Deep Neural Ranking Models for Argument Retrieval

Master's Thesis by Saeed Entezari
Referees: Prof. Stein, PD. Dr. Jakoby
Supervisor: Michael Völske

Faculty of Media
Bauhaus Universität Weimar

September 16, 2020

Bauhaus-
Universität
Weimar

Agenda

Bauhaus-Universität Weimar

Abstract

- Task: Ranking arguments in a collection for the given query
- Contributions
  - **RQ1.** How to shape useful training and validation set fit for the task of ad-hoc retrieval using the collection?
  - **RQ2.** Using neural ranking models that have shown good performance in ad-hoc retrieval tasks in the argument retrieval
    - **RQ2.1.** Interaction-focused vs. representation-focused?
    - **RQ2.2.** Static embedding vs. contextualized embedding?
    - **RQ2.3.** Typical Neural ranking model vs. End-to-End?
  - **RQ3.** How to aggregate model results? Which strategy to use and what we require for doing so?

Outline

Bauhaus-Universität Weimar

Why Argument Retrieval

- Different types of opinions toward controversial topics
- Getting an overview of every opinion is an exhaustive and time consuming task
- Automated decision making
- Opinion Summarization

What is Argument

- Argumentation unit which is composed of a claim (conclusion) and its premise [Rieke et al.(1997)Rieke, Sillars, and Peterson]
- Use the premises of one claim to support or attack other claims
- claims could be a word, phrase or a sentence
- Premises are texts composed of multiple sentences or paragraphs

Bauhaus-Universität Weimar

Argument components

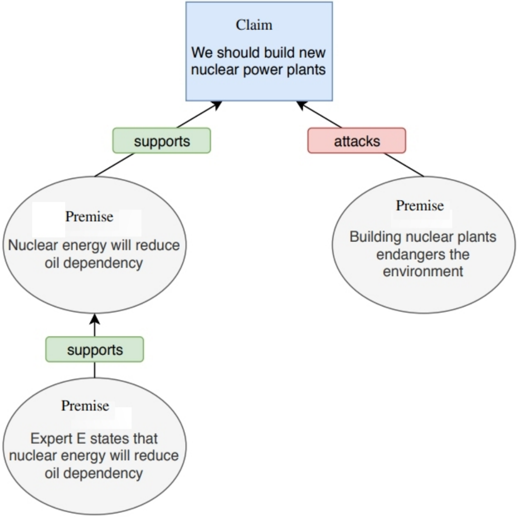


Figure: The relation between the argument units ([Dumani(2019)])

Outline

Bauhaus-
Universität
Weimar

Ad-hoc Retrieval Task

- Heterogeneous Ranking Task
  - Typically queries are of a shorter length
  - Documents are longer texts
- Given the query, the task is to rank the existing documents in the collection
- Query Relevance Files: soft similarity scores for query-document pairs derived from the query log or click through data
  - qrel makes training the models possible

! We do not have the qrel file in our dataset

Bauhaus-Universität Weimar

Outline

Args.me Corpus

387740 annotated arguments in total from crawling 4 debate portals (json format):

- Debatewise (14000 arguments)
- IDebate.org (13000 arguments)
- Debatepedia (21000 arguments)
- Debate.org (338000 arguments)

Information for each argument:

- unique ID
- claim
- premise
- source of crawling
- time of crawling
- stance of premise regard to claim

Bauhaus-Universität Weimar

Outline

Bauhaus-Universität Weimar

Preprocessing and Visualisation: Claims

- Forming normalized claims
  - punctuation removal and case sensitivity
  - stop words removal
- Visualization and Statics
  - 66473 unique claims
  - 29970 unique tokens

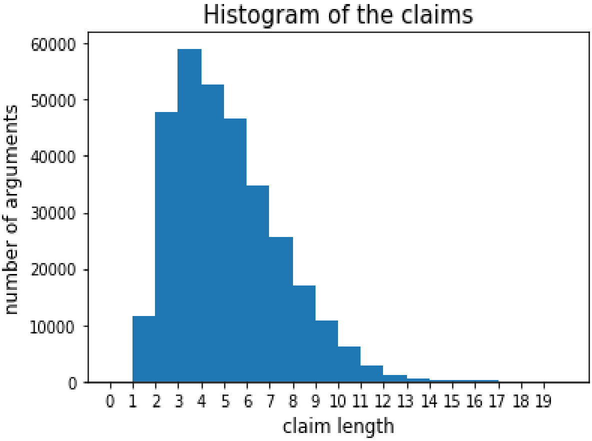


Figure: Histogram of the unique claims based on the number of tokens

Bauhaus-Universität Weimar

Preprocessing and Visualisation: Claims

Table: Normalized claims with the highest number of premises

norm cons	number of premises
abortion	2401
gay marriage	1259
rap battle	1256
god exists	942
death penalty	941

Bauhaus-Universität Weimar

Preprocessing and Visualisation: Premises

- Tokenizing punctuation
  - for static embedding: god exists.$\Rightarrow$ god exists <PERIOD>
  - for contextualized embedding is not required!
- Removing consecutive repetitive tokens
  - !!!!!!!! $\Rightarrow$ <EXCLAMATIONMARK>
  - yes yes yes $\Rightarrow$ yes
- Mapping digits to words
  - 95 $\Rightarrow$ ninety-five
- Removing the URLs
  - http://example.net/achiever.html?boy=armyauthority=beginner

Preprocessing and Visualisation: Premises

- Statistics of the premises:
  - vocabulary size: 586796
  - 85% of the premises have the length of less than 200 words
- Arguments with the premise length of less than 15 tokens are removed

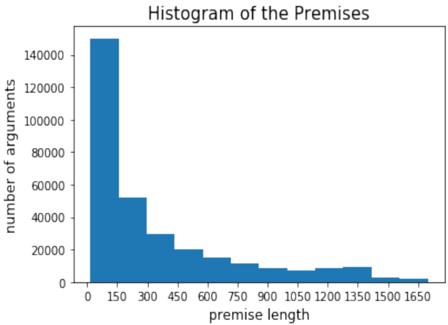


Figure: Histogram of the premises based on their length (number of tokens separated by white space)

Outline

Bauhaus-Universität Weimar

Learning to Rank

- Learning goal: related documents over the unrelated ones
- Pairwise hinge cost function
- Relevant and irrelevant Query-Document pairs are required and are missing in the corpus
- A model to produce the similarity scores (We use Deep ranking models)

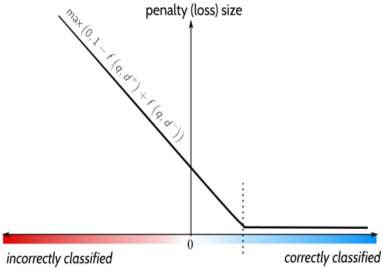


Figure: Hinge as a pairwise cost function

Bauhaus-Universität Weimar

Binary Query Relevance Generation

RQ.1: Useful dataset for ad-hoc task

- Distant Supervision Approach
  - Claims ⇒ Queries
  - Premises ⇒ Related Documents
- Unrelated premise for each query
  - qrel files contain also unrelated query-document pairs
  - similarity measure: fuzzy similarity
  - premise of an unrelated claims could be an unrelated document to our claims
- A binary query relevance is formed ⇒ Exploitation of deep ranking models in the context of argument retrieval is possible now!

Bauhaus-Universität Weimar

Dataset Ready for Ad-hoc Task

Data collection ready for the ad-hoc task (for static and contextualized embedding) with the following columns:
Important Note: Different arguments may have same claims and different premsies

id	claim	norm-claim	premise	unrelated id	unrelated premise
arg_1	...	...	...	...	...
arg_2	...	...	...	...	...

Bauhaus-Universität Weimar

Outline

Bauhaus-Universität Weimar

Training and Validation Sets

- Training set: 312248 arguments with one unrelated documents each
- Validation set: 4885 arguments: 20 unrelated documents each

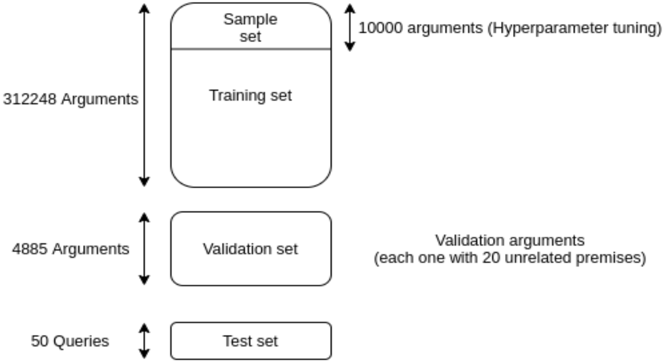


Figure: Different datasets and their number of arguments

Bauhaus-Universität Weimar

Validation Arguments

RQ.1: Forming an appropriate training and validation dataset

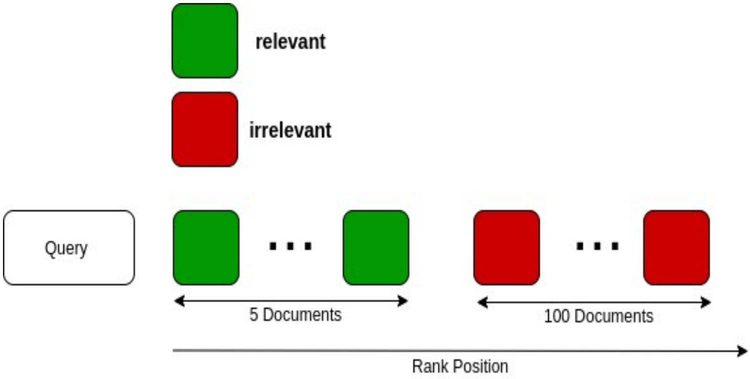


Figure: An ideal ranking for a validation query

Bauhaus-Universität Weimar

Outline

Neural Ranking Models

- **Applications**: ad-hoc retrieval, question answering, automatic conversation
- **Similarity of input pairs (query q, document d)**:

$$f(q,d) = g(\psi(q), \phi(d), \eta(q,d)) \tag{1}$$

  - ψ(q), ϕ(d) and η(q,d) are representation of the texts q, d and the pair of q and d respectively
- Representation-focused and Interaction-focused networks

Exploited Models

Table: Models

Model	type	embedding	re-rank
GRU	rep	static	yes
DRMM	int	static	yes
KNRM	int	static	yes
CKNRM	int	static	yes
Vanilla BERT	int	contx	yes
DRMM BERT	int	contx	yes
KNRM BERT	int	contx	yes
SNRM	rep	static	no

Bauhaus-Universität Weimar

Siamese Network

Model	type	embedding	re-rank
GRU	rep	static	yes
DRMM	int	static	yes
KNRM	int	static	yes
CKNRM	int	static	yes
Vanilla BERT	int	contx	yes
DRMM BERT	int	contx	yes
KNRM BERT	int	contx	yes
SNRM	rep	static	no

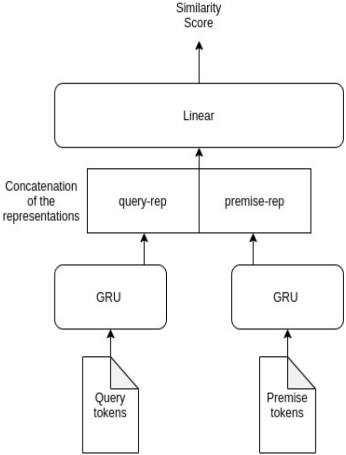


Figure: Similarity scores using recurrent neural network

Bauhaus-Universität Weimar

DRMM: Deep Relevance Matching Model

Model	type	embedding	re-rank
GRU	rep	static	yes
DRMM	int	static	yes
KNRM	int	static	yes
CKNRM	int	static	yes
Vanilla BERT	int	contx	yes
DRMM BERT	int	contx	yes
KNRM BERT	int	contx	yes
SNRM	rep	static	no

- Interaction-focused network
- Matching histogram of the query and document token embedding as the input to a fully connected network for similarity score

Bauhaus-Universität Weimar

KNRM: Kernel-based Neural Ranking Model

Model	type	embedding	re-rank
GRU	rep	static	yes
DRMM	int	static	yes
KNRM	int	static	yes
CKNRM	int	static	yes
Vanilla BERT	int	contx	yes
DRMM BERT	int	contx	yes
KNRM BERT	int	contx	yes
SNRM	rep	static	no

- Another strategy for encoding the input pair interaction
- Forming translation matrix: elements are the cos similarity of the term embedding
- Applying the RBF as the kernels and forming the input features for fully connected network
- A linear layer learns the score similarity of the input pairs

CKNRM: Covolutional KNRM

Model	type	embedding	re-rank
GRU	rep	static	yes
DRMM	int	static	yes
KNRM	int	static	yes
CKNRM	int	static	yes
Vanilla BERT	int	contx	yes
DRMM BERT	int	contx	yes
KNRM BERT	int	contx	yes
SNRM	rep	static	no

- Using **Convolutional** windows to get a representation of document and query n-grams
- Forming cross-match layer instead of translation matrix for encoding the interaction of the n-grams in document and query
- The idea of applying the RBF and linear layer for computing the similarity score remain the same!

Bauhaus-Universität Weimar

Ranking Models with Contextualized Embedding

Model	type	embedding	re-rank
GRU	rep	static	yes
DRMM	int	static	yes
KNRM	int	static	yes
CKNRM	int	static	yes
Vanilla BERT	int	contx	yes
DRMM BERT	int	contx	yes
KNRM BERT	int	contx	yes
SNRM	rep	static	no

- BERT base uncased as the contextualized embedding
- Embedding dimension of the tokens: 768
- Ranking models used with BERT:
  - Vanilla-BERT: linear layer at the top of BERT network
  - BERT and DRMM
  - BERT and KNRM

Bauhaus-Universität Weimar

SNRM: Stand alone Neural Ranking Model

Model	type	embedding	re-rank
GRU	rep	static	yes
DRMM	int	static	yes
KNRM	int	static	yes
CKNRM	int	static	yes
Vanilla BERT	int	contx	yes
DRMM BERT	int	contx	yes
KNRM BERT	int	contx	yes
SNRM	rep	static	no

- All the models up to now require candidate documents to do a re-ranking: Their inference is a 2 step process (candidate selector is BM25 for our case)
- Propagation of the error from the first ranker mode (in our case BM25)
- SNRM as an end-to-end ranking model
  - Hour-glass shape networks for generating representation of the n-grams of the inputs
  - Constructing an inverted index of the documents
  - L1 regularization term in the cost function

Bauhaus-Universität Weimar

SNRM

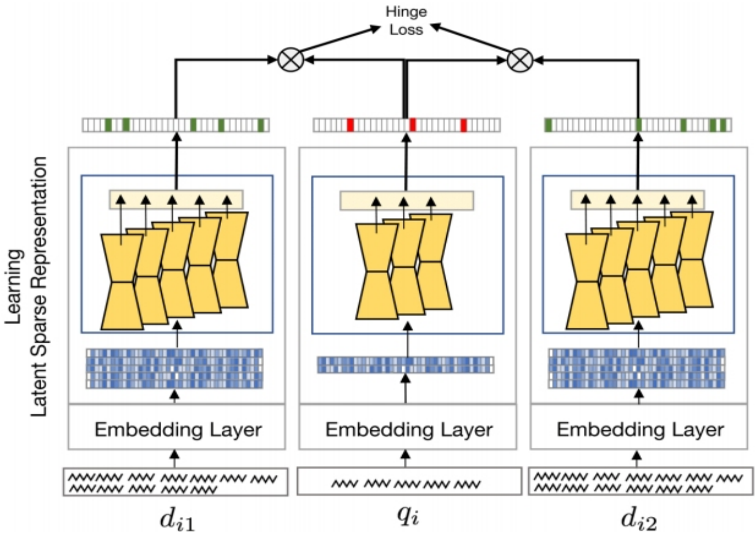


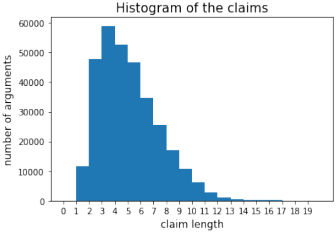
Figure: Training process of SNRM
([Zamani et al.(2018)Zamani, Dehghani, Croft, Learned-Miller, and Kamps])

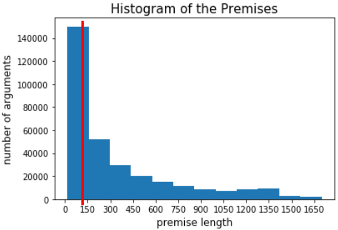
Outline

Bauhaus-Universität Weimar

Train and Validation Phase

- 10000 sample data for hyper-parameter tuning and debug the codes so that the models run correctly
- Query length: 20 and Document length: 100

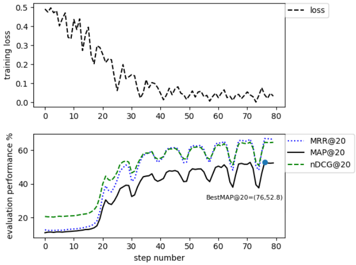




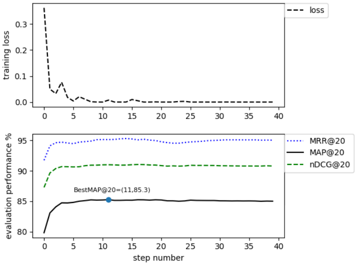
- Each batch: 32 argument
- Train the models
  - static embedding: 10 epochs
  - contextualized embedding: 5 epochs
- Validation run for 8 times within a training epoch
  - Top 20 hits among the 105 validation documents for each query
  - Validation metrics: MRR@20, MAP@20, and nDCG@20
  - For binary qrel: MAP@20 more stable validation scores

Bauhaus-Universität Weimar

Sample Training and Validation Curves



(a) DRMM



(b) Vanilla BERT

Validation Results

- **RQ2.1: Representation-focus vs. interaction-focus**
- **RQ2.2: Contextualized and Static Embedding**
- **RQ2.3:Typical Neural ranking model vs. End-to-End?**

Table: Models

Model	type	embedding	re-rank	MAP@20
GRU	rep	static	yes	0.241
DRMM	int	static	yes	0.528
KNRM	int	static	yes	0.727
CKNRM	int	static	yes	0.733
Vanilla BERT	int	contx	yes	0.88
DRMM BERT	int	contx	yes	0.881
KNRM BERT	int	contx	yes	0.902
SNRM	rep	static	no	0.701

Outline

Bauhaus-Universität Weimar

Re-ranking Candidate Arguments

- 50 test queries provided in the Touché task
- 100 first hits by each model for each test query is saved

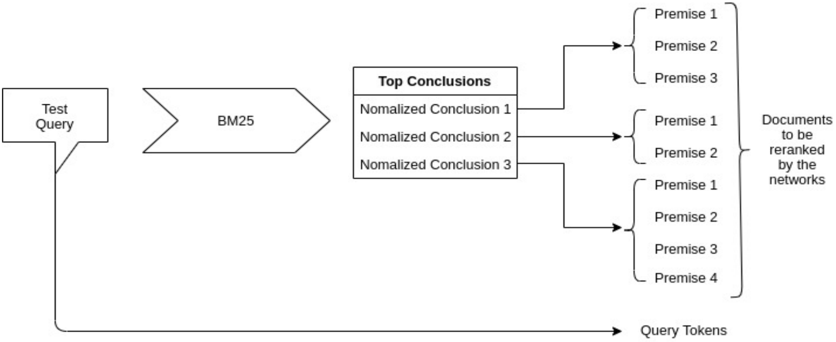


Figure: Candidate documents to be re-ranked in the test phase

Inference in SNRM

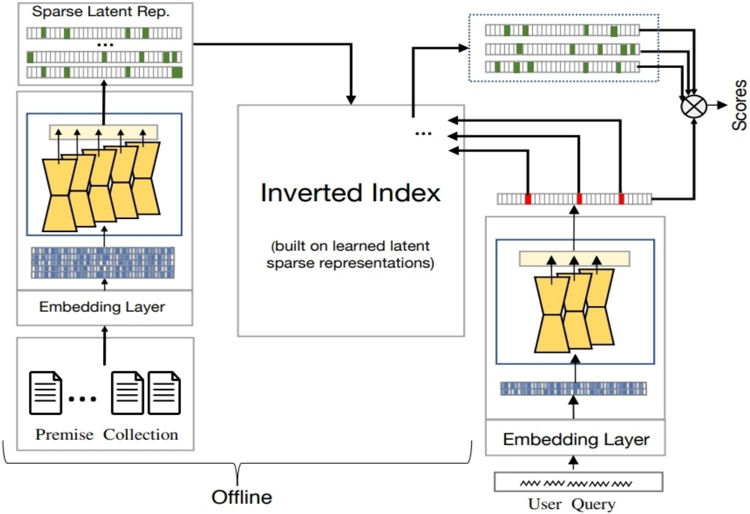


Figure: Document retrieval process ([Zamani et al.(2018)Zamani, Dehghani, Croft, Learned-Miller, and Kamps])

Result Aggregation

RQ3. Aggregation Strategy

- Why to aggregate?
  - Performance improvement
  - Aggregation of the different model principles
- How to aggregate?
  - Using regression between the *normalized* model scores
- What do we need to know before the regression?
  - How diverse the model results are.
  - Models with outlier results. Assumption: Outlier results belong to weak models!

Bauhaus-Universität Weimar

Outline

Bauhaus-Universität Weimar

Model Output Analysis

- The model results are vectors: retrieved documents as dimensions and scores are the values in each dimension retrieved documents are not the same for the models
- Jaccard and Spearman Coefficients for measuring the similarity of the ranking results
  - Jaccard: portion of the documents in common
  - Spearman: correlation of the ranking scores of the common documents
- The average of the coefficients over 50 test queries are calculated

Bauhaus-Universität Weimar

Jaccard Coefficient as Similarity Measure

Jaccard: portion of the documents in common $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$

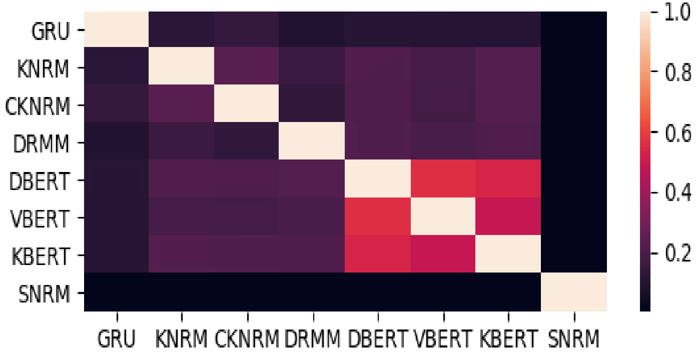


Figure: The heat map of the Jaccard coefficient for the 50 test queries

Outline

Bauhaus-Universität Weimar

Linear Regression as an Aggression Strategy

- We assume SNRM results as outlier data (Based on the similarity results)
- Regression model is trained on validation set (1 related and 1 unrelated document)
  - 2 * 4885 data points for training the regression with the dimension of 7
- union of the retrieved documents by models are scored by the regression model
  - If a model did not retrieve a document, 0 is assigned to the corresponding dimension

Bauhaus-Universität Weimar

Outline

Argument Quality Dimensions

- **Logical**: acceptable and relevant premises to the arguments
- **Rhetorical**: the ability of convince the audiences
- **Dialectical** (utility): the ones by which a stance can be built
- Our concern in this study: Focusing on the Logical aspect

Test Results

- nDCG@5 score is calculated over the retrieved arguments
- Manually annotation is done by human annotators based on the different quality dimensions of the arguments

Model	type	embedding	re-rank	MAP@20	nDCG@5
GRU	rep	static	yes	0.241	x
DRMM	int	static	yes	0.528	x
KNRM	int	static	yes	0.727	0.684
CKNRM	int	static	yes	0.733	x
Vanilla BERT	int	contx	yes	0.88	0.404
DRMM BERT	int	contx	yes	0.881	0.371
KNRM BERT	int	contx	yes	0.902	0.319
SNRM	rep	static	no	0.701	x
Aggregation	x	x	x	x	0.372

Bauhaus-Universität Weimar

Test Results

- KNRM (our best performing model) ranked 4th in the competition
- Most of the competitors got less score than the baseline (Dirichlet LM)
  - Argument retrieval meeting the quality dimensions is not an easy task
- Validation results and test results were not correlated
  - related arguments $\neq$ good arguments (meeting the argument quality dimensions)
  - Relevance is a required but not enough condition for a good argument
- Interaction-focused network outperformed representation-focused networks
  - Representation focused networks' results are not shown in the table
- Aggregation model has been trained on the validation set and its MAP@20 score on the validation set is useless.

Bauhaus-Universität Weimar

Outline

Bauhaus-Universität Weimar

Summary

- **RQ1.** How to shape useful training and validation set fit for the task of ad-hoc retrieval from the collection?
  - ✓ Using distant super vision and assigning unrelated documents with Fuzzy similiarty
  - ✓ Creat validation set with higher number of unrelated documents
- Using neural ranking models that have shown good performance in ad-hoc retrieval tasks in the argument retrieval
  - **RQ2.1.** Interaction-focused vs representation-focused
    - ✓ Representation-focused
  - **RQ2.2.** Static embedding vs. contextualized embedding?
    - ✓ Contextualized embedding
  - **RQ2.3.** Typical Neural ranking model vs. End-to-End?
    - ✓ Improvement needed for end-to-end approach
- **RQ3.** How to aggregate model results? Which strategy to use and what we require for doing so?
  - ✓ Linear regression as an aggregation strategy
  - ✓ Analysis of result similarity is required

Bauhaus-Universität Weimar

Outline

Bauhaus-Universität Weimar

What's next...

- Providing a concrete mathematical definition of the argument quality dimensions to be included in the cost function of the networks
- Working on strategies to map the interaction of the input pairs
- Devising more intuitive structures to create sparse representation for end-to-end models

Bauhaus-Universität Weimar

Thanks!

Evaluation Metrics: Mean Reciprocal Rank (MRR)

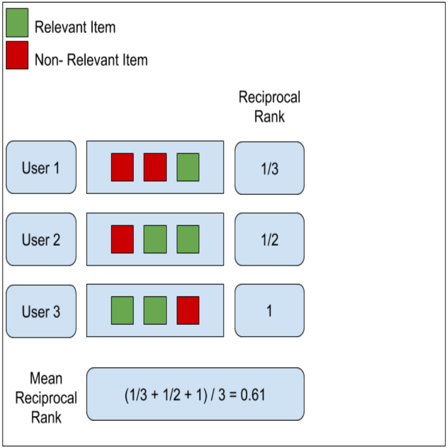


Figure: An example of MRR calculation

Bauhaus-Universität Weimar

Evaluation Metrics: Mean Average Precision (MAP)

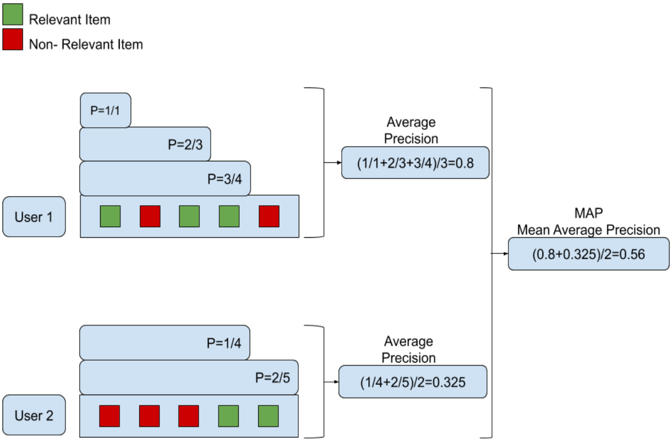


Figure: An example of MAP calculation

Bauhaus-Universität Weimar

Evaluation Metrics: Normalized Discounted Cumulative Gain (nDCG)

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{log_2(i+1)} \tag{2}$$

$$nDCG_p = \frac{DCG_p}{IDCG_p}. \tag{3}$$

Bauhaus-Universität Weimar

Lorik Dumani.
Good premises retrieval via a two-stage argument retrieval model.
In *Grundlagen von Datenbanken*, pages 3–8, 2019.

Richard D Rieke, Malcolm Osgood Sillars, and Tarla Rai Peterson.
Argumentation and critical decision making.
Longman New York, 1997.

Hamed Zamani, Mostafa Dehghani, W Bruce Croft, Erik Learned-Miller, and Jaap Kamps.
From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing.
In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 497–506, 2018.