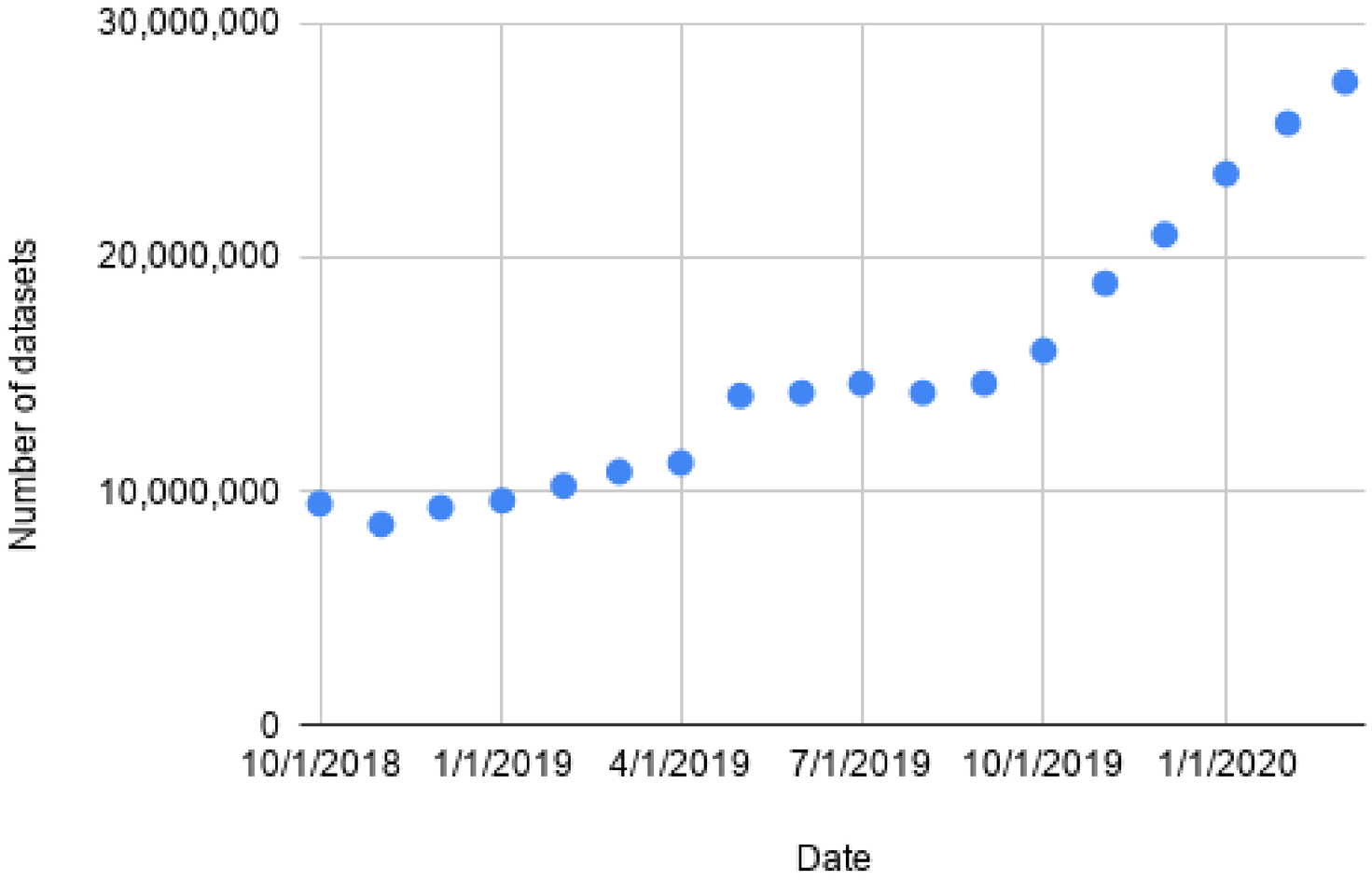
IIC3633

DATAFINDER: SCIENTIFIC DATASET RECOMMENDATION FROM NATURAL LANGUAGE DESCRIPTIONS

Pedro Z, Felipe T, Kahil R

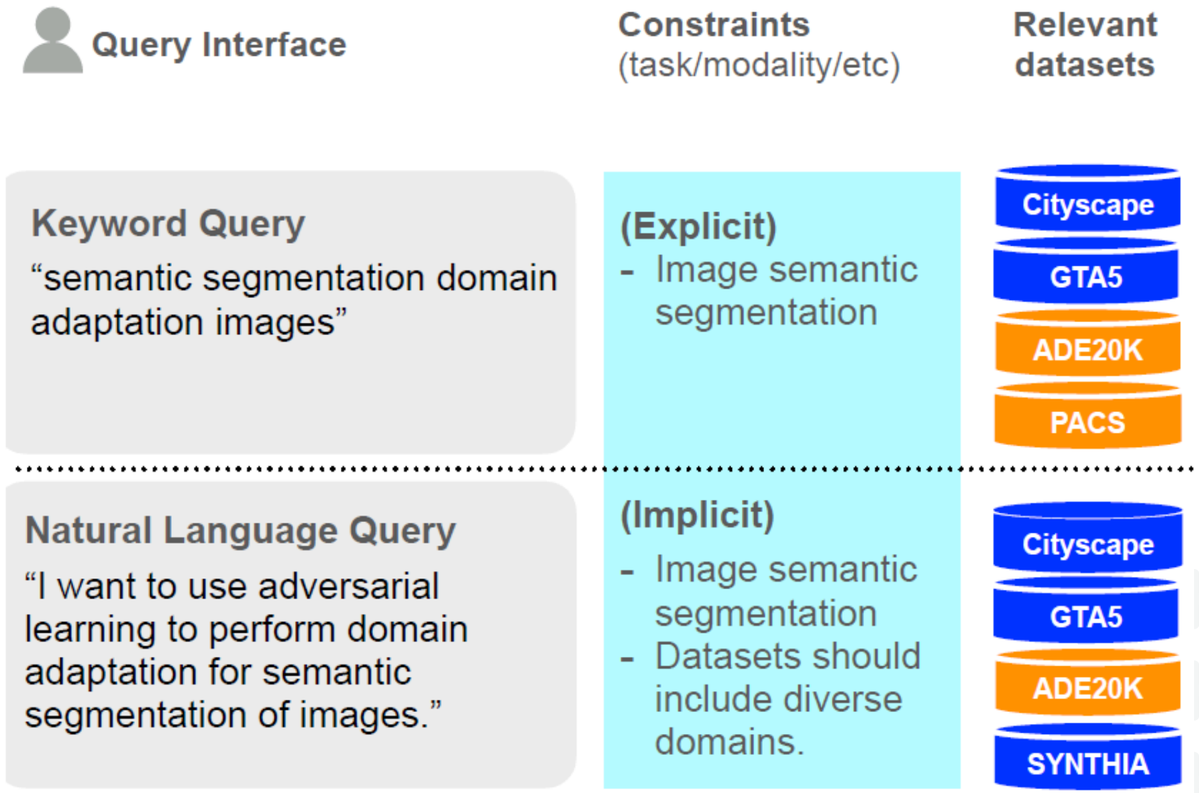
Contexto



Encontrar datasets relevantes para investigaciones específicas se ha vuelto un desafío

Dataset Recommendation

- Consultas a través de **keywords** o **lenguaje natural.**
- Los **requerimientos** para el dataset pueden ser explícitos o implícitos.
- Se deben entregar **datasets relevantes** para el problema presentado.



Estado del arte

Trabajos relacionados

01 Dataset Recommendation for Data Linking: An Intensional Approach (2016)

02 Dataset recommendation via variational graph autoencoder (2019)

03 Dense Passage Retrieval for Open-Domain Question Answering (2020)

04 Recommending Datasets for Scientific Problem Descriptions (2021)

- Existen motores de búsqueda como el **Google Dataset Search, Papers With Code o Kaggle**, limitados algoritmos de coincidencia de palabras y que no comprenden consultas de lenguaje natural.

Contribuciones

- DataFinder aborda el problema desde un **dominio abierto**, por lo que es capaz de manejar consultas complejas y diversas frente a los motores de búsqueda existentes.
- Utilización de **lenguaje natural** para las consultas.
- Utilización de un **modelo bi-encoder** para mapear las consultas y los datasets en un espacio vectorial común, lo que permite capturar mejor la semántica de las consultas en lenguaje natural.
- Creación de un **dataset especializado** anotado por expertos que contiene consultas realistas y específicas.

Métodos Propuestos

Métodos Propuestos

01

DataFinder Dataset

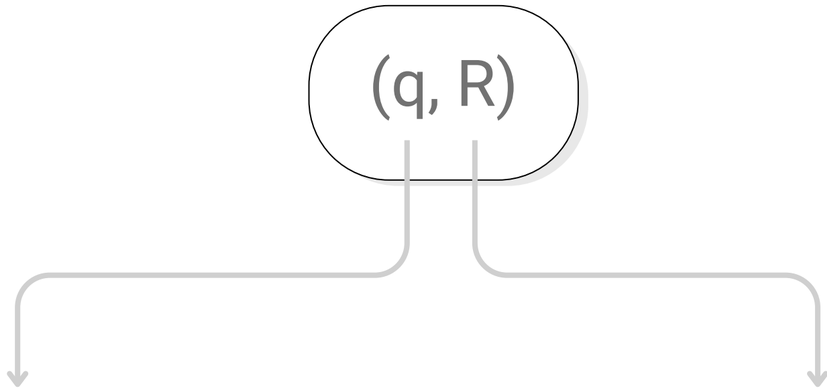
02

Neural bi-encoders

01 DataFinder dataset

- Contiene todos los datasets de **Papers With Code** (PWC)
- El proceso de su **construcción** se dividió en **2 partes**:
  - Colección de queries
  - Identificación de datasets relevantes (Rs)

Fomado por pares (q, R)

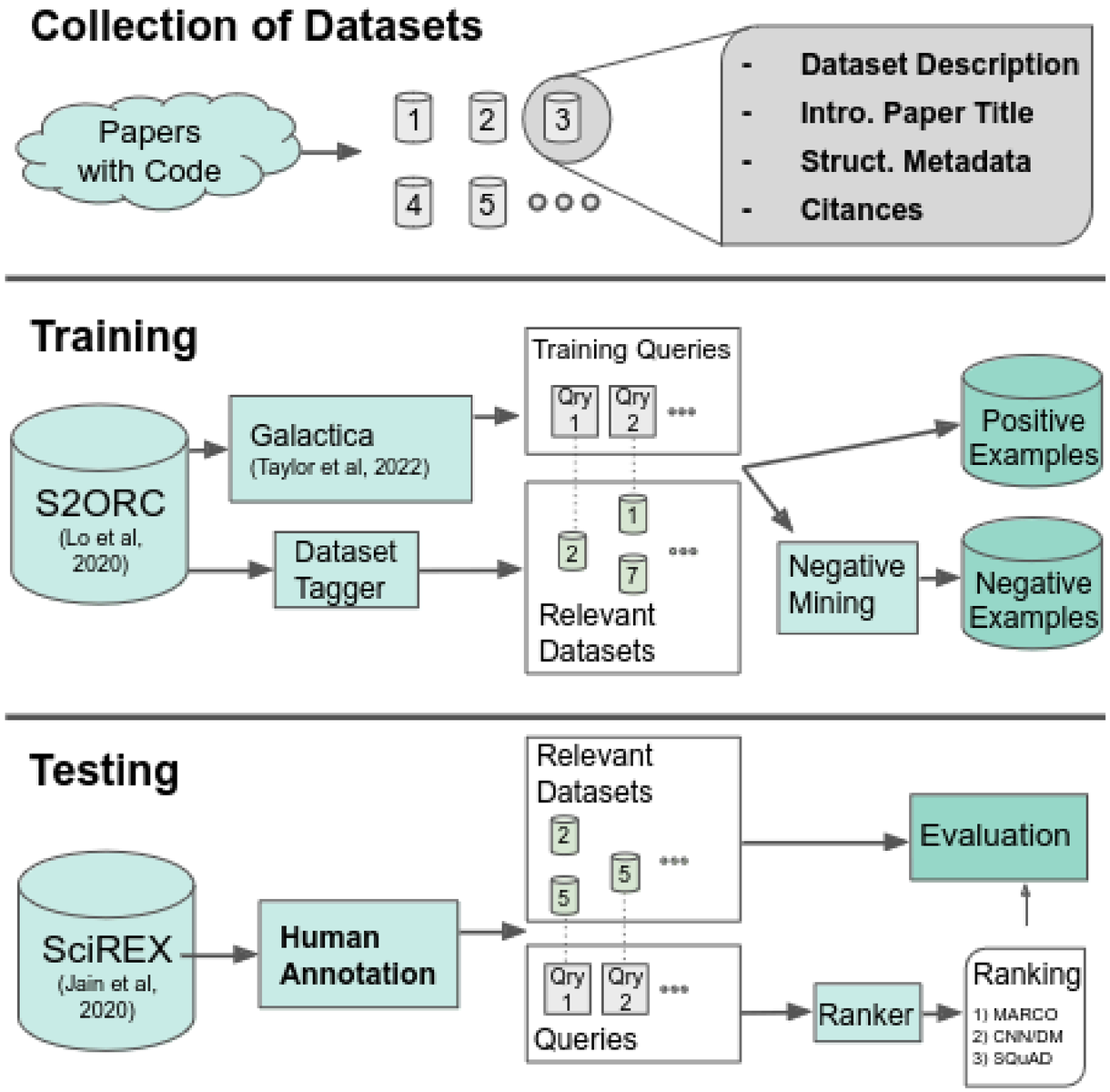


Query como **descripción completa** o por palabras clave (**Keyword**)

Dataset Relevante para q

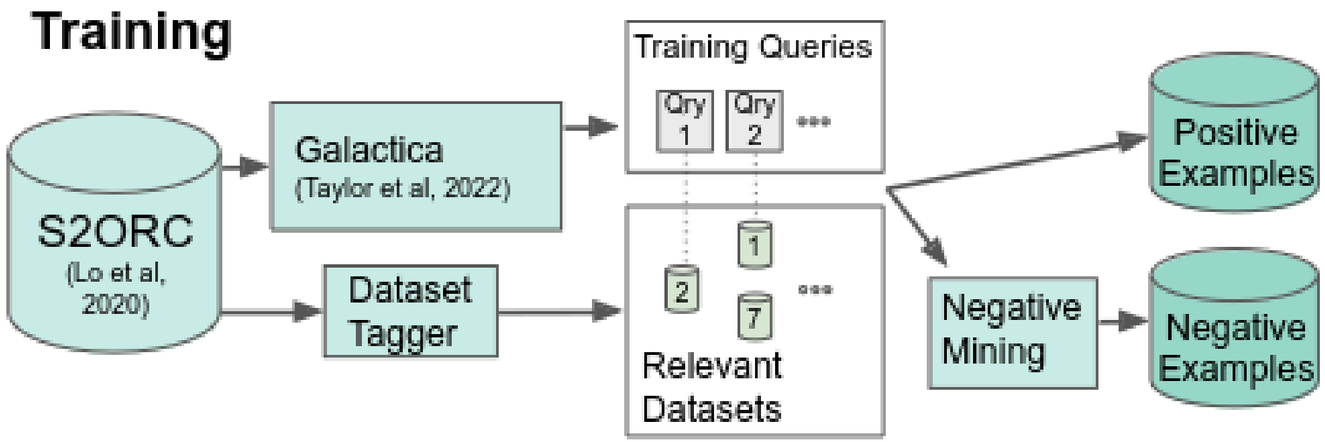
01 DataFinder dataset

- **Training Set** generado **automáticamente** con 17.495 queries.
- **Testing Set** anotado manualmente por **expertos** con 392 queries.



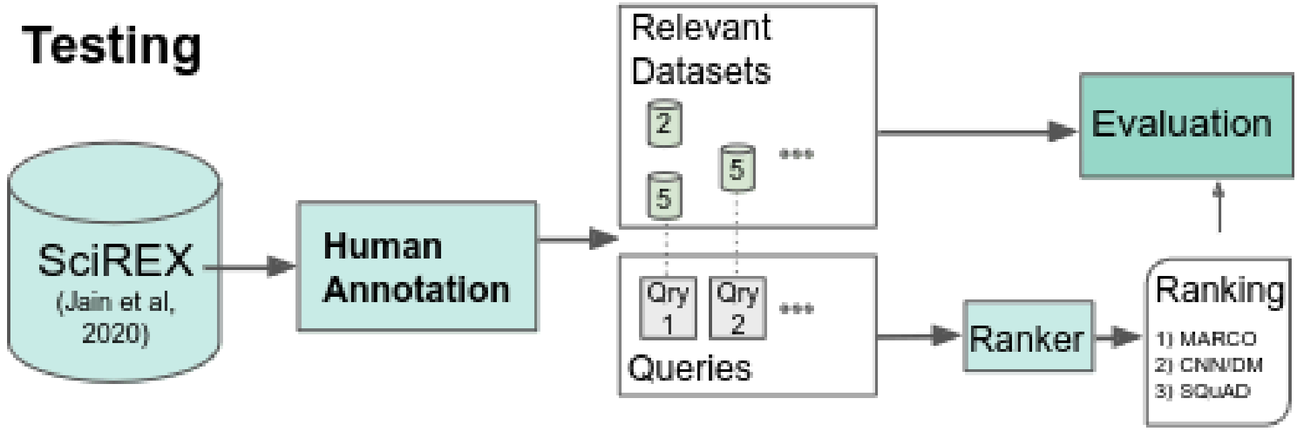
01 DataFinder dataset

- Para obtener las queries y los datasets relevantes se utilizan métodos de **few shot learning** y **rule based** sobre los papers del corpus S2ORC:



  - Obtienen **5 keywords** de los **abstracts** de cada paper con **Galactica** y se generan las **queries**.
  - Un **dataset** se selecciona como relevante si es **mencionado** 2 o más veces en el **body** del paper.

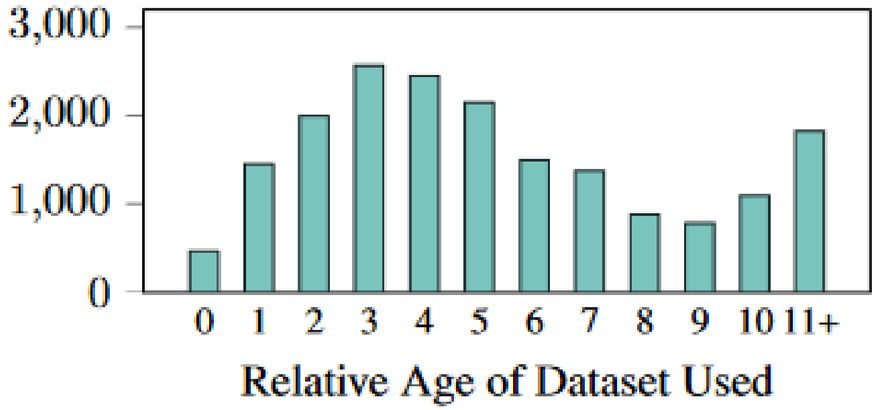
01 DataFinder dataset



- Se utilizaron papers de SciREX, anotados **manualmente**.
  - **Académicos** y estudiantes de **doctorados** formularon queries a partir de **abstracts**.
  - El **dataset** relevante para un paper es información ya **incluida** por SciRex.

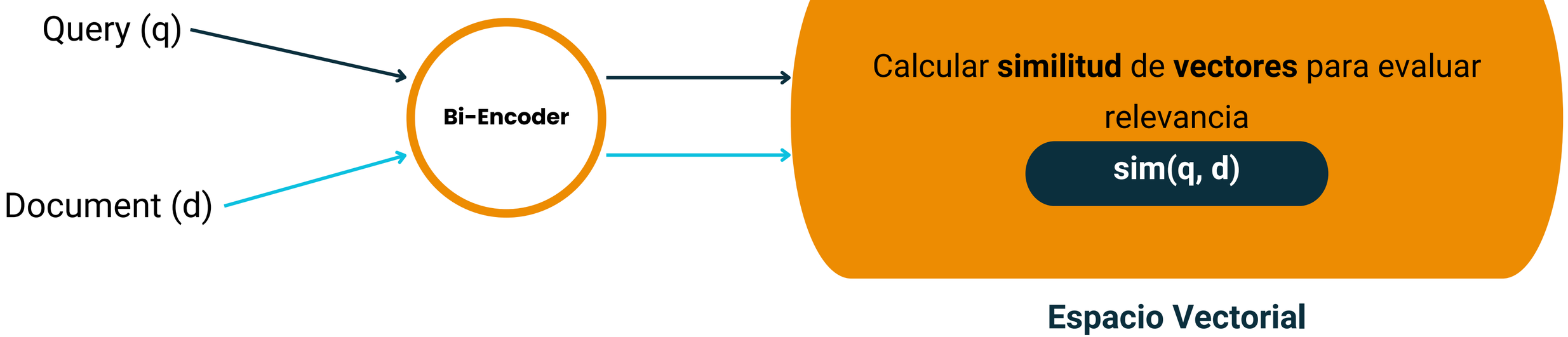
01 DataFinder dataset Análisis

- En **training** set se asigna **un dataset** a una query en promedio. En **testing** set se asignan **3 datasets** en promedio.
- Más de un **50%** de los papers del training set es uno de los **top-5** más **populares** en su área.



- Se evidencia una **tendencia** a utilizar datasets más **recientes** o más **tradicionales**, sugiriendo la necesidad de una recomendación de este tipo.

02 Neural bi-encoders



02 Neural bi-encoders

$$\text{sim}(q, d) = \text{cls}(\text{BERT}(q))^T \text{cls}(\text{BERT}(d))$$

- Utiliza **SciBERT** como modelo de inicio.
- Codifica **q y d** por **separado** mediante codificación de **BERT**, representando los documentos como su **token CLS (clasificación)**.
- El modelo inicial es **ajustado** durante el entrenamiento para **minimizar perdida contrastiva** y seleccionar **hard negatives** con BM25.

02 Evaluación de Neural bi-encoders

Evaluar metodos populares para tarea de ranking

- Term-Based method: BM25 retriever
- Nearest-Neighbor retrieval

Evaluar motores de busqueda

- PWC
- Google Dataset Search (reducido a papers incluidos en PWC)

Evaluación y resultados

0 Filtro basado en el tiempo

Coherencia temporal en las recomendaciones.

- Las **consultas de testeo** provienen de papers introducidos entre **2012 y 2020.**
- **La mitad** de los *datasets* del conjunto de entrenamiento fueron introducidos desde el año 2018 o posterior.

Ranking de datasets filtrado

$$D' = \{d \in D \,|\, year\,(d) \leqslant year\,(q)\}$$

01 Comparacion de técnicas de recuperacion de información

Model	P@5	R@5	MAP	MRR
Full-Sentence Queries				
BM25	4.7 ±0.1	11.6 ±1.7	8.0 ± 1.3	14.5 ±2.0
kNN (TF-IDF)	5.5 ±0.6	12.3 ±1.6	7.8 ±1.1	15.5 ±2.0
kNN (BERT)	7.1 ±0.7	14.2 ±1.5	9.7 ±1.2	21.3 ±2.3
Bi-Encoder	**16.0** ±1.1	**31.2** ±2.2	**23.4** ±1.9	**42.6** ±2.7

Full-Sentence Queries

Model	P@5	R@5	MAP	MRR
Keyphrase Queries				
BM25	6.6 ±0.5	15.3 ±1.1	11.4 ±0.8	19.9 ±1.5
kNN (TF-IDF)	2.7 ±0.4	5.9 ±1.1	3.3 ±0.7	8.2 ±1.6
kNN (BERT)	2.8 ±0.4	5.8 ±1.1	3.3 ±1.1	7.3 ±1.3
Bi-Encoder	**16.5** ±1.0	**32.4** ±2.2	**23.3** ±1.8	**42.3** ±2.6

Keyprase Queries

01 Comparacion de técnicas de recuperacion de información

Model	P@5	R@5	MAP	MRR
Full-Sentence Queries				
BM25	4.7 ±0.1	11.6 ±1.7	8.0 ± 1.3	14.5 ±2.0
kNN (TF-IDF)	5.5 ±0.6	12.3 ±1.6	7.8 ±1.1	15.5 ±2.0
kNN (BERT)	7.1 ±0.7	14.2 ±1.5	9.7 ±1.2	21.3 ±2.3
Bi-Encoder	**16.0** ±1.1	**31.2** ±2.2	**23.4** ±1.9	**42.6** ±2.7

Model	P@5	R@5	MAP	MRR
Keyphrase Queries				
BM25	6.6 ±0.5	15.3 ±1.1	11.4 ±0.8	19.9 ±1.5
kNN (TF-IDF)	2.7 ±0.4	5.9 ±1.1	3.3 ±0.7	8.2 ±1.6
kNN (BERT)	2.8 ±0.4	5.8 ±1.1	3.3 ±1.1	7.3 ±1.3
Bi-Encoder	**16.5** ±1.0	**32.4** ±2.2	**23.3** ±1.8	**42.3** ±2.6

- ***Neural Bi-encoder retrieval*** presenta el **mejor** rendimiento
- ***Term-based retrieval (BM25)*** tiene un **bajo** rendimiento.
- **Term-based kNN** no es efectiva.

01 Comparacion de técnicas de recuperacion de información

Model	P@5	R@5	MAP	MRR
Full-Sentence Queries				
BM25	4.7 ±0.1	11.6 ±1.7	8.0 ± 1.3	14.5 ±2.0
kNN (TF-IDF)	5.5 ±0.6	12.3 ±1.6	7.8 ±1.1	15.5 ±2.0
kNN (BERT)	7.1 ±0.7	14.2 ±1.5	9.7 ±1.2	21.3 ±2.3
Bi-Encoder	**16.0** ±1.1	**31.2** ±2.2	**23.4** ±1.9	**42.6** ±2.7

Model	P@5	R@5	MAP	MRR
Keyphrase Queries				
BM25	6.6 ±0.5	15.3 ±1.1	11.4 ±0.8	19.9 ±1.5
kNN (TF-IDF)	2.7 ±0.4	5.9 ±1.1	3.3 ±0.7	8.2 ±1.6
kNN (BERT)	2.8 ±0.4	5.8 ±1.1	3.3 ±1.1	7.3 ±1.3
Bi-Encoder	**16.5** ±1.0	**32.4** ±2.2	**23.3** ±1.8	**42.3** ±2.6

- ***Neural Bi-encoder retrieval*** presenta el **mejor** rendimiento
- ***Term-based retrieval (BM25)*** tiene un **bajo** rendimiento.
- **Term-based kNN** no es efectiva.

01 Comparacion de técnicas de recuperacion de información

Model	P@5	R@5	MAP	MRR
Full-Sentence Queries				
BM25	4.7 ±0.1	11.6 ±1.7	8.0 ± 1.3	14.5 ±2.0
kNN (TF-IDF)	5.5 ±0.6	12.3 ±1.6	7.8 ±1.1	15.5 ±2.0
kNN (BERT)	7.1 ±0.7	14.2 ±1.5	9.7 ±1.2	21.3 ±2.3
Bi-Encoder	**16.0** ±1.1	**31.2** ±2.2	**23.4** ±1.9	**42.6** ±2.7

Model	P@5	R@5	MAP	MRR
Keyphrase Queries				
BM25	6.6 ±0.5	15.3 ±1.1	11.4 ±0.8	19.9 ±1.5
kNN (TF-IDF)	2.7 ±0.4	5.9 ±1.1	3.3 ±0.7	8.2 ±1.6
kNN (BERT)	2.8 ±0.4	5.8 ±1.1	3.3 ±1.1	7.3 ±1.3
Bi-Encoder	**16.5** ±1.0	**32.4** ±2.2	**23.3** ±1.8	**42.3** ±2.6

- ***Neural Bi-encoder retrieval*** presenta el **mejor** rendimiento
- ***Term-based retrieval (BM25)*** tiene un **bajo** rendimiento.
- **Term-based kNN** no es efectiva.

01 Comparacion de técnicas de recuperacion de información

Model	P@5	R@5	MAP	MRR
Full-Sentence Queries				
BM25	4.7 ±0.1	11.6 ±1.7	8.0 ± 1.3	14.5 ±2.0
kNN (TF-IDF)	5.5 ±0.6	12.3 ±1.6	7.8 ±1.1	15.5 ±2.0
kNN (BERT)	7.1 ±0.7	14.2 ±1.5	9.7 ±1.2	21.3 ±2.3
Bi-Encoder	**16.0** ±1.1	**31.2** ±2.2	**23.4** ±1.9	**42.6** ±2.7

Model	P@5	R@5	MAP	MRR
Keyphrase Queries				
BM25	6.6 ±0.5	15.3 ±1.1	11.4 ±0.8	19.9 ±1.5
kNN (TF-IDF)	2.7 ±0.4	5.9 ±1.1	3.3 ±0.7	8.2 ±1.6
kNN (BERT)	2.8 ±0.4	5.8 ±1.1	3.3 ±1.1	7.3 ±1.3
Bi-Encoder	**16.5** ±1.0	**32.4** ±2.2	**23.3** ±1.8	**42.3** ±2.6

- ***Neural Bi-encoder retrieval*** presenta el **mejor** rendimiento
- ***Term-based retrieval (BM25)*** tiene un **bajo** rendimiento.
- **Term-based kNN** no es efectiva.

01 Comparacion de técnicas de recuperacion de información

Model	P@5	R@5	MAP	MRR
Full-Sentence Queries				
BM25	4.7 ±0.1	11.6 ±1.7	8.0 ± 1.3	14.5 ±2.0
kNN (TF-IDF)	5.5 ±0.6	12.3 ±1.6	7.8 ±1.1	15.5 ±2.0
kNN (BERT)	7.1 ±0.7	14.2 ±1.5	9.7 ±1.2	21.3 ±2.3
Bi-Encoder	**16.0** ±1.1	**31.2** ±2.2	**23.4** ±1.9	**42.6** ±2.7

Model	P@5	R@5	MAP	MRR
Keyphrase Queries				
BM25	6.6 ±0.5	15.3 ±1.1	11.4 ±0.8	19.9 ±1.5
kNN (TF-IDF)	2.7 ±0.4	5.9 ±1.1	3.3 ±0.7	8.2 ±1.6
kNN (BERT)	2.8 ±0.4	5.8 ±1.1	3.3 ±1.1	7.3 ±1.3
Bi-Encoder	**16.5** ±1.0	**32.4** ±2.2	**23.3** ±1.8	**42.3** ±2.6

- ***Neural Bi-encoder retrieval*** presenta el **mejor** rendimiento
- ***Term-based retrieval (BM25)*** tiene un **bajo** rendimiento.
- **Term-based kNN** no es efectiva.

02 DataFinder vs *Commercial Search Engines*

PapersWithCode (PwC)

Google Dataset Search

Datafinder

Model	P@5	R@5	MAP	MRR
PwC (*descriptions*)	0.6	1.7	0.9	1.2
PwC (*keywords*)	3.5	10.0	6.5	9.1
Google (*descriptions*)	0.1	0.1	0.1	0.3
Google (*keywords*)	9.7	19.5	12.3	24.0
Ours (*descriptions*)	16.0	31.2	23.4	42.6
Ours (*keywords*)	16.5	32.4	23.3	42.3

02 DataFinder vs *Commercial Search Engines*

PapersWithCode (PwC)

Google Dataset Search

Datafinder

Model	P@5	R@5	MAP	MRR
PwC (*descriptions*)	0.6	1.7	0.9	1.2
PwC (*keywords*)	3.5	10.0	6.5	9.1
Google (*descriptions*)	0.1	0.1	0.1	0.3
Google (*keywords*)	9.7	19.5	12.3	24.0
Ours (*descriptions*)	16.0	31.2	23.4	42.6
Ours (*keywords*)	16.5	32.4	23.3	42.3

02 DataFinder vs *Commercial Search Engines*

PapersWithCode (PwC)

Google Dataset Search

Datafinder

Model	P@5	R@5	MAP	MRR
PwC (*descriptions*)	0.6	1.7	0.9	1.2
PwC (*keywords*)	3.5	10.0	6.5	9.1
Google (*descriptions*)	0.1	0.1	0.1	0.3
Google (*keywords*)	9.7	19.5	12.3	24.0
Ours (*descriptions*)	16.0	31.2	23.4	42.6
Ours (*keywords*)	16.5	32.4	23.3	42.3

03 Analisis cualitativo de los resultados

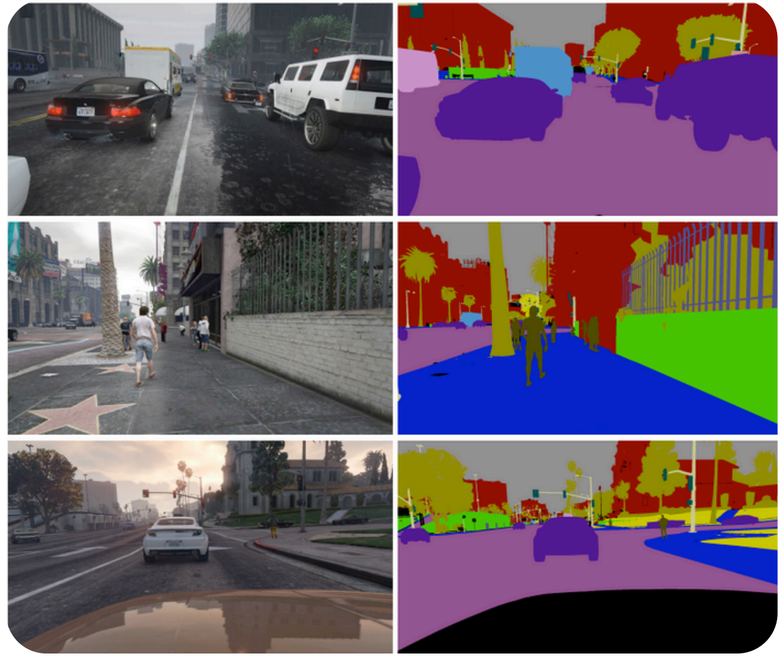
Full-Sentence Query: I want to use adversarial learning to perform domain adaptation for semantic segmentation of images.
Keyword Query: semantic segmentation domain adaptation images

Actual		Google	PWC	Ours
Cityscapes	1	LoveDA	VQA	Cityscapes
GTA5	2	Office-31	RTE	GTA5
SYNTHIA	3	Dark Zurich	VQA 2.0	SYNTHIA

03 Analisis cualitativo de los resultados

Full-Sentence Query: I want to use adversarial learning to perform domain adaptation for semantic segmentation of images.
Keyword Query: semantic segmentation domain adaptation images

Actual		Google	PWC	Ours
Cityscapes	1	LoveDA	VQA	Cityscapes
GTA5	2	Office-31	RTE	GTA5
SYNTHIA	3	Dark Zurich	VQA 2.0	SYNTHIA



Cityscapes dataset. Fuente: Richter, et al.

Cityscapes dataset. Fuente: https://www.cityscapes-dataset.com/

SYNTHIA dataset. Fuente: https://synthia-dataset.net/.

03 Analisis cualitativo de los resultados

Full-Sentence Query: I want to use adversarial learning to perform domain adaptation for semantic segmentation of images.
Keyword Query: semantic segmentation domain adaptation images

Actual		Google	PWC	Ours
Cityscapes	1	LoveDA	VQA	Cityscapes
GTA5	2	Office-31	RTE	GTA5
SYNTHIA	3	Dark Zurich	VQA 2.0	SYNTHIA

LoveDA dataset. Fuente: DatasetNinja.com

VQA dataset. Fuente: https://visualqa.org/

03 Analisis cualitativo de los resultados

Full-Sentence Query: A new deep neural network architecture for machine translation using depthwise separable convolutions.
Keyword Query: machine translation text

Actual		Google	PWC	Ours
WMT 2014	1	WMT 2014	Machine Number Sense	SQuAD
	2		UCI Datasets	WikiText-2
	3		Affective Text	WikiText-103

03 Analisis cualitativo de los resultados

Full-Sentence Query: A new deep neural network architecture for machine translation using depthwise separable convolutions.
Keyword Query: machine translation text

Actual		Google	PWC	Ours
WMT 2014	1	WMT 2014	Machine Number Sense	SQuAD
	2		UCI Datasets	WikiText-2
	3		Affective Text	WikiText-103

03 Analisis cualitativo de los resultados

Full-Sentence Query: A new deep neural network architecture for machine translation using depthwise separable convolutions.
Keyword Query: machine translation text

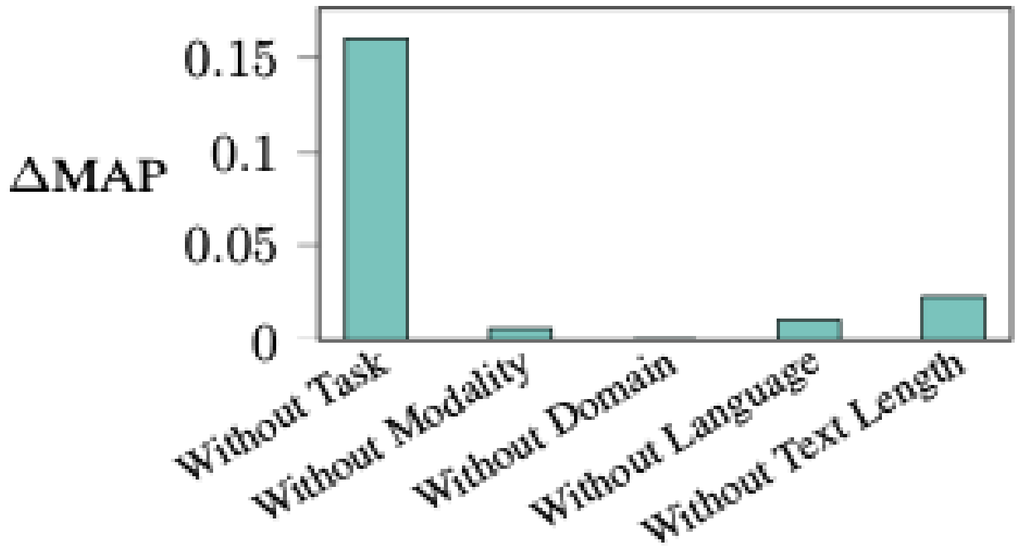
Actual		Google	PWC	Ours
WMT 2014	1	WMT 2014	Machine Number Sense	SQuAD
	2		UCI Datasets	WikiText-2
	3		Affective Text	WikiText-103

04 Factores claves para consultas exitosas

¿Que información de las consultas es **la más importante** para la recomendacion de datos?

¿Como **afecta** la información de las consultas en las **métricas de evaluación**?

04 Factores claves para consultas exitosas



Model	P@5	R@5	MAP	MRR
Full-Sentence Queries				
Description	15.3 ±1.0	30.0 ±2.1	23.0 ± 1.9	42.8 ±2.7
+ Struct. Info	**16.0** ±1.1	**31.2** ±2.2	**23.3** ±1.8	**42.4** ±2.7
+ Citances	15.8 ±1.1	30.8 ±2.2	23.1 ±1.9	42.2 ±2.7
Keyphrase Queries				
Description	13.1 ±1.0	25.6 ±2.0	17.4 ±1.6	33.1 ±2.5
+ Struct. Info	16.6 ±1.1	32.7 ±2.2	23.5 ±1.8	42.8 ±2.8
+ Citances	**16.8** ±1.0	**33.4** ±2.2	**23.6** ±1.8	**43.0** ±2.6

Conclusiones y Hallazgos

01 Conclusiones principales

DataFinder admite **busquedas robustas,** superando a otros motores de busquedas

Neural bi-encoders son **competitivos** para busquedas por **palabras claves** como para busquedas con **frases completas**

01 Conclusiones principales

DataFinder admite **busquedas robustas,** superando a otros motores de busquedas

Neural bi-encoders son **competitivos** para busquedas por **palabras claves** como para busquedas con **frases completas**

02 Hallazgos y limitaciones

1

Dependencia de *datasets* de Papers With Code (PwC)

02 Hallazgos y limitaciones

1

Dependencia de *datasets* de Papers With Code (PwC)

Datasets relevantes que no existen en PwC

02 Hallazgos y limitaciones

1

Dependencia de *datasets* de Papers With Code (PwC)

Datasets relevantes que no existen en PwC

PwC sesga el año de publicacion de los papers

02 Hallazgos y limitaciones

2

Sesgo hacia datasets **populares** en el **conjunto de prueba**

02 Hallazgos y limitaciones

2

Sesgo hacia datasets **populares** en el **conjunto de prueba**

SciREX

02 Hallazgos y limitaciones

2

Sesgo hacia datasets **populares** en el **conjunto de prueba**

SciREX

Numero promedio de citas de un paper en SciREX: ***129***

02 Hallazgos y limitaciones

3

Datos de entrenamiento **sesgados**

4

Consultas solo en **inglés**

Referencias

- [1] Michael Färber and Ann-Kathrin Leisinger. 2021. Recommending Datasets for Scientific Problem Descriptions. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM '21). Association for Computing Machinery, New York, NY, USA, 3014–3018. https://doi.org/10.1145/3459637.3482166

IIC3633

DATAFINDER: SCIENTIFIC DATASET RECOMMENDATION FROM NATURAL LANGUAGE DESCRIPTIONS

Pedro Z, Felipe T, Kahil R