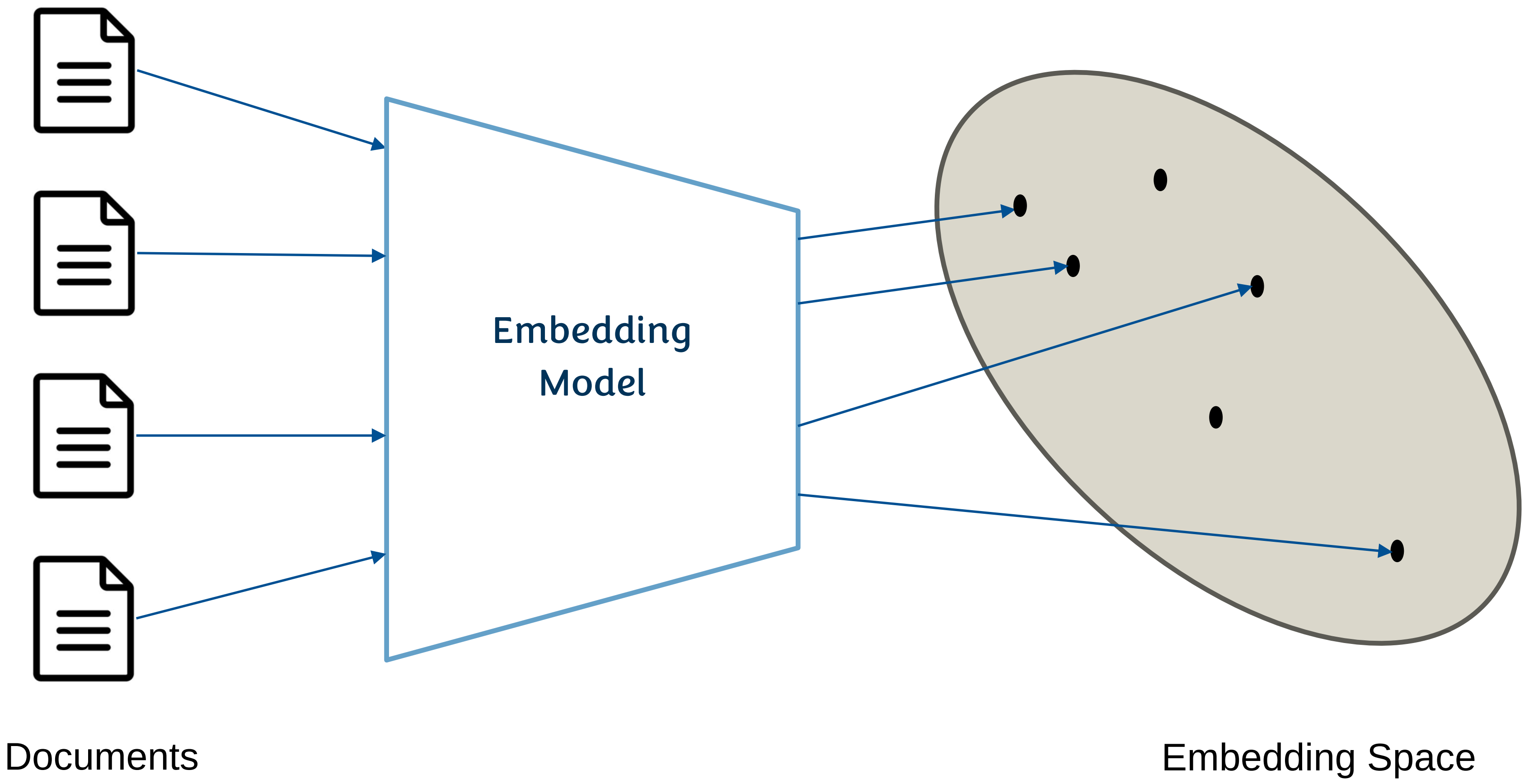
- Classification
- Recommendation
- Information Retrieval
- Clustering
- Information Extraction
- Summarization
- Discourse Analysis

and many more….

Representation Learning



Representation Learning



Objective

To represent scientific documents into vectors/embeddings such that they can be consumed effectively and efficiently for various downstream applications

Semantic Similarity

Present most intuitive way: PLMs like BERT

BERT:

- trained on general natural language such as Wikipedia
- limited by number of tokens (512)

However,

Scientific documents:

- long, complex, special structure, jargon

SciBERT (Beltagy et al., 2019)

BERT trained on 1.14M scientific papers (82% biology, 18% computer science)

SCIVOCAB: 42% overlap with BERT's vocabulary

Others: BioBERT, PubMedBERT, MatSciBERT, OAG-BERT etc.

SciBERT

BERT trained on 1.14M scientific papers (82% biology, 18% computer science)
SCIVOCAB: 42% overlap with BERT’s vocabulary

Others: BioBERT, PubMedBERT, MatSciBERT, OAG-BERT etc.

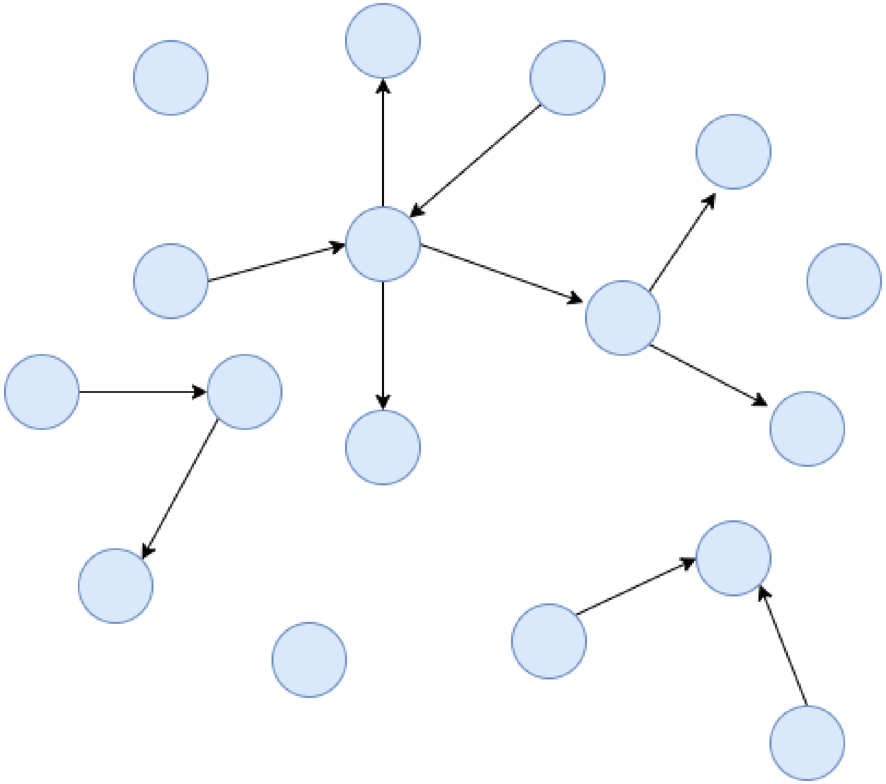
However, still token-level context

Incorporating citation information

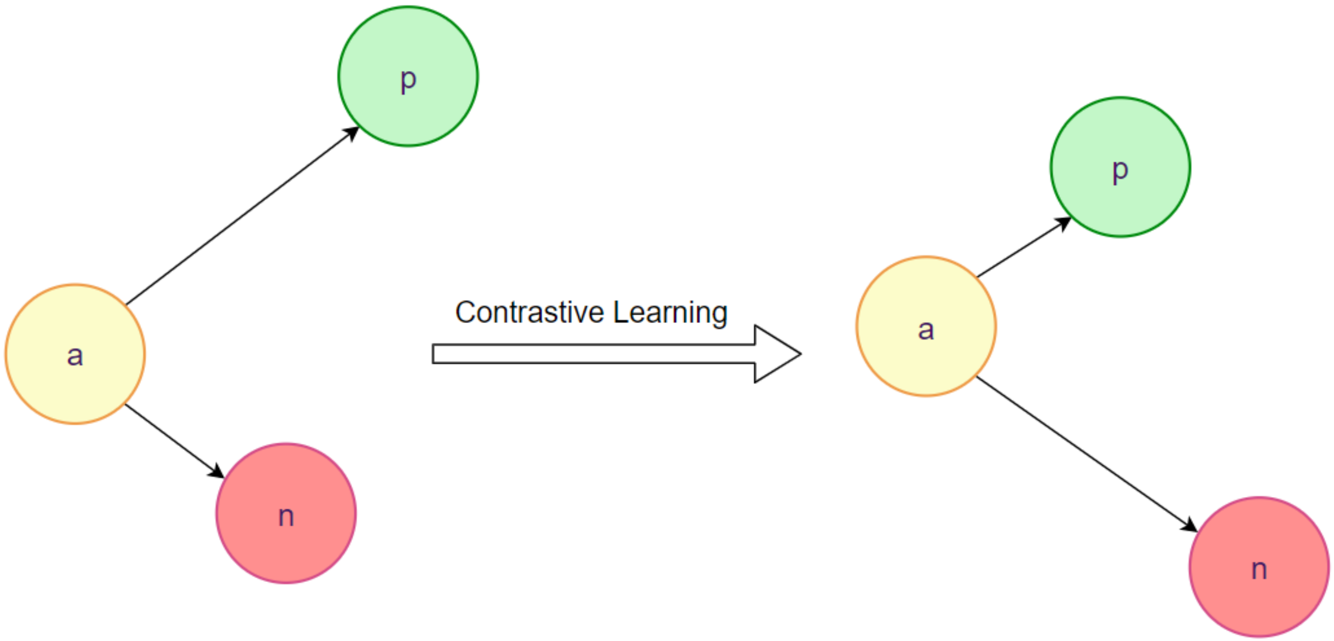
Recent models,

add **corpus-level context** from citation networks

and enhance semantic representations with **contrastive learning**



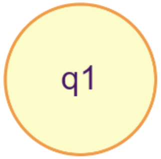
Contrastive Learning



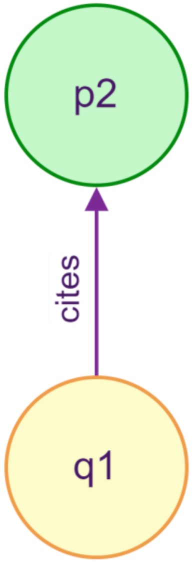
Contrastive Learning

- Anchor/query
- Positives
- Hard negatives
- Easy negatives

SPECTER (Cohan et al., 2020)

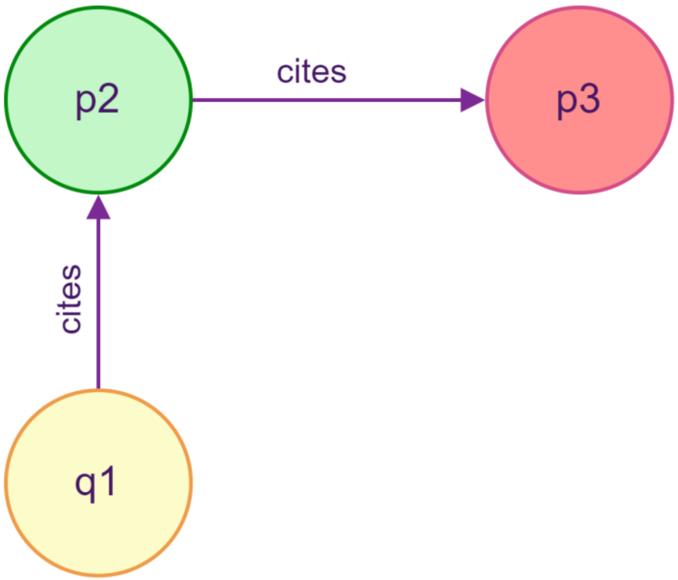


SPECTER



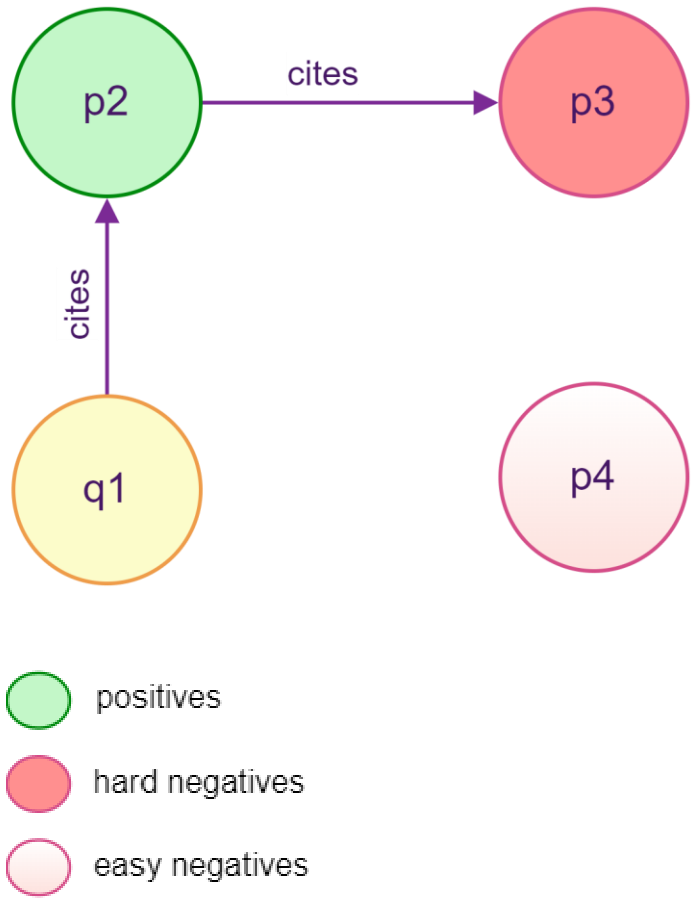
Positives: paper directly cited by the query paper (references)

SPECTER



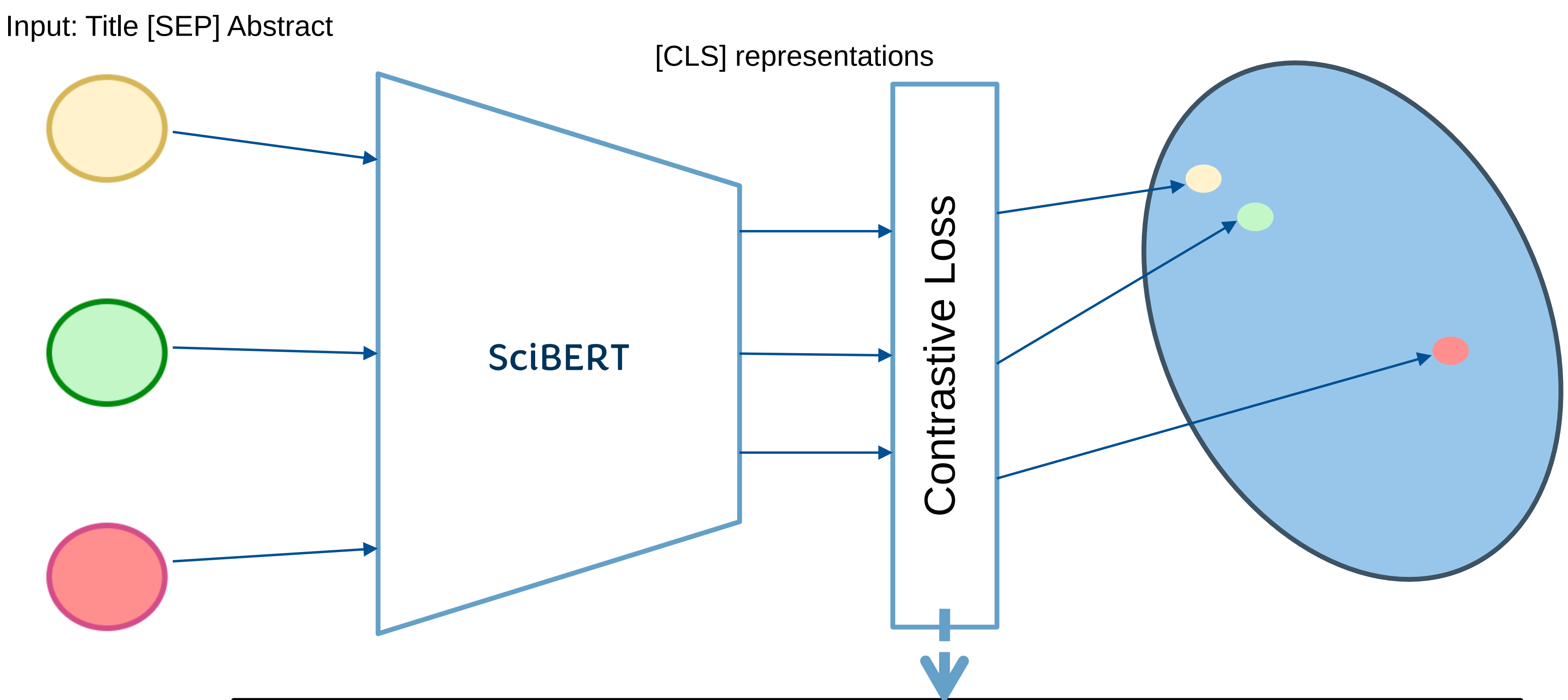
Hard negatives: papers not cited by the query paper but cited by the references of the query paper (references-of-references)

SPECTER



Easy negatives: random papers that are not in the above two categories

SPECTER



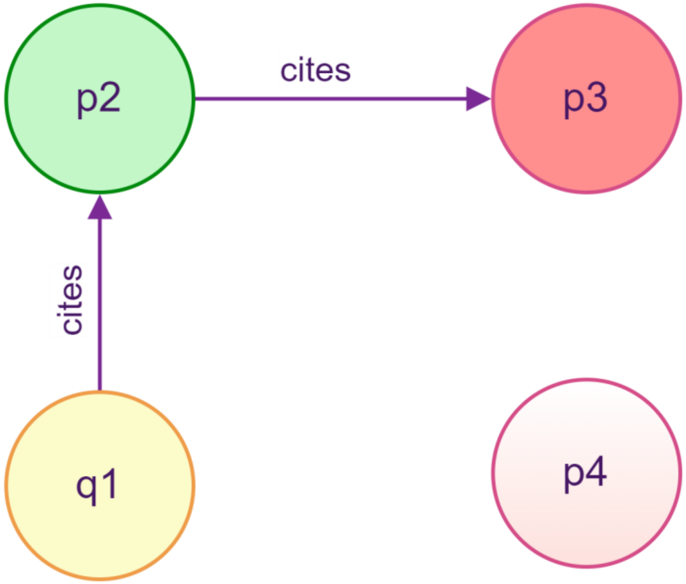
$$\mathcal{L}_{Triplet} = max(0, d(\boldsymbol{a}, \boldsymbol{p}) - d(\boldsymbol{a}, \boldsymbol{n}) + m)$$

SciNCL (Ostendorff et al., 2022)

Issues with SPECTER:

1. Positive and negative information collides between citation directions

Papers citing the query could be treated as easy negatives

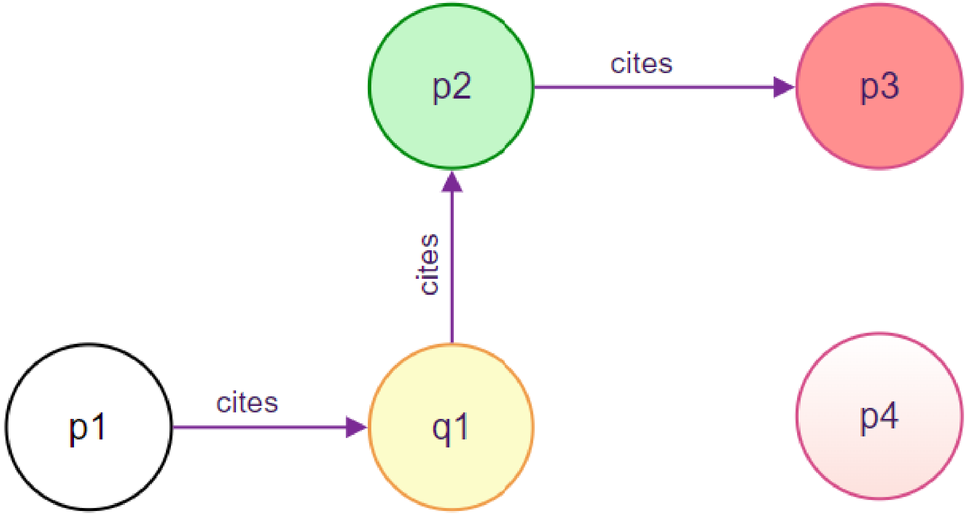


SciNCL

Issues with SPECTER:

1. Positive and negative information collides between citation directions

Papers citing the query could be treated as easy negatives



SciNCL

Issues with SPECTER:

2. Data Leakage

40.5% overlap between training and test data

SciNCL

Issues with SPECTER:

3. Scientific papers can be similar even without a direct citation link between them

Discrete citation relations to generate contrast samples enforce a hard cutoff for similarity and propagate human biases of which papers are similar

SciNCL

Employ controlled **nearest neighbour sampling over citation graph embeddings** for sampling positives and negatives for contrastive learning

SciNCL

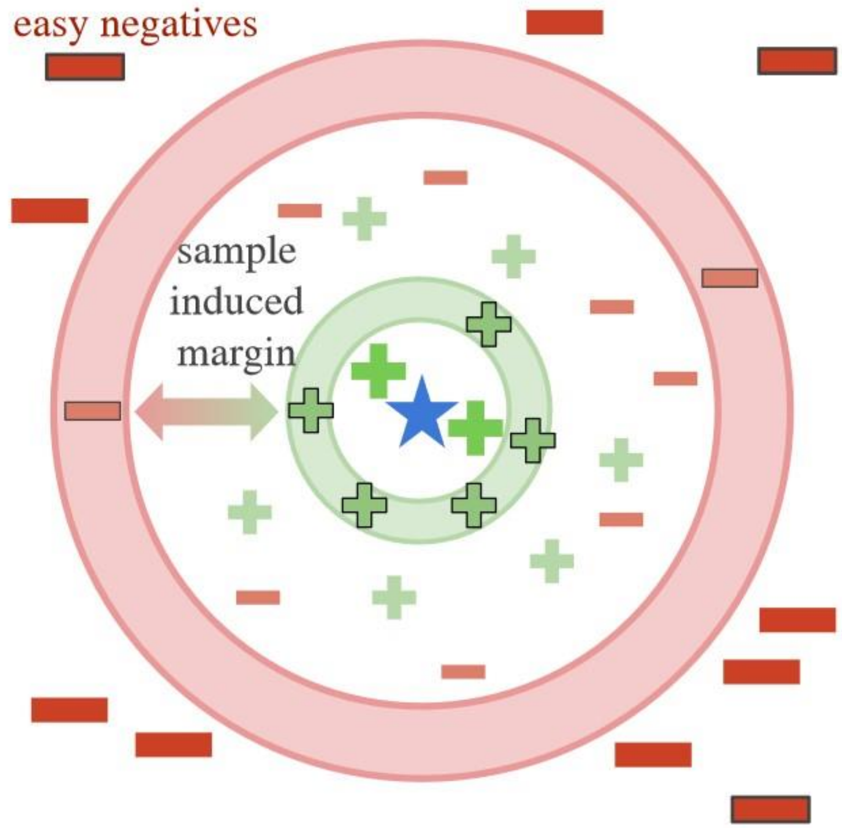
Steps:

1. Remove anchor papers that were present in the SciDocs benchmark and replace those papers with other randomly sampled papers

2. Train a citation embedding model on the whole citation network to get citation graph embeddings of all the papers

3. Re-train SPECTER with positives and negatives sampled using nearest neighbourhood search around the query papers in the citation graph embeddings

SciNCL



Query: (star symbol)

Positives: sampled from the close neighbourhood around the query embedding (+ symbol)

Hard negatives: potential positives but still farther from easy negatives by a certain margin such that they do not collide with positives (- in the red band)

Easy negatives: very distant from the query (- outside the red band)

Our Setup

Data:

We take the same anchor documents used by SciNCL (i.e., SPECTER without leakage) to train and validate our models

Our Setup

Data:

For each anchor document, we queried the Semantic Scholar API for its title, abstract, citations and references

Additionally, for each reference we also query the data of their references.

Thus, creating a local citation graph of each query paper.

$$\text{Citation} \rightarrow \text{Anchor paper} \rightarrow \text{Reference} \rightarrow \text{Reference}$$

Our Setup

Training:

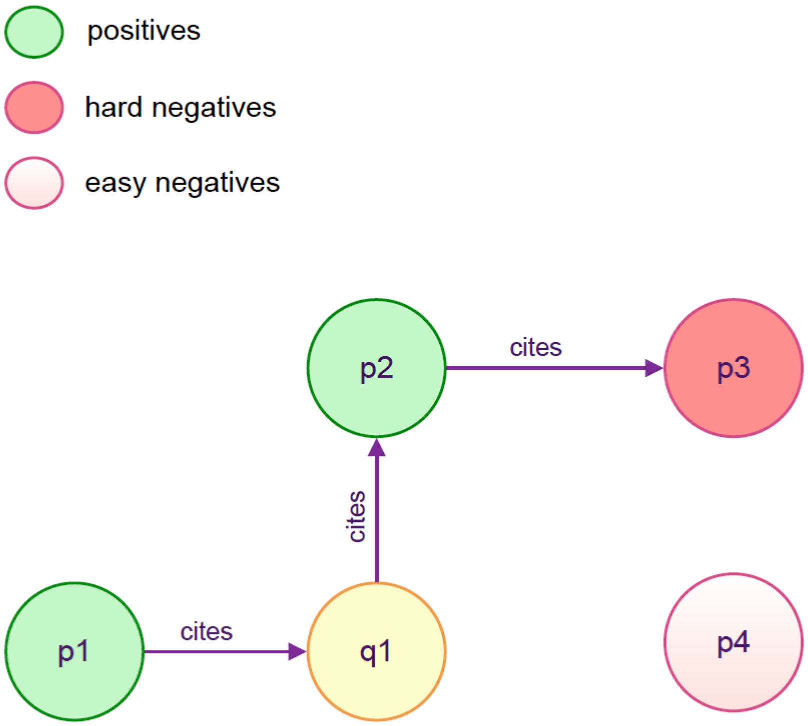
Build up on the SPECTER model:

Take concatenation of title and abstract and pass it through SciBERT for getting initial paper representations

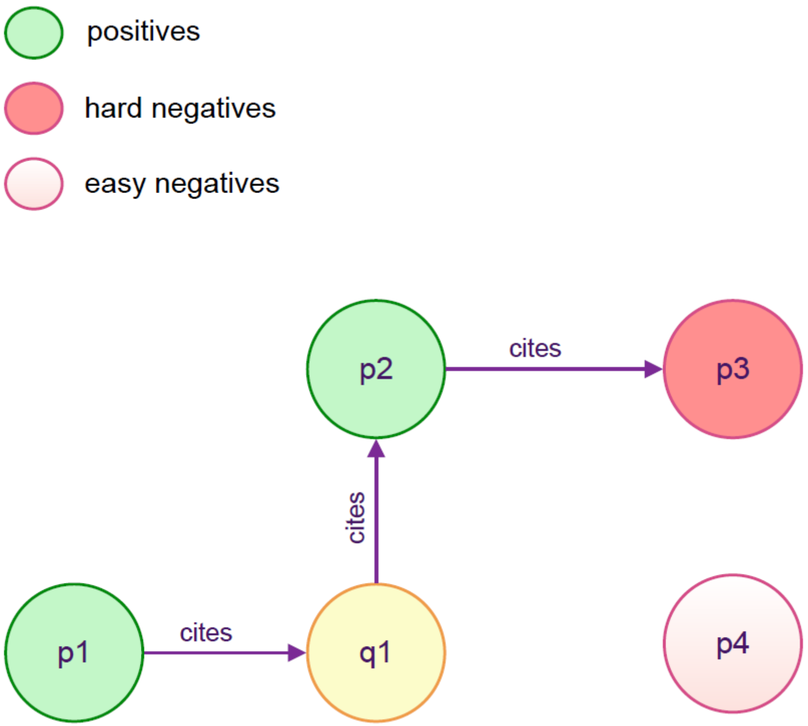
$$z_i = \text{SciBERT}([\text{CLS}]\ \text{title tokens}(i)\ [\text{SEP}]\ \text{abstract tokens}(i)[\text{SEP}])\ [\text{CLS}]$$

Experiment 1: SPECTER(Undirected)

Replication of SPECTER with in-citations as positives and no data leakage



Experiment 1: SPECTER(Undirected)



Positives: All papers directly connected with the query paper (direct citations and references)
Hard negatives: papers not cited by the query paper but cited by the references of the query paper (references-of-references)
Easy negatives: random papers that are not in the above two categories

Experiment 2: SPECTERCL

SPECTER(Undirected) but with general contrastive loss

$$\boxed{\mathcal{L}_{CL} = -\frac{1}{|P(a)|} \sum_{\boldsymbol{p} \in P(a)} \log \frac{\exp(sim(\boldsymbol{a}, \boldsymbol{p})/\tau)}{\sum_{\boldsymbol{p} \in P(a)} \exp(sim(\boldsymbol{a}, \boldsymbol{p})/\tau) + \sum_{\boldsymbol{n} \in N(a)} \exp(sim(\boldsymbol{a}, \boldsymbol{n})/\tau)}}$$

Experiment 2: SPECTERCL

SPECTER(Undirected) but with general contrastive loss

Ablations:

- Different number of positives, hard and easy negatives
- Different temperature values

Best model (According to results on SciDocs):

1 positive, n negatives
Temperature = 0.05

Aspect-Based Representations

- Scientific documents are multi-faceted and can be similar and dissimilar in many aspects
- A single notion of similarity leads the models to assume implicit biases
- Citations have different motivations and should not be treated equally

Aspect-Based Representations

We hypothesize:

Capturing aspect information, while learning the representations themselves, will make them capable of matching specific aspects and they can better capture overall document relatedness

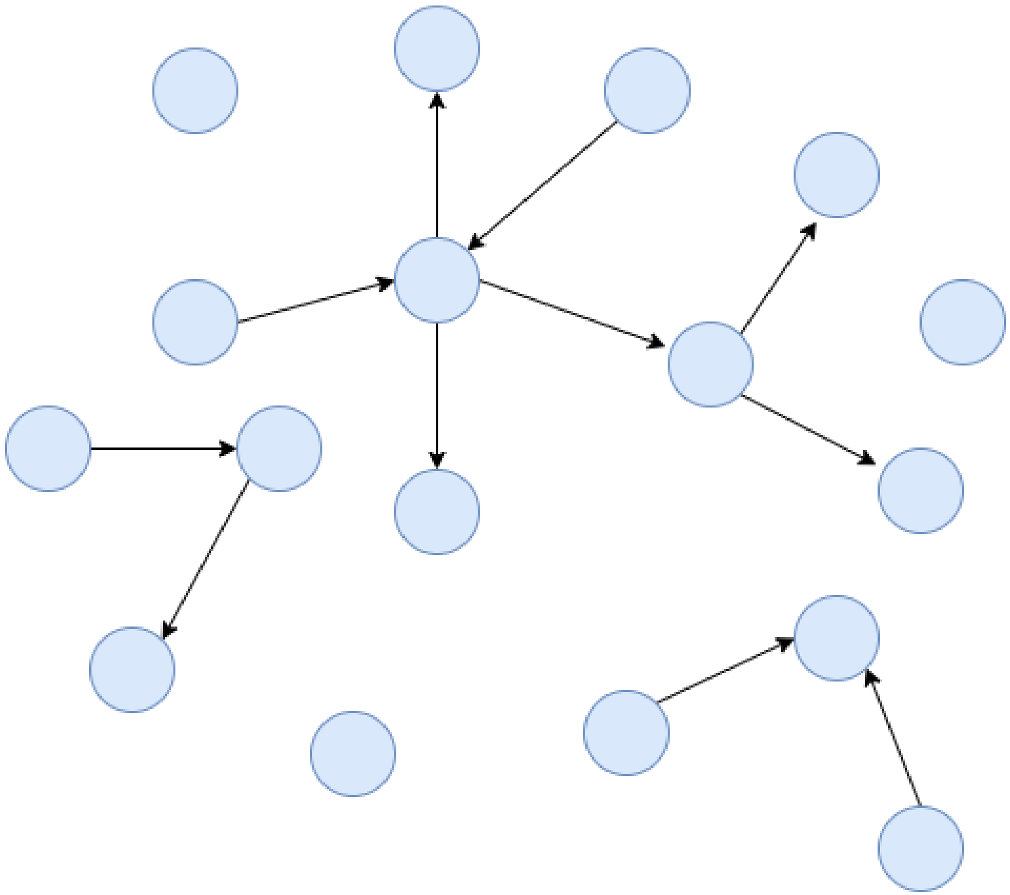
SciAspect

- To generate aspect-aware representations
- Consider, citation intent between two documents as a signal for capturing aspect-specific relatedness
- Employ a general contrastive loss with multiple positives/negatives as opposed to triplet loss
- **Hypothesis:** Papers connected directly with the same citation intents should have closer representations in the respective aspect spaces

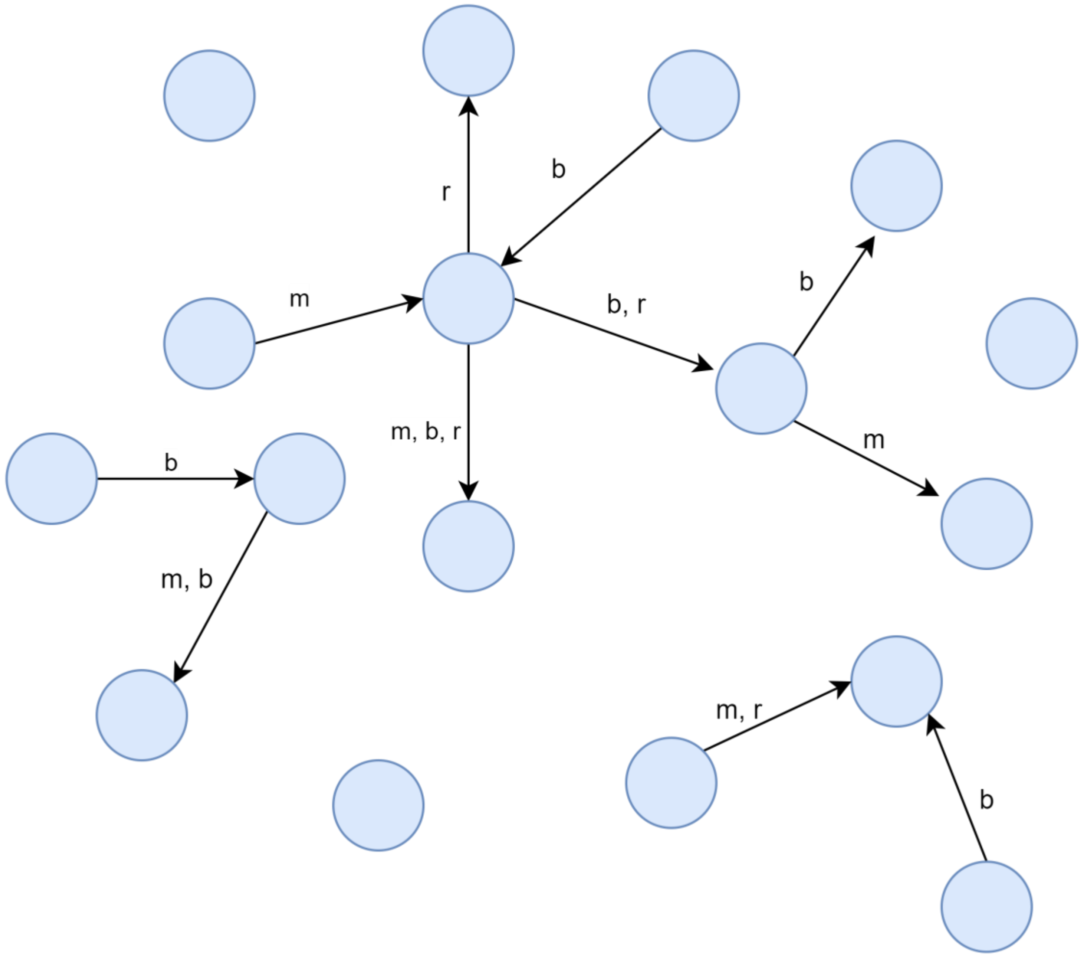
SciAspect

Intent Category	Description
Background information	The citation states, mentions, or points to the background information giving more context about a problem, concept, approach, topic, or importance of the problem in the field
Method	Making use of a method, tool, approach or dataset
Result comparison	Comparison of the paper's results/findings with the results/findings of other work

Aspect-Based Representations



Aspect-Based Representations



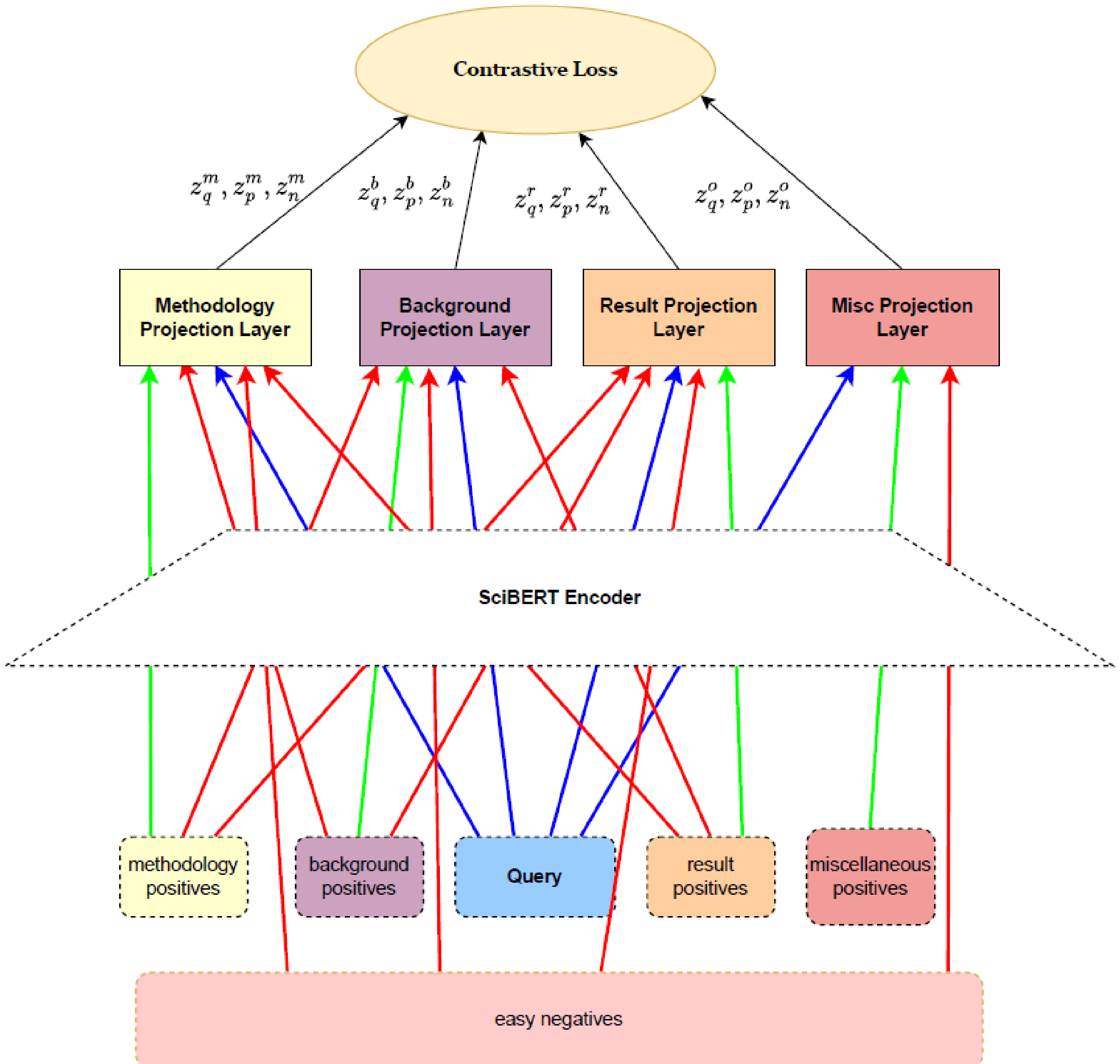
Experiment 3: SciAspect with L1 negatives

Like SPECTERCL but after SciBERT,

the query, positives and negatives are projected into aspect spaces

depending on how they are related for performing contrastive learning in the respective spaces

Experiment 3



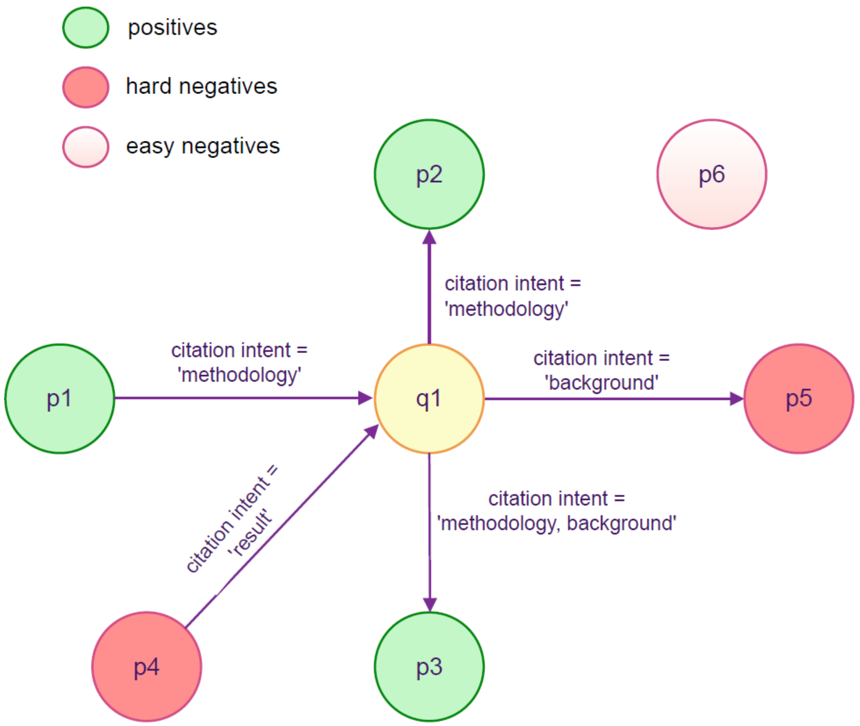
Experiment 3: SciAspect with L1 negatives

Positives: All papers directly connected to the query paper in the given aspect

Hard negatives: All papers directly connected to the anchor paper but in a different aspect (level 1/L1), excluding all anchor papers and positive candidates (as papers can be related in multiple aspects)

Easy negatives: random papers that are not in the above two categories

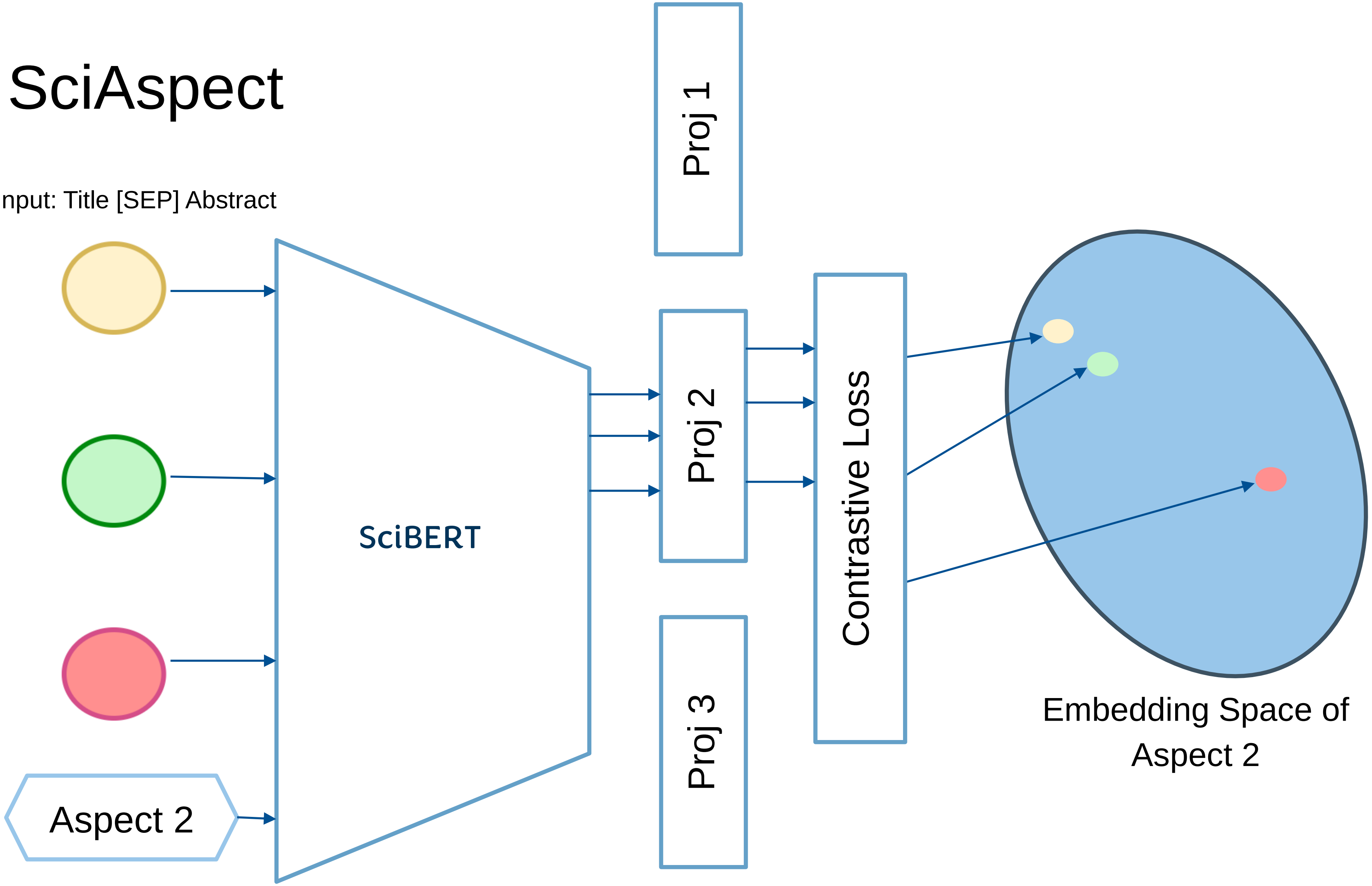
Experiment 3



An example of positives and negatives for the methodology aspect

SciAspect

Input: Title [SEP] Abstract

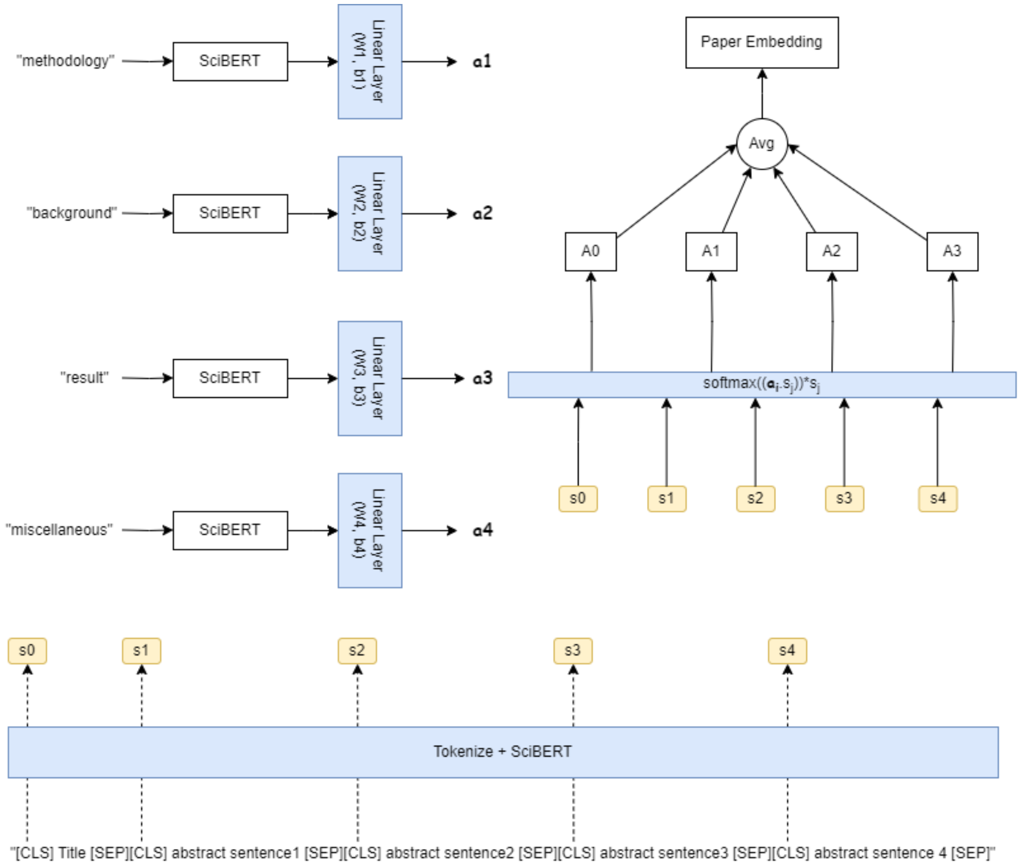


Experiment 3: SciAspect with L1 negatives

Ablations:

1. Different number of positives, negatives
2. Presence and absence of miscellaneous aspect
3. Aspect-specific embeddings
4. SciAspect (Weighted)

SciAspect (Weighted)



Experiment 3: SciAspect with L1 negatives

Best model (According to results on SciDocs):

- 1 positive, n negatives
- no miscellaneous aspect
- general unweighted embeddings

Experiment 4: SciAspect Hybrid

Global Approach (SPECTERCL):

Pull together papers connected through direct citations, otherwise push apart

Local Approach (SciAspect(L1)):

Pull together papers connected through direct citations with a specific citation intent, otherwise push apart

Hybrid:

Strike a balance and learn from both global(too wide) and local(too narrow) similarity signals

Experiment 4: SciAspect Hybrid

$$\mathcal{L}_{hybrid} = \alpha * \mathcal{L}_{global} + (1 - \alpha) * \mathcal{L}_{local}$$

Experiment 4: SciAspect Hybrid

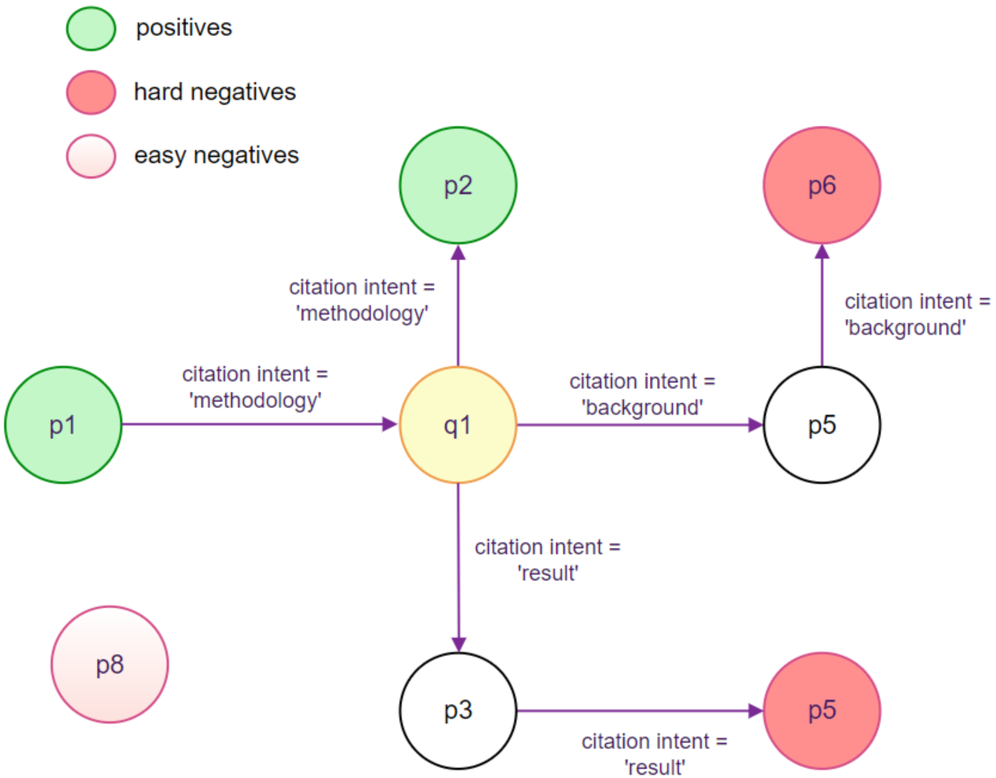
Ablations:

1. Aspect-specific embeddings
2. Different weights (alphas)

Best model (According to results on SciDocs):

$\alpha = 0.7$
generic embeddings

Experiment 5: SciAspect with L2 negatives



Positives: All papers directly connected to the query paper in the given aspect
Hard negatives: All papers not connected with the anchor paper through reference-of-references (level 2/L2) in the aspect that the positive is selected from
Easy negatives: random papers that are not in the above two categories

Evaluation

We use standard scientific document representation benchmarks to evaluate the performance of our resulting embeddings

Each document in these is represented with a title and abstract, which we pass through the encoder layer of our learned models to generate the embeddings for evaluation

Simple linear SVM-based classifiers/regressors and similarity-based recommenders for evaluating the quality of **frozen** embeddings to judge their zero-shot performance on a variety of downstream tasks

Evaluation

- **SciDocs (7 tasks)**
    - **Document Classification (multiclass):**
      Medical Subject Heading Classification (MeSH), Paper Topic Classification (MAG)
    - **Citation Prediction:**
      Direct citation prediction, Co-cited prediction
    - **User Activity:**
      Co-Views, Co-Reads
    - **Recommendation**
- **MDCR (Multi-Domain Citation Recommendation)**

Evaluation

- **<u>SciRepEval (25 tasks)</u>**

  - Ad-Hoc Search (QRY)

    In-train: Search

    Out-of-train: Trec-COVID, Feeds Dataset

  - Proximity (PRX)

    In-train: Same author prediction, citation prediction, influential citation prediction

    Out-of-train: Paper reviewer matching, SciDocs Citation and user Activity Prediction Tasks, Feeds Dataset

  - Classification (CLF)

    In-train: MeSH Descriptors (multiclass)

    Out-of-train: SciDocs MAG and MeSH classification (multiclass), Biomimicry (binary), DRSM (multiclass)

  - Regression (RGN)

    In-train: Citation Count Prediction, Publication Year

    Out-of-train: Tweet Mentions, Peer Review Ratings, Maximum h-Index of authors

SciDocs Results

Task→	Classification		User activity prediction				Citation prediction				Recomm.		Avg
Subtask→	MAG	MeSH	Co-View		Co-Read		Cite		Co-Cite				
Model/Metric↓	F1	F1	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	nDCG	P@1	
SciBERT*	79.43	79.92	59.81	78.1	55.71	75.33	53.17	73.76	57.67	77.33	51.72	17.59	63.295
SPECTER*	81.3	88.4	83.1	91.3	84.0	92.1	86.2	93.9	87.8	94.7	52.2	17.5	79.4
SciNCL*	81.3	89.4	84.3	91.8	85.6	92.8	**91.4**	**96.3**	90.1	95.7	54.3	19.9	81.1
SPECTER(Undirected)	81.78	**89.84**	**84.66**	**92.01**	85.81	93.05	90.45	95.97	89.88	95.6	52.34	17.02	80.70
SPECTERCL (1p, 5n)	81.99	89.25	84.53	91.94	**86.14**	**93.2**	89.76	95.59	**90.26**	**95.82**	**54.66**	**20.56**	**81.14**
SciAspect	82.65	89.0	83.69	91.5	84.2	92.19	87.68	94.59	88.68	95.11	52.36	17.65	79.94
SciAspectHybrid	82.72	89.62	84.33	91.9	84.6	92.32	86.85	94.16	89.34	95.44	53.07	18.85	80.27
SciAspectL2	**82.75**	88.36	84.16	91.76	84.95	92.61	85.62	93.55	89.15	95.36	53.74	18.89	80.07

* The baseline scores taken from [10]

MDCR Results

Models	BM25		SCIBERT		SPECTER		SciNCL		SPECTERCL		SciAspect		SciAspectHybrid	
Fields	MAP	R @ 5	MAP	R @ 5	MAP	R @ 5	MAP	R @ 5	MAP	R @ 5	MAP	R @ 5	MAP	R @ 5
Art	38.2	32.3	22.4	16.6	34.1	28.8	34.7	29.2	**39.5219**	**33.175**	35.8884	29.675	35.3471	29.225
Bio	38.3	**33.6**	20.4	14.0	34.6	30.0	36.8	32.3	**38.3833**	**33.6**	37.8107	32.7	35.6078	29.8
Bus	28.1	22.5	19.1	13.1	27.5	21.8	28.5	24.6	29.614	24.4	29.6102	**24.9**	**29.6716**	24.3
Ch	**38.0**	**32.6**	20.0	13.7	33.7	29.3	36.5	31.5	37.4207	31.9	36.0408	30.4	35.0916	30.0
CS	34.8	30.5	19.5	12.7	35.6	30.4	**37.2**	**32.5**	35.0327	29.8	33.5129	28.6	33.6166	29.6
Eco	**30.5**	**26.0**	21.4	15.4	27.3	21.9	28.3	23.2	29.9341	24.6	29.0597	23.9	29.9081	24.5
Eng	**34.6**	**29.3**	20.5	13.9	31.3	27.3	34.2	28.0	32.7904	26.7	31.7213	26.4	31.9758	27.1
ES	31.6	**26.2**	21.3	15.1	30.1	24.2	31.5	25.5	**31.8268**	25.9	30.5862	24.0	30.4452	24.9
Geog	**31.8**	**27.8**	21.9	16.7	26.4	22.2	29.5	23.8	31.676	25.8	29.6594	23.8	29.0374	25.0
Geol	**33.1**	**28.0**	19.5	13.9	24.8	20.1	25.7	19.9	27.1988	22.0	25.7205	21.4	26.0201	21.2
His	**38.1**	**32.9**	20.8	15.2	27.1	20.6	30.9	23.9	33.6084	26.95	32.0197	25.1	32.134	26.025
MS	36.1	30.7	22.1	15.5	34.1	28.2	35.8	29.6	**36.5694**	**32.4**	34.5561	30.2	33.6648	28.6
Mat	35.3	28.3	22.8	18.3	34.2	28.9	34.9	30.1	**36.6159**	**31.0**	35.2147	29.2	34.8087	28.4
Med	38.6	32.5	22.0	16.4	41.4	36.3	42.7	36.5	**44.1619**	**38.3**	42.9074	37.8	41.8509	36.7
Phi	30.2	**25.7**	19.2	13.3	27.1	21.1	29.9	23.5	**31.6667**	25.55	30.4059	25.35	30.1626	24.65
Phy	**35.1**	**30.2**	23.9	18.1	30.8	26.3	34.5	30.0	32.6382	27.7	33.3889	28.9	33.2899	28.7
PS	28.6	**23.1**	19.4	14.0	24.2	18.0	26.4	21.7	**29.193**	23.0333	27.3483	21.8333	27.646	21.8667
Psy	32.5	28.9	20.3	16.2	32.3	28.1	34.2	30.5	34.6926	30.1	34.2729	29.7	**34.7621**	**30.8**
Soc	26.8	20.5	20.2	15.8	25.2	20.5	26.7	21.9	**29.1627**	**24.4**	28.7196	23.9	28.5986	22.5
Avg	33.7	**28.5**	20.9	15.2	30.6	25.5	32.6	27.3	**33.7741**	28.2794	32.5497	27.2504	32.2968	27.0456

SciRepEval Results

Metric	SciBERT*	SPECTER (w/ leakage)*	SciNCL (w/ leakage)*	SPECTER (Undirected)	SPECTERCL	SciAspectHybrid	SciAspect	SciAspect(L2)
CLF Avg	70.02	73.99	74.32	75.43	75.48	76.13	**76.16**	75.42
REG Avg	23.22	22.15	23.91	24.09	17.35	**25.62**	24.63	25.43
(Excluding SciDocs) PRX Avg	59.99	67.23	**67.71**	67.25	67.08	66.63	66.55	66.78
QRY Avg	72.54	80.36	**80.64**	80.32	80.55	79.04	79.19	79.09
Out of task Avg	49.88	53.74	54.15	54.05	53.39	**54.41**	54.04	54.12
In task Avg	54.71	57.89	59.03	59.43	54.98	59.40	59.45	**59.49**
Scidocs Avg	69.04	89.10	90.83	**89.88**	89.85	89.10	88.86	88.80
All avg	58.05	67.76	**68.82**	68.52	67.25	68.37	68.14	68.16
Avg without SciDocs	51.59	55.20	55.87	55.95	53.95	**56.17**	55.95	56.01

Results

- Our replication of the SPECTER model, modified with a few changes suggested by SciNCL, performed significantly better than SPECTER itself in many tasks
- All our models overall outperformed SPECTER
- All our proposed models also outperformed the reported baselines in the classification tasks.
- SPECTERCL improved baselines in SciDocs and MDCR but not SciRepEval

Results

- Specifically, aspect-aware models outperformed the other aspect-unaware models on the task of classifying MAG Fields in case of SciDocs, and Biomimicry and DRSM classification tasks in SciRepEval
- Among the aspect-based models, in general, SciAspectHybrid performed better than SciAspectL2 which was better than SciAspect
- Other than classification tasks, where SciAspect was better than SciAspectHybrid and SciAspectL2, we learned that adding directly connected papers with a different aspect as negative, harmed learning generalized embeddings

Sources of Error

- Limitations of SciBERT
  - biased towards specific domains (biology and computer science)
  - only title and abstract provide limited information for distinguishing documents

- Evaluation tasks appropriate for evaluating embedding generalizability but not aspect-awareness of the embeddings

- Limitations of the Semantic Scholar model for classifying citation intents
  - Model accuracy: 67.9% F1 on the ACLARC citations benchmark and 84% F1 on the SciCite benchmark

Conclusion

We were able to improve the performance of the SPECTER model by posing its triplet loss function as a general contrastive or hybrid loss function and by proposing newer methods for sampling positives and negatives for the loss.

We found that integrating aspect-specific information into the general structural and semantic information can potentially improve model performance, especially in our case for classification problems.

Overall, we were able to learn good generalizable embeddings that are comparable in performance with the existing approaches and can be productively consumed for downstream applications with minimal fine-tuning.

Future Work

- To use SciNCL based positive/negative sampling method. Citation embeddings can be learned on aspect-specific citation networks for positive/negative mining.
- Better qualitative and quantitative evaluation of aspect-awareness of the embeddings
- Test aspect-aware embeddings on finer-grained intent-based or context-based recommendation systems

Future Work

- Additional metadata, more complex neural network instead of linear projection for representing aspect spaces
- Graph Attention Transformers could be explored for modeling the system
- Hierarchical contrastive learning for even finer-grained representations
- Multi-task learning with added objectives for downstream tasks such as e classification, regression, ad-hoc and proximity

Thank You

Appendix

Task→	Classification		User activity prediction				Citation prediction				Recomm.		Avg
Subtask→	MAG	MeSH	Co-View		Co-Read		Cite		Co-Cite				
Model/Metric↓	F1	F1	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	nDCG	P@1	
Oracle SciDocs	87.1	94.8	87.2	93.5	88.7	94.6	92.3	96.8	91.4	96.4	53.8	19.4	83.0
Doc2Vec	66.2	69.2	67.8	82.9	64.9	81.6	65.3	82.2	67.1	83.4	51.7	16.9	66.6
fastText-sum	78.1	84.1	76.5	87.9	75.3	87.4	74.6	88.1	77.8	89.6	52.5	18.0	74.1
ELMo	77.0	75.7	70.3	84.3	67.4	82.6	65.8	82.6	68.5	83.8	52.5	18.2	69.0
Citeomatic	67.1	75.7	81.1	90.2	80.5	90.2	86.3	94.1	84.4	92.8	52.5	17.3	76.0
SGC	76.8	82.7	77.2	88.0	75.7	87.5	91.6	96.2	84.1	92.5	52.7	18.2	76.9
BERT	79.9	74.3	59.9	78.3	57.1	76.4	54.3	75.1	57.9	77.3	52.1	18.1	63.4
SciBERT	79.7	80.7	50.7	73.1	47.7	71.1	48.3	71.7	49.7	72.6	52.1	17.9	59.6
BioBERT	77.2	73.0	53.3	74.0	50.6	72.2	45.5	69.0	49.4	71.8	52.0	17.9	58.8
CiteBERT	78.8	74.8	53.2	73.6	49.9	71.3	45.0	67.9	50.3	72.1	51.6	17.0	58.8
Random S2ORC	in data (w/o leakage):												
SPECTER	81.3	88.4	83.1	91.3	84.0	92.1	86.2	93.9	87.8	94.7	52.2	17.5	79.4
SciNCL	81.3	89.4	84.3	91.8	85.6	92.8	91.4	96.3	90.1	95.7	54.3	19.9	81.1

Table 6.1.: Values reported in [10] by evaluating SciDocs for different models.

Appendix

Task→	Classification		User activity prediction				Citation prediction				Recomm.		Avg
Subtask→	MAG	MeSH	Co-View		Co-Read		Cite		Co-Cite				
Model/Metric↓	F1	F1	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	nDCG	P@1	
Different number of positives and negatives (t = 0.05)													
1 positive, 1 negative	81.81	88.19	83.88	91.66	84.37	92.18	88.74	95.03	88.94	95.16	54.33	19.95	80.35
1 positive, 5 negatives	81.99	89.25	**84.53**	**91.94**	**86.14**	**93.2**	**89.76**	**95.59**	**90.26**	**95.82**	**54.66**	**20.56**	**81.14**
1 positive, 3 easy negatives	82.68	**89.65**	84.27	91.86	84.87	92.5	88.03	94.75	89.54	95.44	53.94	19.22	80.56
2 positive, 5 negatives	82.51	88.62	83.52	91.48	83.7	91.9	88.15	94.82	88.39	94.85	52.6	17.7	79.85
4 positives, 5 negatives	82.38	88.37	83.65	91.54	83.55	91.82	88.31	94.86	88.2	94.77	52.13	16.93	79.71
5 positives, 5 negatives	**83.02**	89.32	83.45	91.49	83.61	91.86	87.96	94.68	88.11	94.73	52.27	17.61	79.84
Different temperatures (1 positive, 5 negatives)													
$\tau = 0.01$	81.84	88.86	84.1	91.69	84.82	92.52	87.38	94.41	89.15	95.33	54.03	19.57	80.31
$\tau = 0.05$	81.99	**89.25**	**84.53**	**91.94**	**86.14**	**93.2**	**89.76**	95.59	**90.26**	**95.82**	**54.66**	**20.56**	**81.14**
$\tau = 0.1$	**82.87**	89.02	84.43	91.92	85.03	92.68	89.7	**95.62**	89.33	95.44	51.93	17.45	80.45
$\tau = 0.5$	81.85	88.36	79.53	89.31	77.31	88.17	80.18	90.17	81.84	91.35	52.74	17.68	76.54

Table 6.3.: SciDocs results on SPECTERCL variants (Experiment 2)

Appendix

Task→	Classification		User activity prediction				Citation prediction				Recomm.		Avg
Subtask→	MAG	MeSH	Co-View		Co-Read		Cite		Co-Cite				
Model/Metric↓	F1	F1	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	nDCG	P@1	
base model	**82.72**	**89.62**	84.33	91.9	**84.6**	**92.32**	**86.85**	**94.16**	**89.34**	**95.44**	**53.07**	**18.85**	**80.27**
local loss w.r.t. methodology space	81.6	89.14	84.19	91.84	84.34	92.17	86.55	94.07	89.05	95.27	52.19	17.3	79.81
local loss w.r.t. background space	82.12	88.95	84.18	91.77	84.49	92.3	86.52	93.93	89.29	95.38	52.66	17.88	79.96
local loss w.r.t. result space	81.95	89.01	84.35	**91.96**	84.33	92.2	86.57	94.05	89.02	95.26	52.91	18.56	80.01
local loss w.r.t. concatenated embeddings	82.56	89.4	**84.4**	91.93	84.55	92.31	86.66	94.05	89.27	95.41	52.71	18.07	80.11

Table 6.5.: SciDocs results on SciAspectHybrid architecture variants($\tau = 0.05$, number of positives=1, number of negatives=m) (Experiment 4.1)

Task→	Classification		User activity prediction				Citation prediction				Recomm.		Avg
Subtask→	MAG	MeSH	Co-View		Co-Read		Cite		Co-Cite				
Model/Metric↓	F1	F1	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	nDCG	P@1	
alpha = 0.5	**82.91**	89.38	84.21	91.8	84.35	92.2	86.74	94.07	89.04	95.27	52.76	18.34	80.09
alpha = 0.6	82.76	89.54	84.28	91.83	84.48	92.25	**87.13**	**94.27**	89.23	95.38	52.79	18.14	80.17
alpha = 0.7 (base)	82.72	**89.62**	84.33	91.9	84.6	2.32	86.85	94.16	89.34	95.44	53.07	18.85	**80.27**
alpha = 0.8	82.65	89.51	84.37	91.93	84.72	92.41	86.74	94.09	89.36	95.44	52.95	18.4	80.21
alpha = 0.9	82.43	89.47	**84.44**	**91.97**	**84.83**	**92.48**	86.52	93.99	**89.39**	**95.45**	**53.2**	**18.86**	80.25

Table 6.6.: SciDocs results on SciAspectHybrid variants of parameter alpha($\tau = 0.05$, number of positives=1, number of negatives=m) (Experiment 4.2)

Appendix

Task→	Classification		User activity prediction				Citation prediction				Recomm.		Avg
Subtask→	MAG	MeSH	Co-View		Co-Read		Cite		Co-Cite				
Model/Metric↓	F1	F1	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	nDCG	P@1	
SciAspect 1 positive 3 easy negatives	82.64	**89.05**	83.95	91.66	83.93	91.97	**86.77**	**94.09**	88.65	94.98	**53.9**	19.33	**80.08**
SciAspect L2 negatives (5 hard negatives)	82.22	87.42	82.43	90.78	83.54	91.97	80.58	90.93	86.88	94.26	**53.9**	**19.76**	78.72
SciAspect L2 negatives (3 hard negatives, 3 easy negatives)	**82.75**	88.36	**84.16**	**91.76**	**84.95**	**92.61**	85.62	93.55	**89.15**	**95.36**	53.74	18.89	80.07

Table 6.7.: SciDocs results on SciAspect Experiment 5 models with L2 negatives

SciRepEval Full Results

Type	Task		Metric	SciBERT*	SPECTER (w/ leakage)*	SciNCL (w/ leakage)*	SPECTER(Undirected)	SPECTERCL	SciAspectHybrid	SciAspect	SciAspect(L2)
Out-of-Train	**Biomimicry [CLF]**	Complete	F1	73.37	72.87	69.74	72.27	73.05	74.52	72.29	**74.53**
		Few shot - 64 samples 50 runs	F1	37.26	39.62	40.14	42.11	42.06	**45.35**	45.13	44.04
		Few shot - 16 samples 100 runs	F1	16.00	19.50	21.26	22.12	23.02	25.62	**31.14**	19.85
			Wt. F1	50.00	51.22	50.22	52.19	52.79	55.00	**55.21**	53.24
	DRSM [CLF]	Complete	F1	76.84	**77.34**	74.73	74.71	75.27	75.37	76.35	76.11
		Few shot - 64 samples 50 runs	F1	56.31	61.06	61.24	62.04	63.88	64.02	**65.43**	63.50
		Few shot - 24 samples 100 runs	F1	46.05	48.88	49.68	52.61	54.09	55.43	**56.39**	53.78
			Wt. F1	64.01	66.16	65.10	66.02	67.13	67.55	**68.63**	67.37
	Feeds	Paper Query [PRX]	MAP	68.17	81.11	**81.16**	80.15	79.96	78.44	78.54	79.31
		Multi paper query [PRX]	MAP	65.44	74.35	**75.30**	73.44	73.87	72.83	72.52	72.85
		Title query [QRY]	MAP	66.42	**81.23**	80.72	78.65	79.80	76.52	76.96	76.39
	TREC CoVID [QRY]		nDCG	79.73	86.53	87.67	**89.00**	88.67	87.68	88.07	87.83
	Peer Reviewer Matching [PRX]	Hard decision	P@5	26.92	33.27	**34.21**	32.34	30.84	31.59	31.40	31.21
			P@10	24.30	25.51	25.42	25.42	25.05	**25.61**	25.23	25.05
		Soft decision	P@5	60.93	65.79	66.54	**66.92**	66.36	66.73	66.36	65.61
			P@10	54.58	**56.17**	55.42	55.98	55.61	**56.17**	55.79	55.89
			Avg	41.68	45.19	**45.40**	45.17	44.47	45.03	44.70	44.44
	Review Score [RGN]	ICLR 17-22	K Tau	20.26	17.35	18.87	20.60	14.07	**22.08**	21.57	21.40
	Max h-Index [RGN]		K Tau	6.81	10.04	11.30	13.97	11.74	12.82	13.67	**14.12**
	Tweet Mentions [RGN]		K Tau	22.18	24.19	25.78	21.31	21.37	**26.19**	20.50	24.21
In-Train	**MeSH [CLF]**		F1	76.71	85.46	86.17	**87.29**	86.09	85.91	85.80	85.41
	Same Author Prediction [PRX]		MAP	79.48	86.53	87.47	87.39	**87.91**	87.25	87.51	87.40
	Search [QRY/PRX]		nDCG	71.46	73.31	**73.54**	73.31	73.18	72.92	72.54	73.04
	Citation Context [PRX]		MAP	33.72	42.89	43.39	**44.05**	43.11	43.29	43.48	43.63
	Citation Count [RGN]		K Tau	**39.16**	33.21	34.61	36.16	23.70	36.99	36.98	37.10
	Publishing Year [RGN]		K Tau	27.71	25.96	29.00	28.40	15.88	30.01	**30.41**	30.34
SciDocs	**MAG [CLF]**		F1	79.54	79.40	81.11	81.55	81.97	82.50	82.09	**82.62**
	MeSH [CLF]		F1	79.84	87.70	89.00	**90.11**	89.41	89.67	89.08	88.45
	Co-View [PRX]		MAP	59.80	83.40	85.28	**84.63**	84.51	84.33	83.69	84.15
			nDCG	78.10	91.40	92.23	**91.99**	91.93	91.91	91.50	91.76
	Co-Read [PRX]		MAP	55.73	85.10	87.69	85.79	**86.15**	84.59	84.21	84.95
			nDCG	75.34	92.70	94.00	93.05	**93.20**	92.31	92.20	92.61
	Cite [PRX]		MAP	53.20	92.00	93.55	**90.36**	89.69	86.76	87.58	85.51
			nDCG	73.79	96.60	97.35	**95.93**	95.56	94.12	94.55	93.49
	Co-cite [PRX]		MAP	57.71	88.00	91.66	89.84	**90.27**	89.34	88.64	89.14
			nDCG	77.36	94.70	96.44	95.59	**95.83**	95.44	95.08	95.35
			CLF Avg	70.02	73.99	74.32	75.43	75.48	76.13	**76.16**	75.42
			REG Avg	23.22	22.15	23.91	24.09	17.35	**25.62**	24.63	25.43
			(Excluding SciDocs) **PRX Avg**	59.99	67.23	**67.71**	67.25	67.08	66.63	66.55	66.78
			QRY Avg	72.54	80.36	**80.64**	80.32	80.55	79.04	79.19	79.09
			Out of task Avg	49.88	53.74	54.15	54.05	53.39	**54.41**	54.04	54.12
			In task Avg	54.71	57.89	59.03	59.43	54.98	59.40	59.45	**59.49**
			Scidocs Avg	69.04	89.10	90.83	**89.88**	89.85	89.10	88.86	88.80
			All avg	58.05	67.76	**68.82**	68.52	67.25	68.37	68.14	68.16
			Avg without SciDocs	51.59	55.20	55.87	55.95	53.95	**56.17**	55.95	56.01