The 8th International Conference on Computational Data and Social Networks (CSoNet 2019)

Towards an Aspect-based Ranking Model for Clinical Trial Search

Soumyadeep Roy, Nikhil Agrawal, Shamik Sural, **Niloy Ganguly** (IIT Kharagpur, India)
Koustav Rudra (L3S Research Center, Germany)

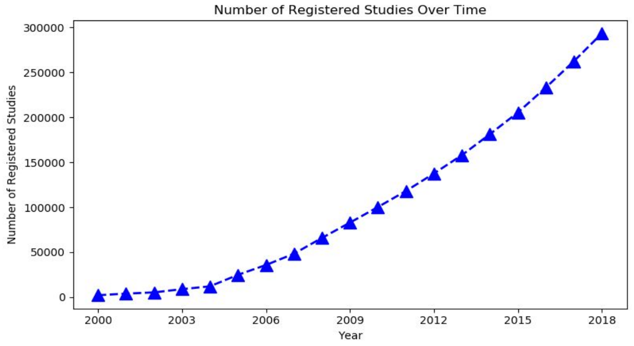
November 20, 2019

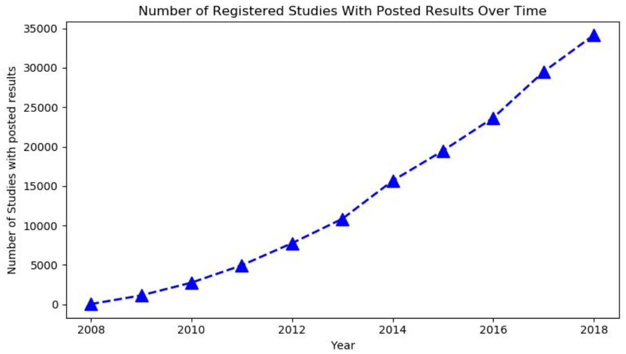
Clinical Trials

- First source of information about new drugs and treatments
- Recruits participants based on published eligibility criteria
    - Usually at a critical stage of disease. Also accepts healthy volunteers
- Many trial search engines exist
    - ClinicalTrials.gov, eTACTs

Clinical Trials Search - What changed?

- Number of trials being registered each year is increasing
- Number of trials that posted results each year is increasing
- Fraction of trials whose result is posted is increasing
    - 12.4% vs. 11.6% (in 2018)





Source: ClinicalTrials.gov, accessed on 16th Nov, 2019

Background - 1

- Large number of Health Information sources are available
  - Free for public use : Medline, ClinicalTrials.gov, Systematic reviews
  - Paid : Synthetic Reviews (DynaMed, UpToDate)
- A large number of search systems exist
  - Trials - eTACTS, ClinicalTrials.gov;
  - Research papers - PubMed, Embase, PMC;
  - Systematic Cochrane Reviews

Background - 2

- Different stakeholders have different expertise level and information need.
  - Patients, Consumers
  - Caregivers(Insurance companies, Nurses), Clinicians, Doctors
- Aspect-based ranking
  - Drug Reviews- aspects(condition, side-effects, dosage & effectiveness)

Problem Statement

- Building a medical search engine applicable to different classes of disease
  - Disease classes : Pathological, Cardiovascular, Nervous System, Nutritional and Metabolic, Immune System Diseases
  - Query : free-form text
  - Output : list of relevant trials ranked by different aspects

Research Challenges

- Different stakeholders have different information needs
  - A trial with high number of adverse events may be highly relevant. But from a participant point of view, such results are highly undesirable
- Existing systems and benchmark datasets focus only on oncology trials and cannot be extended to other disease classes
  - Use cancer knowledge bases
  - TREC Precision Medicine Track (Started from 2014 to till date)

Key contributions

- Address the different stakeholder perspective by introducing different ranking criteria (aspects)
  - Graph-based : relevancy, popularity
  - Metadata-based : recency, adversity
- Applicable to five disease classes, and may be extended

Key contributions

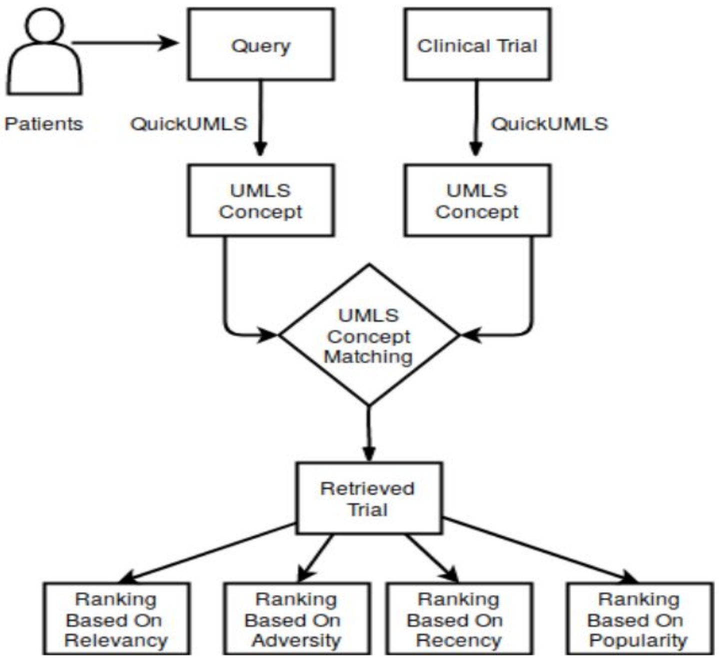
- Address the different stakeholder perspective by introducing different ranking criteria (aspects)
  - Graph-based : relevancy, popularity
  - Metadata-based : recency, adversity
- Applicable to multiple disease classes
- Contribute a ground-truth for evaluation, with detailed annotation mechanism
  - per-query retrieval set (25 queries), open-sourced in Github
- Proposed aspect-based ranking model outperforms the baseline system in almost 90% of cases

Methodology

Methodology overview

- Query representation - UMLS concepts
- Clinical trial representation
- Retrieval of trials
- Ranking trials

- Uses auxiliary information sources
  - PubMed articles



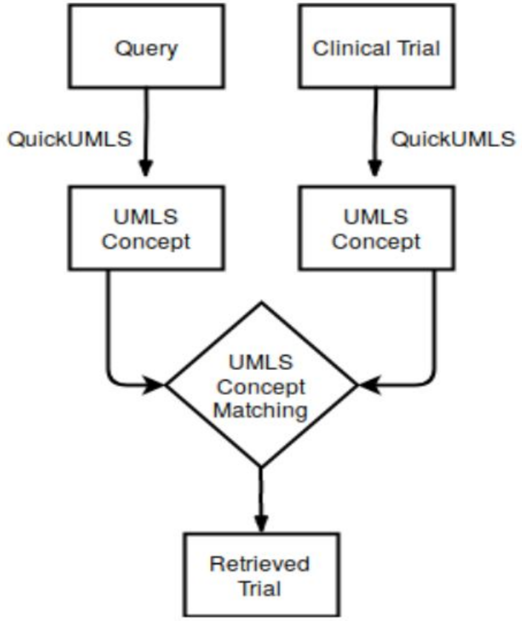
Query and Trial Representation

- Query and Trial representation - Concept extraction
  - Extract UMLS concepts present in brief title and brief summary using QuickUMLS (*Soldaini and Gohrian., 2016*)
  - Same way to represent the query

Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction, In MedIR workshop, SIGIR (2016)

Clinical trial retrieval

- Match-based retrieval
  - For a query 'q', we retrieve all the trials whose brief title contains all the UMLS concept ids that are present in 'q'



Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction

Aspect-based ranking - Relevancy

- PageRank (PGR): Create an undirected graph, G(V, E) where each vertice is a trial
  - Edge weight = *Simpson similarity between UMLS concepts extracted from brief title and brief summary of a trial*
  - Apply PageRank algorithm on G to compute the importance of a trial
  - In terms of the ranking model, trials are ranked in the decreasing order of their 'PageRank' score

Improving Relevancy - background

- Map commonly used patient terms [1] to UMLS concepts
  - 1-word(21.4%), 2-word(55.8%), 3-word queries(19.7%)
- 15% of queries QuickUMLS was unable to extract query concepts
- Problematic queries
  - Aching in limb, Cholesterol levels raised

meddra.org/patient-friendly-term-list

Aspect-based ranking - Relevancy

- Exact term match (ETM): Considers important terms that are not part of an UMLS concept
  - Ranking model: Term frequency in 'brief summary' field > 'official title' > 'brief title';
  - If absent, ranked based on PageRank score
- Synset based term match (STM): Consider 'WordNet synsets' of the extracted terms from a query before exact matching
  - Ranking model: Decreasing order of term frequency in brief summary, followed by official title, brief title, and PageRank score

Aspect-based ranking - Adversity

- May be mapped to 'safety events' category of patient complaints (*Reader et al., 2014*)
  - ‘Adverse events’ is a sub-category
- Data available in the Reported Events table -- Number of subjects affected column
- Ranked in decreasing order of 'Subjects Affected' value
  - If 'Subjects Affected' value is zero, rank those trials in a random order

Reader, T. W., Gillespie, A., Roberts, J.: Patient complaints in healthcare systems: a systematic review and coding taxonomy. BMJ Qual. Saf. 23(8), pp. 678-689 (2014)

Aspect-based ranking - Recency

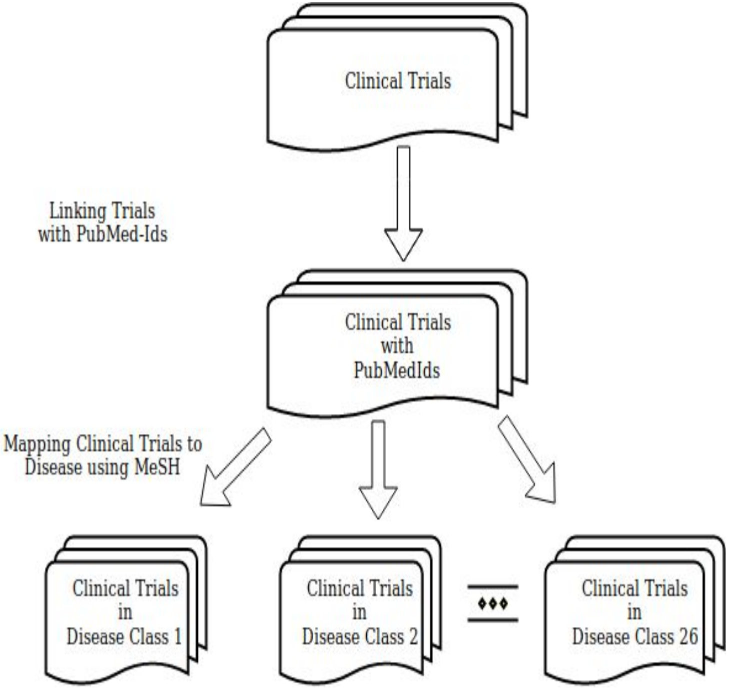
- May want to enroll in a trial or are looking for new treatments or information
  - Existing drugs or treatment methods does not work well for some patients
  - New inventions help medical practitioners to handle such critical patients
- Rank the trials in descending order based on their date of completion
  - Most trials that are going to be completed in the future did not report any tested information or drug

Aspect-based ranking - Popularity

- Measure success of a trial with the 'popularity' of its corresponding PubMed article
  - No. of PubMed articles that cited a given article
- Rank in decreasing order of citation count
  - Break ties by 'relevancy' score

Dataset

- AACT - ClinicalTrials.gov dump on January, 2019
- Select trials mapped to a PubMed id and at least one MeSH term
- Select top 5 disease classes with highest number of trials



https://aact.ctti-clinicaltrials.org/

Ground truth creation

Experimental setup - formulation

- Patients usually query for disease along with some related terms
  - Related terms : syndrome, symptoms,treatment, tests, age-group along the disease (*Patel et al. 2010*)
  - Example queries: Hypertension safe treatments, managing constipation in children
- Retrieve based on user query and then rank the clinical trials
  - Different information like adversity, citation information to rank the trials on different basis

Patel, et.al. "What do patients search for when seeking clinical trial information online?." AMIA Annual Symposium Proceedings. Vol. 2010. American Medical Informatics Association, 2010.

Mapping trials to disease class using MeSH

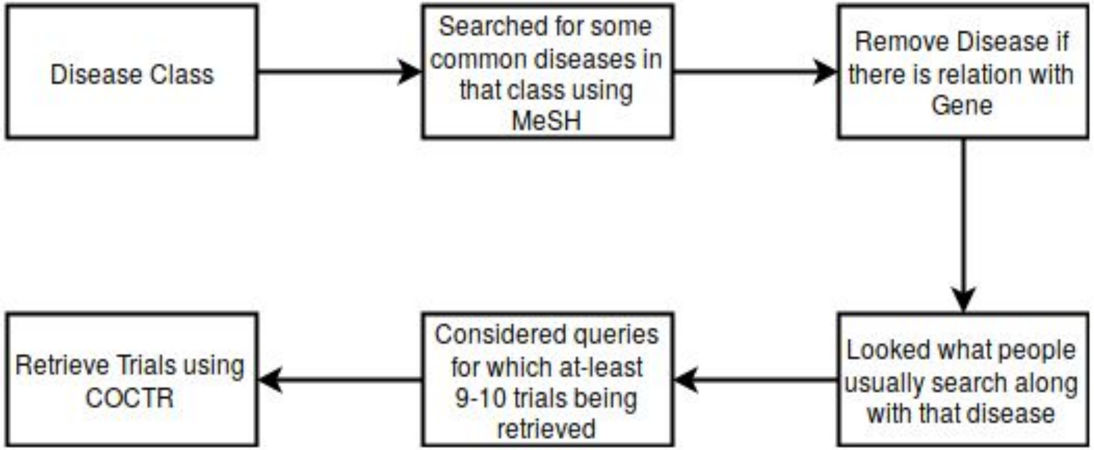
Clinical Trial	Mesh Terms	Trees	Disease
Whole Body Hyperthermia for the Treatment of Rheumatoid Diseases (NCT00000106)	Rheumatic Diseases	C05.799 C17.300.775	1. Musculoskeletal Diseases 2. Skin and Connective Tissue Diseases
	Collagen Diseases	C17.300.200	1. Skin and Connective Tissue Diseases
Safety and Efficacy of ALV003 for the Treatment of Celiac Disease (NCT00959114)	Celiac Disease	C06.405.469.637.250, C18.452.603.250	1. Digestive System Diseases 2. Nutritional and Metabolic Diseases

Query selection

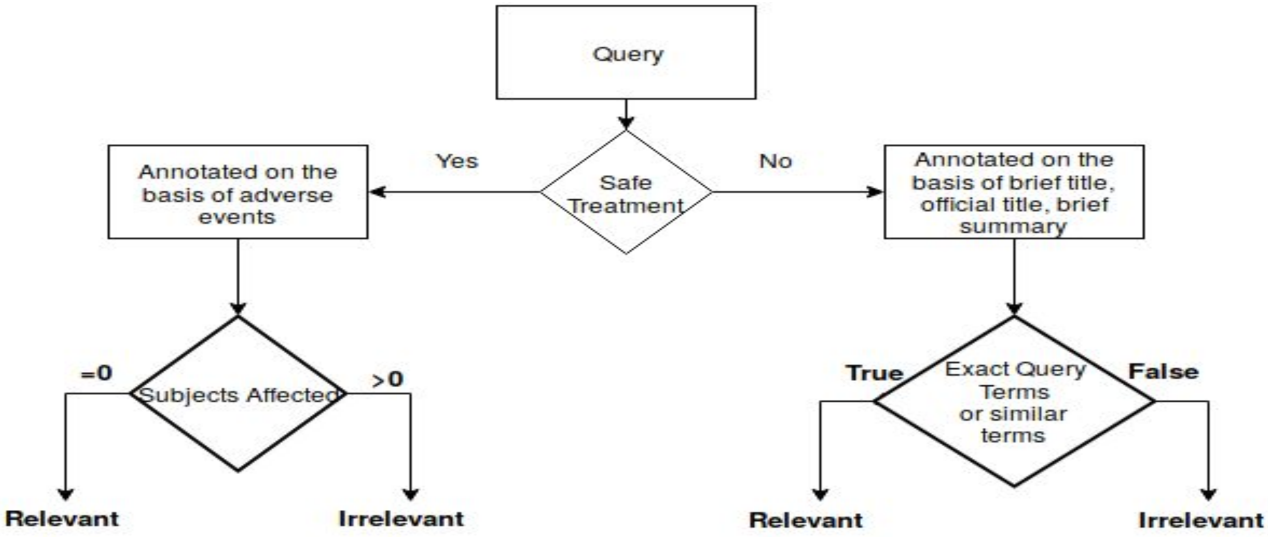
- Set of 5 queries for each of 5 disease classes
- Follow semantic-based query templates proposed by Patel et al.(2010)
  - (disease or syndrome) + (symptom or treatment) *Early Parkinson disease treatment*
  - disease + age group *Managing constipation in children*
  - disease + safety information *Safe treatments for asthma*
  - degree + disease *Serious sleep apnea*
- Prepare final queries after consulting various medical resources
  - MedDRA, CLEF Consumer Health Track 2018, Reddit

Patel, et.al. "What do patients search for when seeking clinical trial information online?." AMIA Annual Symposium Proceedings. Vol. 2010. American Medical Informatics Association, 2010.

Selection of query for each disease class



Relevant retrieval set annotation - adversity



Annotation of retrieved trials

- If query contains safe treatments then trials are marked relevant on the basis of adverse events.
  - E.g. query:- asthma safe treatments
- For queries like serious sleep apnea, serious Rheumatoid arthritis marked retrieve trial relevant due to terms like serious and severe
- Retrieved trial containing children and related to constipation are marked relevant for query
  - managing constipation in children

Experiments and results

Experimental Setup

- Evaluation metric
  - Precision and nDCG score
  - Cannot measure recall due to lack of complete retrieved set of clinical trials for each query
- Three annotators prepared the relevant trial set for each query
  - Followed the annotation scheme for marking a clinical trial 'definitely relevant' for a given topic(in our case, query) for TREC Precision Medicine 2018 Track
- Baseline
  - Thorve et al (TREC 2017) - only system that does not use any disease-specific knowledge bases

Results - 1

Method	P@5	P@10	P@15	P@20
Baseline	0.12	0.08	0.08	0.08
PGR	0.38	0.35	0.35	0.33
ETM	0.53	0.48	0.45	0.42
STM	0.56	0.52	0.47	0.46

STM performs the best since it handles query variations and performs query expansion

Results - 2

- Baseline system retrieves at least five trials for only 3 out of 25 queries
  - Cannot handle query variations and normalization of medical terms
- STM outperforms the baseline system in 90% of queries
  - 'precision@5' value of 0.56
- STM achieves a high precision@10 value for two-third of the queries
  - Using UMLS concept for normalization is useful

Results - 3

Query no.	PAT	CVD	NER	NMT	IM
Q1	0.97	0.88	0.5	0.64	0.8
Q2	0.77	0.65	0.92	0.62	0.95
Q3	0.96	0.66	0.72	0.54	0.96
Q4	1.0	0.88	0.66	0.99	0.71
Q5	0.84	0.52	0.98	0.91	0.8

STM achieves a high nDCG score

Conclusion

- Introduce multi-dimensional ranking of clinical trials - adversity, popularity, recency along with relevancy
- Proposed ranking model, STM, outperforms baseline in 90% of queries
  - Achieves a 'precision@5' value of 0.56
- Codebase and annotated data files publicly available on Github
  - https://github.com/nikhil741/COCTR_multidimensional_ranking

Future Work

- Leverage disease-independent knowledge bases
  - Use stronger baselines by substituting the cancer-specific knowledge bases
  - SOTA systems from the TREC 2017 and 2018 task
- Perform topic expansion and also consider microtext variations
- Apply more sophisticated aspect fusion techniques for creating a single ranked list

Acknowledgements

- SERB-funded project “IoTDiff: Service Differentiation for Personal Hubs”
- IMPRINT-funded project “Development of a Remote Healthcare Delivery System: Early Diagnosis, Therapy, Follow-up and Preventive Care for NCDs (Cardio-pulmonary)”
- European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 832921

References

- Kilicoglu, H., Abacha, A.B., et al.: Semantic annotation of consumer health questions. BMC bioinformatics 19(1), 34 (2018)
- Zuccon, G., Koopman, B., et al.: Choices in knowledge-base retrieval for consumer health search. In: European Conference on Information Retrieval. pp. 72–85 (2018)
- Reader, T.W., Gillespie, A., Roberts, J.: Patient complaints in healthcare systems: a systematic review and coding taxonomy. BMJ Qual Saf 23(8), 678–689 (2014)

Thank you for your attention