University of Rijeka
Faculty of Informatics and Digital Technologies

Construction of a Knowledge Graph in the Climate Research Domain

Author: **Andrija Poleksić**

Mentor: prof. dr. sc. Sanda Martinčić-Ipšić

Croatian Science Foundation

AP is fully supported by Croatian Science Foundation under the project DOK-2021-02.

Motivation

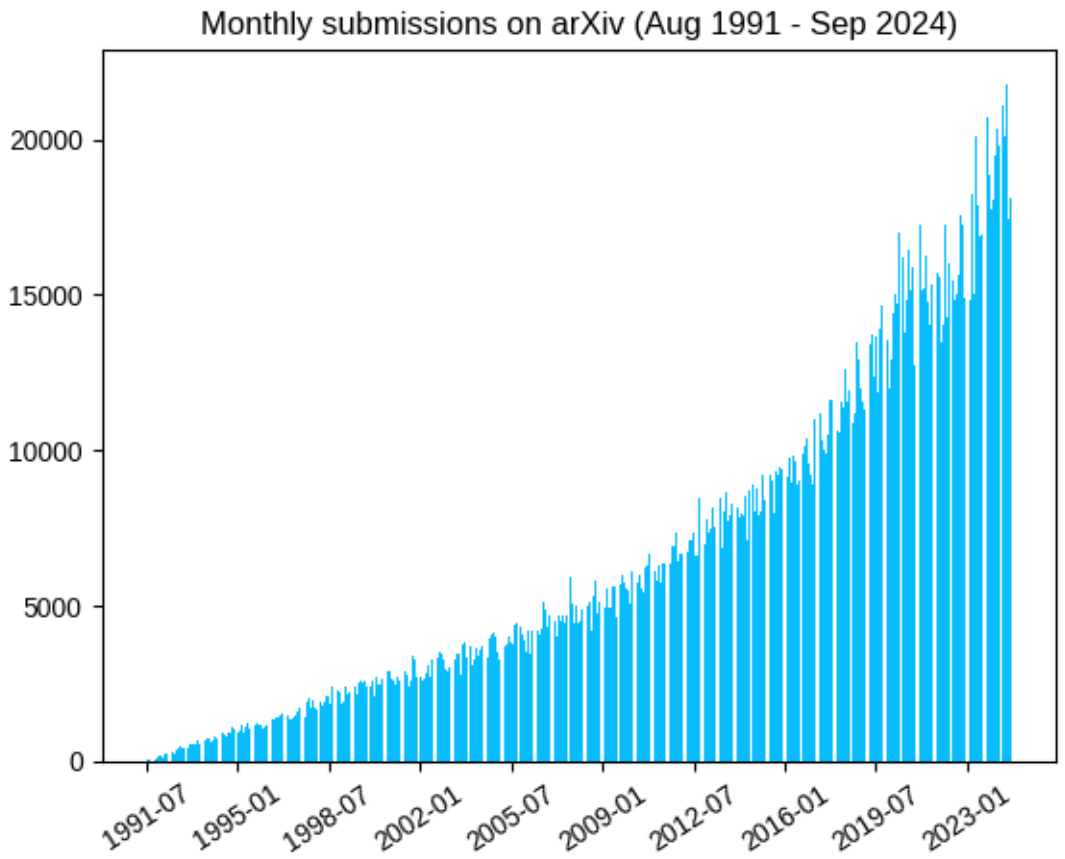
- Global warming and Climate change
  - "Global Warming of 1.5 ºC"[1]
  - profound effects on **global ecosystems**, **weather patterns**, **sea level**, and **human societies**,
  - threat to **planet's biodiversity** and **sustainable future**
- **Climate change denial**
  - Andre et al. [1] – up to **86 %** of individuals acknowledge human-induced Climate change
  - Stems from **misguided beliefs** and **vested corporate interest**
  - Areni [2] – deniers on Reddit rely on **alternative sources**

[1] https://www.ipcc.ch/sr15/

ArXiv - Number of Publications

- Data volume
  - Ever-increasing
  - Information deluge

- **Information extraction** (IE)
  - The task of **automatically** extracting **structured** information from unstructured and/or semi-structured machine-readable documents



Outline

- **Domain corpus**
  - Scientific papers
- **Domain specific language models**
  - Language model pretraining
  - Domain Adaptive pretraining
- **Method for Relation Extraction** using BERT-like architectures
  - Domain specific NER model
  - Domain specific RE model
- **Relation Extraction evaluation**
  - Hand annotated golden dataset
- **Climate Research Knowledge Graph (CliReKG)**
- ...

Data Collection

- Data sources for scientific journals on Climate Change:
  - Scimago Journal & Country Rank (SJR)[1]
  - ScienceWatch Rank[2]
  - MDPI journals
- Total of 29 journals → ~ **194,000** retrieved research papers
  - 77% HTML format and 23% PDF format

[1] https://www.scimagojr.com/
[2] https://clarivate.com/?lid=sciencewatch

Data Info

Journal name	#	Journal name	#	Journal name	#
International Journal of Climatology	3,825	Ecological Applications	4,469	Ecosystem Health and Sustainability	831
Energy Policy	1,023	Journal of Climate	15,325	Climate Dynamics	3,943
Global Change Biology	7,103	Journal of Geophysical Research: Atmospheres	14,512	NPJ Climate and Atmospheric Science	355
NPJ Ocean Sustainability	12	NPJ Climate Action	39	Nature Climate Change	387
Nature Geoscience	560	PNAS	88,534	MDPI water	21,768
MDPI Air	18	MDPI Atmosphere	8.705	MDPI Climate	1,232
MDPI Earth	184	MDPI Ecologies	115	MDPI Energies	8,236
MDPI Hidrology	988	MDPI Forests	10,674	MDPI Fuels	104
MDPI Environments	1,012	MDPI Meteorology	57	MDPI Sustainable Chemistry	116
MDPI Recycling	420	MDPI Oceans	126	**Total**	**185,977**

(Pre)training Data Comparison

Model	Data used	CS	A\#S
BERT [3]	BooksCorpus (800M words) and English Wikipedia (2,500M words)	3.30B	/
SciBERT [4]	Random sample of 1.14M papers from Semantic Scholar	3.17B	154
ClimateBERT [5]	Climate related news articles, climate-related papers abstracts and corporate climate and sustainability reports	0.22B[1]	/
OUR [6]	~ 200,000 climate-related research papers	1.25B[2]	242[3]

[1] Calculated from reported average number of words [6]
[2] Approximation from tokenizer trained od 10,000 papers sample according to The Tokenization pipeline - https://huggingface.co/docs/tokenizers/python/latest/pipeline.html
[3] Approximation from SegtokSentenceSplitter - https://github.com/flairNLP/flair/blob/master/flair/splitter.py

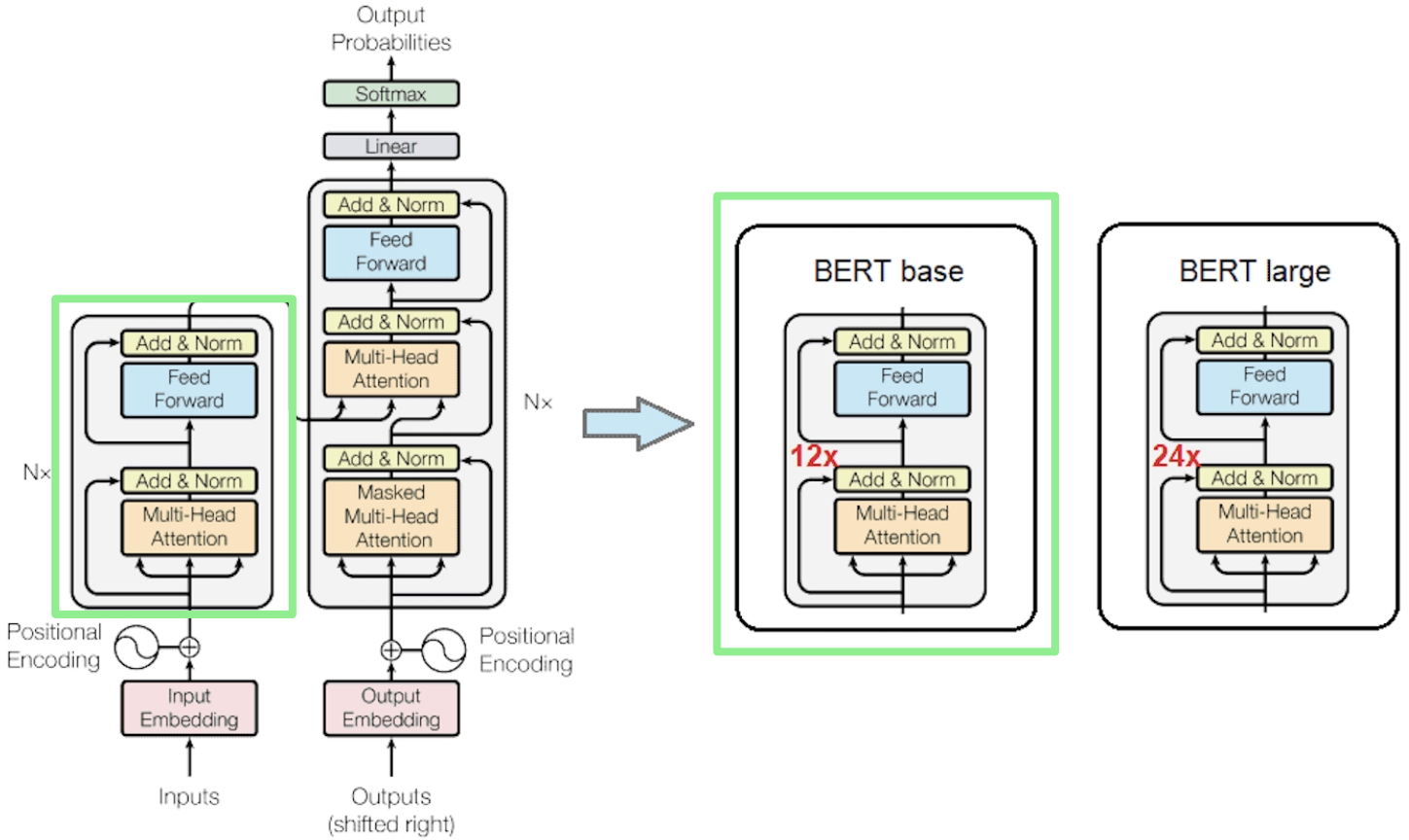
Outline

- **Domain corpus**
  - Scientific papers
- **Domain specific language models**
  - Language model pretraining
  - Domain Adaptive pretraining
- **Method for Relation Extraction** using BERT-like architectures
  - Domain specific NER model
  - Domain specific RE model
- **Relation Extraction evaluation**
  - Hand annotated golden dataset
- **Climate Research Knowledge Graph (CliReKG)**
- ...

(Domain Adaptive) Language Model Pretraining

- LM Pretraining – LM training on a **vast amount of data** in an **self-supervised manner** to learn useful representations of (textual) domain.
  - Usually followed by fine-tuning – Training a model on a downstream task.
- Domain Adaptive Pretraining – Continued pretraining of a LM on a domain specific corpora. [7]
- Vocabulary augmentation:
  - Adding additional vocabulary words [5]
  - Adding additional vocabulary words as an existing token combinations [8]

BERT [3] and (Distil)RoBERTa [9, 10]



[1]

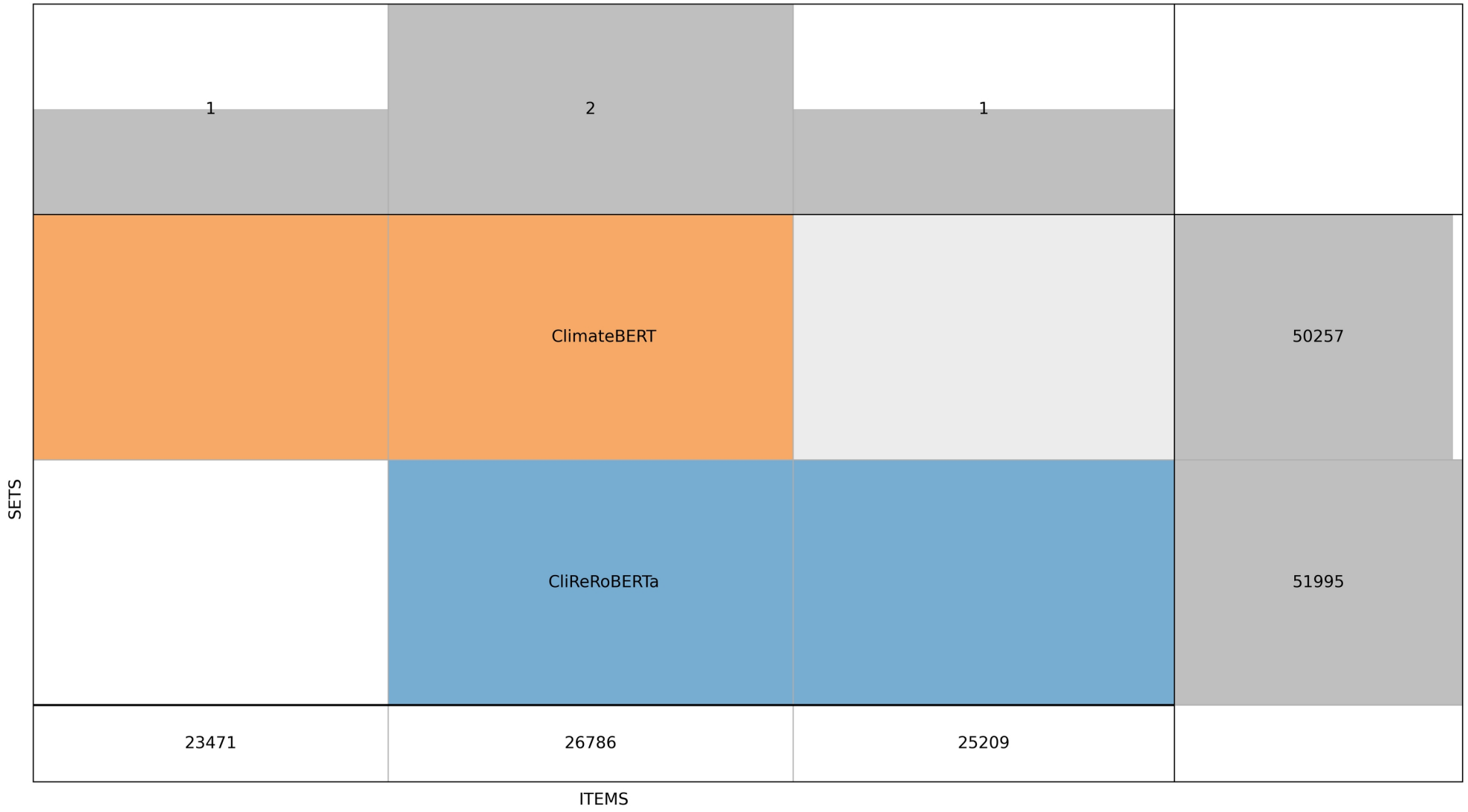
- BERT and RoBERTa follow original architecture implementation from Vaswani et al. [11]

[1] https://sushant-kumar.com/blog/bert

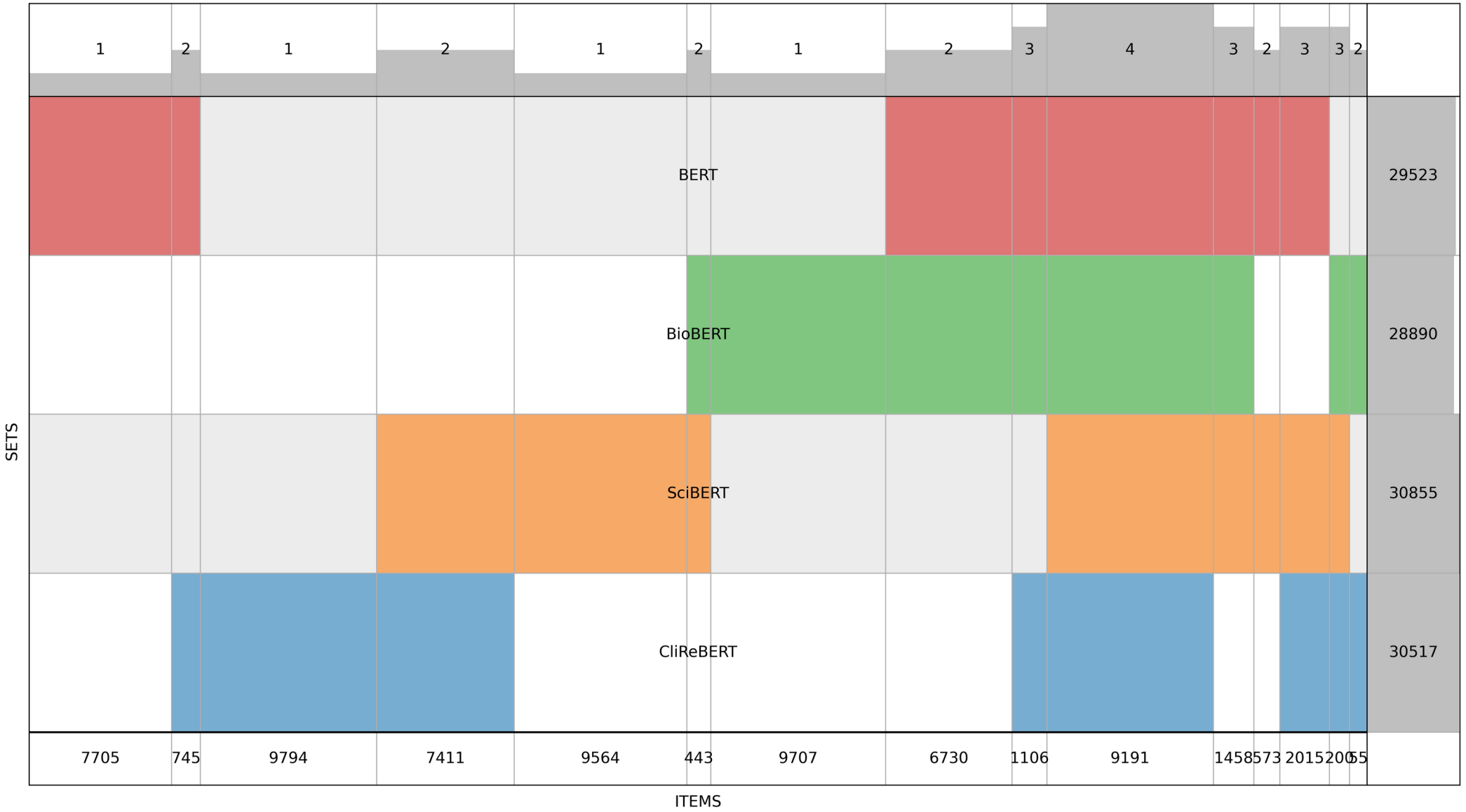
"Why another BERT"?

- BioBERT [12]
- ClinicalBERT [13]
- SciBERT [4]
- LegalBERT [14]
- JuriBERT [15]
- ClimateBERT [5]
- PharmBERT [16]
- ...

ClimateBERT vs OURS



BERT vs BioBERT vs SciBERT vs OURS



DAPT?

	Year	Vocabulary	DAPT/From scratch
BioBERT	2019	Original	DAPT
ClinicalBERT	2019	Original	DAPT
SciBERT	2019	New	From scratch
LegalBERT	2020	Original / New (equal size)	From scratch / DAPT
JuriBERT	2021	New	From scratch
ClimateBERT	2022	Augmented	DAPT
PharmBERT	2023	Original	DAPT

BERT and RoBERTa 2

	BERT	RoBERTa
Parameters	Base: 110M	Base: 125M
Layers / Hidden Dimensions / Self-Attenton Heads	Base: 12 / 768 / 12	Base: 12 / 768 / 12
Pretraining data	BooksCorpus + English Wikipedia = **16 GB**	BERT + CCNews + OpenWebText + Stories = **160 GB**
Method	MLM & NSP	Dynamic MLM ~~& NSP~~
Tokenizer	WordPiece	Byte-level BPE

Trained BERT-like Models

- Domain adaptive pretraining on SciBERT model → **CliSciBERT model** ✓
- Domain adaptive pretraining on ClimateBERT (distill RoBERTa) model → **SciClimateBERT** ✓
- Climate change research model from scratch (BERT) → **CliRe(search)BERT** ✓
- Climate change research model from scratch (distill RoBERTa) model → **CliReRoBERTa** ✓
- SpanBERT [36], DeBERTa [37], ELECTRA [38], … ?

Hardware and Energy

Parameter	Value
GPU	Nvidia Quadro RTX 6000
CPU	AMD Ryzen Threadripper 3960X 24-Core Processor
Power GPU	0.26 kW
Power CPU	0.013 kW
Total Power (**TP**)	0.273 kW
Location	Rijeka, Croatia
Energy Mix (**EM**)	224.71 gCO_2eq/kWh

- Total CO2 Emission (**TCE**) = **TP** * TIME(h) * **EM**

Model Reports

- **CliSciBERT model**
  - Training time: 463h ~ **19 days**
  - Energy report: TP * 463h * EM = 28,403.12 g CO_2 ~ **28kg CO_2 emitted**
- **SciClimateBERT**
  - Training time: 300h ~ **12.5 days**
  - Energy report: TP * 300h * EM = 18,403.75 g CO_2 ~ **18kg CO_2 emitted**
- **CliReBERT**
  - Training time: 463h ~ **19 days**
  - Energy report: TP * 463h * EM = 28,403.12 g CO_2 ~ **28kg CO_2 emitted**
- **CliReRoBERTa**
  - Training time: 300h ~ **12.5 days**
  - Energy report: TP * 300h * EM = 18,403.75 g CO_2 ~ **18kg CO_2 emitted**

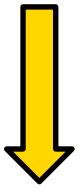
Outline

- **Domain corpus**
  - Scientific papers
- **Domain specific language models**
  - Language model pretraining
  - Domain Adaptive pretraining
- **Method for Relation Extraction** using BERT-like architectures
  - Domain specific NER model
  - Domain specific RE model
- **Relation Extraction evaluation**
  - Hand annotated golden dataset
- **Climate Research Knowledge Graph (CliReKG)**
- ...

Relation Extraction (RE)

- **Relation extraction** (RE) is the subtask of Information extraction (IE) consists of identifying relations between entities in each sentence, paragraph, or larger unit of text.

"***El Niño–Southern Oscillation (ENSO)*** *is another important factor for* ***winter temperature in China***."



(**ENSO**, *affects*, **winter temperature in China**)

Relation Extraction (RE)

- **Relation extraction** (RE) is the subtask of Information extraction (IE) consists of identifying relations between entities in each sentence, paragraph, or larger unit of text.

- Variable number of entities: **Two**
- **Defined (finite)** or undefined set of relations
- **Marked** or unmarked entities
- At **sentence**, few sentence (bag) or document level

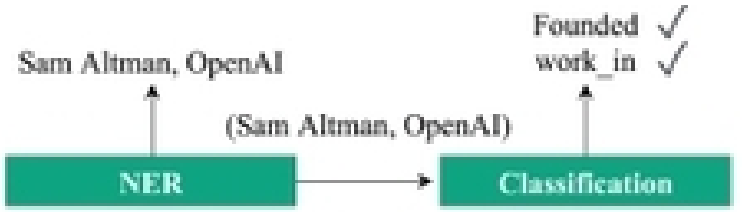
- **Relation Extraction as multi-class sentence level classification.**

Relation Classification

- **Pipeline based approach – NER + Relation classification**
  - NER and RE tasks are **trained separately**, therefore the RE model expects **already extracted entities** in the input text → may be of lower quality, propagating the error
  - **IDEA**: "*Dissect the problem and revisit the paradigm when it works!*"
    - *Relation Extraction: Perspective from Convolutional Neural Networks (2015.) [19]*
    - *Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification (2016.) [20]*
    - *Improving Relation Extraction by Pre-trained Language Representations (2019.) [21]*
    - ***Matching the Blanks: Distributional Similarity for Relation Learning (2019.)** [22]*

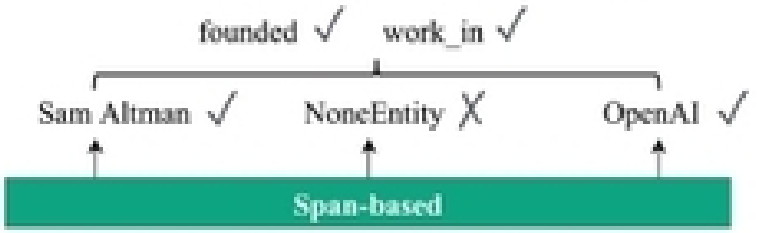
Relation extraction with deep learning methods (2023.) [23]

Joint Extraction Approaches



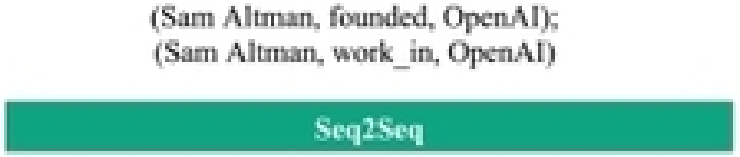
Sam Altman is the co-founder and CEO of OpenAI.

(a) The pipeline approach



Sam Altman is the co-founder and CEO of OpenAI.

(b) The span-based approach

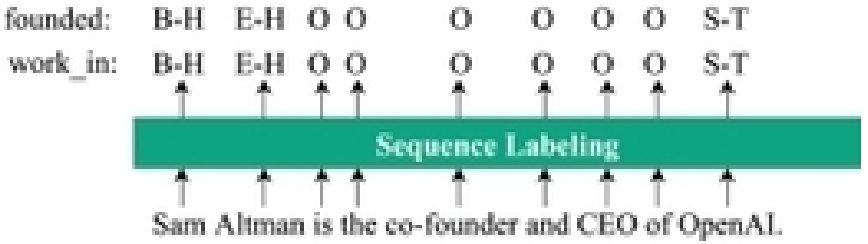


Sam Altman is the co-founder and CEO of OpenAI.

(c) The Seq2Seq approach

Sam Altman is the co-founder and CEO of OpenAI.

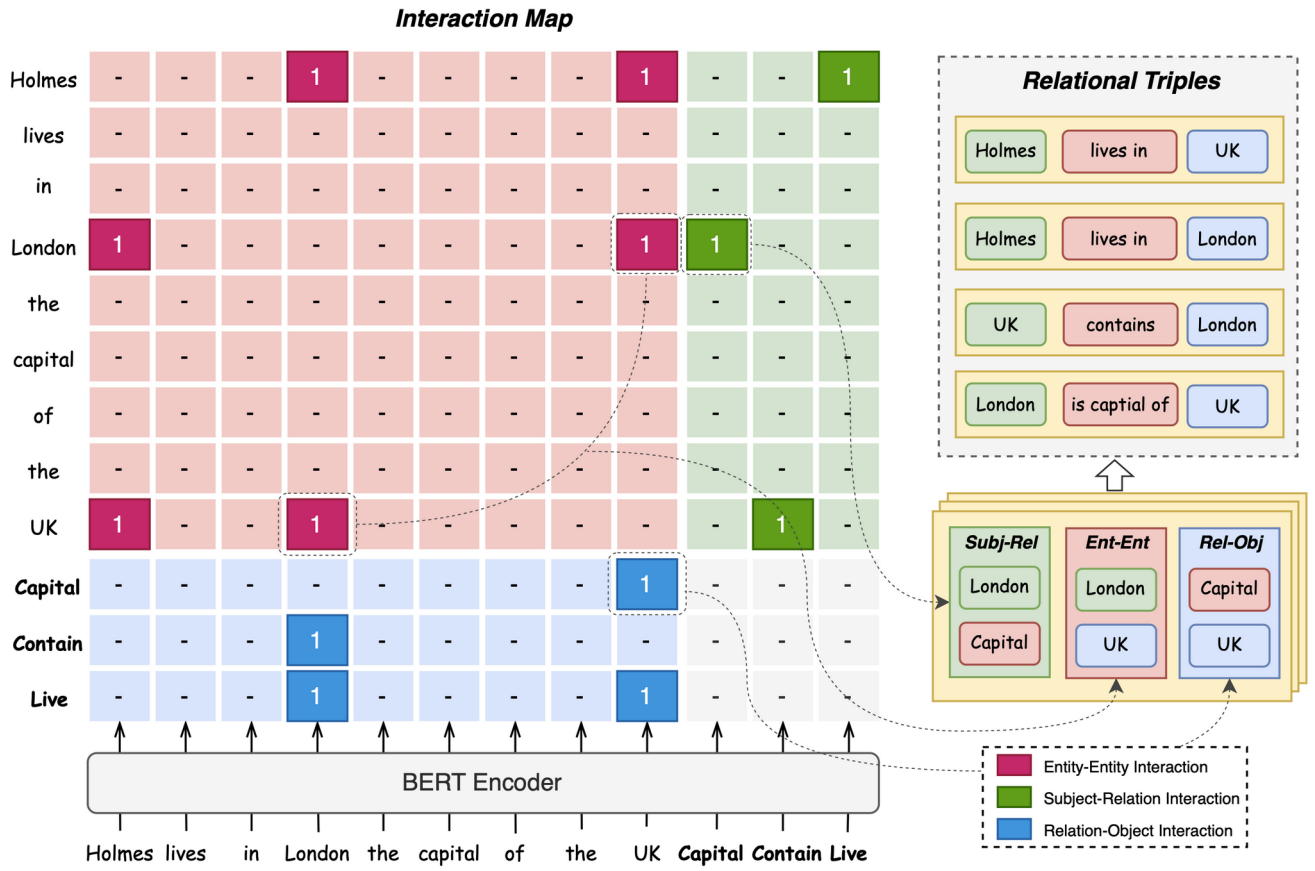
(d) The MRC-based approach



(e) The sequence labeling approach

A comprehensive survey on relation extraction: Recent advances and new frontiers (2024.) [24]

Span-based Approach – "UniRel"

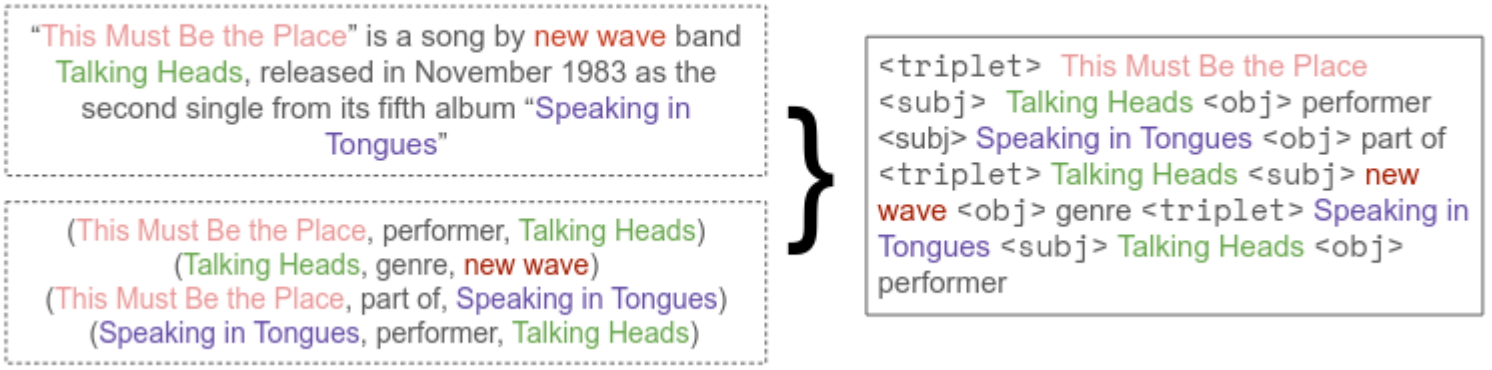


- Process each text into spans – perform classification on spans
- Usually utilize pretrained Transformer encoders

UniRel: Unified representation and interaction for joint relational triple extraction (2022.) [25]

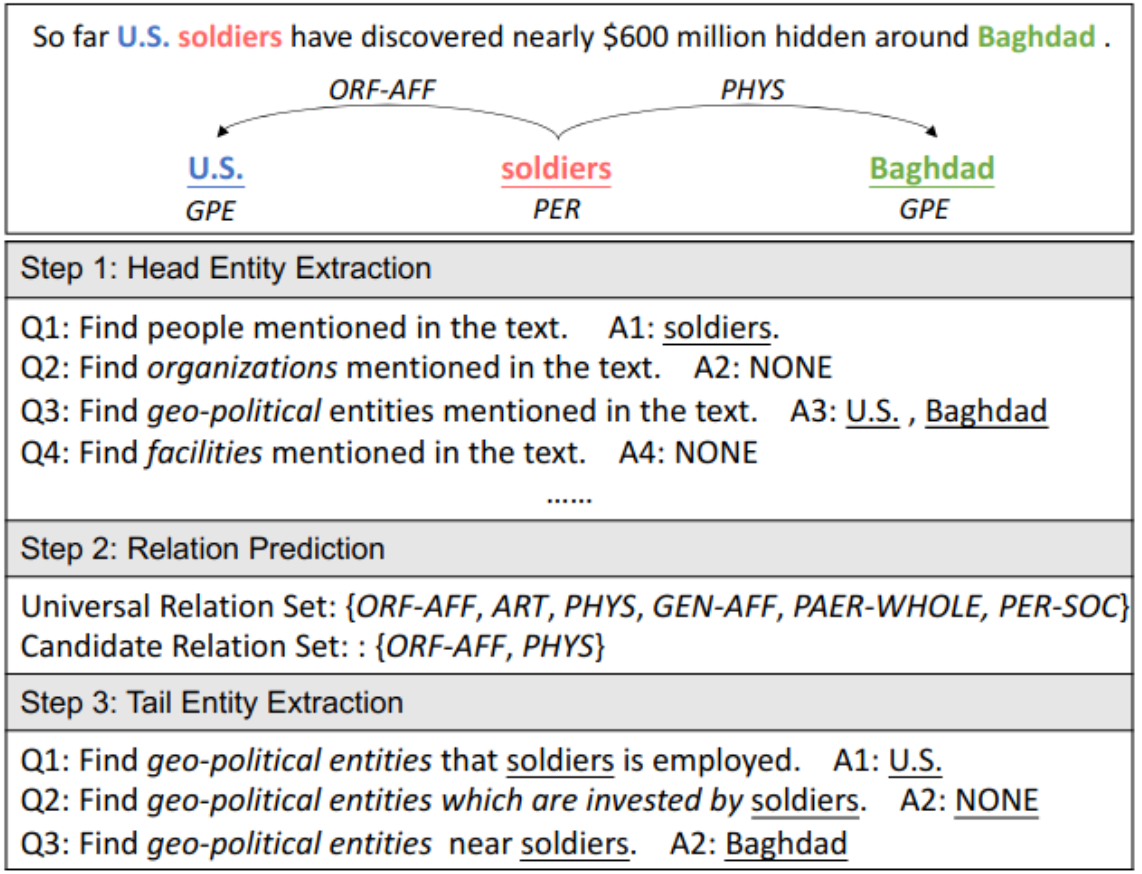
Seq2Seq-based Approach – "REBEL"

- Recieve unstructured text as input and directly generate (head entity – relation – tail entity) as sequential output
- Utilizes translation setup with Encoder-Decoder Transformer model (T5, BART, ...)



REBEL: Relation extraction by end-to-end language generation (2021.) [26]

MRC-based Approach – "Asking Effective and Diverse Questions"



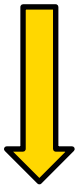
- Utilizes machine reading comprehension (MRC) and multi-turn question answering (QA)
- Exploits well-developed (MRC) models - extract text spans in passages given queries

Asking effective and diverse questions: a machine reading comprehension based framework for joint entity-relation extraction (2020.) [27]

Named Entity Recognition (NER)

- **Named-entity recognition (NER)** is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, etc.[1]

"***El Niño–Southern Oscillation (ENSO)*** *is another important factor for* ***winter temperature in China***."



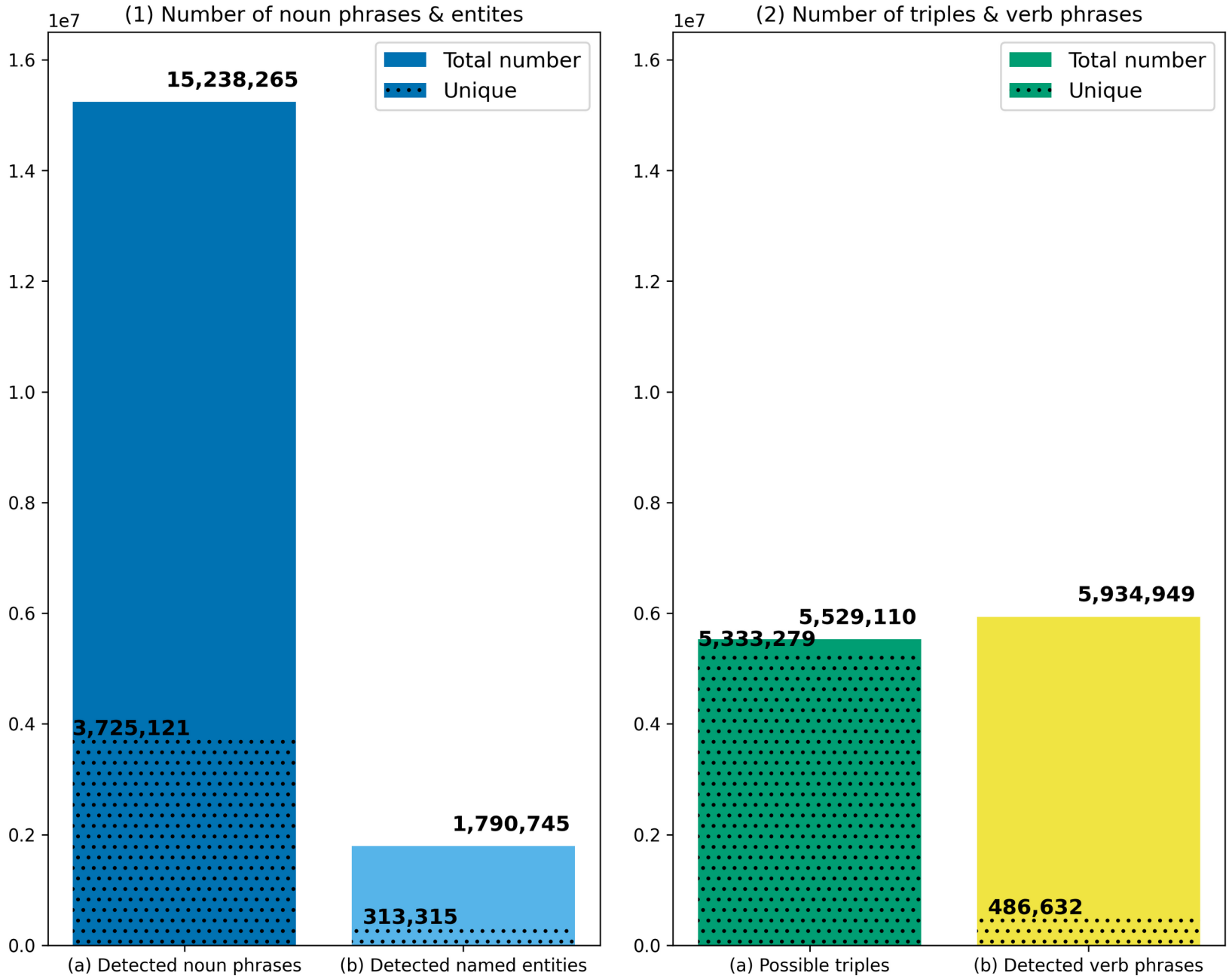
El Niño–Southern Oscillation (ENSO)	– ***Meteorological phenomena [MP]***
winter temperature in China	– ***Meteorological attribute [MA]***
China	– ***Location [LOC]***

[1] https://en.wikipedia.org/wiki/Named-entity_recognition

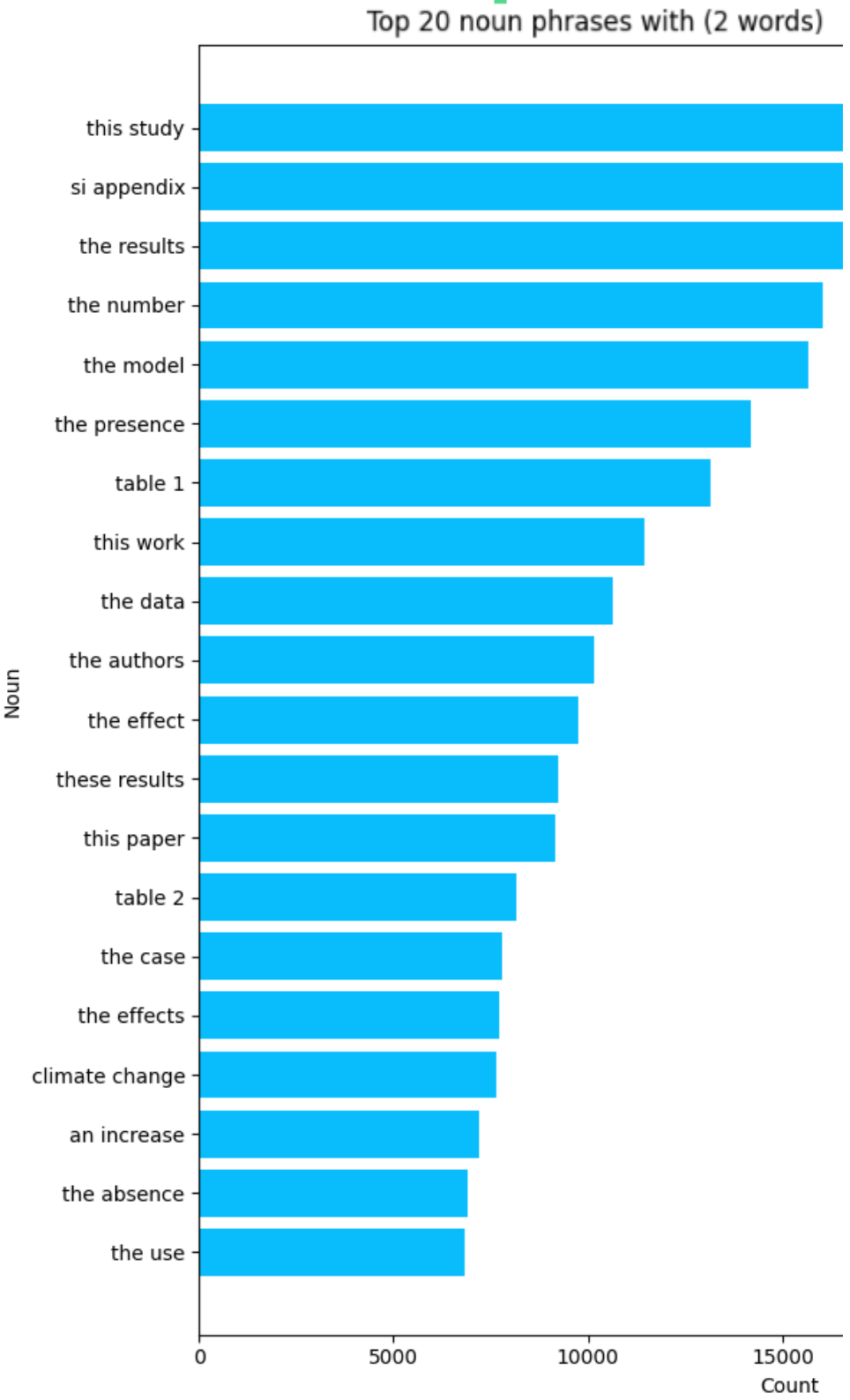
Data Exploration

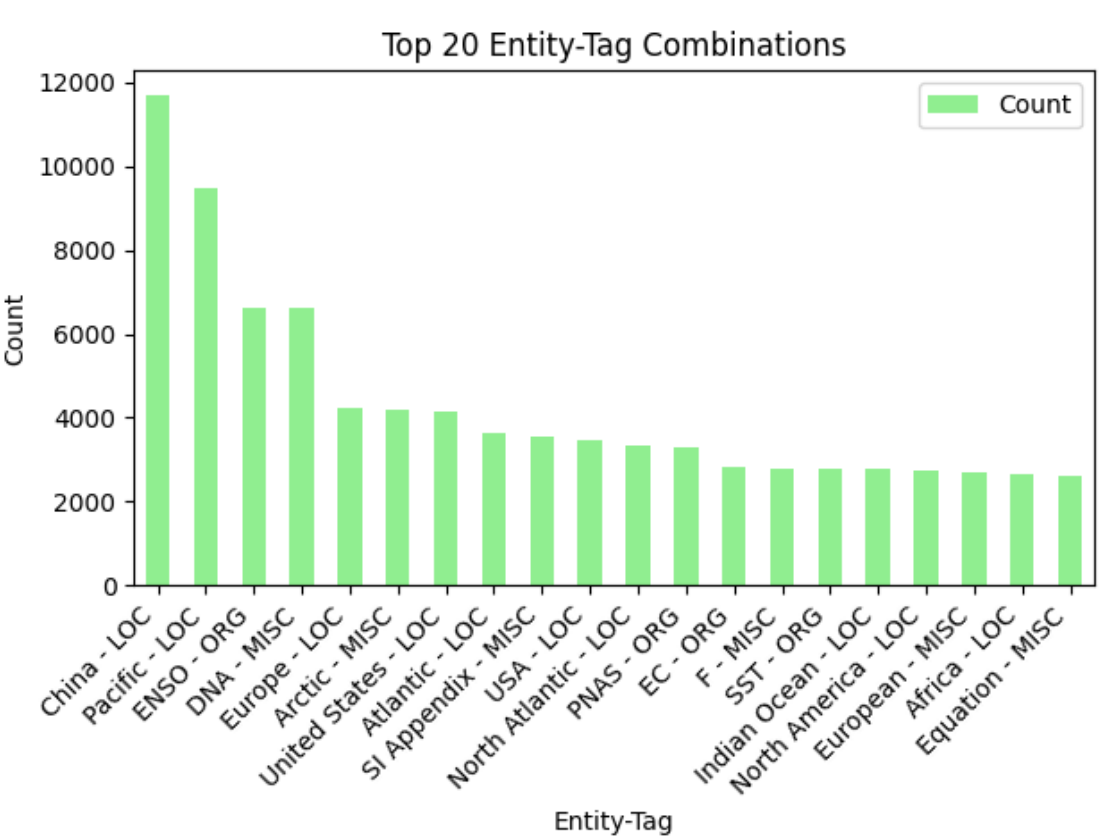
- Sample of **10,000** (~5%) research papers
- Part-Of-Speech (POS) tagger[1]:
  - **Noun phrase (NP)** → potential entity
  - **Verb phrase (VP)** → potential relation
- Named Entity Recognition (NER) model[2]:
  - **Named entity**
- *of-the-shelf* NER & POS from **FLAIR** framework for state-of-the-art NLP

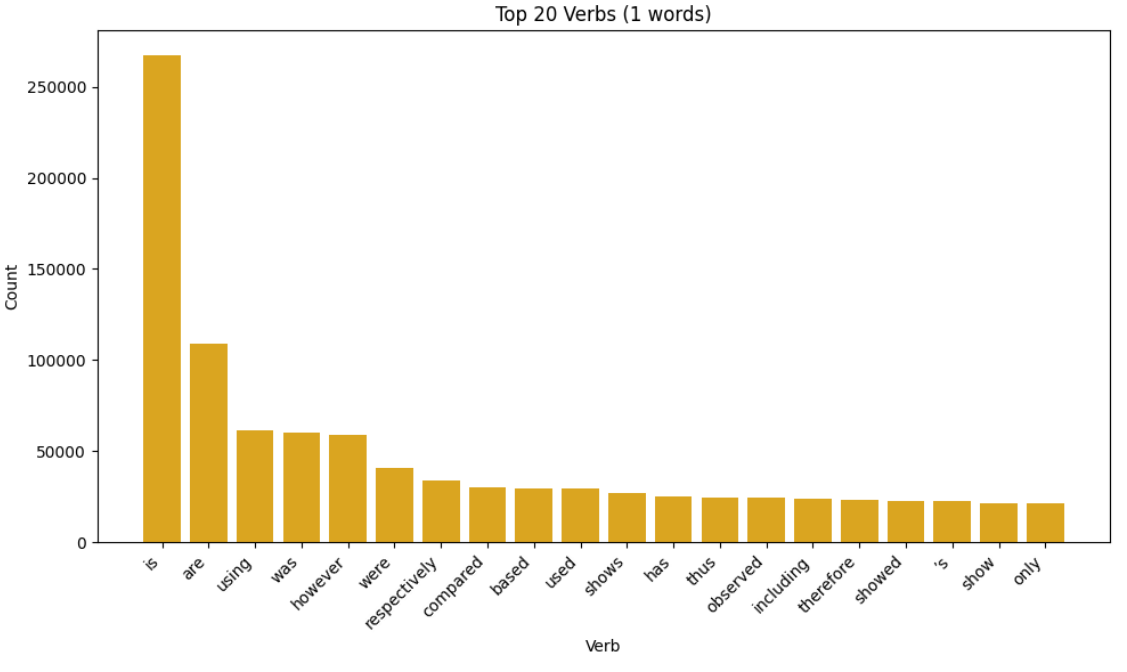
[1] https://huggingface.co/flair/pos-english
[2] https://huggingface.co/flair/ner-english-large



Towards Dataset for Extracting Relations in the Climate-Change Domain (2024.) [39]


Top 20 noun phrases with (2 words)
this study
si appendix
the results
the number
the model
the presence
table 1
this work
the data
the authors
the effect
these results
this paper
table 2
the case
the effects
climate change
an increase
the absence
the use
Noun
0
5000
10000
15000
Count


Top 20 Entity-Tag Combinations
Count
12000
10000
8000
6000
4000
2000
0
Count
China - LOC
Pacific - LOC
ENSO - ORG
DNA - MISC
Europe - LOC
Arctic - MISC
United States - LOC
Atlantic - LOC
SI Appendix - MISC
USA - LOC
North Atlantic - LOC
PNAS - ORG
EC - ORG
F - MISC
SST - ORG
Indian Ocean - LOC
North America - LOC
European - MISC
Africa - LOC
Equation - MISC
Entity-Tag


Top 20 Verbs (1 words)
250000
200000
150000
100000
50000
0
Count
is
are
using
was
however
were
respectively
compared
based
used
shows
has
thus
observed
including
therefore
showed
's
show
only
Verb

Distant Supervision

- Can automatically generate **large scale labeled training dataset** by aligning entities in texts with the entities in knowledge bases (KBs) (eg. Wikidata)
- **The distant supervision (DS) assumption**:

Assume that if a pair of entities has a relation in the KBs, then all sentences that mention the pair of entities will express this relation.

  - Suffers from noisy labeling!

*"The 2000 Dutch Open was an ATP tennis tournament staged in SUBJ{Amsterdam}, OBJ{Netherlands} and played on outdoor clay courts." → (Amsterdam, **capital of**, Netherlands)*

*"Aggregate reviews on OBJ{Amazon.co.uk} and SUBJ{Goodreads} are a little more positive." → (Goodreads, **owned by**, Amazon.co.uk)*

LLM Annotation Setup

- Unconstrained **3-shot learning** prompt with **LLM** to obtain possible (**noisy**) **entites** and **relations**
- **USING: microsoft/Phi-3-mini-4k-instruct**
- **Tested:** qwen:32b-text, mixtral:8x7b, gemma:7b-instruct-q8_0, llama3:8b
- Libraries/Frameworks:
  - **Transformers → using**
  - Ollama – https://ollama.com/
  - Vllm – https://github.com/vllm-project/vllm

LLM Prompt Template

Text:
Rain spells are a key parameter in examining rainfall regime.
Triples (entity, relation, entity):
('Rain spells', 'are_a_key_parameter_in', 'examining'), ('Rain spells', 'are_a_key_parameter_in', 'rainfall regime')

Text:
A rain spell is defined as a series of consecutive days above a certain daily rainfall threshold (DRT).
Triples (entity, relation, entity):
('rain spell', 'is defined as', 'series of consecutive days'), ('series of consecutive days', 'above', 'daily rainfall threshold'), ('daily rainfall threshold', 'has', 'certain DRT')

Text:
For each rain spell, its duration (RSD), yield (RSY) and average intensity (RSI), as well as their total number (NRS) in each year, were calculated.
Triples (entity, relation, entity):
("rain spell", "has attribute", "duration"), ("rain spell", "has attribute", "yield"), ("rain spell", "has attribute", "average intensity"), ("rain spell", "calculation involves", "its duration"),

Text:
{**INPUT_SENTENCE**}
Triples (entity, relation, entity):

LLM Annotation Results

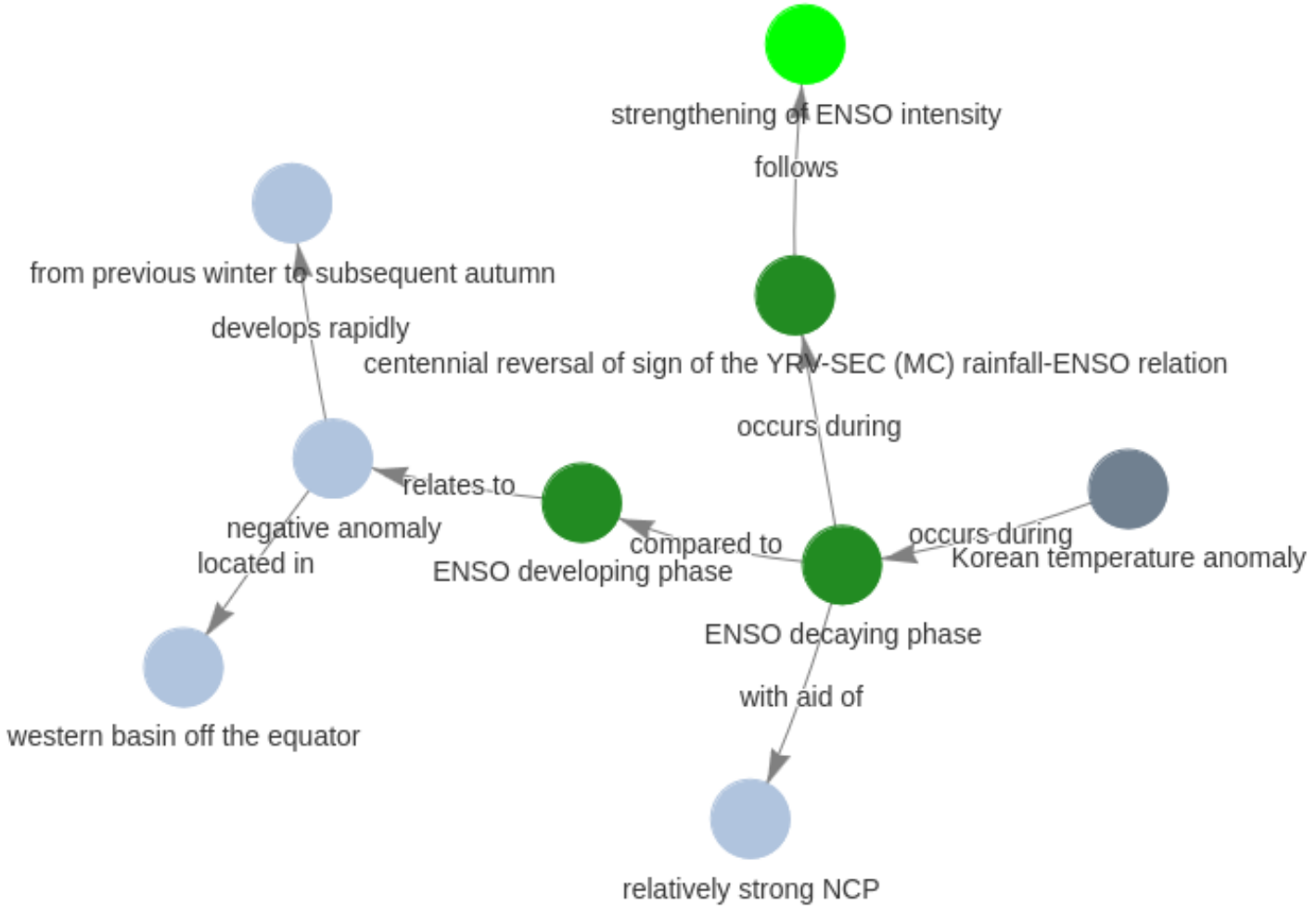
INPUT_SENTENCE: Excavation work primarily attributes to particulate matter (PM) pollution on construction sites.
OUTPUT: **('Excavation work', 'attributes to', 'particulate matter pollution')**

INPUT_SENTENCE: Under these synoptic conditions, rainfall is generally measured in Trentino 1 day out of 2, and extreme rainfall days are more frequent than average in all seasons.
OUTPUT: ("rainfall", "measured under", "synoptic conditions"),
("Trentino", "occurs on", "1 day out of 2"),
("extreme rainfall days", "occur more frequently than", "average")

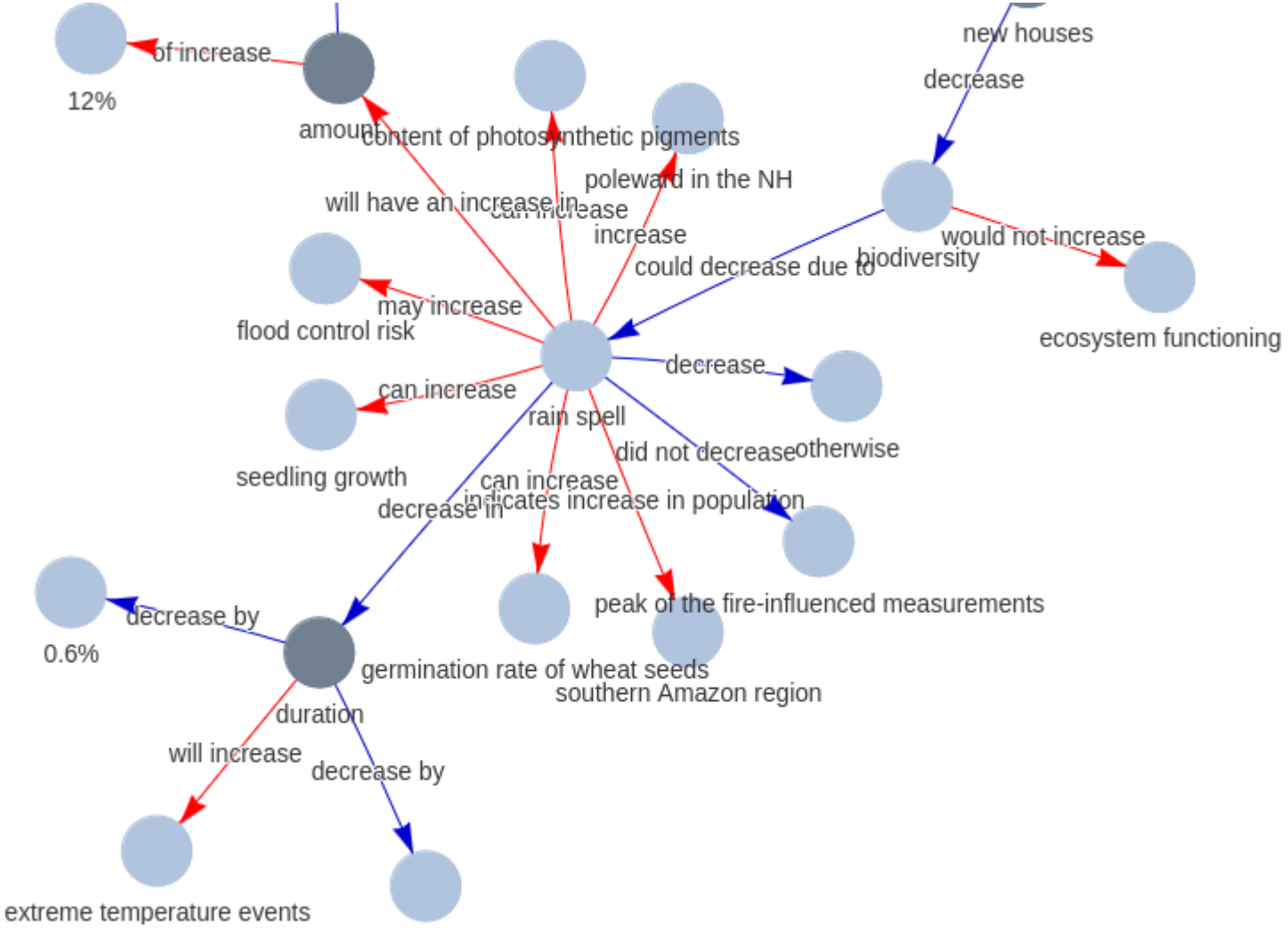
INPUT_SENTENCE: The inversion base height, on the other hand, has an annual cycle similar to that of the inversion top, but it shows also a strong autumn cooling, which is responsible for the large annual cycle in inversion strength observed in Fig. 4b.
OUTPUT: **("inversion base height", "shows", "annual cycle")**
("inversion base height", "shows", "autumn cooling")
("autumn cooling", "causes", "large annual cycle in inversion strength")

INPUT_SENTENCE: Like many tree species, the suitability of Northern China's habitat will benefit from climate change [61,62].
OUTPUT: ('tree species', 'include','many'),
('climate change', 'benefit', 'habitat suitability'),
('Northern China's habitat', 'will benefit from', 'climate change')

LLM Annotation - Graph e.g. 1



LLM Annotation - Graph e.g. 2

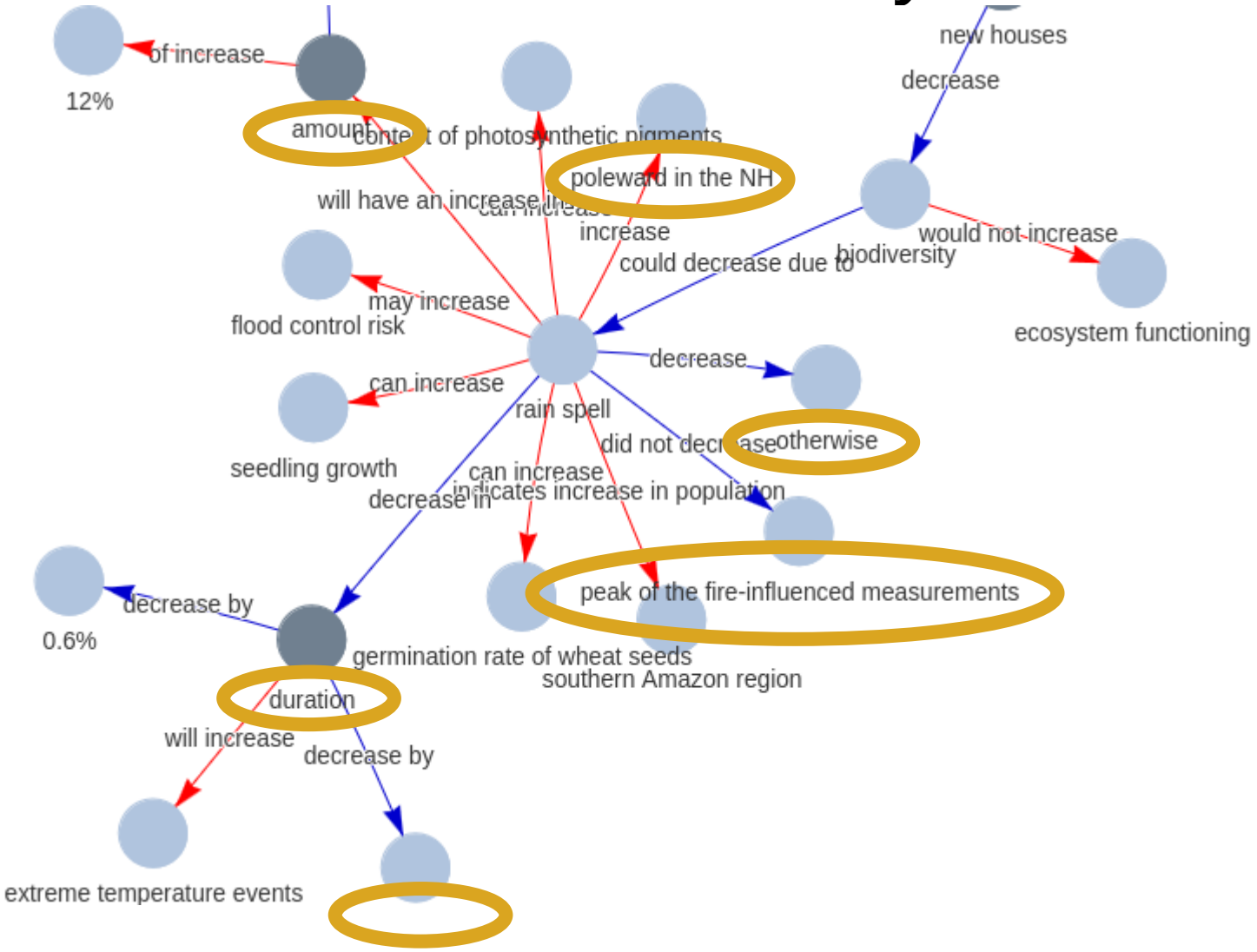


Relation Clustering

"**Unsupervised Relation Extraction** (RE) aims to identify relations between entities in text, **without having access to labeled data during training**. This setting is **particularly relevant for domain specific RE** where no annotated dataset is available and for open-domain RE where the types of relations are a priori unknown. " - [28]

- Related work:
  - SelfORE: Self-supervised Relational Feature Learning for Open Relation Extraction [29]
  - A Relation-Oriented Clustering Method for Open Relation Extraction [30]
  - A Unified Representation Learning Strategy for Open Relation Extraction with Ranked List Loss [31]
  - Unsupervised Relation Extraction: A Variational Autoencoder Approach [32]
  - Entity, Relation, and Event Extraction with Contextualized Span Representations [33]
  - Element Intervention for Open Relation Extraction [34]
  - A Frustratingly Easy Approach for Entity and Relation Extraction [35]

LLM Annotation - Graph e.g. 2 "under an observent eye"



NOTE that this is a cherry-picked graph!

(Expected) Results

- **New corpus:**
  - Scientific papers in Climate Change Domain
- **4 Domain specific language models**
  - 2 Language model pretraining
  - 2 Domain Adaptive pretraining OR
  - 2 BERT
  - 2 RoBERTa
- **Method for Relation Extraction** using BERT-like architectures
  - Pipline approach → Span-based approach
- **Relation Extraction evaluation**
  - (LLM??) Hand annotated golden dataset on ENSO
- **Climate Research Knowledge Graph (CliReKG)**
- ...

Open questions

- Relation Extraction
  - How to model the data? Which **relation types** are relevant and present in the domain?
    - OpenRE approach with clustering to form relation types - **Labels**
- Named Entity Recognition
  - How to model the data once again? Which **entity types** are relevant and present in the domain?
    - Entity Disambiguation approach through Wikidata and usage of existing taxonomy and structure of Wikidata to inherit entity types - **Labels**
- Evaluation – Golden Dataset ENSO
- Which data to use for KG construction?

References

1. Peter Andre, Teodora Boneva, Felix Chopra, and Armin Falk. Globally representative evidence on the actual and perceived support for climate action. Nature Climate Change, 2 2024
2. Charles S. Areni. Motivated reasoning and climate change: Comparing news sources, politicization, intensification, and qualification in denier versus believer subreddit comments. Applied Cognitive Psychology, 38(1), 2024. All Open Access, Hybrid Gold Open Access.
3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the 2 Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis,Minnesota, June 2019. Association for Computational Linguistics.
4. Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.
5. Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold.Climatebert: A pretrained language model for climate-related text. SSRN, September 26 2022.
6. Andrija Poleksi´c and Sanda Martin˘ci´c-Ip˘si´c. Towards dataset for extracting relations in the climate-change domain. In Sanju Tiwari, Nandana Mihindukulasooriya, Francesco Osborne, et al., editors, Joint Proceedings of the 3rd International Workshop on Knowledge Graph Generation from Text (TEXT2KG) and Data Quality Meets Machine Learning and Knowledge Graphs (DQMLKG) co-located with the Extended Semantic Web Conference (ESWC 2024), volume 9, page 15, Heraklion, 2024. CEUR.
7. Suchin Gururangan, Ana Marasovi´c, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Online, July 2020. Association for Computational Linguistics.
8. Vin Sachidananda, Jason Kessler, and Yi-An Lai. Efficient domain adaptation of language models via adaptive tokenization. In Nafise Sadat Moosavi, Iryna Gurevych, Angela Fan, Thomas Wolf, Yufang Hou, Ana Marasovi´c, and Sujith Ravi, editors, Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing, pages 155–165, Virtual, November 2021. Association for Computational Linguistics.
9. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
10. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.
11. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.
12. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, DonghyeonKim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pretrained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240, January 2019. MAG ID: 2911489562.
13. Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2020.
14. Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school, 2020.
15. Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. Juribert: A masked-language model adaptation for french legal text, 2022.
16. Taha ValizadehAslani, Yiwen Shi, Ping Ren, Jing Wang, Yi Zhang, Meng Hu, Liang Zhao, and Hualou Liang. PharmBERT: a domain-specific BERT model for drug labels. Briefings in Bioinformatics, 24(4):bbad226, 06 2023.
17. Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should You Mask 15% in Masked Language Modeling? In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2985–3000, Dubrovnik, Croatia, 2023. Association for Computational Linguistics.
18. Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Towards Climate Awareness in NLP Research, October 2022. arXiv:2205.05071 [cs].
19. Thien Huu Nguyen and Ralph Grishman. Relation extraction: Perspective from convolutional neural networks. In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pages 39–48, Denver, Colorado, June 2015. Association for Computational Linguistics.
20. Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 207–212, Berlin, Germany, August 2016. Association for Computational Linguistics.
21. Christoph Alt, Marc H¨ubner, and Leonhard Hennig. Improving relation extraction by pre-trained language representations, 2019.
22. Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics.
23. Andrija Poleksi´c. Relation extraction with deep learning methods, 2023. Repository of Faculty of Informatics and Digital technologies, University of Rijeka.
24. Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. A comprehensive survey on relation extraction: Recent advances and new frontiers. ACM Comput. Surv., 56(11), July 2024.
25. Wei Tang, Benfeng Xu, Yuyue Zhao, Zhendong Mao, Yifeng Liu, Yong Liao, and Haiyong Xie. UniRel: Unified representation and interaction for joint relational triple extraction. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7087–7099, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
26. Pere-Llu´ıs Huguet Cabot and Roberto Navigli. REBEL: Relation extraction by end-to-end language generation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2370–2381, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
27. Tianyang Zhao, Zhao Yan, Yunbo Cao, and Zhoujun Li. Asking effective and diverse questions: a machine reading comprehension based framework for joint entity-relation extraction. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI'20, 2021.
28. Pierre-Yves Genest, Pierre-Edouard Portier, El¨od Egyed-Zsigmond, and Laurent-Walter Goix. Promptore - a novel approach towards fully unsupervised relation extraction. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, page 561–571, New York, NY, USA, 2022. Association for Computing Machinery.
29. Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip Yu. SelfORE: Self-supervised relational feature learning for open relation extraction. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3673–3682, Online, November 2020. Association for Computational Linguistics.
30. Jun Zhao, Tao Gui, Qi Zhang, and Yaqian Zhou. A relation-oriented clustering method for open relation extraction. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9707–9718, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
31. Lou Renze, Zhang Fan, Zhou Xiaowei, Wang Yutong, Wu Minghui, and Sun Lin. A unified representation learning strategy for open relation extraction with ranked list loss. In Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao, editors, Proceedings of the 20th Chinese National Conference on Computational Linguistics, pages 1096–1108, Huhhot, China, August 2021. Chinese Information Processing Society of China.
32. Chenhan Yuan and Hoda Eldardiry. Unsupervised relation extraction: A variational autoencoder approach. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1929–1938, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
33. David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5784–5789, Hong Kong, China, November 2019. Association for Computational Linguistics.
34. Fangchao Liu, Lingyong Yan, Hongyu Lin, Xianpei Han, and Le Sun. Element intervention for open relation extraction, 2021.
35. Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction, 2021.
36. M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, 'SpanBERT: Improving Pre-training by Representing and Predicting Spans', Transactions of the Association for Computational Linguistics, vol. 8, pp. 64–77, 2020.
37. P. He, X. Liu, J. Gao, and W. Chen, 'DeBERTa: Decoding-enhanced BERT with Disentangled Attention', arXiv [cs.CL]. 2021.
38. K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, 'ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators', arXiv [cs.CL]. 2020
39. Poleksić, Andrija; Martinčić-Ipšić, Sanda Towards Dataset for Extracting Relations in the Climate-Change Domain // Joint proceedings of the 3rd International workshop one knowledge graph generation from text (TEXT2KG) and Data Quality meets Machine Learning and Knowledge Graphs (DQMLKG) co-located with the Extended Semantic Web Conference (ESWC 2024) / Tiwari, Sanju; Mihindukulasooriya, Nandana; Osborne, Francesco et al. (ur.). Heraklion: CEUR, 2024, 9, 15.

University of Rijeka
Faculty of Informatics and Digital Technologies

Construction of a Knowledge Graph in the Climate Research Domain

Author: **Andrija Poleksić**

Mentor: prof. dr. sc. Sanda Martinčić-Ipšić

Croatian Science Foundation

AP is fully supported by Croatian Science Foundation under the project DOK-2021-02.