Automatic Methods for Low-Cost Evaluation and Position-Aware Neural IR Models

-Ph.D. Dissertation Defense-

Kai Hui

Faculty of Mathematics and Computer Sciences
Saarland University

December 4, 2017

Background

QUERY

computer science course Germany

Search Results

1. Institutes in **Germany** provide graduate-level **courses** in **computer science.**

2. MacTrade is an online portal for purchasing personal **computers** in **Germany.**

Information Retrieval

- **Information need** is expressed as a keyword query from a user
- **Search results**. A ranked list of documents from a retrieval system
- **Relevance.** The ranking should satisfy the information need of the user

CONTENT PART I PART II BEGIN END

Motivation

- **Evaluation of the retrieval systems** requires expensive manual labor to provide a ground-truth ranking of a query

CONTENT PART I PART II BEGIN END

Motivation

- **Evaluation of the retrieval systems** requires expensive manual labor to provide a ground-truth ranking of a query

  Automatic methods allow to reduce the number of manual judgments required

Motivation

- **Evaluation of the retrieval systems** requires expensive manual labors to provide a ground-truth ranking relative to a query

  Automatic methods facilitate to reduce the required number of manual judgments

- **Retrieval models** are desired to capture the complicated interactions between a query and a document

Motivation

- **Evaluation of the retrieval systems** requires expensive manual labors to provide a ground-truth ranking relative to a query

  Automatic methods facilitate to reduce the required number of manual judgments

- **Retrieval models** are desired to capture the complicated interactions between a query and a document

  Deep learning models provide instruments to better encode the query-document interactions

Contributions

❑Low-cost evaluation for graded judgments

- Compare different document embedding in terms of their agreement with the cluster hypothesis (WWW16 poster)
- **Max-Rep for low-cost ad-hoc evaluation (SPIRE15 full paper)**
- Lmd-Cascade for low-cost novelty and diversity evaluation (ICTIR17 full paper)

Contributions

❑Low-cost evaluation for preference judgments

- Investigation of the preference judgments with / without ties collected via crowdsourcing (ECIR17 full paper)
- Usage of the ties for low-cost preference judgments (ECIR17 short paper, ICTIR17 short paper)

Contributions

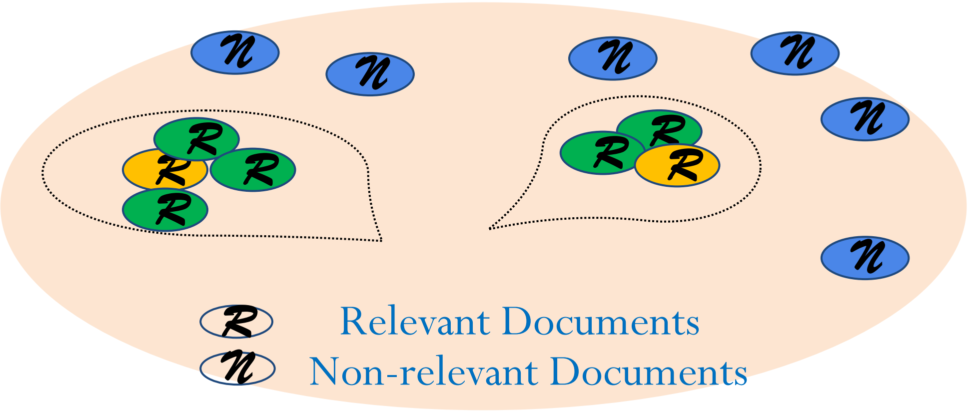
❑Deep retrieval models

- A position-aware representation for ad-hoc retrieval (WWW17 poster)
- **PACRR: a position-aware neural IR model (EMNLP17 full paper)**
- Co-PACRR: encode domain insights from IR into a neural IR model (WSDM18 full paper)

Outline

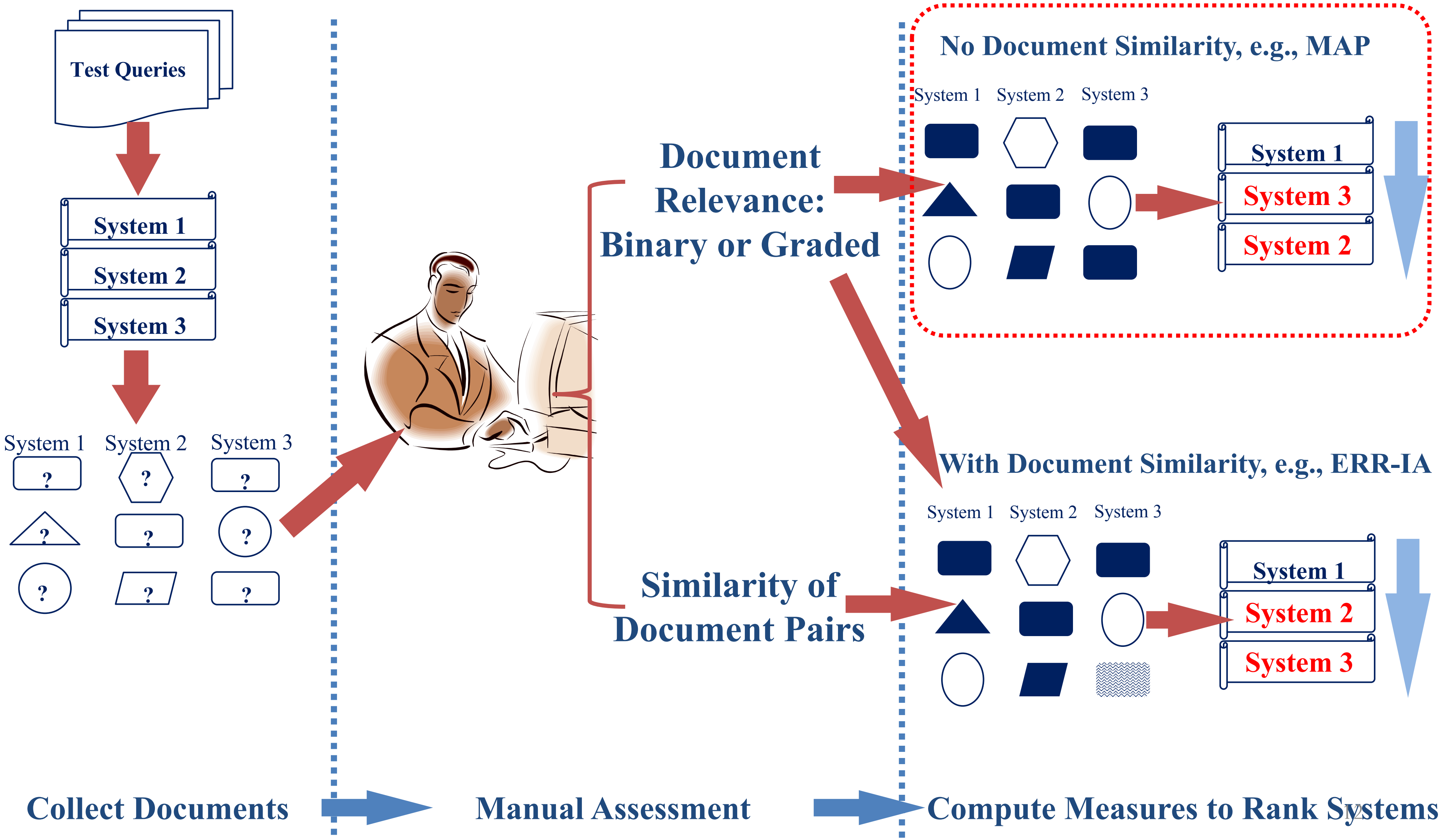
- ❑MaxRep: lost-cost evaluation for binary judgments
- ❑PACRR: a position-aware neural IR model
- ❑Conclusion

Max-Rep: Lost-Cost Evaluation for Binary Judgments

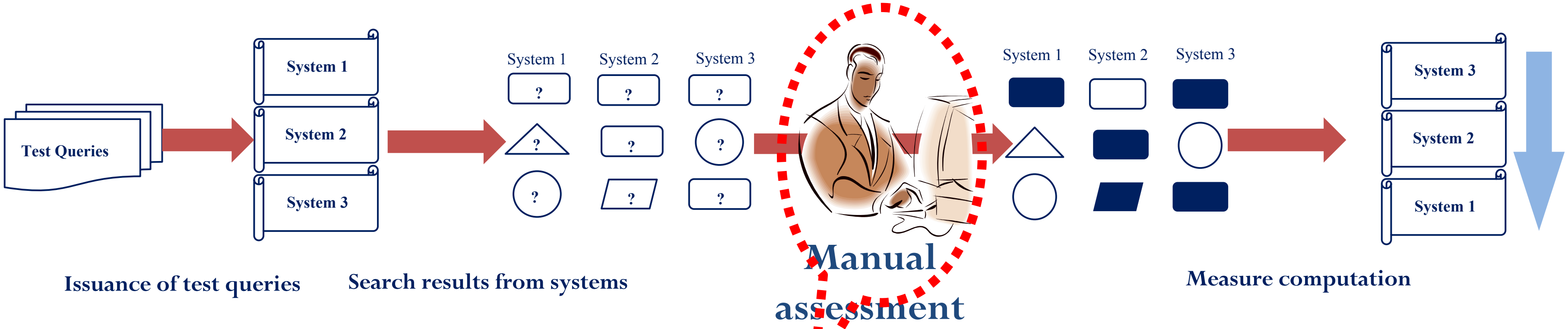


CONTENT PART I PART II BEGIN END

Revisited IR Evaluation Pipeline

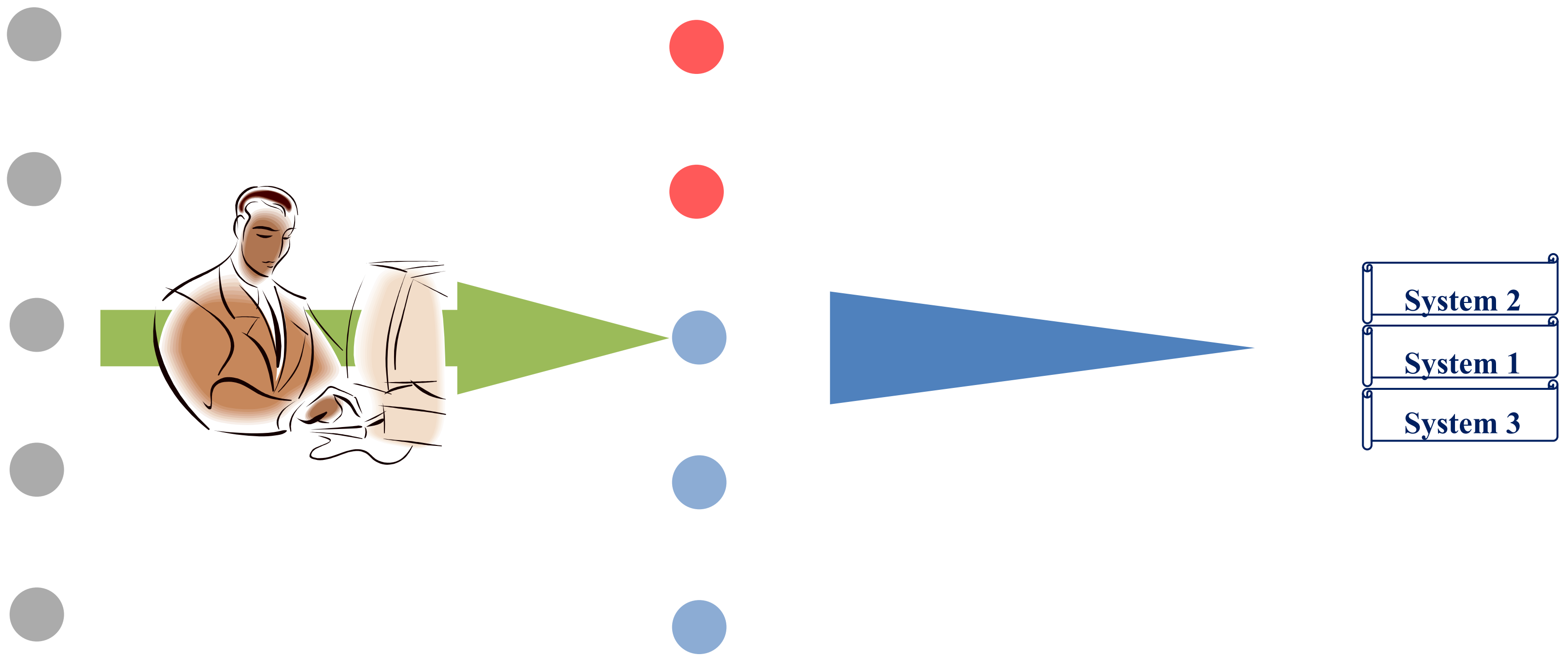


Manual Judgments are Expensive



Statistics of Labels from TREC Web Track ad-hoc Task			
Year	**#Systems**	**Pooling depth**	**#Total labeled doc**
2009	71	20	23,601
2010	55	20	25,330
2011	62	25	19,381
2012	48	20/30	16,055
2013	50	10/20	14,474
2014	27	25	14,432

Low-cost Evaluation

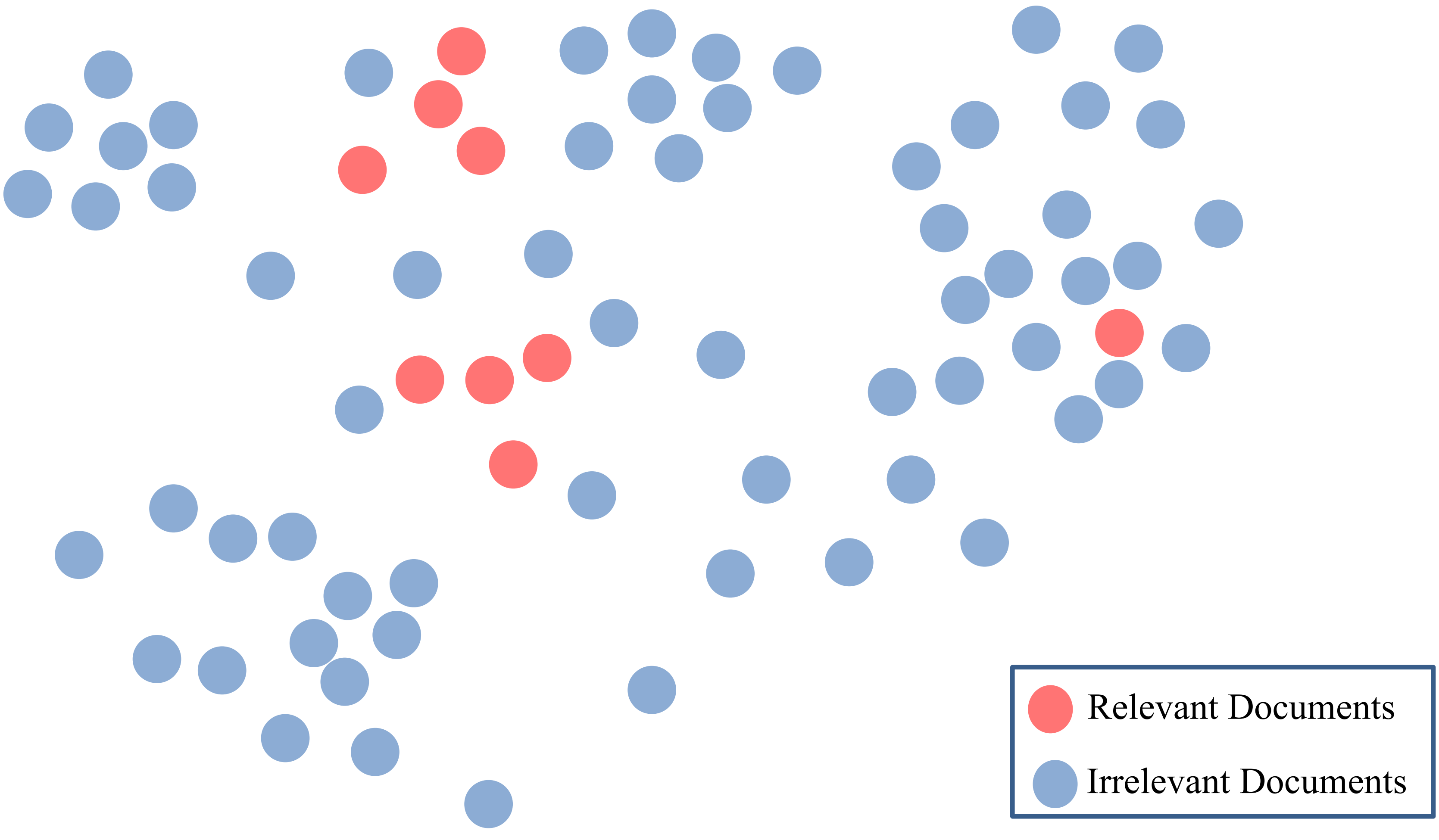


Manual judgments **Ranking of systems**

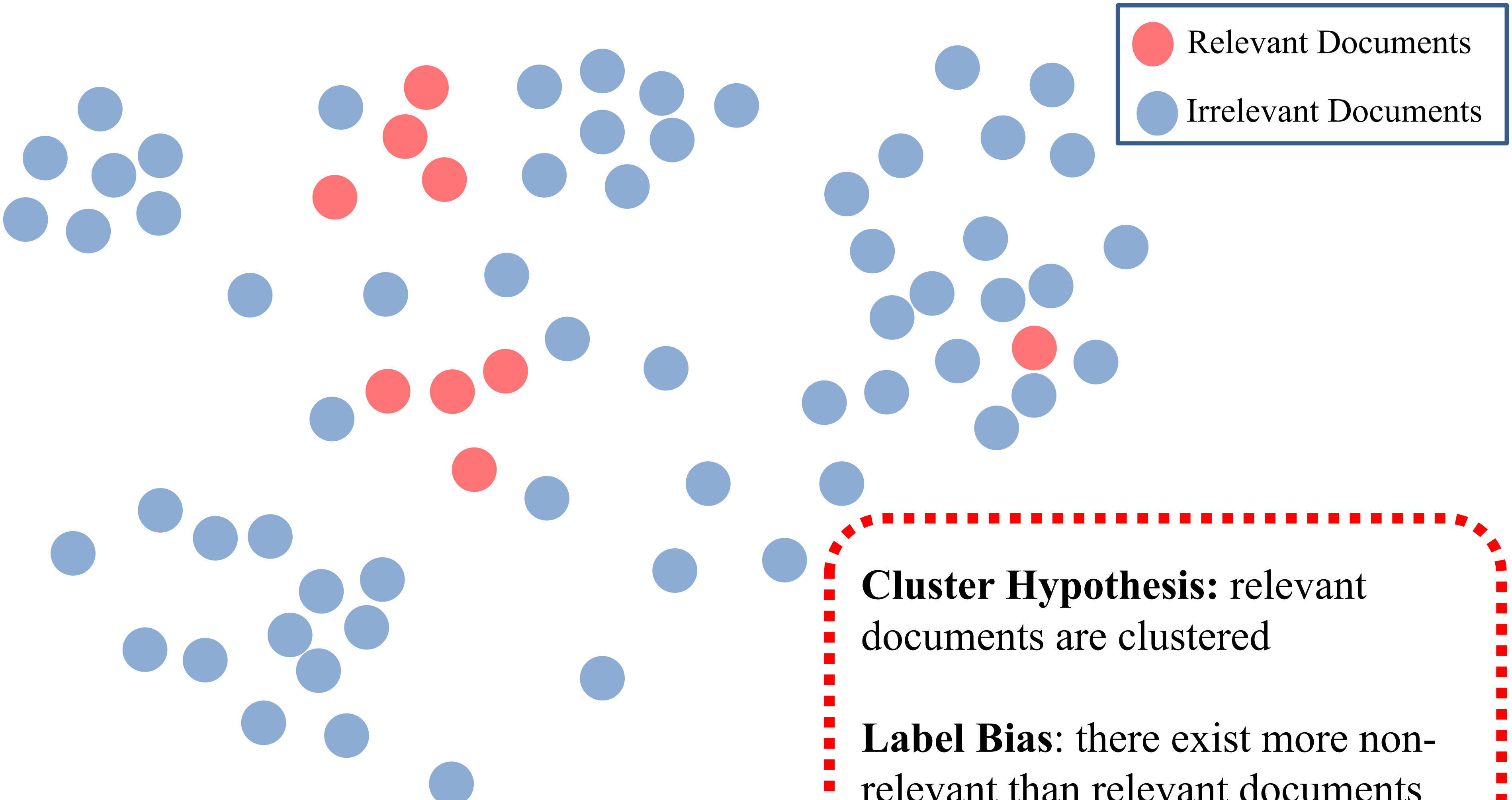
Low-cost Evaluation

Manual judgments + Automatic inference **Ranking of systems**

Document Vector Space in A Search Result



Document Vector Space in A Search Result



Cluster Hypothesis: relevant documents are clustered

Label Bias: there exist more non-relevant than relevant documents

Framework:
Selective Labeling + Label Prediction

Evaluation based on complete judgment

Low-cost evaluation with fewer labels

Collect Documents

Select Subset of Documents

Select representative documents for judgment.

Manual Assessment

Mitigate Missing Labels

Text classification using SVM with linear kernel, trained with the labeled documents.

Measure Computation

MaxRep: Representativeness of Documents

- Document subset L with k documents from document collection D_q

- Representativeness of L is the aggregated maximum coverage of the remaining documents D_q

$$f(L) = \sum_{d_i \in D_q} \max_{d_j \in L} \mathbf{w_i} sim(d_i, d_j)$$

cosine similarity

Prioritize documents that are more likely to be relevant

MaxRep: Select Representative Documents

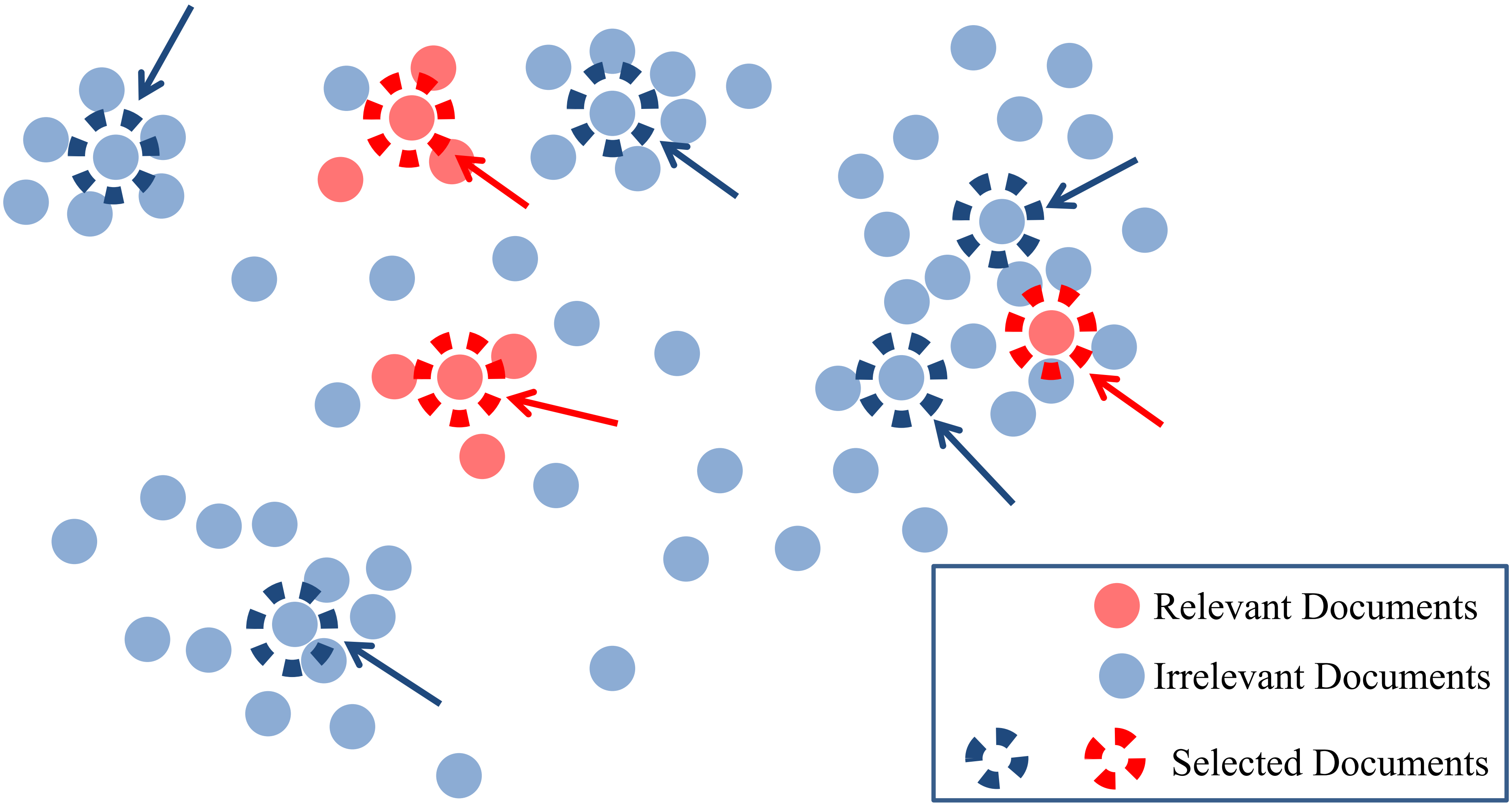
Optimization Target

$$L_k^* = \underset{L_k}{\operatorname{argmax}}\ f(L) \quad \text{s.t.} \quad |L| = k$$

Greedy Algorithm

- Start with L_0 with no document
- In i- th iteration, select a document from $D \backslash L_{i-1}$ to maximize $f(L_i)$
- Stop when k documents are selected and get L_k

Only Label Representative Documents



Experimental Setting

- **Dataset**

TREC Web Track 2011–2014 on ClueWeb 09 & 12, leading to 64 k labeled documents, 200 queries

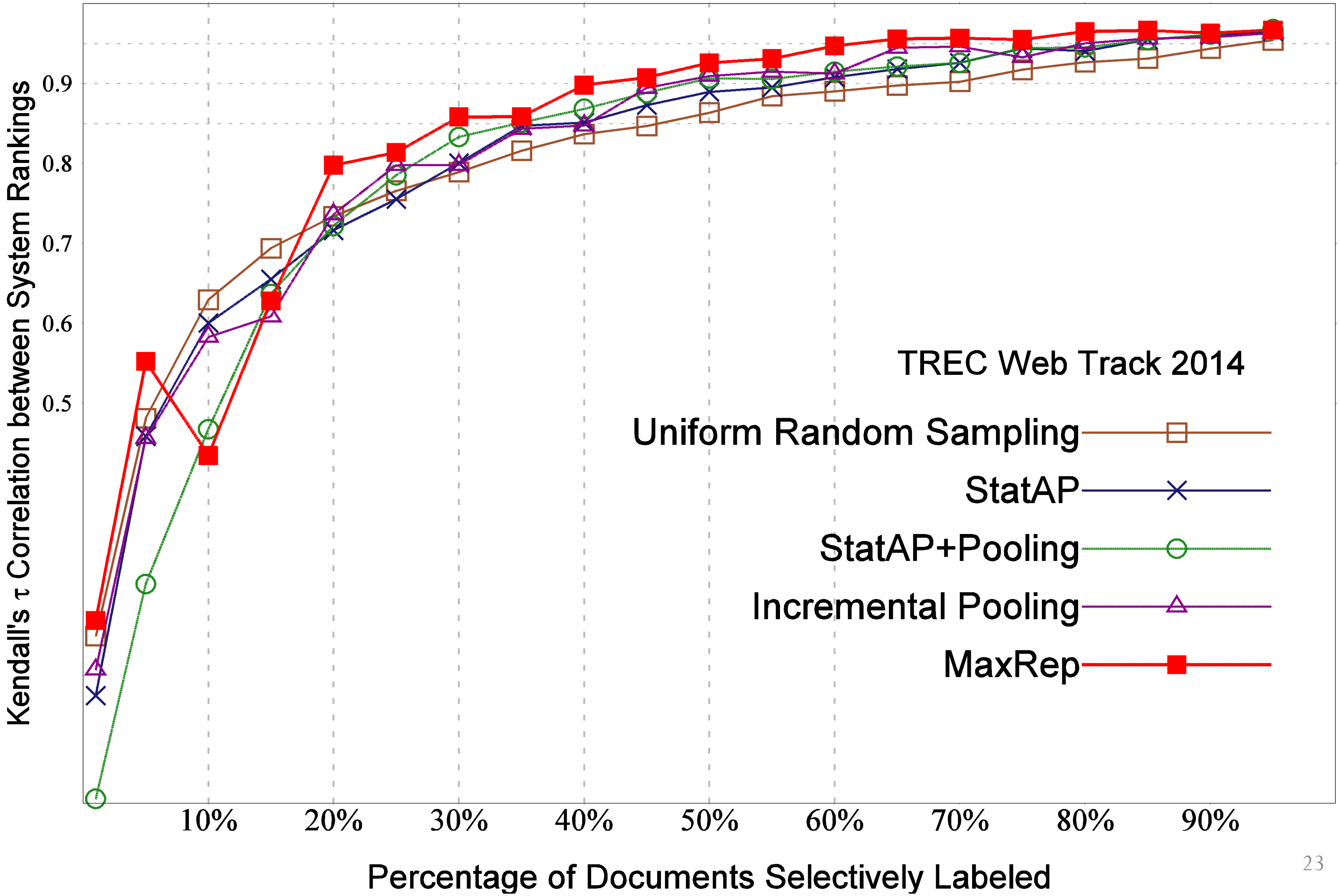
- **Ground-truth measure**

Mean Average Precision (MAP)

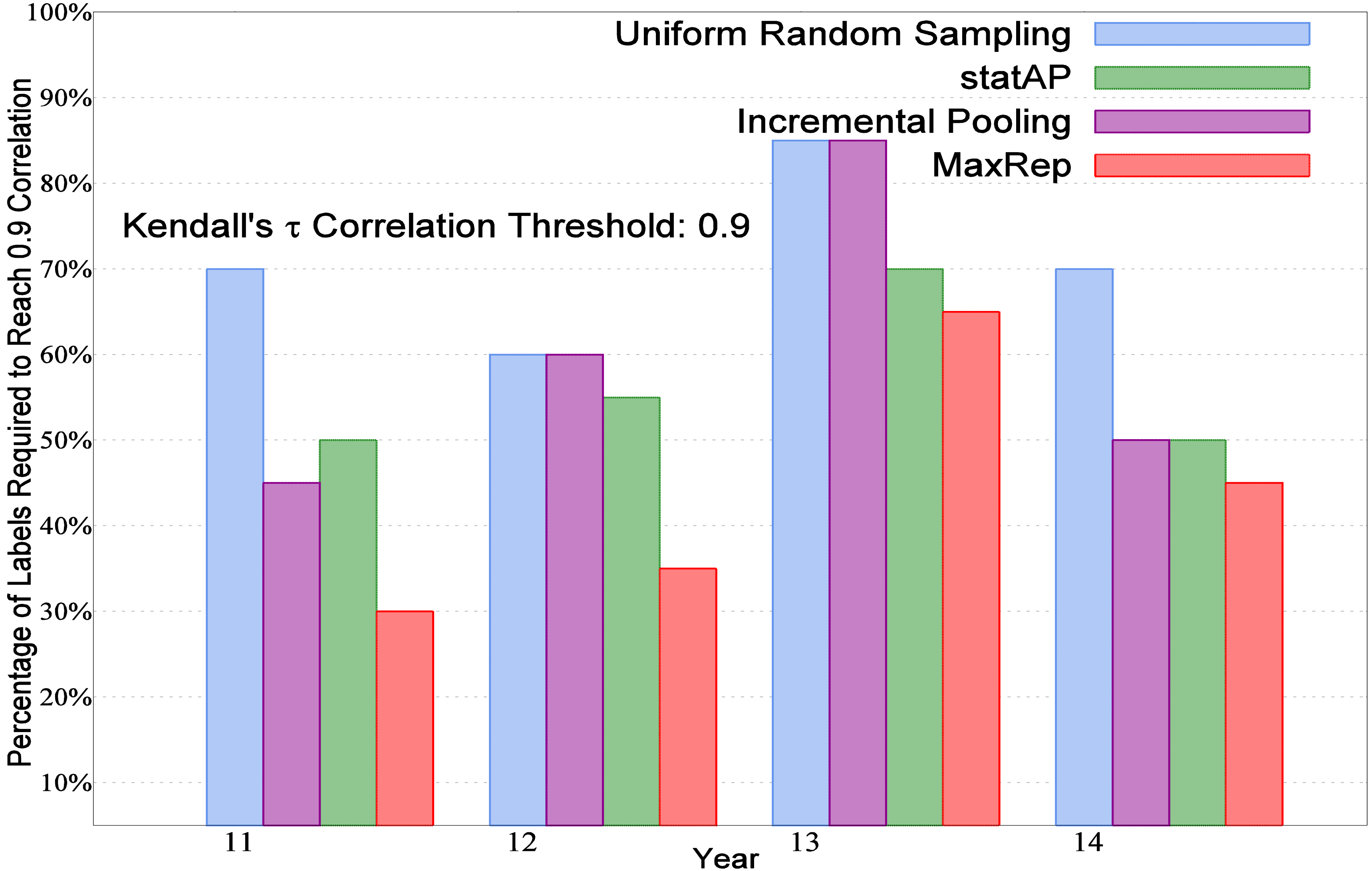
- **Benchmark**

Kendall's τ correlation: Approximation of the system ranking

Approximate System Ranking: Kendall's τ



Summary of Kendall's τ



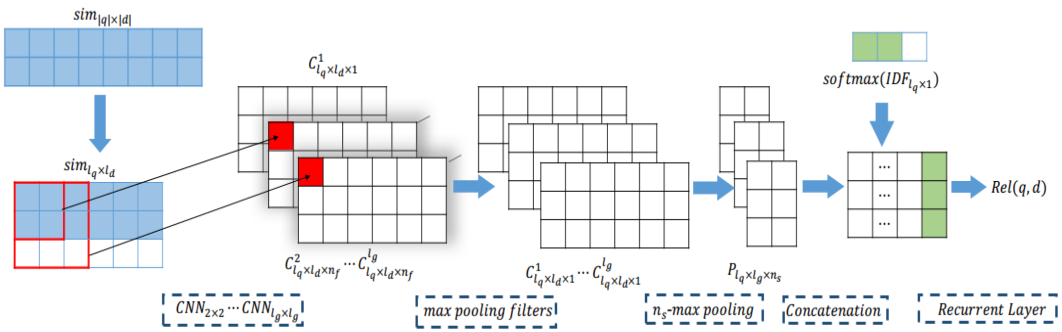
Wrap-up

- **A novel strategy MaxRep is proposed**, considering both ranking information and document contents, selecting a representative subset of documents to label

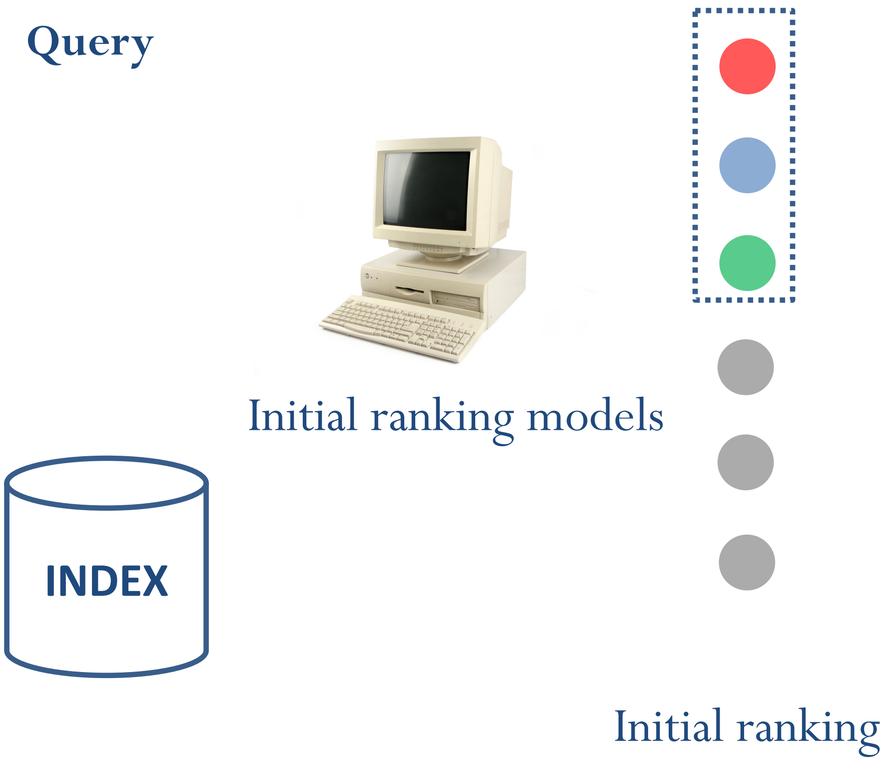
- **Label prediction + MaxRep** can save up to as much as 70% of manual judgments

- **Comparison on TREC Web Track** data confirmed that MapRep outperforms other strategies

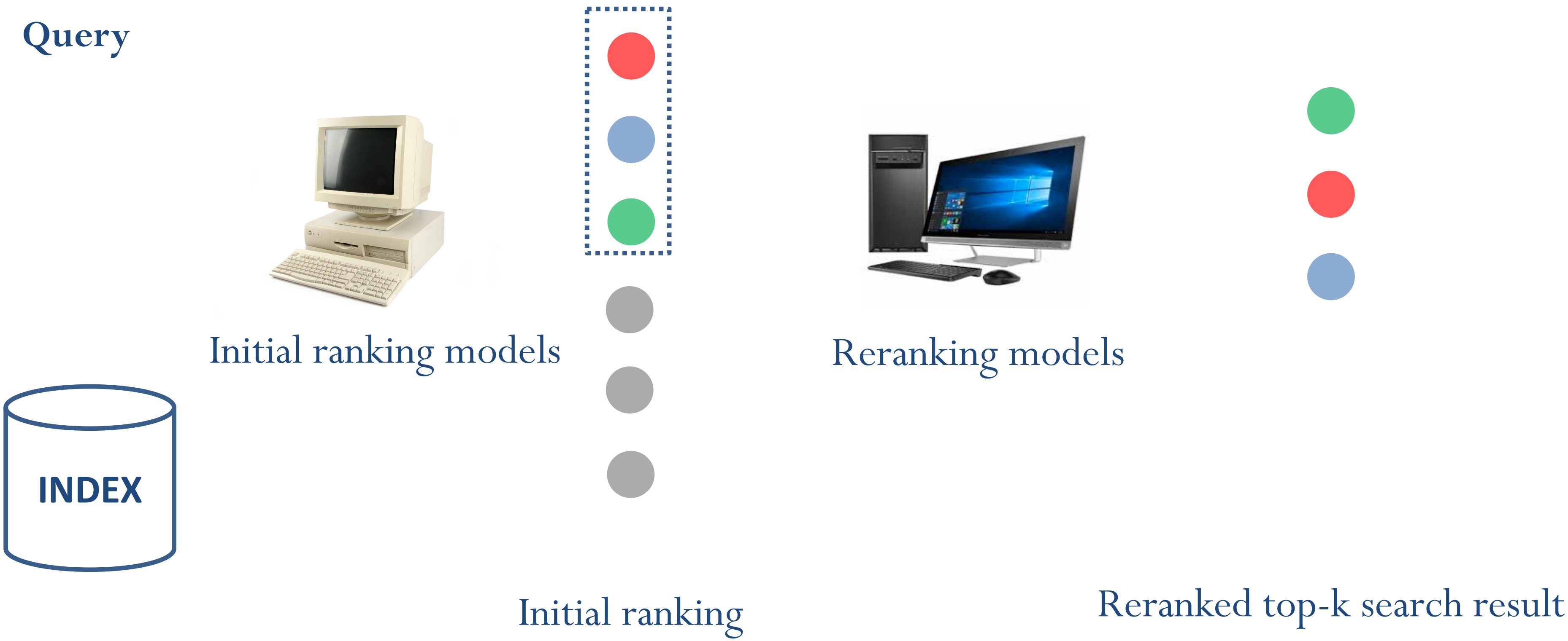
PACRR: A Position-Aware Neural IR model



Reranking Models



Reranking Models



Matching Information to Incorporate

QUERY

computer science course Germany

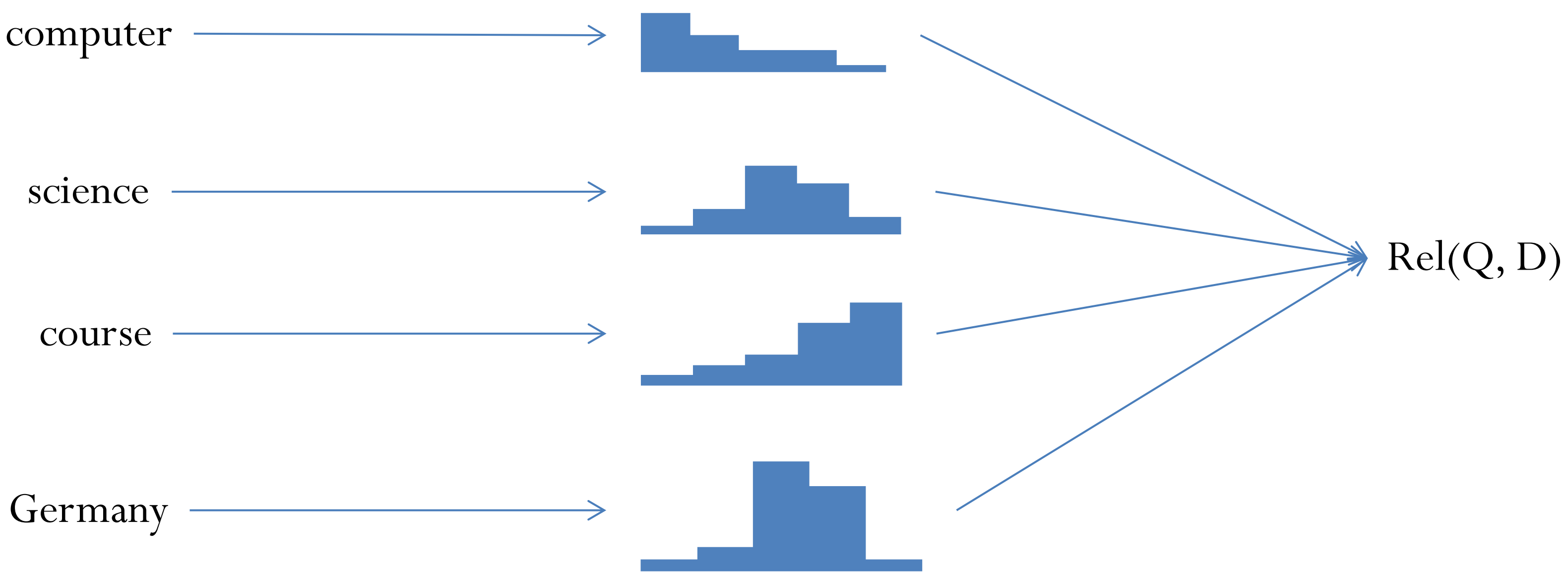
DOCUMENT

1. Institutes in **Germany** provide graduate-level **courses** in **computer science.**
2. MacTrade is an online portal for purchasing personal **computers** in **Germany.**

- Unigram matching: matching individual terms independently
- Term dependency: computer science
- Query proximity: the proximity between different matches

Model Unigram Matching by Counting

- Given a query Q and a document D
- Compute the semantic similarity between each term pair, where one term is from Q and another is from D (via word2vec)
- Group such similarity into bins and model the relevance between Q and D with a histogram



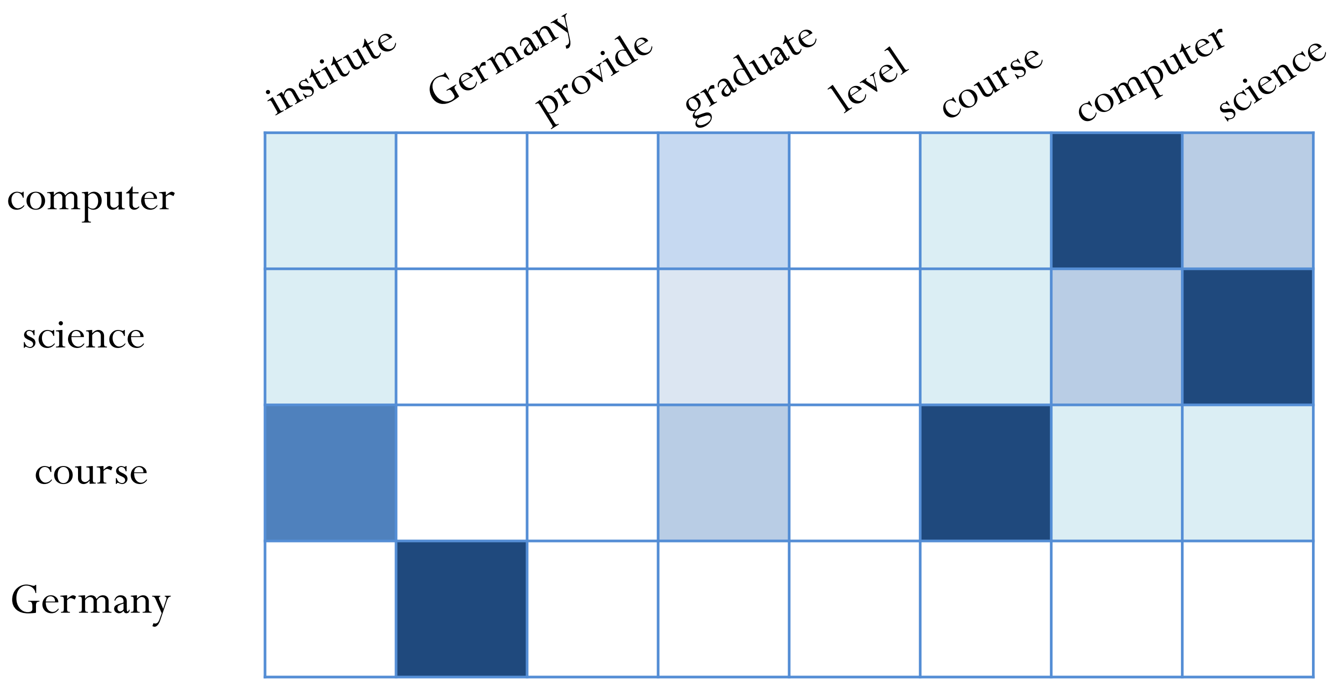
bag-of-word assumption
(independence among terms)

Motivation

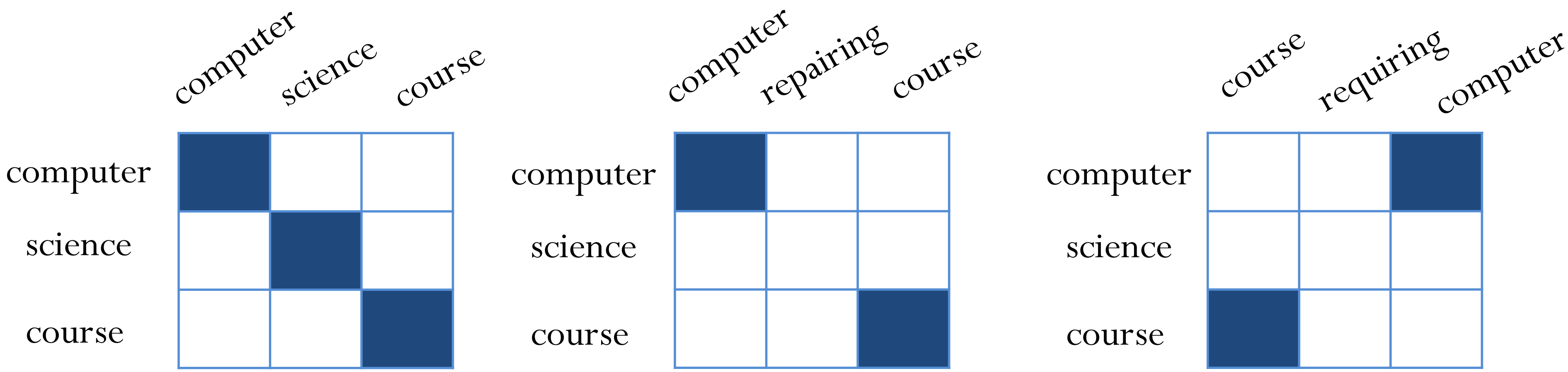
- ❑ Unigram matching signals have been successfully incorporated into neural IR models
- ❑ How to incorporate positional matching information remains unclear

Beyond Unigram Matching: Model Positional Information

1) Retain the positional information by considering a similarity matrix, keeping both similarity and their relative positions

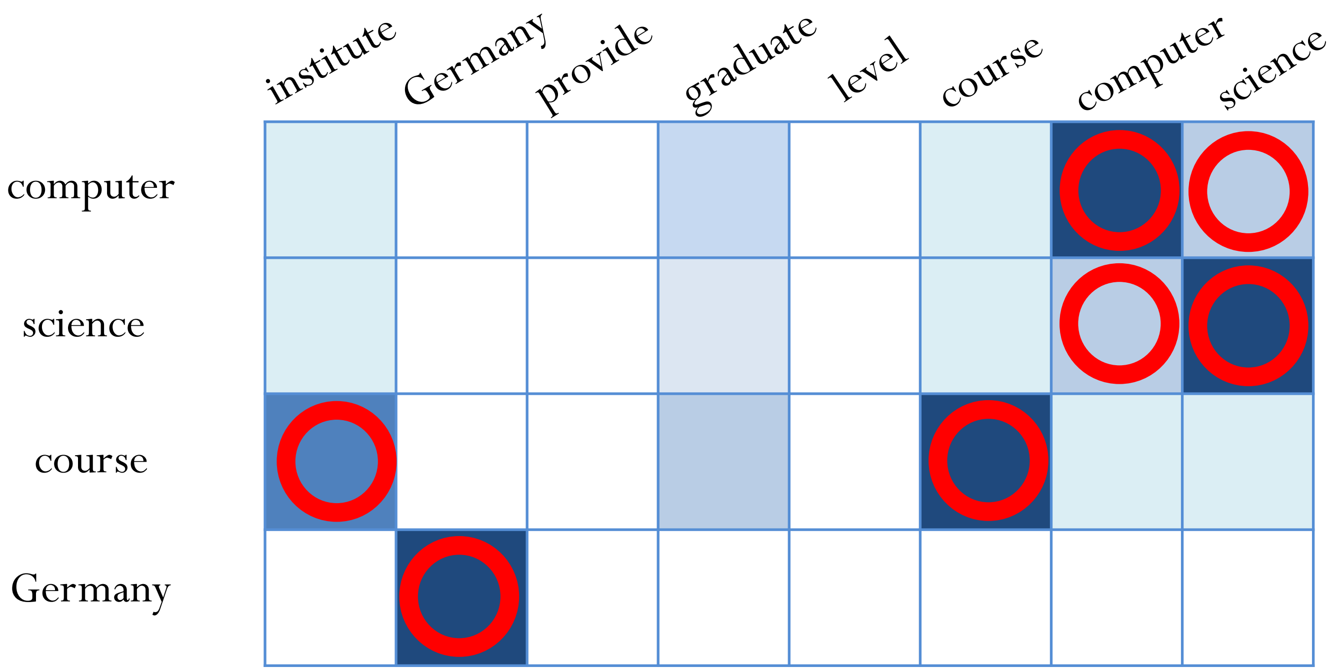


Beyond Unigram Matching: Model Positional Information



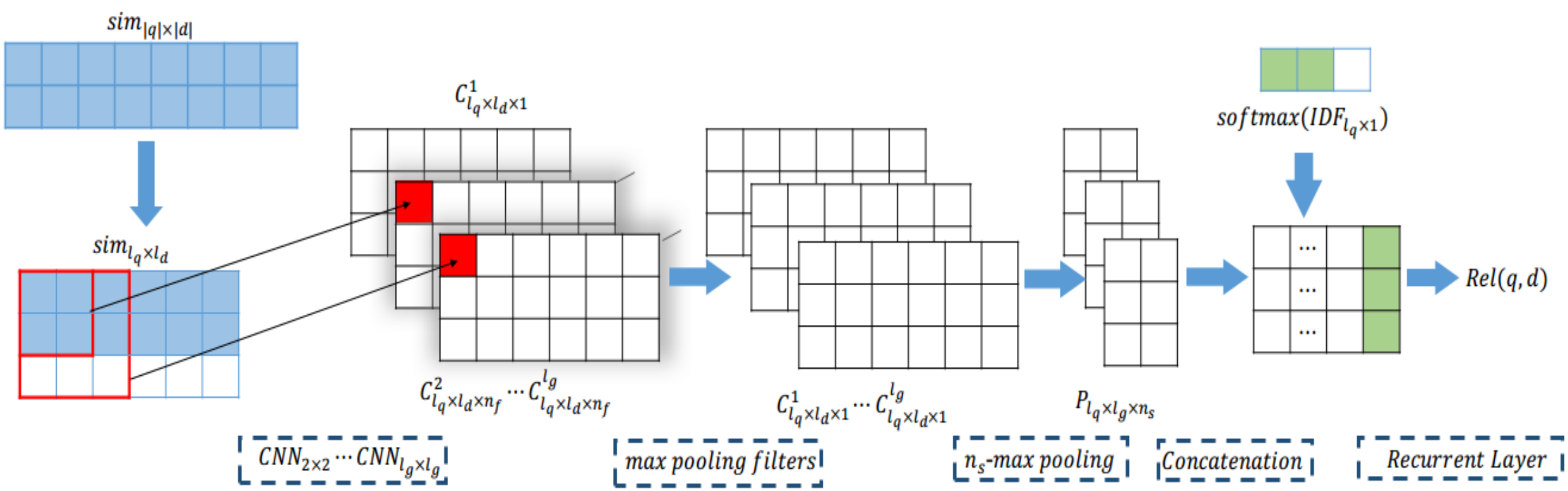
2) Matching could be modeled based on different local patterns in the similarity matrix
3) Individual text windows only include one salient matching pattern

Beyond Unigram Matching: Model Positional Information



4) <u>Only retain the salient matching signals for individual query terms</u>

PACRR: Position-Aware Convolutional Recurrent Relevance Matching



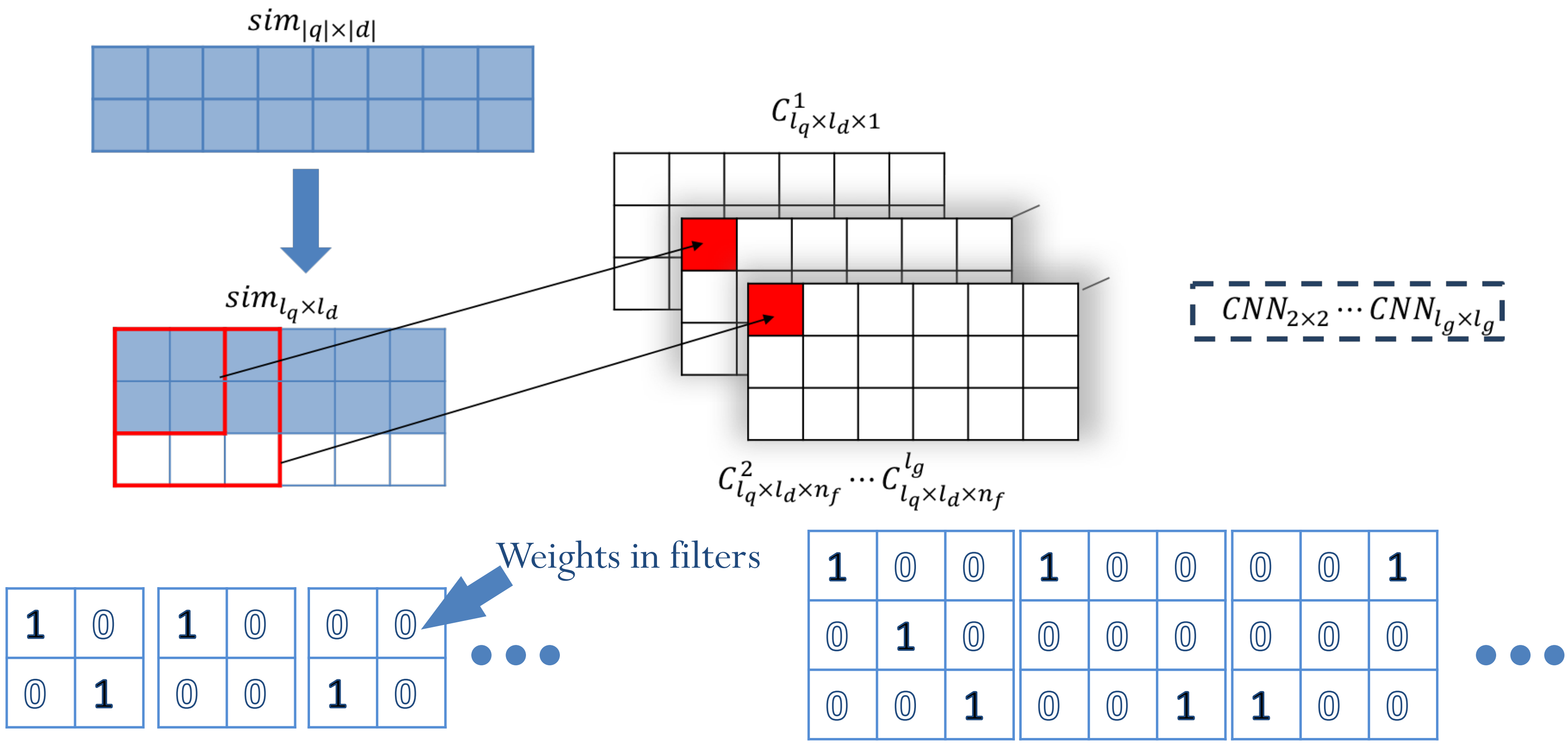
(1) CNN layers with different sizes: 2X2, 3X3, 4X4, etc..

(2) Max-pooling among filters

(3) K-max pooling: retain the k most salient signals for each query term

(4) LSTM layer for combination

PACRR: Parallel Convolutional Layers

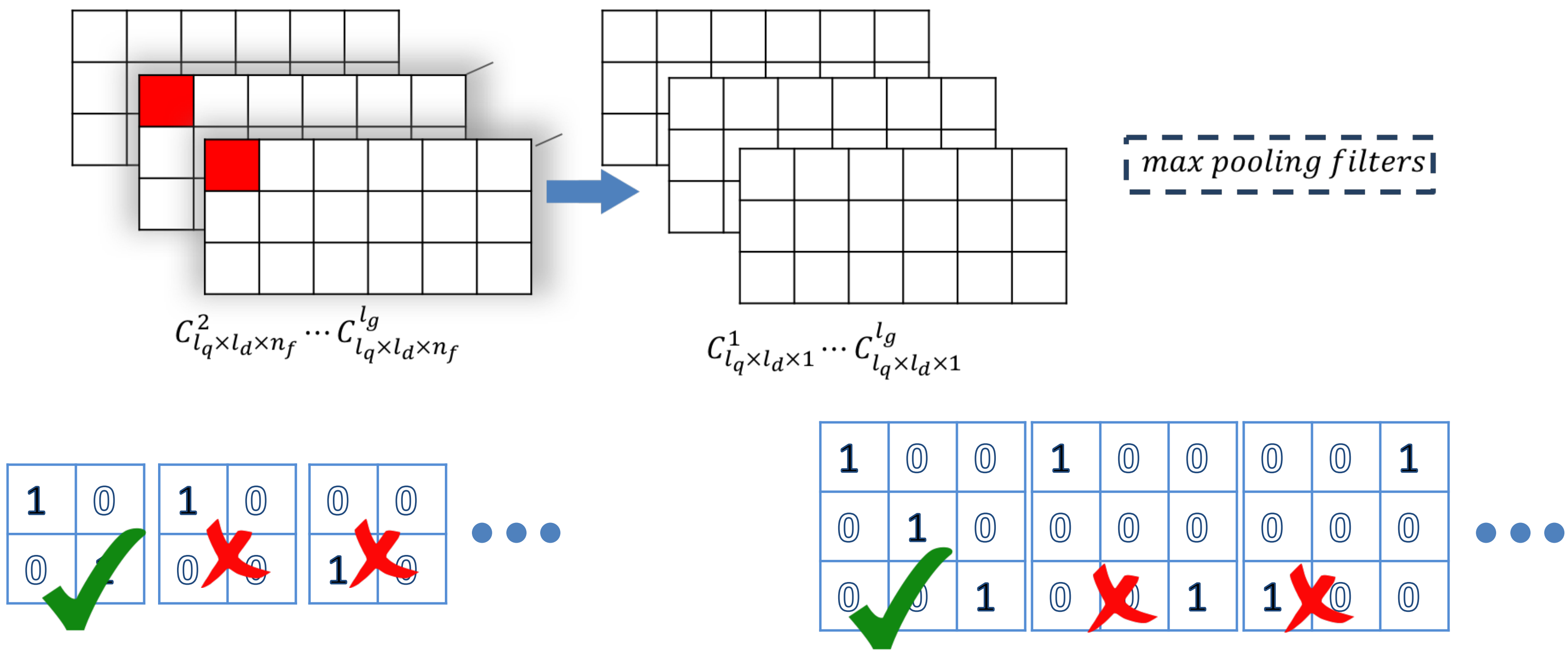


computer science, science course, etc..

computer science course, science course Germany, etc..

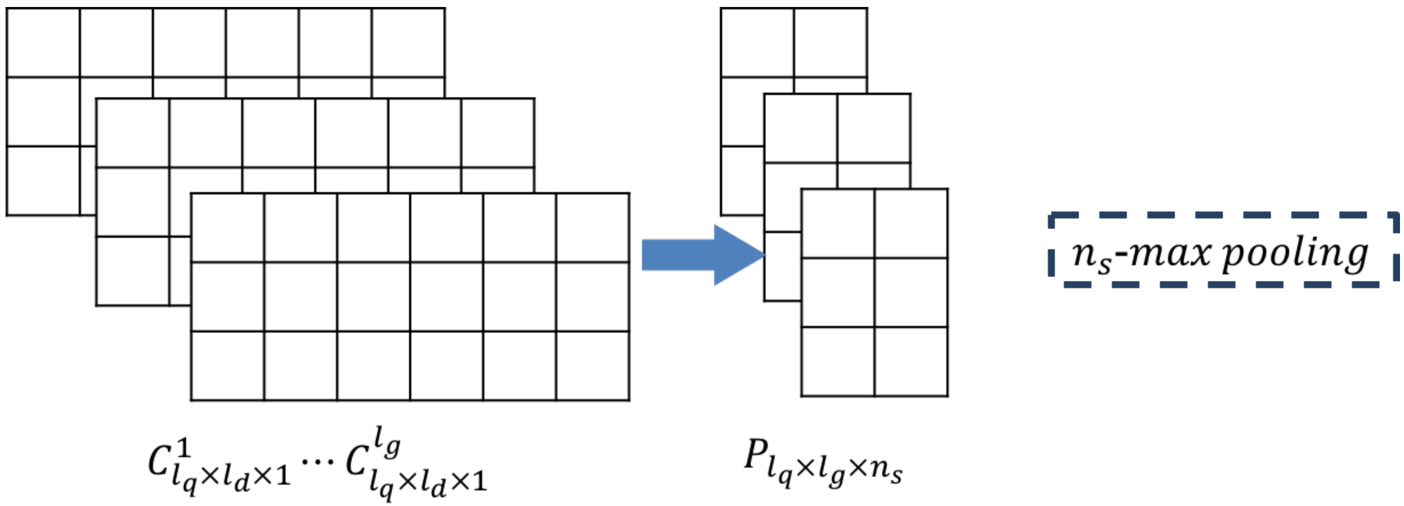
- CNN kernels (dozens of filters) in different sizes, corresponding to text windows with different length

PACRR: Max-Pooling over Filters

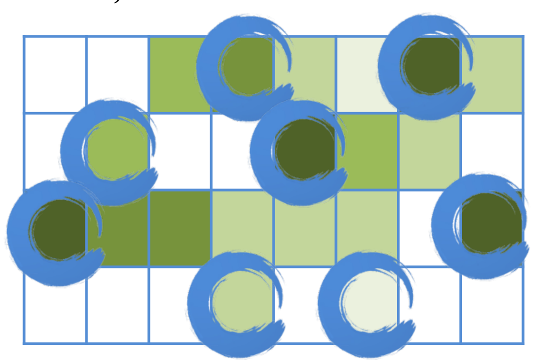


- Max pooling different filters for individual kernels (individual text windows at most include one matching pattern)

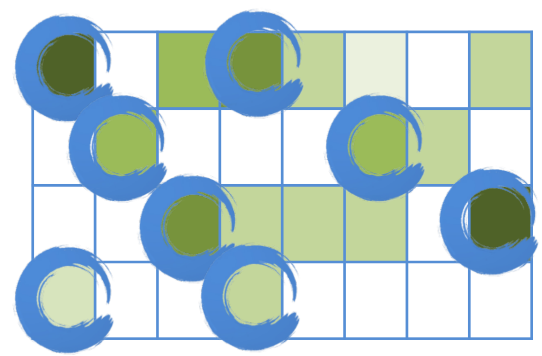
PACRR: K Max-Pooling along Query Terms



K=2, 2X2 kernel

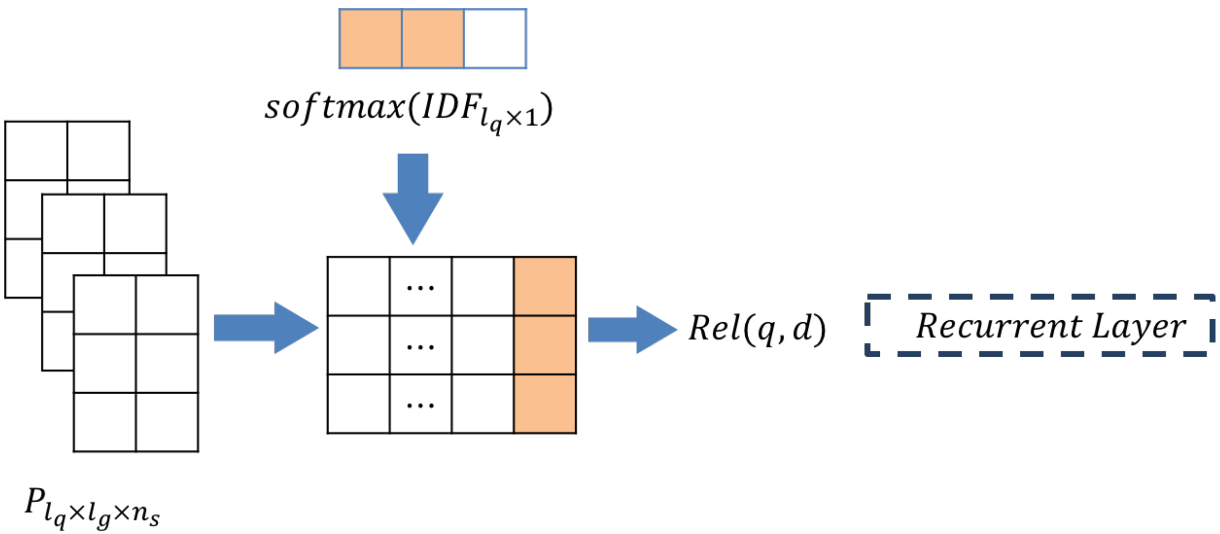


K=2, 3X3 kernel



- K-max pooling for individual query terms, retaining the k most salient signals for individual query terms

PACRR: RNN Layer Along Query Terms



- A LSTM layer combines signals on different query terms

Evaluation

- Based on TREC Web Track ad-hoc task 2009-2014, including 300 queries, 100k judgments and about 50 runs in each year

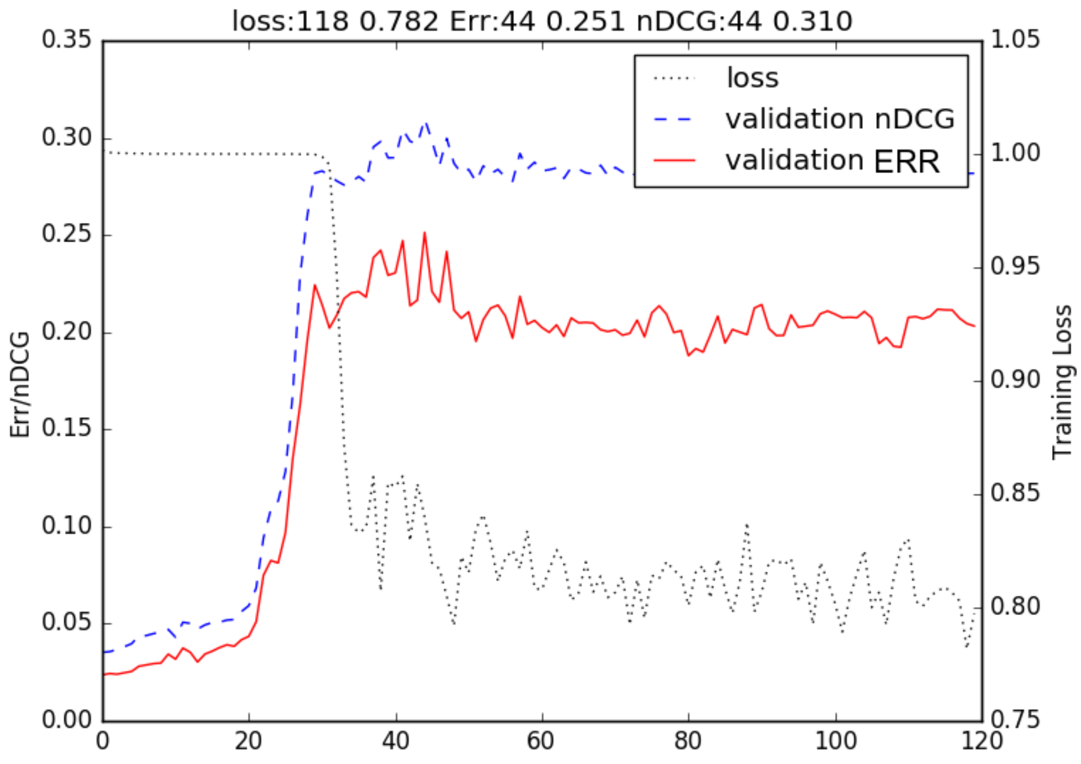
- Measure: ERR@20
  - A real value summarizes the quality of a ranking
  - Lager values are better

- Baseline models: MatchPyramid, DRMM, local model in DUET, and K-NRM

Training and Validation

- Employ five years (250 queries) for training and validation
- Randomly reserve 50 queries from the 250 queries for validation to select models based on ERR@20
- Test on the remaining year (50 queries)

Training and Validation



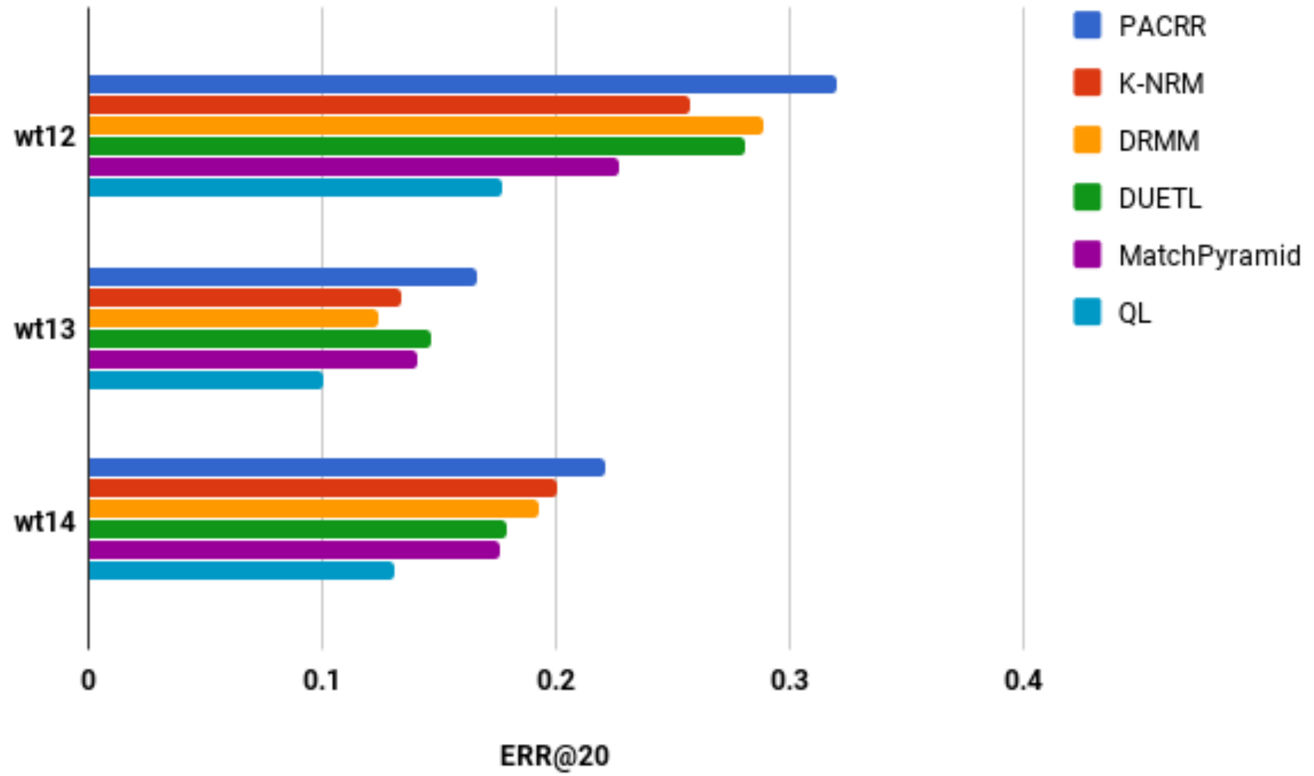
The training loss, ERR@20 and nDCG@20 per iteration on validation data. The x-axis denotes the iterations. The y-axis indicates the ERR@20/nDCG@20 (left) and the loss (right)

Result: RerankSimple

- The Neural IR model is employed as a re-ranker, making improvements by re-ranking top-k (e.g., top-30) search results from initial ranker
- Initial ranker can access the whole collection of documents
- Re-rank search results from a simple ranker, namely, query-likelihood model (QL)

Result: RerankSimple

How good a neural IR model can achieve by reranking QL baseline?



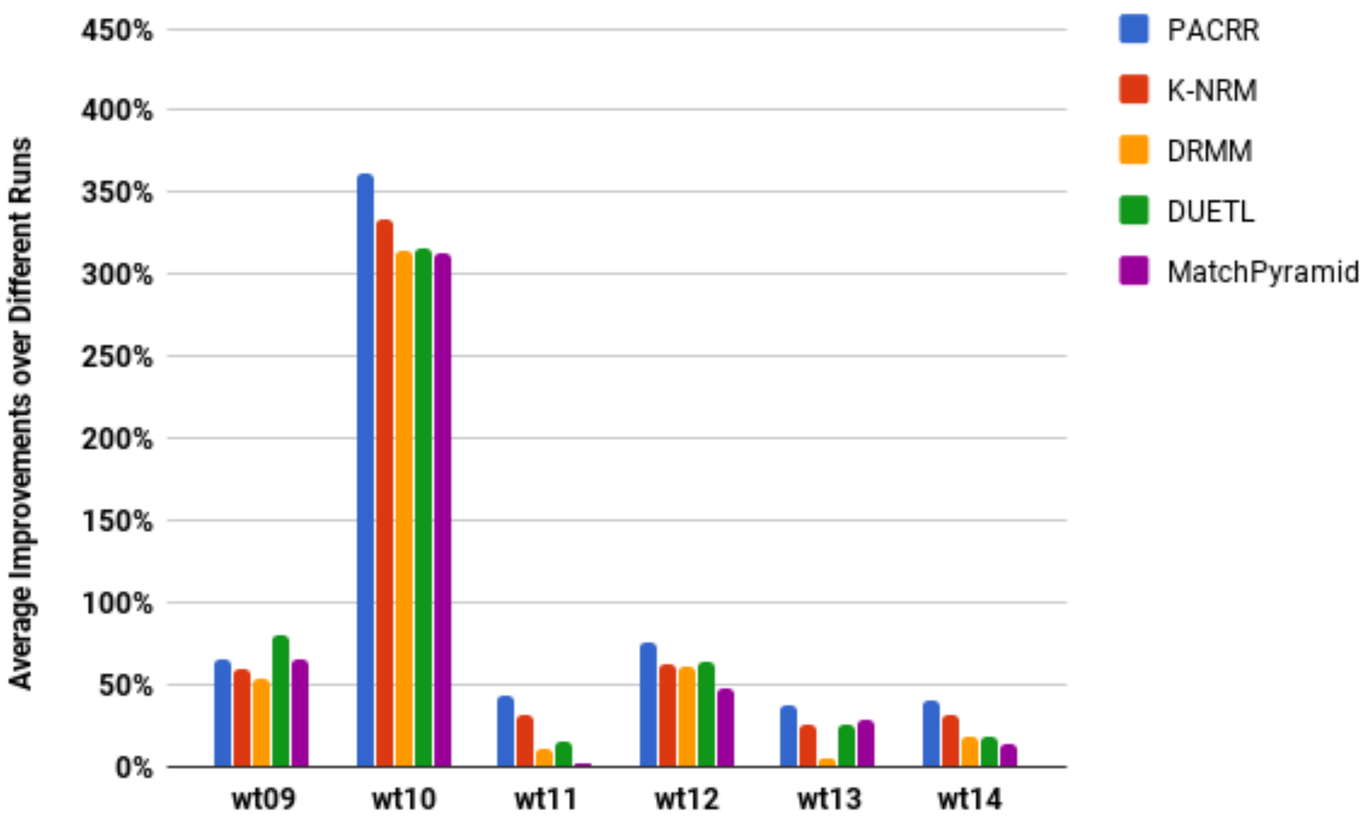
- All neural IR models can improve based on QL search results
- **PACRR can achieve top-3 by solely re-ranking the search results from query-likelihood model**

Result: RerankALL

- Re-rank search results from all runs which participated in TREC
- A neural IR model should work together with diversified initial runs
- Average improvements among all runs in each year
- Percentage of runs that can be improved by a neural IR model

Result: RerankALL

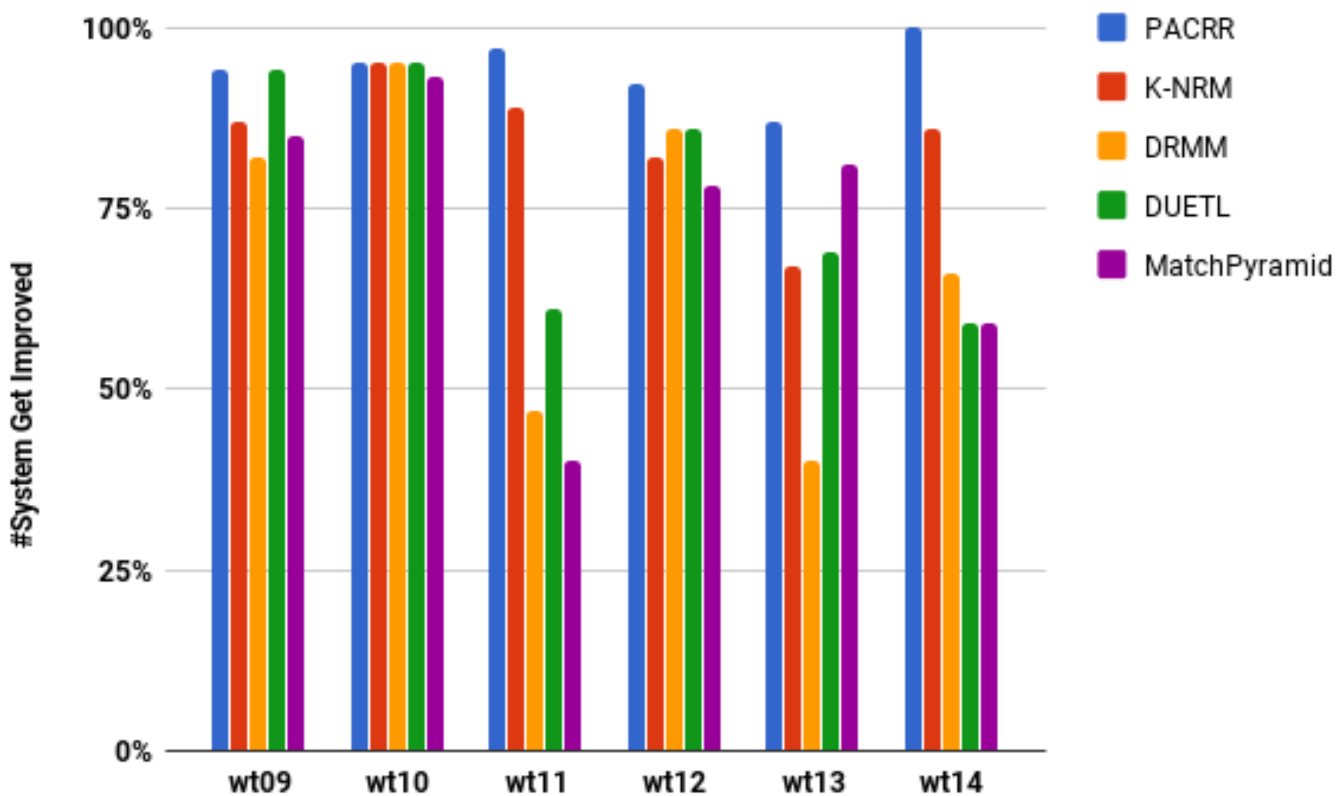
How much a neural IR model can improve on average?



- All neural IR models can improve on average among all years
- **PACRR can at least improve by 37% on average among all different years**

Result: RerankALL

How many runs a neural IR model can improve?



- All neural IR models can improve more than half of the runs
- **PACRR can improve 94% runs on average over six years**

Result: PairAccuracy

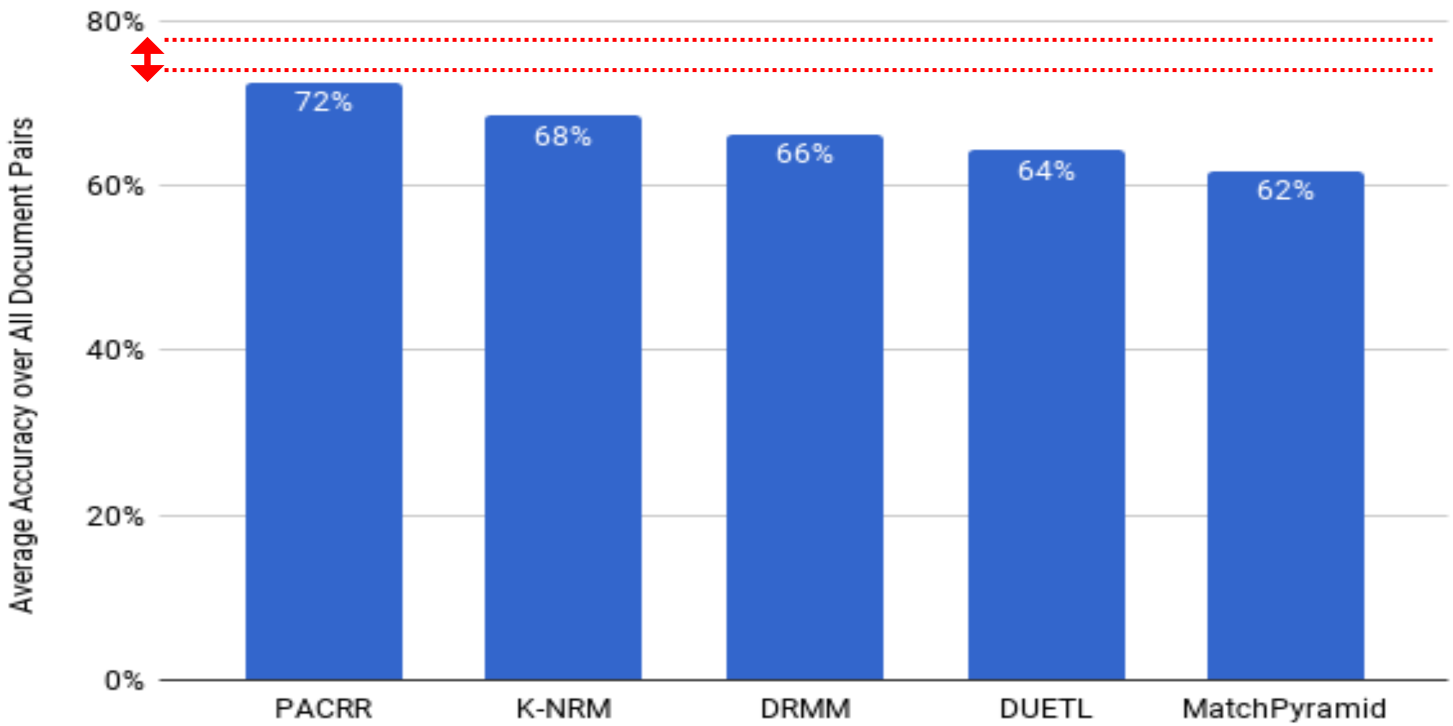
How many doc pairs a neural IR model can rank correctly?

- Evaluate on pairwise ranking benchmark. Given (q, d_1, d_2),
  Is d_1 more relevant or d_2 is more relevant?

- Cover all document pairs that are being predicted
- Calculate the accuracy: the ratio of the concordant pairs

Result: PairAccuracy

How many doc pairs a neural IR model can rank correctly?



- The average accuracy for PACRR among different label pairs is 72%
- As reference, human accessors agree with each other by **74-77%** according to the literature

Wrap-up

- **A novel neural IR model PACRR is proposed,** whose variant (Co-PACRR) performs the best by the time of writing

- **The code/data is published for future comparison:** https://github.com/khui/repacrr

CONTENT PART I PART II BEGIN END

Conclusion

- **MaxRep** selects a representative subset of documents to label. Combining MaxRep with label prediction can save up to 70% label efforts

- **PACRR** encodes positional signals with CNN/max-pooling structures, outperforms all baseline models

Future Work

- ❑ Proper document embedding is desired to better cater for cluster hypothesis

- ❑ Weak supervision of neural IR model is of interest to replace the manual judgments with cheaper label data

Publications

Full papers

[1] **K. Hui**, A. Yates, K. Berberich, G. de Melo:
PACRR: A Position-Aware Neural IR Model for Relevance Matching. EMNLP 2017
[2] **K. Hui**, A. Yates, K. Berberich, G. de Melo:
Co-PACRR: A Context-Aware Neural IR model for Ad-hoc Retrieval. WSDM 2018
[3] **K. Hui**, K. Berberich:
Transitivity, Time Consumption, and Quality of Preference Judgments in Crowdsourcing. ECIR 2017
[4] **K. Hui**, K. Berberich:
Selective Labeling and Incomplete Label Mitigation for Low-Cost Evaluation. SPIRE 2015
[5] **K. Hui**, K. Berberich, I. Mele:
Dealing with Incomplete Judgments in Cascade Measures. ICTIR 2017
[6] Y. Ran, B. He, **K. Hui**, J. Xu, L. Sun:
A Document-Based Neural Relevance Model for Effective Clinical Decision Support. BIBM 2017

Short papers

[1] **K. Hui**, A. Yates, K. Berberich, G. de Melo:
Position-Aware Representations for Relevance Matching in Neural Information Retrieval. WWW 2017
[2] **K. Hui**, K. Berberich:
Cluster Hypothesis in Low-Cost IR Evaluation with Different Document Representations. WWW 2016
[3] **K. Hui**, K. Berberich:
Low-Cost Preference Judgment via Ties. ECIR 2017
[4] **K. Hui**, K. Berberich : Merge-Tie-Judge: Low-Cost Preference Judgments with Ties. ICTIR 2017

Workshop papers

[1] **K. Hui**, A. Yates, K. Berberich, G. de Melo:
RE-PACRR: A Context and Density-Aware Neural Information Retrieval Model. Neu-IR workshop 2017@SIGIR17
[2] S. MacAvaney, **K. Hui**, A. Yates:
An Approach for Weakly-Supervised Deep Information Retrieval. Neu-IR workshop 2017@SIGIR17
[3] A. Yates, **K. Hui**: DE-PACRR:
Exploring Layers Inside the PACRR Model. Neu-IR workshop 2017@SIGIR17
[4] **K. Hui**: Towards Robust & Reusable Evaluation for Novelty & Diversity. PIKM2014@CIKM2014

Thank You!

Email: khui@mpi-inf.mpg.de

CONTENT PART I PART II BEGIN END