CLEF ’25 notes and interesting posters

Benedikt Kantz

Sept 09-12 2025

Contents

1 Introductions

1.1 Keynote: Sameer Antani, AI for Medicine

- Data in medicine is difficult, often biased (i.e. more prevalence of disease vs. natural distribution due to only imaging correct skin cancer)
- AI in medicine must be multimodal (e.g. there is always an order attached to an image!)
- Synthesis as a remedy:
  - Clinical training for people
  - Fill data gaps/sparse data,
  - Problems: Hallucinations, not rule-based (anatomy, diseases, ...)
- Evaluation of synthetic data: what is the specific impact of it being added?
  - Generalization? or just improvements?
  - Hallucinations eval?

Note: CLEF has changed reviews to focus on methodology instead of raw numbers (was good for our submission I guess?)

1.2 Conference Sessions I (Best of CLEF 2024)

1.2.1 Humour Classification According to Genre and Technique by Fine-tuning LLMs

- Add the definitions of the classes into prompts
- Tree-based LM classifier

1.2.2 Language-based Mixture of Transformers for Sexism Identification in Social Networks

- Use ensemble of domain-specific models (models trained on Twitter, same source domain!)
- Model mixture: variation (either half-half, 75 percent or only dominant)
- Some fine-tuning
- Q: how are they mixed? i.e. at what stage, dynamically chosen? based on what??

1.2.3 Robustness of Misinformation Classification Systems to Adversarial Examples Through BeamAttack

- Counterfactuals for classification, which minimal modification has to be done to change output (CheckTHAT task)
- Effectively: how good is the adverserial attack? (Similarity - Levensthein, Effectiveness scored)
- BERT-Attack:
  - Which word: based on word importance (using logits, each word is masked after each other - calculate probability)
  - What to insert: masked AE (e.g. RoBERTa)
- DeepWordBug: replace characters/typos
- Theirs: use beam-search for improved search of replacement
- Tree width + depth of search are hyperparameters
- Disadvantage: needs a lot of evaluation whether they affect the classifier

1.3 Labs Overview

394 Papers, Labs: 13 old + 1 new

1.3.1 Overview of LifeCLEF 2025: Challenges on Species Presence Prediction and Identification, and Individual Animal Identification

- @Simon? iNaturalist :)
- For environmental monitoring
  - BirdCLEF: Sound classification (3k participants - 50k price pool!)
  - PlantCLEF: detection of plants in plots of land
  - GeoCLEF: Multimodal Classification for species
  - AnimalCLEF: Open-Set classification (New! individuals)
  - FungiCLEF: Few-shot classifiction, with multi-modal description
- Paper count not correlated to price pools ;)
- Foundational models were the winners
- Compared to humans: only experts can outperform these models, have strong location prior

1.3.2 Overview of BioASQ 13

- 6 tasks, 6 languages, 3 doc types
- 17 participant in GutBrainIE

1.3.3 Overview of Touché 2025: Argumentation Systems

- Debate simulation,
  - Evaluation Grice's maxims of cooperation
  - Systems often switched sides or admitted defeat!
- analysis
  - ParlaMint: multilingual debates, scores on english best

- image arguments (generation+analysis); eval → core aspects of images are evaluated; best submission extracted aspects and prompted image gen
- Advertisement in RAG: Generate (eval: classifier), and detect ads in responses (AdBlock for LMs) `https://touche.webis.de/clef25/touche25-web/advertisement-detection.html#task` (eval: yes/no)

1.3.4 Overview of the CLEF 2025 JOKER Lab: Humour in Machine

- LMs not able to deal with humor etc.
- humor-aware IR
  - Search for jokes on topics
  - Manual + LM generated jokes, mixed with non-humor (wikipedia)
  - Eval: humor + traditional IR metrics; way better results this year!
- Translate puns
  - Wordplay consistent accross EN-FR translation
  - Q: is the annotation for the "funny word" given to the participants?
  - Eval: consistent meaning of translations, location based of the wordplay
- Onomastic Wordplay Translation
  - e.g. often in Harry Potter, Asterix, ...
  - Used in training sets
  - EN-FR
  - Q: copyright, could GPT have been trained on the source material?

1.3.5 LongEval at CLEF 2025: Longitudinal Evaluation of IR Systems on Web and Scientific Data

- training on evolving information needs over 9 months
- Trending queries and qrels (click models)
- On the TU Wien Research Dataset!

2 LifeCLEF 2025

2.1 Learning from Visual Data in the Wild (Oisín Mac Aodha)

- Growth in Biodiversity data ← iNaturalist,
- Range Maps of Species
  - downside of these citizen scientist approaches: spatially sparse - biased towards human locations, species distribution mismatched with iNaturalist observation
  - Very few expert Range maps...
  - LM generated range maps: only squares, very bad in relation to correct Range maps (interesting research topic?)
  - Idea: Sparse input of observation, output of range maps?
  - Presence detection - based on spatial embeddings $\odot$ species embeddings; need to be compact - fit on phones, improve offline CV species prediction (actually improves it! & is deployed on iNaturalist)
  - Spatial embeddings + species embeddings helps share data between low-observations and high-observation species
  - No absence data, only present data...
  - Visualization of high-dim vector on spatial data: PCA to 3D to RGB
  - Add text to context: as few as 5-10 observations from text works as text quite well
  - Joint training with representation learning for satellite images: Dense Retrieval of Text, Segmentation, ...

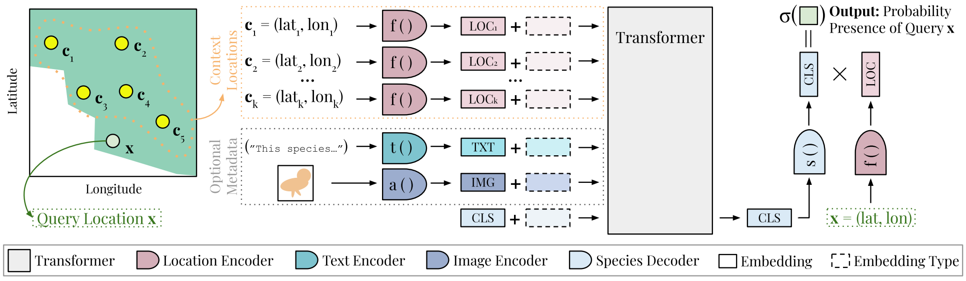


Figure 1: Overview of the FS-SINR [1]

2.2 GeoLifeCLEF Overview

- Absence/Presence data, climate data, time series (climate)
- Very biased observations, people go places
- Test data: not only in-distribution, but OOD with new regions (with presence only)

Participant: Gleb Tikhonov

- Combination of a lot of handcrafted features and encoding systems
- Embedding of images, . . .
- Averaging, cycling,. . .

2.2.1 PlantCLEF Overview

- Earlier: monospecies
- Now: multispecies, in singular images
  - Multiscale, variety of seasons,
  - Train: single plants, monospecies
  - Some non-annotated quadrats
  - Test: multi-label plots → zero shot object detection

Participant: Luciano Dourado

- Approach: filter out background using attention based segmentation using prototype guidance
- Train narrow ViT to match baseline classifier (DinoV2) classification matrix and calculate attention map to find relevant regions
- Use DinoV2 to classify region patches, use grid assembly to search around patch

2.3 Promises and pitfalls of foundation models for the natural world, Lauren Gillespie (MIT)

- Rapid change to environment
- Requires new models: foundation models
- CRISP incorporates multimodal unlabeled data, improves performance across many species detection and range labels (esp. low observation species)

2.3.1 AnimalCLEF Overview

- Challenge: identify *individuals* (e.g. a very specific turtle) given a database of known individuals
- Also: unseen individuals, unclear images, non-overlapping
- Challenges
  - Individual is present, . . .

2.3.2 BirdCLEF Overview

- Bioacoustic surveys: use as restoration markers
- Goals:
  - identify taxonomic groups
  - experiment with limited training data
  - experiment with unlabeled data

2.3.3 FungiCLEF Overview

- Few-Shot ID with few samples
- Data: photos, description, metadata, satelite, climate
- Public leaderboard very different from private, takeaway: be robust!
- Different ensemble types, etc.
- Vision-Only Pipelines, Constrastive learning and prototypes helpful!

Participant: Anthony Miyaguchi, GATECH@LifeCLEF

- DS@Georgia Tech - big Data Science group, a lot of publications!!
- PlantCLEF approach: embeddings in kNN setting, adding GEO-info and Priors help a bit
- FunghiCLEF: vLLM bad with just prompting, better: interpolating embedding subspaces
- BirdCLEF: Best Working notes, Tokenize Audio dataset (spectrogramm), then train on dataset using word2vec+skip-grams, build linear model on top - very efficient, good for deployment!

3 Wednesday

3.1 AI Evaluation Should Make AI Predictable

- Rate LMs by capabilities (as new metrics)
- Taxonomy of LM problems - apply to benchmarks, LM benches measure different things that they claim to (i.e. math tests lang understanding) (Q: the taxonomy is annotated using GPT, isn't this a weakness?)
- Enables the plotting of levels as Spidercharts

3.2 Conference Sessions II

3.2.1 SimpleText Best of Labs in CLEF-2024: Application of Large Language Models for Scientific Text Simplification

3.2.2 Simplified Longitudinal Retrieval Experiments: A Case Study on Query Rewriting and Document Boosting

- Longitudinal evaluation: they provide datasets that can be evaluated for over longer timespan using containers etc.
- Snapshots of datasets
-

3.2.3 Better Call Claude: Can LLMs Detect Changes of Writing Style?

- Identify sentence boundaries
- Goals: benchmark 0-shot on sentence lvl, baselines comparisons, semantic similarity vs. stylistic cues
- Claude has good 0-shot performance, semantic similarity correlated to stylistic changes (?)

3.3 Conference Sessions III + More Labs intro

3.3.1 From Uniform to Unique: Adaptive K-12 Assessment Using Large Language Models

- Generate and asses questions from Kindergarten to 12th Grade
- Use Bloom's taxonomy to instruct model (Remember, Apply, Evaluate) and generate MCQ
- Suppress guessing

3.3.2 Lab Introductions

PAN@CLEF

- AI author attribution[1]
  - Binary classification: AI generation? (with Builder/Breaker (red/blue teams), similar to NLP class of Roman Kern) - text with obfosucation; baseline: binoculars, TF-IDF
  - Classify extent of AI gen
- Multilingual detoxification: classification, de-toxify based on keywords; some varied baselines
- Multi-Author Style change detection
- Generative Plagiarism detection

EXIST@CLEF

- Focuses on Benevolent Sexism (e.g. underlying, cultural stereotypes)
- Human Annotations: very varied annotations, embraced as different opinions -¿ target: soft classification
- Novelty: tiktok videos!
- 300k annotations, bias attention
- Sexism classification (binary), direct/ reported/judgemental, kind of sexism (multilabel!), multilungual, multimodal

SimpleText

- Sentence & Document level simiplification
- Measure hallucinations in sentence outputs from last years

QuantumCLEF

- Eval QC algorithms
- Foster understanding & build community for QC+IR
- quantum annealing: setup qbits and search for energy minimum
- QUBO: quadratic and binary optimisation – set for IR with retrieval metrics
- Tasks: Feature selection, Instance Selection, Clustering
- Task 1 results: 30x faster, about as effective!

3.4 BioASQ 3/4

3.4.1 MultiClinSum

- Summarize (multiple) long clinical reports
- Multilingual, Semi-Automatically generated summarization
- automatic translation for multilingual tasks
- extractive: only smaller models, bigger models abstractive

3.4.2 BioNNE-L

- Nested Entity Linking
- Multilingual challenge, terms missing in some languages – difficult to reconstruct (Russian, . . .)
- Shared dictionary
- Ambigous terms, UMLS coverage limited – joint dictionary with Russian
- Approach: BERGAMOT - BERT+Graph Encoder and bring together in space to align dictionaries

[1] `https://bladerunner.fandom.com/wiki/Voight-Kampff_test`

3.4.3 ElCardio - Clinical Cardiovascular diseases

- Task: coding (ICD-10 system) for multilingual setting, lack in low-resource languages of discharge letters & extracting code mentions
- similar to gutbrainIE → link entities to ICD-10
- identify all ICD-10 mentions within doc (reverse process)

4 Thursday

4.1 Do we co-evolve with what we design? DevOps, AGI, and Human Frailties

- Thoughts about how we co-evolve with AI, bio-inspired
- How does exponential growth affect/interact, or is it sigmoid? - how will this affect policy, how to move to stable society away from exp. growth

4.2 Main Conference Session III

4.2.1 MedAID-ML: A Multilingual Dataset of Biomedical Texts for Detecting AI-Generated

- Fake medical literature detection!
- AI generated text generation for multilungual detection

4.2.2 Selective Search as a First-Stage Retriever

- Make search more efficient, distribute web indices (effectively), sparse search...
- Distribute documents by clusters in distributed search
- Approach: use this only as fist-stage retrieval, but only care about first documents (i.e. 1000)
- Rank biased Recall (how does this differ from nDCG@k)
- Central problem: which shard (cluster) to take - different approaches based on vocab, . . .
- Problem: some selection algs can make shards 'invisible' - documents may not be retrieved as shard index may not expose or represent them correctly.
- Finding: is possible, but efficiency is still a bit lacking

4.3 Labs Overview III

4.3.1 ImageCLEF

- Since 2003 (!)
- Very multimodality-focused, medical tasks
- Datagen, retrieval, classification
- Tasks:
  - MedicalCLEF: caption, generation, VQA
  - ToPicto: image gen (text+speech to pictogram, mostly finetunes)
  - Multimodal VQA
  - Image Retrieval for Arguments
- a lot of participants, 500 runs (expensive)
- A lot of participants used VLMs, explanations: bbox + heatmaps
- Generation: find closest image from training data for generation

4.3.2 eRISK

- Symptom search for depression and detection
- Rank sentences from redding to clinical classes, contextualized detection, and conversational detection (earlier detection better!); LM personality detection task (Problem: jailbreaking. . .)

4.3.3 ELOQUENT

- Voight-Kampff task (AI detection, Blade Runner reference!) as red/blue teams - red team quite good, but none fooled all!
- A lot of misclassified, especially two texts: EU law text + intro to LMs ;)
- Value-Oriented questions, 15 languages - no specific answer; only joint participant report!
- Results: LM have some conservative views regarding live, etc.
- Relevance task: return very concise and relevant output!

4.3.4 CheckTHAT

- Tasks:
  - T1: subjectivity/check whether it should be checked
  - T2: Claim extraction
  - T3: Fact-Checking Numerical Claims
  - T4: Scientific Web Discourse: check and identify mentions

4.3.5 TalentCLEF

- Human Capital Management (??)
- HR: very digital, job portals…
- Tasks: Job Title Matching; Skill Prediction from Job Titles

4.4 ImageCLEF

4.4.1 Training Data Analysis and Fingerprint detection

- Synthetic data generation important for medicine (privacy)
- Problem: generative methods have fingerprints in them…
- Task 1: determine which images were used in training, results poor – interesting divide between tasks, reason not fully clear
- Task 2: link to sets of datasets, very high results??

4.4.2 Medical Concept Detection + Captioning

- Concept detection from images (img2text), evaluation using briefness and correctness
- Then explain with bbox, evaluated using radiologist professional (no formal eval, Likert-Scale) – i.e. GradCAM / IG
- Maybe next year as task? very interesting!!

4.4.3 Visual Question Answering and Synthetic Image Generation for Gastrointestinal Tract

- VQA: what, where, how many (polyps) in image- evaluated using BLEU
- Synthetic Data generation based on prompt

4.4.4 Visual Question Answering: Dermatologistical VQA

- Task 1: Segmentation Maps, solutions mostly finetuned domain models
- Task 2: 'predefined' questions from ontology

4.4.5 ImageCLEFtoPicto

- AAC: augmentative and alternative communication
- Very focused on pictograms, represents ideas & notions
- Currently: a lack of training, and very expensive (+awareness)
- Task: French Text/Speech 2 pictogram
- Very few participants, french-only

4.4.6 Multimodal Reasoning

- Many VQA: very simple questions, images loosely linked to text
- Their benchmark: multilingual (13 languages), multiple-choice, difficulty levels
- Task: Multiple-Choice Questions from student exams within europe
- Some languages test-only!
- Moderately difficulty, parallel data - exactly the same solution across languages, but big diff in languages (e.g. serbian - Cyrillic alphabet!)
- Everyone used VLMs (Qwen Vision)
- Future Work: university-level, are models really reasoning?

4.4.7 Image Retrieval/Generation for Arguments

- Illustrate Argument by images
- Evaluated by aspects contained
- Challenge: combine aspects effectively

5 Friday

5.1 ImageCLEF

5.1.1 ImageCLEFmedical

AUEB NLP Group/Archimedes

- Class Assignment: Multiple Vote strategy of CNNs with ResNET (Union, Intersection, ...)
- Captioning: Q-Former with query assignment,InstructBLIP, Cation Gen + medCLIP scoring (retrieval from generated captions)
- Explainability: assignment based on ChatGPT-drawn boxes on

DS4DH Group

- Concept detection: framed as sequence generation, concepts as tokens (hmmm, CUIs have order; a transformer might be correct) & condition on images
- Caption: InstructBLIP, RAG-based on image retrieval, cluster-based on topics

UMUTeam: Fine-Tuning a Vision-Language Model for Medical Image Captioning and SapBERT-Based Reranking for Concept Detection

5.1.2 MultimodalReasoning (Answers of visual highschool questions)

Ayesha Amjad: Visual Question Answering with Structured Data Extraction and Robust Reasoning

- Approach: Image Captioning using gemini + reasoning modeling for answer generation

ContextDrift: Evaluating VLMs' Multimodal, Multilingual and Multidomain Reasoning Capabilities via Thinking Budget Variations and Textual Augmentation

- Similar Approach, but visual model and prompt design
- A lot of ablation studies[2]

MSA: Multilingual Multimodal Reasoning with Ensemble Vision-Language Models

- OCR + vLLM
- + Ensembling

[2] https://www.dei.unipd.it/~faggioli/temp/clef2025/paper_194.pdf

5.1.3 MEDIQA-MAGIC

DS@GT

- Emulate Collaborative Reasoning of Physicians
- 7 vLLM + orchestrators, combination of reasoning ...

IReL, IIT(BHU): Tackling Multimodal Dermatology with CLIPSeg-Based Segmentation and BERT-Swin Question Answering

5.1.4 MEDVQA

Gaurav Parajuli (JKU, Linz): Querying GI Endoscopy

- LoRA finetuned vLLM

Sujata Gaihre

- Similar approach

Krishna Tewari

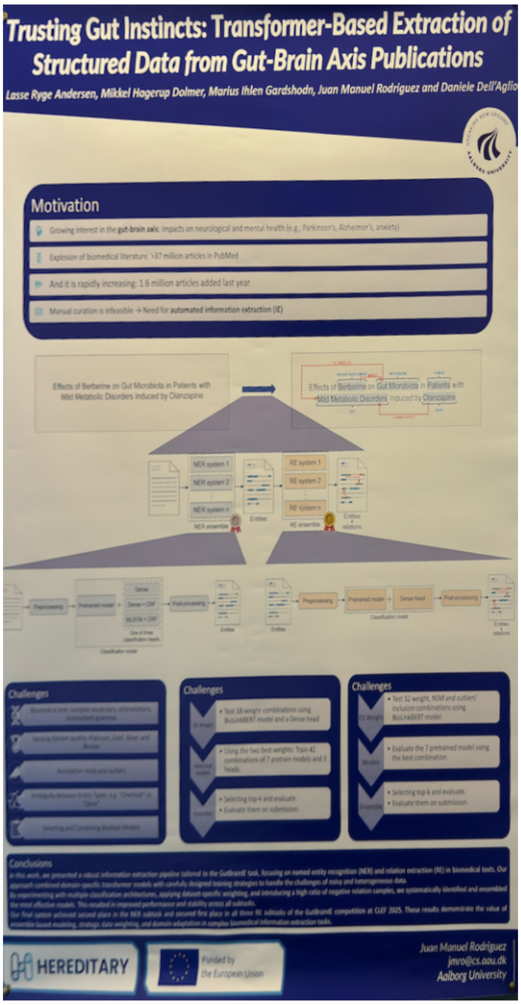
- Data Augmentation/Preprocessing!

5.2 Closing Ceremony

- New CLEF challenges ;)

6 Posters

Table 1:

<table>
<tr><th>Poster</th><th>Information</th></tr>
<tr><td>


Trusting Gut Instincts: Transformer-Based Extraction of Structured Data from Gut-Brain Axis Publications

Lasse Ryge Andersen, Mikkel Hagerup Dolmer, Marius Ihlen Gardshodn, Juan Manuel Rodriguez and Daniele Dell'Aglio

Motivation

- Growing interest in the **gut-brain axis**: impacts on neurological and mental health (e.g., Parkinson's, Alzheimer's, anxiety)
- Explosion of biomedical literature: >37 million articles in PubMed
- And it is rapidly increasing: 1.6 million articles added last year
- Manual curation is infeasible → Need for **automated information extraction (IE)**

Challenges

- Biomedical text: complex vocabulary, abbreviations, inconsistent grammar
- Varying dataset quality: Platinum, Gold, Silver and Bronze
- Annotation noise and outliers
- Ambiguity between Entity Types: e.g. "Chemical" or "Gene"
- Selecting and Combining Multiple Models

Challenges

DS Weight
- Test 18 weight combinations using BioLinkBERT model and a Dense head

Head and models
- Using the two best weights: Train 42 combinations of 7 pretrain models and 3 heads.

Ensemble
- Selecting top-k and evaluate.
- Evaluate them on submission.

Challenges

DS Weight
- Test 32 weight, NSM and outliers' inclusion combinations using BioLinkBERT model

Models
- Evaluate the 7 pretrained model using the best combination.

Ensemble
- Selecting top-k and evaluate.
- Evaluate them on submission.

Conclusions

In this work, we presented a robust information extraction pipeline tailored to the GutBrainIE task, focusing on named entity recognition (NER) and relation extraction (RE) in biomedical texts. Our approach combined domain-specific transformer models with carefully designed training strategies to handle the challenges of noisy and heterogeneous data.
By experimenting with multiple classification architectures, applying dataset-specific weighting, and introducing a high ratio of negative relation samples, we systematically identified and ensembled the most effective models. This resulted in improved performance and stability across all subtasks.
Our final system achieved second place in the NER subtask and secured first place in all three RE subtasks of the GutBrainIE competition at CLEF 2025. These results demonstrate the value of ensemble-based modeling, strategic data weighting, and domain adaptation in complex biomedical information extraction tasks.

HEREDITARY

Funded by the European Union

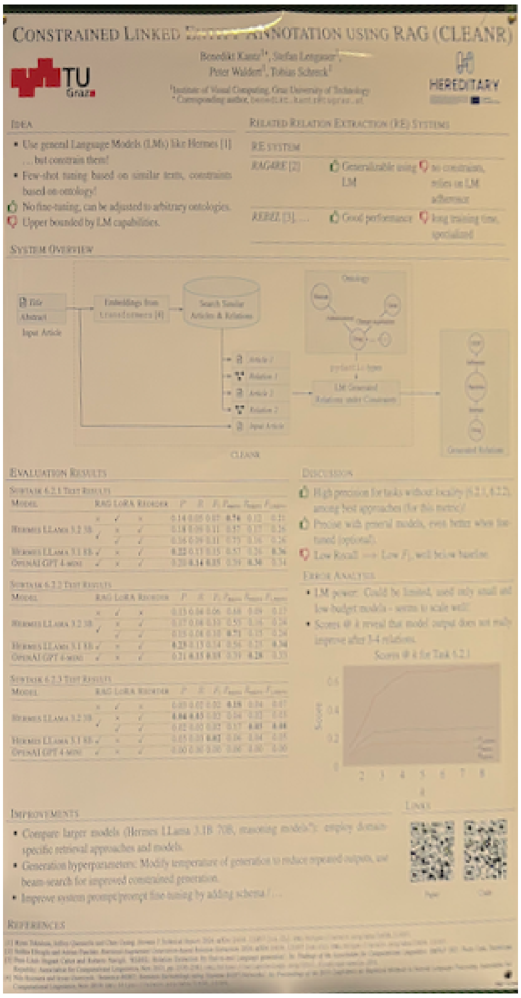
Juan Manuel Rodriguez
jmro@cs.aau.dk
Aalborg University
</td><td>

Trusting Gut Instincts: Transformer-Based Extraction of Structured Data from Gut-Brain Axis Publications *Lasse Ryge Andersen, Mikkel Hagerup Dolmer, Marius Ihlen Gardshodn, Juan Manuel Rodriguez and Daniele Dell'Aglic*

`https://www.dei.unipd.it/~{}faggioli/temp/clef2025/paper_6.pdf`
</td></tr>
</table>

Continued on next page

Table 1: (Continued)

	**Constrained Linked Entity ANnotation using RAG (CLEANR)** *Upper bounded by I.M capablises.* `https://www.dei.unipd.it/~{}faggioli/temp/clef2025/paper_23.pdf`

Constrained Linked Entity Annotation using RAG (CLEANR)

Benedikt Kantz[1*], Stefan Lengauer[1], Peter Waldert[1], Tobias Schreck[1]

[1] Institute of Visual Computing, Graz University of Technology
* Corresponding author, benedikt.kantz@tugraz.at

TU Graz

HEREDITARY

Idea

- Use general Language Models (LMs) like Hermes [1] … but constrain them!
- Few-shot tuning based on similar texts, constraints based on ontology!

👍 No fine-tuning, can be adjusted to arbitrary ontologies.

👎 Upper bounded by LM capabilities.

Related Relation Extraction (RE) Systems

RE system		
RAG4RE [2]	👍 Generalizable using LM	👎 no constraints, relies on LM adherence
REBEL [3], …	👍 Good performance	👎 long training time, specialized

System Overview

Evaluation Results

Subtask 6.2.1 Test Results

Model	RAG	LoRA	Reorder	P	R	F_1	P_{macro}	R_{macro}	F_{macro}
Hermes LLama 3.2 3B	×	✓	×	0.14	0.05	0.07	**0.76**	0.12	0.11
	✓	×	✓	0.18	0.09	0.11	0.57	0.17	0.26
	✓	✓	✓	0.16	0.09	0.11	0.73	0.16	0.26
Hermes LLama 3.1 8B	✓	×	✓	**0.22**	0.11	0.15	0.57	0.26	**0.36**
OpenAI GPT 4-mini	✓	×	✓	0.20	**0.14**	**0.15**	0.59	**0.36**	0.34

Subtask 6.2.2 Test Results

Model	RAG	LoRA	Reorder	P	R	F_1	P_{macro}	R_{macro}	F_{macro}
Hermes LLama 3.2 3B	×	✓	×	0.15	0.04	0.06	0.68	0.09	0.17
	✓	×	✓	0.17	0.08	0.10	0.55	0.16	0.24
	✓	✓	✓	0.15	0.08	0.10	**0.71**	0.15	0.24
Hermes LLama 3.1 8B	✓	×	✓	**0.28**	0.13	0.14	0.56	0.25	**0.34**
OpenAI GPT 4-mini	✓	×	✓	0.21	**0.15**	**0.15**	0.59	**0.28**	0.33

Subtask 6.2.3 Test Results

Model	RAG	LoRA	Reorder	P	R	F_1	P_{macro}	R_{macro}	F_{macro}
Hermes LLama 3.2 3B	×	✓	×	0.03	0.02	0.02	**0.19**	0.04	0.07
	✓	×	✓	**0.04**	**0.45**	0.02	0.04	0.02	0.03
	✓	✓	✓	0.02	0.00	0.02	0.17	**0.05**	**0.08**
Hermes LLama 3.1 8B	✓	×	✓	0.05	0.00	**0.02**	0.06	0.04	0.05
OpenAI GPT 4-mini	✓	×	✓	0.00	0.00	0.00	0.00	0.00	0.00

Discussion

👍 High precision for tasks without locality (6.2.1, 6.2.2), among best approaches (for this metric)!

👍 Precise with general models, even better when fine-tuned (optional).

👎 Low Recall ⟶ Low F_1, well below baseline

Error Analysis

- LM power: Could be limited, used only small and low-budget models – seems to scale well!
- Scores @ k reveal that model output does not really improve after 3-4 relations.

Scores @ k for Task 6.2.1

Improvements

- Compare larger models (Hermes LLama 3.1B 70B, reasoning models?); employ domain-specific retrieval approaches and models.
- Generation hyperparameters: Modify temperature of generation to reduce repeated outputs, use beam-search for improved constrained generation.
- Improve system prompt/prompt fine-tuning by adding schema / …

Links

Paper

Code

References

[1] Ryan Teknium, Jeffrey Quesnelle and Chen Guang. Hermes 3 Technical Report. 2024. arXiv: [illegible]

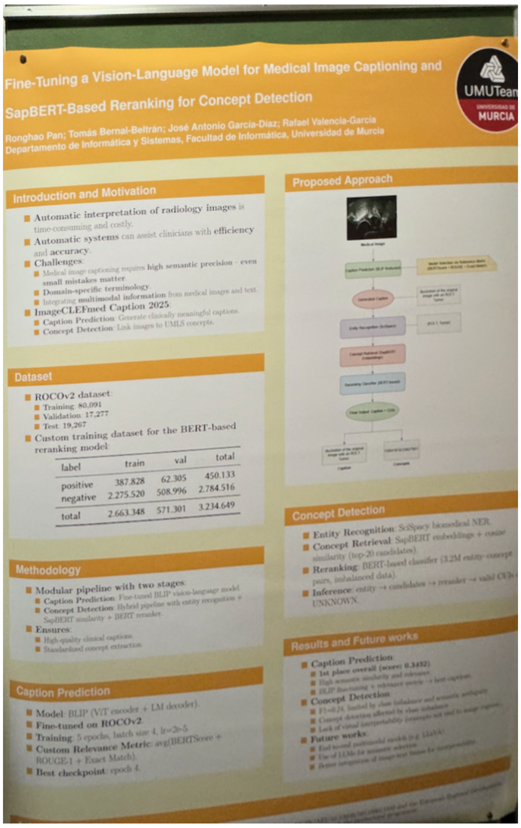
[2] Sefika Efeoglu and Adrian Paschke. Retrieval-Augmented Generation-based Relation Extraction. 2024. arXiv: [illegible]

[3] Pere-Lluís Huguet Cabot and Roberto Navigli. "REBEL: Relation Extraction By End-to-end Language generation". In: Findings of the Association for Computational Linguistics: EMNLP 2021. [illegible] Republic: Association for Computational Linguistics, [illegible]

[4] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, [illegible]

Continued on next page

Table 1: (Continued)

	**UMUTeam at ImageCLEF 2025: Fine-Tuning a Vision-Language Model for Medical Image Captioning and SapBERT-Based Reranking for Concept Detection** # *Standardied concept, extraction.* `https://www.dei.unipd.it/~{}faggioli/temp/clef2025/paper_200.pdf`

Fine-Tuning a Vision-Language Model for Medical Image Captioning and SapBERT-Based Reranking for Concept Detection

Ronghao Pan; Tomás Bernal-Beltrán; José Antonio García-Díaz; Rafael Valencia-García
Departamento de Informática y Sistemas, Facultad de Informática, Universidad de Murcia

UMUTeam
UNIVERSIDAD DE MURCIA

Introduction and Motivation

- Automatic interpretation of radiology images is time-consuming and costly.
- Automatic systems can assist clinicians with efficiency and accuracy.
- Challenges:
  - Medical image captioning requires high semantic precision – even small mistakes matter.
  - Domain-specific terminology.
  - Integrating multimodal information from medical images and text.
- ImageCLEFmed Caption 2025:
  - Caption Prediction: Generate clinically meaningful captions.
  - Concept Detection: Link images to UMLS concepts.

Dataset

- ROCOv2 dataset:
  - Training: 80,091
  - Validation: 17,277
  - Test: 19,267
- Custom training dataset for the BERT-based reranking model:

label	train	val	total
positive	387.828	62.305	450.133
negative	2.275.520	508.996	2.784.516
total	2.663.348	571.301	3.234.649

Methodology

- Modular pipeline with two stages:
  - Caption Prediction: Fine-tuned BLIP vision-language model.
  - Concept Detection: Hybrid pipeline with entity recognition + SapBERT similarity + BERT reranker.
- Ensures:
  - High-quality clinical captions.
  - Standardized concept extraction.

Caption Prediction

- Model: BLIP (ViT encoder + LM decoder).
- Fine-tuned on ROCOv2.
- Training: 5 epochs, batch size 4, lr=2e-5.
- Custom Relevance Metric: avg(BERTScore + ROUGE-1 + Exact Match).
- Best checkpoint: epoch 4.

Proposed Approach

Medical Image
Caption Prediction (BLIP finetuned)
Model Selection via Relevance Metric (BERTScore + ROUGE + Exact Match)
Generated Caption
Entity Recognition (SciSpacy)
Concept Retrieval (SapBERT Embeddings)
Reranking Classifier (BERT-based)
Final Output: Caption + CUIs
Caption
Concepts

Concept Detection

- Entity Recognition: SciSpacy biomedical NER.
- Concept Retrieval: SapBERT embeddings + cosine similarity (top-20 candidates).
- Reranking: BERT-based classifier (3.2M entity-concept pairs, imbalanced data).
- Inference: entity → candidates → reranker → valid CUIs [illegible] UNKNOWN.

Results and Future works

- Caption Prediction:
  - 1st place overall (score: 0.3432)
  - High semantic similarity and relevance.
  - BLIP fine-tuning + relevance metric → best captions.
- Concept Detection:
  - F1=0.24, limited by class imbalance and semantic ambiguity.
  - Concept detection affected by class imbalance.
  - Lack of visual interpretability (concepts not tied to image regions).
- Future works:
  - End-to-end multimodal models (e.g. LLaVA).
  - Use of LLMs for semantic selection.
  - Better integration of image-text fusion for interpretability.

References

[1] C. Lange et al., “Feedforward Few-shot Species Range Estimation,” *arXiv*, Feb. 2025. DOI: 10.48550/arXiv.2502.14977. eprint: 2502.14977.