Knowledge Graph Construction from Radiology Reports with LLMs

Master Thesis Proposal

Name: Hanbin Chen
Matr.-Nr.: 421714
Study Program: Computer Science (Master)
Supervisor(s): Prof. Dr. Stefan Decker
Advisor(s): Yongli Mou, M.Sc., Dr. Sulayman Sowe

Computer Science 5 - Information Systems and Databases
RWTH AACHEN UNIVERSITY

Overview

- **Introduction**
- Background and Related Work
- Problem Statements
- Methods
- Evaluation Plan
- Project Plan

Computer Science 5 - Information Systems and Databases
RWTH AACHEN UNIVERSITY

Introduction

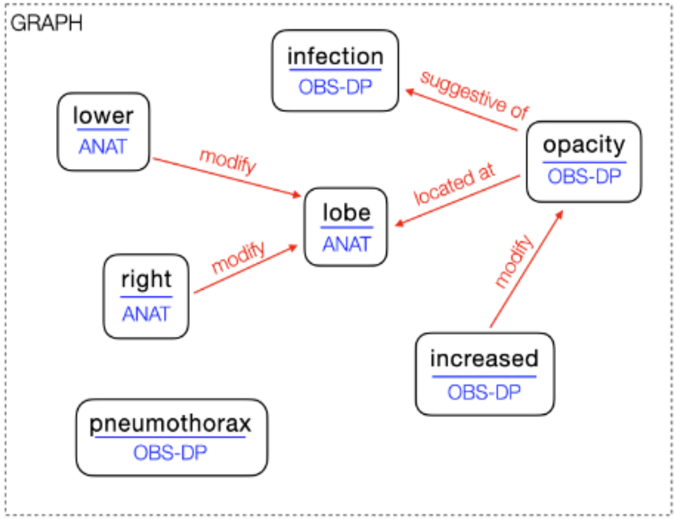
Motivation

- Radiology Report:
  - Unstructured textual reports
  - Extensive medical terminology
  - Difficulty in extracting and analyzing critical clinical information
  - Automated information extraction needed

- Knowledge Graph:
  - Intuitive visualization and expression than text
  - Connects entities and relationships
  - Structured form
  - Quick access to patient conditions

[report]

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



A report and the associated knowledge graph [1]

[1] 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.

Overview

- Introduction
- **Background and Related Work**
- Problem Statements
- Methods
- Evaluation Plan
- Project Plan

Background and Related Work

Task definition

Knowledge Graph Construction (KGC)

- Named Entity Recognition (NER)
- Relation Extraction (RE)
- *(Event Extraction)*
- Named Entity Normalization (NEN)

Information Extraction (IE) [1]

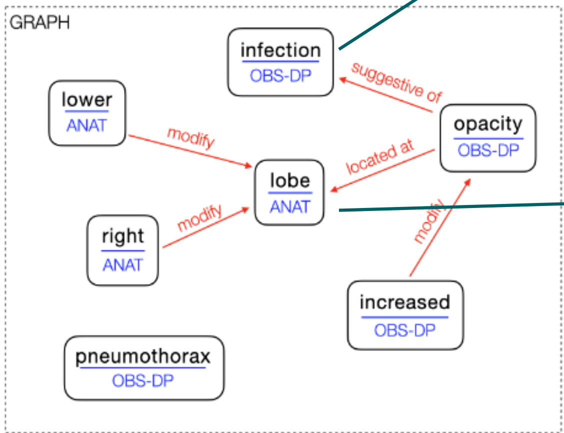
[1] David Wadden et al., “Entity, Relation, and Event Extraction with Contextualized Span Representations” (arXiv, September 9, 2019), http://arxiv.org/abs/1909.03546.

Background and Related Work

Thesis goal

[report]

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



Preferred Name:	infection
RadLex ID:	RID3710
PURL:	http://www.radlex.org/RID/RID3710
Definition:	Invasion and multiplication of microorg… result in local cellular injury. A local infe… subacute, or chronic clinical infection or… microorganisms gain access to the lym…
Preferred_name_German:	Infektion
UMLS_ID:	C0021311
May_Cause:	http://radlex.org/RID/RID35403, http:/… http://radlex.org/RID/RID35451, http:/…
Source:	Playbook
Is_A:	http://radlex.org/RID/RID3381

Term ”infection” in RadLex [3]

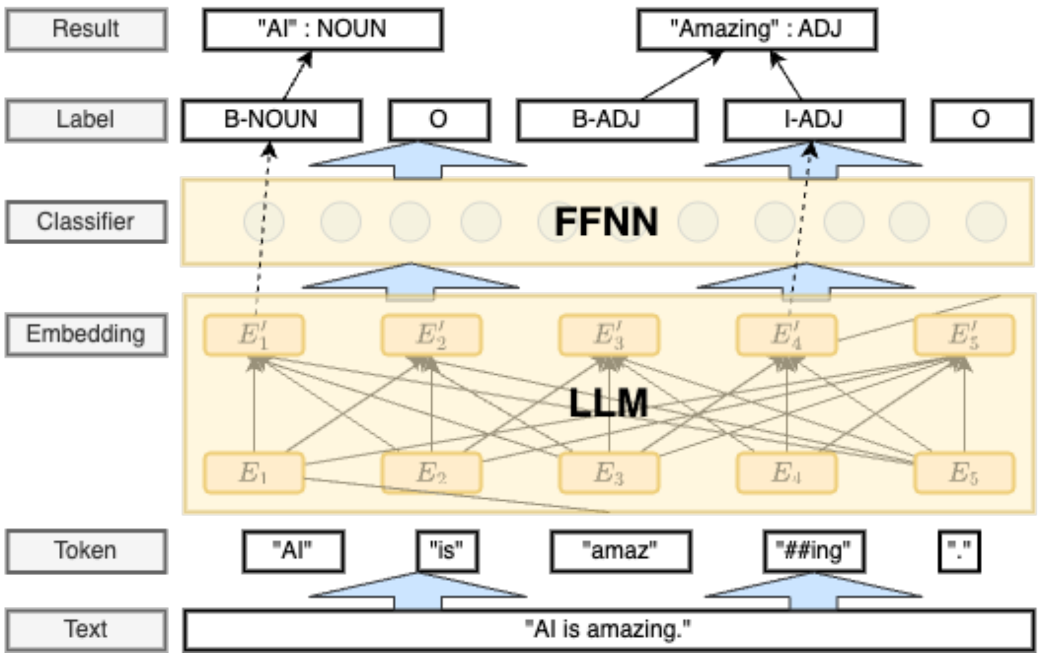
Preferred Name:	lobe
RadLex ID:	RID5967
PURL:	http://www.…
Preferred_name_German:	Lappen
Is_A:	http://radlex…

Term ”lobe” in RadLex [3]

[1] 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.
[2] https://uts.nlm.nih.gov/uts/umls/concept/C0796494
[3] https://radlex.org/RID/RID5967

NER with LLMs (BERT token embedding)

- NER Task Definition:
  - X : Sentence consisting of n tokens
    - $\mathrm{X} = [x_1, x_2, \dots, x_n]$
  - L : set of pre-defined entity labels / types
    - $L = \{l_1, l_2, \dots, l_k\}$
    - IOB tagging schema (Inside, Outside, Begin)
  - O_e : tuple set as output
    - $O_e = \{(\mathrm{x_{start}}, \mathrm{x_{end}}, \mathrm{l}) \mid \mathrm{x_{start}}, \mathrm{x_{end}} \in \mathrm{X}, \mathrm{l} \in \mathrm{L}\}$
- Tokens → Labels (Classification)



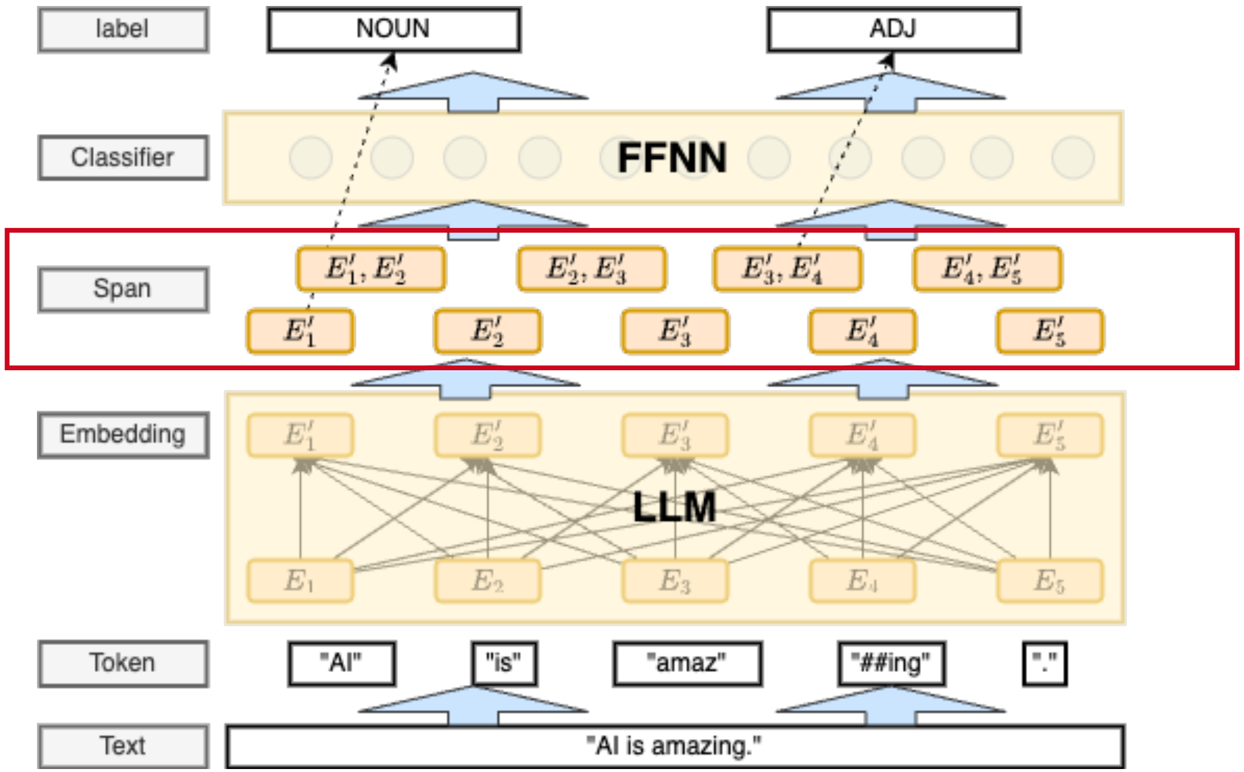
Example of Token Classification in LLMs

Computer Science 5 - Information Systems and Databases
RWTH AACHEN UNIVERSITY

Background and Related Work

Span Representation in NER task

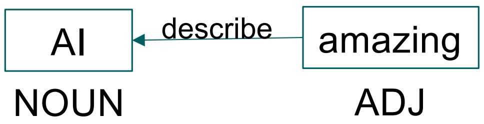
- NER Task Definition new:
  - X : Sentence consisting of n tokens
    - $\mathrm{X} = [x_1,\ x_2,\ \dots,\ x_n]$
  - $\boldsymbol{S}$ **: all possible spans**
    - $\boldsymbol{S} = \{\boldsymbol{s_1},\ \boldsymbol{s_2},\ \dots,\ \boldsymbol{s_m}\}$
      where $\boldsymbol{s_i} = [\boldsymbol{x_{start}},\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



Example of Span Classification

Relation Extraction

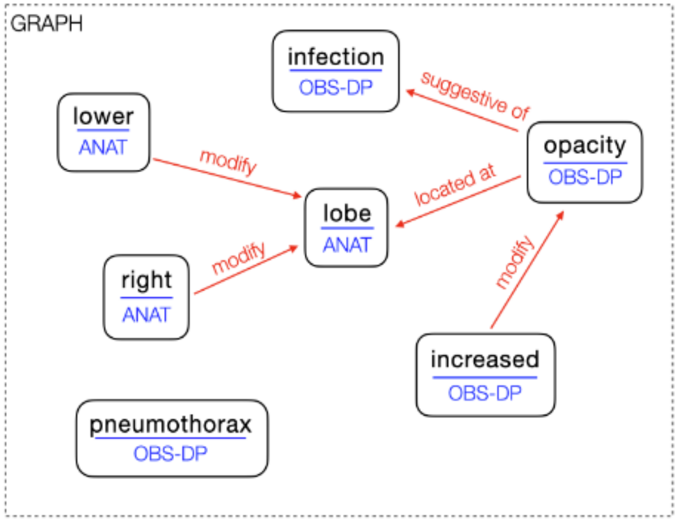
- RE Task Definition:
  - S : all possible spans
    - $\mathrm{S} = \{s_1, s_2, \dots, s_m\}$
  - R : set of pre-defined relation types.
    - $\mathrm{R} = \{r_1, r_2, \dots, r_p\}$
  - O_r : tuple set as output
    - $O_\mathrm{r} = \{(\mathrm{s_i}, \mathrm{s_j}, r) \mid s_i, s_j \in \mathrm{S}, r \in \mathrm{R}\}$



- Complexity
  - Comparison between each pair leads to $O(n^2)$.

[report]

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



A report and the associated knowledge graph [1]

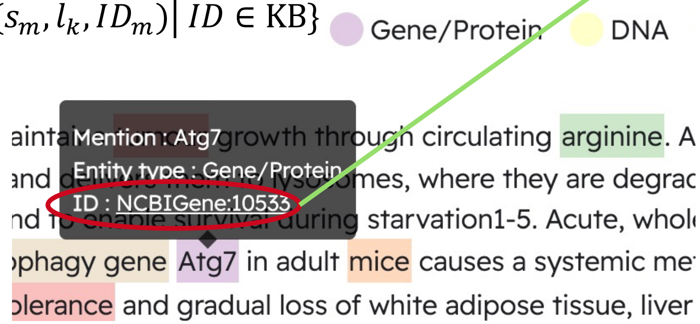
[1] Saahil Jain et al., "RadGraph: Extracting Clinical Entities and Relations from Radiology Reports" (arXiv, August 29, 2021), https://doi.org/10.48550/arXiv.2106.14463.

Background and Related Work

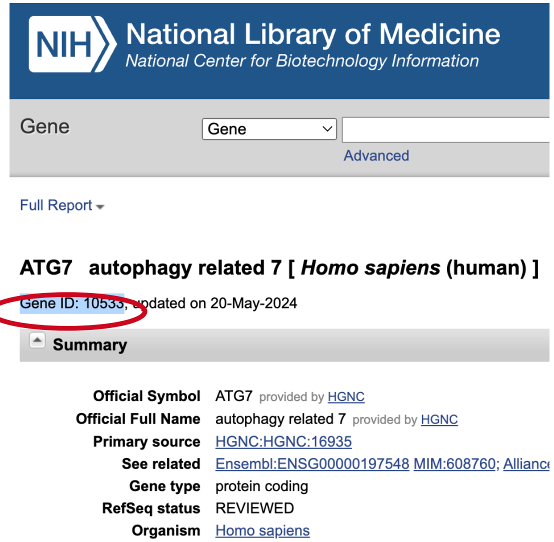
Named Entity Normalization (NEN)

- Linking to Standard Knowledge Base (KB) / Dictionary / Database

- NEN Task Definition:
  - X : set of spans with labels
    - $\mathrm{X} = \{(s_i, l_1),\ (s_j,\ l_2), \dots ,\ (s_m, l_k)\}$
  - O_1 : tuple set as output
    - $O_1 = \{(s_i, l_1, ID_i),\ (s_j,\ l_2, ID_j), \dots ,\ (s_m, l_k, ID_m) \mid ID \in \mathrm{KB}\}$



Example of a tool named BERN2 [1]



Term “Atg7” in NCBI [2]

[1] 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.
[2] https://www.ncbi.nlm.nih.gov/gene/10533

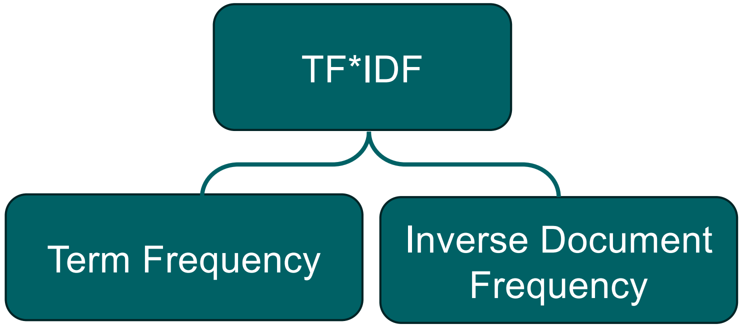
Background and Related Work

NEN Approaches

- **Rule-Based Methods**
  - Dictionary Matching
  - Regular Expressions
  - Morphological Analysis
    - prefixes, suffixes, roots

<Levenshtein distance>
hormonerelated protein
Hormone-related protein

- **Learning-Based Methods**
  - TF-IDF Representations
  - Word Embeddings
    - Word2Vec
  - Deep Learning Models
    - BERT for contextual information



Overview

- Introduction
- Background and Related Work
- **Problem Statements**
- Methods
- Evaluation Plan
- Project Plan

Computer Science 5 - Information Systems and Databases
RWTH AACHEN UNIVERSITY

Problem Statements

Challenges

- Limited Resource: only 500 annotated reports as development data from board-certified radiologist.
- Traditional rule-based methods for NEN can only perform character-level morphological matching.
  - LLMs embedding improved quality of representations, making better semantic matching possible.

Research Questions

- RQ1: How is the performance of LLMs with different architecture in solving NER & RE tasks?
- RQ2: How could LLMs be effectively integrated into the NEN task to achieve efficient and accurate entity retrieval?
- RQ3: How is the performance of LLMs in settings with very limited labeled radiology reports using semi-supervised learning?

Overview

- Introduction
- Background and Related Work
- Problem Statements
- **Methods**
- Evaluation Plan
- Project Plan

Methods

Overview

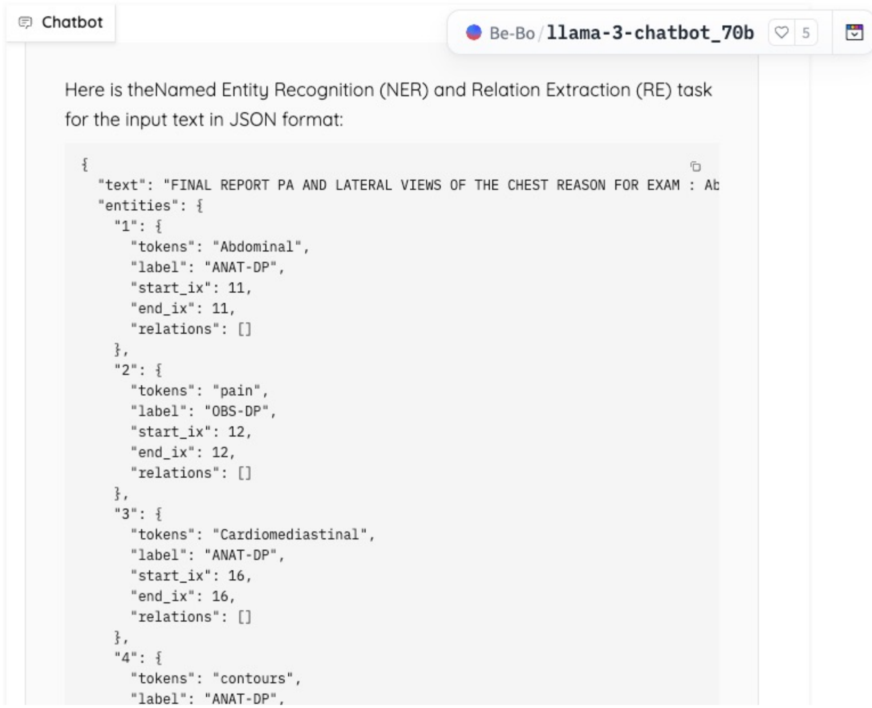
- **NER & RE Tasks**
  - Using Decoder's Generative Capability with prompt engineering
    - Decoder-only model (e.g., LLaMA 3)
    - Encoder-decoder model (e.g., T5)
  - Using LLMs' Embeddings with FFNN for Classification
    - Decoder-only model (e.g., LLaMA 3)
    - Encoder-only model (e.g., BERT-based model)

- **NEN**
  - Constructing a vector database from the original knowledge base using Embeddings from LLMs
  - Using semantic similarity for an efficient and accurate entity retrieval
    - Similarity search
    - Hybrid search (combine string match)

Methods

Prompt Engineering

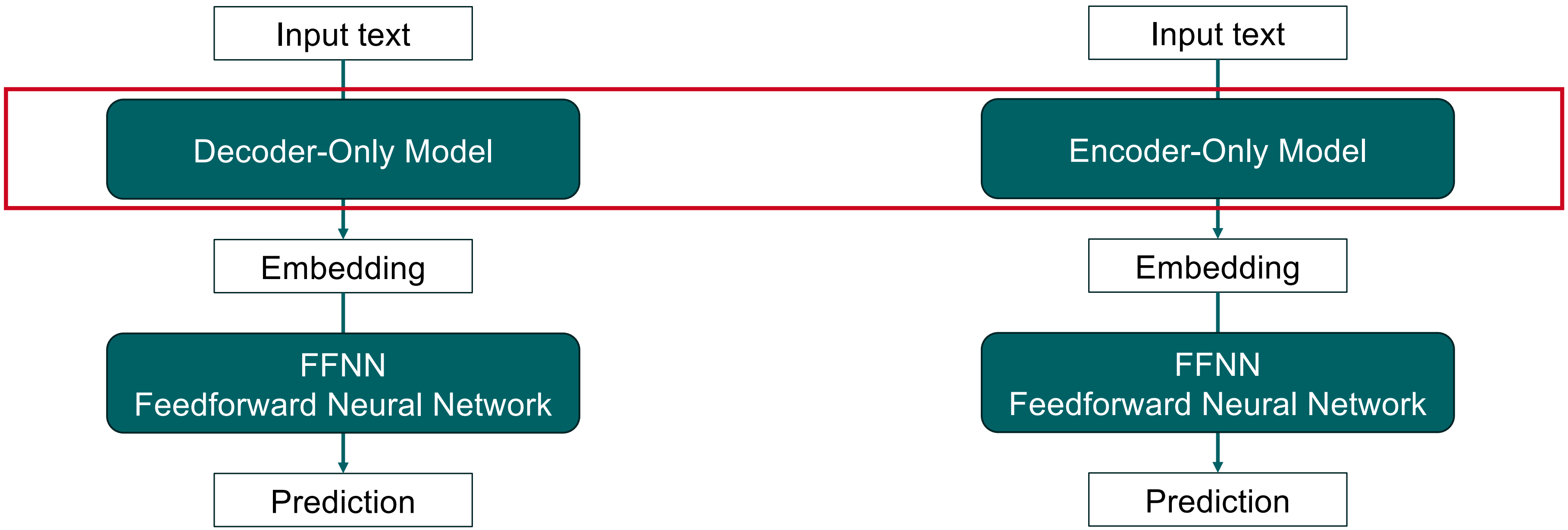
<Prompt>
Please do named entity recognition (NER) and relation extraction (RE) task for following text:
'''
FINAL REPORT PA AND LATERAL VIEWS OF THE CHEST REASON FOR EXAM : Abdominal pain . Cardiomediastinal contours are normal . The lungs are clear . There is no pneumothorax or pleural effusion . There are mild degenerative changes in the thoracic spine . IMPRESSION : No evidence of pneumonia .
'''
output in JSON format, as the examples:
'''...... ''', , '''......''', '''......''', '''......''', '''......''', '''......'''



Example of Llama 3 70B

Methods

Decoder-Only Model **VS** **Encoder-Only Model**



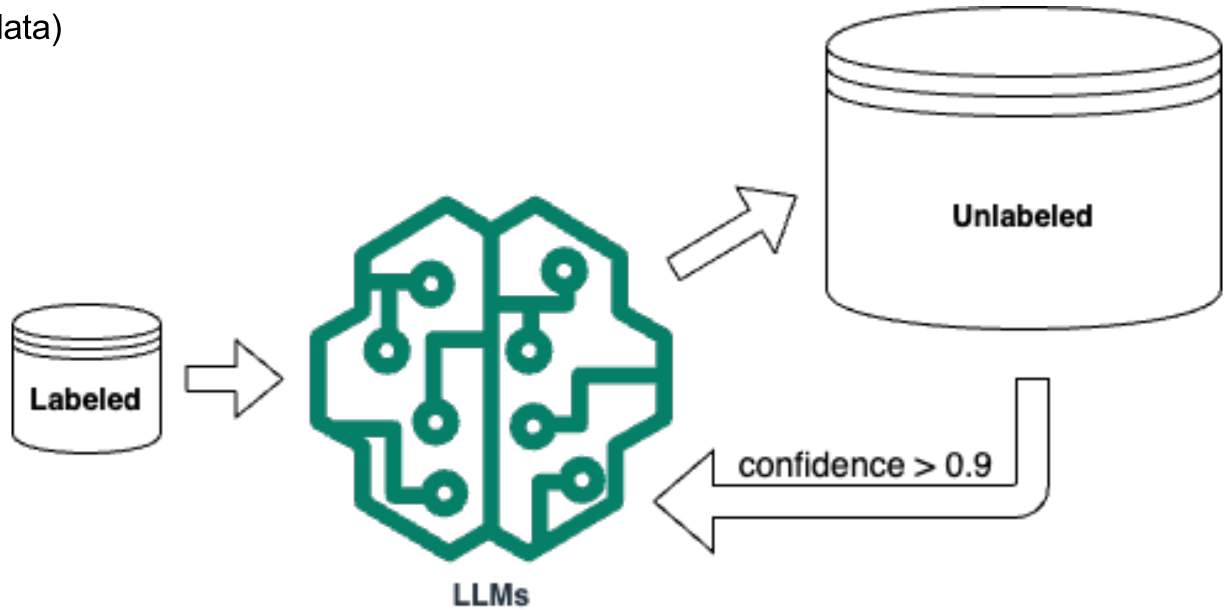
Methods

Finetune with Semi-supervised self-training

- **Dataset**
  – 500 labeled report (very limited training data)
  – 220,763 raw report

- **Framework**
  – PURE [1] as pipeline approach
  – DyGIE [2] as dynamic graph approach

- **LLMs**
  – BlueBERT
  – ClinicalBERT
  – Bio+ClinicalBERT



[1] Zexuan Zhong and Danqi Chen, "A Frustratingly Easy Approach for Entity and Relation Extraction" (arXiv, March 23, 2021), http://arxiv.org/abs/2010.12812.
[2] Yi Luan et al., "A General Framework for Information Extraction Using Dynamic Span Graphs" (arXiv, April 5, 2019), http://arxiv.org/abs/1904.03296.

Methods

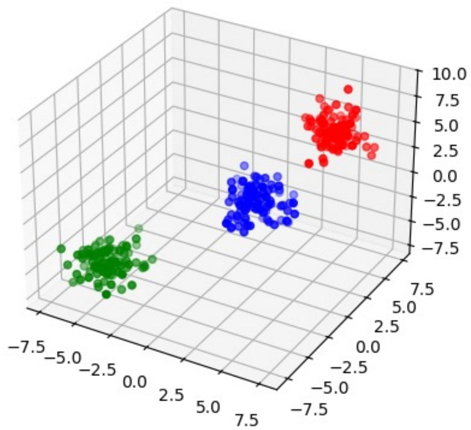
Construct a vector DB and apply advanced retrieval algorithms

- Encoding all concept with BERT in vector database N.
  - e.g., lung, pulmo

- String-level morphologically similarity of mention n and item $n \in N$.
  - $\boldsymbol{S}_{morphologically}(m,n) = f(m,n)$

- BERT embedding representation encodes the semantic similarity.
  - $\boldsymbol{S}_{semantic}(m,n) = f_{embedding}(m,n)$
  - Maximum inner product search (MIPS) for retrieving the nearest synonym

- Hybrid similarity function S
  - $\boldsymbol{S}_{Hybrid}(m,n) = S_{morphological}(m,n) + S_{semantic}(m,n)$



Example of vector representation

Overview

- Introduction
- Background and Related Work
- Problem Statements
- Methods
- **Evaluation Plan**
- Project Plan

Evaluation Plan

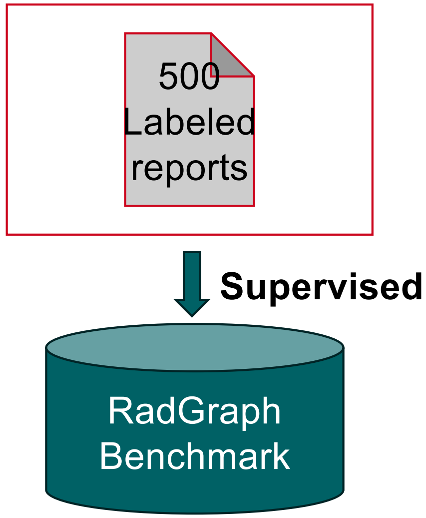
Experimental Settings

- **Baseline**
  - RadGraph Benchmark, a BERT-based state-of-art model, with highest F1 scores of NER & RE in radiology report.

- **Datasets for NER & RE**
  - Development dataset: 500 annotated radiology reports from the MIMIC-CXR dataset
  - Test dataset: 100 annotated radiology reports MIMIC-CXR and CheXpert dataset

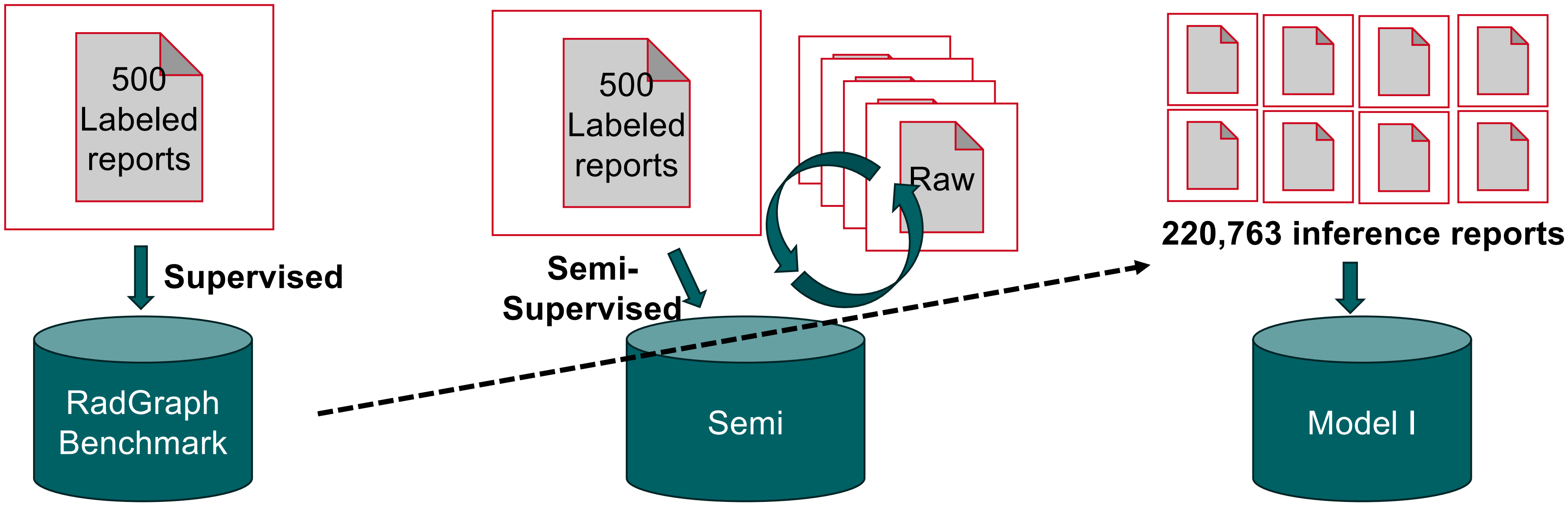
- **Evaluation Metrics**
  - Entity F1
    - $(start,\ end,\ label)$
  - Relation F1
    - $(entity_1,\ entity_2,\ relation)$

- $recall = \frac{TP}{TP+FN}$
- $precison = \frac{TP}{TP+FP}$
- $F1 = 2 * \frac{reall+precision}{recall * precision}$



Evaluation

Semi-supervised self-training



Evaluation

Retrieval Entity in Vector DB of BERT Embedding

- **Knowledge Base**
  - RadLex (RadLex radiology lexicon)
  - 46,657 concepts

- **Test set**
  - Verified annotated report as gold normalized concept
    - have inquired, not received a response yet
    - (maybe need help from radiologist from Uniclinik Aachen)

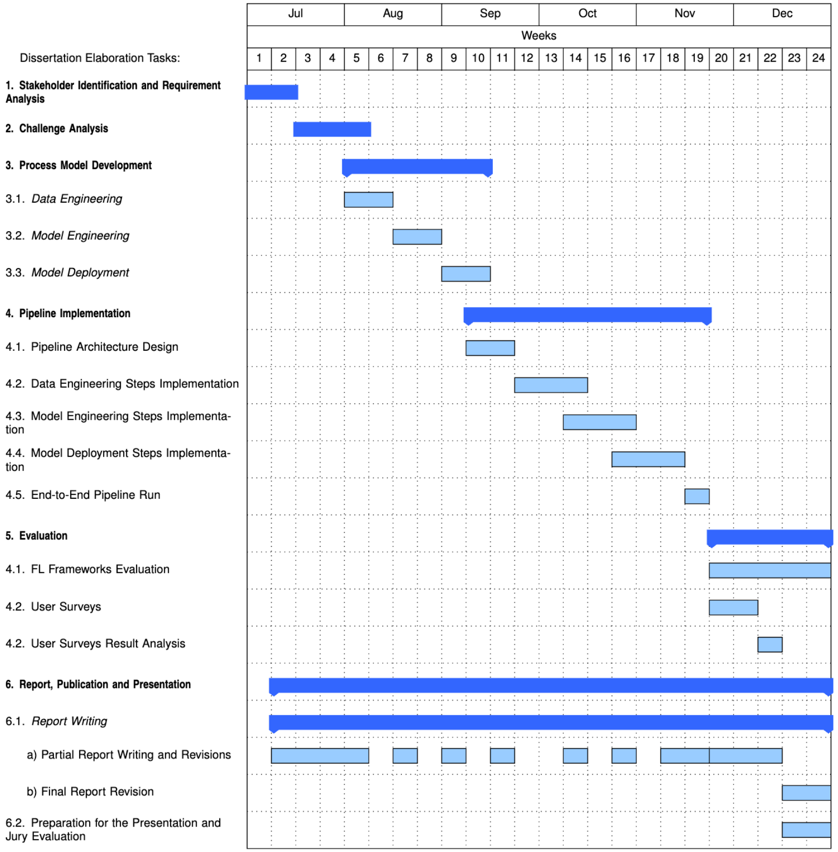
- **Evaluation Metrics**
  - F1 Score
    - (entity, RID), e.g., (lobe, RID 5967)

- $recall = \frac{TP}{TP+FN}$
- $precison = \frac{TP}{TP+FP}$
- $F1 = 2 * \frac{reall+precision}{recall * precision}$

- **Baseline**
  - A deep learning-based methods with F1-score 0.7593 [1]
    - No access to source code and dataset

[1] Surabhi Datta, Jordan Godfrey-Stovall, and Kirk Roberts, "RadLex Normalization in Radiology Reports," *AMIA Annual Symposium Proceedings* 2020 (January 25, 2021): 338–47.

Project Plan



References

1. 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.
2. Surabhi Datta, Jordan Godfrey-Stovall, and Kirk Roberts, “RadLex Normalization in Radiology Reports,” *AMIA Annual Symposium Proceedings* 2020 (January 25, 2021): 338–47.
3. Zexuan Zhong and Danqi Chen, “A Frustratingly Easy Approach for Entity and Relation Extraction” (arXiv, March 23, 2021), http://arxiv.org/abs/2010.12812.
4. Yi Luan et al., “A General Framework for Information Extraction Using Dynamic Span Graphs” (arXiv, April 5, 2019), http://arxiv.org/abs/1904.03296.
5. 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.
6. Surabhi Datta, Jordan Godfrey-Stovall, and Kirk Roberts, “RadLex Normalization in Radiology Reports,” *AMIA Annual Symposium Proceedings* 2020 (January 25, 2021): 338–47.
7. 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.
8. Shang Gao et al., “A Pre-Training and Self-Training Approach for Biomedical Named Entity Recognition,” ed. Nicolas Fiorini, *PLOS ONE* 16, no. 2 (February 9, 2021): e0246310, https://doi.org/10.1371/journal.pone.0246310.

Thank you
for your attention!

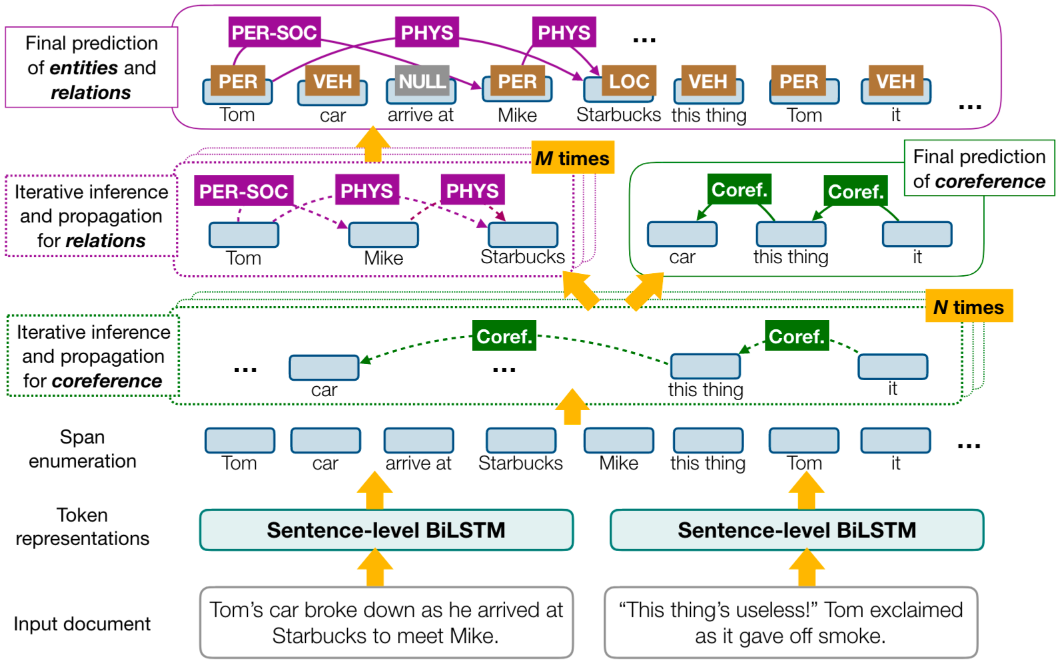
Methods

NER & RE Framework

- Dynamic Graphs Information Extraction (DyGIE)
  - Take current best guess of the graph then update
    - Entity as node
    - Relation as arc
  - Jointly extraction
  - End-to-end

- Princeton University Relation Extraction (PURE)
  - Pipeline extraction
  - Information from NER is helpful for RE
  - “Disney” refers to a person or an organization before trying to understand the relations. [1]

- Limitation: square complexity for relation extraction



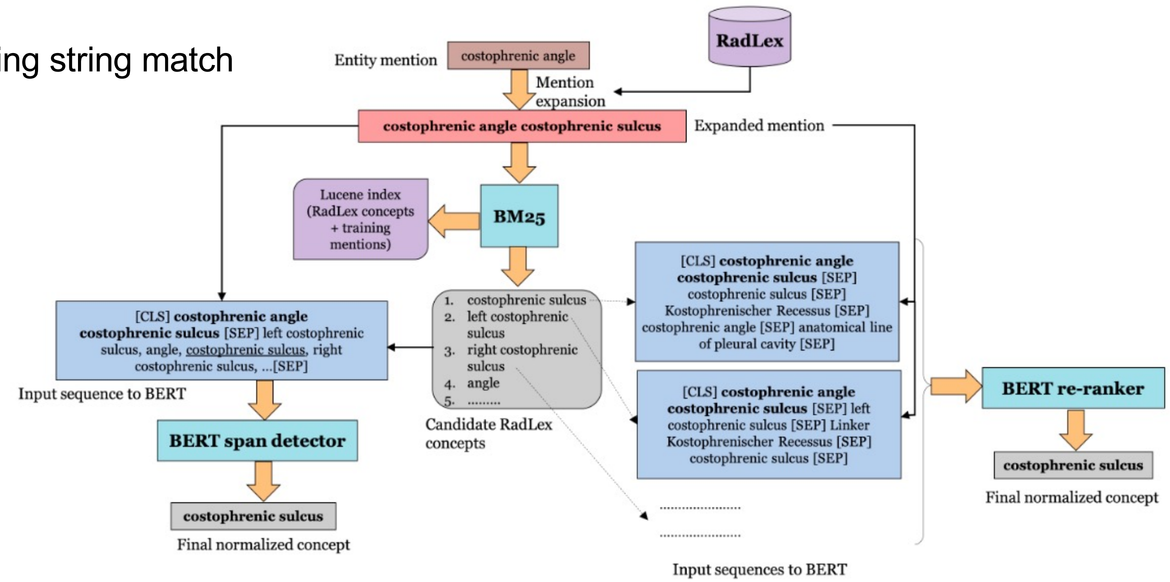
NER & ER jointly with DyGIE [1]

[1] Yi Luan et al., “A General Framework for Information Extraction Using Dynamic Span Graphs” (arXiv, April 5, 2019), http://arxiv.org/abs/1904.03296.

Background and Related Work

Named Entity Normalization (NEN)

- Approaches for Solving NEN task
  - Previously as Matching problem, solve it using string match
  - Now for LLMs as solving Mapping problem



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.
Surabhi Datta, Jordan Godfrey-Stovall, and Kirk Roberts, “RadLex Normalization in Radiology Reports,” *AMIA Annual Symposium Proceedings* 2020 (January 25, 2021): 338–47.