Crash Course IR – Test Collections

Sebastian Hofstätter

sebastian.hofstaetter@tuwien.ac.at
/s_hofstaetter

Today

Crash Course IR – Test Collections

1. **Existing Test Collections**
   - What is a test collection?
   - MSMARCO & TREC-DL
2. **Creating a Test Collection**
   - How to run your own annotation campaign
   - Pooling multiple IR system results
3. **Analyzing Biases**
   - Social biases
   - Position bias in passages and documents

Slides partially adapted from Markus Zlabinger & Aldo Lipani

The Need for Datasets

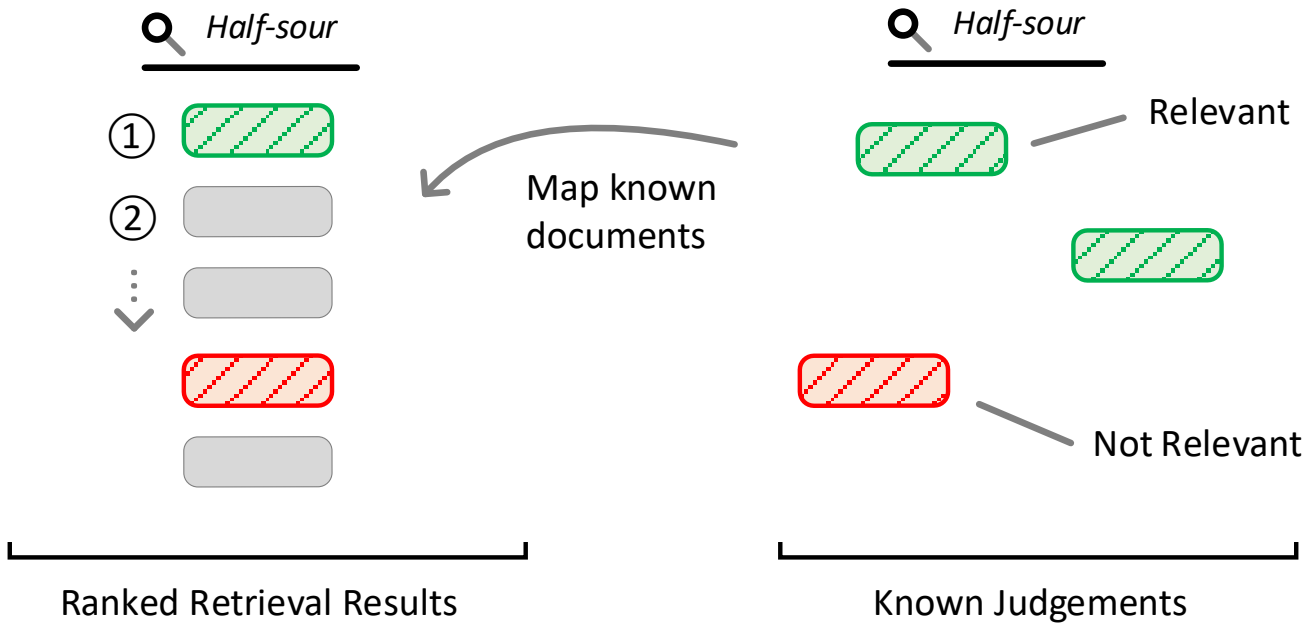
- Machine Learning is driven by data
- Every task needs a custom task-specific dataset
  - For empirical evaluation of model effectiveness
    Is our system better than the baseline?
  - Most of the time also for training or fine-tuning of a pre-trained model
  - The more the merrier!

- Many tasks = many datasets
  - However, the more specific a task we want to solve
    the scarcer datasets become

A great hub for text-based datasets: https://huggingface.co/datasets

Offline IR Evaluation Setup

- Quality of systems, that produce ranked list of documents
- Compared by a pool of judgements (does not necessarily cover the whole list)
  - Missing judgements are often considered as non-relevant



Existing IR Test Collections

The foundation for all IR research

A Test Collection

- Offline evaluation with a fixed *test collection*
  - Fixed set of documents
  - Fixed set of queries
  - Fixed set of judgements
    - Does not necessarily cover all query-doc combinations
    - Can be binary or graded (f.e. 0,1,2,3)
- Query Source:
  - Handcrafted queries for a set of documents (Many TREC collections)
  - Sampled queries from actual users (MS MARCO)
- Document Source:
  - Is very task specific (well parsed web, news articles, Wikipedia, etc..)

Existing IR Datasets

- Annual annotation campaigns producing datasets
  - TREC https://trec.nist.gov/data.html
  - NTCIR http://research.nii.ac.jp/ntcir/data/data-en.html

- Example of datasets:
  - CORD-19 (COVID related scientific publication search)
  - ClueWeb (Web search on open-source web-crawl data)
  - TREC Robust 04 (Topic-based news article search; influential for a long time)
  - Newsgroups-20
  - Reuters-21578

Brand new hub to get IR data: https://ir-datasets.com/ by MacAvaney et al.

MS MARCO Microsoft MAchine Reading COmprehension Dataset

- First re-ranking dataset with **a lot of** training data
  - Scale of training & evaluation data now is an issue (luxury problem!)
- Real-world web search queries and passage-level answers from Bing
  - Initial focus on QA answer generation
  - Released by Microsoft Research
- Sparse Judgement labels for 500 thousand (!) queries
  - Queries selected with high occurrence count -> impossible to leak personal data
  - Human annotated
  - Only ~1 relevant judged document per query

Official website (dataset & leaderboard): http://www.msmarco.org/

Importance of MS MARCO

- Fueled the paradigm shift towards neural networks
  - Allowed academic researchers to participate in large-scale IR
- Leaderboard quickly became very popular (> 100 entries)
- Most neural IR papers make use of the MSMARCO data in some form
  - Including work from Google

MS MARCO — Home | Document Ranking | **Passage Ranking** | Updates | Submissions | About

Passage Ranking Leaderboard(10/26/2018-Present) ranked by MRR on Eval

Rank	Model	Ranking Style	Submission Date	MRR@10 On Eval	MRR@10 On Dev
1	**RocketQA + ERNIE** Baidu NLP - [Qu et al.]	Full Ranking	September 18th, 2020	0.426	0.439
2	**UED-Large** Anonymous	Full Ranking	August 12th, 2020	0.424	0.436
3	**DR-BERT** X.W. S of Meituan-Dianping NLP-KG Group	Full Ranking	May 20th, 2020	0.419	0.420
4	**expando-mono-duo-T5** Ronak Pradeep, Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin - University of Waterloo	Full Ranking	May 19th, 2020	0.408	0.420
5	**DeepCT + TF-Ranking Ensemble of BERT, ROBERTA and ELECTRA** (1) Shuguang Han, (2) Zhuyun Dai, (1) Xuanhui Wang, (1) Michael Bendersky and (1) Marc Najork - 1) Google Research, (2) Carnegie Mellon - Paper and Code	Full Ranking	June 2nd, 2020	0.407	0.421
6	**UED** Anonymous	Full Ranking	May 5th, 2020	0.405	0.414
7	**UED-Large** Anonymous	Full Ranking	August 11th, 2020	0.405	
8	**TABLE Model** X.W. S of Meituan-Dianping NLP-KG Group	Full Ranking	May 11th, 2020	0.401	0.412
9	**TABLE Model** X.W. S of Meituan-Dianping NLP-KG Group	Full Ranking	January 21th, 2020	0.400	0.401
10	**TABLE Model** X.W. S of Meituan-Dianping NLP-KG Group	Full Ranking	May 8th, 2020	0.400	0.401
11	**Knowledge Distilled Student + Teacher Ensemble** Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, Allan Hanbury of TU Wien - [Hofstätter et al., '20] and [Code]	Full Ranking	November 30th, 2020	0.399	0.407

Leaderboard from 5.3.2021

MS MARCO – Data Example

- Training triples
  - **Query:** *what fruit is native to Australia*
  - **Relevant:** *Passiflora herbertiana. A rare passion fruit native to Australia. Fruits are green-skinned, …*
  - **Non-Relevant:** *The kola nut is the fruit of the kola tree, a genus (Cola) of trees that are native to the tropical rainforests of Africa.*
- Re-ranking Evaluation tuples
  - **Ids:** 837202 1000252
  - ***Query:*** *what is the nutritional value of oatmeal*
  - ***Document:*** *Oats make an easy, balanced breakfast. One cup of cooked oatmeal contains about 150 calories, four grams of fiber (about half soluble and half insoluble), and six grams of protein. …*

MS MARCO Passages & Documents

- Initial MSMARCO-Passage dataset consist of 8 million passages (3GB plaintext)
  - Passages have been pre-selected from a web-document by Bing
- MSMARCO-Document is an additional collection
  - The same queries and judgements as MSMARCO-Passage
  - Includes now full web pages, titles, and URLs from corresponding passages
  - Slight time-mismatch and non-perfect matching (but not a huge issue)
  - Documents are long >2.000 words on average
  - Plaintext size increased to ~20GB

Limitations of MS MARCO

- Cookie cutter selection of passages from the full Bing index
  - Started from sampled queries -> take top-10 passages from Bing per query to form the passage corpus
  - Might make the task to easy in comparison with a web-scale index
- Focus on natural language question style queries
  - Initial project was focused on QA and answer generation
- Only for the English language
- Sparse judgements
  - Only 1 judgement per query (misses a lot of additional relevant results)
  - Somewhat mitigated with large query count for evaluation

TREC – Text Retrieval Conference

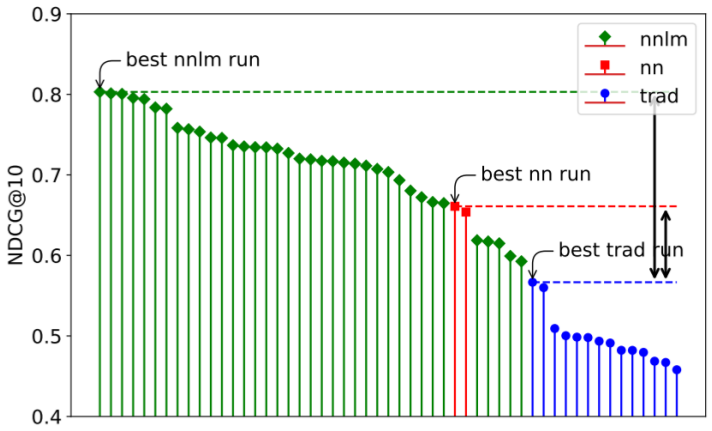
- Annual Event organized by NIST including:
  - Various tasks (tracks) with a challenge for participating systems
    - IR-oriented, but with many interesting twists in domains and settings
  - Many teams try to solve the tasks and submit solutions (runs)
    - In the summer months
  - TREC has budget and expertise for annotations
    - Organizers create judgements for all participating systems
  - Conference (in November) to present results and for teams to connect
- Established many IR standards and test collections for all
- Awesome culture for information sharing and fun competition

TREC Deep Learning Track

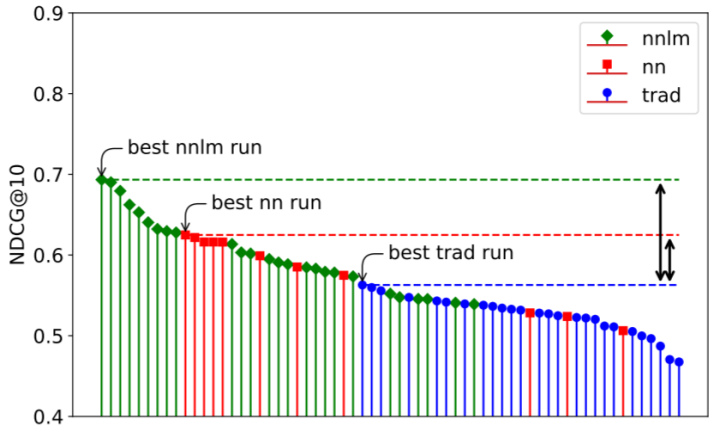
- Started in 2019 (continued in 2020)
- Based on MSMARCO Passage & Document collections
- Setup to evaluate a diverse set of systems trained on large-data
- Selected a small subset of MSMARCO queries (43 in 2019 & 54 in 2020)
  - Participants created runs for those queries
  - TREC pooled runs & created as many graded judgements as possible per query
- Main results so far:
  - Neural methods are as good as we thought before
  - Sparse judgements are a good approximator (for certain evaluations)

Official website: https://microsoft.github.io/msmarco/TREC-Deep-Learning

TREC-DL 2020: Neural vs. Traditional Methods



Passage Task

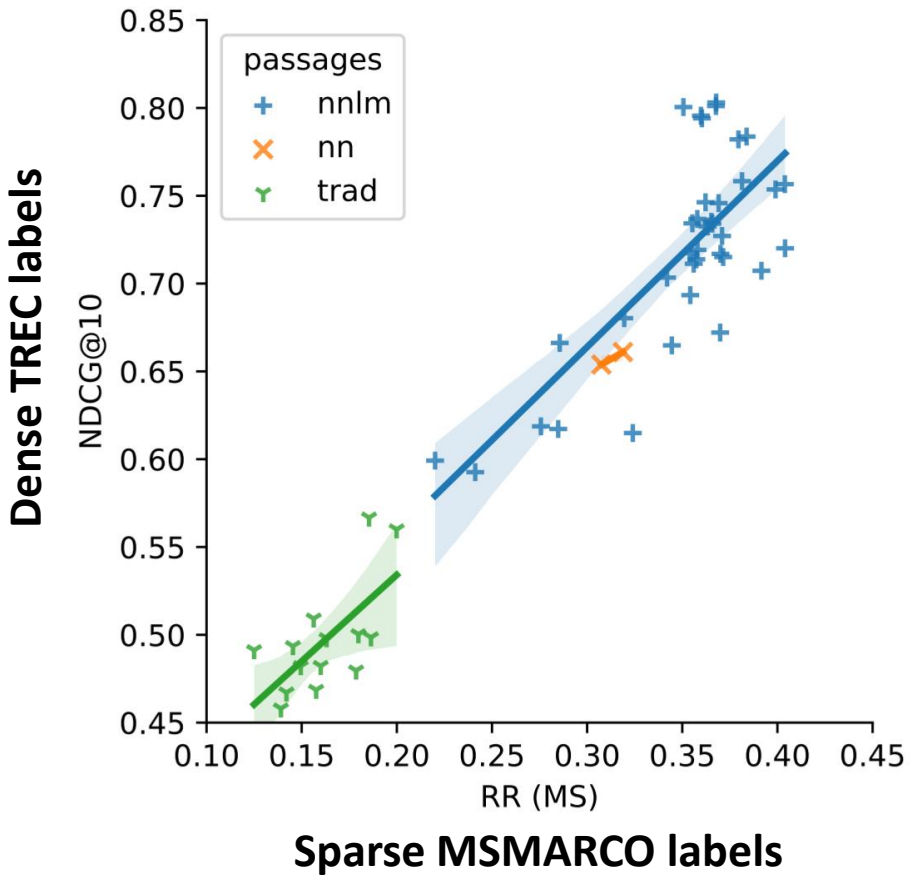


Document Task

- Passage Task shows a very large spread between best nnlm (pre-trained language model) , nn (neural model from scratch) & trad (traditional IR models)
  - Strong validation of the neural hype!
  - *Note:* top nnlm spots are slow ensembles
- Documents have same trend with less spread and nn does better
  - Passage & Doc. Numbers are not directly comparable
  - More investment in passage task by the community

From the 2020 TREC-DL overview paper: https://arxiv.org/pdf/2102.07662.pdf

TREC-DL 2020: Sparse vs. Dense Judgements



- Analysis using both MSMARCO & TREC passages labels found:
  - A clear correlation between sparse and dense judgements
  - Caveat: Small differences, esp. at the top can not be used for clear system ordering using the 50 queries
  - We can continue to use the sparse labels, if we not overclaim small differences
- For documents the correlation is lower and less clear (esp. for trad. models)

From the 2020 TREC-DL overview paper: https://arxiv.org/pdf/2102.07662.pdf

Creating IR Test Collections

How to know what we don’t know?

Who creates the dataset?

- Every decision has tradeoffs
- Initial assessment of pros & cons:

	On your own	External annotators
Recruitment	Not required	Requires recruitment procedure
Compensation	Not required	Needs compensation
Preparation	You (should) know how the task is performed	Need training to prepare workers for the task
Scaling	Does not scale well	Scales very well
Expertise	You might not have required expertise (e.g. medical domain)	Can recruit expert annotators

Task Design

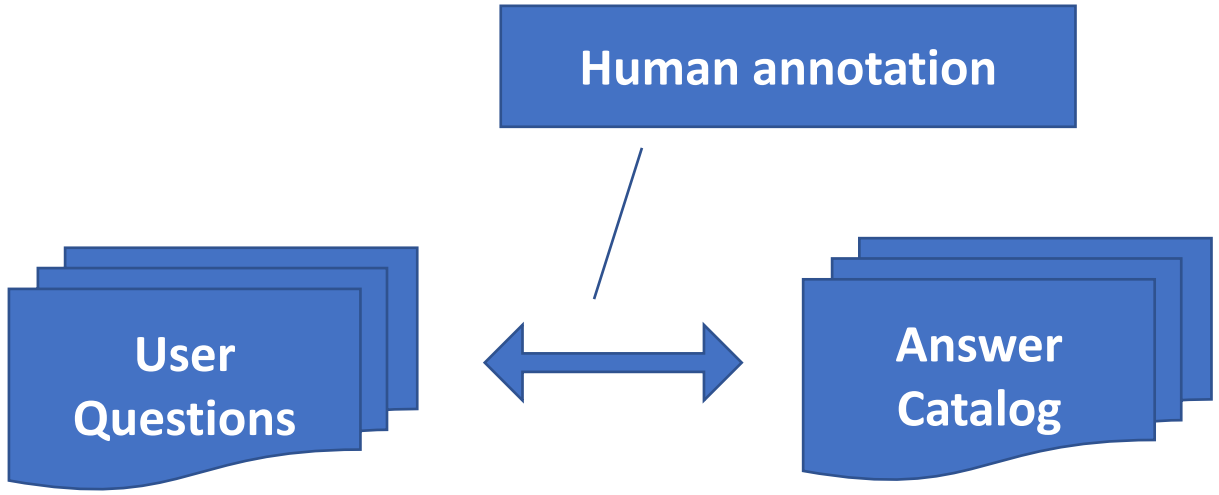
- Before you start annotating:
  - Have a clear picture of what you want to accomplish
  - Think about caveats of the task you want to solve;
    and how to handle them
  - Does it make sense what you planned? Try it out yourself!

- Make the task simple (or decompose complex ones)
  - Less ambiguity -> better agreement between annotators
  - Easier for annotators to understand
    - Reading a 20-page manual before starting is not motivational
  - Easier to analyze afterwards
  - Annotations might be used for more than one goal if re-combined

Task Design – Example FAQ Annotation

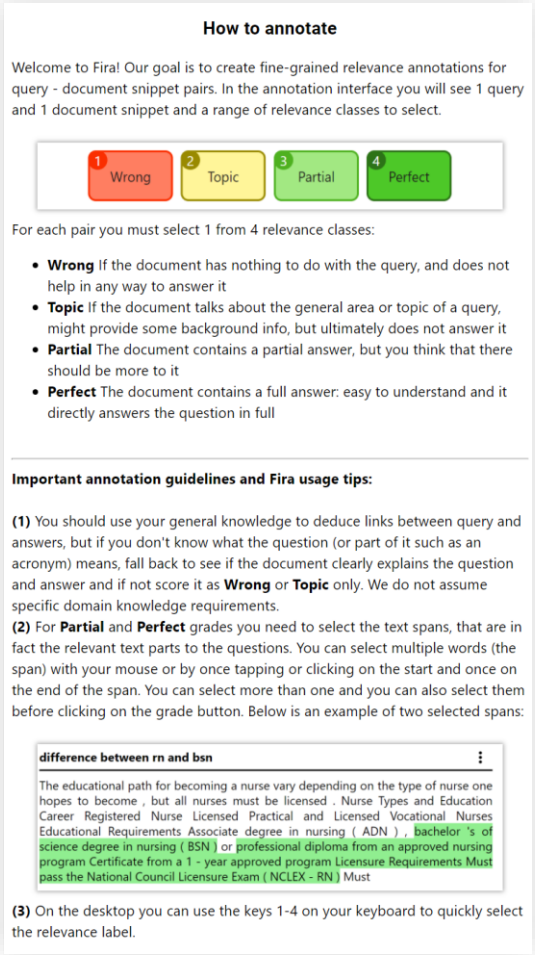
- The answer catalog can contain hundreds of entries, making an answer look-up time-consuming
- Idea: Annotate questions with the same-intent as a group
  - Groups of questions are found by unsupervised semantic text similarity
- Result: More efficient annotation

Annotation Time	
ONE-BY-ONE Avg.	42 seconds
GROUP-WISE Avg.	25 seconds



Task Guidelines and Examples

- Guidelines are a textual description of what an annotator should do
  - Clear description of the overall goal and individual tasks
  - Short and concise
- Incorporate examples for different scenarios
  - Practically demonstrate of what should be done



How to annotate

Welcome to Fira! Our goal is to create fine-grained relevance annotations for query - document snippet pairs. In the annotation interface you will see 1 query and 1 document snippet and a range of relevance classes to select.

1 Wrong 2 Topic 3 Partial 4 Perfect

For each pair you must select 1 from 4 relevance classes:

- **Wrong** If the document has nothing to do with the query, and does not help in any way to answer it
- **Topic** If the document talks about the general area or topic of a query, might provide some background info, but ultimately does not answer it
- **Partial** The document contains a partial answer, but you think that there should be more to it
- **Perfect** The document contains a full answer: easy to understand and it directly answers the question in full

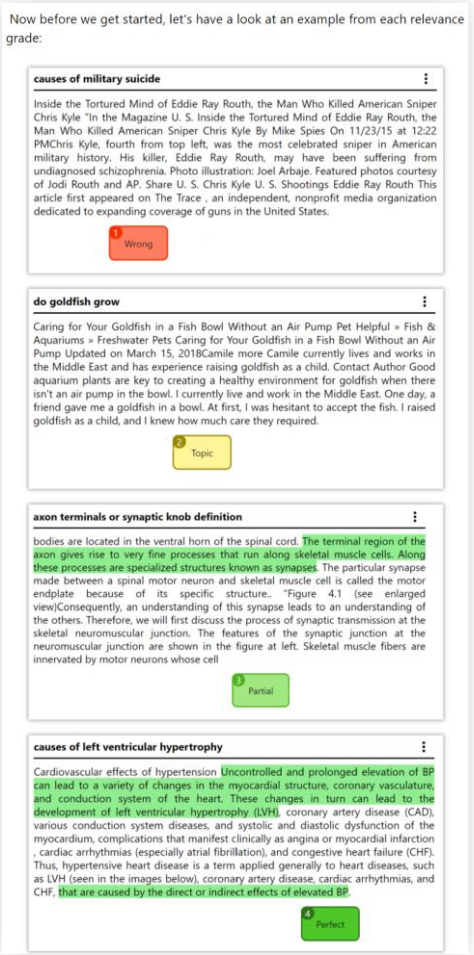
Important annotation guidelines and Fira usage tips:

(1) You should use your general knowledge to deduce links between query and answers, but if you don't know what the question (or part of it such as an acronym) means, fall back to see if the document clearly explains the question and answer and if not score it as **Wrong** or **Topic** only. We do not assume specific domain knowledge requirements.
(2) For **Partial** and **Perfect** grades you need to select the text spans, that are in fact the relevant text parts to the questions. You can select multiple words (the span) with your mouse or by once tapping or clicking on the start and once on the end of the span. You can select more than one and you can also select them before clicking on the grade button. Below is an example of two selected spans:

difference between rn and bsn

The educational path for becoming a nurse vary depending on the type of nurse one hopes to become , but all nurses must be licensed . Nurse Types and Education Career Registered Nurse Licensed Practical and Licensed Vocational Nurses Educational Requirements Associate degree in nursing (ADN) , bachelor 's of science degree in nursing (BSN) or professional diploma from an approved nursing program Certificate from a 1 - year approved program Licensure Requirements Must pass the National Council Licensure Exam (NCLEX - RN) Must

(3) On the desktop you can use the keys 1-4 on your keyboard to quickly select the relevance label.



Now before we get started, let's have a look at an example from each relevance grade:

causes of military suicide

Inside the Tortured Mind of Eddie Ray Routh, the Man Who Killed American Sniper Chris Kyle "In the Magazine U. S. Inside the Tortured Mind of Eddie Ray Routh, the Man Who Killed American Sniper Chris Kyle By Mike Spies On 11/23/15 at 12:22 PMChris Kyle, fourth from top left, was the most celebrated sniper in American military history. His killer, Eddie Ray Routh, may have been suffering from undiagnosed schizophrenia. Photo illustration: Joel Arbaje. Featured photos courtesy of Jodi Routh and AP. Share U. S. Chris Kyle U. S. Shootings Eddie Ray Routh This article first appeared on The Trace , an independent, nonprofit media organization dedicated to expanding coverage of guns in the United States.

1 Wrong

do goldfish grow

Caring for Your Goldfish in a Fish Bowl Without an Air Pump Pet Helpful » Fish & Aquariums » Freshwater Pets Caring for Your Goldfish in a Fish Bowl Without an Air Pump Updated on March 15, 2018Camile more Camile currently lives and works in the Middle East and has experience raising goldfish as a child. Contact Author Good aquarium plants are key to creating a healthy environment for goldfish when there isn't an air pump in the bowl. I currently live and work in the Middle East. One day, a friend gave me a goldfish in a bowl. At first, I was hesitant to accept the fish. I raised goldfish as a child, and I knew how much care they required.

2 Topic

axon terminals or synaptic knob definition

bodies are located in the ventral horn of the spinal cord. The terminal region of the axon gives rise to very fine processes that run along skeletal muscle cells. Along these processes are specialized structures known as synapses. The particular synapse made between a spinal motor neuron and skeletal muscle cell is called the motor endplate because of its specific structure.. "Figure 4.1 (see enlarged view)Consequently, an understanding of this synapse leads to an understanding of the others. Therefore, we will first discuss the process of synaptic transmission at the skeletal neuromuscular junction. The features of the synaptic junction at the neuromuscular junction are shown in the figure at left. Skeletal muscle fibers are innervated by motor neurons whose cell

3 Partial

causes of left ventricular hypertrophy

Cardiovascular effects of hypertension Uncontrolled and prolonged elevation of BP can lead to a variety of changes in the myocardial structure, coronary vasculature, and conduction system of the heart. These changes in turn can lead to the development of left ventricular hypertrophy (LVH), coronary artery disease (CAD), various conduction system diseases, and systolic and diastolic dysfunction of the myocardium, complications that manifest clinically as angina or myocardial infarction , cardiac arrhythmias (especially atrial fibrillation), and congestive heart failure (CHF). Thus, hypertensive heart disease is a term applied generally to heart diseases, such as LVH (seen in the images below), coronary artery disease, cardiac arrhythmias, and CHF, that are caused by the direct or indirect effects of elevated BP.

4 Perfect

Conduct a Test Run

- To see, if everything works as intended:
  - Does my setup work?
  - Is the output data really in the right format and not all over the place?
  - Can annotators understand the task?
  - Is the resulting label quality sufficient?

- Performed on a small percentage of all samples
- Possible with some annotators or all annotators
- After the test-run the full-scale run can be started

Crowdsourcing

- Marketplace platforms (f.e. Amazon Mechanical Turk, Figure Eight) allow you to post tasks; workers can decide to complete them
  - You can set requirements on country, language, education level etc..
- Allows for very large scale of annotations
  - Scaling costs a lot of money, and time to set up and supervise
- You need a lot of focus on quality control
  - Cheating is common
  - Task design needs to be done so that random cheating is not possible
  - Larger number of majority voting is necessary (drives up costs)

Majority Voting

- Let different annotators judge the same examples
  - Pro: Higher quality of resulting data
  - Con: x N more cost in terms of effort
- With randomness, don't create groups of annotators
- Allows us to monitor the quality of individual annotators
  - Some disagreement is always there, but if one always disagrees, this might warrant more investigation
- Allows us to better understand the task and how to interpret agreement ratios
  - Is overall agreement high or low?

Evaluate Annotation Quality

- Measure the label-quality of annotators
  - Based on a few samples labeled by an expert
  - Based on the inter-annotation agreement between annotators
- Inter-annotator agreement (IAA) tells us if humans agree on results
  - Important prerequisite for an algorithmic solution!
  - Cohen's Kappa is usually used to compute the IAA between 2 annotators.
  - Fleiss' Kappa can be used to compute the IAA between 2 or more annotators.
- Kappa result interpretation:
  - 80%+ → strong agreement; 60%-79% → moderate; 40%-59% → weak
  - Caution: Interpretation is subjective + depends on the task

Pooling in Information Retrieval

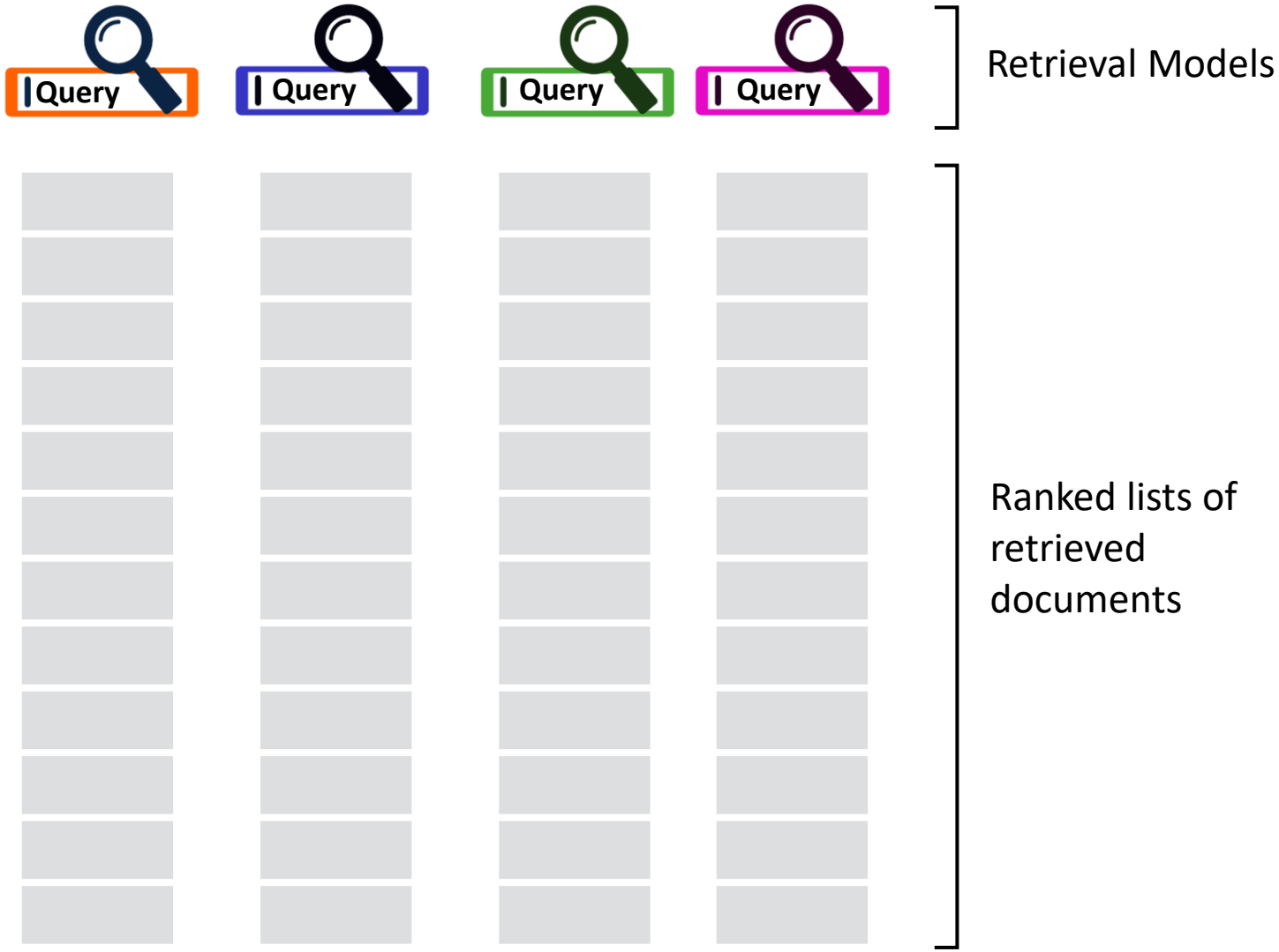
- In IR we need systems to retrieve from a full (large) collection
  - Otherwise, we would be too far away from real-world conditions
- We can’t annotate all documents per query (this would mean millions of annotations (where most are not relevant)
- So, we get retrieval models to help us!
  - Retrieval systems create candidate documents for us to annotate
  - A diverse pool of different systems is better for re-use
  - Using proven models, gives us confidence, that we have at least some relevant results in those results
  - Allows us to drastically reduce annotation time

Pooling Process

Current Step:

- We gather a diverse set of retrieval models and index the same collection
- We select a list of queries and repeat this pooling process

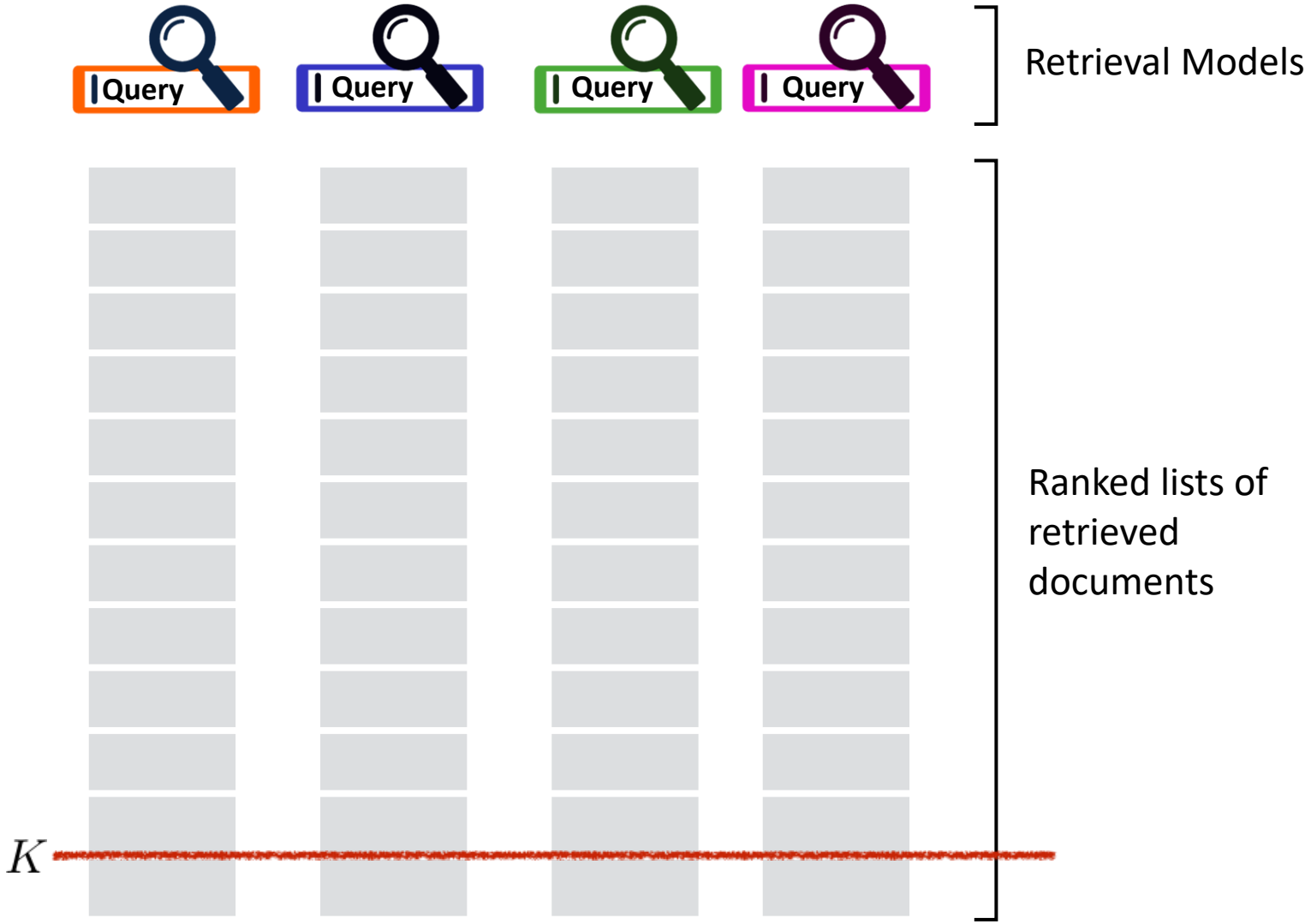
- For each query:
  - Let each system retrieve their top ranked documents / passages / elements



Pooling Process

Current Step:

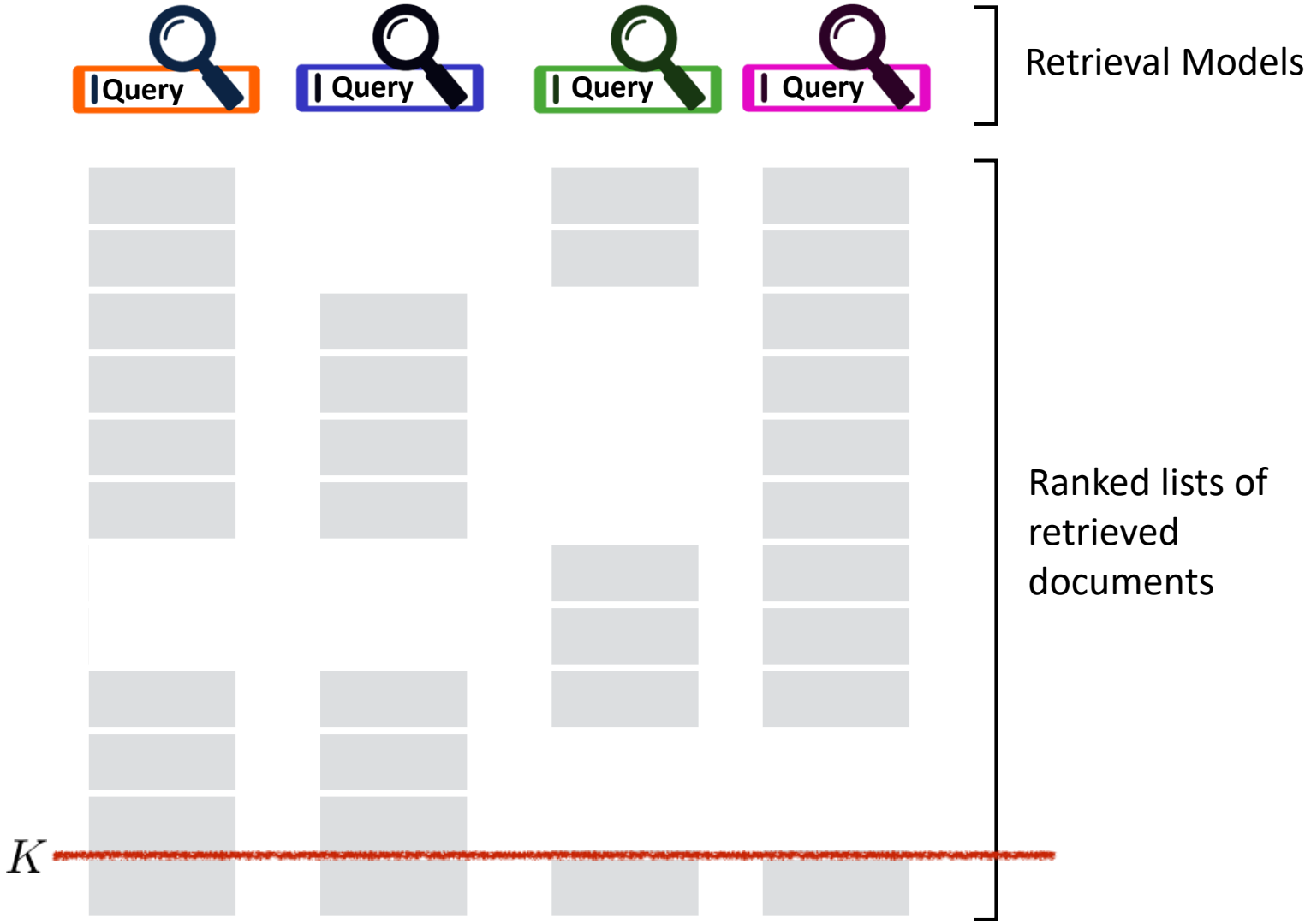
- Choose a cutoff point K, where we guarantee to include results per system
- This is a simple method that allows us to fully compare systems for all metrics @K
- More complicated and improved methods exist, but we keep it simple for now



Pooling Process

Current Step:

- Creating our pool to annotate, by removing duplicates
  - This makes pooling more effective than simply relying on a single system at a time
- This set of query-doc pairs can now be annotated

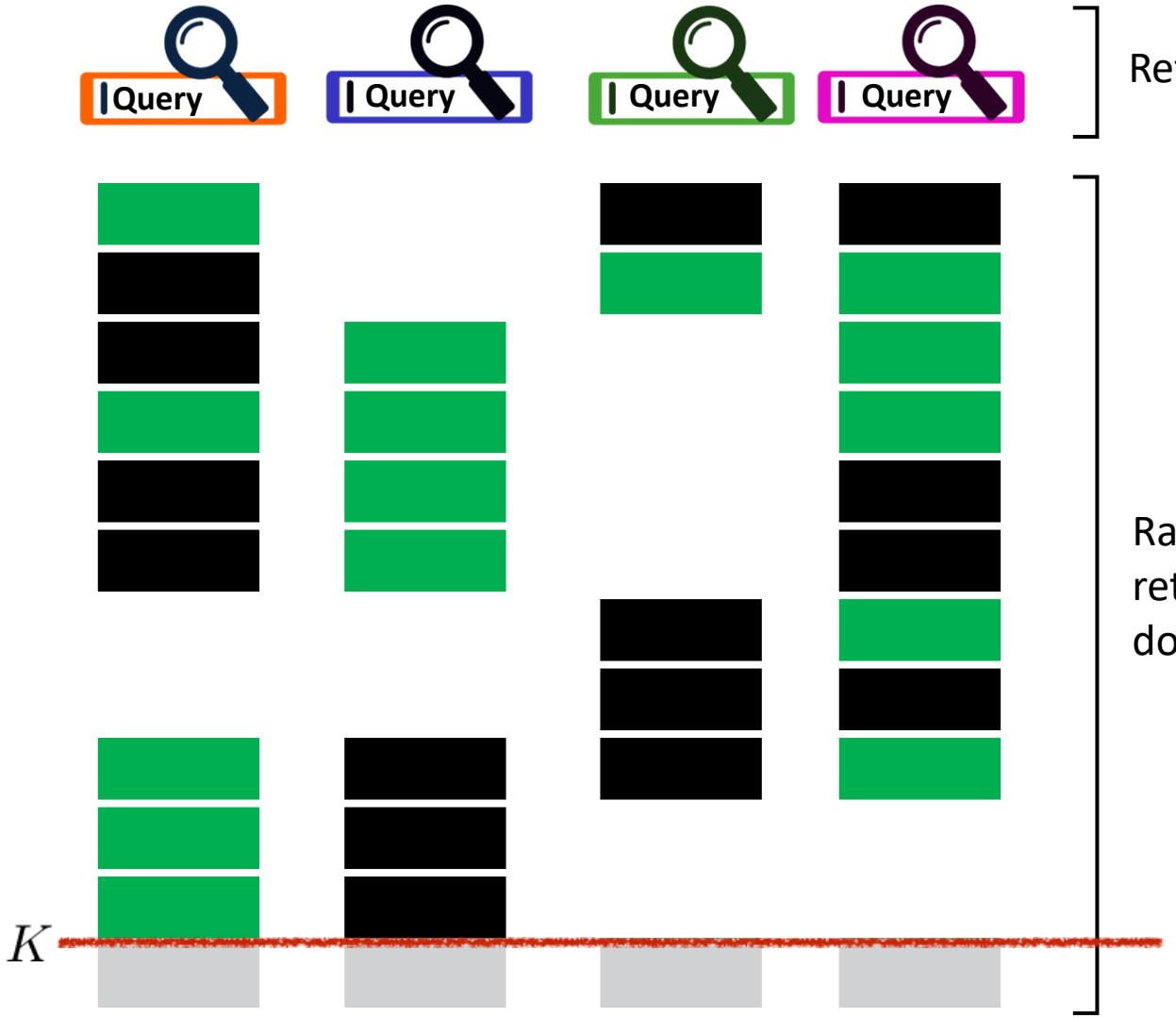


Pooling Process

Current Step:

- We annotate pairwise (query, document)

- The exact task depends on your task design (relevance grading, text selection, description, etc…)

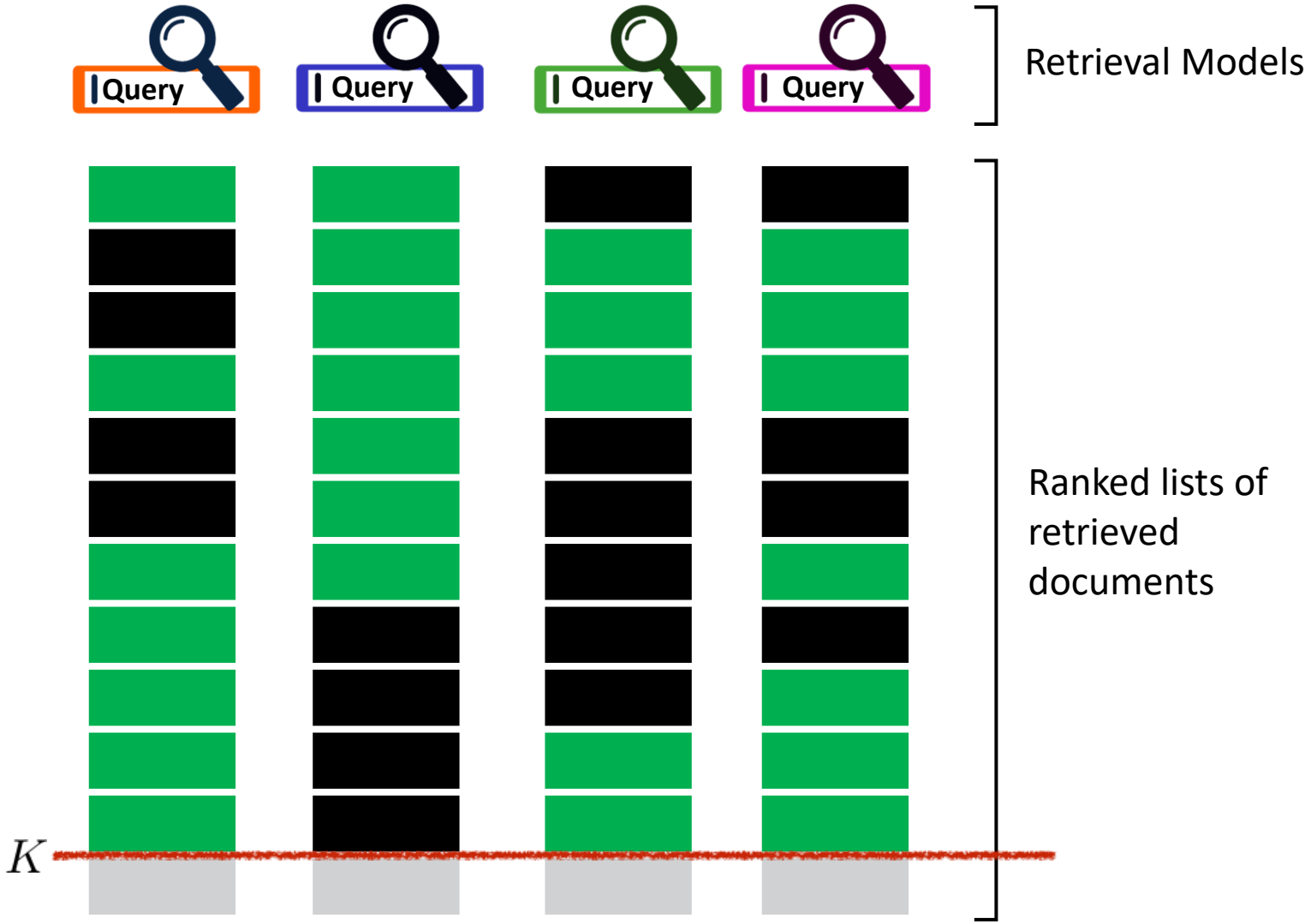


Pooling Process

Current Step:

- After every pooled pair is annotated, we map the labels back to the duplicated documents and receive our full list per system
- Now we can use ranked list evaluation metrics to compare our systems

Yeah 🎉 !



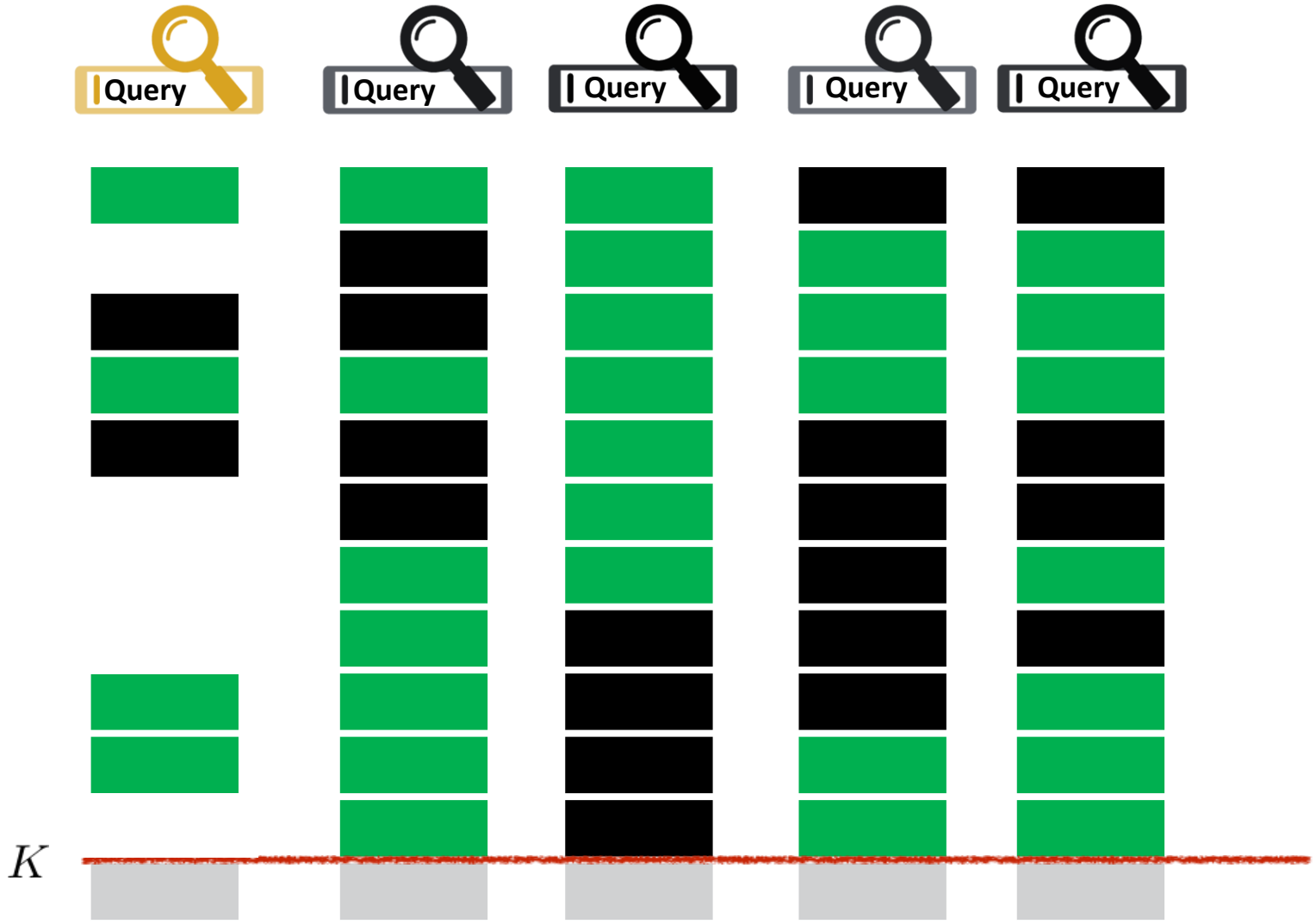
Pool Bias

Scenario:

- After the annotation campaign you create a new retrieval model and want to re-use the annotations
  - Without re-annotating anything

Result:

- If new documents are retrieved (likely) than we have missing pairs, and we must assume them to be non-relevant

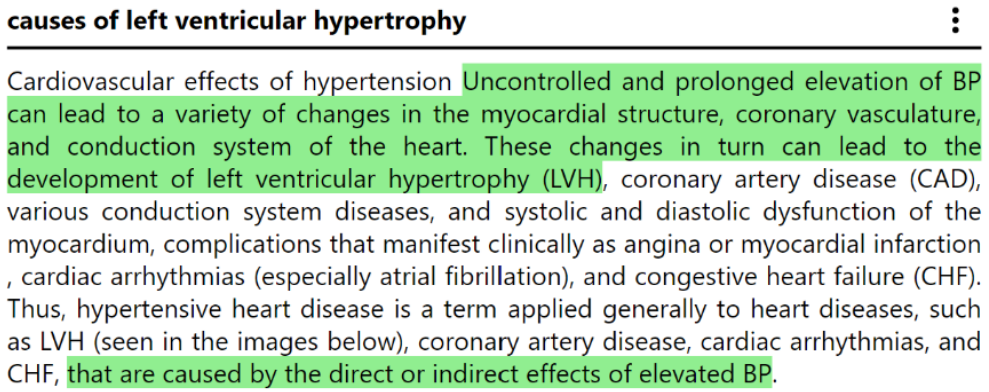


Annotating for Recall

- Recall depends on the judgement of “all” relevant items
- What happens if we don’t know that there is a relevant item
  - Test collections depend on pre-selection of candidate documents
  - Relevant items might be missing for example because of vocabulary mismatch
- Either we actively work on this problem with iterative annotation cycles
  - + potential active learning -> HiCAL is an annotation system that integrates it
- Or: If our test collection is not prepared for high recall, we should at least be aware of its limitations when interpreting results

The HiCAL tool https://hical.github.io/index.html

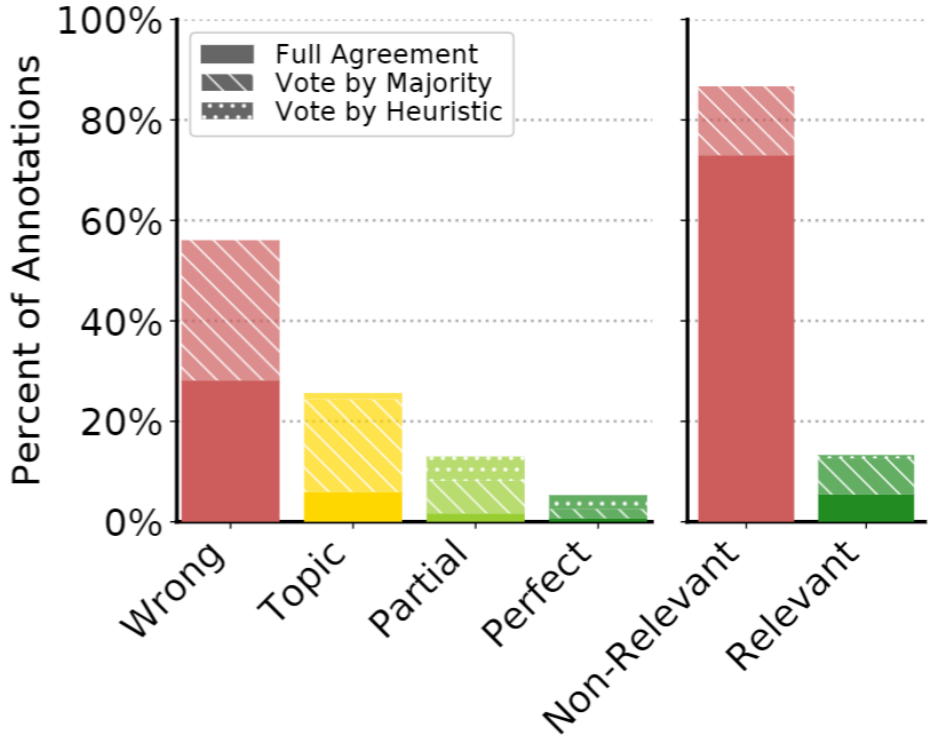
FiRA: Fine-Grained Relevance Annotations



- We support fulltext based neural ranking models with fine-grained relevance annotations
- FiRA is an augmentation of TREC'19 DL document ranking annotations
- We annotated every highly relevant document of the 43 TREC queries
  - Total: 24,197 query-passage pairs
  - Using 4 classes graded relevance

Hofstätter, Zlabinger, Sertkan, Schröder, Hanbury
Fine-Grained Relevance Annotations for Multi-Task Document Ranking and Question Answering In Proc. of CIKM 2020

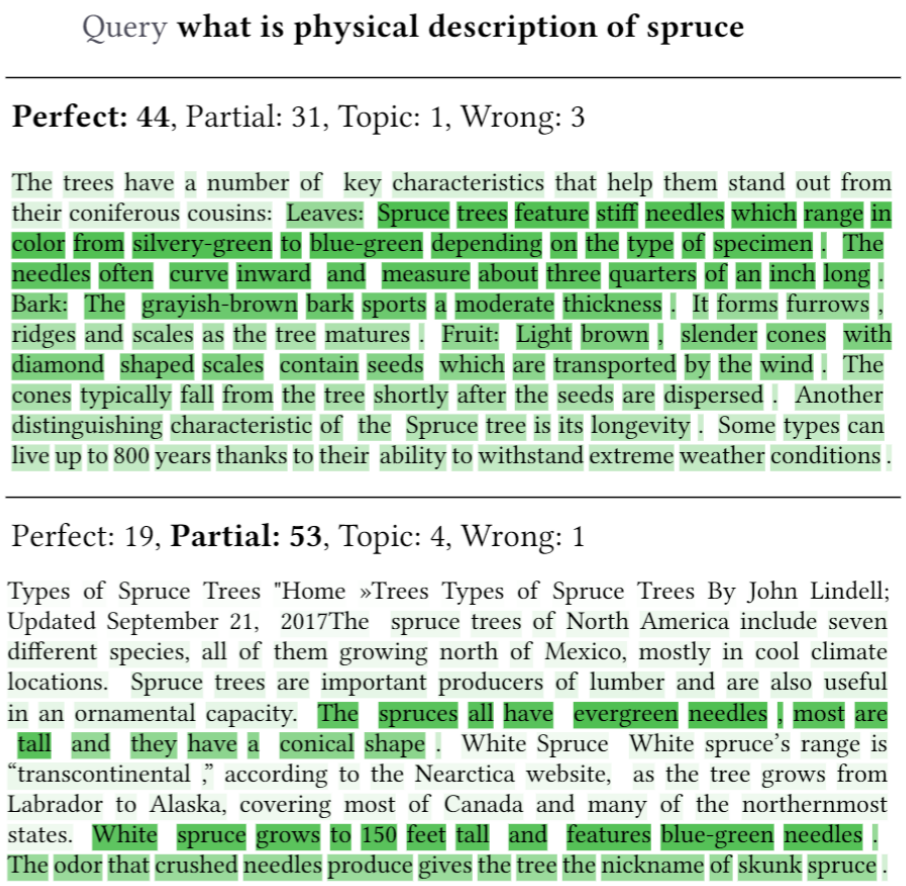
FiRA-DL‘19: Relevance Classes & Majority Voting



- Every document is relevant overall (judged by TREC)
- We find many passages are not relevant
  - Intuition for passage score aggregation techniques
- Majority voting is important, as most relevant passages are not unanimously decided

Hofstätter, Zlabinger, Sertkan, Schröder, Hanbury
Fine-Grained Relevance Annotations for Multi-Task Document Ranking and Question Answering In Proc. of CIKM 2020

FiRA-DL'19: Relevance Uncertainty



Query **what is physical description of spruce**

Perfect: 44, Partial: 31, Topic: 1, Wrong: 3

The trees have a number of key characteristics that help them stand out from their coniferous cousins: Leaves: Spruce trees feature stiff needles which range in color from silvery-green to blue-green depending on the type of specimen . The needles often curve inward and measure about three quarters of an inch long . Bark: The grayish-brown bark sports a moderate thickness . It forms furrows , ridges and scales as the tree matures . Fruit: Light brown , slender cones with diamond shaped scales contain seeds which are transported by the wind . The cones typically fall from the tree shortly after the seeds are dispersed . Another distinguishing characteristic of the Spruce tree is its longevity . Some types can live up to 800 years thanks to their ability to withstand extreme weather conditions .

Perfect: 19, **Partial: 53**, Topic: 4, Wrong: 1

Types of Spruce Trees "Home »Trees Types of Spruce Trees By John Lindell; Updated September 21, 2017The spruce trees of North America include seven different species, all of them growing north of Mexico, mostly in cool climate locations. Spruce trees are important producers of lumber and are also useful in an ornamental capacity. The spruces all have evergreen needles , most are tall and they have a conical shape . White Spruce White spruce's range is "transcontinental ," according to the Nearctica website, as the tree grows from Labrador to Alaska, covering most of Canada and many of the northernmost states. White spruce grows to 150 feet tall and features blue-green needles . The odor that crushed needles produce gives the tree the nickname of skunk spruce .

- Selected passages are annotated by all
  - We can qualitatively study the subjectivity of relevance annotations
- Most agree on the general relevance (between 2 classes)
  - Shows relevance as a distribution not as fixed class definition
- Heatmap of selected regions shows:
  1. Uncertainty in upper example
  2. Certain what is relevant in lower passage

Hofstätter, Zlabinger, Sertkan, Schröder, Hanbury
Fine-Grained Relevance Annotations for Multi-Task Document Ranking and Question Answering In Proc. of CIKM 2020

Analyzing Biases

Analyzing the downstream impact of existing bias in the data

The Many Types of Biases

- Taking biased text as input produces bias representations
- Bias can take many forms, such as gender or racial bias
- Can have a decisive effect on peoples lives (unnoticed to responsible people)
  - Hiring decision
  - Predictive policing
  - Recommendation algorithms that marginalize minorities
- Thinking about downstream effects of ML is important
- Can also come in less dangerous forms such as learning to ignore positions of text in a document

Bias in Word Embeddings

- Word embeddings are trained on large scale unlabeled text
  - For example: Wikipedia
- If training data is biased -> vectors are biased as well

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

- Word2Vec trained on Wikipedia contains significant gender bias

Bolukbasi, Tolga, et al. *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings* NeurIPS 2016

- Various methods have been proposed to debias word embeddings
  - Might not be truly effective and only cover up the bias

Gonen and Goldberg; *Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them*; NAACL 2019

Bias in (Large) Language Models

- *Stochastic Parrot* paper and subsequent refusal by Google to let it publish created quite a stir in the community in early 2021
- Opinionated paper with a great survey of the state of the field
- Bender et al. argue:
  - Exploding dataset sizes & outdated text result in unchecked biases
  - Static stop-gap measures to remove keywords hides minority communities
  - LMs are only seemingly coherent – not true language understanding
  - LMs amplify training bias in real world use
  - For a mindset that weaves in responsibility in the whole research process

Bender, Gebru, McMillan-Major, Shmitchell On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT 2021 http://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf

Social Impact of Ranking

- Recommendations are optimized for time spent on a platform (= revenue)
- Easy to fall down the rabbit hole – because scandalous & “click here to find the truth” videos keeps you on the platform
- Multilingual problems: Manually blocking English content does not automatically translate to other languages

How YouTube Radicalized Brazil https://www.nytimes.com/2019/08/11/world/americas/youtube-brazil.html

Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, Ed Chi. 2019. Recommending What Video to Watch Next: A Multitask Ranking System. RecSys ’19

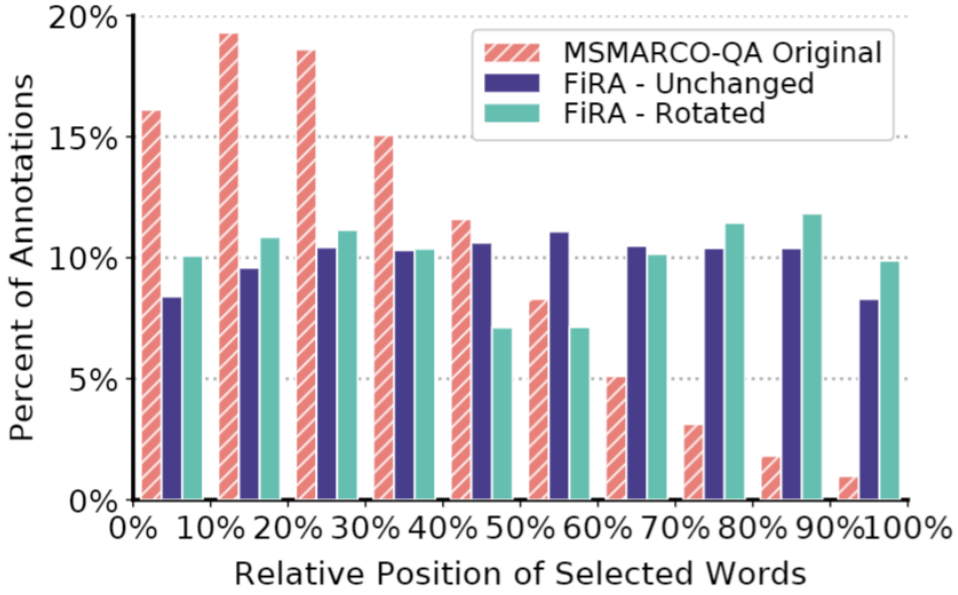
Technical Term-Position Bias

- In a bag-of-words model terms don’t have positions
  - This problem didn’t exist there
- In most neural models: a term has positional information associated
- Models can learn to over-emphasizes the position information
  - Which puts the same information later in a passage/document at a disadvantage
  - A problem for generalization & a limitation that we should be aware of

---

- *Different from*: Result Position Bias
  - Where Users tend to click on first results
    (even if relevant ones are further down the list

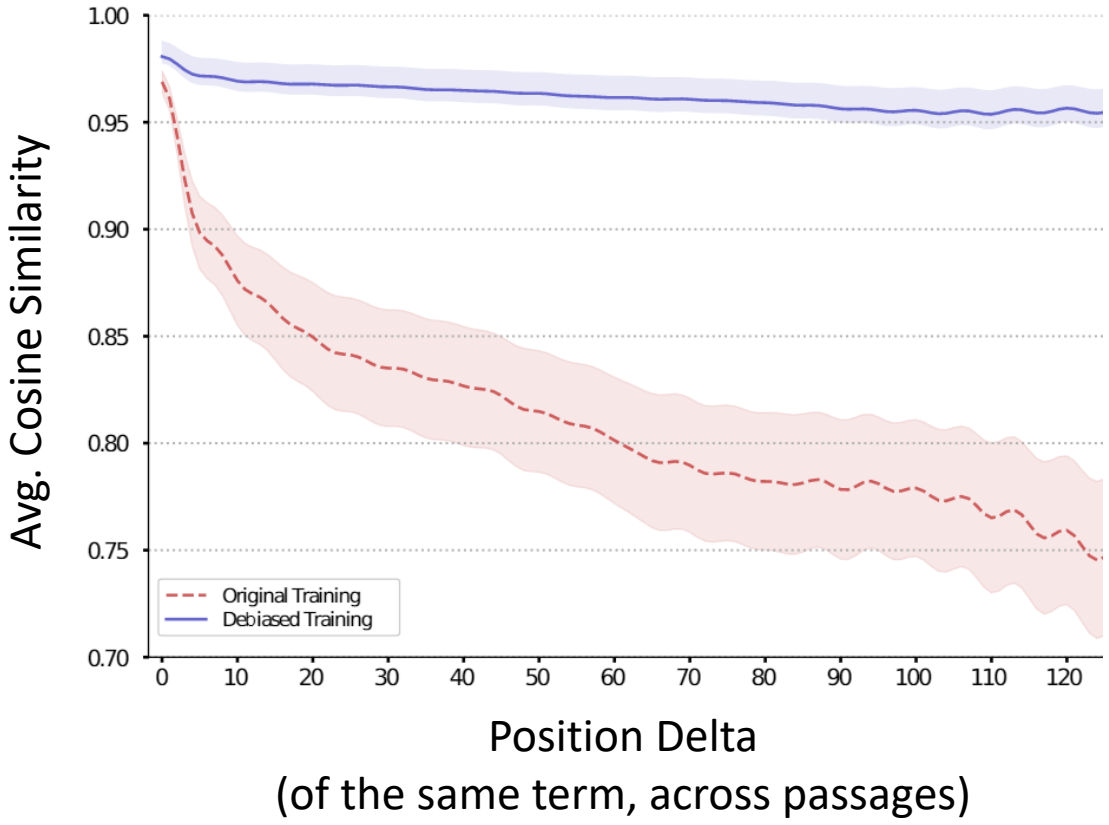
Where is the passage-level relevance?



- We observe a bias in MSMARCO-QA (the basis for TREC DL)
  - Answers tend to involve words at the beginning of a passage
- We conducted a rotation-based controlled experiment to see if we can replicate this phenomenon
  - Disclaimer: Not the same passages, but same domain
- Results: almost uniformly selected relevant words
  - Slight drop around the corners

Hofstätter, Zlabinger, Sertkan, Schröder, Hanbury
Fine-Grained Relevance Annotations for Multi-Task Document Ranking and Question Answering In Proc. of CIKM 2020

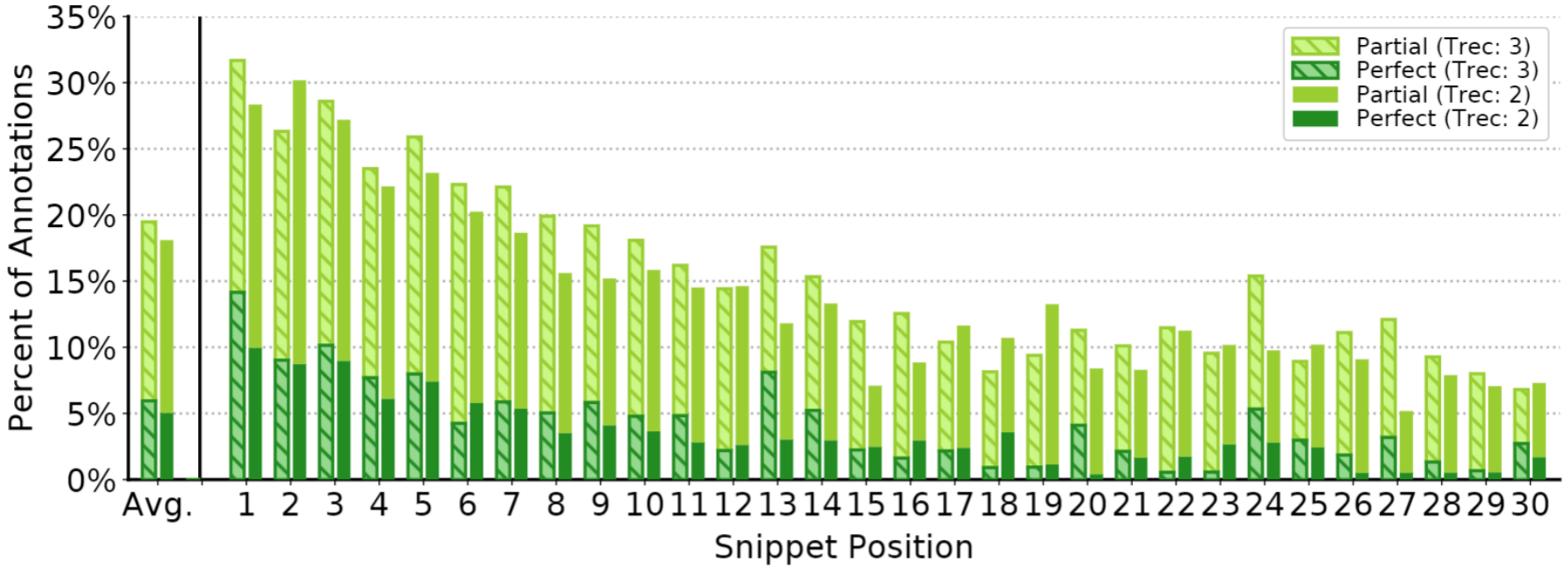
Tracking Position Influence in Transformers



- Transformers use positional information of terms (via an embedding)
- Influence is trained (TK & BERT models)
- If data is biased, we find that:
  - Models overfit on importance of earlier position
  - Models (TK) much stronger incorporate position in term representation
- De-bias data: Phenomenon vanishes

Hofstätter et al. *Mitigating the Position Bias of Transformer Models in Passage Re-Ranking* In Proc. of ECIR 2021

Where is the document-level relevance?



- Likelihood of a relevant passage is higher at the beginning
  - Although not exclusively for the first passage

Hofstätter, Zlabinger, Sertkan, Schröder, Hanbury
Fine-Grained Relevance Annotations for Multi-Task Document Ranking and Question Answering In Proc. of CIKM 2020

Summary: Test Collections

1. IR Test Collections are incomplete, we must deal with uncertainty
2. Judgement pairs should use pooling of many diverse system results
3. Bias exists everywhere in different forms, important to be aware of it

1. IR Test Collections are incomplete, we must deal with uncertainty
2. Judgement pairs should use pooling of many diverse system results
3. Bias exists everywhere in different forms, important to be aware of it

Thank You