Multi-Perspective Similarity Modeling with Hierarchical ConvNets for Social Media Search

Wei (Victor) Yang

March 28, 2018

Overview

- Task: Social Media Search
- Proposed Approach
- Experiments
- Summary

Overview

Social Media Search

- **Shorter length**: Twitter posts are limited to 140 characters.

Social Media Search

- **Shorter length**: Twitter posts are limited to 140 characters.
- **Informality**: Abbreviations, misspellings, typos and emojis

Social Media Search

- **Shorter length**: Twitter posts are limited to 140 characters.
- **Informality**: Abbreviations, misspellings, typos and emojis
- **Heterogeneous relevance signals**
    - Many real-world news and events

Social Media Search

- **Shorter length**: Twitter posts are limited to 140 characters.
- **Informality**: Abbreviations, misspellings, typos and emojis
- **Heterogeneous relevance signals**
  - Many real-world news and events
  - URLs or hastags

Social Media Search

- **Shorter length**: Twitter posts are limited to 140 characters.
- **Informality**: Abbreviations, misspellings, typos and emojis
- **Heterogeneous relevance signals**
  - Many real-world news and events
  - URLs or hastags
  - Retweets

Related Work

- **Traditional non-neural methods for Ad-hoc Retrieval**
  - Probabilistic models: BM25, query likelihood and RM3
  - Exact term match

Related Work

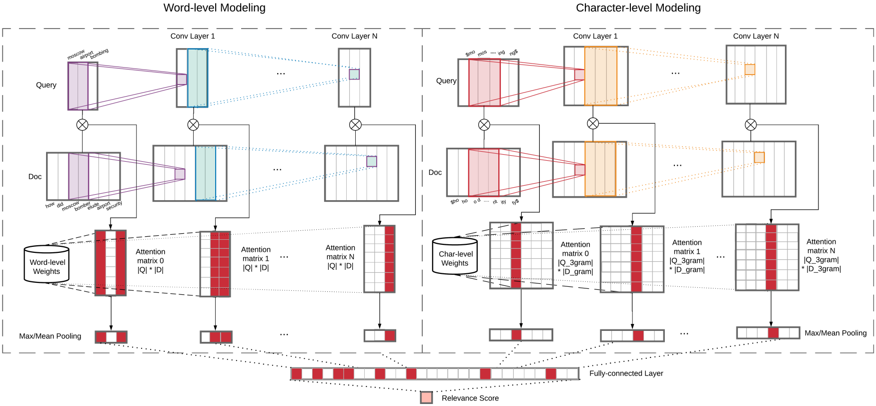
- **Traditional non-neural methods for Ad-hoc Retrieval**
  - Probabilistic models: BM25, query likelihood and RM3
  - Exact term match
- **Domain-specific Matching Pattern Design**
  - Li et al. (2003) tried to understand the temporal behavior of a query on the web search logs dataset.
  - Lee et al. (2015) concentrated on medical terminologies for medical documents.

Related Work

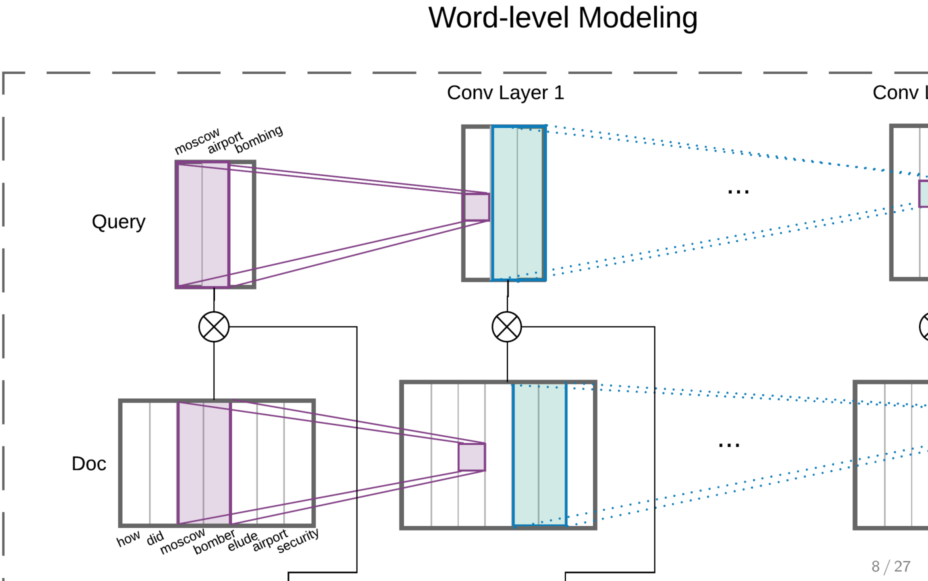
- **Traditional non-neural methods for Ad-hoc Retrieval**
  - Probabilistic models: BM25, query likelihood and RM3
  - Exact term match
- **Domain-specific Matching Pattern Design**
  - Li et al. (2003) tried to understand the temporal behavior of a query on the web search logs dataset.
  - Lee et al. (2015) concentrated on medical terminologies for medical documents.
- **Word Embedding and Character Embedding in NN**
  - It has shown an increasing trend to use both word embeddings and character embeddings in neural models.

Overview

Model Framework



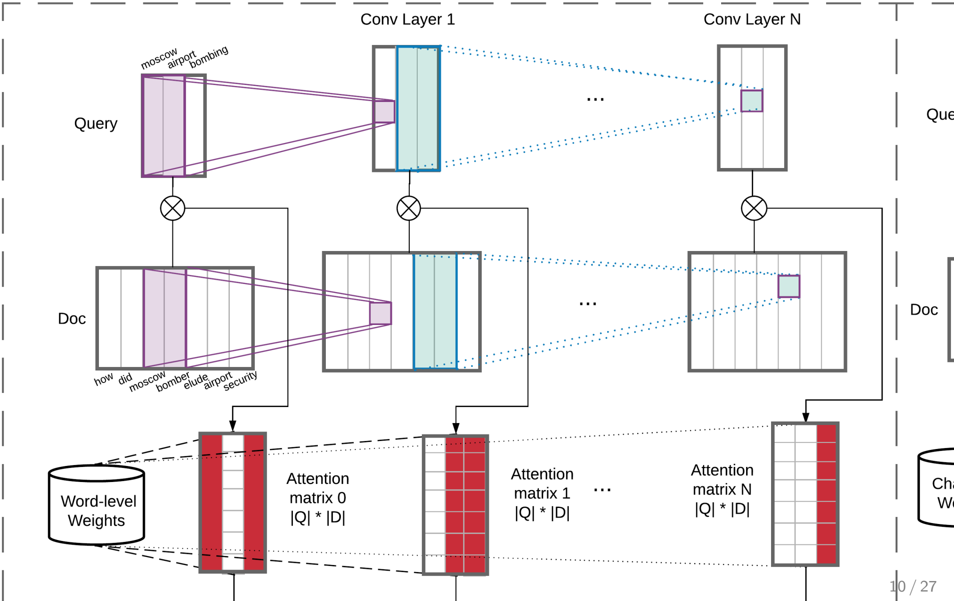
A. Hierarchical ConvNets



A. Hierarchical ConvNets

- CNN v.s. RNN
  - Multiple consecutive convolution layers can extract phrase-level semantic representation
  - CNN based network often converge faster than LSTM or GRU based network

Word-level Modeling



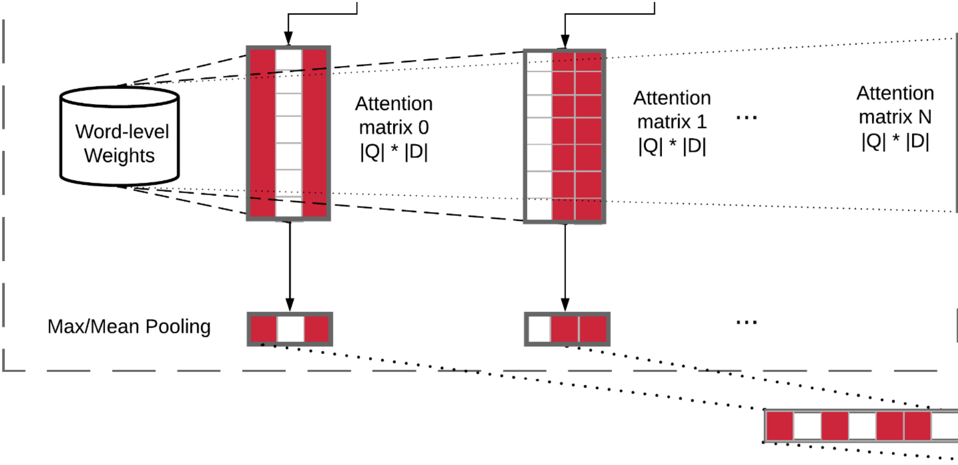
A. Hierarchical ConvNets

- CNN v.s. RNN
  - Multiple consecutive convolution layers can extract phrase-level semantic representation
  - CNN based network often converge faster than LSTM or GRU based network
- Similarity Matrix (or Attention Matrix)
  - Pair-wise dot product of query terms and tweet terms

$$S_i = softmax(Q_i \otimes D_i) \tag{1}$$

$$softmax(X) = [\frac{\mathbf{e^{x_1}}}{\sum_i e^{x_{1i}}}, \frac{\mathbf{e^{x_2}}}{\sum_i e^{x_{2i}}}, ..., \frac{\mathbf{e^{x_{l_q}}}}{\sum_i e^{x_{l_q i}}}] \tag{2}$$

B. From Attention Matrix to Relevance Score



B. From Attention Matrix to Relevance Score

- Max/Mean Pooling

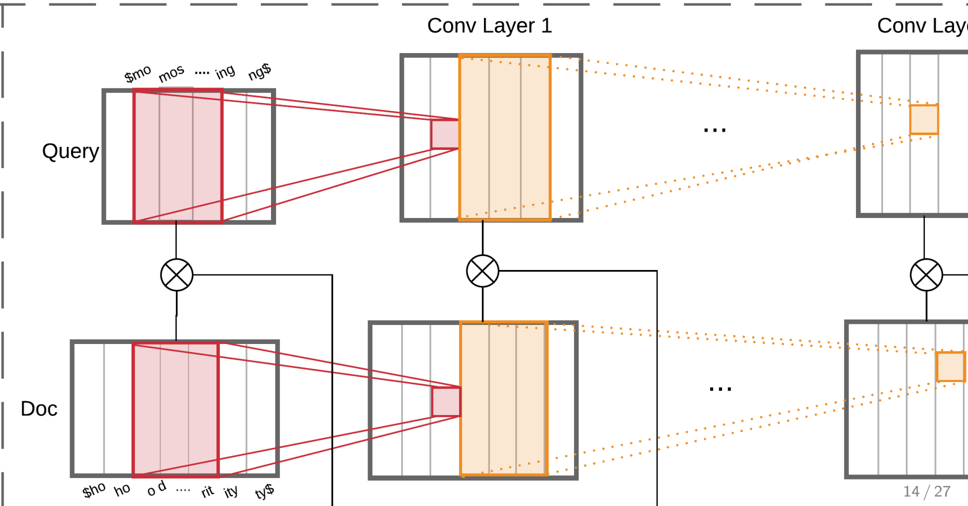
$$\begin{aligned} Max(S) &= [\max(\mathbf{s_1}), \max(\mathbf{s_2}), ..., \max(\mathbf{s_{l_q}})] \\ Mean(S) &= [mean(\mathbf{s_1}), mean(\mathbf{s_2}), ..., mean(\mathbf{s_{l_q}})] \end{aligned} \tag{3}$$

- IDF weighting

$$Score(Q, D) = \sigma(W \cdot \mathbf{IDF\_weights} \cdot \mathbf{Pooled\_features} + b) \tag{4}$$

C. *k*-gram at Character Level

Character-level Modeling



C. *k*-gram at Character Level

- It allows us to match most OOV words and rare words
    1. **Loanword and non-English words**: emociones (Spainish, emotions), desgostosa (Portuguese, disgusted), insomma (Italian, for heaven's sake), hayatm (Turkish, sweetheart)

C. *k*-gram at Character Level

- It allows us to match most OOV words and rare words
  1. **Loanword and non-English words**: emociones (Spainish, emotions), desgostosa (Portuguese, disgusted), insomma (Italian, for heaven's sake), hayatm (Turkish, sweetheart)
  2. **Portmanteaus and compounds**: netizen, candieapple, freshideas, actor-director, scranton-based, early-stage

C. *k*-gram at Character Level

- It allows us to match most OOV words and rare words
  1. **Loanword and non-English words**: emociones (Spainish, emotions), desgostosa (Portuguese, disgusted), insomma (Italian, for heaven's sake), hayatm (Turkish, sweetheart)
  2. **Portmanteaus and compounds**: netizen, candieapple, freshideas, actor-director, scranton-based, early-stage
  3. **Abbreviations**: CHTE (Center for Heat Treating Excellence), b-day (birthday)

C. *k*-gram at Character Level

- It allows us to match most OOV words and rare words
  1. **Loanword and non-English words**: emociones (Spainish, emotions), desgostosa (Portuguese, disgusted), insomma (Italian, for heaven's sake), hayatm (Turkish, sweetheart)
  2. **Portmanteaus and compounds**: netizen, candieapple, freshideas, actor-director, scranton-based, early-stage
  3. **Abbreviations**: CHTE (Center for Heat Treating Excellence), b-day (birthday)
  4. **URLs, hashtags, and mentions**: #nokia, @barackobama
     - http://www.huffingtonpost.com/2013/03/22/world-water-day-2013-facts_n_2927389.html

C. *k*-gram at Character Level

- It allows us to match most OOV words and rare words
  1. **Loanword and non-English words**: emociones (Spainish, emotions), desgostosa (Portuguese, disgusted), insomma (Italian, for heaven's sake), hayatm (Turkish, sweetheart)
  2. **Portmanteaus and compounds**: netizen, candieapple, freshideas, actor-director, scranton-based, early-stage
  3. **Abbreviations**: CHTE (Center for Heat Treating Excellence), b-day (birthday)
  4. **URLs, hashtags, and mentions**: #nokia, @barackobama
     - http://www.huffingtonpost.com/2013/03/22/world-water-day-2013-facts_n_2927389.html
     - However, in previous social media search models, the URLs are often simply dropped or transformed into binary feature.

C. *k*-gram at Character Level

- It allows us to leverage the information of most OOV words
  1. **Loanword and non-English words**: emociones (Spainish, emotions), hayatm (Turkish, sweetheart)
  2. **Portmanteaus and compounds**: netizen, candieapple, freshideas, actor-director, scranton-based, early-stage
  3. **Abbreviations**: CHTE (Center for Heat Treating Excellence), b-day (birthday)
  4. **URLs, hashtags, and mentions**: #nokia, @barackobama
  5. **Typos**: begngen (beggen), yawnn (yawn), chomeos (Chrome OS), tansport (transport), afternoo (afternoon), foreverrrr (forever), yupppppppp, woooooow

C. *k*-gram at Character Level

- It allows us to leverage the information of most OOV words
  1. **Loanword and non-English words**: emociones (Spainish, emotions), hayatm (Turkish, sweetheart)
  2. **Portmanteaus and compounds**: netizen, candieapple, freshideas, actor-director, scranton-based, early-stage
  3. **Abbreviations**: CHTE (Center for Heat Treating Excellence), b-day (birthday)
  4. **URLs, hashtags, and mentions**: #nokia, @barackobama
  5. **Typos**: begngen (beggen), yawnn (yawn), chomeos (Chrome OS), tansport (transport), afternoo (afternoon), foreverrrr (forever), yupppppppp, woooooow
  6. **Domain-specific words**: utf-8, vlookup

Overview

Experimental Setup

- Dataset: TREC Microblog Tracks in 2011-2014

*: after tokenization

Test Set	2011	2012	2013	2014
# of query topics	49	60	60	55
# of query-doc pairs	39,780	49,879	46,192	41,579
# of relevant docs	1,940	4,298	3,405	6,812
# of unique words*	21,649	27,470	24,546	22,099
# of unique OOV words*	13,067	17,190	15,724	14,331
# of URLs	20,351	25,405	23,100	20,885

Experimental Setup

- Dataset: TREC Microblog Tracks in 2011-2014
- Word Vectors

Experimental Setup

- Dataset: TREC Microblog Tracks in 2011-2014
- Word Vectors
  - GloVe 300-dimension word vectors pre-trained from Common Crawl Dataset with 840B tokens

Experimental Setup

- Dataset: TREC Microblog Tracks in 2011-2014
- Word Vectors
  - GloVe 300-dimension word vectors pre-trained from Common Crawl Dataset with 840B tokens
  - OOV word vectors and trigram vectors are uniformly sampled between $[-0.05, 0.05]$.

Experimental Setup

- Dataset: TREC Microblog Tracks in 2011-2014
- Word Vectors
  - GloVe 300-dimension word vectors pre-trained from Common Crawl Dataset with 840B tokens
  - OOV word vectors and trigram vectors are uniformly sampled between $[-0.05, 0.05]$.
- Length Padding

Experimental Setup

- Dataset: TREC Microblog Tracks in 2011-2014
- Word Vectors
  - GloVe 300-dimension word vectors pre-trained from Common Crawl Dataset with 840B tokens
  - OOV word vectors and trigram vectors are uniformly sampled between $[-0.05, 0.05]$.
- Length Padding
  - Each query is padded to 10 words and 51 trigrams
  - Each document is padded to 68 words and 140 trigrams

Experimental Setup

- Dataset: TREC Microblog Tracks in 2011-2014
- Word Vectors
  - GloVe 300-dimension word vectors pre-trained from Common Crawl Dataset with 840B tokens
  - OOV word vectors and trigram vectors are uniformly sampled between $[-0.05, 0.05]$.
- Length Padding
  - Each query is padded to 10 words and 51 trigrams
  - Each document is padded to 68 words and 140 trigrams
- 4-fold Cross-validation: training on three years' datasets and testing on the rest one

Experiment Procedure

1. Applied query likelihood (QL) algorithm[1] implementation provided by the TREC Microblog API[2] to retrieve up to 1000 tweets per topic.

[1] QL is the original and basic method for using language models in IR. It evaluates the relevance score by Bayes rule and the multinomial unigram language model.

[2] https://github.com/lintool/twitter-tools

Experiment Procedure

1. Applied query likelihood (QL) algorithm[1] implementation provided by the TREC Microblog API[2] to retrieve up to 1000 tweets per topic.
2. Rerank the tweet candidates by NN models

[1] QL is the original and basic method for using language models in IR. It evaluates the relevance score by Bayes rule and the multinomial unigram language model.

[2] https://github.com/lintool/twitter-tools

Experiment Procedure

1. Applied query likelihood (QL) algorithm[1] implementation provided by the TREC Microblog API[2] to retrieve up to 1000 tweets per topic.
2. Rerank the tweet candidates by NN models
3. Compute and compare the mean average precision (MAP) and precision at 30 (P@30)

[1] QL is the original and basic method for using language models in IR. It evaluates the relevance score by Bayes rule and the multinomial unigram language model.

[2] https://github.com/lintool/twitter-tools

Baselines

- QL and previous NN models for text ranking
  1. **Query Likelihood**
  2. **DSSM** (2013)
  3. **C-DSSM** (2014)
  4. **MatchPyramid** (2016)
  5. **DRMM** (2016)
  6. **DUET** (2017)
  7. **K-NRM** (2017)

Baselines

- QL and previous NN models for text ranking
  1. **Query Likelihood**
  2. **DSSM** (2013)
  3. **C-DSSM** (2014)
  4. **MatchPyramid** (2016)
  5. **DRMM** (2016)
  6. **DUET** (2017)
  7. **K-NRM** (2017)
- Hybrid Models
  1. **RM3** is a hybrid model combining the QL model and the relevance model with the pseudo relevance feedback technique through the interpolation strategy.
  2. **Hybrid NN models** are the baselines (2)-(7) interpolated with QL.

Experiments 1: Overall Results

Table: MAP

Model	2011	2012	2013	2014
QL	0.3576	0.2091	0.2532	0.3924
DSSM	0.1742	0.1087	0.1434	0.2566
C-DSSM	0.0887	0.0803	0.0892	0.1884
MatchPyramid	0.1967	0.1334	0.1378	0.2722
DRMM	0.2635	0.1777	0.2102	0.3440
DUET	0.1533	0.1325	0.1380	0.2680
K-NRM	0.2519	0.1607	0.1750	0.3472
Our Model	**0.3699**	**0.2313**	**0.2900**	**0.4144**

Table: P@30

Model	2011	2012	2013	2014
QL	0.4000	0.3311	0.4450	0.6182
DSSM	0.2340	0.1791	0.2772	0.4261
C-DSSM	0.1122	0.1525	0.1717	0.2752
MatchPyramid	0.2259	0.2390	0.2561	0.4491
DRMM	0.3095	0.3169	0.4061	0.5424
DUET	0.2109	0.2356	0.2528	0.4091
K-NRM	0.3034	0.2966	0.3178	0.5388
Our Model	**0.4014**	**0.3757**	**0.5161**	**0.6236**

Experiments 2: Results of Hybrid Models

Table: MAP

Model	2011	2012	2013	2014
RM3	0.3824	0.2342	0.2738	**0.4480**
DSSM+	0.3589	0.1777	0.2435	0.3742
C-DSSM+	0.3380	0.2091	0.2532	0.3924
MatchPyramid+	0.3707	0.2170	0.2594	0.3915
DRMM+	0.3477	0.2213	0.2639	0.4042
DUET+	0.3576	0.2243	0.2779	0.4219
K-NRM+	0.3576	0.2277	0.2721	0.4137
Our Model+	**0.4040**	**0.2482**	**0.2937**	0.4412

Table: P@30

Model	2011	2012	2013	2014
RM3	0.4211	0.3452	0.4476	0.6339
DSSM+	0.4143	0.2989	0.4328	0.5885
C-DSSM+	0.3952	0.3311	0.4450	0.6182
MatchPyramid+	0.4190	0.3362	0.4556	0.6085
DRMM	0.4034	0.3537	0.4772	0.6139
DUET+	0.4000	0.3644	0.4878	**0.6467**
K-NRM+	0.4000	0.3520	0.4756	0.6358
Our Model+	**0.4435**	**0.3915**	**0.5250**	0.6345

Experiments 3: Ablation Test

Table: MAP

Model	2011	2012	2013	2014
QL	0.3576⋆	0.2091⋆	0.2532⋆	0.3924⋆
Our Base Model	0.3699	**0.2313**	**0.2900**	**0.4144**
–Doc_URL	0.3610⋆	0.2204⋆	0.2820	0.4052⋆
–Doc_Word	**0.3713**	0.2188⋆	0.2812⋆	0.4012⋆
–IDF weighting	0.3594⋆	0.2174⋆	0.2712⋆	0.4031⋆
–CL	0.3623	0.2181⋆	0.2763⋆	0.4050⋆
–WL	0.1651⋆	0.0762⋆	0.0987⋆	0.1849⋆

Table: P@30

Model	2011	2012	2013	2014
QL Baseline	0.4000	0.3311⋆	0.4450⋆	0.6182⋆
Our Base Model	0.4014	**0.3757**	**0.5161**	0.6236
–Doc_URL	0.4075	0.3458⋆	0.5117	**0.6242**
–Doc_Word	**0.4190***	0.3537⋆	0.5122	0.6121⋆
–IDF weighting	0.3980	0.3582⋆	0.4950⋆	0.6109⋆
–CL	0.3878⋆	0.3412⋆	0.5011⋆	0.6091⋆
–WL	0.1293⋆	0.1119⋆	0.1517⋆	0.2048⋆

Overview

Summary

- We summarize our proposed model as **multi-perspective architecture** in social media search models because it
  - adopts attention matrix based NN architecture, which is widely applied and proved effective in text matching tasks

Summary

- We summarize our proposed model as **multi-perspective architecture** in social media search models because it
  - adopts attention matrix based NN architecture, which is widely applied and proved effective in text matching tasks
  - understands short but informative text by ConvNet

Summary

- We summarize our proposed model as **multi-perspective architecture** in social media search models because it
  - adopts attention matrix based NN architecture, which is widely applied and proved effective in text matching tasks
  - understands short but informative text by ConvNet
  - takes advantage of the URL and other OOV information by character-level modeling

Summary

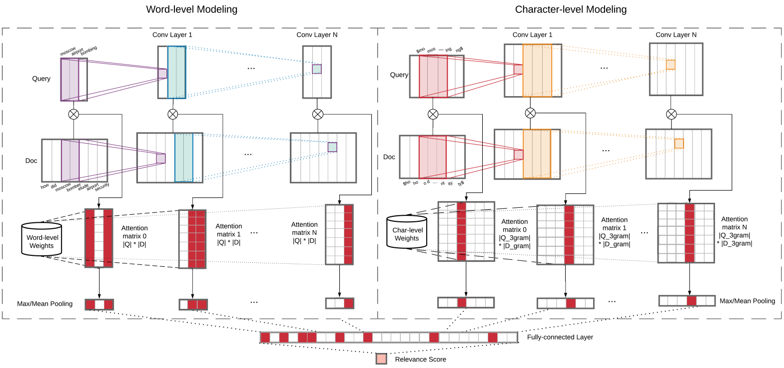
- We summarize our proposed model as **multi-perspective architecture** in social media search models because it
  - adopts attention matrix based NN architecture, which is widely applied and proved effective in text matching tasks
  - understands short but informative text by ConvNet
  - takes advantage of the URL and other OOV information by character-level modeling
- Our proposed model:
  - requires much less training data to obtain a well-performed model compared to the previous NN models

Summary

- We summarize our proposed model as **multi-perspective architecture** in social media search models because it
  - adopts attention matrix based NN architecture, which is widely applied and proved effective in text matching tasks
  - understands short but informative text by ConvNet
  - takes advantage of the URL and other OOV information by character-level modeling
- Our proposed model:
  - requires much less training data to obtain a well-performed model compared to the previous NN models
  - is the first work to build an effective NN model in social media domain without query feedback or external data.

Q & A

- Thank You



Word-level Modeling
Character-level Modeling
Conv Layer 1
Conv Layer N
Query
Doc
Attention matrix 0 |Q| * |D|
Attention matrix 1 |Q| * |D|
Attention matrix N |Q| * |D|
Word-level Weights
Char-level Weights
Attention matrix 0 |Q_3gram| * |D_gram|
Attention matrix 1 |Q_3gram| * |D_gram|
Attention matrix N |Q_3gram| * |D_3gram|
Max/Mean Pooling
Fully-connected Layer
Relevance Score