A Deep Relevance Matching Model for Ad-hoc Retrieval

Jiafeng Guo\textsuperscript{1}, Yixing Fan\textsuperscript{1}, Qingyao Ai\textsuperscript{2}, W. Bruce Croft\textsuperscript{2}

\textsuperscript{1}CAS Key Lab of Web Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences

\textsuperscript{2}Center for Intelligent Information Retrieval, University of Massachusetts Amherst
Outline

• Motivation
• Problem Analysis
• Our Approach
• Experiments
• Conclusions
Outline

• Motivation
• Problem Analysis
• Our Approach
• Experiments
• Conclusions
Machine learning to Rank

- Machine learning methods have been successfully applied to Information retrieval (IR)

- Human defined features
  - Time-consuming
  - Incomplete
  - Over-specified

\[
\begin{align*}
q^{(1)} &= (d_{1}^{(1)}, 4), (d_{1}^{(1)}, 2), (d_{1}^{(1)}, 1) \\
q^{(2)} &= (d_{1}^{(1)}, 4), (d_{1}^{(1)}, 2), (d_{1}^{(1)}, 1)
\end{align*}
\]

Learning Systems

Min Loss

Model \( f(x, w) \)

\( x \): feature representation of \(<q,d>\)
Success of deep learning

- Deep neural networks have led to exciting breakthroughs in a variety of applications
  - Discover the hidden structures and features at different levels of abstraction from the training data that are useful for the tasks

Computer Vision  Speech Recognition  Machine Translation

- Natural to apply deep learning methods for representation learning in IR
Ad-hoc Retrieval as a Matching Problem

• The core problem in ad-hoc retrieval can be formalized as a text matching problem

\[
\text{match}(T_1, T_2) = F(\phi(T_1), \phi(T_2))
\]

Scoring function based on the interaction between texts

Map each text to a representation vector
Ad-hoc Retrieval as a Matching Problem

• The core problem in ad-hoc retrieval can be formalized as a text matching problem

\[ \text{match}(T_1, T_2) = F(\phi(T_1), \phi(T_2)) \]

Scoring function based on the interaction between texts

Map each text to a representation vector

• A general formalization

<table>
<thead>
<tr>
<th>Task</th>
<th>T₁</th>
<th>T₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad-hoc Retrieval</td>
<td>Query</td>
<td>Document</td>
</tr>
<tr>
<td>Paraphrase Identification</td>
<td>String A</td>
<td>String B</td>
</tr>
<tr>
<td>Question Answering</td>
<td>Question</td>
<td>Answer</td>
</tr>
<tr>
<td>Automatic Conversation</td>
<td>Dialog</td>
<td>Response</td>
</tr>
</tbody>
</table>
Existing Deep Matching Models

• **Representation-focused Models**
  • build a good representation for a single text with a deep neural network
  • Conduct matching between two compositional and abstract text representations
  • Representative methods: DSSM [Huang et al. 2013], C-DSSM [Shen et al. 2014], ARC-I [Hu et al. 2014]
Existing Deep Matching Models

• **Interaction-focused Models**
  • Build the local interactions between two texts based on some basic representations
  • Use deep neural networks to learn the hierarchical interaction patterns for matching
  • Representative methods: DeepMatch [Lu et al. 2014], ARC-II [Hu et al. 2014], MatchPyramid [Pang et al. 2016]

\[
match(T_1, T_2) = F(\phi(T_1), \phi(T_2))
\]

A simple mapping function, e.g. sequence of word vectors

Complex deep model, e.g. CNN, DNN

hierarchical deep architecture over the local interaction matrix
Observations

• However...
  • Mostly demonstrated effective on a set of NLP tasks
    • e.g. paraphrase identification, QA
  • Tested on non-typical ad-hoc retrieval setting
    • e.g., DSSM and C-DSSM, were only evaluated on the <query, doc title> pairs
  • Directly apply on benchmark retrieval collections: relatively poor performance!
    • TREC collections;
    • Compared with language model and BM25;

A General Formalization

<table>
<thead>
<tr>
<th>Task</th>
<th>$T_1$</th>
<th>$T_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad-hoc Retrieval</td>
<td>Query</td>
<td>Document</td>
</tr>
<tr>
<td>Paraphrase Identification</td>
<td>String A</td>
<td>String B</td>
</tr>
<tr>
<td>Question Answering</td>
<td>Question</td>
<td>Answer</td>
</tr>
<tr>
<td>Automatic Conversation</td>
<td>Dialog</td>
<td>Response</td>
</tr>
</tbody>
</table>
Research Questions

1. Is matching in ad-hoc retrieval really the same as that in NLP tasks?
2. Are the existing deep matching models suitable for the ad-hoc retrieval task?

<table>
<thead>
<tr>
<th>Task</th>
<th>$T_1$</th>
<th>$T_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad-hoc Retrieval</td>
<td>Query</td>
<td>Document</td>
</tr>
<tr>
<td>Paraphrase Identification</td>
<td>String A</td>
<td>String B</td>
</tr>
<tr>
<td>Question Answering</td>
<td>Question</td>
<td>Answer</td>
</tr>
<tr>
<td>Automatic Conversation</td>
<td>Dialog</td>
<td>Response</td>
</tr>
</tbody>
</table>
Outline

• Motivation
• Problem Analysis
• Our Approach
• Experiments
• Conclusions
Problem Analysis

• Matching in many NLP tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>$T_1$</th>
<th>$T_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad-hoc Retrieval</td>
<td>Query</td>
<td>Document</td>
</tr>
<tr>
<td>Paraphrase Identification</td>
<td>String A</td>
<td>String B</td>
</tr>
<tr>
<td>Question Answering</td>
<td>Question</td>
<td>Answer</td>
</tr>
<tr>
<td>Automatic Conversation</td>
<td>Dialog</td>
<td>Response</td>
</tr>
</tbody>
</table>

• **Semantic matching**: identifying the semantic meaning and inferring the semantic relations between two pieces of text
  • Homogeneous texts
  • Consist of a few natural language sentences
Semantic Matching

- **Similarity Matching Signals**
  - Important to capture the semantic similarity/relatedness;

Automatic conversation:

S: Where do you come from?
R: I am from Madrid, Spain.
Semantic Matching

• **Similarity Matching Signals**
  • Important to capture the semantic similarity/relatedness;

• **Compositional meanings**
  • Natural language sentences;
  • Compositional meaning based on their grammatical structures;

Automatic conversation:
S: Where do you come from?
R: I am from Madrid, Spain.
Semantic Matching

• **Similarity Matching Signals**
  • Important to capture the semantic similarity/relatedness;

• **Compositional meanings**
  • Natural language sentences;
  • Compositional meaning based on their grammatical structures;

• **Global matching requirement**
  • Limited lengths and concentrated topic scope;
  • Treat the text as a whole to infer the semantic relations

**Automatic conversation:**
S: Where do you come from?
R: I am from Madrid, Spain.
R: I am a student.

S: Where do you come from?
R: I am from Madrid, Spain.

S: Where do you come from?
R: I am a student.
Problem Analysis

- Matching in ad-hoc retrieval

<table>
<thead>
<tr>
<th>Task</th>
<th>$T_1$</th>
<th>$T_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad-hoc Retrieval</td>
<td>Query</td>
<td>Document</td>
</tr>
<tr>
<td>Paraphrase Identification</td>
<td>String A</td>
<td>String B</td>
</tr>
<tr>
<td>Question Answering</td>
<td>Question</td>
<td>Answer</td>
</tr>
<tr>
<td>Automatic Conversation</td>
<td>Dialog</td>
<td>Response</td>
</tr>
</tbody>
</table>

- **Relevance matching**: identifying whether a document is relevant to a given query.
  - The query is typically short and keyword based;
  - The document can vary considerably in length, from tens of words to thousands or even tens of thousands of words.
Relevance Matching

• **Exact matching signals**
  • The exact matching of terms is still the most important signal in ad-hoc retrieval;
  • Indexing and search paradigm in modern search engines;
Relevance Matching

• **Exact matching signals**
  • The exact matching of terms is still the most important signal in ad-hoc retrieval;
  • Indexing and search paradigm in modern search engines;

• **Query term importance**
  • Query: mainly short and keyword based without complex grammar
  • Critical to take into account term importance
Relevance Matching

• **Exact matching signals**
  - The exact matching of terms is still the most important signal in ad-hoc retrieval;
  - Indexing and search paradigm in modern search engines;

• **Query term importance**
  - Query: mainly short and keyword based without complex grammar
  - Critical to take into account term importance

• **Diverse matching requirement**
  - Long document: Different hypotheses concerning document length in ad-hoc retrieval

1. **Verbosity Hypothesis**
   - Global relevance

2. **Scope Hypothesis**
   - Relevance matching could happen in any part of a document
Semantic Matching vs. Relevance Matching

• There are large differences between relevance matching in ad-hoc retrieval and semantic matching in NLP tasks
  • Affect the design of deep model architectures
  • No “one-fit-all” matching models

<table>
<thead>
<tr>
<th>Semantic Matching</th>
<th>Relevance Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Similarity Matching Signals</td>
<td>• Exact matching signals</td>
</tr>
<tr>
<td>• Compositional meanings</td>
<td>• Query term importance</td>
</tr>
<tr>
<td>• Global matching requirement</td>
<td>• Diverse matching requirement</td>
</tr>
</tbody>
</table>
Re-Visit Deep Matching Models

- Most existing models are about **semantic matching** rather than **relevance matching**.
  - **Representation-focused models**:  
    - Focus on the compositional meaning of the texts;
    - Fit the global matching requirement;
    - Exact matching signals are lost.

![Diagram of Deep Matching Models]

- DSSM [Huang et al. 2013]
- C-DSSM [Shen et al. 2014]
- ARC-I [Hu et al. 2014]
- ...
Re-Visit Deep Matching Models

• Most existing models are about **semantic matching** rather than **relevance matching**.
  • **Interaction-focused models:**
    • Preserve detailed matching signals but do not differentiate exact and similarity matching signals in model;
    • Do not address the query term importance;
    • Cannot meet the diverse matching requirement;

DeepMatch [Lu et al. 2014]
ARC-II [Hu et al. 2014]
MatchPyramid [Pang et al. 2016]
...
Outline

• Motivation
• Problem Analysis
• Our Approach
• Experiments
• Conclusions
A Deep Relevance Matching Model

• Specifically designed for relevance matching in ad-hoc retrieval by explicitly addressing the three factors
  • Exact matching signals
  • Query term importance
  • Diverse matching requirement

• Our model employs a joint deep architecture at the query term level over the local interactions between query words and document words
  • Similar to interaction-focused models rather than representation-focused models
Model Architecture

Matching Score

Score Aggregation

Feed Forward Matching Network

Matching Histogram Mapping

Local Interaction

Term Gating Network

Matching Score

Matching Network

Model Architecture
Model Architecture

Score Aggregation

① Matching Histogram Mapping

② Feed Forward Matching Network

③ Term Gating Network

Matching Score

Local Interaction

q

d
1. Matching Histogram Mapping

- Matching histogram: map the varied-size interactions into a fixed-length representation
  - Groups local interactions according to different strength levels
  - position-free but strength-focused representation

\[
z_i^{(0)} = h\left( w_i^{(q)} \otimes d \right), \quad i = 1, \ldots, M
\]

- Different mappings \( h() \):
  - Count-based histogram: frequency
  - Normalized histogram: relative frequency
  - LogCount-based histogram: logarithm
2. Feed forward Matching Network

- FFMN
  - Extract hierarchical matching patterns from different levels of interaction signals

\[
\begin{align*}
  z_i^{(0)} &= h \left( w_i^{(q)} \otimes d \right), & i &= 1, \ldots, M \\
  z_i^{(l)} &= \tanh \left( W^{(l)} z_i^{(l-1)} + b^{(l)} \right), & i &= 1, \ldots, M, l = 1, \ldots, L
\end{align*}
\]
3. Term Gating Network

- Control how much relevance score on each query term contribute to the final relevance score

\[ s = \sum_{i=1}^{M} g_i z_i^{(L)} \quad g_i = \frac{\exp(w_g x_i^{(q)})}{\sum_{j=1}^{M} \exp(w_g x_j^{(q)})}, \quad i = 1, \ldots, M \]

- Input:
  - Term vector [Zheng et al. SIGIR2015]
  - Inverse document frequency

![Diagram of Term Gating Network]

Matching Score

Score Aggregation

Term Gating Network
Model Discussion

Existing Interaction-focused models

Matching matrix
- position preserving
- zero-padding
- signals are equal

Our Model

Matching histogram
- strength preserving
- no need for padding
- distinguish exact/similarity signals
Model Discussion

- Existing Interaction-focused models: CNNs base on matching matrix
  - Learn positional regularities in matching patterns
    - Suitable for image recognition and global matching requirement (i.e., all the positions are important)
    - Not suitable for diverse matching requirement (i.e., no positional regularity)

- Our method: DNN on matching histogram
  - Learn position-free but strength-focused patterns
  - Explicitly model term importance
Model Training

• As a ranking task, we employ pairwise ranking loss such as hinge loss for training our model

\[ \mathcal{L}(q, d^+, d^-; \Theta) = \max(0, 1 - s(q, d^+) + s(q, d^-)) \]

• Back-propagation
• SGD with AdaGrad
• Mini-batch (20 in size)
• Early stopping
Outline

• Motivation
• Problem Analysis
• Our Approach
• Experiments
• Conclusions
Experimental Settings

• Dataset:
  • Robust04: news collection
  • ClueWeb09-Cat-B: Web collection

• Pre-processing:
  • Word stemming using Krovetz stemmer
  • Stop word remove in queries using INQUERY stop list

• Evaluation Methodology:
  • 5-fold cross validation
  • Tuned towards MAP
  • Evaluated by MAP, nDCG@20, P@20

<table>
<thead>
<tr>
<th></th>
<th>Robust04</th>
<th>ClueWeb09-Cat-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
<td>0.6M</td>
<td>38M</td>
</tr>
<tr>
<td>Document Count</td>
<td>0.5M</td>
<td>34M</td>
</tr>
<tr>
<td>Collection Length</td>
<td>252M</td>
<td>26B</td>
</tr>
<tr>
<td>Query Count</td>
<td>250</td>
<td>150</td>
</tr>
</tbody>
</table>

The ClueWeb-09-Cat-B collection has been filtered to the set of documents in the 60th percentile of spam scores.
Baseline Methods

• Traditional retrieval models
  • QL: Query likelihood model based on Dirichlet smoothing
  • BM25: The classical probabilistic retrieval model
Baseline Methods

• Traditional retrieval models
  • **QL**: Query likelihood model based on Dirichlet smoothing
  • **BM25**: The classical probabilistic retrieval model

• Representation-focused deep matching models
  • **DSSM<sub>T</sub>/DSSM<sub>D</sub>**: A state-of-the-art deep matching model for Web search using feed forward neural network *(released model trained on large clickthrough data)*
  • **C-DSSM<sub>T</sub>/C-DSSM<sub>D</sub>**: A similar deep matching model to DSSM for Web search, replacing DNN with CNN *(released model trained on large clickthrough data)*
  • **ARC-I**: A general representation-focused deep matching model based on CNN *(Implement according to the original paper)*
Baseline Methods

- Traditional retrieval models
  - **QL**: Query likelihood model based on Dirichlet smoothing
  - **BM25**: The classical probabilistic retrieval model

- Representation-focused deep matching models
  - **DSSM_T/DSSM_D**: A state-of-the-art deep matching model for Web search using feedforward neural network *(released model trained on large clickthrough data)*
  - **C-DSSM_T/C-DSSM_D**: A similar deep matching model to DSSM for Web search, replacing DNN with CNN *(released model trained on large clickthrough data)*
  - **ARC-I**: A general representation-focused deep matching model based on CNN *(Implement according to the original paper)*

- Interaction-focused deep matching models
  - **ARC-II**: A hierarchical model over local interactions using CNN *(Implement according to the original paper)*
  - **MP**: MatchPyramid is another state-of-the-art interaction-focused deep matching model. There are three variants of the model based on different interaction operators, denoted as **MP_{IND}, MP_{COS}, and MP_{DOT}** *(Released codes)*
Baseline Methods

• Traditional retrieval models
  • QL: Query likelihood model based on Dirichlet smoothing
  • BM25: The classical probabilistic retrieval model

• Representation-focused deep matching models
  • DSSM$_T$/DSSM$_D$: A state-of-the-art deep matching model for Web search using feed forward neural network (released model trained on large clickthrough data)
  • C-DSSM$_T$/C-DSSM$_D$: A similar deep matching model to DSSM for Web search, replacing DNN with CNN (released model trained on large clickthrough data)
  • ARC-I: A general representation-focused deep matching model based on CNN (Implement according to the original paper)

• Interaction-focused deep matching models
  • ARC-II: A hierarchical model over local interactions using CNN (Implement according to the original paper)
  • MP: MatchPyramid is another state-of-the-art interaction-focused deep matching model. There are three variants of the model based on different interaction operators, denoted as MP$_{IND}$, MP$_{COS}$, and MP$_{DOT}$ (Released codes)

• Our Approach: DRMM (6 variants)
  • DRMM$_{CHxIDF}$ we refer to DRMM with Count-based histogram and term gating network using inverse document frequency
Other Configurations

• **Term Embeddings**
  - For all deep models, we used 300-dimensional term vectors trained with the CBOW [T. Mikolov et al., NIPS2013] on the Robust04 and ClueWeb-09-Cat-B collections, respectively.
  - Context window size: 10 Negative samples: 10
  - Preprocess: HTML tags removed and stemming
  - Vocabulary Size: Robust04 (0.1M), ClueWeb-09-Cat-B(4.1M)
Other Configurations

• **Term Embeddings**
  - For all deep models, we used 300-dimensional term vectors trained with the CBOW [T. Mikolov et al., NIPS2013] on the Robust04 and ClueWeb-09-Cat-B collections, respectively.
  - Context window size: 10
  - Negative samples: 10
  - Preprocess: HTML tags removed and stemming
  - Vocabulary Size: Robust04 (0.1M), ClueWeb-09-Cat-B(4.1M)

• **Network Configurations**
  - ARC-I, ARC-II and MatchPyramid
    - Tried both the default configurations and other settings
    - ARC-I and ARC-II: 3-word windows, 64 feature maps and 6 layers
    - MatchPyramid: 3x3 kernel size, 8 feature maps, and 4 layers
  - DRMM
    - 30-bin matching histogram. 4 layers
## Retrieval Performance

### Robust-04 collection

Significant improvement or degradation with respect to QL is indicated (+/-)

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Model Name</th>
<th>Topic Titles</th>
<th>Topic Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAP</td>
<td>nDCG@20</td>
</tr>
<tr>
<td>Traditional Retrieval Baslines</td>
<td>QL</td>
<td>0.253</td>
<td>0.415</td>
</tr>
<tr>
<td></td>
<td>BM25</td>
<td>0.255</td>
<td>0.418</td>
</tr>
<tr>
<td>Representation-Focused Matching Baslines</td>
<td>DSSM(_D)</td>
<td>0.095(^-)</td>
<td>0.201(^-)</td>
</tr>
<tr>
<td></td>
<td>CDSSM(_D)</td>
<td>0.067(^-)</td>
<td>0.146(^-)</td>
</tr>
<tr>
<td></td>
<td>ARC-I</td>
<td>0.041(^-)</td>
<td>0.066(^-)</td>
</tr>
</tbody>
</table>

All the representation-focused models perform significantly worse than the traditional retrieval models
## Retrieval Performance

**Robust-04 collection**

Significant improvement or degradation with respect to QL is indicated (+/-)

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Model Name</th>
<th>Topic Titles</th>
<th>Topic Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAP</td>
<td>nDCG@20</td>
</tr>
<tr>
<td>Traditional Retrieval Baselines</td>
<td>QL</td>
<td>0.253</td>
<td>0.415</td>
</tr>
<tr>
<td></td>
<td>BM25</td>
<td>0.255</td>
<td>0.418</td>
</tr>
<tr>
<td>Representation-Focused Matching Baselines</td>
<td>DSSM(_D)</td>
<td>0.095</td>
<td>0.201</td>
</tr>
<tr>
<td></td>
<td>CDSSM(_D)</td>
<td>0.067</td>
<td>0.146</td>
</tr>
<tr>
<td></td>
<td>ARC-I</td>
<td>0.041</td>
<td>0.066</td>
</tr>
<tr>
<td>Interaction-Focused Matching Baselines</td>
<td>ARC-II</td>
<td>0.067</td>
<td>0.147</td>
</tr>
<tr>
<td></td>
<td>MP(_{IND})</td>
<td>0.169</td>
<td>0.319</td>
</tr>
<tr>
<td></td>
<td>MP(_{COS})</td>
<td>0.189</td>
<td>0.330</td>
</tr>
<tr>
<td></td>
<td>MP(_{DOT})</td>
<td>0.083</td>
<td>0.159</td>
</tr>
</tbody>
</table>

1. Interaction-focused models cannot compete with the traditional retrieval models either
2. By preserving detailed matching signals, interaction-focused models can work better than representation-focused matching models
## Retrieval Performance

### Robust-04 collection

Significant improvement or degradation with respect to QL is indicated (+/-)

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Model Name</th>
<th>Topic Titles</th>
<th>Topic Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAP</td>
<td>nDCG@20</td>
</tr>
<tr>
<td><strong>Traditional Retrieval Baselines</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QL</td>
<td>0.253</td>
<td>0.415</td>
<td>0.369</td>
</tr>
<tr>
<td>BM25</td>
<td>0.255</td>
<td>0.418</td>
<td>0.370</td>
</tr>
<tr>
<td><strong>Representation-Focused Matching Baselines</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSSM&lt;sub&gt;D&lt;/sub&gt;</td>
<td>0.095&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.201&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.171&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>CDSSM&lt;sub&gt;D&lt;/sub&gt;</td>
<td>0.067&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.146&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.125&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>ARC-I</td>
<td>0.041&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.066&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.065&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Interaction-Focused Matching Baselines</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARC-II</td>
<td>0.067&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.147&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.128&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>MP&lt;sub&gt;IND&lt;/sub&gt;</td>
<td>0.169&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.319&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.281&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>MP&lt;sub&gt;COS&lt;/sub&gt;</td>
<td>0.189&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.330&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.290&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>MP&lt;sub&gt;DOT&lt;/sub&gt;</td>
<td>0.083&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.159&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.155&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Our Approach</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRMM&lt;sub&gt;CHXTV&lt;/sub&gt;</td>
<td>0.253</td>
<td>0.407</td>
<td>0.357</td>
</tr>
<tr>
<td>DRMM&lt;sub&gt;NHXTV&lt;/sub&gt;</td>
<td>0.160&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.293&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.258&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>DRMM&lt;sub&gt;LCHXTV&lt;/sub&gt;</td>
<td>0.268&lt;sup&gt;+&lt;/sup&gt;</td>
<td>0.423</td>
<td>0.381</td>
</tr>
<tr>
<td>DRMM&lt;sub&gt;CHXIDF&lt;/sub&gt;</td>
<td>0.259</td>
<td>0.412</td>
<td>0.362</td>
</tr>
<tr>
<td>DRMM&lt;sub&gt;NHXIDF&lt;/sub&gt;</td>
<td>0.187&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.312&lt;sup&gt;-&lt;/sup&gt;</td>
<td>0.282&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>DRMM&lt;sub&gt;LCHXIDF&lt;/sub&gt;</td>
<td>0.279&lt;sup&gt;+&lt;/sup&gt;</td>
<td>0.431&lt;sup&gt;+&lt;/sup&gt;</td>
<td>0.382&lt;sup&gt;+&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

1. The best performing DRMM is significantly better than all the existing baselines
2. The performance on topic descriptions can be comparable to that on topic titles
1. LCH-based histogram > CH-based histogram > NH-based histogram
   - CH-based > NH-based: Document length information is important in ad-hoc retrieval
   - LCH-based best: input signals with reduced range, and non-linear transformation useful for learning multiplicative relationships

Retrieval Performance

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Model Name</th>
<th>Topic Titles</th>
<th>Topic Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAP</td>
<td>nDCG@20</td>
</tr>
<tr>
<td>Our Approach</td>
<td>DRMM_{CHXTV}</td>
<td>0.253</td>
<td>0.407</td>
</tr>
<tr>
<td></td>
<td>DRMM_{NHXTV}</td>
<td>0.160</td>
<td>0.293</td>
</tr>
<tr>
<td></td>
<td>DRMM_{LCHXTV}</td>
<td>0.268</td>
<td>0.423</td>
</tr>
<tr>
<td></td>
<td>DRMM_{CHXIDF}</td>
<td>0.259</td>
<td>0.412</td>
</tr>
<tr>
<td></td>
<td>DRMM_{NHXIDF}</td>
<td>0.187</td>
<td>0.312</td>
</tr>
<tr>
<td></td>
<td>DRMM_{LCHXIDF}</td>
<td>0.279</td>
<td>0.431</td>
</tr>
</tbody>
</table>

Robust-04 collection

Significant improvement or degradation with respect to QL is indicated (+/-)
1. LCH-based histogram > CH-based histogram > NH-based histogram
   • CH-based > NH-based: Document length information is important in ad-hoc retrieval
   • LCH-based best: input signals with reduced range, and non-linear transformation useful for learning multiplicative relationships

2. IDF-based Term Gating > Term vector-based Gating
   • Term vectors do not contain sufficient information
   • Model using term vectors introduces too many parameters to be learned sufficiently

### Retrieval Performance

#### Robust-04 collection

Significant improvement or degradation with respect to QL is indicated (+/-)

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Model Name</th>
<th>Topic Titles</th>
<th></th>
<th>Topic Descriptions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>MAP</td>
<td>nDCG@20</td>
<td>P@20</td>
</tr>
<tr>
<td>Our Approach</td>
<td>DRMM\textsubscript{CHXTV}</td>
<td>0.253</td>
<td>0.407</td>
<td>0.357</td>
<td>0.247</td>
</tr>
<tr>
<td></td>
<td>DRMM\textsubscript{NHXTV}</td>
<td>0.160</td>
<td>0.293</td>
<td>0.258</td>
<td>0.132</td>
</tr>
<tr>
<td></td>
<td>DRMM\textsubscript{LCHXTV}</td>
<td>0.268</td>
<td>0.423</td>
<td>0.381</td>
<td>0.265</td>
</tr>
<tr>
<td></td>
<td>DRMM\textsubscript{CHXIDF}</td>
<td>0.259</td>
<td>0.412</td>
<td>0.362</td>
<td>0.255</td>
</tr>
<tr>
<td></td>
<td>DRMM\textsubscript{NHXIDF}</td>
<td>0.187</td>
<td>0.312</td>
<td>0.282</td>
<td>0.145</td>
</tr>
<tr>
<td></td>
<td>DRMM\textsubscript{LCHXIDF}</td>
<td>0.279</td>
<td>0.431</td>
<td>0.382</td>
<td>0.275</td>
</tr>
</tbody>
</table>
## Retrieval Performance

**ClueWeb-09-Cat-B collection**

Significant improvement or degradation with respect to QL is indicated (+/-)

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Model Name</th>
<th>Topic Titles</th>
<th>Topic Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAP</td>
<td>nDCG@20</td>
</tr>
<tr>
<td>Traditional Retrieval Baselines</td>
<td>QL</td>
<td>0.100</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>BM25</td>
<td>0.101</td>
<td>0.225</td>
</tr>
<tr>
<td>Representation-Focused Matching Baselines</td>
<td>DSSMₜ</td>
<td>0.054−</td>
<td>0.132−</td>
</tr>
<tr>
<td></td>
<td>DSSMₐ</td>
<td>0.039−</td>
<td>0.099−</td>
</tr>
<tr>
<td></td>
<td>CDSSMₜ</td>
<td>0.064−</td>
<td>0.253−</td>
</tr>
<tr>
<td></td>
<td>CDSSMₐ</td>
<td>0.054−</td>
<td>0.134−</td>
</tr>
<tr>
<td></td>
<td>ARC-I</td>
<td>0.024−</td>
<td>0.073−</td>
</tr>
<tr>
<td>Interaction-Focused Matching Baselines</td>
<td>ARC-II</td>
<td>0.033−</td>
<td>0.087−</td>
</tr>
<tr>
<td></td>
<td>MPₐIND</td>
<td>0.056−</td>
<td>0.139−</td>
</tr>
<tr>
<td></td>
<td>MPₐCOS</td>
<td>0.066−</td>
<td>0.158−</td>
</tr>
<tr>
<td></td>
<td>MPₐDOT</td>
<td>0.044−</td>
<td>0.109−</td>
</tr>
<tr>
<td>Our Approach</td>
<td>DRMMₜCHXTV</td>
<td>0.103</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td>DRMMₐNHXTV</td>
<td>0.065−</td>
<td>0.151−</td>
</tr>
<tr>
<td></td>
<td>DRMMₜLCHXTV</td>
<td>0.111+</td>
<td>0.250+</td>
</tr>
<tr>
<td></td>
<td>DRMMₜCHXIDF</td>
<td>0.104</td>
<td>0.252+</td>
</tr>
<tr>
<td></td>
<td>DRMMₐNHXIDF</td>
<td>0.066−</td>
<td>0.151−</td>
</tr>
<tr>
<td></td>
<td>DRMMₜLCHXIDF</td>
<td>0.113+</td>
<td>0.258+</td>
</tr>
</tbody>
</table>

Similar results can be found on ClueWeb-09-Cat-B collection
Impact of Different Model Components

Comparison of several simpler versions of DRMM over topic titles of the two test collections in terms of MAP

- DRMLCHxIDF
- DRMLCHxUNI
- DRMMxIDF
- DRMMKMAXxIDF

Original DRMM

remove term gating network (no term importance)
Impact of Different Model Components

Comparison of several simpler versions of DRMM over topic titles of the two test collections in terms of MAP

- DRMMLCHxIDF
- DRMMLCHxUNI
- DRMMMDYNxIDF
- DRMMKMAXxIDF

Original DRMM

replace matching histogram with matching sequence, using dynamic pooling strategy (position-related input)
Impact of Different Model Components

Comparison of several simpler versions of DRMM over topic titles of the two test collections in terms of MAP

- DRMLCHxIDF
- DRMLCHxUNI
- DRMMDYNxIDF
- DRMKMAXxIDF

Original DRMM

replace matching histogram with matching sequence, using k-max pooling strategy (top K strength related input)
Impact of Term Embeddings

Performance comparison of DRMM over different dimensionality of term embeddings trained by CBOW on the Robust04 collection

<table>
<thead>
<tr>
<th>Topic</th>
<th>Embedding</th>
<th>MAP</th>
<th>nDCG@20</th>
<th>P@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titles</td>
<td>CBOW-50d</td>
<td>0.268</td>
<td>0.421</td>
<td>0.375</td>
</tr>
<tr>
<td></td>
<td>CBOW-100d</td>
<td>0.270</td>
<td>0.427</td>
<td>0.379</td>
</tr>
<tr>
<td></td>
<td>CBOW-300d</td>
<td>0.279</td>
<td>0.431</td>
<td>0.381</td>
</tr>
<tr>
<td></td>
<td>CBOW-500d</td>
<td>0.277</td>
<td>0.430</td>
<td>0.381</td>
</tr>
<tr>
<td>Descriptions</td>
<td>CBOW-50d</td>
<td>0.268</td>
<td>0.431</td>
<td>0.365</td>
</tr>
<tr>
<td></td>
<td>CBOW-100d</td>
<td>0.271</td>
<td>0.433</td>
<td>0.367</td>
</tr>
<tr>
<td></td>
<td>CBOW-300d</td>
<td>0.275</td>
<td>0.437</td>
<td>0.371</td>
</tr>
<tr>
<td></td>
<td>CBOW-500d</td>
<td>0.274</td>
<td>0.435</td>
<td>0.370</td>
</tr>
</tbody>
</table>

1. With lower dimensionality, the similarity between term embeddings might be coarse and hurt the relevance matching performance.
2. With larger dimensionality, one may need more data to train reliable term embeddings.
Outline

• Introduction
• Problem Analysis
• Our Approach
• Experiments
• Conclusions
Major Contributions

• We point out three major differences between semantic matching and relevance matching, which may lead to significantly different architecture design of the deep matching models.

• We proposed a novel deep relevance matching model for ad-hoc retrieval by explicitly addressing the three key factors of relevance matching.

• We conduct rigorous comparisons over state-of-the-art retrieval models to demonstrate the effectiveness of our model.
Future Work

• Leverage larger training data to train deeper DRMM to further explore the potential of the proposed model
  • e.g. click-through logs

• Include phrase embeddings so that phrases can be treated as a whole rather than separate terms
  • local interactions can better reflect the meaning by using the proper semantic units in language
Thanks!

Email: guojiafeng@ict.ac.cn
Data and Codes can be found at
http://www.bigdatalab.ac.cn/benchmark/