Knowledge Graph Construction from Radiology Reports using Large Language Models

Master Thesis

Name: Hanbin Chen
Matr.-Nr.: 421714
Study Program: Computer Science (Master)

First Supervisor: Prof. Dr. Stefan Decker
Second Supervisor: Prof. Dr. med. Daniel Truhn
Advisor(s): Yongli Mou M.Sc., Dr. Sulayman Sowe

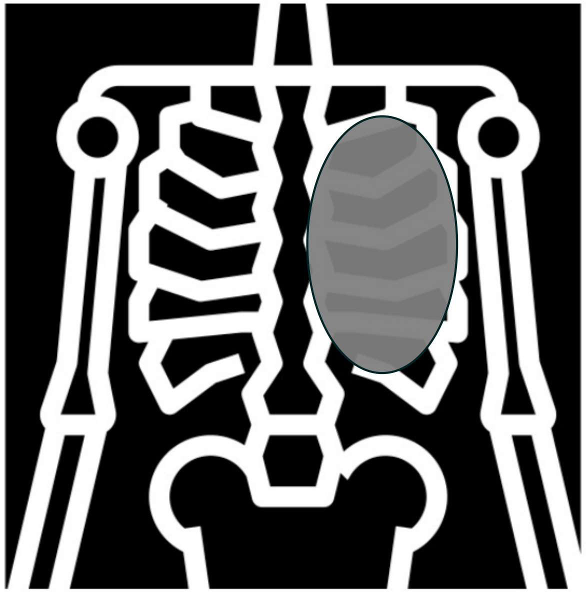
Computer Science 5 - Information Systems and Databases
RWTH AACHEN UNIVERSITY

Overview

- **Background & Related Work**
- Research Challenges & Objectives
- Methodology
- Results & Evaluation
- Discussion
- Conclusions & Future Work

Background & Related Work

Radiology Report



[report]

Increased right lower lobe opacity, concerning for infection.

No evidence of pneumothorax.

A report from MIMIC [1]

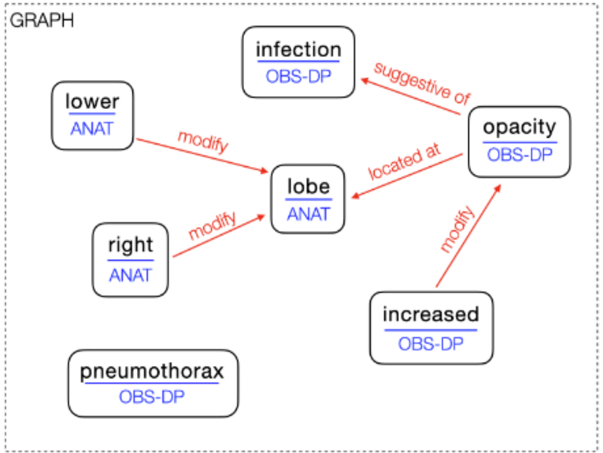
- Valuable clinical information
- But in flat textual form
- Difficult to analyze computationally

Background & Related Work

RadGraph as prior work

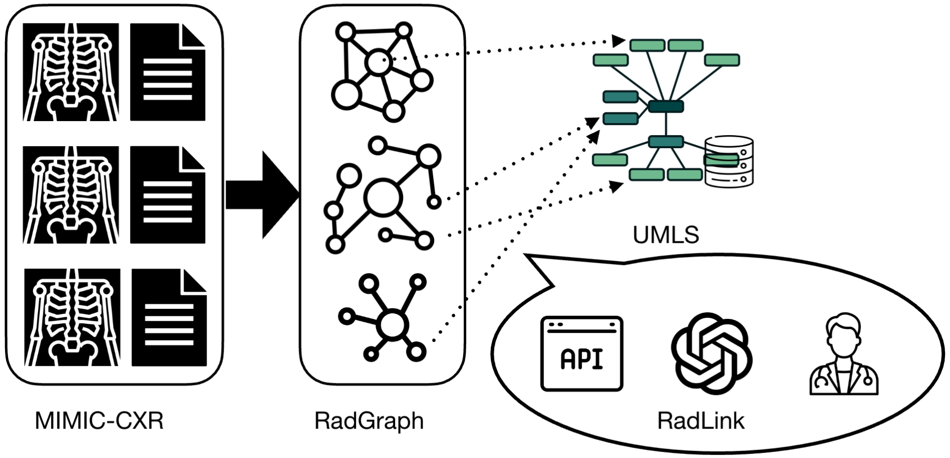
Named Entity Recognition, Relation extraction

[report]
1. *Increased right lower lobe opacity, concerning for infection.*
2. *No evidence of pneumothorax.*



A report and the associated knowledge graph [2]

Named Entity Normalization (NEN)



- Unified Medical Language System (UMLS). [7]
- Entity normalization
- Information consistency

Computer Science 5 - Information Systems and Databases
RWTH AACHEN UNIVERSITY

RadGraph Overview

- 500,000 chest radiograph reports
  - MIMIC-CXR
  - CheXpert
- **RadGraph**
  - **600 annotated reports**, but lacks normalization

Dataset	MIMIC-CXR	CheXpert	Total
train	425	0	425
dev	75	0	75
test	50	50	100
Total	550	50	600


```
{
    "p10/p10003412/s59172281.txt": {
        "text": "FINAL REPORT EXAMINATION : lungs ...",
        "entities": {
            "1": {
                "tokens": "lungs",
                "label": "ANAT-DP",
                "start_ix": 35,
                "end_ix": 35,
                "relations": []
            },
            "2": {},
            "3": {},
            "...": {}
        },
        "data_source": "MIMIC-CXR",
        "data_split": "dev"
    },
    "...": {},
    ...
}
```


An annotation from RadGraph [2]

Background & Related Work

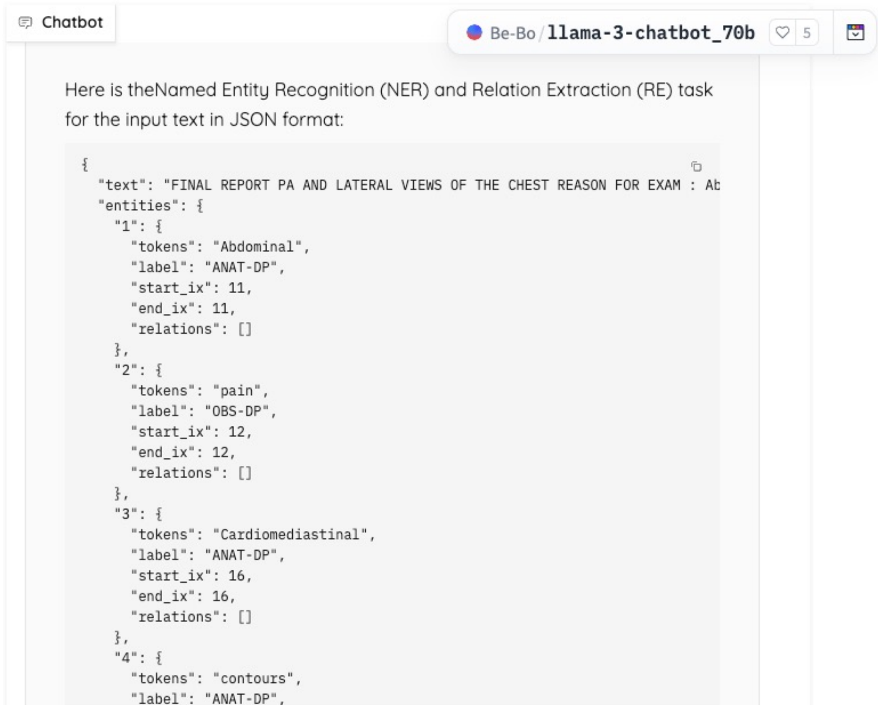
Prompt Engineering

<Prompt>
Please do **named entity recognition (NER)** and **relation extraction (RE)** task for following text:

'''
Increased right lower lobe opacity, concerning for infection.

No evidence of pneumothorax.
'''

output in JSON format,
as the **examples**:
"'......"', , "'......"', "'......"', "'......"', "'......"', ...



Example of Llama 3 70B

Overview

- Background
- **Challenges & Objectives**
- Methodology
- Results & Evaluation
- Discussion
- Conclusions & Future Work

Challenges & Objectives

Challenges

- **Terminological Complexity**
  – Medical terms with synonyms, abbreviations, and context-dependent variations [3]
- **Institutional Variation**
  – Inconsistent documentation practices across healthcare systems [8]

Requirements

- **Precise Boundary Detection**
  - Accurate entity span identification in medical text
- **Semantic Normalization**
  - Consistent mapping to standardized medical concepts

Research Questions

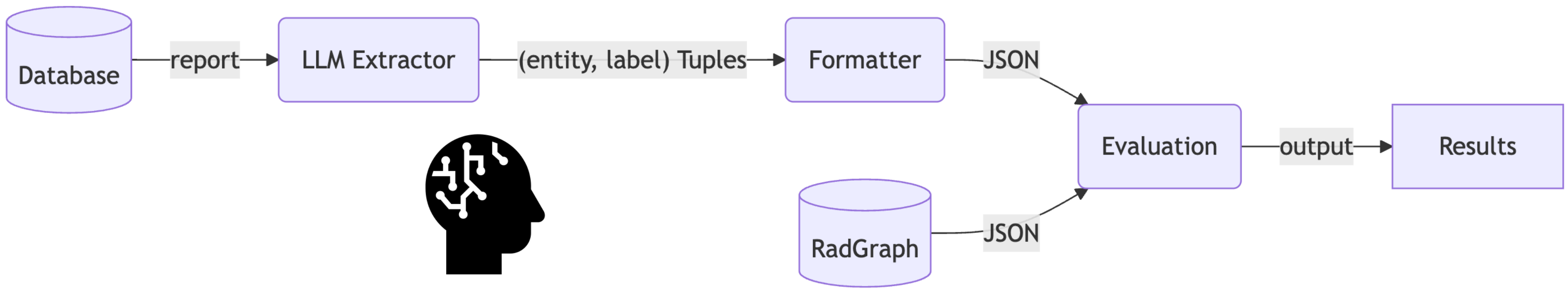
- **Entity Recognition:**
  – How do LLMs perform in medical entity recognition?
- **Entity Normalization:**
  – How can LLMs be effectively integrated with medical ontologies for consistent entity representation?
- **Resource Limitation:**
  – What strategies improve LLM's performance with limited labeled radiology data?

Overview

- Background
- Challenges & Objectives
- **Methodology**
  - **Named Entity Recognition (NER)**
  - Named Entity Normalization (NEN)
- Results & Evaluation
- Discussion
- Conclusions & Future Work

Methodology

Named Entity Recognition (NER) approach pipeline



- Modular pipeline design
- Component-level optimization and evaluation
- Comparing different model architectures and prompting strategies.

Methodology

Named Entity Recognition Prompt

You are a **radiologist** performing **clinical term extraction** from the FINDINGS, PA AND LATERAL CH
RADIOGRAPH and IMPRESSION sections in the radiology report.

Here a clinical term can be either anatomy or observation that is related to a finding or an impression.
The **anatomy** term refers to an anatomical body part such as a 'lung'.
The **observation** terms refer to observations made when referring to the associated radiology image
Observations are associated with visual features, identifiable pathophysiologic processes, or diagnostic
classifications. For example, an observation could be 'effusion' or description phrases like 'increased'.

You also need to assign a label to indicate whether the clinical term is **present, absent or uncertain**.

<OUTPUT>
*ANSWER: tuples separated by newlines. Each **tuple** has the format:*
*(<**clinical term text**>, <**label**: obs-dp | obs-da | obs-u | ana-dp>).*

If there are no extraction related to findings or impression, return ()
</OUTPUT>


```json
{
    "p10/p10003412/s59172281.txt": {
        "text": "FINAL REPORT EXAMINATION : lungs ...",
        "entities": {
            "1": {
                "tokens": "lungs",
                "label": "ANAT-DP",
                "start_ix": 35,
                "end_ix": 35,
                "relations": []
            },
            "2": {},
            "3": {},
            "...": {}
        },
        "data_source": "MIMIC-CXR",
        "data_split": "dev"
    },
    "...": {},
    ...
}
```

Methodology

Few-shots learning

You are a **radiologist** performing **clinical term extraction** from the FINDINGS, PA AND LATERAL CHEST RADIOGRAPH and IMPRESSION sections in the radiology report.

Here a clinical term can be either anatomy or observation that is related to a finding or an impression.
The **anatomy** term refers to an anatomical body part such as a 'lung'.
The **observation** terms refer to observations made when referring to the associated radiology image. Observations are associated with visual features, identifiable pathophysiologic processes, or diagnostic disease classifications. For example, an observation could be 'effusion' or description phrases like 'increased'.
You also need to assign a label to indicate whether the clinical term is **present, absent or uncertain**.

<OUTPUT>
ANSWER: tuples separated by newlines. Each **tuple** has the format:
(<**clinical term text**>, <**label**: obs-dp | obs-da | obs-u | ana-dp>).

If there are no extraction related to findings or impression, return ()
</OUTPUT>

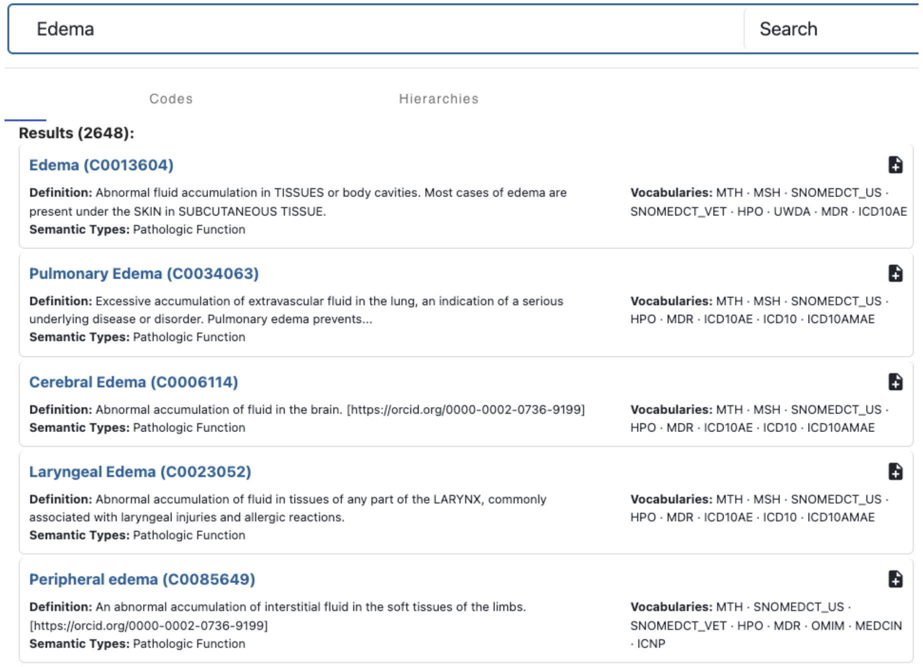
\+


```
{
    "p10/p10003412/s59172281.txt": {
        "text": "FINAL REPORT EXAMINATION : lungs ...",
        "entities": {
            "1": {
                "tokens": "lungs",
                "label": "ANAT-DP",
                "start_ix": 35,
                "end_ix": 35,
                "relations": []
            },
            "2": {},
            "3": {},
            "...": {}
        },
        "data_source": "MIMIC-CXR",
        "data_split": "dev"
    },
    "...": {},
    ...
}
```

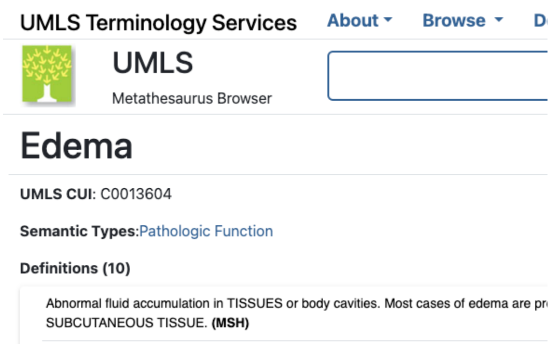

Annotations from RadGraph [2]

Methodology

Named Entity Normalization (NEN)



LLMs (GPT-4)



Entity from UMLS [7]

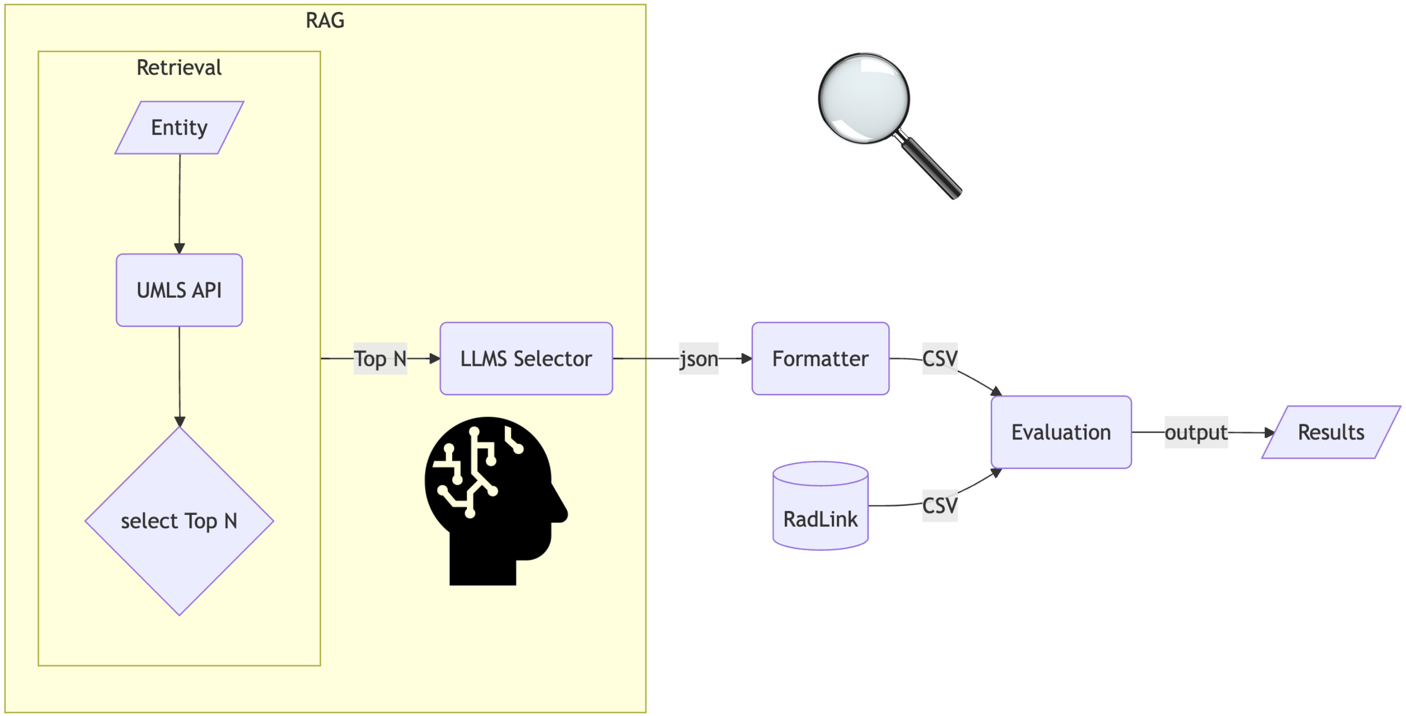
Morphological Matching

Semantic

Methodology

NEN Retrieval-Augmented Generation (RAG) approach pipeline

- Morphological Matching
  – String similarity
- RAG approach
  – Semantic understanding
- Hybrid approach



Named Entity Normalization (NEN) Prompt

You are **radiologist** in named entity normalization for medical terms using the UMLS ontology. Your task is to analyze the given entity and search results, then select the most appropriate normalized form or the most likely UMLS concept.

Search Results:

{ "results": [
{ "ui": "C0024109", "name": "Lung", "score": 0.95, "source": "UMLS" },
{ "ui": "C0225694", "name": "Lung structure", "score": 0.88, "source": "UMLS" },
{ "ui": "C1278908", "name": "Entire lung", "score": 0.82, "source": "UMLS" },
{ "ui": "C0819757", "name": "Structure of lung", "score": 0.75, "source": "UMLS" }
]
}

Overview

- Background
- Challenges & Objectives
- Methodology
- **Evaluation & Results**
- Discussion
- Conclusions & Future Work

NER Evaluation Matrix

- **Entity-level F1**

Scenario	Ground Truth	Prediction	Classification
I: Exact match	("right lung", ANATOMY), ("pneumonia", FINDING)	("right lung", ANATOMY), ("pneumonia", FINDING)	True Positive
II: Extra entities	("right lung", ANATOMY), ("opacity", -)	("right lung", ANATOMY), ("opacity", FINDING)	False Positive (for opacity)
III: Missing entities	("right lung", ANATOMY), ("pneumonia", FINDING)	("right lung", ANATOMY), ("pneumonia", -)	False Negative (for pneumonia)
IV: Correct position, incorrect label	("pneumonia", FINDING)	("pneumonia", DISEASE)	False Positive + False Negative
V: Partial overlap, same label	("right lower lobe of the lung", ANATOMY)	("right lower lobe", ANATOMY)	False Positive + False Negative
VI: Partial overlap, different label	("ground glass opacity", FINDING)	("ground", ANATOMY), ("glass opacity", FINDING)	False Positive + False Negative

Results & Evaluation

NER experimental results

- **Performance Gap**:
  - LLMs (max 61.9 F1) underperform compared to RadGraph benchmark (94.0 F1)
- **Few-Shot Learning**:
  - Performance scales significantly with example quantity
    - Llama 3.1: 1.8→61.9 F1 as shots increase from 1→100
- **Model Efficiency**:
  - GPT-4o demonstrates superior few-shot utilization
    - 55.9 F1 with 10 examples vs. Llama's 43.0 F1
- **Key Finding**:
  - LLMs show promising few-shot capabilities for clinical NER tasks

Model	MIMIC	CheXpert
RadGraph Benchmark	94.0	90.5

RadGraph Benchmark [2,4]

Model	MIMIC	CheXpert
Llama3.1:70B (0)	-	-
Llama3.1:70B (1)	1.8	0.8
Llama3.1:70B (5)	34.6	35.9
Llama3.1:70B (10)	43.0	48.4
Llama3.1:70B (100)	61.9	49.8
GPT-4o (10)	55.9	42.2

(n): number of random shots

Results & Evaluation

NEN Evaluation Matrix

Scenario	Ground Truth	Prediction	Classification
I: Exact match	CUI: C0024109 (Lung)	CUI: C0024109 (Lung)	**True Positive (TP)**
II: Incorrect concept	CUI: C0024109 (Lung)	CUI: C0263494 (Pulmonary tissue)	**False Positive (FP)**
III: Incorrectly assigned	N/A (Not normalizable)	CUI: C0024109 (Lung)	**False Positive (FP)**
IV: Failed to normalize	CUI: C0205148 (Surface)	N/A (Not normalizable)	**False Negative (FN)**
V: I: Exact match (TN)	N/A (Not normalizable)	N/A (Not normalizable)	**True Negative (TN)**

Results & Evaluation

NEN experimental results

- **Comparison**: morphological vs. semantic matching approaches

- **Dataset**: 1,250 radiology entities (73.44% UMLS-linkable)

- **Performance metrics:**
  - Accuracy: 95.84% vs. 66.08% (+29.76%)
  - Precision: 97.97% vs. 99.40% (-1.43%)
  - Recall: 96.24% vs. 54.04% (+42.20%)
  - F1: 97.10% vs. 70.01% (+27.09%)

Metric	Morphological Matching	Semantic Matching
Accuracy	0.6608	0.9584
Precision	0.9940	0.9797
Recall	0.5404	0.9624
F1 Score	0.7001	0.9710

TABLE I
EVALUATION RESULTS OF MORPHOLOGICAL MATCHING AND SEMANTIC MATCHING

- **Conclusion**: LLM-based semantic matching significantly enhances normalization quality

Results & Evaluation

RadLink Dataset

1	name	ui	normalized_name
2	Lungs	C0024109	Lungs
3	clear	C2963144	clear
4	Normal	C0205307	Normal
5	cardiomediastinal		
6	hilar	C0205150	hilar
7	silhouettes		
8	pleural	C1522720	pleural
9	surfaces		
10	Endotracheal	C0599554	Endotracheal
11	tube	C1561954	tube
12	tip	C3282898	tip
13	approximately a 4.6 cm		

......

1238	Both	C1706086	Both
1239	no	C1298908	no
1240	sequela	C0543419	Sequela of disorder
1241	traumatic event	C4751223	traumatic event
1242	amount of		
1243	AC		
1244	joint	C0022417	joint
1245	resorption	C2985494	resorption
1246	Overlying		
1247	EKG	C0013798	Electrocardiogram
1248	Minimally		
1249	Bronchovascular	C2326513	Bronchovascular bundle
1250	Chronic	C0205191	Chronic
1251	separation		

Overview

- Background
- Challenges & Objectives
- Methodology
- Results & Evaluation
- **Discussion**
- Conclusions & Future Work

Discussion

Error from NER Task, RadGraph annotation inconsistencies affecting downstream tasks

- Errors in entity recognition (NER) affect NEN accuracy.
- Complex clinical language and abbreviations create normalization challenges. [6]
- Meaningless words occasionally misclassified as entities (e.g., '1', 'the’)

NEN in Radiology

- Radiologists tend to use precise, consistent terminology.
- Hight F1 score

LLM stability issues and performance variability

- Difficulty reproducing consistent results across runs
- Intermittent failures resolved through re-execution

Ontology coverage limitations in radiology

- Gaps in UMLS coverage for radiology-specific terminology
  - (about 75% UMLS-linkable, 25% remaining)
- Need for specialized ontologies like RadLex to supplement coverage [2]

Overview

- Background
- Challenges & Objectives
- Methodology
- Results & Evaluation
- Discussion
- **Conclusions & Future Work**

Conclusions & Future Work

Conclusions & Contributions

- **Comparative model evaluation**:
  - Benchmarked specialized biomedical vs. general-purpose LLMs (GPT-4, LLaMA 3)
- **Enhanced metrics framework**:
  - Developed precise evaluation protocols for radiological NER/NEN assessment
- **UMLS-aligned dataset**:
  - Created manually annotated corpus (1,250 entities) with standardized concept mapping
- **Retrieval-augmented generation**:
  - Implemented hybrid architecture combining LLMs with medical ontologies
- **Clinical deployment insights**:
  - Evaluated optimal model selection across varied data availability scenarios

Conclusions & Future Work

Future Work

- **Agent-Based Systems**:
  - Specialized collaborative agents with self-monitoring capabilities for error correction

- **Advanced RE Techniques**:
  - Cross-sentence relationship extraction with uncertainty quantification

- **Multimodal Integration**:
  - Incorporate imaging data alongside reports for enhanced entity disambiguation

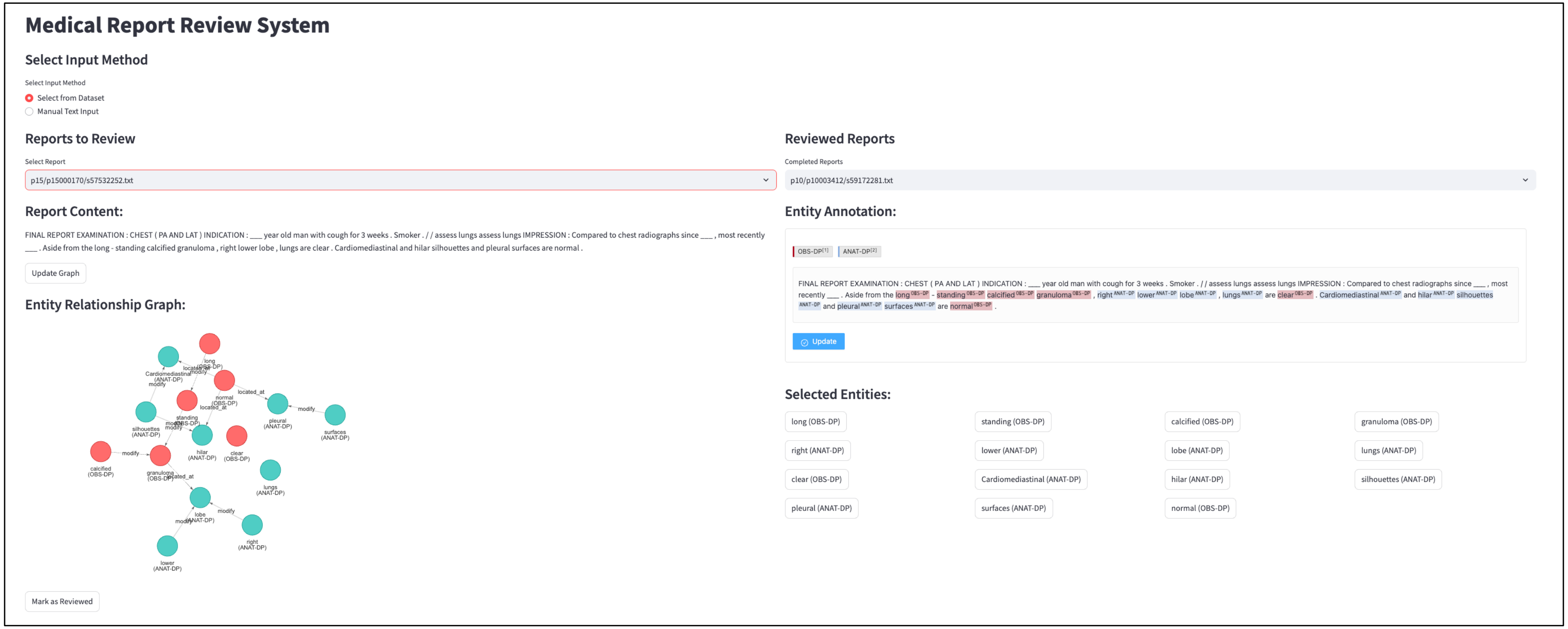
References

1. Johnson, Alistair EW, et al. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports." *Scientific data* 6.1 (2019): 317.
2. Saahil Jain et al., “RadGraph: Extracting Clinical Entities and Relations from Radiology Reports” (arXiv, August 29, 2021), https://doi.org/10.48550/arXiv.2106.14463.
3. Surabhi Datta, Jordan Godfrey-Stovall, and Kirk Roberts, “RadLex Normalization in Radiology Reports,” *AMIA Annual Symposium Proceedings* 2020 (January 25, 2021): 338–47.
4. Zexuan Zhong and Danqi Chen, “A Frustratingly Easy Approach for Entity and Relation Extraction” (arXiv, March 23, 2021), http://arxiv.org/abs/2010.12812.
5. Yi Luan et al., “A General Framework for Information Extraction Using Dynamic Span Graphs” (arXiv, April 5, 2019), http://arxiv.org/abs/1904.03296.
6. Mujeen Sung et al., “BERN2: An Advanced Neural Biomedical Named Entity Recognition and Normalization Tool,” ed. Karsten Borgwardt, *Bioinformatics* 38, no. 20 (October 14, 2022): 4837–39, https://doi.org/10.1093/bioinformatics/btac598.
7. Bodenreider, Olivier. "The unified medical language system (UMLS): integrating biomedical terminology." Nucleic acids research 32.suppl_1 (2004): D267-D270.
8. Evan French and Bridget T. McInnes, “An Overview of Biomedical Entity Linking throughout the Years,” *Journal of Biomedical Informatics* 137 (January 2023): 104252, https://doi.org/10.1016/j.jbi.2022.104252.

Thank you
for your attention!

Demo

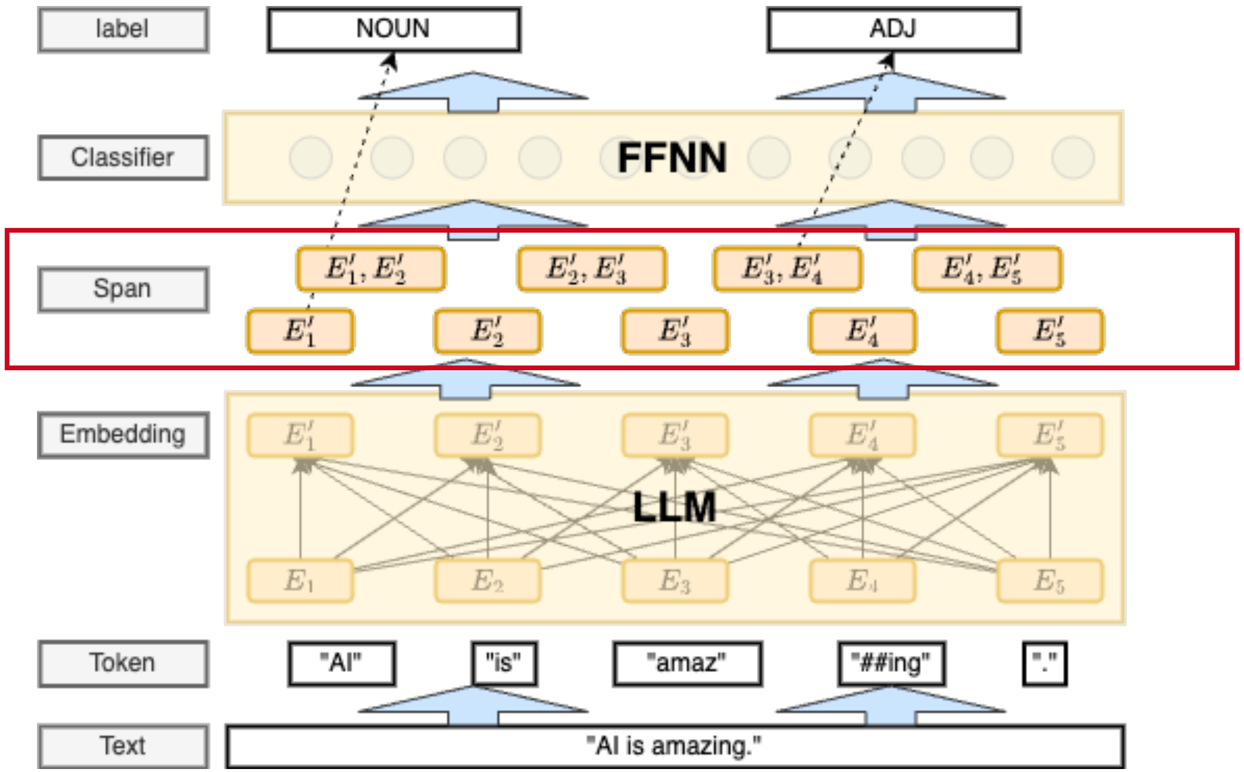
Demo Overview



Background and Related Work

Span Representation in NER task

- NER Task Definition new:
  - X : Sentence consisting of n tokens
    - $X = [x_1, x_2, \ldots, x_n]$
  - ***S*** **: all possible spans**
    - $\boldsymbol{S} = \{\boldsymbol{s_1}, \boldsymbol{s_2}, \ldots, \boldsymbol{s_m}\}$
      where $\boldsymbol{s_i} = [\boldsymbol{x_{start}}, \boldsymbol{x_{end}}]$
  - L : set of pre-defined entity labels
    - $L = \{l_1, l_2, \ldots, l_k\}$
  - O : tuple set as output
    - $O_e = \{(s, l) \mid s \in S, l \in L\}$
- Tokens → Spans
- Spans → Labels (Classification)



Example of Span Classification