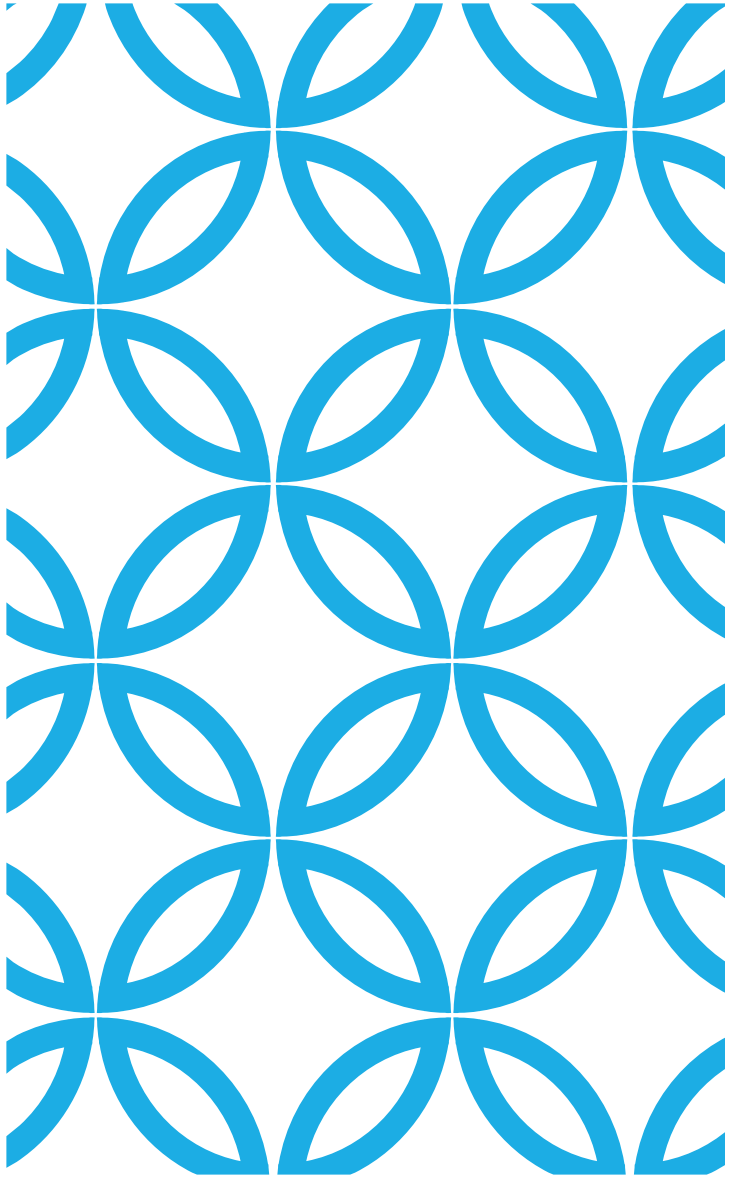


TRANSLATIONAL UMLS VOCABULARY ALIGNMENT

By:
Bernal Jimenez

MOTIVATION

- UMLS is a medical knowledge base which combines over 200 medical terminologies.
  - Valuable repository for medical knowledge and useful resource for inter-operability
- UMLS contains
  - ~4 million concepts
  - ~16 million atoms (1 atom = 1 string from a specific source)
- UMLS grows larger every year
  - Approx. a million atoms are added every year

MOTIVATION

- UMLS editors integrate new atoms into UMLS
  - Every new atom needs to be associated with atoms already in UMLS.
  - Task is called UMLS Vocabulary Alignment or UVA
  - The task can range from very simple to extremely challenging and often requires in-depth domain knowledge.

  - Simple Examples:
    - “ascorbic acid” => “vitamin C”
    - “lung cancer” => ”pulmonary carcinoma”

  - Complex Example:
    - “SPRL1B” => “LCE2B gene”

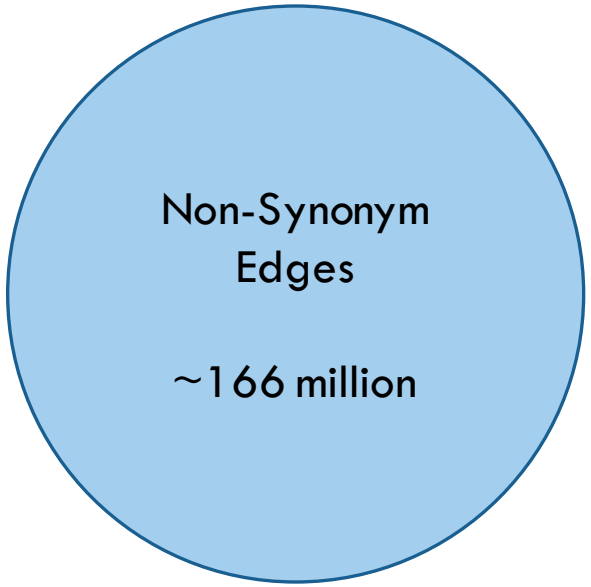
MOTIVATION

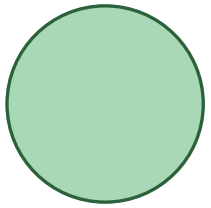
- Given that there are ~10^{10} atom comparisons to be done for each new batch of terms, updating UMLS can become extremely expensive.
  - A medium sized team is unable to keep up with the task by itself.
- Currently, UMLS editors rely on rule-based tools.
  - Thus, our team has been exploring deep learning methods to alleviate the burden on UMLS editors and improve UMLS quality.

PREVIOUS WORK

- Our lab's previous UVA work is formulated as a task where every possible pair would need to be classified.
- Due to the massive size of this task (10^{10} – 10^{14} pairs), current work is done on idealized (but still large-scale) subsets (10^{8} pairs).
  - 166 million negative edges when there are 10^{10} possible negative edges.

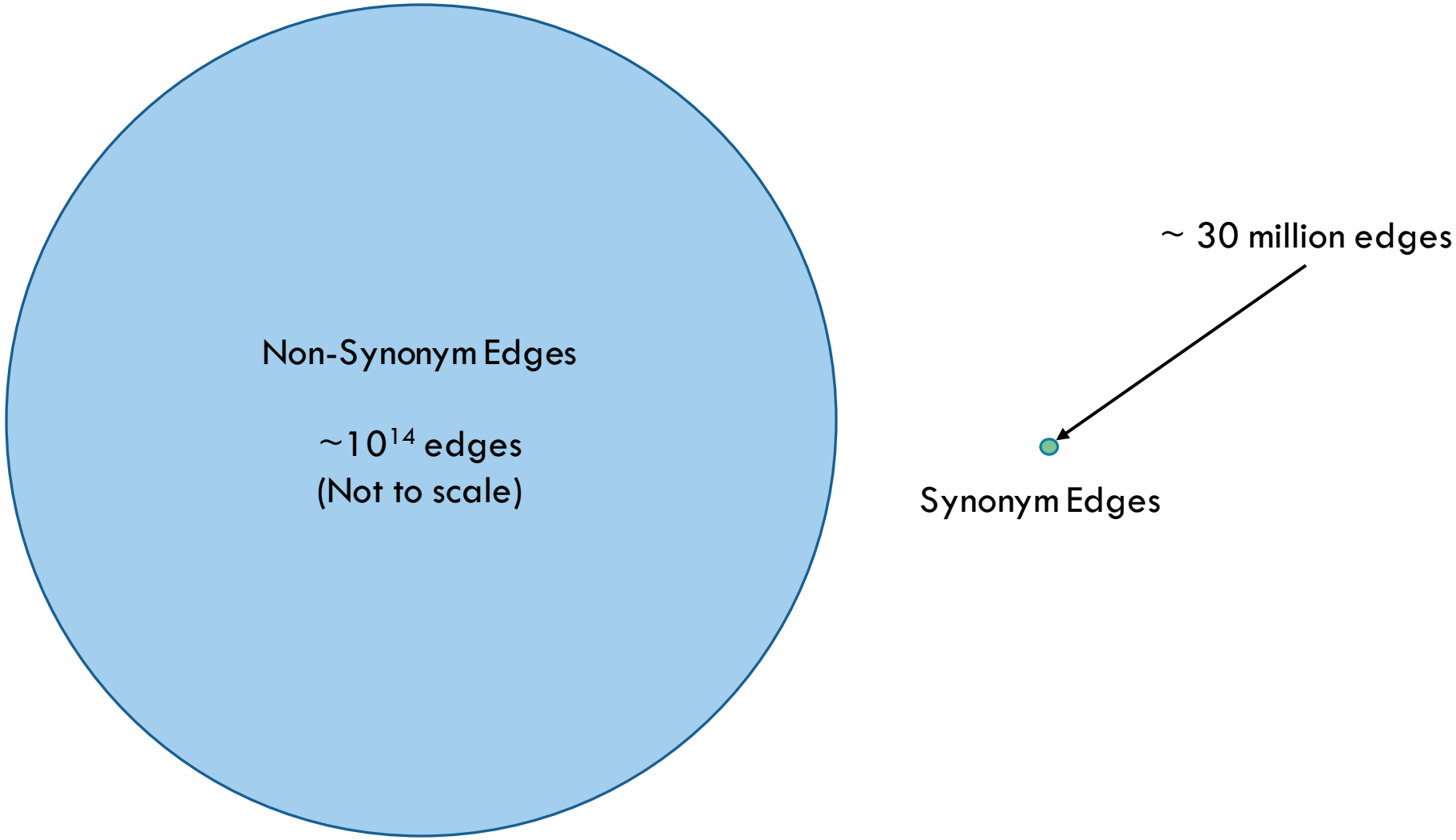
CURRENT TEST DATASET





Synonym Edges
~5 million

REAL WORLD DATASET



PREVIOUS WORK

- Current datasets use a prevalence which is 10^6 times lower than real world datasets
  - Prevalence: % of positive samples in dataset
- This prevalence gap means that even models which perform well in current datasets would likely yield poor results in the real-world scenario.
  - Many negative edges would be incorrectly predicted as positive.

TRANSLATIONAL UVA GOALS

Main Goal: Build a system which can be directly deployed to support UMLS editors for UMLS construction and updating.

Research Aims:

1. Define **task** and **datasets** which faithfully represent the real-world task.
2. Design a **baseline system** that can address the task and dataset designed.
3. Perform **quantitative** and **qualitative evaluation** of baseline system.

TRANSLATIONAL UVA GOALS

Research Aims:

1. Define **task** and **datasets** which faithfully represent the real-world task.
   - **Task Definition:** For each new atom to be introduced to UMLS, find all synonymous atoms in the current UMLS. (analogous to real-world task)
   - **Evaluation Dataset:**
     - 430k new atoms were introduced between the first and second version of (2020AA vs 2020AB).
     - For each of these 430k new atoms in 2020AB, we are looking to determine which atoms are likely to be its synonyms in UMLS 2020AA.
2. Design a **baseline system** that can address the task and dataset designed.
3. Perform **quantitative** and **qualitative evaluation** of baseline system.

TRANSLATIONAL UVA GOALS

Research Aims:

1. Define **task** and **datasets** which faithfully represent the real-world task.
2. Design a **baseline system** that can address the task and dataset designed.
   - Two-step system:
     - High recall candidate generation (fast but misses few potential synonyms)
     - High precision synonymy classification (slower but more discriminative)
3. Perform **quantitative** and **qualitative evaluation** of baseline system.

TRANSLATIONAL UVA GOALS

Research Aims:

1. Define **task** and **datasets** which faithfully represent the real-world task.
2. Design a **baseline system** that can address the task and dataset designed.
3. Perform **quantitative** and **qualitative evaluation** of baseline system.
   - Quantitative Evaluation
     - High Recall Step
       - Recall at K - % of true synonyms that can be found within the first K atoms retrieved from the original UMLS.
     - High Precision Step
       - F1, Precision and Recall on true synonym pairs

TRANSLATIONAL UVA GOALS

Research Aims:

1. Define **task** and **datasets** which faithfully represent the real-world task.
2. Design a **baseline system** that can address the task and dataset designed.
3. Perform **quantitative** and **qualitative evaluation** of baseline system.
   - Qualitative Evaluation
     - Sample output should be carefully examined by biomedical experts
     - UMLS has some ambiguities and errors, thorough analysis is required to ascertain the quality of the predictions compared to the “gold standard”.

METHODOLOGY

- High recall candidate generation (fast but gets many false positives)
- High precision synonymy classification (slower but more discriminative)

METHODOLOGY

- High recall candidate generation (fast but gets many false positives)
- High precision synonymy classification (slower but more discriminative)

HIGH RECALL STEP: FORMULATION

- Task formulation:
  - 430k query terms
  - ~10 million term database
  - Retrieve a limited # of candidates from the database for each query which hopefully contain relevant candidates.
- Similar tasks:
  - Information retrieval (IR) (finding relevant documents with respect to a query)
  - Entity linking (finding relevant concepts with respect to a term mentioned in text)

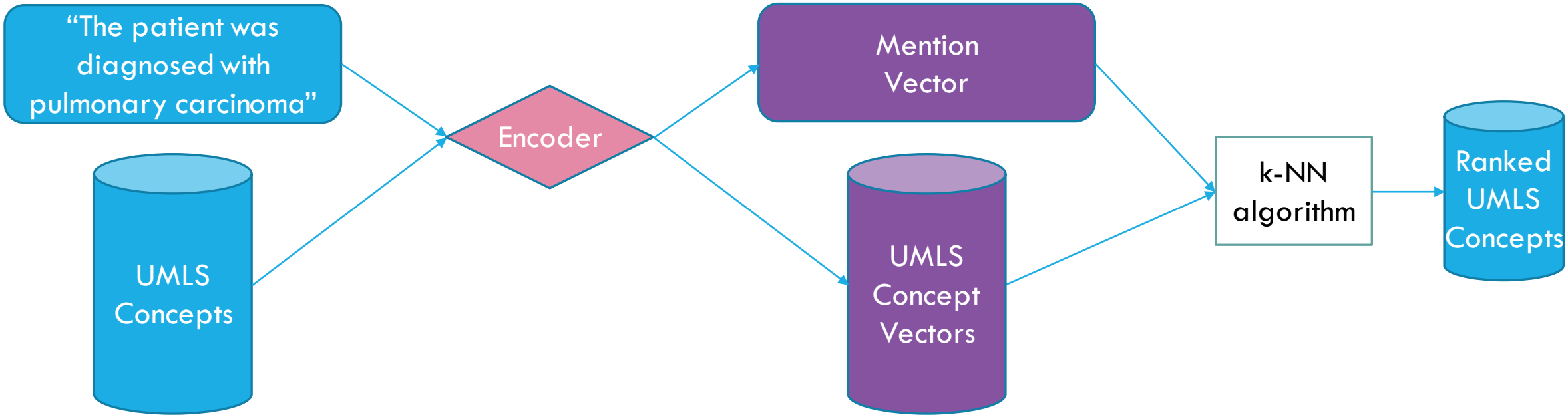
HIGH RECALL STEP: FORMULATION

- Both IR and entity linking use modern textual encoders (often pre-trained language models) and a fast implementation of the k-nearest neighbors (k-NN) algorithm to achieve a fast and high recall candidate retrieval step.
- Spurred on by PLMs as well as k-NN speedups using GPUs (Johnson et al. 2017)

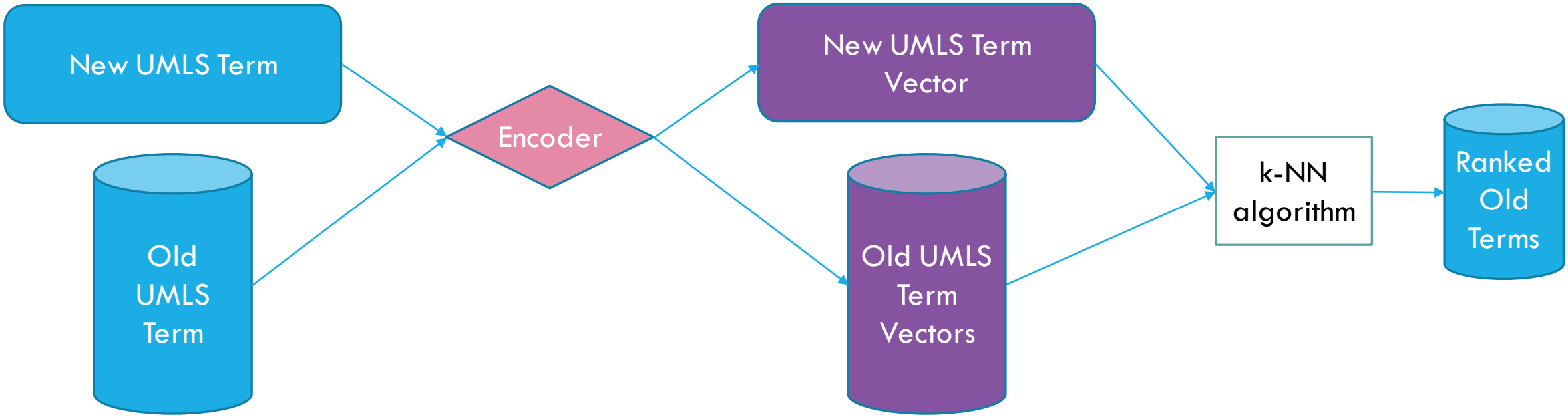
HIGH RECALL STEP: FORMULATION

- We model our high recall approach directly on biomedical entity linking:
  - In this task, a concept is mentioned within a sentence:
    - "The patient was diagnosed with pulmonary carcinoma."
  - We then link this mention to a UMLS entity which represents "pulmonary carcinoma".

HIGH RECALL STEP: BIOMEDICAL ENTITY LINKING



HIGH RECALL STEP: OUR APPROACH



HIGH RECALL STEP: ENCODERS

- Any system that produces a dense vector from text can be an encoder.
- Examples:
  - Current UVA models like LexLM, ConLM and UBERT.
  - BioWordVec embeddings
  - Pretrained language models (PLMs) like BERT, RoBERTa
  - Biomedical PLMs (PubMedBERT)
  - Biomedical PLMs with infused UMLS information (SAPBERT and KRISSBERT)
    - (Liu et al. 2020, Zhu et al. 2020, Bhowmik et al. 2021, Zhang et al. 2022, Xu et al. 2022)

HIGH RECALL STEP: GPU K-NN SPEEDUP

- Benchmarking k-NN for LexLM on GPUs vs CPU
  - Database: 8,521,220 AUIs
  - Queries: 430,135 AUIs
  - Dimension: 50
- CPU Time: 3 hours (180 minutes)
- GPU Time: 3 minutes
- GPU offers around a 60 times speedup to the k-NN

Model	R @ Source Synonymy	R@1	R@5	R@10	R@50	R@100	R@200	R@500	R@1000	R@2000
PubMedBERT	-	9%	16%	18%	23%	25%	28%	31%	34%	37%
LexLM	-	10%	22%	28%	42%	47%	51%	56%	59%	62%
KRISSBERT	-	13%	25%	30%	43%	48%	53%	59%	64%	68%
SAPBERT	-	20%	44%	53%	71%	76%	81%	86%	88%	89%
PubMedBERT (Source Syn)	35%	41%	46%	48%	52%	53%	55%	57%	59%	61%
LexLM (Source Syn)	35%	42%	51%	56%	66%	70%	73%	77%	79%	81%
KRISSBERT (Source Syn)	35%	43%	51%	55%	64%	68%	71%	76%	79%	82%
SAPBERT (Source Syn)	35%	46%	63%	70%	83%	86%	90%	93%	95%	95%

HIGH RECALL STEP: RESULTS

Model	R @ Source Synonymy	R@1	R@5	R@10	R@50	R@100	R@200	R@500	R@1000	R@2000
PubMedBERT	-	9%	16%	18%	23%	25%	28%	31%	34%	37%
LexLM	-	10%	22%	28%	42%	47%	51%	56%	59%	62%
KRISSBERT	-	13%	25%	30%	43%	48%	53%	59%	64%	68%
SAPBERT	-	20%	44%	53%	71%	76%	81%	86%	88%	89%
PubMedBERT (Source Syn)	35%	41%	46%	48%	52%	53%	55%	57%	59%	61%
LexLM (Source Syn)	35%	42%	51%	56%	66%	70%	73%	77%	79%	81%
KRISSBERT (Source Syn)	35%	43%	51%	55%	64%	68%	71%	76%	79%	82%
SAPBERT (Source Syn)	35%	46%	63%	70%	83%	86%	90%	93%	95%	95%
SAPBERT (Source Syn + LUI)	58%	25%	60%	71%	88%	91%	94%	96%	97%	98%

HIGH RECALL STEP: RESULTS

HIGH RECALL STEP: TAKEAWAYS

- **SAPBERT** is by far the most effective encoder.
- Leveraging **basic lexical similarity** and **source synonymy** greatly improves candidates obtained from only dense representations.
  - Throwing away rule-based signal is detrimental to performance.
- We achieve above **90% recall at above 100 candidates** with the best system.
  - This is high enough for a useful real-world system to support UMLS editors (humans in the loop are still vital).

METHODOLOGY

- High recall candidate generation system (fast but gets many false positives)
- High precision synonymy classification (slower but more discriminative)

HIGH PRECISION STEP: FORMULATION

- Output from first step:
  - ~100-200 (query term, candidate term) pairs for each query term
  - **Imbalanced distribution:** only 5-10% of these pairs are synonymous (much higher prevalence than natural one)
- Task formulation:
  - Same formulation used by previous UVA methods (LexLM, UBERT))
  - Binary synonymy classification for each (query term, candidate term) pair.
- Approach:
  - Given the success of PLM fine-tuning in a wide range of NLP tasks, we leverage PLMs.
  - To deal with the heavy class imbalance, we sample a balanced number of positive and negative pairs for training.

HIGH PRECISION STEP: DATASETS

- Dev and Test Datasets
  - Top 100 SAPBERT candidates from the 430k new 2020AB term dataset.
  - Set aside 1000 and 2000 concepts for dev and test sets, respectively.
  - Use all 100 candidates for each concept to create dev and test sets
    - 100,000 dev set and 200,000 test set pairs
    - For this setting, we also add whatever synonyms are missing from the candidate list (not fully realistic but upper bound on performance)

HIGH PRECISION STEP: DATASETS

- Training Datasets
  - Ideal Distribution
    - Rest of 2020AB new terms (Same semantic group distribution as dev and test set)
    - By using this training set, we are inadvertently introducing information about the new terms that we would not have in the real-world setting.
  - Realistic Distribution
    - We separate 400k UMLS 2020AA terms as a different “new” dataset.
    - Find 100 k-NNs for each of these terms within what remains of 2020AA.
    - Create 40 million training dataset to sample balanced datasets from.

HIGH PRECISION STEP: DATASETS

- Training Dataset Types
  - Balanced
  - Stratified
  - Dev Set Equivalent: 100 candidates for each query term (not shuffled)
- Training Dataset Sizes
  - 10k
  - 100k
  - 200k
  - 500k

HIGH PRECISION STEP: MODELS

- Fine-tuned Models
  - PubMedBERT
- Baselines
  - UBERT (Original + SAPBERT)
  - LexLM
  - ConLM

HIGH PRECISION STEP: RESULTS

Model	Training Dataset	Training Data Type	Training Data Size	F1	Precision	Recall
SAPBERT + UBERT Synonymy Prediction	UVA Train	Stratified	166 M	27.4%	16.3%	87.4%
UBERT MLM + Synonymy Prediction	UVA Train	Stratified	166 M	35.0%	22.0%	85.0%
PubMedBERT Fine Tuning	Ideal	Balanced	10k	38.2%	24.6%	85.4%
PubMedBERT Fine Tuning	Ideal	Stratified	10k	40.6%	38.1%	43.3%
PubMedBERT Fine Tuning	Ideal	Balanced	100k	46.7%	31.7%	88.8%
PubMedBERT Fine Tuning	Ideal	Stratified	100k	33.4%	49.1%	25.3%
PubMedBERT Fine Tuning	Ideal	Dev Set Equivalent	100k	25.9%	54.7%	17.0%

HIGH PRECISION STEP: TAKEAWAYS

- Both UBERT versions (which outperform other models in UVA work) underperform small-scale fine-tuning
  - Small (100k samples) but more realistic datasets yield better real-world performance than training on millions of synonym pairs.

- Training set distribution drastically affects performance
  - Balanced datasets (1:1) yield high recall but low precision
  - Stratified datasets (1:~10) yields higher precision but very low recall

HIGH PRECISION STEP: RESULTS

Model	Training Dataset	Training Data Type	Training Data Size	F1	Precision	Recall
PubMedBERT Fine Tuning	Ideal	Balanced	100k	46.7%	31.7%	88.8%
PubMedBERT Fine Tuning	Realistic	Balanced	100k	41.5%	27.3%	86.6%
PubMedBERT Fine Tuning	Ideal	Balanced	200k	42.8%	32.0%	64.7%
PubMedBERT Fine Tuning	Realistic	Balanced	200k	43.8%	29.0%	90.2%
PubMedBERT Fine Tuning	Ideal	Balanced	500k	52.5%	37.1%	90.0%
PubMedBERT Fine Tuning	Realistic	Balanced	500k	37.1%	23.1%	94.6%

HIGH PRECISION STEP: TAKEAWAYS

- Original distribution training datasets underperform ideal distribution training.
- The correlation between dataset size and performance is not as strong as expected.
- Training is quite noisy
  - Training metrics keeps increasing but dev set performance drops after epoch 1 in most cases

- Precision is only at ~30-40%, not high enough for deployable system
  - 2/3 of all predicted synonym pairs are not synonymous according to gold labels
  - Qualitative evaluation is necessary to determine how this model performs in practice

QUALITATIVE EVALUATION: FALSE POSITIVES

- Low precision problem is due to the high number of false positives
- For every 1 synonym pair predicted correctly, 2 are incorrect according to UMLS
- Unfortunately, or fortunately, it is very difficult to determine whether each of these false positives is a true error or a UMLS error
  - The amount of time spent on each term would be very large (even for a person with some biomedical training)

QUALITATIVE EVALUATION: FALSE POSITIVES

Query	Candidate	Label	Pred
arginine/serine-rich protein 1	SRA1	0	1
ATP synthase, H+ transporting, mitochondrial Fo complex, subunit F2 pseudogene 3	ATP5MC3 gene	0	1
SDYS	SDYS	0	1
BENZALKONIUM CHLORIDE 1 mg in 1 g TOPICAL CLOTH [Antiseptic Towel Benzalkonium Chloride]	Pro Pet Dental Wipes 0.1 % Medicated Pad	0	1
transfer RNA tyrosine 1 (anticodon GUA)	TRT-AGT2-1 gene	0	1
WW domain binding protein 1-like pseudogene 4	WBP4 gene	0	1
protirelin	Thyrotropin-releasing factor, prepro-	0	1
sodium hyaluronate 23 MG/ML Injectable Solution	HYALURONATE NA (GEL-ONE) 10MG/ML SYR 3ML	0	1
Sf9 Cell	SR cell line	0	1
fragile histidine triad diadenosine triphosphatase	Hemin-Controlled Translational Repressor	0	1
Meningismus	hemiballismus	0	1
Endocervix	Endocervical epithelium structure (body structure)	0	1

CONCLUSION

Complete:

1. Define **task** and **datasets** which faithfully represent the real-world task.
2. Design a **baseline system** that can address the task and dataset designed.

In Progress:

1. Perform **quantitative** and **qualitative evaluation** of baseline system.
    1. More baselines are necessary for classification system
    2. Thorough qualitative evaluation is crucial

FUTURE CHALLENGES

- Addressing the more moderate but still important class imbalance problem
  - Distributionally robust optimization (Levy et al. 2020) or other similar techniques
- UMLS is noisy and synonymy task is often ambiguous
- Data scarcity
  - Only data point for determining synonymy is a short phrase and the source it comes from.
  - This is unrealistic since other data points are used by humans to make determination (other source synonyms, source semantic categories, descriptions, hierarchical structure, etc.).
  - Adding this information is crucial for models to perform better.

REFERENCES

- Bhowmik, R., Stratos, K., & de Melo, G. (2021). Fast and Effective Biomedical Entity Linking Using a Dual Encoder. *ArXiv, abs/2103.05028*.
- Jeff Johnson, Matthijs Douze, Hervé Jégou: "Billion-scale similarity search with GPUs", 2017; [http://arxiv.org/abs/1702.08734 arXiv:1702.08734].
- Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, Shaoping Ma: "RepBERT: Contextualized Text Embeddings for First-Stage Retrieval", 2020; [http://arxiv.org/abs/2006.15498 arXiv:2006.15498].
- Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, Shaoping Ma: "Optimizing Dense Retrieval Model Training with Hard Negatives", 2021; [http://arxiv.org/abs/2104.08051 arXiv:2104.08051].
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, Arnold Overwijk: "Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval", 2020;
- Liu, F., Shareghi, E., Meng, Z., Basaldella, M., & Collier, N. (2021). Self-Alignment Pretraining for Biomedical Entity Representations. *NAACL*.
- Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, Hoifung Poon: "Knowledge-Rich Self-Supervision for Biomedical Entity Linking", 2021; [http://arxiv.org/abs/2112.07887 arXiv:2112.07887].
- Xu D, Miller T. A simple neural vector space model for medical concept normalization using concept embeddings. J Biomed Inform. 2022 Jun;130:104080. doi: 10.1016/j.jbi.2022.104080. Epub 2022 Apr 23. PMID: 35472514.
- Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, Haifeng Wang: "RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering", 2020;