End-to-end Neural Information Retrieval

MMath thesis

Wei Yang

Cheriton School of Computer Science
University of Waterloo

April 2019

Table of Contents

UNIVERSITY OF WATERLOO

Table of Contents

UNIVERSITY OF WATERLOO

Types of NLP Tasks

- Sequence classification
- Sequence pair classification (text matching)
- Sequence labeling
- Sequence-to-sequence generation

UNIVERSITY OF WATERLOO

Types of NLP Tasks

- Sequence classification
- **Sequence pair classification (text matching)**
- Sequence labeling
- Sequence-to-sequence generation

UNIVERSITY OF WATERLOO

Types of NLP Tasks

- Sequence classification
- **Sequence pair classification (text matching)**
- Sequence labeling
- Sequence-to-sequence generation

UNIVERSITY OF WATERLOO

Types of NLP Tasks

- Sequence classification
- **Sequence pair classification (text matching)**
- Sequence labeling
- Sequence-to-sequence generation

UNIVERSITY OF WATERLOO

Information Retrieval

Tasks	Text 1	Text 2	Objective
Paraphrase Indentification	string 1	string 2	C
Textual Entailment	text	hypothesis	C
Question Answering	question	answer	C/R
Conversation	dialog	response	C/R
Information Retrieval	query	document	R

Table: Typical text matching tasks (C: classification; R: ranking)

UNIVERSITY OF WATERLOO

Problem Definition

- **Search microblogs**:
  - **Query**: 2022 fifa soccer
  - **Relevant document**: #ps3 best sellers: fifa soccer 11 ps3 #cheaptweet https://www.amazon.com/fifa-soccer-11-playstation-3
- **Search newswire articles**:
  - **Query**: international organized crime
  - **Relevant document**: The past few years have been characterized by an unprecedented growth in crime, changes in its characteristics, and for all practical purposes the loss of state and public control over the crime situation...

    More than 40 international smuggling crime groups have been identified. More than 130 "Russian" stores selling Russian antiques have been found abroad...

UNIVERSITY OF WATERLOO

Problem Definition

- **Search microblogs**:
  - **Query**: 2022 fifa soccer
  - **Relevant document**: #ps3 best sellers: fifa soccer 11 ps3 #cheaptweet https://www.amazon.com/fifa-soccer-11-playstation-3
- **Search newswire articles**:
  - **Query**: international organized crime
  - **Relevant document**: The past few years have been characterized by an unprecedented growth in crime, changes in its characteristics, and for all practical purposes the loss of state and public control over the crime situation...

    More than 40 international smuggling crime groups have been identified. More than 130 "Russian" stores selling Russian antiques have been found abroad...

UNIVERSITY OF WATERLOO

Why Neural Network

Why NN for NLP?

- Low-dimensional semantic space
- Thousands of variations
- Hierarchical structure
- Hardware developments

Why NN for IR?

- Relevance judgments are based on a complicated human cognitive process

UNIVERSITY OF WATERLOO

Why Neural Network

Why NN for NLP?

- Low-dimensional semantic space
- Thousands of variations
- Hierarchical structure
- Hardware developments

Why NN for IR?

- Relevance judgments are based on a complicated human cognitive process

UNIVERSITY OF WATERLOO

Why Neural Network

Why NN for NLP?

- Low-dimensional semantic space
- Thousands of variations
- Hierarchical structure
- Hardware developments

Why NN for IR?

- Relevance judgments are based on a complicated human cognitive process

UNIVERSITY OF WATERLOO

Why Neural Network

Why NN for NLP?

- Low-dimensional semantic space
- Thousands of variations
- Hierarchical structure
- Hardware developments

Why NN for IR?

- Relevance judgments are based on a complicated human cognitive process

UNIVERSITY OF WATERLOO

Why Neural Network

Why NN for NLP?

- Low-dimensional semantic space
- Thousands of variations
- Hierarchical structure
- Hardware developments

Why NN for IR?

- Relevance judgments are based on a complicated human cognitive process

UNIVERSITY OF WATERLOO

Why Neural Network

Why NN for NLP?

- Low-dimensional semantic space
- Thousands of variations
- Hierarchical structure
- Hardware developments

Why NN for IR?

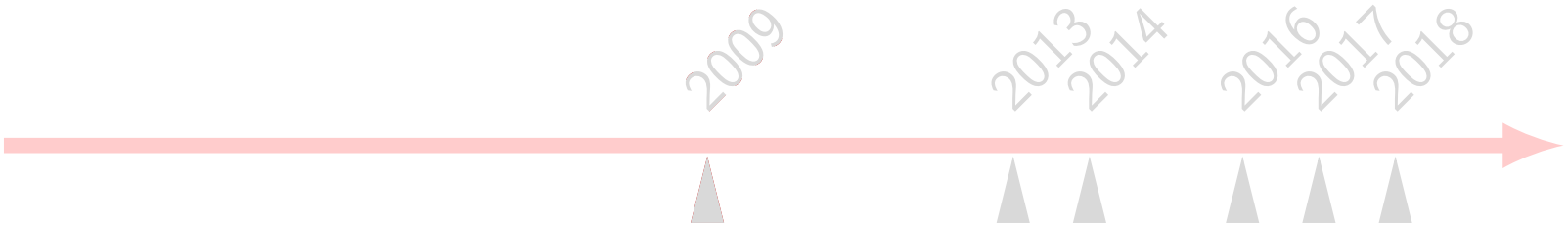
- Relevance judgments are based on a complicated human cognitive process

UNIVERSITY OF WATERLOO

Table of Contents

UNIVERSITY OF WATERLOO

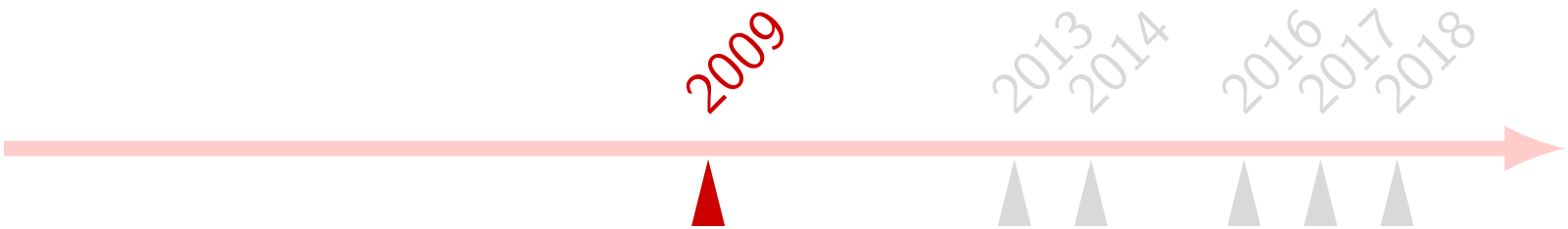
Related Work



- before 2009: vector space models and probabilistic models (Query likelihood (QL), BM25, RM3...)
- 2009: Learning to rank models (tens of hand-crafted features)
- 2013: Deep Structured Semantic Model (DSSM)
- 2014: CDSSM, ARC-I, ARC-II (mainly for short text ranking)
- 2016: MatchPyramid, DRMM ...
- 2017: KNRM, DUET, DeepRank, PACRR ...
- 2018: HINT, MP-HCNN (hierachical matching patterns) ...

UNIVERSITY OF WATERLOO

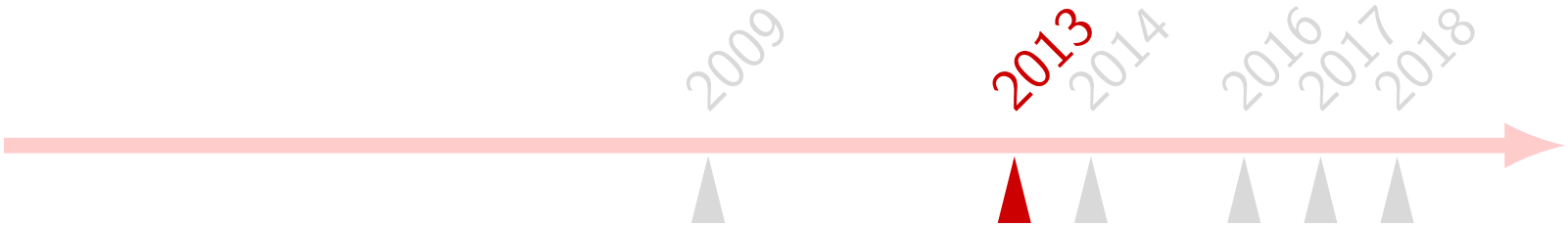
Related Work



- before 2009: vector space models and probabilistic models (Query likelihood (QL), BM25, RM3...)
- 2009: Learning to rank models (tens of hand-crafted features)
- 2013: Deep Structured Semantic Model (DSSM)
- 2014: CDSSM, ARC-I, ARC-II (mainly for short text ranking)
- 2016: MatchPyramid, DRMM ...
- 2017: KNRM, DUET, DeepRank, PACRR ...
- 2018: HINT, MP-HCNN (hierachical matching patterns) ...

UNIVERSITY OF WATERLOO

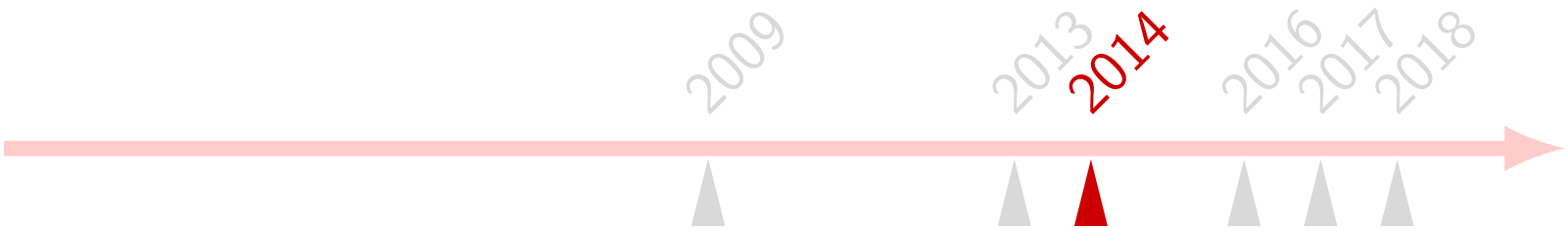
Related Work



- before 2009: vector space models and probabilistic models (Query likelihood (QL), BM25, RM3...)
- 2009: Learning to rank models (tens of hand-crafted features)
- 2013: Deep Structured Semantic Model (DSSM)
- 2014: CDSSM, ARC-I, ARC-II (mainly for short text ranking)
- 2016: MatchPyramid, DRMM ...
- 2017: KNRM, DUET, DeepRank, PACRR ...
- 2018: HINT, MP-HCNN (hierachical matching patterns) ...

UNIVERSITY OF WATERLOO

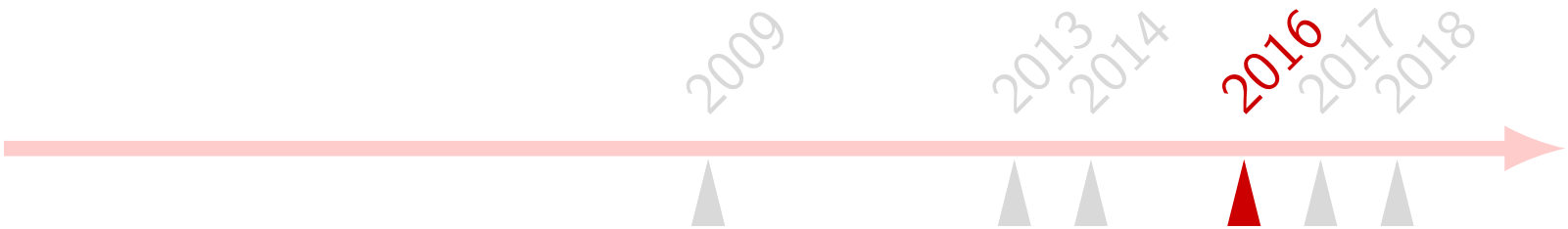
Related Work



- before 2009: vector space models and probabilistic models (Query likelihood (QL), BM25, RM3...)
- 2009: Learning to rank models (tens of hand-crafted features)
- 2013: Deep Structured Semantic Model (DSSM)
- 2014: CDSSM, ARC-I, ARC-II (mainly for short text ranking)
- 2016: MatchPyramid, DRMM ...
- 2017: KNRM, DUET, DeepRank, PACRR ...
- 2018: HINT, MP-HCNN (hierachical matching patterns) ...

UNIVERSITY OF WATERLOO

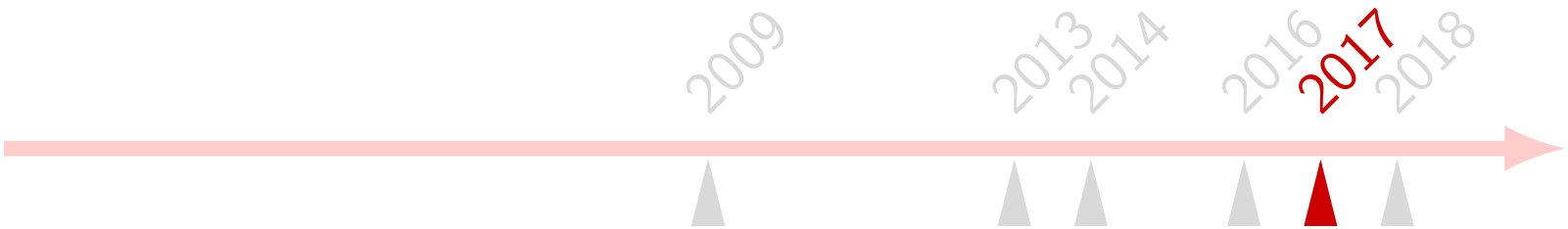
Related Work



- before 2009: vector space models and probabilistic models (Query likelihood (QL), BM25, RM3...)
- 2009: Learning to rank models (tens of hand-crafted features)
- 2013: Deep Structured Semantic Model (DSSM)
- 2014: CDSSM, ARC-I, ARC-II (mainly for short text ranking)
- 2016: MatchPyramid, DRMM ...
- 2017: KNRM, DUET, DeepRank, PACRR ...
- 2018: HINT, MP-HCNN (hierachical matching patterns) ...

UNIVERSITY OF WATERLOO

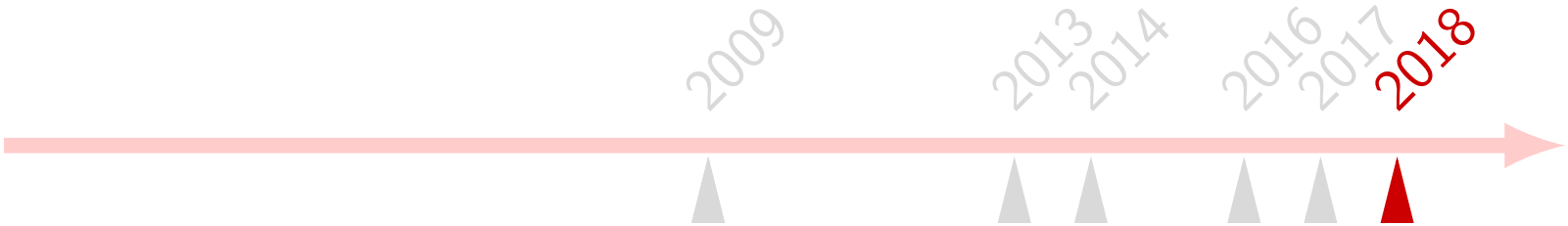
Related Work



- before 2009: vector space models and probabilistic models (Query likelihood (QL), BM25, RM3...)
- 2009: Learning to rank models (tens of hand-crafted features)
- 2013: Deep Structured Semantic Model (DSSM)
- 2014: CDSSM, ARC-I, ARC-II (mainly for short text ranking)
- 2016: MatchPyramid, DRMM ...
- 2017: KNRM, DUET, DeepRank, PACRR ...
- 2018: HINT, MP-HCNN (hierachical matching patterns) ...

UNIVERSITY OF WATERLOO

Related Work



- before 2009: vector space models and probabilistic models (Query likelihood (QL), BM25, RM3...)
- 2009: Learning to rank models (tens of hand-crafted features)
- 2013: Deep Structured Semantic Model (DSSM)
- 2014: CDSSM, ARC-I, ARC-II (mainly for short text ranking)
- 2016: MatchPyramid, DRMM ...
- 2017: KNRM, DUET, DeepRank, PACRR ...
- 2018: HINT, MP-HCNN (hierachical matching patterns) ...

UNIVERSITY OF WATERLOO

Contribution

- **An end-to-end retrieval and reranking system** to allow the user to apply different retrieval models and neural reranking models on different datasets.
- **State-of-the-art** performance on two benchmark datasets (Robust04 and Microblog) for document retrieval.
- Prove the effectiveness and additivity of a **strong baseline** for neural reranking methods.
- Co-design the **MP-HCNN** model for social media post searching.

UNIVERSITY OF WATERLOO

Contribution

- **An end-to-end retrieval and reranking system** to allow the user to apply different retrieval models and neural reranking models on different datasets.
- **State-of-the-art** performance on two benchmark datasets (Robust04 and Microblog) for document retrieval.
- Prove the effectiveness and additivity of a **strong baseline** for neural reranking methods.
- Co-design the **MP-HCNN** model for social media post searching.

UNIVERSITY OF WATERLOO

Contribution

- **An end-to-end retrieval and reranking system** to allow the user to apply different retrieval models and neural reranking models on different datasets.
- **State-of-the-art** performance on two benchmark datasets (Robust04 and Microblog) for document retrieval.
- Prove the effectiveness and additivity of a **strong baseline** for neural reranking methods.
- Co-design the **MP-HCNN** model for social media post searching.

UNIVERSITY OF WATERLOO

Contribution

- **An end-to-end retrieval and reranking system** to allow the user to apply different retrieval models and neural reranking models on different datasets.
- **State-of-the-art** performance on two benchmark datasets (Robust04 and Microblog) for document retrieval.
- Prove the effectiveness and additivity of a **strong baseline** for neural reranking methods.
- Co-design the **MP-HCNN** model for social media post searching.

UNIVERSITY OF WATERLOO

Table of Contents

UNIVERSITY OF WATERLOO

Architecture

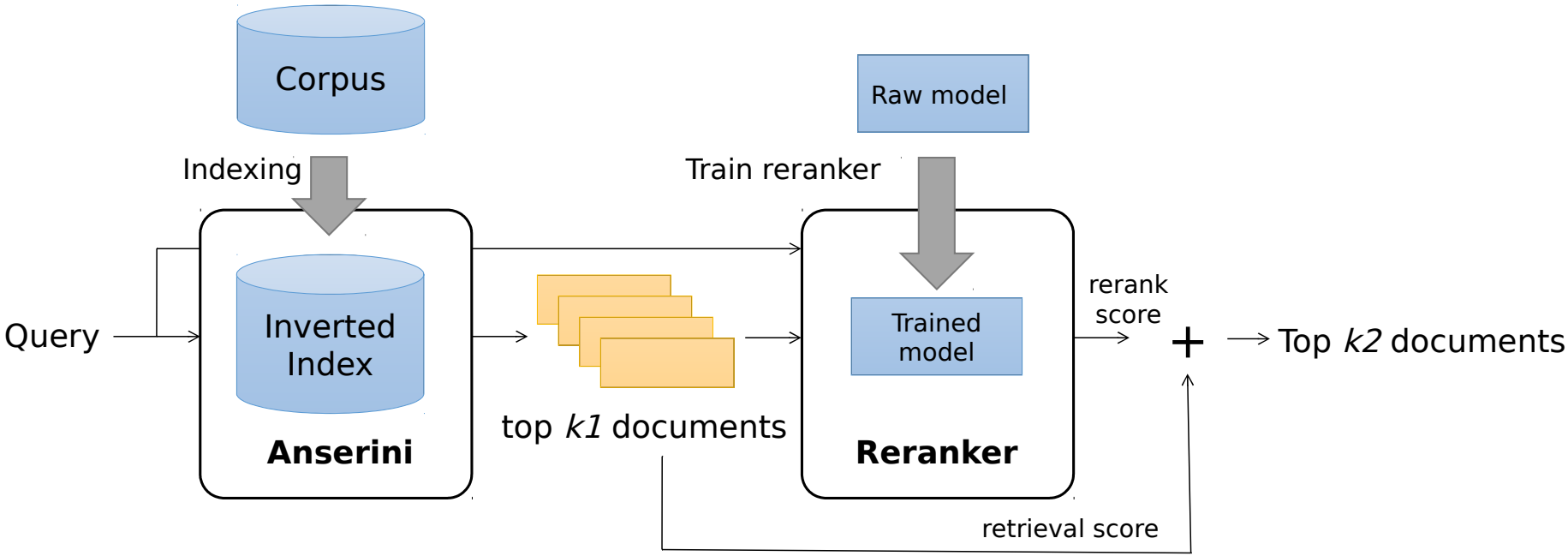


Figure: The architecture of the Retrieval-rerank framework.

Retrieval, Rerank and Aggregation

- Retrieval: Anserini (QL, QL+RM3, BM25, BM25+RM3)
- Rerank:
  - MatchZoo models (DSSM, CDSSM, DUET, KNRM, DRMM)
  - MP-HCNN
  - BERT
- Aggregation:

$$\text{rel}(q, d) = \lambda * \text{Reranker}(q, d) + (1 - \lambda) * \text{Retriever}(q, d) \quad (1)$$

UNIVERSITY OF WATERLOO

Retrieval, Rerank and Aggregation

- Retrieval: Anserini (QL, QL+RM3, BM25, BM25+RM3)
- Rerank:
  - MatchZoo models (DSSM, CDSSM, DUET, KNRM, DRMM)
  - MP-HCNN
  - BERT
- Aggregation:

$$\text{rel}(q, d) = \lambda * \text{Reranker}(q, d) + (1 - \lambda) * \text{Retriever}(q, d) \quad (1)$$

UNIVERSITY OF WATERLOO

Retrieval, Rerank and Aggregation

- Retrieval: Anserini (QL, QL+RM3, BM25, BM25+RM3)
- Rerank:
  - MatchZoo models (DSSM, CDSSM, DUET, KNRM, DRMM)
  - MP-HCNN
  - BERT
- Aggregation:

$$\text{rel}(q, d) = \lambda * \text{Reranker}(q, d) + (1 - \lambda) * \text{Retriever}(q, d) \quad (1)$$

UNIVERSITY OF WATERLOO

Retrieval, Rerank and Aggregation

- Retrieval: Anserini (QL, QL+RM3, BM25, BM25+RM3)
- Rerank:
  - MatchZoo models (DSSM, CDSSM, DUET, KNRM, DRMM)
  - MP-HCNN
  - BERT
- Aggregation:

$$\text{rel}(q, d) = \lambda * \text{Reranker}(q, d) + (1 - \lambda) * \text{Retriever}(q, d) \quad (1)$$

UNIVERSITY OF WATERLOO

Retrieval, Rerank and Aggregation

- Retrieval: Anserini (QL, QL+RM3, BM25, BM25+RM3)
- Rerank:
  - MatchZoo models (DSSM, CDSSM, DUET, KNRM, DRMM)
  - MP-HCNN
  - BERT
- Aggregation:

$$\text{rel}(q, d) = \lambda * \text{Reranker}(q, d) + (1 - \lambda) * \text{Retriever}(q, d) \quad (1)$$

UNIVERSITY OF WATERLOO

MatchZoo

Model	Task	Dataset	NN Architecture		
			Encoding	Hidden	Combination
DSSM (2013)	Web Search	clickthrough data	Word Hashing of Letter-Trigram	MLP	Dot + softmax
CDSSM (2014)	Web Search	clickthrough data	Word Hashing of Letter-Trigram	Conv1D + MLP	Dot + softmax + Max Pooling
DRMM (2016)	Ad-hoc Retrieval	Robust04 and ClueWeb09B	Query: Word Embedding; Doc: local interac-tion+matching histogram	MLP	Dot
DUET (2017)	Web Search	Bings search logs	LM: one-hot vector; DM: word embedding	Conv1D	LM: intersection; DM: entrywise product
K-NRM (2017)	Ad-hoc Retrieval	search logs of Sogou.com	Word embedding	Kernel pooling	Cosine

Figure: Details of five neural information retrieval models in MatchZoo

UNIVERSITY OF WATERLOO

MP-HCNN

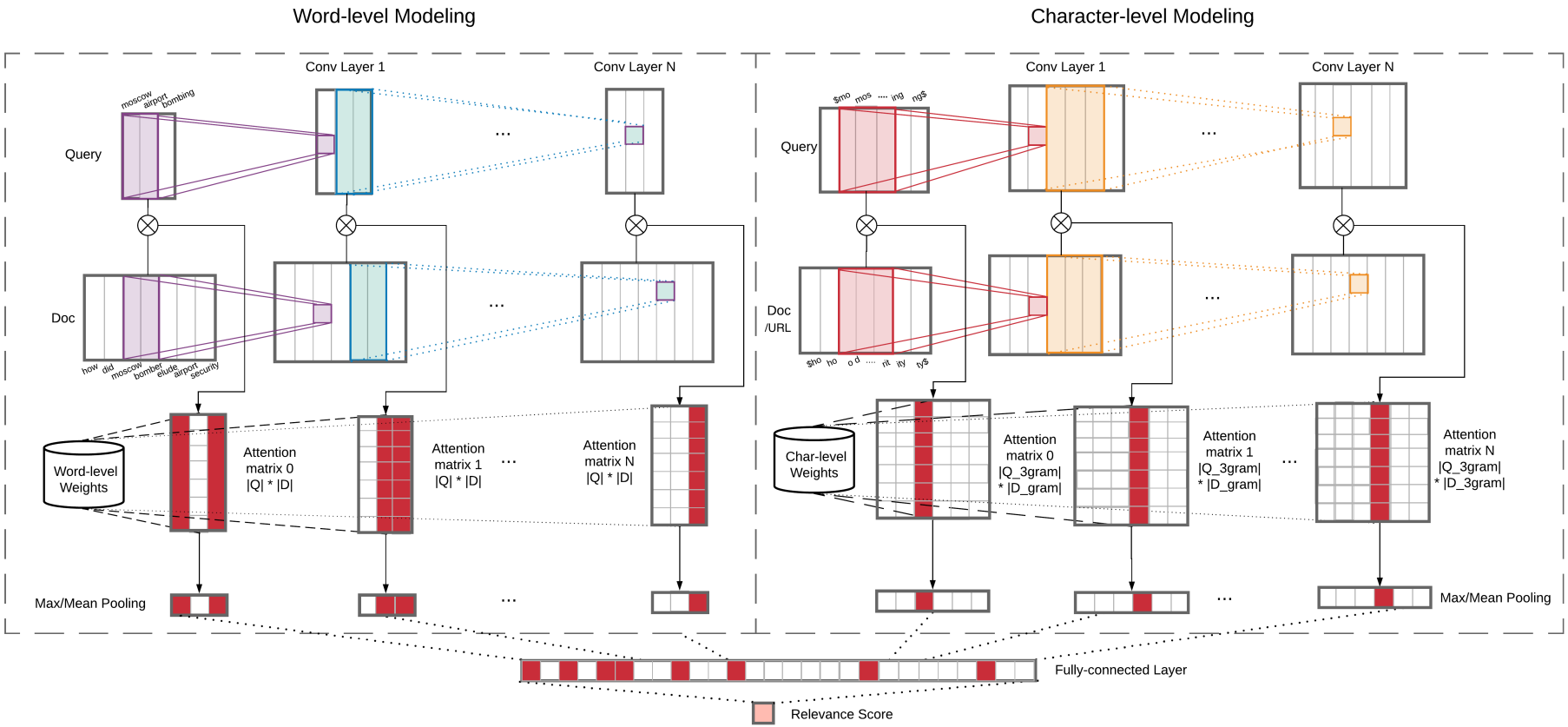


Figure: Overview of our Multi-Perspective Hierarchical Convolutional Neural Network (MP-HCNN), which consists of two parallel components for word-level and character-level modeling between queries, social media posts, and URLs.

BERT

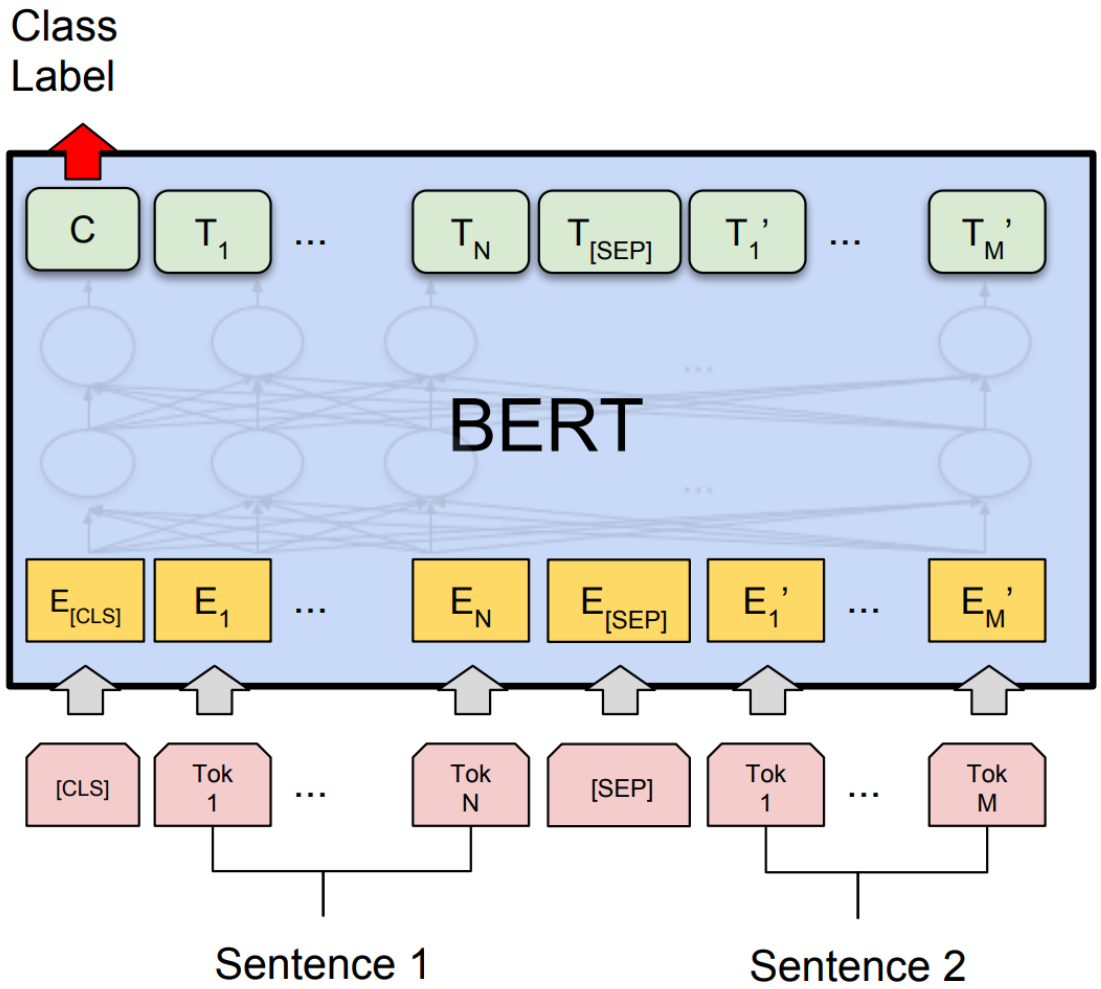


Figure: The architecture of the BERT model for text matching.

Table of Contents

UNIVERSITY OF WATERLOO

Evaluation Metrics

$$Precision_q = \frac{\sum_{\langle i,d \rangle \in R_q} rel_q(d)}{|R_q|} \tag{2}$$

$$AP_q = \frac{\sum_{\langle i,d \rangle \in R_q} Precision_{q,i} \times rel_q(d)}{\sum_{d \in D} rel_q(d)} \tag{3}$$

$$\mathrm{N}DCG_q = \frac{DCG_q}{IDCG_q} \tag{4}$$

$$\mathrm{D}CG_q = \sum_{\langle i,d \rangle \in R_q} \frac{2^{rel_q(d)} - 1}{\log_2(i+1)} \tag{5}$$

Evaluation Metrics

$$Precision_q = \frac{\sum_{\langle i,d \rangle \in R_q} rel_q(d)}{|R_q|} \tag{2}$$

$$AP_q = \frac{\sum_{\langle i,d \rangle \in R_q} Precision_{q,i} \times rel_q(d)}{\sum_{d \in D} rel_q(d)} \tag{3}$$

$$\mathrm{N}DCG_q = \frac{DCG_q}{IDCG_q} \tag{4}$$

$$\mathrm{D}CG_q = \sum_{\langle i,d \rangle \in R_q} \frac{2^{rel_q(d)} - 1}{\log_2(i+1)} \tag{5}$$

Evaluation Metrics

$$Precision_q = \frac{\sum_{\langle i,d \rangle \in R_q} rel_q(d)}{|R_q|} \tag{2}$$

$$AP_q = \frac{\sum_{\langle i,d \rangle \in R_q} Precision_{q,i} \times rel_q(d)}{\sum_{d \in D} rel_q(d)} \tag{3}$$

$$\mathrm{N}DCG_q = \frac{DCG_q}{IDCG_q} \tag{4}$$

$$\mathrm{D}CG_q = \sum_{\langle i,d \rangle \in R_q} \frac{2^{rel_q(d)} - 1}{\log_2(i+1)} \tag{5}$$

Datasets

Test Set	2011	2012	2013	2014
# query topics	49	60	60	55
# query-doc pairs	49,000	60,000	60,000	55,000

Table: Statistics of the TREC Microblog Track datasets

# query topics	250
# query-doc pairs	250,000

Table: Statistics of the Robust04 datasets

Datasets

Test Set	2011	2012	2013	2014
# query topics	49	60	60	55
# query-doc pairs	49,000	60,000	60,000	55,000

Table: Statistics of the TREC Microblog Track datasets

# query topics	250
# query-doc pairs	250,000

Table: Statistics of the Robust04 datasets

UNIVERSITY OF WATERLOO

Experimental Setup

- Train/test splits:
  - Microblog: train on 2011, 2012 and 2013, test on 2014
  - Robust04: five-fold cross validation
- Hyper-parameter tuning: 10% of the training data
- Models:
  - Microblog: MatchZoo models, MP-HCNN, BERT
  - Robust04: MatchZoo models

UNIVERSITY OF WATERLOO

Experimental Setup

- Train/test splits:
  - Microblog: train on 2011, 2012 and 2013, test on 2014
  - Robust04: five-fold cross validation
- Hyper-parameter tuning: 10% of the training data
- Models:
  - Microblog: MatchZoo models, MP-HCNN, BERT
  - Robust04: MatchZoo models

UNIVERSITY OF WATERLOO

Experimental Setup

- Train/test splits:
  - Microblog: train on 2011, 2012 and 2013, test on 2014
  - Robust04: five-fold cross validation
- Hyper-parameter tuning: 10% of the training data
- Models:
  - Microblog: MatchZoo models, MP-HCNN, BERT
  - Robust04: MatchZoo models

UNIVERSITY OF WATERLOO

Results of Baselines: Robust04

Model	AP	P@20	NDCG@20
QL (Guo et al.)	0.253	0.369	0.415
BM25 (Guo et al.)	0.255	0.370	0.418
DRMM (Guo et al.)	0.279	0.382	0.431
MatchPyramid (Pang et al.)	0.232	0.327	0.411
BM25 (Mcdonald et al.)	0.238	0.354	0.425
PACRR (Mcdonald et al.)	0.258	0.372	0.443

Table: Previous Results on the Robust04 dataset

	QL	QL+RM3	BM25	BM25+RM3
AP	0.2465	0.2743	0.2515	0.3033
P@20	0.3508	0.3639	0.3612	0.3973
NDCG@20	0.4109	0.4172	0.4225	0.4514

Table: Our results of retrieval models on the Robust04 dataset

UNIVERSITY OF WATERLOO

Results of Baselines: Robust04

Model	AP	P@20	NDCG@20
QL (Guo et al.)	0.253	0.369	0.415
BM25 (Guo et al.)	0.255	0.370	0.418
DRMM (Guo et al.)	0.279	0.382	0.431
MatchPyramid (Pang et al.)	0.232	0.327	0.411
BM25 (Mcdonald et al.)	0.238	0.354	0.425
PACRR (Mcdonald et al.)	0.258	0.372	0.443

Table: Previous Results on the Robust04 dataset

	QL	QL+RM3	BM25	BM25+RM3
AP	0.2465	0.2743	0.2515	0.3033
P@20	0.3508	0.3639	0.3612	0.3973
NDCG@20	0.4109	0.4172	0.4225	0.4514

Table: Our results of retrieval models on the Robust04 dataset

UNIVERSITY OF WATERLOO

Results of End-to-end Neural IR Models: Robust04

Models	MAP	P@20	NDCG@20
BM25+RM3	0.3033	0.3973	0.4514
DSSM	0.0982^{-}	0.1331^{-}	0.1551^{-}
CDSSM	0.0641^{-}	0.0842^{-}	0.0772^{-}
DRMM	0.2543^{-}	0.3405^{-}	0.4025^{-}
KNRM	0.1145^{-}	0.1480^{-}	0.1512^{-}
DUET	0.1426^{-}	0.1561^{-}	0.1946^{-}
DSSM+RM3	0.3026	0.3946	0.4491
CDSSM+RM3	0.2995	0.3944	0.4468
DRMM+RM3	$\mathbf{0.3151}^{+}$	$\mathbf{0.4147}^{+}$	$\mathbf{0.4717}^{+}$
KNRM+RM3	0.3036	0.3928	0.4441
DUET+RM3	0.3051	0.3986	0.4502

Table: Results of retrieval and reranking on the Robust04 dataset. RM: retrieval model. NRM: neural re-ranking model. Significant improvement or degradation with respect to the retrieval model is indicated (+/-) (p-value ≤ 0.05).

Results of Baselines: Microblog

Method	AP	P@30
QL (Rao et al.)	0.3924	0.6182
RM3 (Rao et al.)	0.4480	0.6339
L2R (Rao et al.)	0.3943	0.6200
MP-HCNN (Rao et al.)	0.4409	0.6612
BiCNN (Shi et al.)	0.4563	0.6806

Table: Previous Results on Microblog datasets

	QL	QL+RM3	BM25	BM25+RM3
AP	0.4181	0.4676	0.3931	0.4374
P@30	0.6430	0.6533	0.6212	0.6442

Table: Our results of retrieval models on Microblog datasets

UNIVERSITY OF WATERLOO

Results of Baselines: Microblog

Method	AP	P@30
QL (Rao et al.)	0.3924	0.6182
RM3 (Rao et al.)	0.4480	0.6339
L2R (Rao et al.)	0.3943	0.6200
MP-HCNN (Rao et al.)	0.4409	0.6612
BiCNN (Shi et al.)	0.4563	0.6806

Table: Previous Results on Microblog datasets

	QL	QL+RM3	BM25	BM25+RM3
AP	0.4181	0.4676	0.3931	0.4374
P@30	0.6430	0.6533	0.6212	0.6442

Table: Our results of retrieval models on Microblog datasets

UNIVERSITY OF WATERLOO

Results of End-to-end Neural IR Models: Microblog

Models	AP	P@30
QL+RM3	0.4676	0.6533
DSSM	0.2634−	0.3836−
CDSSM	0.1936−	0.2636−
DRMM	0.4477−	0.6127−
KNRM	0.3432−	0.5121−
DUET	0.2713−	0.3533−
MP-HCNN	0.4497	0.6219
BERT	0.4646	0.6509
DSSM+RM3	0.4666	0.6539
CDSSM+RM3	0.4703	0.6624
DRMM+RM3	0.4862+	0.6703
KNRM+RM3	0.4848+	0.6624
DUET+RM3	0.4844+	0.6594
MP-HCNN+RM3	0.4902+	0.6712
BERT+RM3	**0.5011**+	**0.6842**+

UNIVERSITY OF WATERLOO

Per-topic Analysis: Microblog

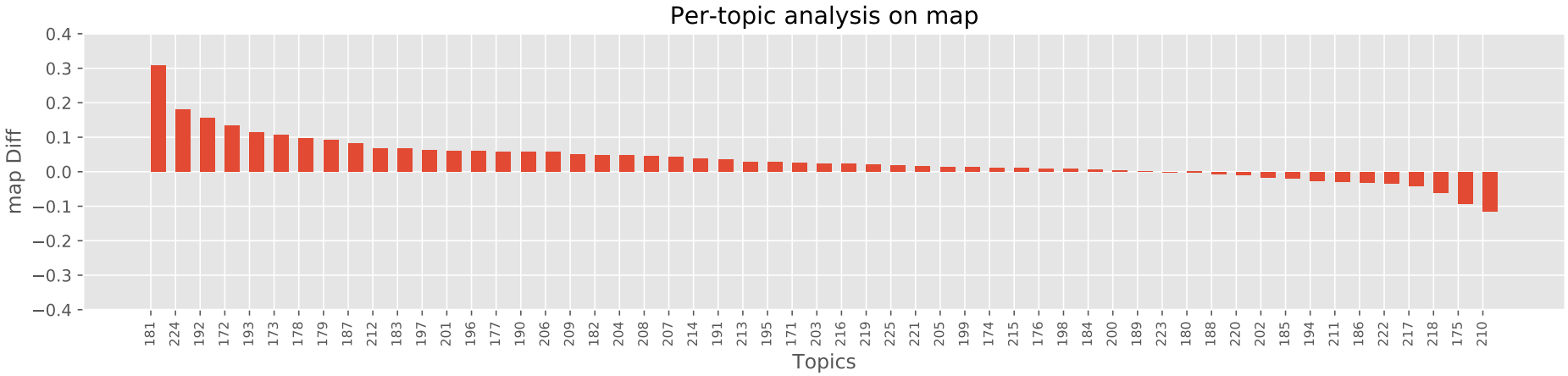


Figure: Per-topic differencesbetween BERT+RM3 and QL+RM3

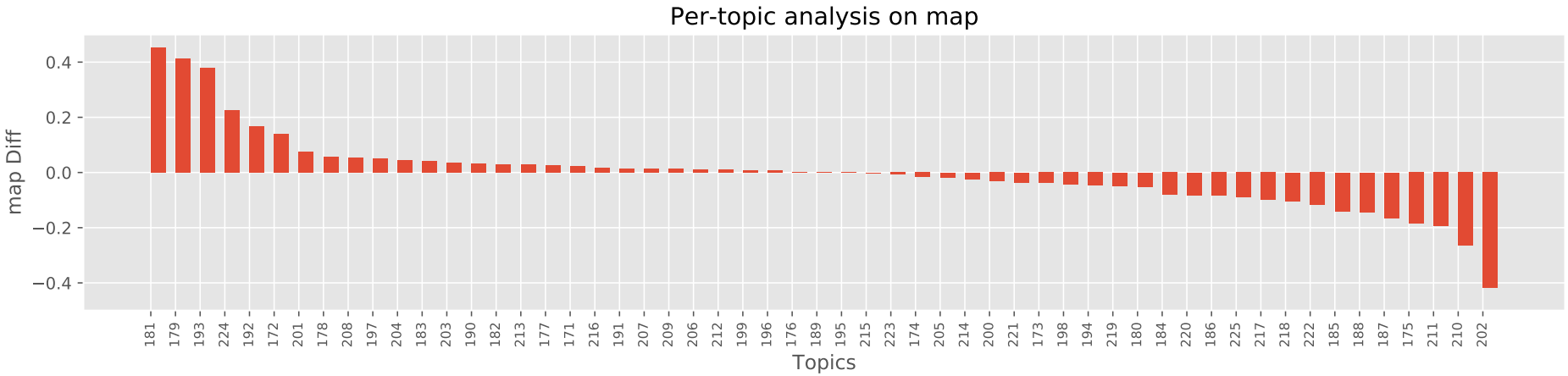


Figure: Per-topic differences between BERT and QL+RM3

WATERLOO

Per-topic Analysis: Robust04

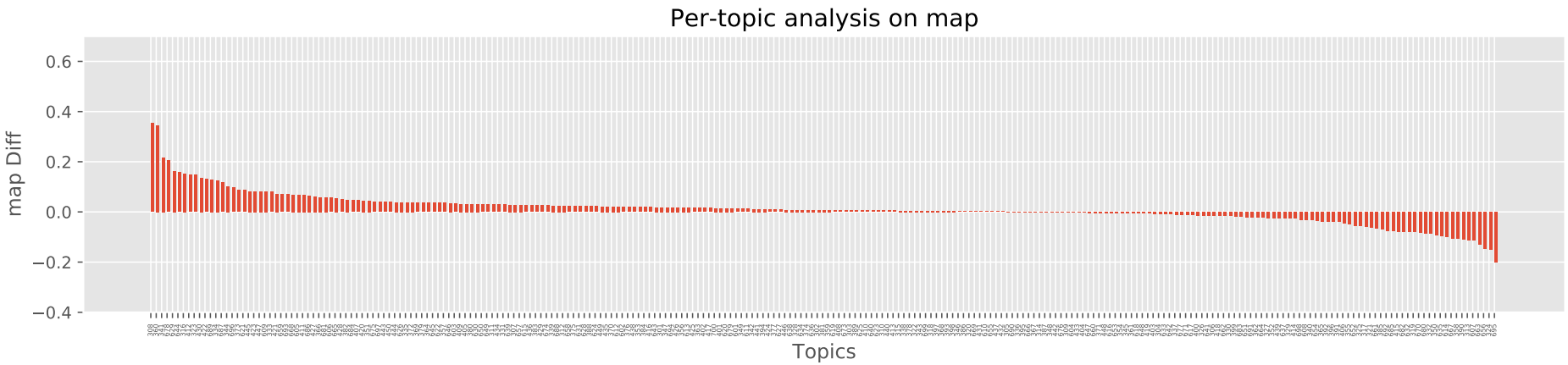


Figure: Per-topic differences between DRMM+RM3 and BM25+RM3

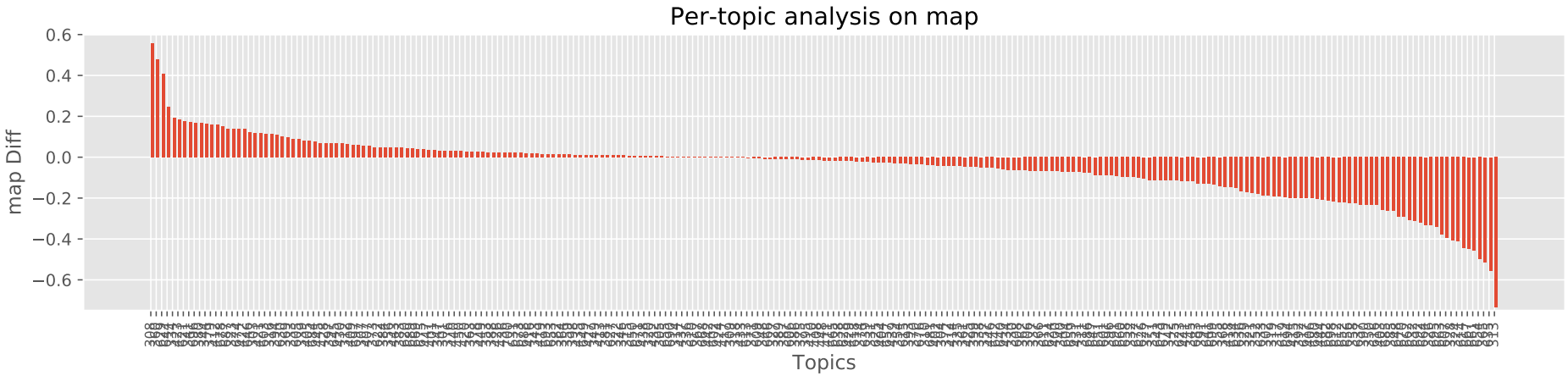


Figure: Per-topic differences between DRMM and BM25+RM3

Sample Analysis: Microblog

Category	Percentage (%)
Exact word match	100
Exact phrase match	41
Partial paraphrase match	64
Partial URL match	24

Table: Matching evidence breakdown by category based on manual analysis of the top 100 tweets for the five best-performing topics with MP-HCNN on the Microblog dataset.

UNIVERSITY OF WATERLOO

Table of Contents

UNIVERSITY OF WATERLOO

Conclusion and Discussion

- 4, 7, and 2
- SOTA
- Discussion:
  - relevance v.s. similarity
  - exact matching v.s. semantic matching
  - effectiveness v.s. efficiency
  - external knowledge v.s. domain-specific design

UNIVERSITY OF WATERLOO

Conclusion and Discussion

- 4, 7, and 2
- SOTA
- Discussion:
  - relevance v.s. similarity
  - exact matching v.s. semantic matching
  - effectiveness v.s. efficiency
  - external knowledge v.s. domain-specific design

UNIVERSITY OF WATERLOO

Conclusion and Discussion

- 4, 7, and 2
- SOTA
- Discussion:
  - relevance v.s. similarity
  - exact matching v.s. semantic matching
  - effectiveness v.s. efficiency
  - external knowledge v.s. domain-specific design

UNIVERSITY OF WATERLOO

Conclusion and Discussion

- 4, 7, and 2
- SOTA
- Discussion:
  - relevance v.s. similarity
  - exact matching v.s. semantic matching
  - effectiveness v.s. efficiency
  - external knowledge v.s. domain-specific design

UNIVERSITY OF WATERLOO

Conclusion and Discussion

- 4, 7, and 2
- SOTA
- Discussion:
  - relevance v.s. similarity
  - exact matching v.s. semantic matching
  - effectiveness v.s. efficiency
  - external knowledge v.s. domain-specific design

UNIVERSITY OF WATERLOO

Conclusion and Discussion

- 4, 7, and 2
- SOTA
- Discussion:
  - relevance v.s. similarity
  - exact matching v.s. semantic matching
  - effectiveness v.s. efficiency
  - external knowledge v.s. domain-specific design

UNIVERSITY OF WATERLOO