max planck institut informatik

RE-PACRR: A Context and Density-Aware Neural Information Retrieval Model

Kai Hui[1], Andrew Yates[1], Klaus Berberich[1], Gerard de Melo[2]
[1]Max Planck Institute for Informatics
{khui, kberberi, ayates}@mpi-inf.mpg.de
[2] Rutgers University, New Brunswick
gdm@demelo.org

max planck institut informatik

Motivation

- Decades of research in ad-hoc retrieval provides insights about the effective measures to boost the performance.
- Implementation of such insights into neural IR models is under-explored.
- More importantly, building blocks to encode different insights should work together.

max planck institut informatik

Insights to Incorporate

Query: **Jaguar SUV price**

- Unigram matching.

  All occurrences of "jaguar", "suv" or "price" are regarded as relevance signals.

- Vocabulary mismatch and sense mismatch (e.g., ambiguity).

  Occurrences of "F-face", "sport cars" or "discount" could also lead to relevance signals; "jaguar" referring to one kind of big cat should not be considered as relevant.

- Positional information, e.g., term dependency and query proximity.

  Co-occurrences of "jaguar price" or "jaguar suv price" indicate stronger signals.

- Query coverage.

  "jaguar", "suv" and "price" should all be covered by a relevant document.

- Cascade reading model.

  Earlier occurrences of relevant information are preferred, given that users are inpatient, resulting in information in the end being neglected due to an early stop.

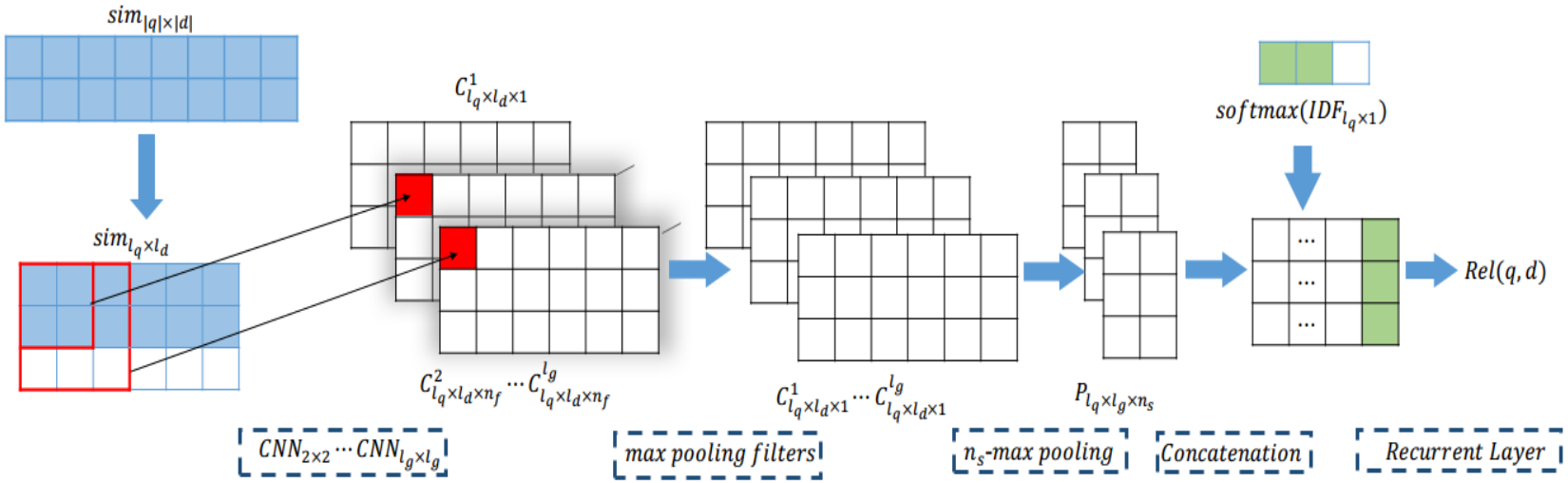
max planck institut informatik

Insights to Incorporate

- Unigram matching.
  - → Counting, as in DRMM and K-NRM.
- Vocabulary mismatch and sense mismatch (e.g., ambiguity).
  - → Similarity in place of exact match, as in DUET distributed model etc..
- Positional information, e.g., term dependency and query proximity.
  - → CNN filters as in DUET, MatchPyramid and PACRR.
- Query coverage.
  - → Combination of relevance signals from different query terms, as in DRMM etc..
- Cascade reading model.
  - → ?

Recap PACRR Model

- Four building blocks are proposed and plugged into an established neural IR model: PACRR (Hui et al., 2017).



Kai Hui, Andrew Yates, Klaus Berberich, Gerard de Melo: PACRR: A Position-Aware Deep Model for Relevance Matching. EMNLP 2017

max planck institut informatik

Design of Modular

- Sense mismatch (e.g., ambiguity).

For individual relevance signals, examine whether their contexts are also relevant, e.g., if context of "jaguar" is distant with a car but close to an animal, …

- Query proximity.

Consider co-occurrences of multiple query terms in a large text window.

- Query coverage.

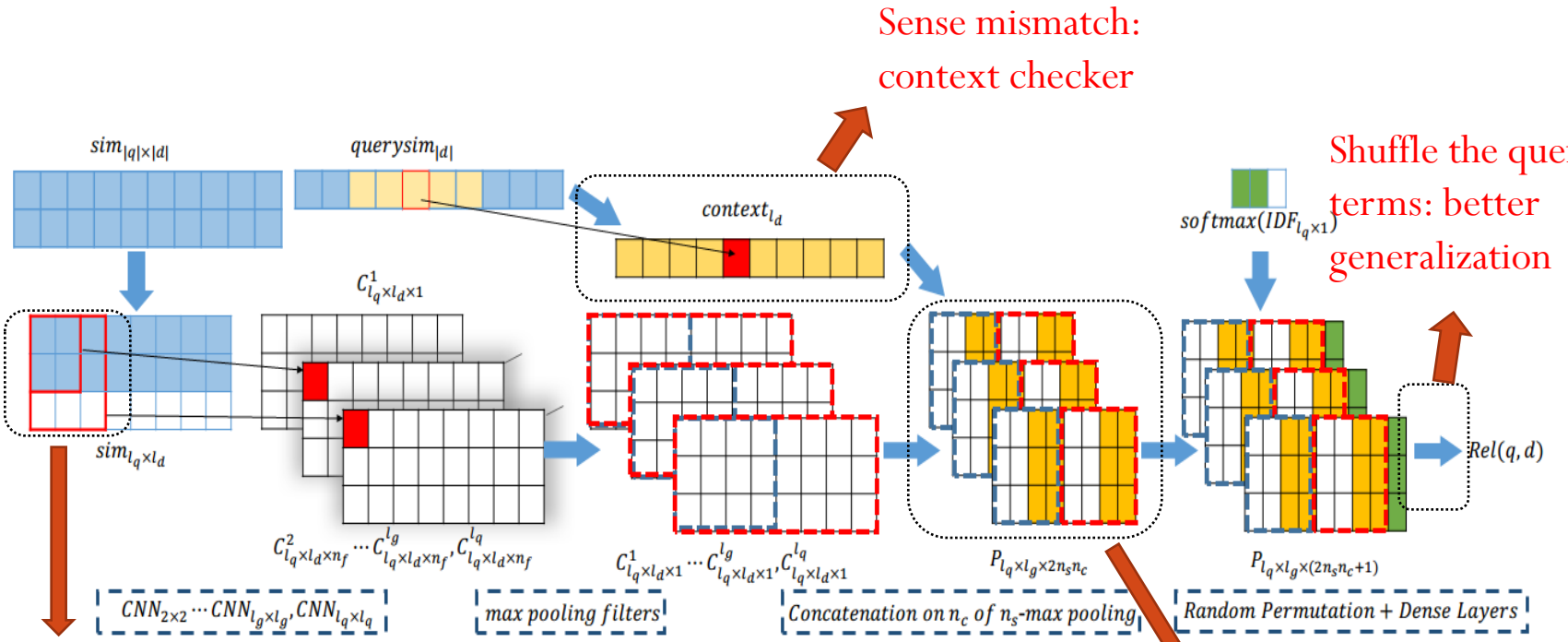
Cover of all query terms, meanwhile, assume relevance signals for individual query terms are independent, so that the relevance signals could be shuffled before combination.

- Cascade reading model.

Max-pool salient signals in cascade manners.

Design of Modular

- Please refer to our paper and poster for more technical details.



max planck institut informatik

Evaluation

- Based on TREC Web Track ad-hoc task 2009-2014.
- Measures: nDCG@20 and ERR@20.
- Benchmarks:

RerankSimple: re-rank search results from a simple ranker, namely, query-likelihood model.
RerankALL: re-rank different runs from TREC, examining the applicability and the improvements.
PairAccuracy: cast as classification problems on individual document pairs.

- Baseline models: DRMM, local model in DUET, PACRR and MatchPyramid.

Training and Validation

- Split the six years into four years for training, one year for validation and one year for test.
- In total, there are 15 such train/validation/test combinations.
- For each year, there are five predictions based on different training/validation combinations.
- Significant tests are based on these five predictions for individual comparisons.

Result: RerankSimple

Compare RE-PACRR with baselines. P/p, D/d, L/l and M/m indicate significant differences at 95% or 90% statistical level.

Rank relative to original TREC runs.

Measures	Year	RE-PACRR	PACRR	MatchPyramid	DUETL	DRMM
ERR@20	wt12	0.390 (p↑D↑L↑M↑) 121% 1	0.347 (d↑L↑M↑) 96% 1	0.309 (P↓L↑) 74% 5	0.218 (P↓D↓M↓) 23% 18	0.314 (p↓L↑) 78% 4
	wt13	0.190 (p↑D↑L↑M↑) 89% 1	0.175 (D↑L↑M↑) 74% 3	0.135 (P↓D↓) 34% 15	0.137 (P↓D↓) 36% 14	0.155 (P↓L↑M↑) 54% 7
	wt14	0.246 (P↑D↑L↑M↑) 88% 1	0.223 (D↑L↑M↑) 70% 1	0.183 (P↓) 40% 12	0.174 (P↓D↓) 33% 16	0.195 (P↓L↑) 49% 8

ERR@20.

Improvements relative to QL.

- All neural IR models can improve based on QL search results (omitted here).
- RE-PACRR can achieve top-1 by solely re-ranking the search results from query-likelihood model.

Result: RerankALL

----How many runs could be improved by a neural IR model?

Measures	Year	RE-PACRR	PACRR	MatchPyramid	DUETL	DRMM
ERR@20	wt09	91% (D↑L↑)	92% (D↑L↑m↑)	86% (p↓D↑l↑)	77% (P↓m↓)	72% (P↓M↓)
	wt10	98% (P↑D↑L↑M↑)	95% (D↑L↑)	95% (D↑L↑)	69% (P↓D↓M↓)	91% (P↓L↑M↓)
	wt11	98% (P↑D↑L↑M↑)	69% (D↑L↑M↑)	43% (P↓L↑)	26% (P↓D↓M↓)	49% (P↓L↑)
	wt12	98% (P↑d↑L↑M↑)	92% (L↑)	93% (L↑)	68% (P↓D↓M↓)	95% (L↑)
	wt13	94% (P↑D↑L↑M↑)	85% (L↑M↑)	64% (P↓d↓)	61% (P↓D↓)	83% (L↑m↑)
	wt14	96% (P↑D↑L↑M↑)	84% (L↑M↑)	58% (P↓)	52% (P↓)	68%

Percentage of runs that get improved.

- RE-PACRR significantly outperforms all baselines on five years.
- More than 95% of runs are improved by RE-PACRR.

Result: RerankALL

----By how much a neural IR model can improve?

Measures	Year	RE-PACRR	PACRR	MatchPyramid	DUETL	DRMM
ERR@20	wt09	43% (D↑L↑M↑)	40% (D↑L↑M↑)	31% (P↓D↑L↑)	22% (P↓M↓)	20% (P↓M↓)
	wt10	98% (P↑D↑L↑M↑)	74% (D↑L↑M↑)	54% (P↓d↑L↑)	23% (P↓M↓)	44% (P↓m↓)
	wt11	33% (P↑D↑L↑M↑)	11% (D↑L↑M↑)	-4% (P↓)	-11% (P↓D↓)	-0% (P↓L↑)
	wt12	89% (P↑D↑L↑)	66% (L↑)	68% (L↑)	22% (P↓D↓M↓)	70% (L↑)
	wt13	36% (P↑D↑L↑M↑)	27% (L↑M↑)	9% (P↓D↓)	8% (P↓D↓)	20% (L↑M↑)
	wt14	29% (P↑D↑L↑M↑)	16% (d↑L↑M↑)	5% (P↓)	2% (P↓)	8% (p↓)

Average differences on all runs between the measure scores before and after re-ranking.

- RE-PACRR significantly outperforms all baselines on four years.
- At least 29% of improvements are observed on individual years.

Result: PairAccuracy

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

Pairs of different labels in the ground truth.

Percentage of the number of document pairs with the particular labels.

Label Pair	volume (%)	# queries	Year	RE-PACRR
HRel-NRel	23.1%	262	wt09	0.715 (P↑D↑L↑M↑)
			wt10	0.846 (D↑L↑M↑)
			wt11	0.837 (P↑D↑L↑M↑)
			wt12	0.826 (P↑D↑L↑M↑)
			wt13	0.758 (D↑L↑M↑)
			wt14	0.766 (D↑L↑M↑)
HRel-Rel	8.4%	257	wt09	0.531
			wt10	0.587 (p↑D↑L↑)
			wt11	0.582 (P↑d↓L↑)
			wt12	0.671 (D↑L↑M↑)
			wt13	0.572 (D↑L↑)
			wt14	0.602 (P↑D↑L↑M↑)
Rel-NRel	63.5%	290	wt09	0.682 (P↑D↑L↑M↑)
			wt10	0.799 (D↑L↑M↑)
			wt11	0.782 (D↑L↑M↑)
			wt12	0.741 (P↑D↑L↑M↑)
			wt13	0.707 (p↑D↑L↑M↑)
			wt14	0.700 (P↓D↑L↑M↑)

- RE-PACRR performs better on Hrel-NRel and Rel-NRel, and gets close to other models on Hrel-Rel.
- The overall accuracy is beyond 70%.

max planck institut
informatik

Thank You!

MAX-PLANCK-GESELLSCHAFT